DATA WRANGLING AND MASSAGING

USING ETL TOOL

by

Kishen Sivakumar

16000775

Dissertation submitted in partial fulfilment of

The requirement for the

Bachelor of Information Technology (Hons)

SEPTEMBER 2021

Universiti Teknologi PETRONAS,
32610 Seri Iskandar,
Perak Darul Ridzuan.

CERTIFICATION OF APPROVAL

Data Wrangling and Massaging using ETL Tools

By

Kishen Sivakumar

16000775

A project dissertation submitted to the
Information Technology Programme
Universiti Teknologi PETRONAS
In partial fulfillment of the requirement for the
Bachelor of Information Technology (Hons)

Approved by,

_____

(Dr Helmi B M Rais)

UNIVERSITI TEKNOLOGI PETRONAS

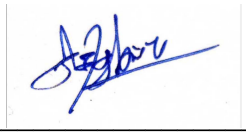
32610, SERI ISKANDAR,

PERAK DARUL RIDZUAN

SEPTEMBER 2021

CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in the references and acknowledgements, and that the original work contained herein have not been undertaken or done by unspecified sources or persons.



KISHEN SIVAKUMAR

ABSTRACT

It appears that the accounting and finance profession is progressing at a quick pace, in the same way that the rest of the commercial world is. In most finance teams, the daily duties, transactional tasks, and record keeping no longer account for the vast bulk of the time spent. A larger focus than ever before is being placed on financial planning and analysis in accounting and finance roles, which is an exciting development for both professionals and employers alike. Incorporating the capabilities of current technology to reduce audit firm workloads, such as leveraging KNIME Analytics with the application of linear regression to create an autonomous audit process and Power BI for futuristic data visualization, are examples of how to improve efficiency. As a result, the project's goal is to transform unstructured data into the appropriate conclusion and represent it in accordance with the requirements of the target user, a Malaysian finance analyst working for a small or medium-sized business or even a large multinational corporation. Accounting businesses can benefit from the application of data transformation and data aggregation to free up time and bandwidth for accountants to pursue more exciting work and create value and spend less time on the low-value jobs that must be performed before the high-value activity can begin. Data visualization may be beneficial in that it presents information in the most effective manner possible. By making the data more natural to examine for accountants, this makes it easier to discover trends, patterns, and outliers in massive data sets than it would otherwise be. For audit firms, the use of linear regression in machine learning will be the most effective method of forecasting the future outcome of the audit. Due to the fact that machine learning technology for auditing is still in its early stages of development, machine learning systems are being developed for accountants who work for larger CPA firms, and smaller businesses should begin to benefit as the feasibility of the technology improves and auditing standards are adjusted.

ACKNOWLEDGMENT

In the first place, I'd like to use this time to offer my heartfelt gratitude to every single person who has supported me during this process. I would want to express my gratitude to my parents for their unwavering support and blessings, which have enabled me to reach this point in my life. Next, I'd want to express my gratitude to my university, Universiti Teknologi PETRONAS (UTP), as well as the professors who have educated and directed me throughout these past three years with essential knowledge.

In addition, I would like to express my gratitude to Dr Helmi B M Rais, my supervisor, for accommodating my request to work on a project including data wrangling and massaging using ETL tools under his supervision. He had been a kind and understanding supervisor who, despite his hectic schedule, had always made time for me to discuss my concerns. I am eternally grateful to him for the countless hours and efforts he put in to guiding me and ensuring that I was completely aware of the tasks that I was performing during this project's duration. I am also grateful to him for providing me with honest evaluations, feedback, and recommendations throughout each of our meetings.

My Final Year Project would also have been impossible to complete without the assistance of my Co-Supervisor, Madam Shakirah Binti Mohd Taib, who provided me with invaluable guidance throughout the duration of my project and generously donated her valuable time to evaluate my project.

Thank you.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION AND SYMBOLS

FYP 1	: Final Year Project 1
FYP 2	: Final Year Project 2
Power BI	: Power Business Intelligence
ETL	: Extract, Transform and Load
CSV	: Comma-Separated Values
API	: Application Program Interface
PDF	: Portable Document Format
AI	: Artificial Intelligence
CPA	: Certified Public Accountant

CHAPTER 1

INTRODUCTION

1.1 Background of study

Financial analysts oversee a company's or corporation's financial planning, analysis, and projections. They anticipate future income and expenditures in order to develop cost structures and project capital budgets. As technology lessens the necessity for some traditional accounting abilities like bookkeeping, transaction processing, and maintaining extensive records, other skills like analysis, forecasting, and financial strategy are likely to become more important over time. These skills must be learned by professionals who want to be competitive in the future employment market. Some of the tasks previously handled by finance analyst, such as gathering transactions and applying them to prepare tax returns and financial statements, are gradually becoming automated and AI capable.

Is it possible that finance analyst loses their job as a result of this? While technology can take over some accounting activities, humans will still be required to double-check the work done by automated technology. Finance professionals who can adapt to these new technologies will continue to be in high demand in this environment, as they will be the ones assisting firms with the implementation and usage of technology. However, most of them find these new changes and adaptations difficult because learning programming or adjusting to automated technologies might be difficult.

As a result, I devised a project entitled "Data Wrangling and Massaging Using ETL Tools," for my Final Year Project which will be built specifically for finance analysts to help them with their accounting, data collection & data wrangling chores and data visualization for clear understanding. Following extensive research and consultations with a few financial analysts, I decided to use multiple ETL (Extract, Transform & Load) tools which are KNIME Analytic, Power BI and Linear Regression in Machine Learning in KNIME will be heavily implemented in this Final Year Project. KNIME allows finance analyst to visually create data flows (or pipelines), selectively execute some or all analysis steps, implementing Linear Regression and later inspect the results, models, using interactive widgets and views. This ETL Tool helps to create pipelines for CSV data editing such as merging multiple CSV, deleting data, pivoting data, and many more. Power BI will be implemented for data visualization and data collection from the KNIME automated workflow.

Since day one, I researched and making improvement on this project from the planning phase until implementing it. Throughout the time, I followed Gantt chart and prepared milestones to work on this project in organized manner. I distributed automation tasks evenly and collected all the requirements needed for the development. I even set few appointments with few finances analyst to understand and note their requirement and insights for this project. I encountered multiple challenges and issues while developing this project. This project was very insightful, and I gained knowledge and abilities up to date with the most recent innovations in the finance and accounting industry.

1.2 Problem Statement

Finance analysts have had difficulty completing their daily stacked tasks manually or in the traditional manner, which entails many working hours, human errors, and task complexity. Financial analysts work an average of over forty hours each week, with the majority working between fifty and seventy hours. Many newcomers to the area may need to devote more time to studying for their professional and licensing exams. Through literature review, I need to comprehend

the day-to-day tasks and challenges that finance analysts confront, and I need to write down many factors based on their causes.

1.2.1 Time Consumption

Finance analysts have had to execute their routines manually until now, such as aggregating transactions and using them to prepare tax returns and documenting financial statements, because they do not have an automated method in the first place. To complete their given jobs by the deadline, they had to remain in the same place for lengthy periods of time, doing all the documentation, data collection, and data updates internally, resulting in them working about seventy hours each week. This also consumes up all their free time, leaving them physically and mentally exhausted.

1.2.2 Risk of human error

Because it is performed manually, the traditional function is prone to errors. As a result, there have been numerous human errors, necessitating more time to correct them. This never-ending loop has resulted in the squandering of several working hours that could have been better spent on other projects. As a result, this new automation invention aims to make everyone's job easier, save time from unneeded hassle, and eliminate common errors.

1.2.3 Data Transformation

Financial Analyst’s daily tasks involves large number of finance data such as assets, liabilities, equity, income, expenses, and cash flow. One of the issues is that it requires a huge amount of data, which necessitates many working hours to complete each tedious job manually. Assets are what the company owns, liabilities are what the company owes, and equity is what is left for the owners of the company after the value of the liabilities are subtracted from the value of the assets. As a result, in order to update the internal organization system, all different types of data must be converted to the same value and compiled in one CSV.

1.2.4 Data Wrangling

Finance analysts who work with data spend nearly 80% of their time wrangling data, leaving only 20% of their time for exploration and modelling. Once the code and infrastructure base are in place, the data handling process will offer immediate outcomes for as long as the process is relevant. Finance analysts, on the other hand, tend to bypass crucial data wrangling procedures, resulting in big downfalls, missed opportunities, and erroneous models that harm the organization's analysis reputation. Data wrangling software has evolved into a critical component of data processing. The essential value of using data wrangling tools, which may be summarized as making raw data usable, is often overlooked by finance analysts. Most finance analysts are unable to comprehend the appropriately wrangled data, ensuring that only high-quality data is entered into downstream analysis. They also fail to recognize the need of consolidating data from numerous sources into a single spot where it can be utilized. Due to a lack of experience and understanding, Finance analysts fail to deploy automated data integration tools as data wrangling procedures that clean and convert source data into a standard format that can be utilized repeatedly according to end requirements. They fail to remove noise or flawed, missing parts from their data. As a result, data wrangling serves as a preliminary step in the data mining process, which entails obtaining and analyzing data. As a result, understanding the principles of data wrangling and how to use data wrangling tools is critical for finance analysts.

1.3 OBJECTIVE

The primary goal of this research for my final year project is to provide the best automated data transformation for finance analysts to achieve the desired results. The aims and objectives of this tasks are listed below:

- To utilize ETL tools for optimum Data Wrangling and Linear Regression in audit task

  Using the KNIME Analytic platform, all unstructured data type must be transformed to structured data type and compiled in global spreadsheet with the implementation of Linear Regression.

- To develop forecasting in Data Visualization

  Embed rich personalized dashboard and implement machine learning to extract a greater number of future insights from visualized data.

- To Perform User Acceptance Testing on developed KNIME and Power BI dashboard

  Provide best chance to find and fix broken functionality and usability issues. Quality assurance testing and ensure goals at the task and requirement level helps testers to notice and observe a lot more, and even do things that aren't in the developer's scope.

1.4 SCOPE OF STUDY

I discovered a few critical concepts that allowed me to meet the requirements and design the project based on my findings. By utilizing ETL tools such as KNIME Analytic and Power BI, which are the most relevant tools I discovered via my research, the purpose of this project is to bridge the gap between data transformation and data analysis. Following the conversion of raw data to structured data using Linear Regression, the database will be exported to Power BI for automatic dashboard. Moreover, the results of my research into Linear Regression from machine learning provided me with ideas into how to employ that technique in KNIME Analytic software to automate data collection and visualization, as well as to

estimate sales data. Because of this, the purpose of this project was to translate raw data into the desired conclusion and visualize it according to the demands of a Malaysian finance analyst who works for any small or medium-sized business or even for a major organization.

Thus, my final year project would help finance analysts with their accounting, data collection, and data manipulation tasks, I've decided to design it exclusively for them. This Final Year Project will use numerous ETL (Extract, Transform, and Load) tools, including KNIME Analytic, Power BI, and Linear Regression in Machine Learning in KNIME, following thorough research and conversations with a few financial analysts. Data flows (or pipelines) can be created, executed, and inspected using interactive widgets and views in KNIME. Linear Regression can be implemented, and the results can be inspected using interactive widgets and views. With the help of this ETL tool, it can build pipelines for CSV data editing tasks including merging, removing, and pivoting. Data from the KNIME automated workflow will be imported into Power BI for visualization and wrangling.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

It is not uncommon for finance analysts to overlook the use of automated data integration such like data wrangling techniques that clean and transform source data into a standard format that can be used repeatedly according to the end requirements. How technology is changing the future of accounting and the challenges faced by finance analyst will be discussed in this chapter. Since many of the previously manual, routine-based accounting duties are now handled by non-human help. Thus, the most used and recommended software platforms also will be discussed to support my research.

2.2 Modern Technology in Accounting

The days of the financial employee crunching CSV and carrying a business calculator are long gone. As technology becomes a more integral part of our daily lives, companies in areas ranging from consumer electronics to financial services are looking for more tech-savvy finance workers. finance analysts are increasingly immersing themselves in higher-level work as a result of technological progress, taking on major decision-making and managerial duties within organizations. Disruptive technologies will continue to have an impact on the finance department. Finance professionals can use advanced technologies like robotic process automation (RPA), artificial intelligence (AI), and block-chain to replace labor-intensive financial reporting chores with more strategic high-value-add efforts.

Monthly closure operations, basic budgeting, and standard reporting are just a few of the normal and simple finance activities that will inevitably be automated. RPA and sophisticated AI technologies, for example, can automate human accounting operations, handle transactional data from many finance systems, and reduce the need for critical standardized financial reports to be created manually. As a result, we anticipate that future finance professionals will devote less time to regular audits and reconciliations and more time to advanced analytic and partner management. While technology advancements have rendered some jobs obsolete, accounting's essential principles remain the same, and the role is as important as it has ever been. Finance analyst have always offered information to stakeholders, primarily management, but also shareholders and other interested parties. This has not changed, but the manner in which this data is collected and wrangled has.

2.3 Digital Transformation in Modern Accounting

According to an IMA research, one-third of accounting teams spend between 51 percent and 75 percent of their time on low-value, repetitive duties. Furthermore, 56 percent of the finance analyst polled claimed they require automation just to keep up with their growing workloads. This means that, despite how far technology has progressed, finance analyst is still notoriously sluggish to adapt. We're still operating at a snail's pace, particularly when it comes to financial closure processes. Organizations are now searching for an entirely different skill set because of the emergence of purpose-built technology for Accounting and Finance. CSV and spreadsheets are still my favorite tools as a finance analyst. However, according to Accounting Today, CFOs no longer consider CSV to be a necessary competency for new hires. Instead, they're searching for people who can adapt to new technologies, are knowledgeable with KPIs and modelling, and can work well with others. Businesses operate around the clock, seven days a week, 365 days a year. To deal with all of the data that is continuously pouring in and present it in a meaningful way, a fresh mentality and strategy are required.

Finance analysts are now required to have a better awareness of the technologies available, which can assist the organization in becoming significantly more effective in their close operations. Their position will become increasingly

cross-functional, requiring better reporting and data-driven decisions, as well as the use of rules-based automation to automate the most time-consuming manual operations. Nowadays, the introduction of APIs, or application programming interfaces, has made accounting automation easier. APIs allow other pieces of software to interact with the accounting system. Connecting two systems before APIs generally necessitated coding. it may now develop automated workflows quickly and easily by largely plugging and playing. For example, APIs are used by apps like Expensify and Receipt Bank to feed data into the general ledger from PDFs and images of invoices and receipts. APIs also enable bank feeds, which allow transactions to be downloaded directly into an accounting system. Accounting businesses utilize automation to extract data from tax papers and feed it into tax returns immediately. To construct the journal entries and footnotes required by accounting rules, finance professionals employ automation to extract data from contracts and leases. Automation allows business owners to keep an eye on their needs in real time. Many accounting tasks can be automated according to the graphic from McKinsey as below.

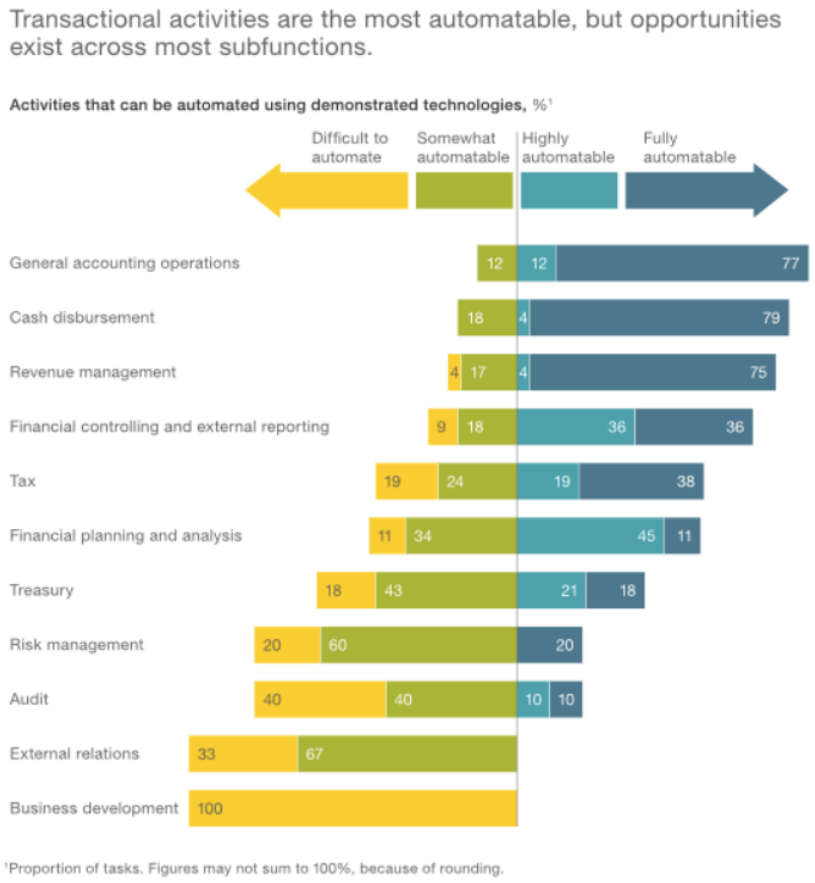


Figure 1 Automating Accounting Operations

Setting up rules in the accounting software to classify transactions that come in via bank feeds is a simple example of automation. Adding software tools to perform specific accounting chores is the next step. Order to cash software, for example, can take orders from a customer interface and route them to the appropriate locations for fulfillment and collecting payment. Automation in accounting frees up time and bandwidth for finance analyst to pursue more exciting work and add value. They will devote less time to the low-value job that must be completed before the high-value activity can begin.

2.4 Data Aggregation and Automation using KNIME

After grasped the significance of any advancements that may occur in accounting firm, my main goal during the early stages of my final year project research was to discover the ideal alternative automation platform that would be easy to use for audit firms. KNIME Analytic Platform is the free, open-source software for creating data science and allows users to visually create data

flows or pipelines, selectively execute some or all analysis steps, and later inspect the results, models, using interactive widgets and views. How is it a solution for accounting firm in this digital world? As an instance, each month, travel and event expenses are collected, summarized, and analyzed in order to match costs with successful sales in an organization. Previously, the accounting manager in that organization gathered CSV files from every employee and copied and pasted these in to a global CSV file. When the repository of these CSV files was moved to a shared folder.

Throughout my research I found that above issue can be solved using KIME. Thus, I discovered that the data wrangling abilities of KNIME could be used to create the summary and, with the Table to PDF node, create a very simple expenses report. Utilizing KNIME Analytic Platform to collect and summarize employee expenses. Nodes like CSV Reader, CSV Writer, Table to PDF, and Copy/Move make the work of an accounting manager easier and less prone to errors. The workflow, deployed via KNIME Server, is available as an analytical application in the KNIME Web Portal. It provides a simple interface for users to enter expenses and the executed workflow automatically generates the required reports. A previously defined CSV template simplifies accounting file filling by each employee. Data Analysis team of that organization was able to create KNIME workflow that can be deployed as an analytical application in the KNIME Web Portal.

Therefore, later Accounting Manager selects the starting folder, which contains different sub-folders (one for each month) with the CSV files of each employee and the summary CSV files generated from the previously executed workflow. The most recent summary CSV file is identified and split from the other files. Individual CSV files are renamed to avoid future reprocessing. After concatenating all the individual accounting files, the workflow appends the resulting table to the split summary CSV file and writes a new summary CSV file with the process date in the name. Lastly, a simple PDF report is sent to the HR manager with a list of all the reimbursements. As a result, the global accounting CSV file is updated within a few seconds every time the workflow runs which is once a month. At the same time, a summary of the total reimbursements is sent to the HR manager and Managing Director. This results in time savings from a couple of hours to a

couple of minutes, as well as repetitive operations and avoids mismatches due to copy and paste.

Finance and Business Analysts are expected to provide data-based insights and guidance for where the company is heading. Unfortunately, they often don't have the time for analyzing and making sense of the data. Instead, they are busy getting data in shape that are fragmented, inconsistent, and from many sources. KNIME provides solutions for FP&A teams to repeatedly aggregate data for reports, insights, and audit. The extremely flexible, no-code, visual workflows allow safe access to the right level of complexity. KNIME allows finance analysts get a no-code software that automates their repetitive financial analytic tasks and makes data more reliable and more insightful – much more quickly. It allows Finance Managers see their teams becoming significantly more productive, motivated, and enabled to generate reports faster and get deeper insights and better forecasts. Nonetheless, KNIME also supports Shared services such as Centers of Excellence, IT, and Business Consultants are able to provide packaged data solutions that Finance Teams can use without further technical or expert support.

KNIME Analytic Platform makes it easy to seamlessly blend core functionality for data manipulation with highly complex and bespoke operations for data collection and cleaning using Linear Regression. The visual workflow builder enables workflow creators to explain the process to the media team and others outside data specialist roles who don’t necessarily have the technical data science knowledge, drawing parallels to how data was treated in the manual process. This means the solution is less of a “black box”, but rather something that those who use the data can have confidence in because they understand how the parts fit together. (Rob Blanford, 2019)

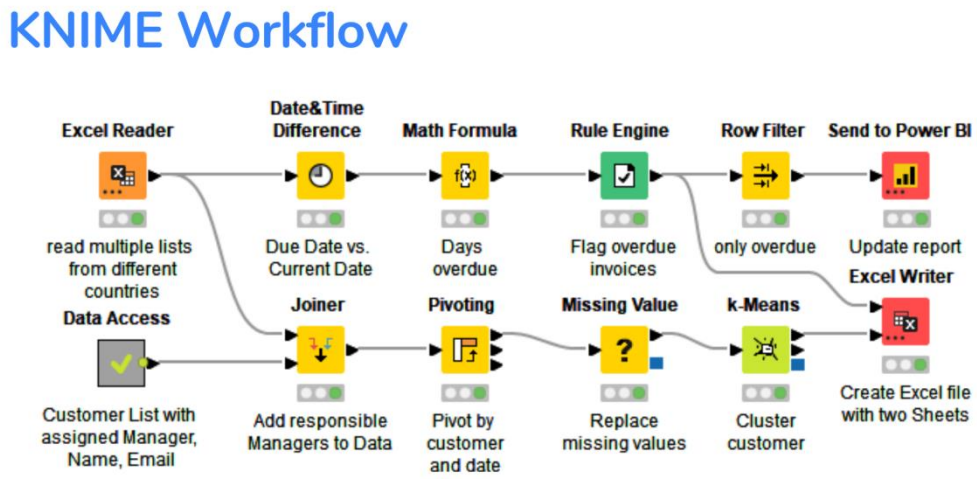


Figure 2 KNIME Analytic Example Workflow

2.5 Combining Power of Power BI

Data visualization is an important phase in the business intelligence process because it takes raw data, models it, and delivers it so that conclusions may be drawn. Machine learning algorithms are being developed by data scientists in advanced analytic to better assemble critical data into visualizations that are easier to grasp and analyze. Data visualization, in particular, employs visual data to express information in a universal, quick, and effective manner. This method can assist businesses in determining which areas require improvement, which factors influence customer satisfaction and dissatisfaction, and what to do with certain items. Stakeholders, business owners, and decision-makers can better estimate sales volumes and future growth using visualized data.

Internal audit functions are no exception to the trend of using data to drive innovation and transformation. The current financial crisis has put a greater emphasis on data analysis and visualization, and we've grown accustomed to using intelligent dashboards to guide our own decision-making. Internal audit must follow suit, leveraging the potential of data analytic to bring more value and insight to the organization in a more efficient manner. With so much data being acquired through data analysis in today's corporate environment, we need a way to visualize that data

so we can understand it. Therefore, I decided to bring in Microsoft Power BI as visualization tool since most of the organization utilizing it. Microsoft Power BI is used to find insights within an organization's data. Power BI can help connect disparate data sets, transform, and clean the data into a data model and create charts or graphs to provide visuals of the data. By placing data in a visual context, such as maps or graphs, Power BI helps us understand what it means. This makes the data more natural to interpret for the human mind, making it easier to see trends, patterns, and outliers in vast data sets. By conveying data in the most effective manner possible, Power BI may assist.

Finance analysts are frequently responsible with monitoring their company's performance. To get insight into the tendencies of the firm, its industry, and even its competitors, this approach frequently necessitates consulting several sources. Critical information about an organization's pulse can be summarized in a customized format in one or more visualizations. To find the better and efficient tool for my final year project and to prioritize finance professional’s concern, I made comparison between CSV which is widely used till date for traditional tasks.

Because the two Microsoft tools complement each other, Power BI and CSV operate nicely together. CSV is an excellent data source for Power BI, which can receive data from a wide range of sources. Power BI is an excellent solution for a wide range of visualization possibilities, higher-level analytic, automated updates when source data changes, very huge datasets, and user involvement, among other things. While CSV is mostly used for basic analytic, Power BI allows users to construct dashboards and reports with at-a-glance views based on data kept on-premises or in the cloud, as detailed in the aforementioned "CSV vs. Tableau" article. The software provides the user with a "canvas" as well as other visualization tools for displaying data to users. Data from millions of entries can be quickly summarized using visualizations. Power BI provides a shorter learning curve than CSV for new users. The synergistic combination of data used to make sound business decisions is called business intelligence. Certified Public Accounting Firm have found business intelligence useful for multiple purposes such as budget analysis and reporting, key performance indicator analysis and reporting, predictive analytic in auditing,

forecasting, and reporting, analysis of audit samples and tests and many more audit tasks.

Internal audit departments frequently use text-heavy reporting, such as word processing documents and PowerPoint presentations. Very few would employ visualization, and what we're seeing is a significant expectation of increased use of visualization to report and connect with stakeholders. to stakeholders considerably more dynamic. Internal audit chiefs and departments employ visualization approaches to communicate with auditees and stakeholders. The use of visualization makes discussions with stakeholders and reporting to stakeholders considerably more dynamic. Internal audit chiefs and departments employ visualization approaches to communicate with auditees and stakeholders. It enables extra insights to be derived from internal audit's reports, and it enables internal audit to increase its impact and influence. (Hatherell, 2016)

Figure 3 Power BI Example Dashboard

2.6 Machine Learning in KNIME

Machine learning is a subset of artificial intelligence (AI) that was born out of the idea that machines might be trained to learn in the same way that humans do. While humans are just now beginning to appreciate machine learning's dynamic potential, the notion has been around for decades. Machine learning has become a crucial part of modern life because of the explosion of data, especially due to the rise of the Internet and advances in computer processing speed and data storage. Machine learning can be found in e-mail spam filters and credit monitoring software, as well

as in technology companies' news feeds and targeted advertising features such as Facebook and Google. Machine learning technology for auditing is still in its early stages of development. Machine learning systems are being developed by several of the larger CPA companies, and smaller businesses should start to profit as the feasibility of the technology improves, auditing standards adjust, and teaching program expand. (Julia Kokina and Thomas H. Davenport, 2017)

Moreover, machine learning algorithm in data visualization tool, Power BI could allow CPA firms to detect patterns that currently might otherwise go unnoticed. Because of the inherent limitations of machine pattern-finding, auditors will continue to need an understanding of the individual business and its industry, as well as the external business environment and societal forces. For example, user accounts might be the best predictor of revenues for companies and therefore should be given the appropriate weighting in the internal algorithm. Without judgment as to what to specifically look for, the authenticity of accounts and the presence of “bots” may not be detectable by machines and could lead auditors to reach incorrect conclusions.

Auditors will need to understand and validate the completeness and accuracy of the input data in order to reach an appropriate conclusion on the output. Furthermore, there will always be potential blind spots when evaluating empirical evidence; therefore, an auditor’s intuition will likely continue to be an important source of knowledge. Thus, my research is solely for the finance analyst in any firms to request for the help in automation to the amount of workload they have in everyday life. The implementation of machine learning in Power BI improves the perceptibility of a client's data, allowing them to gradually discover business knowledge. It provides a wide range of pre-built representations, allows to customize existing ones, and allows to explore an ever-growing list of in-built perceptions in the network display. (Julia Kokina and Thomas H. Davenport, 2017)

Therefore, based on thorough research, application of linear regression in KNIME will be ideal way of predicting the future outcome for the audit firms. Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between

variables and forecasting. Different regression models differ based on the kind of relationship between dependent and independent variables, they are considering, and the number of independent variables being used. The database which will be automated in KNIME and visualized in Power BI will be used as dependent and independent variables to sales forecast prediction modeling. The reason for choosing linear regression in the research is because these models can be trained easily and efficiently even on systems with relatively low computational power when compared to other complex algorithms. Linear regression has a considerably lower time complexity when compared to some of the other machine learning algorithms. The mathematical equations of Linear regression are also fairly easy to understand and interpret which also will be perfectly applicable in an audit firm tasks. The graphic below shows hypothesis function of linear regression which is Y= a + bX , where Y is the dependent variable (that's the variable that goes on the Y axis), X is the independent variable (i.e. it is plotted on the X axis), b is the slope of the line and a is the y-intercept.

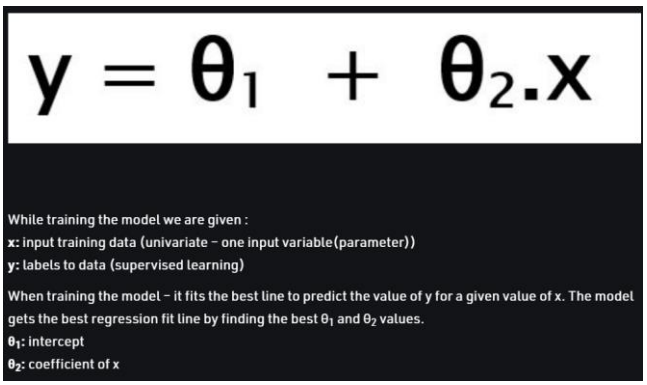


Figure 4 Linear Regression Formula

CHAPTHER 3

METHODOLOGY

3.1 Introduction

The project's major functionalities are discussed in this section of the report. Based on my research on this project, agile methodology is decided to be implemented to fulfill both functional and non-functional requirements. The main objective of this project is to develop a complete automation process for finance analyst with their everyday audit tasks. It involves multiple discussions and meet-ups to make with finance analyst to understand and collect respective data to develop a process based on their requirements. Thus, agile methodology will the best fit to follow up with the progress of this project. This is because, agile software development will be a collection of iterative software development approaches in which project requirements and solutions emerge from cooperation among self-organizing cross-functional teams. Agile methods promote a disciplined project management process that encourages frequent inspection and adaptation, a leadership philosophy that encourages teamwork, self-organization, and accountability, a set of engineering best practices that allow for rapid delivery of high-quality software, and a business approach that aligns development with finance analyst’s needs and audit firm's goal.

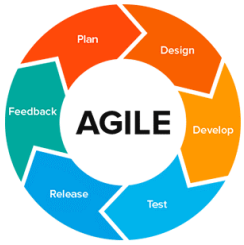


Figure 5 Agile Methodology

3.2 Agile Methodology

The Agile methodology will help to collects the value adaptability and flexibility in this project. This also helps and aims to provide better responsiveness to changing project needs and therefore focuses on enabling schedule to deliver in workable increments for the finance analysts too. Agile consists of six main elements in a proper order to obey. The elements are listed as below.

3.2.1 Phase 1: Planning and gathering data/information

The goal of this project is to perform the daily tasks as an automation engineer(myself), using all the available tools, skills, and time to create a feasible solution ready for KNIME automation, Power BI dashboard and linear regression algorithm in KNIME. The preparation for the entire execution is essential because of the large amount of data and tedious accounting processes. The finance analyst or financial analyst and I will need to have a thorough discussion about their audit tasks and the difficulties they are encountering with standard auditing approaches. At the discussion, we shared respective insights on how we going to improve the execution of automation.

In order to ensure a smoother implementation of the project at its first stage, the finance analyst and I discussed the data source that would be used for the automated processing. During the meeting, finance analyst presented two independent e-commerce sales data files in CSV format for consideration. As noted by the finance analyst, they will need to perform data wrangling, data cleaning, and data visualization in order to update those CSV files in order to fulfil their auditing responsibilities. They must take immediate action in order to update the information in their system and communicate it to their upper management team. These data represent sales of their products sold in the years 2015, 2016, 2017, and 2018 respectively. All four years of sales, which include Furniture, Electronics, and Accessories sales, as well as the quantity ordered and client ID, are included in this report. The second CSV file contains the status of the product sold, as well as the product code and the customer ID. It is expected that finance analysts will integrate both CSV files into a single CSV file and use data wrangling to edit or eliminate

noise from their data, as well as faulty or missing elements from their data. Also expected are VLOOKUP, renaming of columns, data merging, row filtering, column filtering, and other necessary editing of the CSV files, in addition to other required editing.

Throughout the project discussion with them, I was able to comprehend the data wrangling they needed to do in these CSV files in order to fulfil their other tasks moving forward. As a result, the finance analyst provided the necessary information and data that was used to automate the process of configuring these data sources as the input data for the project later on. Because obtaining the relevant information is necessary in order to meet the needs of the development phase later on, it is important to do it now in planning phase.

3.2.1.1 Implementation of Software Requirements Engineering

Software requirements was implemented to identify the functionality that is needed by a system in order to satisfy the finance analyst's requirements. Enhancing the clarity of project requirements can dramatically improve the outcome of the created automation process. It can prevent project failure, uncover latent errors early enough and reduce miscommunication throughout the discussion.

Table 1 System Context Object Table

No	System Context Object	Rationales	System context facet	Properties
1	CSV	The data of CSV will be stored and analyzed	Subject facet	Raw Data/Data
2	Nodes	its inputs and outputs, and which parameters it has and determined by the independent variable	Usage facet	Variables
3	Database	Database is used to store data and its analysis.	Usage facet	CSV, KNIME project, Software
4	Internet	Internet is fundamental to access,	IT System facet	Internet

		develop and retrieve data to process and analyze		connection
5	KNIME Analytic	KNIME Analytic Platform is the open source software for creating data science	IT System facet	Nodes, database, internet, interface
6	Finance Analyst	Finance analyst who are facing issue and will perform the system once development completed	Subject facet	Required data, information to start the project
7	KNIME Server	KNIME Server is the enterprise software for team-based collaboration, automation, management, and deployment of data science workflows as analytical applications and services	IT System facet	Internet connection
8	Power BI	Using the Microsoft Power BI programme, employees or coworkers can build a clear visual representation of the studied data that is easy to interpret for sales forecasting purposes.	Usage facet	Data Visualization

Table 2 Development Context Object

No	System Context Object	Rationales
1	KNIME Analytic Platform	KNIME Analytic Platform is the open source software for creating data science
2	KNIME Server	KNIME Server is the enterprise software for team-based collaboration, automation, management, and deployment of data science workflows as analytical applications and services
3	Microsoft CSV	Microsoft CSV features calculation, graphing tools, pivot tables, and a macro

		programming language called Visual Basic for Applications
4	Power BI	Using the Microsoft Power BI programme, employees or coworkers can build a clear visual representation of the studied data that is easy to interpret for sales forecasting purposes.

Table 3 Requirement Sources & Elicitation Technique

No	System Context Object	Requirement source	Suitable elicitation technique
1	CSV	Finance Analyst	Observation System specific document
2	Nodes	KNIME	Observation
3	Database	KNIME Local Database	Perspective-based reading
4	Internet	Article/Journal	Perspective-based reading
5	KNIME	KNIME software	Predecessor systems
6	Finance Analyst	Organization/corporate	Interview & Organization specific document
7	KNIME Server	KNIME	Predecessor systems
8	Sales forecasting	Power BI	Predecessor system

The opinion of the finance analyst can assist in bridging the gap between what is expected and what is delivered. It is possible that their advice will assist me in decreasing development time and project costs by providing useful insight into technical areas of the requirements and so minimizing the amount of rework that will

be required for the project. Putting together the necessary information and putting it into practise software requirement engineering assisted me in developing a high-quality automation project that resulted in a significantly better outcome.

3.2.2 Phase 2: Design

This is where all knowledge and information have carefully analyzed, structured, and designed on the basis of what one has, would have, and all their hands can get. Based on the knowledge gained and data that has been collected about the audit tasks, I will need to design automation process based on how to execute it based on desired outcome. By reviewing KNIME online forums, as well as learning from other configurators, I researched the statements and priorities of the problem to meet most, if not all, of the client's expectations. An automated workflow on KNIME will be built and designed following the collection and analysis of insights and information during the planning phase. This will be accomplished by identifying relevant KNIME function nodes that will match the criteria and incorporating them into the workflow. It is possible to locate thousands and hundreds of KNIME function nodes in the programme, but it takes time and effort to investigate and properly analyse each function node in order to find the most appropriate function nodes to use in my automated workflow. Based on the study I conducted, I discovered a plethora of relevant function nodes that would meet the needs of my automation workflow.

Prior to anything else, in order to upload and configure both CSV files, I needed to find an appropriate function node, which turned out to be CSV Reader, and that is exactly what I did. As a result of the path flow variable being supported by this node, it is able to read CSV files, in particular. However, if I notice that this node is unable to read the CSV file, there is another node titled File Reader node that I might use instead of this one. It provides a higher number of options for reading difficult files than previous versions. Integrating the File reader node into my project, on the other hand, will result in issues and failures while reading CSV files. Using the CSV Reader node in the workflow, which may be used in a server or batch environment and if the input files' structure varies between invocations, this node will read a CSV file and show the contents of the file. Among other things, it has a

variable number of input columns, among other characteristics. The node will scan the input file in order to determine the number and types of columns, and then output a table with the structure decided by the auto-guessing algorithm.

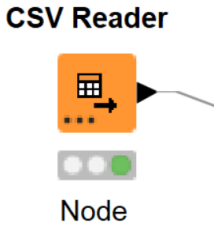


Figure 6 KNIME CSV Reader Node

Second, for the process section of the workflow, I needed to locate a suitable node for merging the two CSV files, which is where I discovered the Joiner node. This node combines two tables in a manner similar to that of a database. The join is based on the joining of columns, tables, or data in the aggregate form of the join. It is possible to handle this issue in a number of ways if data from the first CSV Reader node cannot be connected with a row from the bottom CSV Reader and vice versa. Only rows that fulfil the criteria of the Inner Join will appear in the result table. If there is no matching row in the bottom CSV Reader, a Left Outer Join will fill in the blanks in the columns of data that came from the bottom table with values from the top table. Similarly, if there is no matching row in the top CSV Reader, a Right Outer Join will fill in the blanks in the columns of the top CSV Reader with values from the missing rows. A Full Outer Join will fill in the blanks in the columns of both the CSV Reader nodes and the bottom table with values from the missing rows. To meet the need of merging two distinct CSV files, I chose the Joiner node to be implemented in my automation workflow, which I found to be the most appropriate.

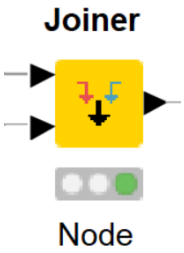


Figure 7 KNIME Joiner Node

Following that, I looked into finding an appropriate KNIME function node to rename column title since a finance analyst had requested that the column name of QTR ID be changed to Customer ID in order to prevent confusion when viewing the information. As a result, I chose the Column Rename node. Column Rename is a node that allows to rename column names or change the types of columns. This node allows to alter the names of individual columns by editing the text field, or can change the kind of column by selecting one of the available types from the combo box on the right. Compatible types are those into which the cells in a column can be securely cast or changed without causing damage to the cells in the column. A configuration with a red border indicates that the column for which the configuration was created no longer exists. As a result, I decided on the Column Rename node to be included in my KNIME process.

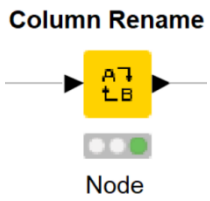


Figure 8 KNIME Column Rename Node

Besides that, I continued my investigation to discover an appropriate KNIME function node to duplicate row filter, which aids in the removal of duplicated data from CSV files using the KNIME tool. This is due to the fact that duplicated data in a CSV file must be cleaned by deleting duplicate data, which later on aids in the proper implementation of Linear Regression and the avoidance of unexpected outcomes when applying linear regression. So the Duplicate Row Filter node is responsible for identifying duplicate rows that have the same values in specific columns as the original row. For each group of duplicates, the node selects a single row from the table. In the input table, this node has two options: it can either eliminate all duplicate rows and keep only the rows that are unique and chosen, or it can annotate the rows with additional information about their duplication status.

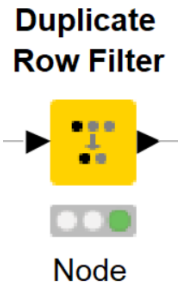


Figure 9 KNIME Duplicate Row Filter Node

It is necessary for me to be able to receive an acceptable overview of the data that has been configured from the preceding function nodes after the Duplicate Row Filter node. To check the data, it was decided to use the Data Explorer node because it offers a variety of options for presenting aspects of the input data in an interactive manner. By checking the boxes next to the numbers it want to see, it can display the most common and least frequent values. The nominal tab of the interactive view will generate two columns for each given number n: one column containing the n most common nominal values and another column containing the n most infrequent nominal values, respectively. If the selected number is equal to the total number of different values in a column, only one column with all values will be created in that column. Unless otherwise mentioned, all data are presented in decreasing order of frequency, unless otherwise noted.

The frequency distribution also includes the number of the most frequent and least frequent frequencies. For any number n that is supplied, the nominal tab of the interactive view will list the n most frequent nominal values in one column and the n most infrequent nominal values in another column. It also displays the median, which allows to calculate the mean of the numerical values in the input data set by comparing them to the median. This tab also contains information about the Row ID, which will result in the development of a column with the Row ID of each value in the data that is displayed on this page. As an added feature, it highlights missing values in histograms with an additional bar to make them more noticeable. Automatically adjusting the number of histogram bars for each numeric data column is also possible, with the number of histogram bars for each numeric data column changing depending on the values that appear in the column, for example, The ability

to use a global number format for all double values while viewing them in the interactive view is another benefit of this feature.

The number of decimal places also dictates how many decimal places are used for all of the numbers in the interactive display, which is based on the number of decimal places selected. The max number of nominal values, in addition, sets the maximum number of distinct values that can be considered in a single nominal column and is presented. As a result, I chose Data Explorer to provide an overview of processed data because it helps to apply linear regression later on.

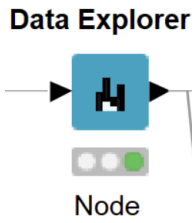


Figure 10 KNIME Data Explorer Node

The Data Explorer node enabled me to identify the needs for data wrangling process for output data that had not been met and to automate those criteria within that node. This node assists in taking a list of user-defined rules and attempting to match each row in the input data with one of the rules. If a rule is found to be applicable, the value of the rule's outcome is entered into a new column. The outcome is determined by the first matching rule that appears in the sequence of definition. Each rule is represented by a single line in the diagram. To include comments, begin each line with the symbol. (comments can not be placed in the same line as a rule). Anything that comes after the / will not be considered a rule. When a rule is met, it is composed of two parts: a condition component (antecedent), which must evaluate to true or false, and an outcome part (consequent, following the => symbol), which is placed into the new column if the rule is met. The outcome of a rule can be anything, such as a string (between 0 and 255 characters) "or /), a number, a boolean constant, a reference to another column, or the value of a flow variable value are all valid characters.

The type of the outcome column is the super type that applies to all conceivable outcomes in a given situation (including the rules that can never match). If no rule is found to match the input, the result is a missing value. Columns are denoted by their names enclosed by the dollar sign ($), and numbers are denoted using the standard decimal representation. It is important to note that strings cannot contain (double-)quotes. Flow variables are represented by the characters $$TypeCharacter and $$FlowVarName$$, respectively. The TypeCharacter for double (real) values should be 'D,' for integer values should be 'I,' and for strings should be 'S.' This node allows to manually enter column names or flow variables into the dialogue box, or can click on the appropriate lists in the dialogue box. With the use of parenthesis, logical expressions can be joined together. The following are the rules of precedence that apply to them: The NOT operator binds the most, followed by AND, XOR, and finally OR. Comparison operators always take precedence over logical connectives in the syntax of a programme. All operators (as well as their names) have case-sensitive names and values. The ROWID field is the row key string, the ROWINDEX field represents the row index (the first row has a value of 0), and the ROWCOUNT field reflects the total number of rows in the table.

As a result, I chose Rule Engine to create a new column named Product Status, which would identify whether a purchase process initiated by a customer is complete or incomplete by categorizing column data named 'Shipped', 'Canceled', and 'Resolved' as Completed and column data named "Disputed', 'On Hold', and 'In Process' are all marked as Incomplete, and a new column named Status is created for them.

Figure 11 KNIME Rule Engine Node

When performing sales forecasting in Power BI, it is important to have a datasets that has been applied using linear regression in order to perform the

forecasting later on. The implementation of linear regression in KNIME will therefore be the most successful approach of anticipating future outcomes for audit organizations, as determined by thorough study. As a machine learning technique, Linear Regression relies on the principle of supervised learning and can be used to predict future outcomes. In this operation, sales data for accessories, electronics, and furniture are transformed into information that may be used in a linear regression formula later on in the process, which is performed by the regression operator. It is a mathematical model that predicts a target value from independent variables using mathematical formulas. This technique is most commonly used to determine the relationship between variables and forecasting. In addition to the type of relationship between dependent and independent variables that they take into consideration, different regression models differ in the number of independent variables that they employ, which in my project dataset would include accessories, electronics and furniture among other things.

In addition, the estimated coefficients and error statistics are displayed in this node. A scatter plot is used to display the input data along with the regression line. The y-coordinate is fixed to the answer column that has been estimated, however the x-column can be picked from any of the independent variables that have numerical values associated with them. It will set the value of each variable that is not displayed in the view to the mean of the set of variables. This view is therefore only useful if the dataset contains only a small number of input variables after the KNIME setting has been completed.

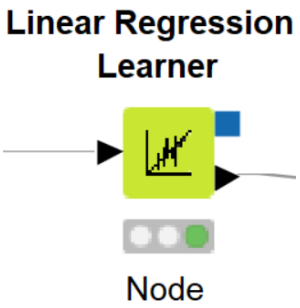


Figure 12 KNIME Linear Regression Node

It is essential that I am able to export the data into Power BI for data visualization once it has been properly structured and organized. As a result, I

selected the Send to Power BI node to export the defined data from the KNIME process to Power BI without having to perform any manual operations. It only uploads data that is compatible with Power BI's capabilities. The rest of the columns are disregarded. The following data types are supported such as String, Number (integer), Number (long), Number (double), Boolean value, Local Date, and Local Date Time (in local time zone). The node uploads rows to Microsoft Power BI in pieces, and each chunk contains one row. Despite the fact that the node has been cancel, the rows that have already been uploaded will stay in the Power BI datasets. It is necessary to authenticate in order to export data with Power BI in order to export the configured data. During the data export, a page will be opened in browser that will prompt to sign into Microsoft account in order to complete the process.

If I have not previously accepted the permission requests from the KNIME Analytic Platform application, I will need to do so once I have logged in. Once the authentication process has been completed successfully, the user will be led to a page that states that the verification code has been successfully received. As a result, I selected this node because it would be more appropriate for exporting the configured data to Power BI datasets.

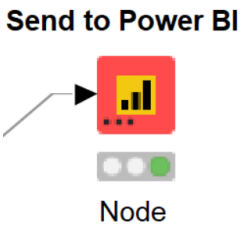


Figure 13 KNIME Send to Power BI Node

In addition, the output process must export the configuration data to a CSV file separately because finance analyst has requested that this be done in order to preserve a copy of the specified data as a backup copy of the output process. After implementing Linear Regression, it is possible that a File Writer node will have difficulties reading the configured data; as a result, I explicitly chose the CSV Writer node to export the configured data into a CSV file. The path flow variable is supported by this node. A file or a remote destination defined by a URL is written out by this node after the input data table has been processed. Because my data

consists of e-commerce sales and various IDs, the input data table can only contain string or integer columns. Other types of columns will not be supported.

Figure 14 KNIME CSV Writer Node

3.2.2.1 Implementation of Software Architecture

Software architecture serves as a plan for both the system and the project it is intended to support. The work assignments that must be completed by the design and implementation teams are outlined in this document. Scalability, performance, modifiability, and security are all characteristics of a system that are carried primarily by its architecture, and none of these characteristics can be attained in this project without a unified architectural vision. In my FYP, architecture served as an artefact for early analysis to ensure that a design approach will result in a system that is acceptable. By developing effective architecture, I was able to identify design risks and reduce them early in the development process, which was beneficial.

3.2.2.2 System Goals in Structured Natural Language

A sequence diagram is a sort of interaction diagram that shows how a group of items interacts and in what order. In this case, it defines how finance analyst and automation engineer will be playing their role to understand the project requirements for the automated system and to describe an existing process. It shows the process flow from the beginning as finance analyst explaining about the difficulties and every tedious technical process of manual audit tasks. From the discussion, the KNIME automation will be developed based on the requirements, structured data will be visualized on Power BI and implementation of machine learning algorithm as well. This sequence diagram displays how the whole process executed, starting with uploading unstructured data files in KNIME to be processed as structured. Then structured data will be uploaded to Power BI automatically once KNIME run

successfully. If CSV of structured data is uploaded unsuccessfully, the alternative is to upload to the Power BI manually. Based on the structured database, it will be visualized with machine learning algorithm to collect required outcome.

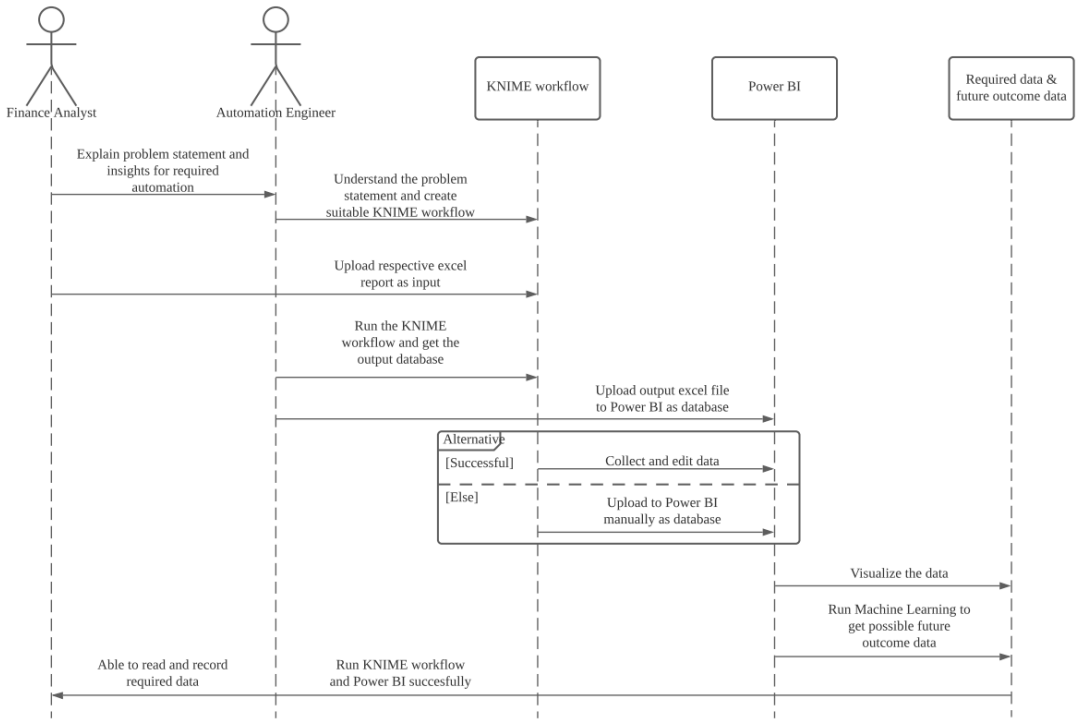


Figure 15 Sequence Diagram

3.2.2.3 Entity Relationship Diagram

The relationships between entity sets contained in a database are depicted in an entity relationship diagram (ERD). The object, or the data component of this project has been classed in an entity in this context. These entities of this project can have qualities defined via attributes. The logical structure of this project databases is illustrated by defining entities, their attributes, and displaying relationships, as per graphic below. Before a structured database of CSV is automated, the process of unstructured data will be executed. Unstructured data consist of multiple datasets as shown in the ERD diagram below such as customer, order, product and transaction. KNIME automated process will compile all the datasets in global CSV file and processed further for amendment to retrieve as structured data before proceeded for data visualization in Power BI. The properties between each entity are also shown in

the ER diagram below, where numerous unstructured data will be compiled in a single global CSV and run in a single KNIME and Power BI.

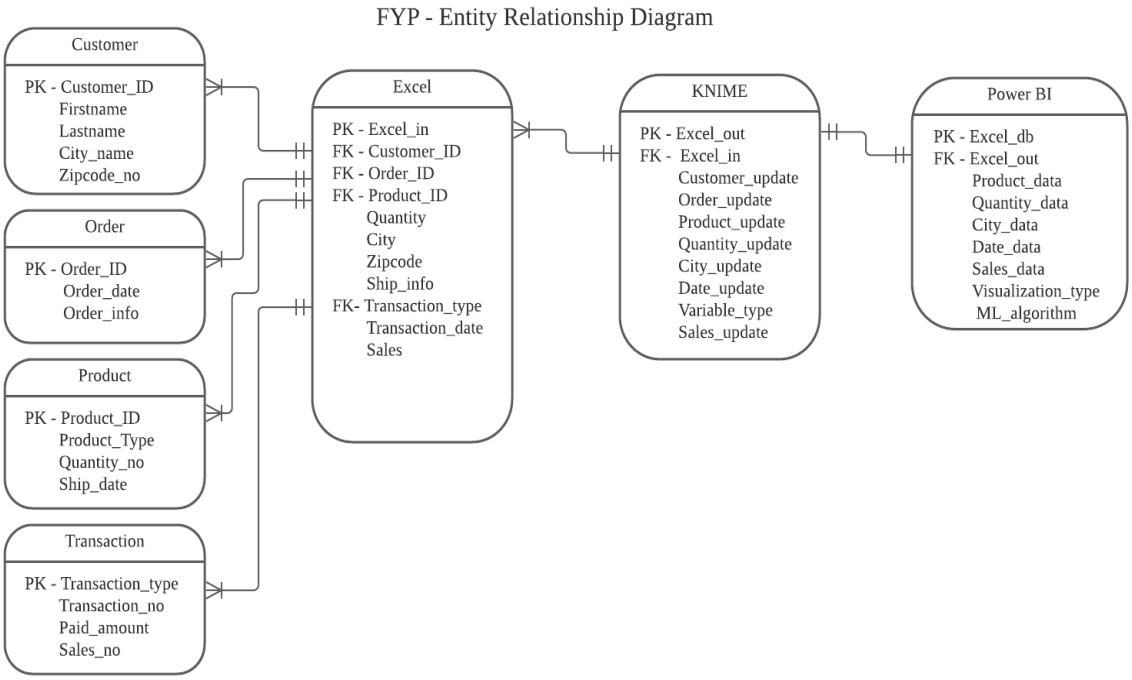


Figure 16 Entity Relationship Diagram

3.2.2.4 Data Flow Diagram

A data flow diagram (DFD) maps out the flow of information for any process or system. It is used to define the data flows from the beginning till the end of this automated process. The purpose of using this diagram for this project can frequently graphically “say” things that are difficult to describe in words, and it can be used by both technical and nontechnical audiences who are finance analyst and automation engineer based on this project. The graphic below shows the data flow of unstructured data, structured data, database, execution process until reach the end goal of the automation. It displays how the database will be shared between both finance analyst and automation engineer without complicating the technical perspective. Automation engineer develops and enhances the automation process

while finance analyst will executes and get the desired outcome for their daily audit tasks.

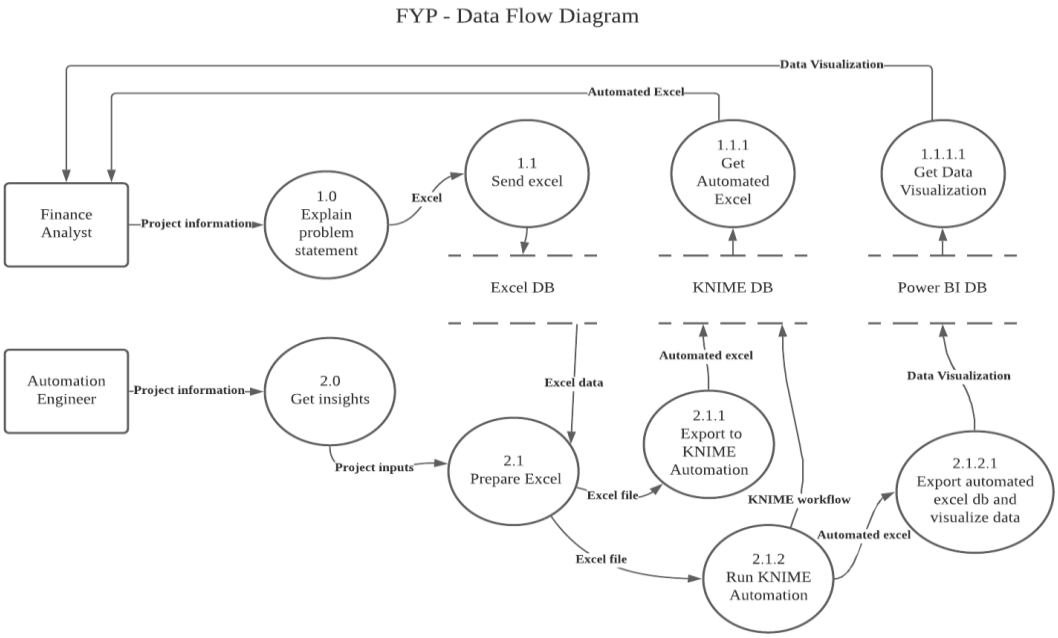


Figure 17 Data Flow Diagram

By implementing Software Architecture, I am able to build a strong foundation for my FYP while also making my platform more scalable. In addition, by incorporating a vision, the platform's overall performance is improved. The architecture is a useful tool for assessing the overall state of information technology and developing a vision of where the finance analyst needs to or wants to go with its automated structure, among other things. The architecture must provide me with the big picture and sell the vision throughout the entire software development lifecycle, evolving it as needed throughout the project and taking responsibility for ensuring that it is delivered successfully. This is critical in order to see the project through to its successful conclusion.

3.2.3 Phase 3: Development

This is where schemes and goals come to fruition. This is the logical conclusion following assessment, commitment, observation, preparation, and exploration of the resources of a project. After thorough research and agreement with the finance analyst’s agreement on proposed solution, I develop the automation process and visualization according to the requirements and how it should work based on standard procedures.

It will be necessary for me to select a file system that has the information seeking for in order to upload the data I've obtained. There are four fundamental file system options to choose from, the first of which is local file system, which allows me to select a file or folder from my local system. The other three possibilities are shared file system, shared folder system, and shared folder system. The other two systems are the shared file system and the shared folder system, respectively. The implementation of mountpoint by this node, which allows it to read from a mounted file system, is also an option. Furthermore, it has the capability of resolving the route in reference to the current mountpoint, the current workflow, or the current workflow's data area, among other things. When the option is selected, a new drop-down menu appears, from which the user can choose one of the three options available.

With the addition of the Custom/KNIME URLs feature, may now provide a custom URL when uploading CSV Reader files. Because finance analyst has provided my data, the uploading of CSV files through the local file system will be implemented, making it simple and quick to upload those e-commerce sales data that are identified as Sales Data 1 and Sales Data 2, respectively. In order to accomplish this, I set up the uploading process by importing data by selecting the appropriate CSV files, which would be Sales data 1 in the first CSV Reader node and Sales data 2 in the second CSV Reader node. In order to proceed, I first checked the preview area and verified that the data uploaded with the correct variables had been successfully imported by viewing the data displayed in the associated CSV files, and then I proceeded to the next step.

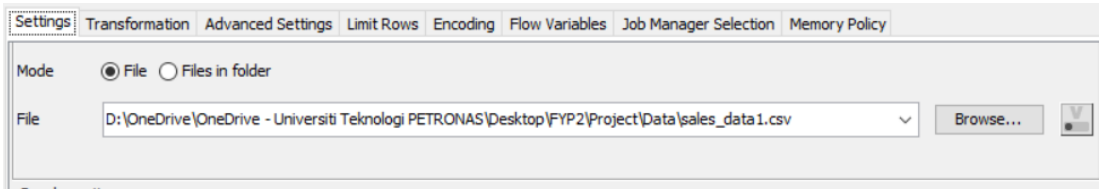


Figure 18 CSV Files Importing

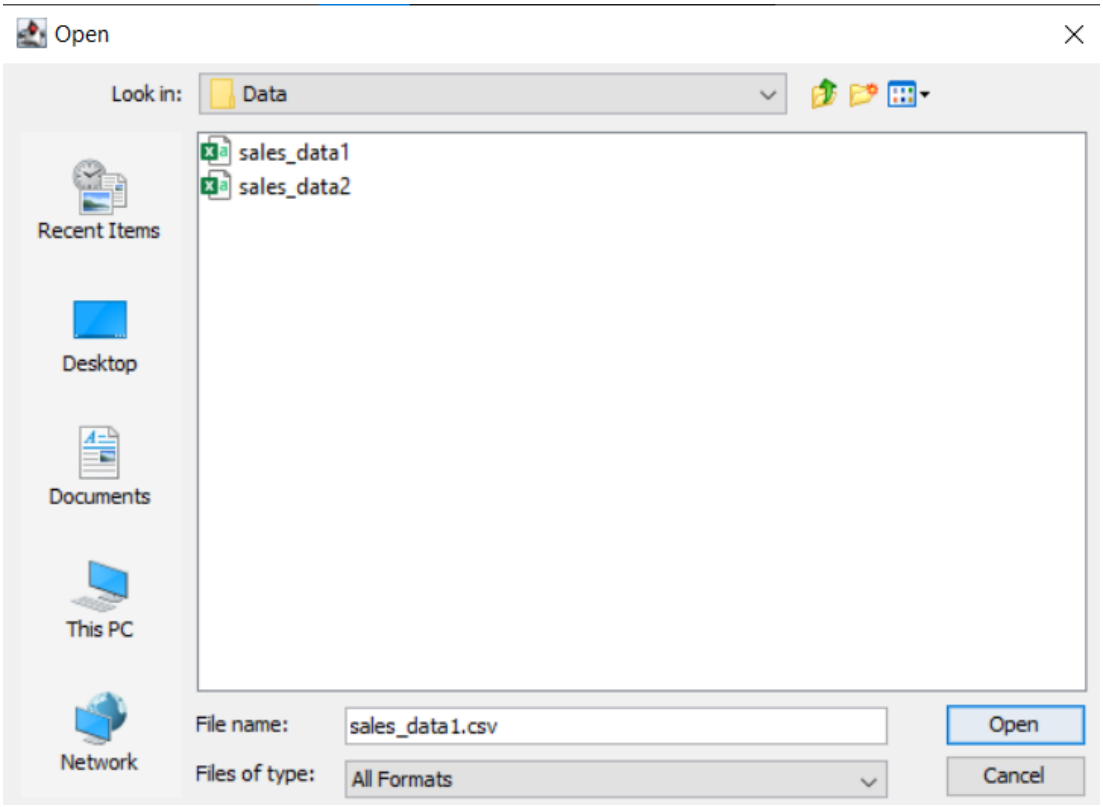


Figure 19 Selecting Respective CSV Files

Preview

The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	Date	Electronics	Furnitures	Access...	Sales	QUANTITYORDERED	QTR_ID
Row0	Jan-15	230.1	37.8	69.2	22.1	30	100001
Row1	Jan-15	44.5	39.3	45.1	10.4	34	100002
Row2	Jan-15	17.2	45.9	69.3	9.3	41	100003
Row3	Jan-15	151.5	41.3	58.5	18.5	45	100004
Row4	Jan-15	180.8	10.8	58.4	12.9	49	100005
Row5	Feb-15	8.7	48.9	75	7.2	36	100006
Row6	Feb-15	57.5	32.8	23.5	11.8	29	100007
Row7	Feb-15	120.2	19.6	11.6	13.2	48	100008
Row8	Feb-15	8.6	2.1	1	4.8	22	100009
Row9	Mar-15	199.8	2.6	21.2	10.6	41	100010
Row10	Mar-15	66.1	5.8	24.2	8.6	37	100011
Row11	Mar-15	214.7	24	4	17.4	23	100012
Row12	Mar-15	23.8	35.1	65.9	9.2	28	100013
Row13	Apr-15	97.5	7.6	7.2	9.7	34	100014

Figure 20 Preview On the Imported Data

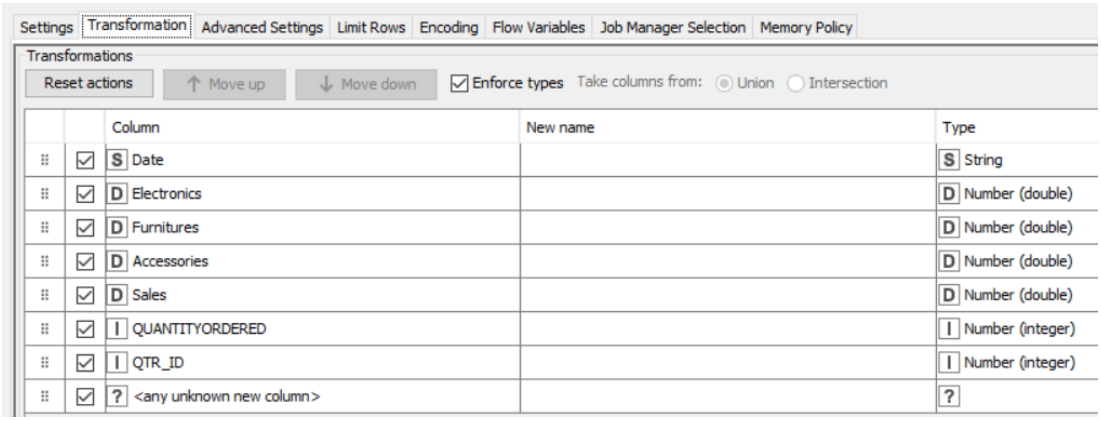


Figure 21 Datasets with Verified Variables

After data has been successfully imported into the KNIME platform, the automated workflow process can be initiated. The Joiner node would be used in the following setup to join the two CSV files that were imported. Using a Full Outer Join, may fill in the blanks in the columns of the first CSV Reader nodes and the second CSV Reader nodes with values from the missing rows as well as values from the entire data set. Only the rows that satisfy the conditions of the Outer Join will appear in the result table of the query. As a result, I used the Full Outer Join method to combine the data from both CSV files by matching the columns that were comparable in both files. In this scenario, the QTR ID column will be the most appropriate one to employ in order to merge and customize the entire set of data. Apart from that, I configured the column selection tab to choose all of the essential columns for data wrangling and data cleanup, and I saved the changes.

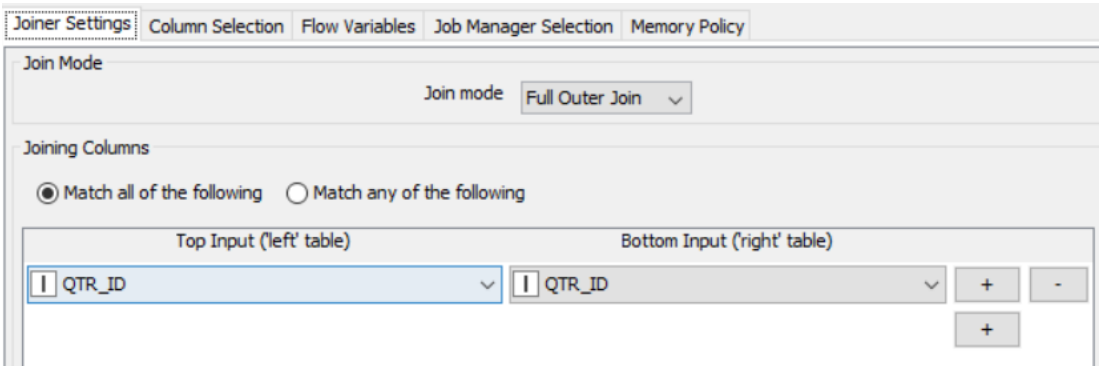


Figure 22 Joiner Node Configuration

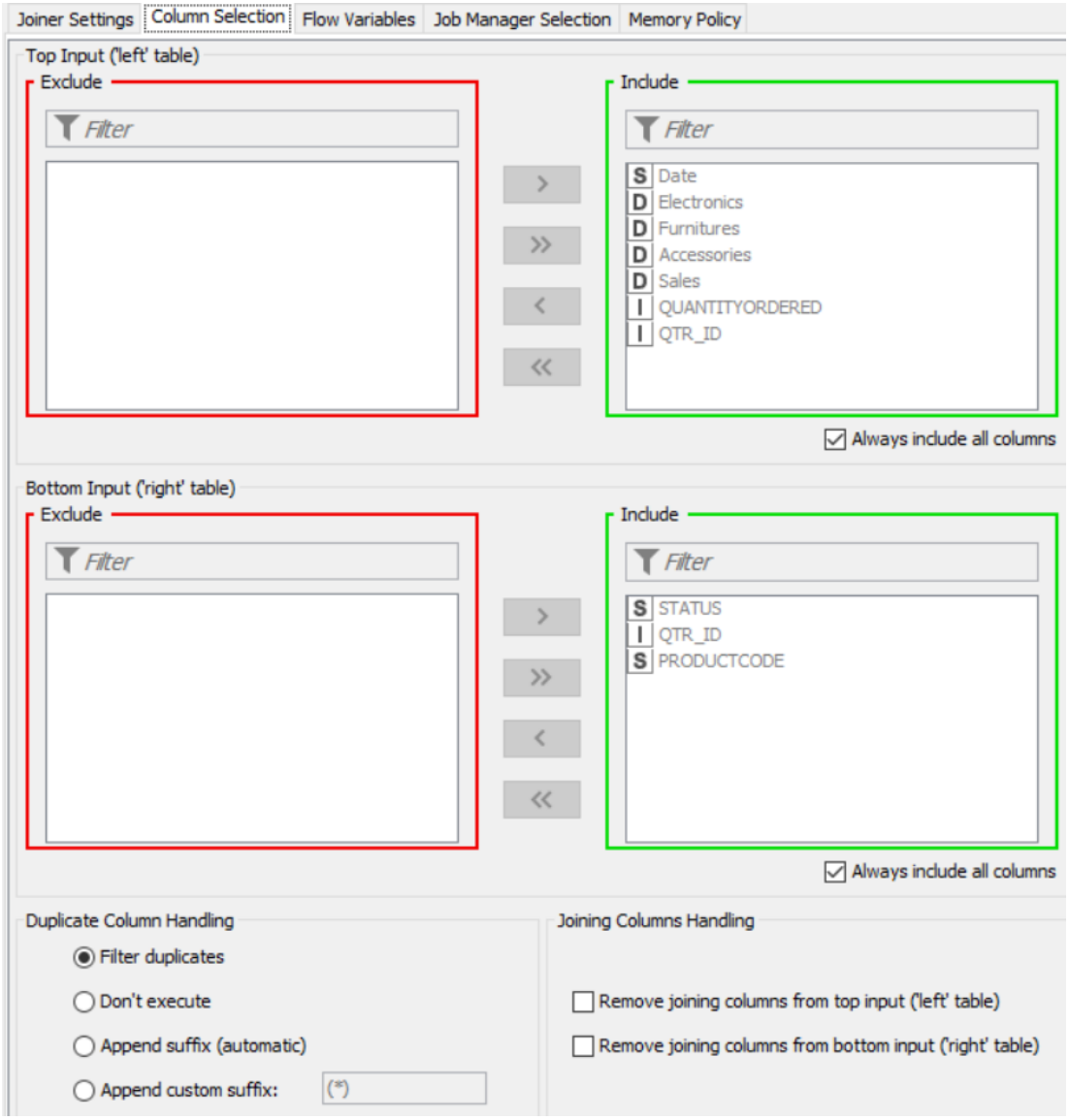


Figure 23 Column Selection Configuration

In order to avoid confusion when viewing the information, I looked into configuring the data to merge CSV and column data successfully. This was done because the finance analyst had requested that the column name of QTR ID be changed to Customer ID so that there would be no confusion when viewing information. As a second thought, I changed the name of the column Product Status to Status in order to create a new column named Product Status in order to avoid a column configuration problem later on. In order to prevent data from modifying, the automation must ensure that variables such as Customer ID and Status do not change from their initial values of Int and String, respectively, in the data.

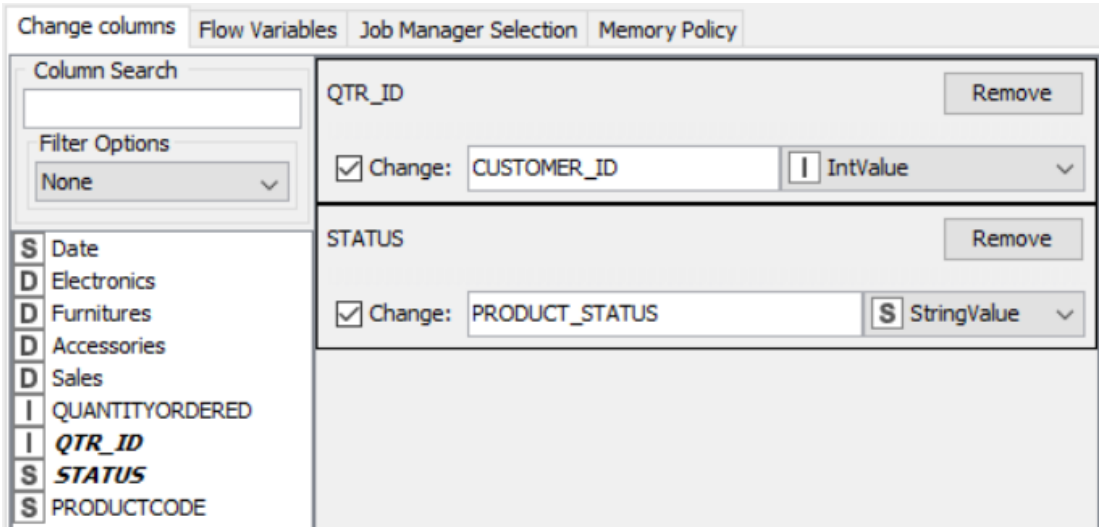


Figure 24 Change Column Configuration and Automation

Once the Column Rename automation has been completed successfully, the next step would be to setup the entire data set to eliminate duplicate data by leveraging the Column Rename node to complete the process. This is owing to the fact that duplicated data in a CSV file must be cleaned up by eliminating duplicate data, which later on aids in the effective implementation of Linear Regression and the avoidance of unexpected consequences when applying linear regression to the data set. As a result, the Duplicate Row Filter node is in charge of finding duplicate rows that have the same values in particular columns as the original row and removing them from the graph. As a result, my configuration reveals that the Status column has numerous instances of the same data, prompting me to customise it by inserting the appropriate variable into the column.

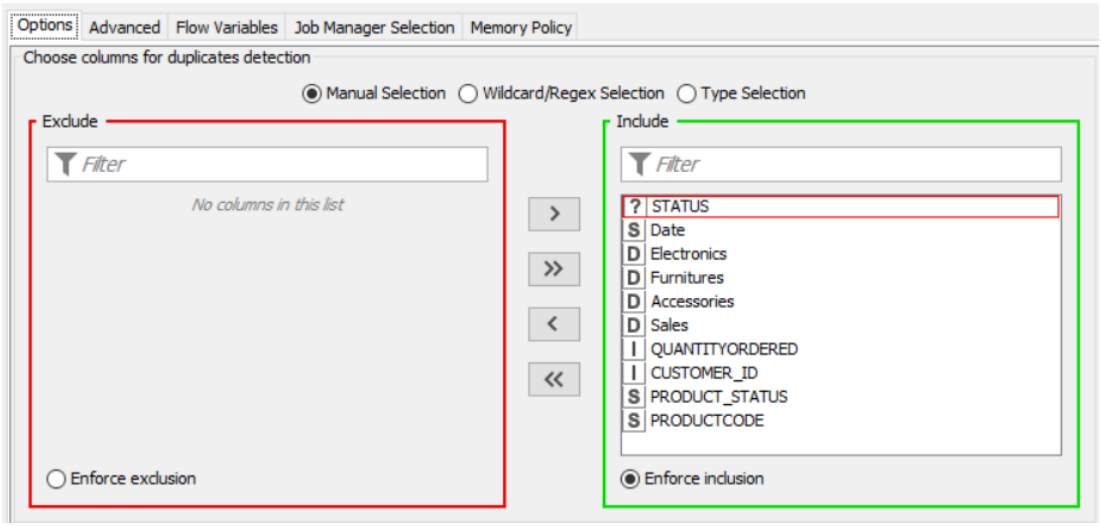


Figure 25 Duplicate Row Filter Configuration

Following the successful removal of duplicated data, the next step would be to review the data in an overview using the Data Explorer node, which is accessible through the Data Explorer node. It was decided to utilize the Data Explorer node to check the data because it provides a variety of options for displaying portions of the supplied data in an interactive manner. By selecting the checkboxes next to the numbers want to see, I could see the most common and least frequent values, respectively.

For each number n entered into the interactive view, the nominal tab will generate two columns: one column carrying the n most common nominal values, and another column containing the n most infrequent nominal values, and so on for each number n in the interactive view. In order to perform linear regression with the least amount of error, this setup assists in the initialization of the essential data for the execution of linear regression. Using the settings, it was possible to see the minimum and maximum values, as well as the mean, standard deviation, variance, and skewness of all numeric column data. It also displays all of the data from the nominal columns of dates, product statuses, and product codes (in the nominal tab), among other things.

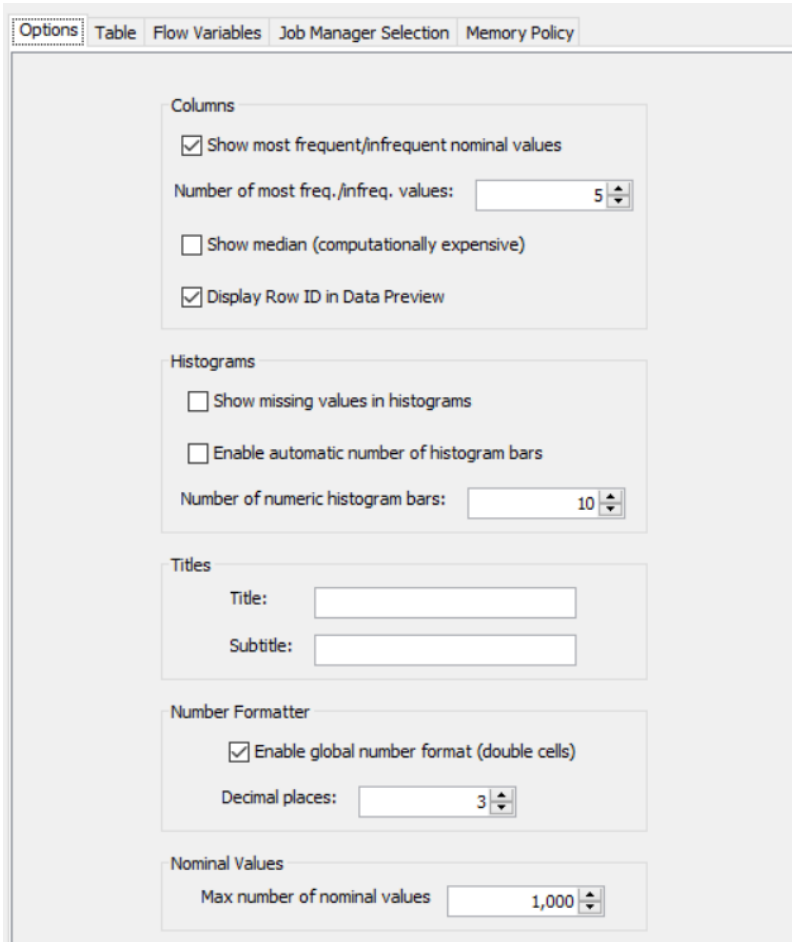
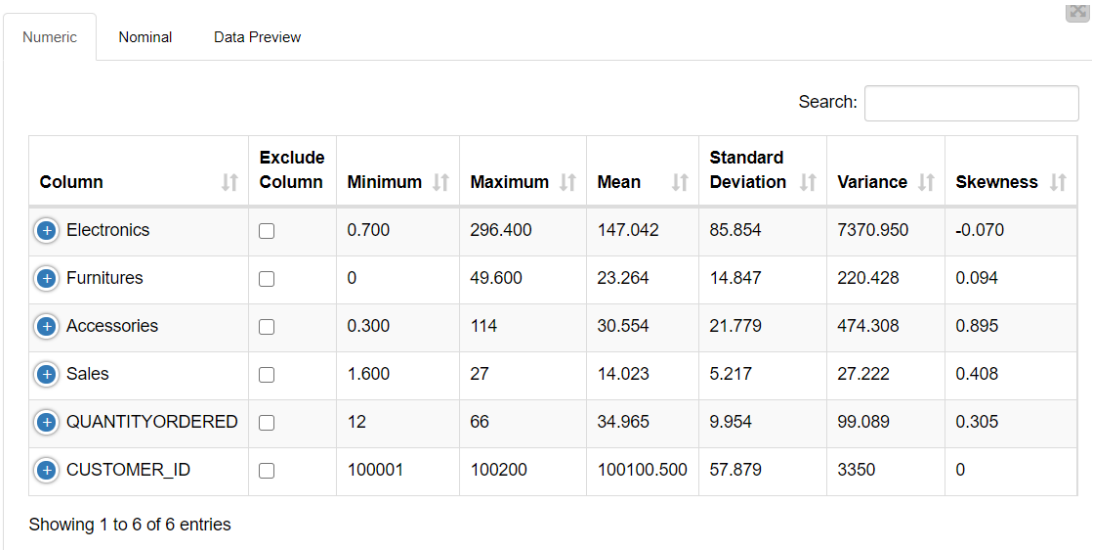


Figure 26 Data Explorer Configuration



Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness
Electronics		0.700	296.400	147.042	85.854	7370.950	-0.070
Furnitures		0	49.600	23.264	14.847	220.428	0.094
Accessories		0.300	114	30.554	21.779	474.308	0.895
Sales		1.600	27	14.023	5.217	27.222	0.408
QUANTITYORDERED		12	66	34.965	9.954	99.089	0.305
CUSTOMER_ID		100001	100200	100100.500	57.879	3350	0

Figure 27 Numeric Data Display

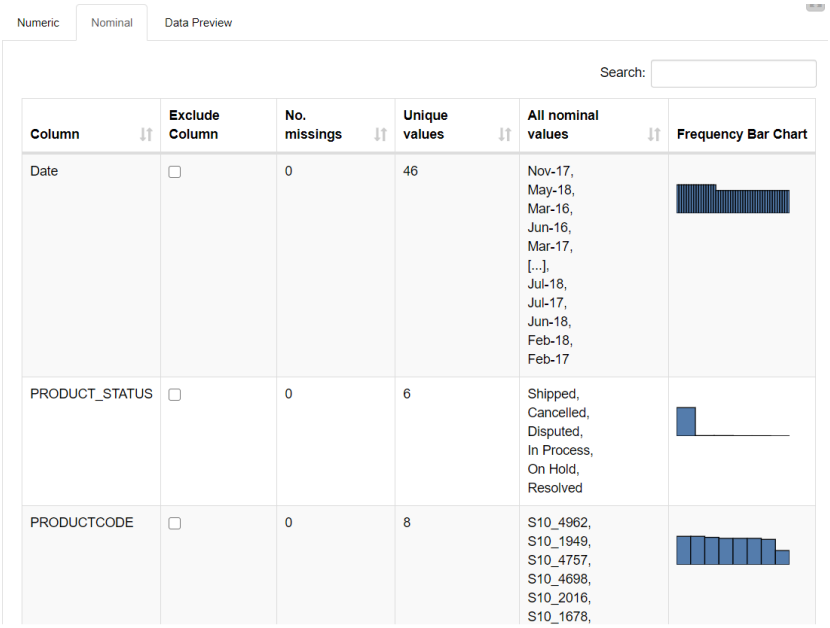


Figure 28 Nominal Data Display

After obtaining an overview of the total data through the use of Data Explorer, this step will initialise all of the data in order to prepare it for the Linear Regression implementation. This is due to the fact that linear regression automation must be able to read the automated data and effectively apply the linear regression configuration. As a result of the idea of supervised learning, Linear Regression is a technique that can be used to predict future outcomes. The regression operator performs this operation on the sales data for accessories, electronics, and furniture. The information obtained from the sales data is then utilized to create a linear regression formula, which is then utilized in the last step of the process. Using mathematical formulas in the automation, it is possible to anticipate a goal value from independent factors.

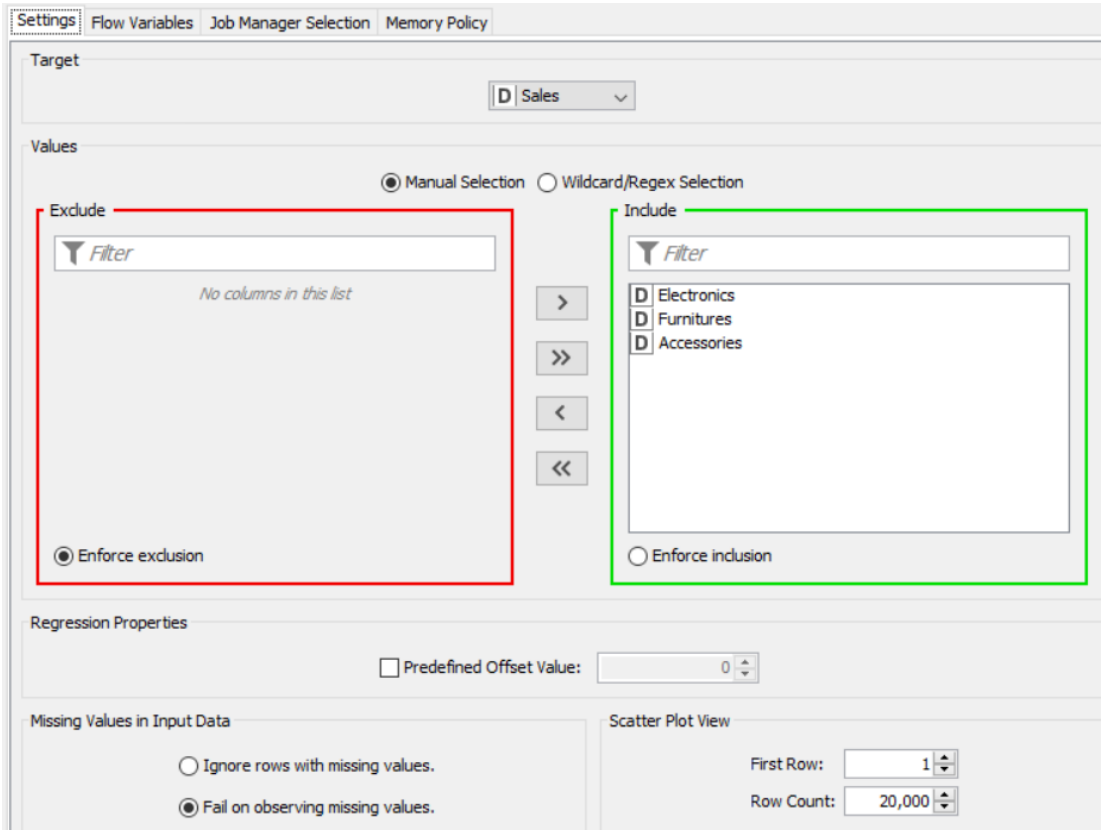


Figure 29 Linear Regression Configuration and Automation

The output component was my primary emphasis when I confirmed that linear regression was being used on the appropriate sales data. This included exporting the wrangled and automated sales data to Power BI datasets using the Send to Power BI node. The node uploads rows to Microsoft Power BI in chunks of one row each, with each chunk containing a total of one row. Despite the fact that the node has been terminated, the rows that have already been uploaded will remain in the Power BI datasets until they are deleted. To successfully export the configured data for data visualization, it is important to authenticate with Power BI in order to export the data using Power BI properly. As a result, I exported the configured datasets titled 'Regression Model' into the Power BI datasets and imported them into Power BI.

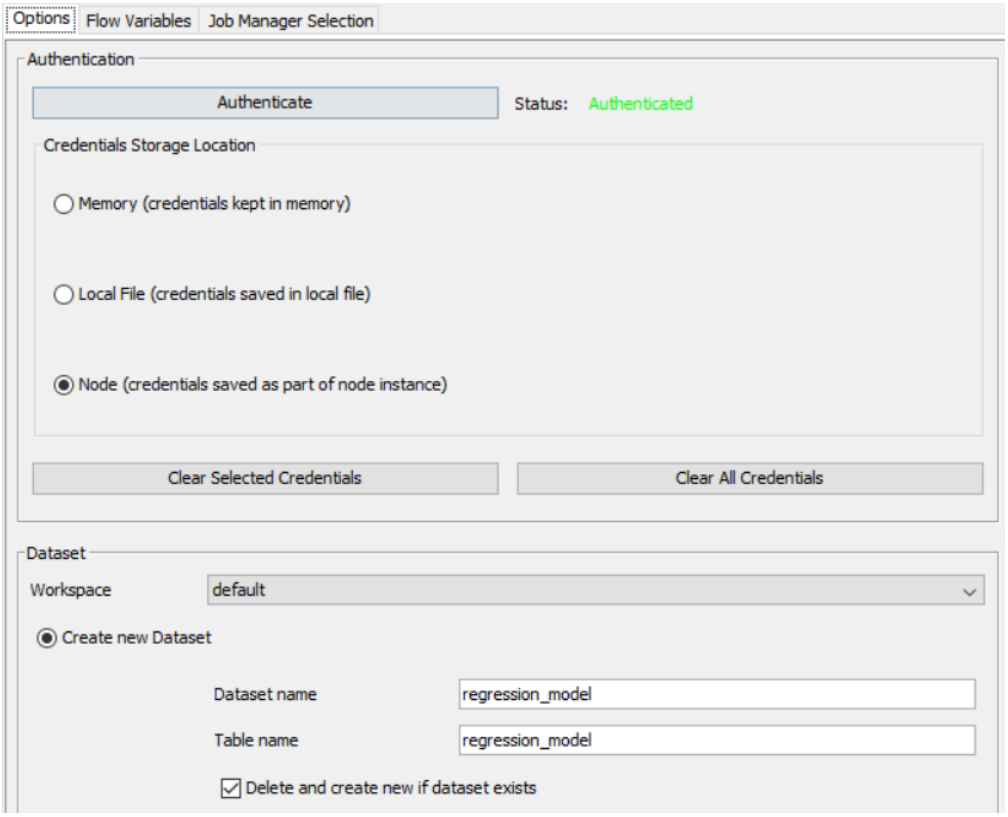


Figure 30 Send to Power BI Configuration

Lastly, but certainly not least, I worked on creating a simple and user-friendly data visualization in Power BI to display sales predictions and the entire automated data set. Finance analysts are typically in charge of keeping track of their company's sales success. This technique typically entails reviewing a number of different sources in order to gain insight into the trends in sales data from the past, present, and future. I was able to represent the key information about sales forecasts that may be summed in a custom format in multiple visualizations, which was a significant accomplishment. I was able to prioritize finance professionals' concerns by utilizing Power BI, which is a better and more efficient tool for my final year project. I also made a comparison between automated CSV sales data with linear regression and Power BI dashboard to optimize data reading to prioritize finance professionals' concerns.

While the linear regression did change the specific data, it also supplied coefficient values that were exported to Power BI in a tabular form, where they could be used in the dashboard equations. The predicted sales were visualized in a newly generated module named Predicted Return on Investment, which I developed using

Power BI equations. In addition, I built numerous What-If parameters where I could type in targeted sales in furniture, electronics, and accessories, each of which differed in the Power BI formula module that I had previously created and implemented. I designed a revenue trend dashboard, sales performance dashboard, sales table dashboard, and sales prediction dashboard to provide a summary of the previous year's sales. With the use of this data visualization, finance analyst was able to identify sales regions that needed improvement, factors that influence customer contentment and dissatisfaction, and what to do with certain products that were performing poorly.

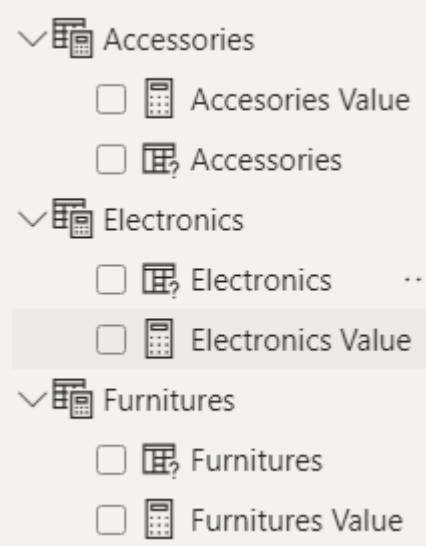


Figure 31 What-IF Parameters

```
Predicted Return on Investment =
var accessories_coef = CALCULATE (SUM( 'test (2)'[value]), 'test (2)'[component] ="Accessories")
var electronics_coef = CALCULATE (SUM( 'test (2)'[value]), 'test (2)'[component]="Electronics")
var furnitures_coef =CALCULATE (SUM( 'test (2)'[value]), 'test (2)'[component]="Furnitures")
var intercept = CALCULATE(SUM('test (2)'[value]), 'test (2)'[component] = "intercept")
var formula = intercept +(accessories_coef * [Accesories Value])+(electronics_coef * [Electronics Value])+(furnitures_coef * [Furnitures Value])
return
IF(AND([Accesories Value]>0,AND ([Electronics Value], [Furnitures Value])), formula,0)
```

Figure 32 Power BI Formula in Module

3.2.4 Phase 4: User Acceptance Testing

This is where any setbacks with the KNIME automation or Power BI visualization get resolved to wrap up the development and implementation process. The project's longevity is also being checked here, as I implemented multiple black box testing techniques. This testing phase helped me to identify issues and errors from the KNIME automation and Power BI dashboard. While implementing testing on the created automation and data wrangling, I received feedback after executing the automated process few times, errors will be amended and resolved based on what finance analyst executed it.

3.2.4.1 Black-Box Testing - Use Case Testing Technique

Using a Use Case as a tool for describing the required user interaction and for developing new applications, automating processes, or making changes to existing software functionalities is a common practice. Use Case Testing is typically performed as part of black box testing, and it assisted me in identifying test scenarios that exercised the entire system on a transaction-by-transaction basis from start to finish for my Final Year Project. It is a functional testing technique that assisted me in identifying and testing scenarios across the entire system or in performing start to finish data cleansing and manipulation. In the course of executing the Use Case testing technique, I established a number of sub-goals(G2-G5) to meet the major goals(G1) in order to ensure that the FYP was a success in its entirety.

Table 4 Use Case Table

No	Goals
G1	Data Wrangling automation based on finance analyst requirement
G2	Get respective e-commerce CSV report and import to KNIME
G3	Run KNIME workflow to edit CSV report, apply linear regression and export to Power BI
G4	Upload automated database to Power BI
G5	Visualize data, analyze and predict sales forecast in Power BI

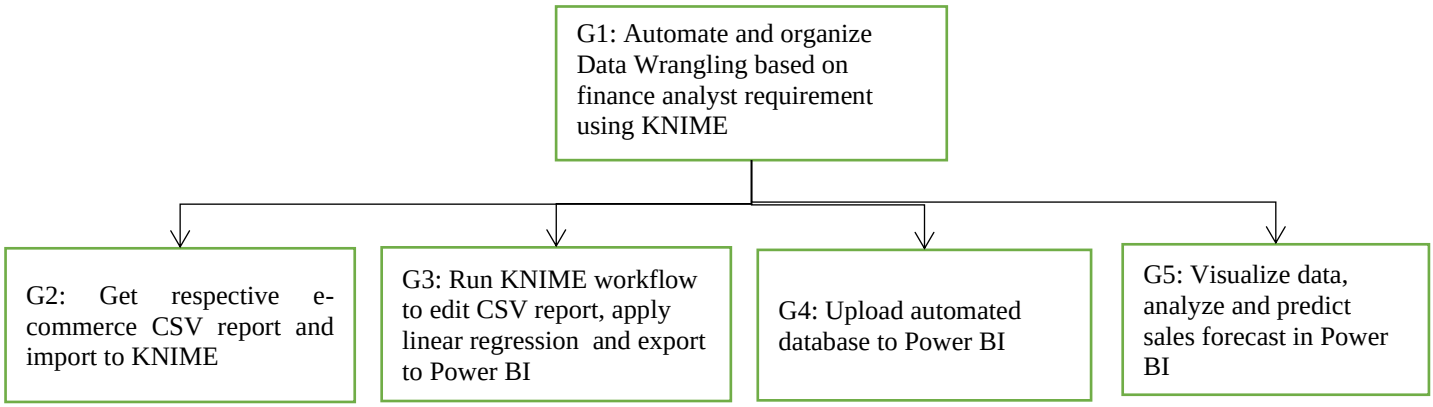


Table 5 Table G1

Scenario ID	G1
Name	Kishen Sivakumar
Associated Goal	Able to automate and organize KNIME automation based on project admin/finance analyst's request
Actor	Automation engineer, Project admin/Finance analyst
Precondition	Having meet up session with finance analyst to understand problem statement
Post condition	Plan and initiate project based on finance analyst's requirement
Result	Able to automate KNIME workflow, Data visualization on Power BI and predict sales forecast
Main Steps	1. Set up discussion session with finance analyst to understand their request and problem statement 2. Plan suitable KNIME automation workflow based on and their requirements and initiate 3. Get CSV report from finance analyst or download from source if given by finance analyst 4. Import that CSV report/ data to KNIME 5. Automate efficient KNIME workflow to edit the CSV data 6. Export the edited CSV report 7. Upload those data into Power BI and create suitable data visualization

	8. Run machine learning to predict sales forecast data 9. Send over the completed project to finance analyst/project admin
Alternative Steps	- Automate efficient KNIME workflow to edit the CSV data - Automate KNIME workflow using suitable function node - Automate KNIME workflow by finding suitable function nodes by browsing - Automate KNIME workflow while having discussion with finance analyst to get quick insights from them
Exception Steps	- KNIME cannot read the CSV report/data - KNIME finds the CSV data is incorrect or blank - KNIME asks to import different CSV report/data

Table 6 Table G2

Scenario ID	G2
Name	Kishen Sivakumar
Associated Goal	Get respective CSV report and import to KNIME
Actor	Automation engineer, Project admin/Finance analyst
Precondition	Prepare correct CSV report/data or download from website link if given by finance analyst
Post condition	Able to import CSV report/data to KNIME workflow
Result	Successfully able to read the CSV data
Main Steps	1. Get CSV report/data from finance analyst/project admin or download from any website link if they agree to provide 2. Able to import and read report data on KNIME successfully
Alternative Steps	- Get CSV report/data from finance analyst/project admin on chat

	- Get CSV report/data from finance analyst/project admin by emailing them
Exception Steps	- KNIME cannot read the CSV report/data - KNIME finds the CSV data is incorrect or blank - KNIME asks to import different CSV report/data

Table 7 Table G3

Scenario ID	G3
Name	Kishen Sivakumar
Associated Goal	Run KNIME workflow to edit CSV report, apply linear regression and export to Power BI
Actor	Automation engineer, Project admin/Finance analyst
Precondition	Run KNIME workflow with suitable function nodes with no errors
Post condition	Able to run KNIME workflow successfully and get the expected outcome of CSV file
Result	Successfully run the automation workflow
Main Steps	1. Make sure the KNIME workflow generated with no errors 2. Make sure KNIME workflow able to read the input data before run the workflow 3. Apply linear regression successfully
Alternative Steps	- Run the workflow multiple times while creating the KNIME workflow to understand and modify the errors - Fix the errors with suitable function nodes

Exception Steps	- KNIME cannot read the CSV report/data - KNIME workflow may face configuration error - KNIME function nodes may not be able to process required function

Table 8 Table G4

Scenario ID	G4
Name	Kishen Sivakumar
Associated Goal	Upload CSV outcome database to Power BI
Actor	Automation engineer, Project admin/Finance analyst
Precondition	Able to run KNIME workflow successfully and retrieve the output data in CSV
Post condition	Upload the database to Power BI for data visualization
Result	Successfully upload to Power BI and able to read the database
Main Steps	1. Retrieve end result database from KNIME workflow 2. Upload the database successfully on Power BI
Alternative Steps	- Use Power BI function node on KNIME to upload the database to Power BI - Extract and import the data from KNIME and upload to Power BI manually
Exception Steps	- KNIME cannot read the imported data - KNIME workflow may have error for output - Power BI may not read the database

Table 9 Table G5

Scenario ID	G5
Name	Kishen Sivakumar
Associated Goal	Visualize data, analyze and predict sales forecast in Power BI
Actor	Automation engineer, Project admin/Finance analyst
Precondition	KNIME successfully upload output database to Power BI
Post condition	Able to visualize sales forecast database outcome
Result	Successfully visualize, analyze current database, and predict the sales forecast.
Main Steps	1. Visualize final data with proper representation of data 2. Able to analyze and get required data 3. Power BI predicts and visualize potential sales forecast
Alternative Steps	- Find alternative ways on visualizing the data - Represent the data in many visualization options to help Finance analyst/ Project Admin data collection process easy - Record required data manually or extract data straight from the Power BI database - Find alternative ways of machine learning to use to predict the sales forecast dashboard
Exception Steps	- KNIME cannot visualize specific data type - KNIME not able to extract required data - Machine learning error to process the sales forecast data

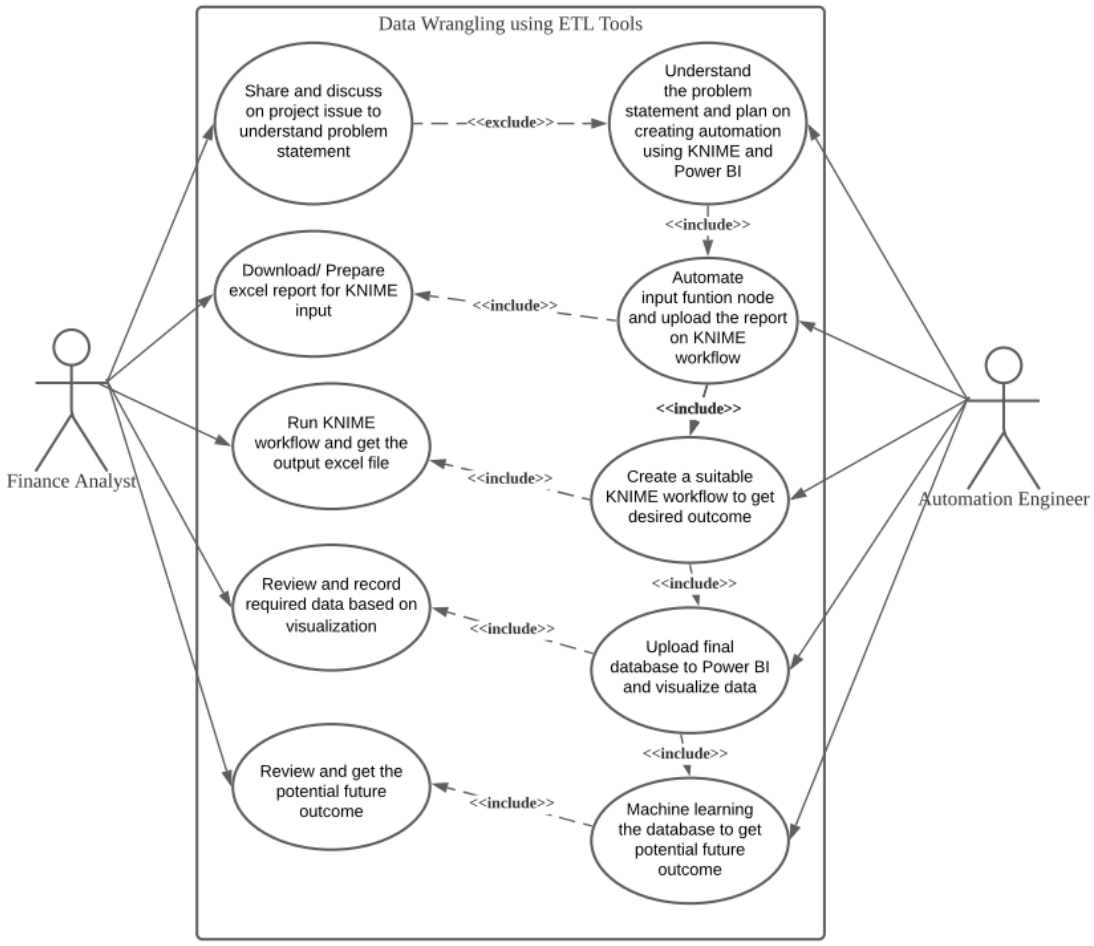


Figure 33 Use Case Diagram

3.2.4.2 Black-Box Testing - Decision Tree Testing Technique

One of the predictive modelling methodologies used in statistics, data mining, and machine learning is decision tree technique. It goes from observations about an item to inferences about the item's goal value using a decision tree as a predictive model. Classification trees are tree models in which the goal variable can take a discrete set of values represent feature combinations that lead to those class labels. A decision tree can be used to describe decisions and decision-making visually and explicitly in decision analysis. A decision tree is a data mining tool that can be used in data wrangling as well in my FYP to represent data, but the resulting classification tree can be an input for decision making.

The decision tree technique is beneficial for my final year project model, which is based on the KNIME Analytic tool. Software testing technique called decision table testing is used to test the automated behavior for various input combinations using a variety of different CSV input combinations. This is a systematic approach in which the various input CSV file combinations and KNIME's settings, as well as the accompanying functionality, are recorded in a tabular format. Hence the name "Cause Effect Table," which is used to record the causes and effects of input data as well as configuration settings in order to improve test coverage and efficiency. As a result of the KNIME configuration, a Decision Table with two excel files containing e-commerce sales data versus rules of each KNIME node function with the required test conditions is tested for effective decision making at the end of the process.

On the basis of the tabular data I generated, I was able to conclude that if a single KNIME node is not successfully executed at the beginning of the process, it will have an impact on the succeeding corresponding nodes. Based on the FYP KNIME model figure attached, for example, the Joiner node is not executed properly (there is a exclamation sign in the node) due to incorrect configuration, and this has an impact on the subsequent KNIME function nodes (which display red color in the subsequent nodes), indicating that the node is not configurable to produce successful output data. The configuration of the KNIME automation process will be affected regardless of which node is involved or where it occurs in the process flow. As a result, the entire process will be failed at the completion of the process.

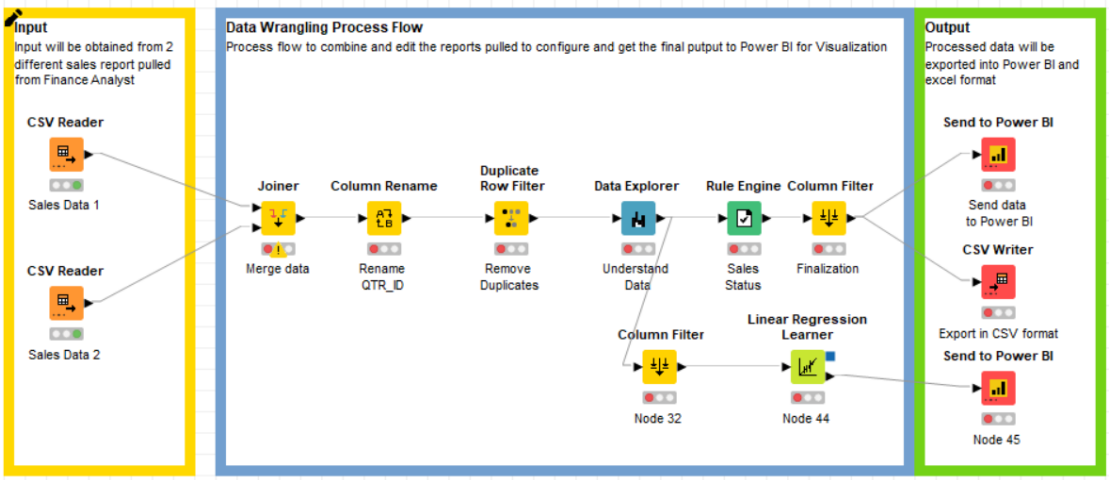


Figure 34 Failed KNIME workflow model

Table 10 Decision Tree Table

Condition	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10	Case 11	Case 12
Input Data **CSV Reader Node - Sales Data 1**	Configuration Executed	**Configuration Not Executed**	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed
CSV Reader Node - Sales Data 2	Configuration Executed	Not Configurable	**Configuration Not Executed**	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed
Process - KNIME Automation Data Wrangling **Joiner Node**	Configuration Executed	Not Configurable	Not Configurable	**Configuration Not Executed**	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed
Column Rename Node	Configuration Executed	Not Configurable	Not Configurable	Not Configurable	**Configuration Not Executed**	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed
Duplicate Row Filter Node	Configuration Executed	Not Configurable	Not Configurable	Not Configurable	Not Configurable	**Configuration Not Executed**	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed
Data Explorer Node	Configuration Executed	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	**Configuration Not Executed**	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed
Rule Engine Node	Configuration Executed	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	**Configuration Not Executed**	Configuration Executed	Configuration Executed	Configuration Executed	Configuration Executed
Column Filter Node	Configuration Executed	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	**Configuration Not Executed**	Configuration Executed	Configuration Executed	Configuration Executed
Linear Regression Learner Node	Configuration Executed	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	**Configuration Not Executed**	Configuration Executed	Configuration Executed
Output Data **Send to Power BI Node**	Configuration Executed	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	**Configuration Not Executed**	Configuration Executed
CSV Writer	Configuration Executed	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	Not Configurable	**Configuration Not Executed**
Overall KNIME Automation	**Successful**	**Unsuccessful**	**Unsuccessful**	**Unsuccessful**	**Unsuccessful**	**Unsuccessful**	**Unsuccessful**	**Unsuccessful**	**Unsuccessful**	**Unsuccessful**	**Unsuccessful**	**Unsuccessful**

3.2.5 Phase 5: Project Deployment

Once the automation processes are created by the evolutionary development phase it is ready to be put into production. The finance analyst will be running the automated processes based on their designated tasks and schedule. When the project is not deployed, it's only seen by the finance analyst and me. On the other hand, when the product is deployed, a wide range of finance analyst will be using it, and the feedback will be more reliable to look at.

3.2.5.1 KNIME Automation

The KNIME automation that I had developed for one of the finance analysts was deployed to him after the development phase had been successfully finished. Although finance analysts can run the automation on their own, I accompanied him on the first deployment to train and provide essential concepts as well as an overall picture of how to run the automation. As a result, I had to explain to him how each of the function nodes in the KNIME workflow was automated, as well as how each of his detailed requests was automated. Whenever the circular green light is illuminated on each function node, we may be certain that the function node is correctly executing. If any of the circular lights turns red, it indicates that the function node is not executable and that it requires more setup before it can be executed. If the colour orange is displayed, it indicates that the function node must be executed in order to make changes to the sales data files.

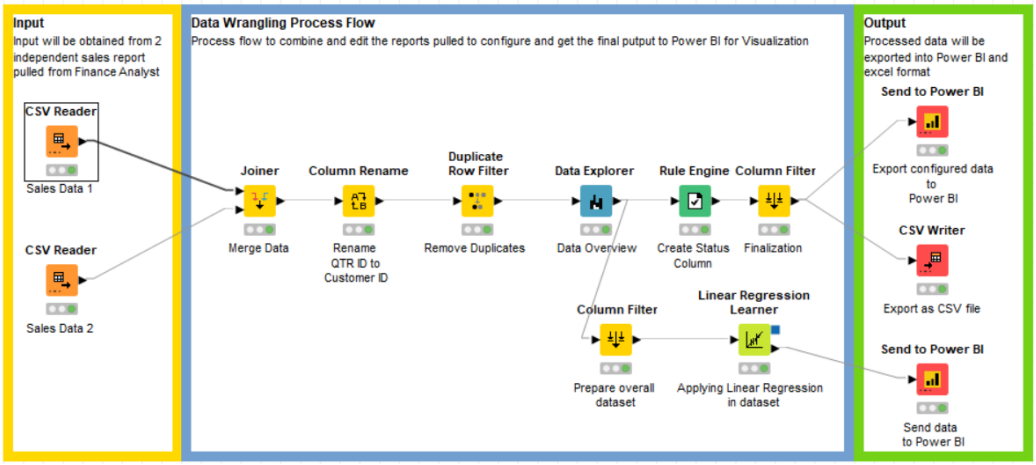


Figure 35 Successful KNIME Workflow Model

3.2.5.2 Power BI Visualization

After the development phase had been completed successfully, the Power BI application that I had created for one of the finance analysts was made available to him. He asked me to accompany him on his first deployment so that I could train him and explain the fundamental ideas, as well as provide an overall picture of how the visualization works and how to use it. Therefore, I had to explain to him how each What-If parameter in Power BI was used, as well as how each of his precise demands was converted into a visual representation. When an individual's input value is entered into the What-If parameter that has been generated, it displays the expected return in investment value, which would be the sales forecasting value for the financial analyst to use. While carrying out the assignment, he gained an understanding of how to manage Power BI, which they had previously employed in their daily tasks.

When I was satisfied that the Power BI dashboard was executable and that the finance analyst's device could run it, I gave him permission to use the dashboard on his device. The deployment of the Power BI dashboard went off without any hitches. After successfully exploring with and refining the Power BI dashboard for sales forecasting, and as a result, I have high confidence in the effectiveness of the Power BI Dashboard that I created.

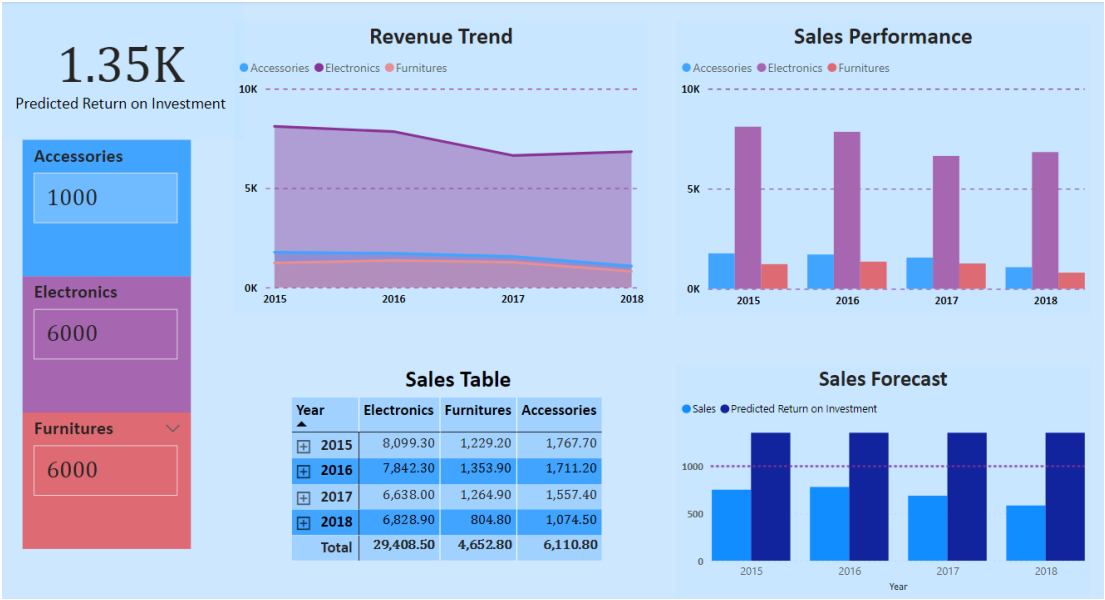


Figure 36 Successful Power BI Dashboard

3.2.6 Phase 6: Feedback & Documentation

The deployed project allows finance analyst the opportunity to submit early and ongoing feedback. This solves a fundamental issue in which the client could only see the finished product of KNIME workflow and Power BI dashboard at the end. Thus, it allows me to incorporate most new changes into the product development process, i.e. in future iterations, by collecting frequent finance analyst' input as the project develops in each iteration. This helps to identify issues at the early stage and resolving them. Every detailed information of the phase will be documented throughout the completion of project.

3.3 Gantt Chart

Throughout my study, I use a Gantt Chart to keep track of my work on both the project and the course. I was able to develop and complete every element of my research process with adequate planning and scheduling. This the Gantt chart below shows my progress throughout my Final Year Project 1 and Final Year Project 2 which was from May till December.

Task/Week	FINAL YEAR PROJECT 1												FINAL YEAR PROJECT 2														
	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15
Planning Phase																											
Project Title Selection																											
Preliminary Research																											
Literature Research																											
Discussion with Finance Analyst																											
Data Gathering & Analysis																											
Software Requirement Enginering Implementation																											
Design Phase																											
Understanding KNIME function nodes																											
Understanding Power BI Visualization																											
Implementing Software Architecture																											
Proposal Defense																											
Submission of progress assessment 1																											
Interim Report Draft Submission																											
Interim Report Submission																											
Development Phase																											
KNIME Automation																											
Power BI Dashboard																											
Testing Phase																											
Decision Tree Testing																											
Use case Testing																											
Deployement Phase																											
User Acceptance Testing																											
Feedback & Documntation Phase																											
Finance Analyst Feedback																											
Project dissertation																											
Presentation Slides and Prototype Video Submission																											
Viva Presentation																											

	Completed
	In Progress
	Yet to be completed

CHAPTER 4

RESULT AND DISCUSSION

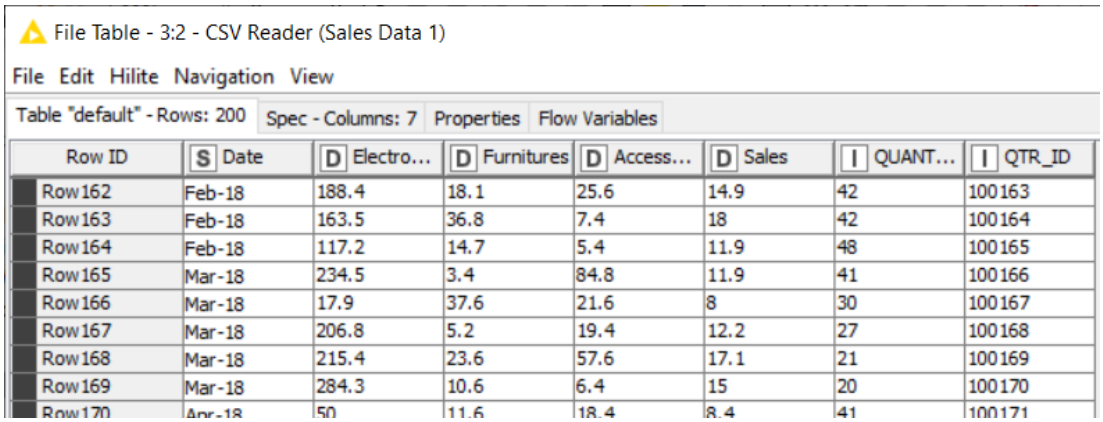
4.1 Result and Discussion

The configuration and automation were working well after a rigorous development, testing, and deployment phase, which resulted in the elimination of nearly all syntax errors and errors. This testing and deployment step allows me to concentrate first and foremost on the results I acquired and to clearly outline what occurred throughout my literature review study and testing without having to worry about the ramifications of those results. Both the KNIME workflow and the Power BI dashboard were successfully completed, as I will discuss in more detail in the next sections.

4.2 KNIME Automation and Linear Regression

When I was satisfied that the KNIME process was executable and that the finance analyst's device was capable of running it, I provided him permission to use his device to run the workflow. The workflow was organized into three basic sections: the input section, the processing section, and the output section. The input portion was divided into two sections: the processing section and the output section. All three stages of the workflow were carried out one after the other in order to exclude any possibility of human error during the execution process itself. A total of two automated nodes (two CSV Reader nodes) are present in the input area to facilitate the import and reading of Sales Data 1 and Sales Data 2. After applying the automation and configuration, I was able to successfully execute those CSV Reader nodes, which read the table column and rows, respectively, from the CSV file. As shown in the attached figure 34, sales data 1 has been imported in its entirety, including all the data columns and data rows, in preparation for subsequent configuration activities. However, as shown in Figure 35, which is attached below, sales data 2 was properly imported into the system.

As a result of my ability to effectively complete the input section, the process section is completed. The procedure part, as well as the entire section, is effectively completed.

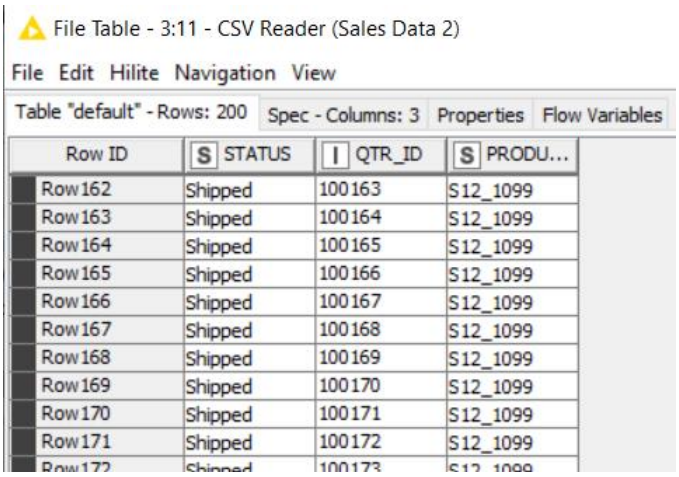

File Table - 3:2 - CSV Reader (Sales Data 1)

File Edit Hilite Navigation View

Table "default" - Rows: 200 | Spec - Columns: 7 | Properties | Flow Variables

Row ID	Date	Electro...	Furnitures	Access...	Sales	QUANT...	QTR_ID
Row162	Feb-18	188.4	18.1	25.6	14.9	42	100163
Row163	Feb-18	163.5	36.8	7.4	18	42	100164
Row164	Feb-18	117.2	14.7	5.4	11.9	48	100165
Row165	Mar-18	234.5	3.4	84.8	11.9	41	100166
Row166	Mar-18	17.9	37.6	21.6	8	30	100167
Row167	Mar-18	206.8	5.2	19.4	12.2	27	100168
Row168	Mar-18	215.4	23.6	57.6	17.1	21	100169
Row169	Mar-18	284.3	10.6	6.4	15	20	100170
Row170	Apr-18	50	11.6	18.4	8.4	41	100171

Figure 37 Imported Sales Data 1


File Table - 3:11 - CSV Reader (Sales Data 2)

File Edit Hilite Navigation View

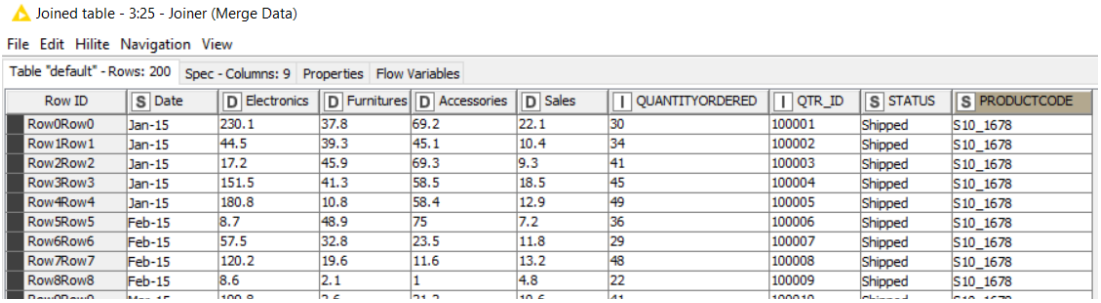
Table "default" - Rows: 200 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	STATUS	QTR_ID	PRODU...
Row162	Shipped	100163	S12_1099
Row163	Shipped	100164	S12_1099
Row164	Shipped	100165	S12_1099
Row165	Shipped	100166	S12_1099
Row166	Shipped	100167	S12_1099
Row167	Shipped	100168	S12_1099
Row168	Shipped	100169	S12_1099
Row169	Shipped	100170	S12_1099
Row170	Shipped	100171	S12_1099
Row171	Shipped	100172	S12_1099
Row172	Shipped	100173	S12_1099

Figure 38 Imported Sales Data 2

First and foremost, the Joiner node must be able to read the unique data/primary key data column, which is the QTR ID column in both CSV Reader nodes, in order to merge the two sets of data together. This demonstrates that the Joiner node has been successfully executed. Following that, the Column Rename node was successfully executed, resulting in the renaming of two columns: QTR ID to Customer ID and Product Status to Status, respectively. This column renaming process was successful in changing the names of the respective columns without

encountering any syntax errors. Furthermore, the Duplicate Row Filter node was successful in removing all of the duplicated data as well as any missing values from the columns and rows that were set.



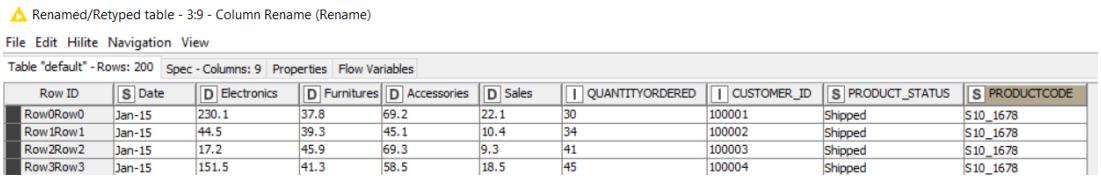
Joined table - 3:25 - Joiner (Merge Data)

File Edit Hilite Navigation View

Table "default" - Rows: 200 | Spec - Columns: 9 | Properties | Flow Variables

Row ID	Date	Electronics	Furnitures	Accessories	Sales	QUANTITYORDERED	QTR_ID	STATUS	PRODUCTCODE
Row0Row0	Jan-15	230.1	37.8	69.2	22.1	30	100001	Shipped	S10_1678
Row1Row1	Jan-15	44.5	39.3	45.1	10.4	34	100002	Shipped	S10_1678
Row2Row2	Jan-15	17.2	45.9	69.3	9.3	41	100003	Shipped	S10_1678
Row3Row3	Jan-15	151.5	41.3	58.5	18.5	45	100004	Shipped	S10_1678
Row4Row4	Jan-15	180.8	10.8	58.4	12.9	49	100005	Shipped	S10_1678
Row5Row5	Feb-15	8.7	48.9	75	7.2	36	100006	Shipped	S10_1678
Row6Row6	Feb-15	57.5	32.8	23.5	11.8	29	100007	Shipped	S10_1678
Row7Row7	Feb-15	120.2	19.6	11.6	13.2	48	100008	Shipped	S10_1678
Row8Row8	Feb-15	8.6	2.1	1	4.8	22	100009	Shipped	S10_1678

Figure 39 Merged Sales Data 1 and Sales Data 2



Renamed/Retyped table - 3:9 - Column Rename (Rename)

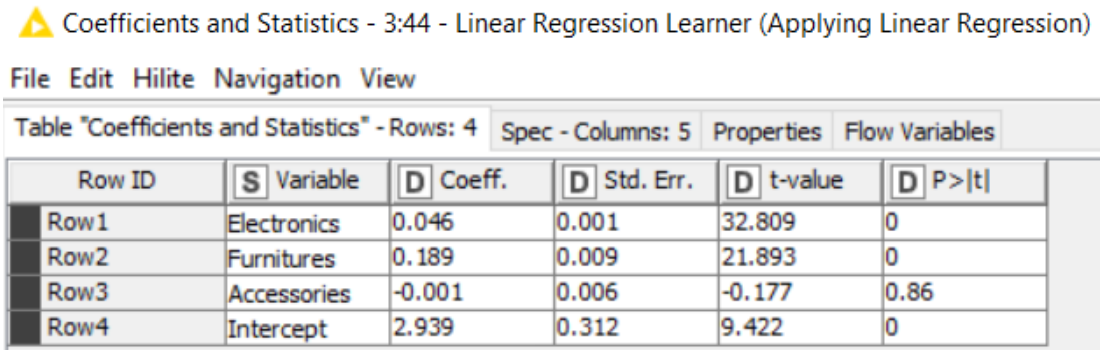
File Edit Hilite Navigation View

Table "default" - Rows: 200 | Spec - Columns: 9 | Properties | Flow Variables

Row ID	Date	Electronics	Furnitures	Accessories	Sales	QUANTITYORDERED	CUSTOMER_ID	PRODUCT_STATUS	PRODUCTCODE
Row0Row0	Jan-15	230.1	37.8	69.2	22.1	30	100001	Shipped	S10_1678
Row1Row1	Jan-15	44.5	39.3	45.1	10.4	34	100002	Shipped	S10_1678
Row2Row2	Jan-15	17.2	45.9	69.3	9.3	41	100003	Shipped	S10_1678
Row3Row3	Jan-15	151.5	41.3	58.5	18.5	45	100004	Shipped	S10_1678

Figure 40 Column Rename Configured

Following its successful execution, the Data Explorer Node followed suit, running without any errors as it read the whole data and displayed it in numeric and nominal form to provide the user a sense of how linear regression might be applied later. It chooses all the needed columns such as Accessories column, Electronics column, and Furniture column data to prepare for linear regression application once the column filter is applied once the data explorer node has been selected. The Linear Regression node was able to complete its task effectively since the column filter gave sufficient data for it to comprehend the columns and calculate the coefficient values for Accessories, Electronics, and Furniture, among other things.



Row ID	Variable	Coeff.	Std. Err.	t-value	P>|t|
Row1	Electronics	0.046	0.001	32.809	0
Row2	Furnitures	0.189	0.009	21.893	0
Row3	Accessories	-0.001	0.006	-0.177	0.86
Row4	Intercept	2.939	0.312	9.422	0

Figure 41 Linear Regression Coefficient Table

As a result, the coefficient values in a table were transferred to Power BI datasets for analysis. While another automation path for the Rule Engine node was successfully executed to create a new column for Product status derived from the Status column, which successfully classified column data named 'Shipped,' 'Canceled,' and 'Resolved' as Completed, while column data named 'Disputed,' 'On Hold,' and 'In Process' are all marked as Incomplete, another automation path for the Rule Engine node was successfully executed to create a new column for Product status derived To finish it all off, the output section worked flawlessly, as it successfully exported all of the configured datasets to Power BI, using two separate Send to Power BI nodes, with the top node being responsible for overall data and the bottom node being responsible for exporting coefficient value tables configured from linear regression node to Power BI datasets. In conclusion, all of the KNIME function nodes were configured and successfully executed, allowing us to go on to the data visualization step. He was able to successfully implement the KNIME workflow, and as a result, I am confident in the efficacy of the KNIME automation that I built.

```
// enter ordered set of rules, e.g.:
// $double column name$ > 5.0 => "large"
// $string column name$ LIKE "*blue*" => "small and blue"
// TRUE => "default outcome"
$PRODUCT_STATUS$ = "Shipped" OR $PRODUCT_STATUS$ = "Cancelled" OR $PRODUCT_STATUS$ = "Resolved" => "Complete"
TRUE => "Incomplete"
```

Figure 42 Rule Engine Automation

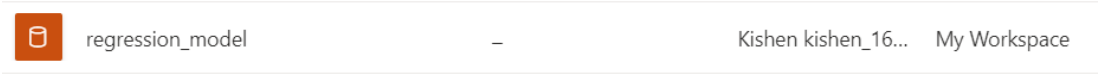


Figure 43: Exported Regression Model to Power BI Datasets

4.3 Power BI Dashboard and Sales Forecasting

Following confirmation that the Power BI dashboard was functional and executable by the deployment completed with the financial analyst, I was satisfied with the results. The Power BI dashboard I created was divided into five key sections: the What-If parameter for each product, the revenue trend, the sales performance, the sales table, and the sales forecasting. The What-If parameter for each product was divided into two sections: the average and the maximum. The dashboard's five components were all functioning well, and the finance analyst was a big fan of the overall design of the dashboard. In the first place, the revenue trend chart depicts the previous year's sales of accessories, electronics, and furniture for the years 2015, 2016, 2017, and 2018.

Revenue trends are displayed to help finance analysts comprehend the revenue curve based on the products that are offered for sale. Sales performance illustrates the same previous sales of accessories, electronics, and furniture in the years 2015, 2016, 2017, and 2018 in a clustered column chart graph to provide a clear visual representation of the profits of each product in the following years: As a result of this, each product's sales table displays the actual quantity as well as the overall amount in currencies for each product based on the year in which it was sold. Finally, we get at the most crucial and fascinating component of the dashboard: the navigation bar. They were permitted to enter any precise value based on their assumptions about the accessories, gadgets and furniture. The What-If parameters box was created.

By entering the assumed value in the box, predicted sales in the future will be displayed in the predicted return on investment card and sales forecasting clustered column graph to show the amount of sales that could be achieved in the future, or in another word, sales forecasting based on the value entered by the finance analyst in each product What-If parameter box will be displayed. To view sales projecting more than a thousand-ringgit profit, for example, they can alter the dot line on the clustered column graph to provide a baseline for comparison. After that, users can key in the required value for each item. When you use the What-If parameter boxes, you can see how the predicted return on investment card and sales forecasting clustered column graph change. The predicted return on investment card will display

the amount of profit they will receive, while sales forecasting can ensure that the predicted sales value reaches or exceeds the baseline value and beyond. As illustrated in the figures 40 and 41 linked below, the output display varies depending on the input provided.

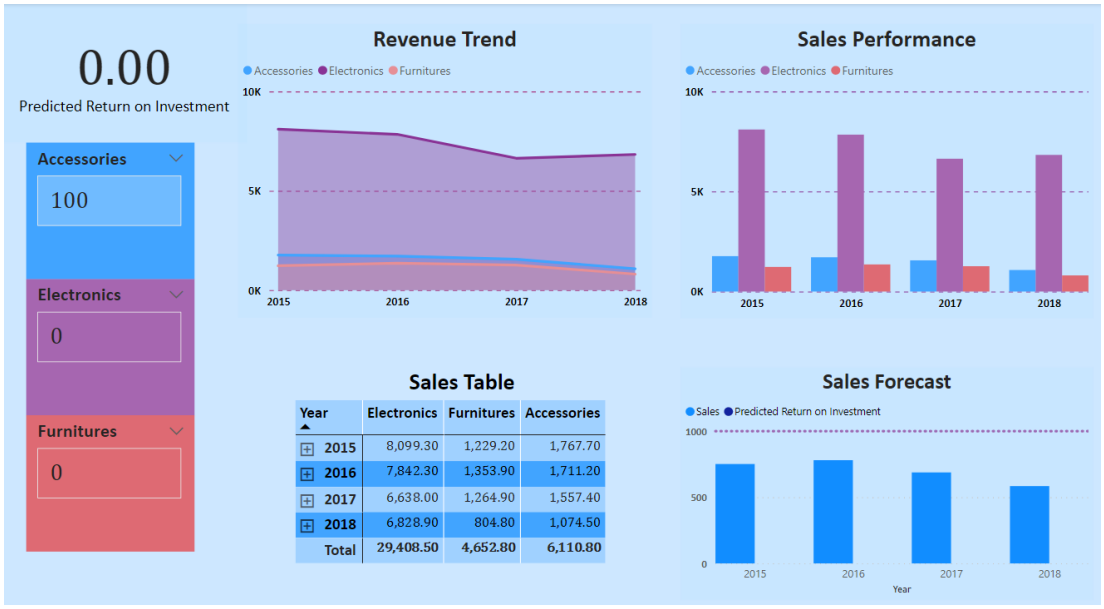


Figure 44 Power BI Dashboard Without Input

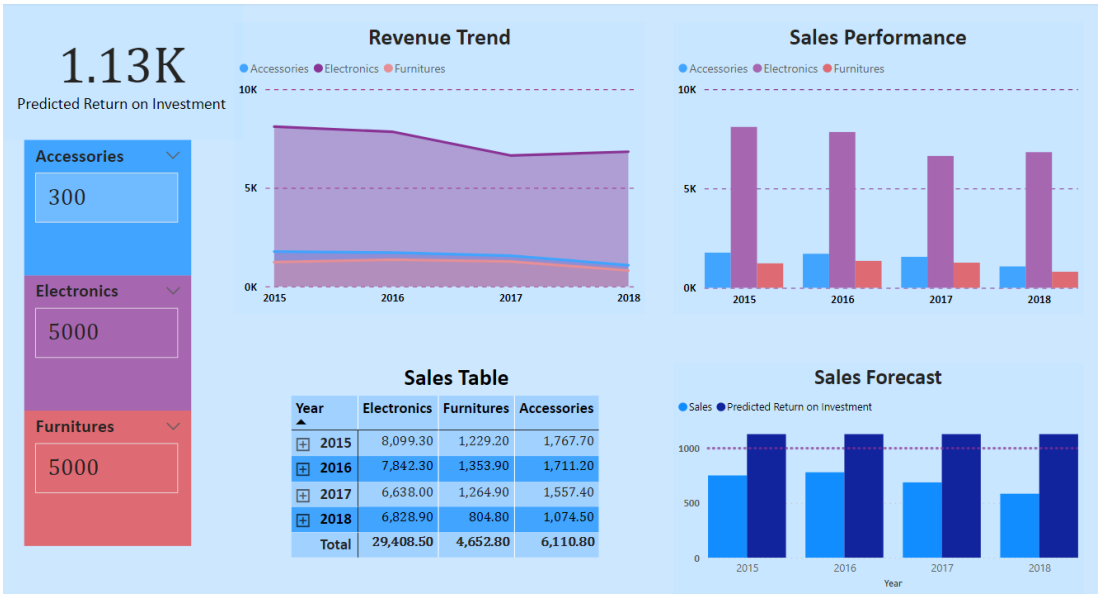


Figure 45 Power BI Dashboard with Desired Input

The future marketing and sales of e-commerce products can be improved by financial analysts, allowing them to boost the revenues and profits of their company or organization. Marketing and sales teams may increase the accuracy of their projections by harnessing the insights provided by the Power BI dashboard, which can account for seasonal demand, product promotions, slow-moving goods, causative variables, outliers, and other factors. When the appropriate information is available to support the underlying analysis, the forecasting effort can be significantly reduced, and the dependability of the forecast can be significantly increased. Using this dashboard, decision-makers can get ready for access to information that presents a complete portrayal of the company's sales performance. Easy and direct access to past sales information helps to improve forecast accuracy while also allowing for better and faster procurement and inventory decision-making.

CHAPTER 5

CONCLUSION & FUTURE WORK

5.1 Conclusion

I identified numerous essential concepts of using the KNIME Analytic platform and Power BI Visualization business intelligence tool in order to meet the requirements and construct the project based on my discoveries. I am overjoyed that I was able to complete this project. The goal of this project is to use ETL technologies such as KNIME Analytic and Power BI to bridge the gap between data transformation and data analysis. These are the most relevant tools that I discovered through my research and that I was able to complete. After conducting research, I was able to successfully apply my knowledge and expertise to design an automated workflow in the data wrangling process, which I believe will play a significant part in the Finance sector in all large, medium-sized, and even small firms throughout the world.

This will assist finance analysts in understanding the significance of technology adaptation in the finance sector, as their traditional job scope and responsibilities will be transitioning to information technology and data analytic adaptation as a result of the transition to information technology and data analytic adaptation. Future or sooner technology adoption in their organization will see most audit or finance tasks databases being sent to Power BI for automated visualization after the raw data has been automated to structured data through the KNIME Analytic tool or other ETL tools, according to the organization's technology adoption plan. The use of ETL tools in any organization allows a corporation to study quarterly reports, many years' worth of balance sheets, measure ROI, and compare the company's pattern to that of their competitors to identify their competitive advantage. Manually searching the data from a variety of sources might result in uncertainty in the information.

Additionally, when entering hundreds of data points repeatedly, human error is unavoidable. Data extraction and collecting technologies make it possible to automate the process in a short period of time, allowing the organization to make better-informed business decisions. The amount of data generated by every financial organization is enormous. We can see how critical it is to maintain control of one's finances in order to run a successful business. Using ETL technologies, businesses would be able to greatly improve the architecture of their financial models by incorporating real-time financial data extraction and management into their workflows. This also helps to expedite the financial analysis process while also lowering the likelihood of fraud. Implementation of artificial intelligence and machine learning technologies in order to extract data and create one-of-a-kind solutions tailored to a customer's individual requirements.

By utilizing ETL technologies, any firm may stay one step ahead of their competitors. On addition, my research into machine learning was perfectly suited to leveraging that technique by applying linear regression in the KNIME platform to show data for future forecasting processes, which was a fantastic fit. Because of this, the project's goal is to transform unstructured data into the desired outcome and visualize it according to the needs of its target user, Malaysian finance analysts working for a small or medium-sized business or even a large multinational corporation, by utilizing modern automation technology to replace traditional audit tasks.

5.2 Future Work

Given that my research isn't finished yet, I'm still looking for more convenient resources that can be applied to the finance analyst's task or to the finance industry. The research on developing the automation process using KNIME, Power BI, and machine learning is still in progress, as the workplaces' adaptation to technology is changing daily. Artificial intelligence and machine learning in finance are topics I'd like to learn more about because they're playing an increasingly important role in transforming the way people interact with data. From credit decisions to quantitative trading and financial risk management, artificial intelligence (AI) is assisting the financial industry in streamlining and optimizing procedures.

There are a few things that can be included in the KNIME automation itself, such as machine learning function processes, that will allow you to automate unstructured data. In the finance and accounting industries, advancements in artificial intelligence have the potential to transform the industry by removing tedious tasks from the hands of human finance professionals, allowing them to focus on higher-level and more lucrative analysis and counselling for their clients.

Organizations, on the other hand, are hesitant to integrate artificial intelligence into their workforce because of uncertainties surrounding the business case and return on investment. Artificial intelligence-based technologies will completely automate finance job scopes such as accounting tasks such as tax preparation and payroll, auditing, and banking. This disruption in ways never imagined will bring both enormous opportunities and significant challenges to the accounting industry. Artificial intelligence (AI) has the potential to increase both productivity and the quality of outputs while also allowing for greater transparency and audibility. Apart from providing a broad range of possibilities and alleviating the regular responsibilities of the finance team, artificial intelligence will also save time and provide accounting professionals with the opportunity to conduct critical research on a variety of topics, which will be extremely beneficial.

Aside from that, artificial intelligence will easily forecast accurate financial statements. The underlying concept is that accounting professionals would be able to predict future data based on past data/records if they used machine learning. To that end, I'm interested in gaining a deeper understanding and gaining more hands-on experience in how machine learning applications and artificial intelligence services can help finance professionals complete their regular responsibilities at a faster rate. Furthermore, there are opportunities to automate data visualization based on the database and datasets that have been uploaded to Power BI, if desired. These suggestions will be considered for inclusion in my final year project because of this, with the goal of achieving the best possible result. Remember to set up the best autonomous processes for finance analysts in order to reduce their daily audit workloads and eliminate unnecessary stress caused by worrying about implementing current accounting technologies around the world, particularly here in Malaysia.

REFERENCE

- Singh, N. (2020, July 27). *Advantages and Disadvantages of Linear Regression*. OpenGenus IQ: Computing Expertise & Legacy. https://iq.opengenus.org/advantages-and-disadvantages-of-linear-regression/.

- Dickey, G., Blanke, S., & Seaton, L. (2019, June 18). Machine Learning in Auditing. Retrieved from https://www.cpajournal.com/2019/06/19/machine-learning-in-auditing/

- Davidiseminger. (n.d.). Tutorial: Build a Machine Learning model in Power BI - Power BI. Retrieved from https://docs.microsoft.com/en-us/power-bi/connect-data/service-tutorial-build-machine-learning-model

- *How Digital Transformation Enables Modern Accounting: BlackLine Magazine*. BlackLine. (n.d.). https://www.blackline.com/blog/digital-transformation-enables-modern-accounting/.

- *Combining the Power of KNIME and PowerBI for Automated Sentiment Analysis*. KNIME. (n.d.). https://www.knime.com/solutions/success-story/automated-sentiment-analysis.

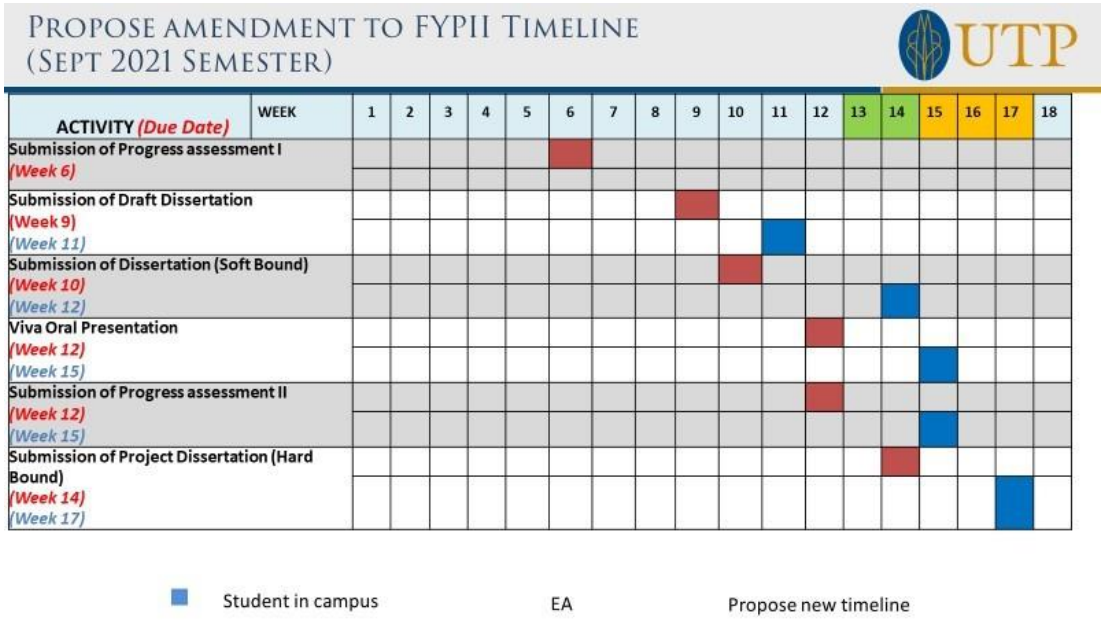
- Cohn, M. (2016, August 11). *Internal Audit Leveraging Data Visualization Tools*. Accounting Today. https://www.accountingtoday.com/opinion/internal-audit-leveraging-data-visualization-tools.

- Editor@HostBooks, P. (2020, March 2). *How AI will impact the accounting and finance industry?* My blog. Retrieved from https://www.hostbooks.com/us/blog/how-ai-will-impact-the-accounting-and-finance-industry/.

- Sundararajan, A. (2021, May 18). *Why are ETL tools necessary for financial data analysis and reporting?* Blog | Xtract.Io. https://xtract.io/blog/why-are-etl-tools-necessary-for-financial-data-analysis-and-reporting/

APPENDICES

i. Propose Amendment to FYPII Timeline



ii. Example KNIME Linear Regression output for reference

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
	-0.0	NaN	NaN	NaN
	-4.00E17	NaN	NaN	NaN
	-0.0	NaN	NaN	NaN
	1.1138	NaN	NaN	NaN
	0.2242	NaN	NaN	NaN
	-0.1355	NaN	NaN	NaN
	-0.2546	NaN	NaN	NaN

iii. Final Year Project Poster