13th May 2018

Technical Report

Cashless Payment Analyst

Siuk Ching Tan

16103653

BSC IN COMPUTING

DATA ANALYSTICS

2017/2018

National College of Ireland

Table of Contents

Executive Summary

This project was completed for the BSc (Honours) in Computing course at National College of Ireland. The objectives of the project is to analyse the frequency of cash and cashless payment in Europe countries. The result of this project ascertain the gap of cash and cashless payment spread over Europe countries. Main priority parts are identifies the frequency of each country using cashless payment and figure out the correlation relationship between payment frequency with population of Ireland and number of tourist to Ireland.

Prediction has also involved in this project by using Data Mining Machine Learning techniques. It based on some attributes such as payment method, time period, value of transactions and so on to make predict of payment method in Ireland.

The datasets for this project are retrieved from European Central Bank (ECB) and Central Statistics Office (CSO) and store in a database which connect to RStudio for analysis purpose. R programming language was mainly used in this project to generate results and Tableau has been exploit for visualization.

1 Introduction

1.1 Background

This analyst is to identify the trends of using cashless payment. This is idea is come with when I am in a long queue at grocery shop. I figure out the spending time of payment process is much different when people used cash and card. Most of the people will only take out the cash when they had been noticed by cashers the amount of the things they buy. It would make a long queue especially at the peak hour. Due of this issue, I want to observe and analyst people are likely to pay with cash or cashless and how people frequently used card to make payment also which area they would like to spent with card. For example, people may prefer to pay with card when purchase a high cost product such as washing machine because some of the shops are allows the consumer to make instalment for few months.

Moreover, cashless payment is highly praise nowadays. Cashless payment describes not only credit/debit card payment but also used virtual cash such as Bitcoin, Paypal or mobile application to make payment. Through cashless payment, it will increase the efficiency of payment process and indirectly improve the quality of life for people to spend their time in more meaningful ways [7].

In Sweden nowadays is the most cash-free society in the world. The reason why Sweden government encouraged their citizen used cashless to make payment is because it can decreases the budget for printing money and reduced crime rate. Even though the homeless vendors that sold magazine in the capital of Sweden are using electronic payment to deal their business [6].

Cashless payment would be a tendency within this few years, so that my project is work to generate an accurate and reliable result to prove the frequency of cashless payment.

1.2 Aims

The aims of Cashless Payment Analyst are to identify the trends of the public make payment without using cash. Cashless payment can be used to reduce the corruption of cash medium takes place and reduce the burden of the cost of printing currency. While using Cashless Payment, the movement of the money is able to track completely how the money has been used.

For Cashless Payment Analyst, it is able to specify the population of people used cashless to payment and predict the trends of using cashless payment for next few years. It will bring a new transformation for the way of consume and improve the economic consumption.

1.3 Technologies

R Studio

In this project, R Studio as an open-source integrated development environment and perform statistical computing and graphical in R Language. R Studio operates as retrieval data from Ms Excel and store as a dataset in RStudio. It will work with KDD methodology to select, process, transform and visualize the data and information.

Microsoft Excel

Ms.Excel is a spreadsheet tools used to calculate, provides graphic diagram and pivot table. Dataset represents in comma separated values file(.csv) format and import into MySQL database for analysis.

Tableau

Tableau is a data visualization tools that used to preform data in visualize format which allow user easy to understand the meaning of the results that carry out in the data. It brings a lot of convenient for analyst and end user to implement the visualization and absorb all findings of the data.

1.4 Structure

The scope of this project is to develop a system that able to generate an accurate report about cashless payment and predict the flows of public using cashless payment in future.

To achieve this project completely and accurately, a few steps need to be take such as meet the project functional and non-functional requirements, understand what to present on the interfaces and comprehend why those interfaces is necessary to show, clearly describe the system architecture and flow of system evolution.

In the functional requirements, we will split a system into few functions and provide a use case diagram for each use case. Every use case would explain the scope of the use case, description, precondition, activation, main flow, alternate flow, exceptional flow, termination and post condition. Non-functional requirements would describe the minor features involved in the system.

The interfaces of this project would be generate by collecting data, import data, clean data and run the script of data to show plot or diagram of report analysis. The ways we produce was find the relevant dataset from open source, used the knowledge about R and Python programming to remove any irrelevant value and write the script to generate results.

Moreover, KDD methodology is the architecture of this project which throughout few steps such as selection, processing, transformation, data mining and interpretation or evaluation to generate the final reports. Other than that, system evolution is used to describe how this project evolves over time to meet the project requirements.

1.5 Definitions, Acronyms, and Abbreviations

Dropbox: A cloud-based storage that allows people uploads their documents or folder into cloud storage. It able to sharing with other via send link or add email.

GitHub: A web-based hosting service, offers storage to store source code and able to keep track on changes of code.

Google Drive: A cloud-based storage, it provides a personal data storage that can be access anytime throughout any digital devices.

KDD: Knowledge Discovery in Databases is a methodology that describes the process of data source transforms to information and become a useful knowledge.

RStudio: A programming application that used to implement data for analyst graphic and statistical computing.

European Central Bank: The central bank for the euro and administers monetary policy of the Eurozone which consists of 19 EU member states and one of the largest currency areas in the world.

Central Statistics Office: Government body that compiles official statistics in Ireland.

Machine Learning: A field of computer science that uses algorithm and statistical techniques to allow computer system get the extra information that hidden in the data. It able to learning the value of data and make an accuracy prediction.

2 System

2.1 Requirements

Requirements specifications for this project are divided into functional and non-functional requirements. Functional requirements represent a few major process of this project such as import data, clean data, generate statistic and show prediction for future. These processes will show useful information about cashless payment. In the other way, non-functional requirements represent the features that are not major functions but might indirect effect this system. For examples, availability, recover and security are the non-functional requirements in this project.

2.1.1 Data Requirement

There are one main data and two small data involves in this project.

The main dataset from ECB is an open source datasets from National Central Banks in European Union. Datasets include statistics on access and usage of payment service and terminals. Currency has be counted as EUR for euro area Member States and domestic currencies (non-euro area Member States). All data are collected and compiled by EU NCBs. Reference period of data was recorded from 2000 to 2016.

Table 1: Payment Statistics Dataset [8]

Name	Type	Description
MainKEY	varchar(100)	Primary Key
FREQ	varchar(10)	Frequency Dimension
REF_AREA	varchar(10)	Country codes
PSS_INFO_TYPE	varchar(10)	PSS information type
PSS_INSTRUMENT	varchar(10)	PSS instrument
PSS_SYSTEM	varchar(10)	PSS entry point
DATA_TYPE_PSS	varchar(10)	PSS data type

COUNT_AREA	varchar(10)	Counterpart area
COUNT_SECTOR	varchar(10)	Counterpart sector
CURRENCY_TRANS	varchar(10)	Currency of transaction
SERIES_DENOM	varchar(10)	Series denominat/ spec calcul
TIME_PERIOD	varchar(10)	Year of time period
OBS_VALUE	Numeric(10)	Observation value
OBS_STATUS	varchar(10)	Observation status
OBS_CONF	varchar(10)	Observation confidentiality
OBS_PRE_BREAK	varchar(10)	Pre-break observation value
OBS_COM	varchar(10)	Observation comment
TIME_FORMAT	varchar(10)	Time Series-level attribute
BREAKS	varchar(10)	Time Series-level attribute
COLLECTION	varchar(10)	Collection indicator
COMPILING_ORG	varchar(10)	Compiling organisation
DISS_ORG	varchar(10)	Data dissemination orgranisation
DOM_SER_IDS	varchar(10)	Domestic series ids
PUBL_ECB	varchar(10)	Source publication (ECB only)
PUBL_MU	varchar(10)	Source publication (Euro area only)
PUBL_PUBLIC	varchar(10)	Source publication (public)
COMPILATION	varchar(10)	Sibling-level attribute
DECIMALS	numeric(10)	Sibling-level attribute
METHOD_REF	varchar(10)	Methodology reference
NAT_TITLE	varchar(10)	National language title
SOURCE_AGENCY	varchar(10)	Sibling-level attribute

TITLE	varchar(255)	Sibling-level attribute
TITLE_COMPL	varchar(255)	Title complement
UNIT	varchar(10)	Sibling-level attribute
UNIT_MULT	varchar(10)	Unit multiplier

Another data collect from Central statistics Office (CSO) was Enumerated Population 1926 to 2016 by Census Year. It shows every five year of population in Ireland.

Table 2: Population in Ireland Dataset [2]

Name	Type	Description
Year	integer(4)	Number of year
Population	numeric(10)	Number of population

Third dataset was also collect from CSO with the Number of Overseas Visitors to Ireland and Number of Visits Abroad by Irish Residents.

Table 3: Travel Dataset [3]

Name	Type	Description
Year	integer(4)	Number of year
NumOverseasVisitorstoIreland	numeric(10)	Number of oversea visitors visit to Ireland
NumofVisitsAbroadbyIrishResidents	numeric(10)	Number of Irish residents travel abroad

Note：Second and third datasets were used to check correlation with main datasets.

2.1.2 Functional requirements

The functional requirements represent the ways users interact with the system. In this system, there are different roles in the requirements such as user and system. User is the person who works on the system. System plays an important role that has to implement the data.

2.1.2.1 Use Case Diagram

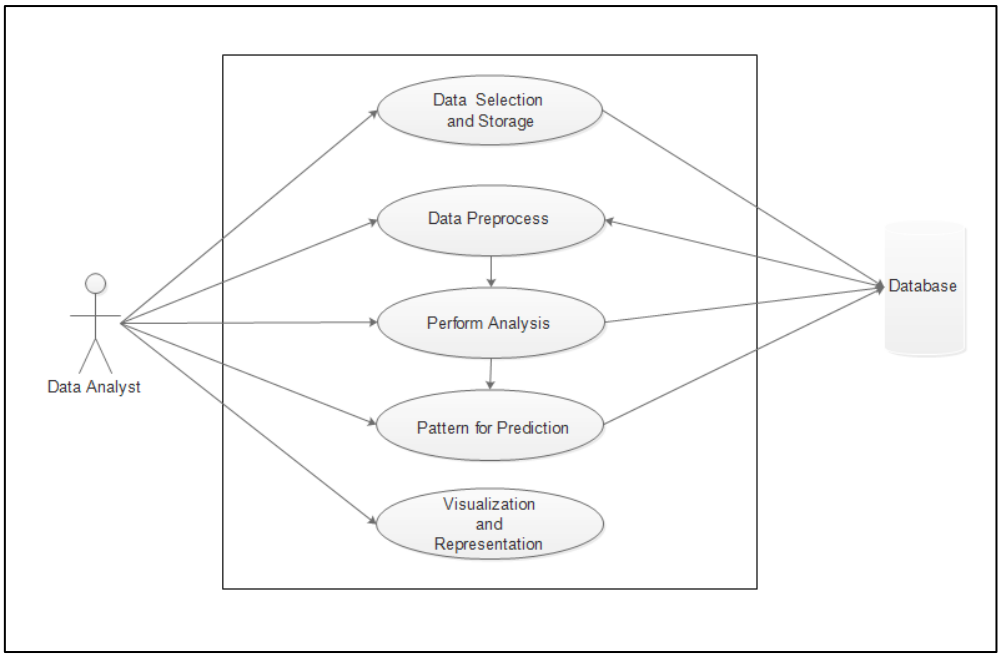


Figure 1 : Use Case of the project

2.1.2.2 Requirement 1: Data Selection and Storage

Description & Priority

This is the first step that extract or collect all datasets and store in database.

Use case

Score

The scope of this use case is to collect all the relevant information and import into data storage application.

Description

This use case describes the flow of import the data into data analyst application. First step is download datasets from open source, continue with create storage and import data to database.

Use Case Diagram

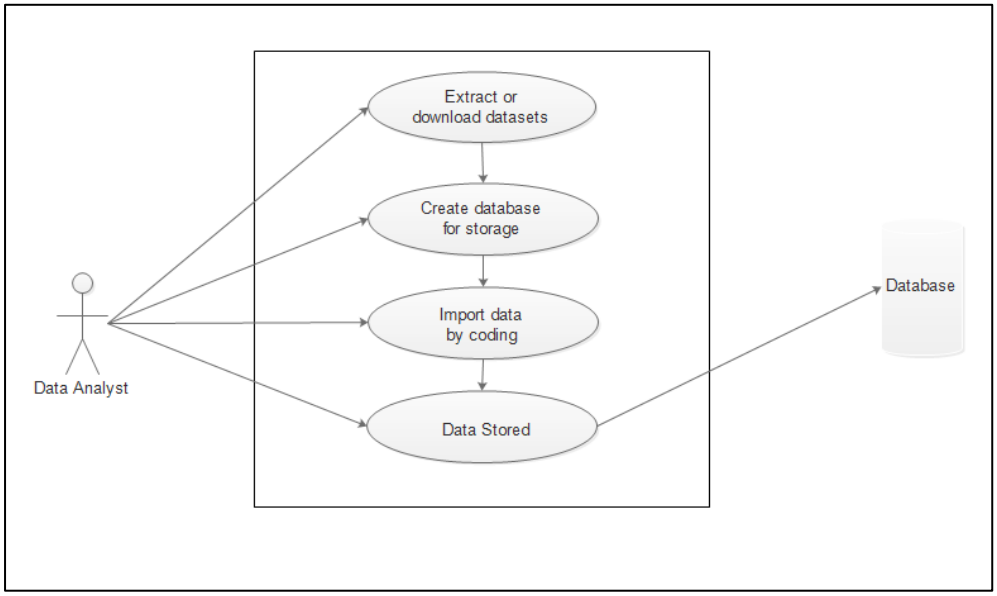


Figure 2 : Use Case of Data Selection and Extraction

Flow Description

Precondition

The user must able to access into the particular website to extract or download the relevant data.

Activation

This use case starts when an analyst access into relevant website.

Main flow

1. The <Analyst> accesses into open source website and downloads the datasets.
2. The <Analyst> creates a new storage to store datasets.
3. The system accepts the request to create a new storage.
4. The <Analyst> imports datasets into storage by coding.

5. The system stores a new datasets in database.

Exceptional flow

1. Particular website disable to access.
2. Failed to create storage.
3. The datasets are corrupted.
4. Failed to import data.

Termination

The system presents the next use case when storage is created and datasets are imported.

Post condition

The datasets in storage is ready for analyze.

2.1.2.3 Requirement 2: Data Preprocess

Description & Priority

This requirement involves the process of cleaning, integrating and transforming the data to a depth analysis stage.

Use Case

Scope

Preprocessed data improves the ability for various type of analysis.

Description

This use case describes the process of eliminates useless data and integrate with more relevant information.

Use Case Diagram

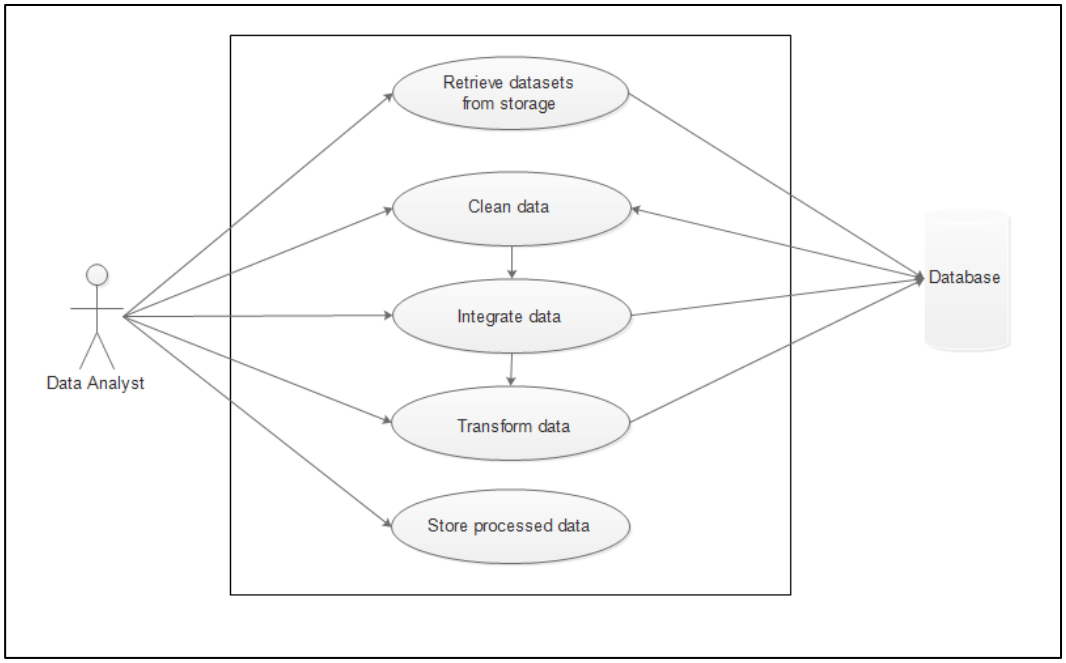


Figure 3 : Use Case of Data Preprocess

Flow Description

Precondition

The database is accessible and data are retrievable.

Activation

This use case starts when an analyst access into database storage.

Main flow

1. The <Analyst> retrieves all datasets from database.
2. The data is cleaned and integrated.
3. Transformed data based on the analysis requirement.
4. The <Analyst> store processed data into database.

Exceptional flow

The data failed to retrieve.

Termination

Preprocess data terminates when data is stored in database.

Post condition

The datasets is ready to perform analysis.

2.1.2.4 Requirement 3: Perform Analysis

Description & Priority

This requirement is the priority requirement use to generate the analysis and fulfill the project objectives.

Use Case

Scope

Retrieved preprocess data to perform various analysis by using R programming script. Regression and correlation calculation will be perform in this stage.

Description

This use case describes the steps of analysis will be carrying out.

Use Case Diagram

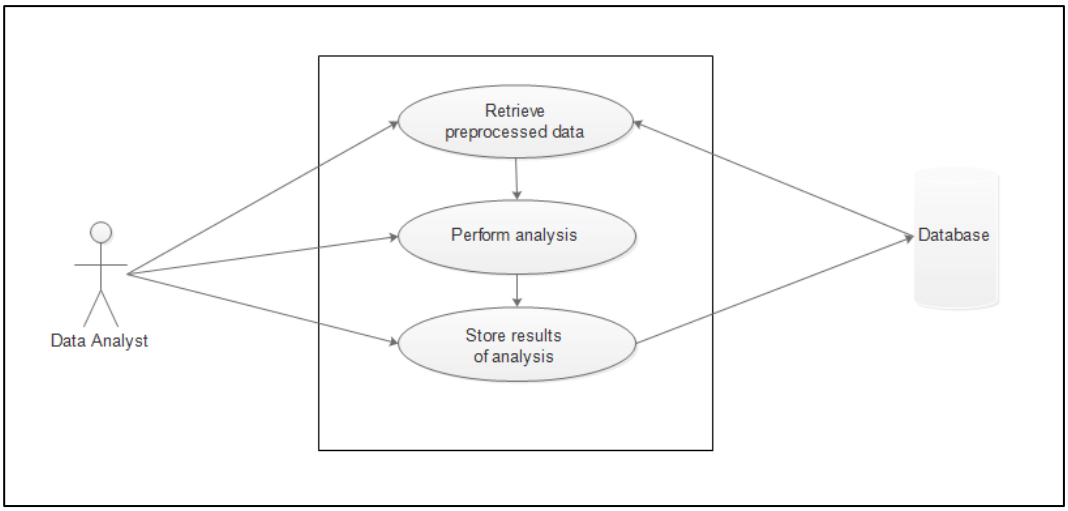


Figure 4 : Use Case of Perform Analysis

Flow Description

Precondition

Data analyst able to access preprocessed data from database.

Activation

This use case starts when an <Analyst> calls data from database storage.

Main flow

1. The < Analyst > retrieved preprocessed data from database.
2. Some calculation will be perform before analyse data.
3. Data is analysed using relevant script.
4. Results of analysis are stored in database.

Exceptional flow

Analyse of data is not working properly.

Termination

This use case terminated when analyse is accomplish.

Post condition

The results of analysis can be used for further analysis.

2.1.2.5 Requirement 4: Pattern for Prediction

Description & Priority

This requirement is use to predict the trend of using cashless payment in future.

Use Case

Scope

The scope of this use case is to generate the prediction for this project and the final result will be stored.

Description

This use case describes the process of prediction data.

Use Case Diagram

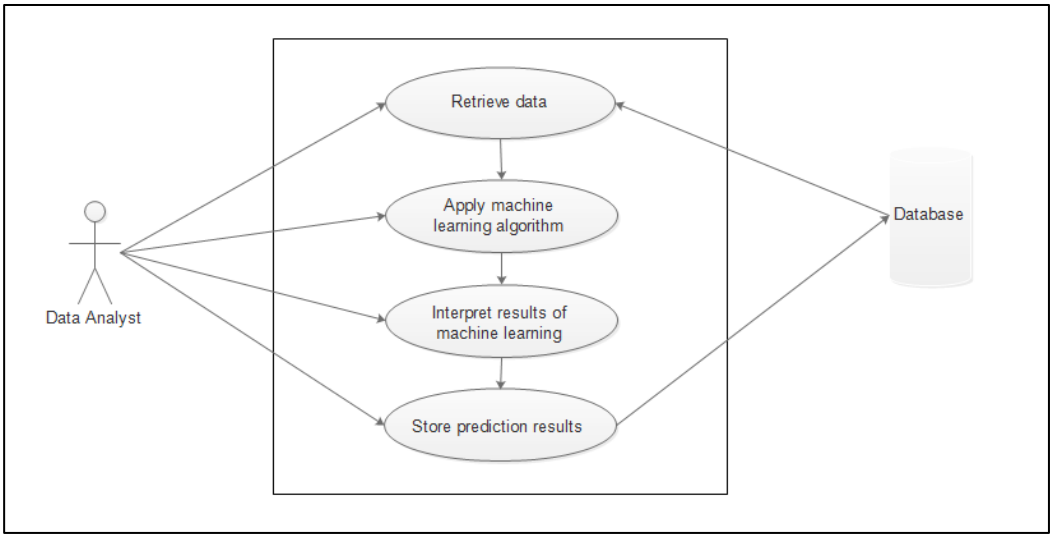


Figure 5 : Use Case of Pattern for Prediction

Flow Description

Precondition

Data is accessible.

Activation

This use case starts when data analyst calls the data.

Main flow

1. The <Analyst> retrieves the data.
2. Apply decision tree pattern into selection data.
3. Results are analysed and interpreted.
4. Prediction results stored.

Exceptional flow

Decision tree may not performed well due to low correlation attributes.

Termination

The system terminated after the results of prediction stored.

2.1.2.6 Requirement 5: Visualisation and Representation

Description & Priority

Visualisation and representation is the last requirement for this project which to display the results in easy to understand and meaningful way.

Use Case

Scope

Visual diagrams shows the results in a simple and understandable.

Description

This use case describes the process of visualisations and representations.

Use Case Diagram

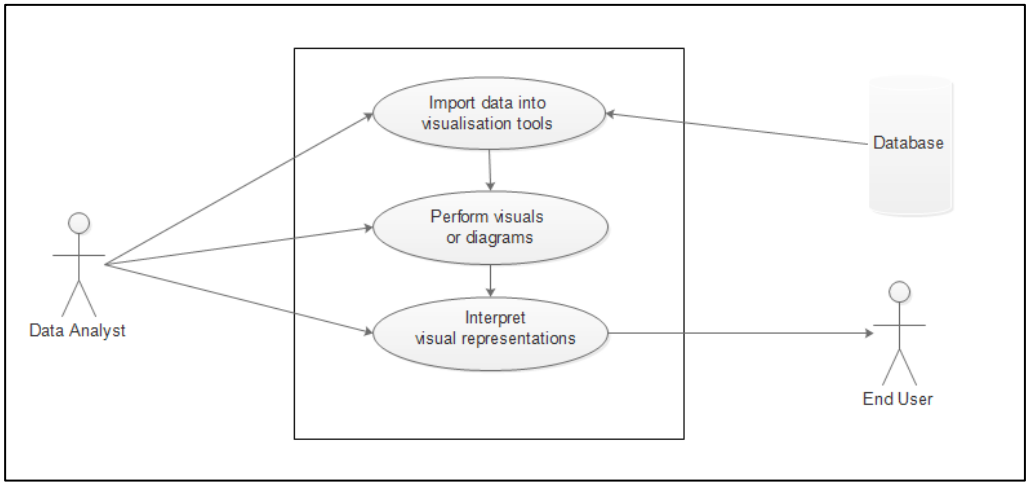


Figure 6 : Use Case of Visualisation and Representation

Flow Description

Precondition

Data is accessible and all analysis should be accomplish in this stage.

Activation

This use case starts when data is imported into visualisation tools.

Main flow

1. Data is imported into visualisations tools.
2. Design various visual or diagrams of results.
3. All diagrams able to present for end user.

Exceptional flow

May failed to imported data into data visualisation tools.

Termination

This use case terminates when all visualisations and representations are accomplish.

2.1.3 Non-Functional Requirements

1. Performance/Response time requirement

High performance and response time are not necessary in this project, it might depend on speed of internet when downloading the datasets or the user' selection on the system. The response time of analysis process may take shorter and prediction may take few more second to complete it.

2. Availability requirement

Data will be available all the time in the system and the output can be export to a file document for storage purpose.

3. Recover requirement

To avoid the occurrence of system errors or server down, this project is able fully recovery all data. Cloud based website such as Dropbox, Google Drive and GitHub can be used to store all data and report file as a backup document.

4. Security requirement

The data sources for this project were taken from open source – European Central Bank and Central Statistics Office so that does not have any security issues. Data in this project does not involved any personal information and working in personal laptop. The system can only be accesses after decrypted personal laptop.

5. Maintainability requirement

No maintainability for this project but may need some modification for the script of database to generate different type of output.

6. Extendibility requirement

This project can be extending for future if collects the data and regenerate the analyst again but there is not necessary to do that.

7. Resource utilization requirement

The project's output could be utilized for public to understand the trend of cashless payment has been used and there are many intangible advantages of using cashless payment such as decrease the number of crime rate.

2.2 Design and Architecture

To carry out this analyst, I do some research of population of cashless payment and whose countries are pursuing cashless payment. I also figure out European Central Bank have bulk of open dataset that are related to my analyst. It reduces the difficulty for me to form the data from various website. For the main technical approach in this project was KDD methodology [4].

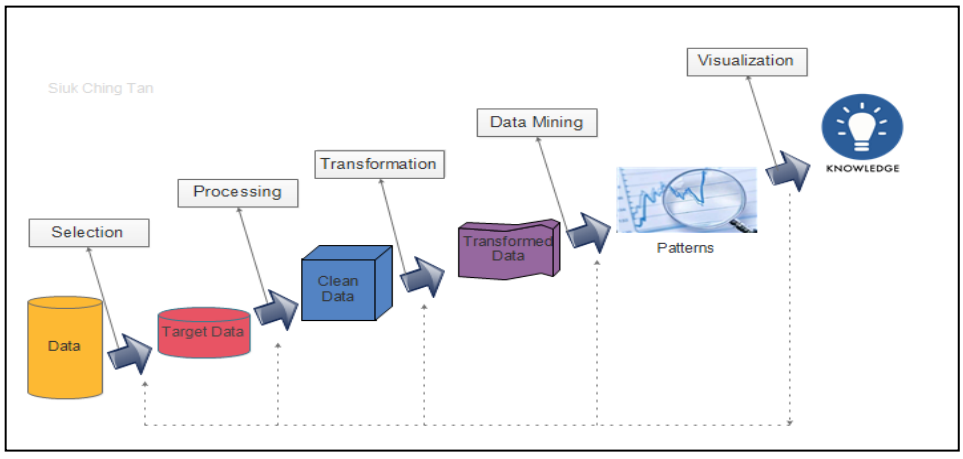


Figure 7 : KDD Methodology

In this project, we used Knowledge Discovery in Databases (KDD) as system architecture. KDD is a process of discovering useful knowledge from a collection of data. It represents

the process of finding knowledge in data and highlights the different level of data mining methods. Useful knowledge can be taken throughout the process of selection, processing, transformation, data mining and visualization.

Selection

In a pre-defined stage, a storage was created to store the relevant datasets that extract or retrieve from open source such as European Central Bank and Central Statistics Office.

Processing

In this stage, data will be filter and remove any outliers or irrelevant data. It can drop unnecessary data or concatenate data if needed. Missing data or null value will be handle in this stage in order to generate more accurate results.

Transformation

Processed data will be used to perform several calculations and analysis. Some datasets are concatenate to implement regression calculation and correlation relationship test. The transformation process also include an analysis of the payment method.

Data Mining

This stage describes the pattern of machine learning algorithm that apply in this project. Decision tree pattern will be execute to get the results of prediction with the time period and value involved.

Visualization

Last stage of KDD is the process of display the diagrams in various ways and interpret the findings of the results in clear and concise method.

2.3 Implementation

In this stage, all the functions and analysis was conducted and the represent the results.

2.3.1 R Packages used:

- DBI
- RMySQL
- C50
- dbConnect
- dplyr
- gmodels
- rpart

2.3.2 R Functions used:

- subset()
- rpart()
- CrossTable()
- c5.0()
- sort()
- filter()
- for()
- lm()
- predict()
- pie()
- plot()
- dbGetQuery()
- read.csv()
- set.seed()
- cor()
- dbConnect()
- merge()
- data.frame()

2.3.3 Initial stage

Analysis process is the main priority of this project but before this the setup of database is a compulsory steps to store the data. While the datasets are store in the MySQL Workbench, it need to connect with RStudio in order to produce analysis.

```
drop database if exists analysis;
create database analysis;
Use analysis;

drop table if exists FirstData;
create table FirstData (
        MainKEY VARCHAR(100),
        ……..
    );
```

Figure above shows the scheme that used to create database and table to store the relevant datasets into database.

```
load data local infile 'c:\\data.csv'
into table Firstdata
fields terminated by ','
lines terminated by '\n'
ignore 1 lines;
```

After the table is created, few scripts in written to import the datasets into database.

```
con <- dbConnect(RMySQL::MySQL(),
                 ...
```

Figure 8: Script of connect RStudio with SQL

Figure above shows the code that used in RStudio to connect with SQL database therefore we able to retrieve the data directly from RStudio and perform the analysis.

2.3.4 Payment Method Summary

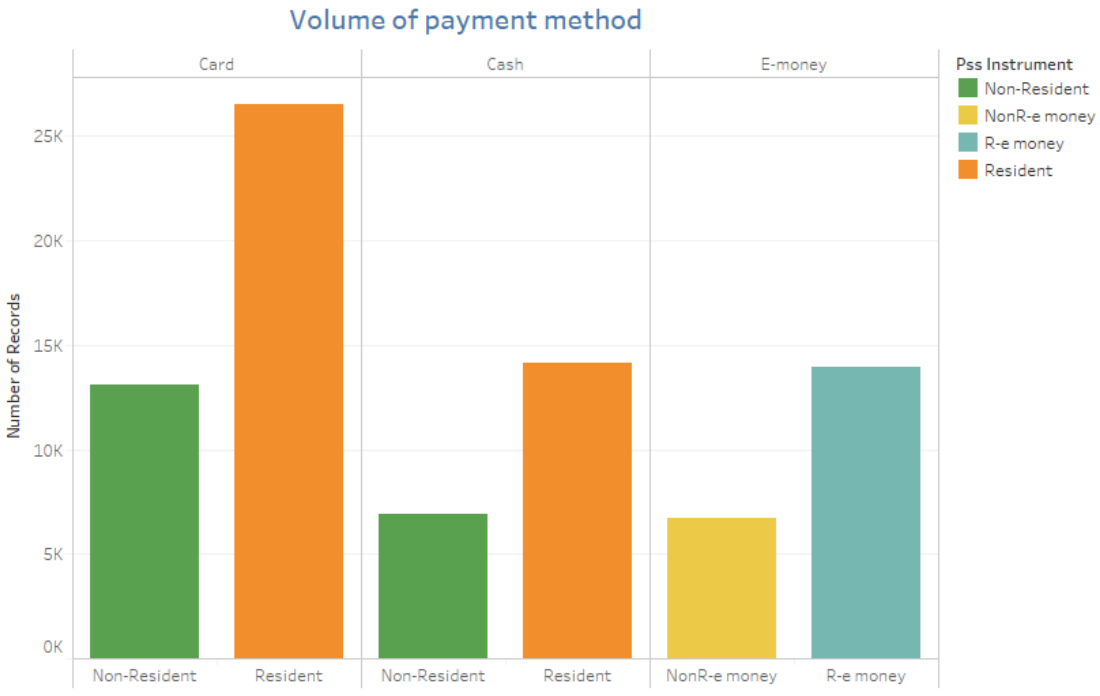


Figure 9: Volume of Payment method

Figure 9 shows the number of record for three payment methods which include cash, card and e-money payment. From the diagram we can know that card has the highest number of record and there is twice for cash and e-money. Orange and blue colour represent

residents which means residents consume in their own countries. Green and yellow colour shows people consumer outside their countries. Obviously, the number of resident is almost twice for non-resident. From this point of view, we can estimate around 50% of people willing to travel to another country. Moreover, e-money can cash has the same level of number in this diagram, therefore we can assume people willing to try new payment methods such as e-money, virtual wallet, virtual coins and so on. As the technologies keep improves, we reserve the possibility the flow of e-money will be increase.

2.3.5 Observation Value in Europe Country

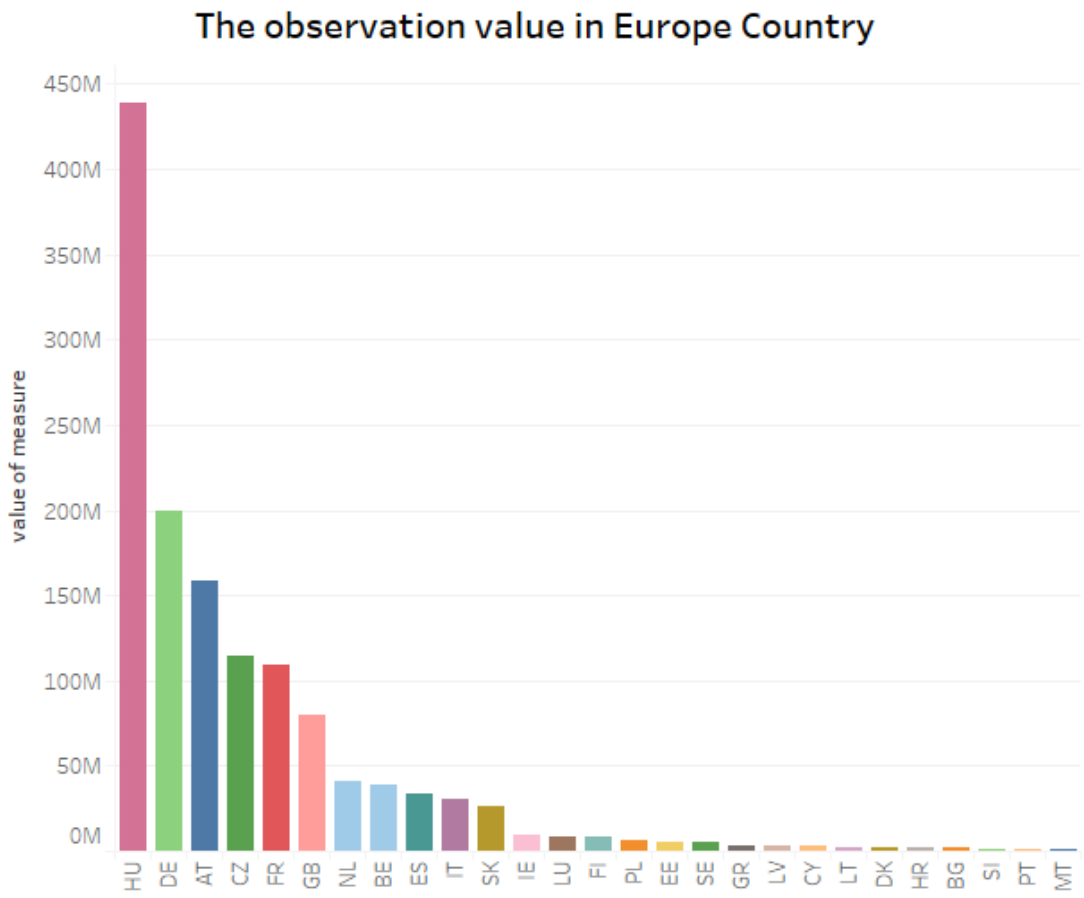


Figure 10 : Observation value for each country

Figure above shows the observation value in Europe country. Hungary stand out with the highest number from those Europe country. In the data we have, we get the value of

approximate 440 million are spend in Hungary and following by Germany which only have 200 million that is less than half of Hungary. From Germany, there are just a small difference between those countries.

2.3.6 Correlation of number visitor to Ireland

For the assumption above, the relationship of number of trips and frequency transactions was tested by using correlation coefficient. We pick the attributes of number of visitor to Ireland to check the correlation with our data.

```
timeperiod <- table(projectData$TIME_PERIOD, dnn=c("Year"))
mergeData2 <- merge(timeperiod, travelData, by="Year")
```

Figure 11: Merge function for Correlation

```
xFreq <- mergeData2$Freq
yNum <- mergeData2$NumOverseasVisitorstoIreland
cor.test(xFreq,yNum, method="pearson")
```

Figure 12: Pearson Correlation Coefficient

First, I summarise the time period of the main data by using table function and store in a variables called timeperiod. Merge function has been used to concatenate timeperiod and travel dataset by year. Pearson correlation coefficient was tested by using two attributes which is freq refer to the volume of each year and number oversea visitors travel to Ireland.

```
> cor.test(xFreq,yNum, method="pearson")

        Pearson's product-moment correlation

data:  xFreq and yNum
t = -1.5881, df = 10, p-value = 0.1433
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8132347  0.1685124
sample estimates:
       cor
-0.4487917
```

Figure 13: Output of Correlation

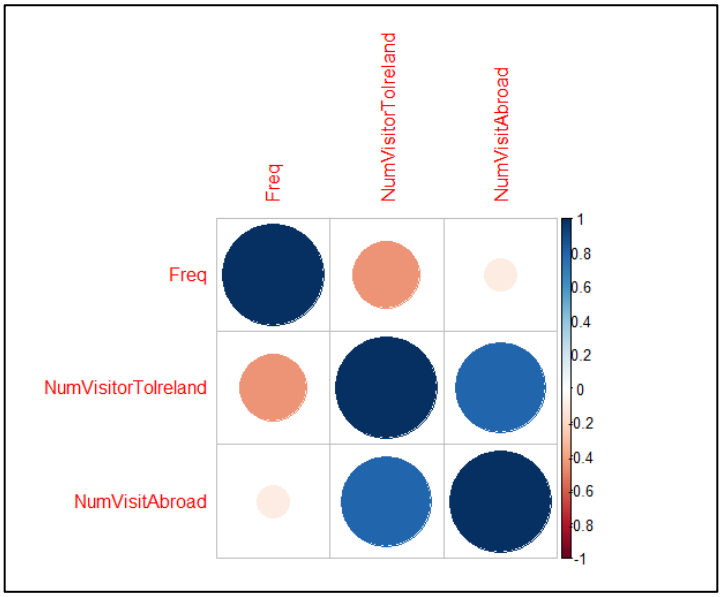


Figure 14: Matrix Correlation plot

From the correlation calculation, the value of -0.44879 was found to prove there is a negative medium relationship between them. The Figure 14 shown the correlation plot between frequency and number of visitor are medium correlation with melon colour and even weaker correlation with the number of people visit abroad of Ireland. In overall, the number of visitor to Ireland has no directly relationship with the frequency of transactions. Therefore, if the number of visitor increases, the frequency of transactions will not be affected.

2.3.7 Correlation of population in Ireland

```
for(i in 2000:2013) {
  count <- (-35772961) + (19855 * i)
  newrow <- c(i, count)
  PopulationReg <- rbind(PopulationReg,newrow)
  print(paste(i, count))
}
```

Figure 15: For loop function for Regression

Furthermore, the population of Ireland also been tested the correlation relationship with the frequency of transactions. From the datasets extracted from Central Statistics Office (CSO), the number of population has shown every five years. Therefore regression calcul ation has implement to get the number for each year. The value of y-intercept was found with -3772961 and 19855 for year. For loop function also calls in this part to get exactly value for each year.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -35772961    5929312  -6.033 3.07e-05 ***
pop$Year        19855       3000   6.619 1.15e-05 ***
---
```

Figure 16: Output of Regression

A list of number of population for each year was generate after this and store as a data frame to another correlation calculation. Merge function and cor function has perform again as above.

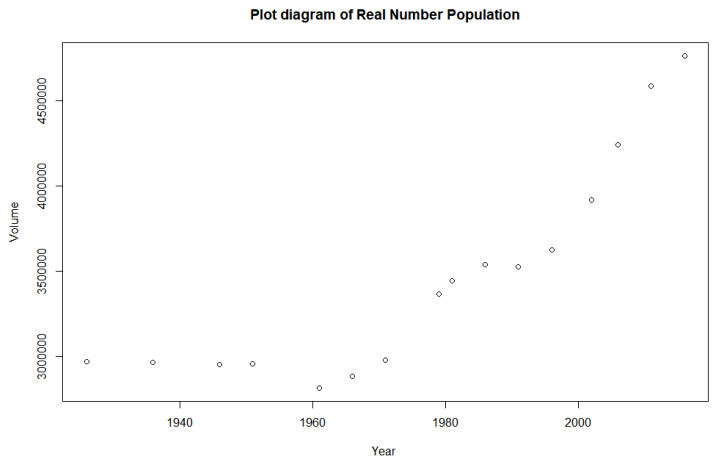


Figure 17: Plot Diagram of Real Number Population

```
> cor.test(xFrequency,yPopu, method="pearson")

        Pearson's product-moment correlation

data:  xFrequency and yPopu
t = 2.4156, df = 12, p-value = 0.03258
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.05944583 0.84585975
sample estimates:
      cor
0.5719845
```

Figure 18: Output of second Pearson Correlation

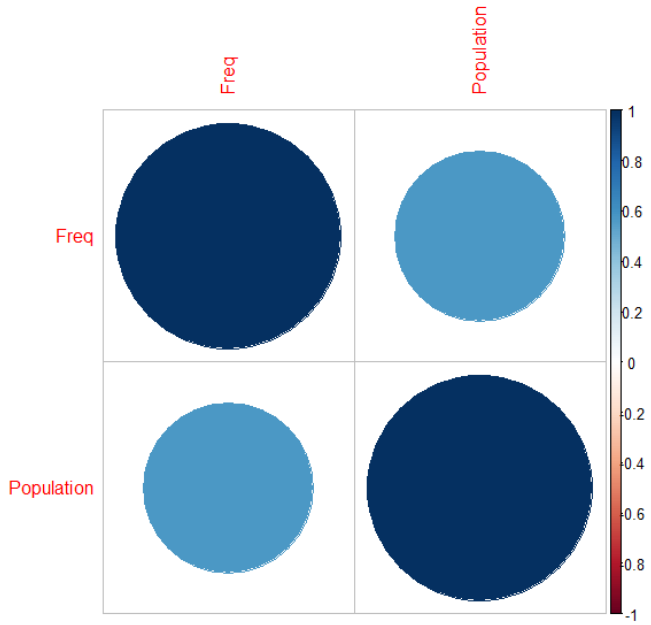


Figure 19: Plot for Matrix Correlation

From the results, the value of 0.57198 was found to say that there is medium relationship between them. From Figure 17 the plot shows the number of population if Ireland increas es constantly every year and reach approximate 4.8 million.

2.3.8 Machine Learning Algorithm: Decision Tree

Table 4: Description of attributes for Decision Tree

Attributes	Description
PSS_INFO_TYPE	Types of transactions. Eg: cash, card e-money.
PSS_INSTRUMENT	Type of payment service for resident or non-resident. It can be using card service by debit or credit function.
DATA_TYPE_PSS	Data type each transactions, most of the transactions was recorded as number or value. Average and percentage also consider in this category but did not use in this dataset.
COUNT_AREA	Transactions reported by specific country.
COUNT_SECTOR	Unit of manage those transactions. All transactions were under Monetary Financial Institution (MFI).
CURRENCY_TRANS	Currency that used in the transactions. It only categories as euro and euro are national currencies or all currencies combined.
SERIES_DENOM	Value that denominate as either euro or not applicable.
TIME_PERIOD	Year of transactions.
OBS_VALUE	Amount that observe in each transactions.

```
getData <- filter(projectData, (REF_AREA == "IE"))
getData <- getData[!is.na(getData$OBS_VALUE),]
getData <- filter(getData, (PSS_INFO_TYPE == "F104"
                  |PSS_INFO_TYPE == "F105" | PSS_INFO_TYPE == "F200"))
```

Figure 20: Clean Data for Decision Tree

```
Decision tree:

PSS_INSTRUMENT = I20: F104 (146)
PSS_INSTRUMENT = I10:
:...OBS_VALUE <= 0:
    :...TIME_PERIOD <= 2013: F104 (41/14)
    :   TIME_PERIOD > 2013: F105 (305/91)
    OBS_VALUE > 0:
    :...COUNT_AREA in {AT,BE,BG,CY,CZ,D9,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IT,LT,
        :              LU,LV,MT,NL,PL,PT,RO,SE,SI,SK,U6,Z9}: F104 (167/4)
        COUNT_AREA = X0:
        :...OBS_VALUE > 149301: F105 (3)
            OBS_VALUE <= 149301:
            :...DATA_TYPE_PSS in {ND,NM,NR,VD,VM,VR}: F104 (23/9)
                DATA_TYPE_PSS = NT:
                :...OBS_VALUE <= 150.575: F105 (5)
                :   OBS_VALUE > 150.575: F104 (3)
                DATA_TYPE_PSS = VT:
                :...OBS_VALUE <= 6690.92: F105 (3)
                    OBS_VALUE > 6690.92: F104 (4)
```

Figure 21: Output of Decision Tree

	data_pred		
data_test$PSS_INFO_TYPE	F104	F105	Row Total
F104	180	28	208
	17.090	29.947	
	0.865	0.135	0.693
	0.942	0.257	
	0.600	0.093	
F105	11	81	92
	38.639	67.707	
	0.120	0.880	0.307
	0.058	0.743	
	0.037	0.270	
Column Total	191	109	300
	0.637	0.363	

Figure 22: Cross Table for Decision Tree

Machine Learning Algorithm has implement in this project in the final stage. After all the datasets are collected and analysed, Decision Tree has apply a strategy of dividing the data into smaller portions to identify the pattern of prediction.

Decision Tree has built by using C5.0 package in RStudio in this part to identify the payment method that will be used. There are various types of payment in the dataset but only three payment methods are pick and sort to use in this prediction. F104 represents cash payment, F105 represents card payment and F200 represent e-money payment. There are few rules has states in Figure 20 one of the rules proves people who are residents has spent less than or equal to 6690.92 million in card payment. From the 2013 onward, the number of using card payment has increases and before that people prefer using cash payment.

From the table above, this model classifies 87% accuracy of the test group which is randomly selected. 11 out of 191 with F104 was incorrectly classified as F105 while 28 out of 109 with F105 was incorrectly classified as F104. Based on F104, 180 transactions fell into true positive area which means it exactly using cash payment and 81 transactions fell into true negative area which shows it using card payment. Besides that, 28 transactions fell into false negative which means the classifier predicted the transactions will using card payment but using cash in real and 11 transactions fell into false positive which is predicted as cash payment but in card anyway.

2.4 Testing

2.4.1 Unit Testing

Unit testing is an application testing where the individual units of an applications are tested. The aims is to ensure each units are perform well as designed. Basically, unit test was test the correctness of single input and output. Testing mainly explains the purpose, actions, expected results and actual results will do.

2.4.2 Database Server Testing

As I mentioned above, RMySQL has been used to connect SQL with RStudio in order retrieved data from database directly. Therefore a testing has executed by using tryCatch function.

```
tryCatch({
    con <- dbConnect(RMySQL::MySQL(),
                     ....},
    warning = function(w) {
        (print(paste("Warning: ", conditionMessage(w))))},
    error = function(e) {
        (print(paste("Error: ",conditionMessage(e))))
    })
```

Figure 23: tryCatch function for Database Server Testing

```
[1] "Error:  Failed to connect to database: Error: Can't connect to MySQL server on 'localhost' (0)\n"

[1] "Warning:  Decimal MySQL column 27 imported as numeric"
```

Figure 24: Error and Warning message returned

Most of the time the database server will be operating normally but in case the server down or lost connect, tryCatch function will avoid this issues come out and affect the implement of the application. I was testing this features when I turn off the SQL server and Figure 23 shows the error message that notice I failed to connect to database. Moreover, after the database server is connected and runs the script of retrieve data from database it shows the warning message to warn me that the data type of column 27 has changed to numeric from decimal. In order to get the message from system, conditionMessage() was used to get the warning or error message and display in the print().

Until this stage, tryCatch function has performed well to catch an error or warning but some problems occurs in within this function. Even though the warning message shown the imported message but retrieve datasets from database unable to work. To solve this problem, some research has done and I figure out the reason of not working properly is because getQuery() will not working in tryCatch() due to how R code was written. A new package called RPostgreSQL has suggested to replace RMySQL, getQuery() but is shows few errors and also not working for my datasets.

It is so frustrated to say that it come out more problems after trying to solve a problem. I decided to manually run the getQuery() outside of tryCatch() and shown that is working. Therefore, tryCatch() is working well in checking the database connection but failed to get query and datasets from database.

2.4.3 Validate Data Testing

As I mentioned above, script of retrieve datasets from database was not perform well and execute manually outside the tryCatch(). The datasets consisted huge number of variables and insert into database tables by comma separated value. The accuracy of variables has to be consider. Therefore a test of checking the rows and columns of the datasets was conducted by using assertr package.

```
projectData %>%
  assertr::verify(nrow(.)==423890) %>%
  assertr::verify(ncol(.)==35)
```

Figure 25： Data Validation Testing

Figure 24 shows the script of assertr() to check the rows and columns was exactly match with the expected 423890 rows and 35 columns. When this test was TRUE, it will return the datasets and shows the error message when error occurred.

```
> projectData %>%
+   assertr::verify(nrow(.) > 423890) %>%
+   assertr::verify(ncol(.) > 35)
verification [nrow(.) > 423890] failed! (1 failure)

    verb redux_fn        predicate column index value
1 verify       NA nrow(.) > 423890     NA     1    NA

Error: assertr stopped execution
```

Figure 26: Error returned from Data Testing

Return error when there are limited number of data in the datasets but the function was execute to test more than the limited number. Assertr function was very useful in testing the rows and columns in specified prediction.

Note: Both testing was refer to Section 2.1.1.2 use case requirement 1

2.5 Evaluation

In overall, all the aims and results of analysis were presented in 2.3 Implementation section. From the recourse of background until the implementation by using KDD methodology has been discussed and included in this document. KDD methodology guide me the flow the project lifecycle and results in an effective ways to complete this project. Throughout entire project, KDD ensures the project follows the processes of selecting, processing, transforming data and generate data mining to present the results of analysis.

Key analysis of this project was get the frequency of cash and cashless payment to analysis when the cashless payment rise and which country has the higher number of cashless payment. The scalability of this project was uncertainty because the convenience and security of cashless payment did not mention in this project. Among Europe country, different country has different strategy on the mode of doing business. In my results, Hungary has involved with the highest observation value but did not means this country was the high number of cashless payment. From the background I mentioned above, Sweden has the higher number of cashless payment and cash transactions was only accounted for 2% of the value of all payment made in 2015 (guardian).

3 Conclusion

Overall of this project was getting fluently in planning and implementation process. It makes me more familiar with R programming and the structure of development a project. The advantages of this project was to prove people start changing payment landscape and make easy to their life as cash is not only option for doing business. More options will come out in future to replace the usage of cash to make life easier. Big disadvantages of this project was the huge amounts of data did not brought me any benefits but more time consuming in cleaning data. It prove more did not means better.

Limitation I faced in this project were transactions record was confidential level for each party therefore it increases the difficulty for me to get the datasets to support my evidence. Even though European Central Bank (ECB) was provided open source data but still have many restriction in the open source data.

4 Further development or research

This project would be more perfect if include the comparison of cost for various payment method and maintenance of security. Cost of using cash and card might be differ due to the volume of transactions and time consuming either in people or system might also bring effect on the results. I believe that if more element has involve in this, results might be changes and be more accurate. As the technologies keep improving and advances, some of the technique should participate and make a connection between people throughout payment.

5 References

1. Cookbook-r.com. (2018). *Converting between data frames and contingency tables*. [online] Available at: http://www.cookbook-r.com/Manipulating_data/Converting_between_data_frames_and_contingency_tables/ [Accessed 24 Apr. 2018].
2. Data.gov.ie. (2018). *E3001 - Enumerated Population 1926 to 2016 (Number) by Age Group, Sex and CensusYear - data.gov.ie*. [online] Available at: https://data.gov.ie/dataset/enumerated-population-1926-to-2016-number-by-age-group-sex-and-censusyear [Accessed 23 Feb. 2018].
3. Data.gov.ie. (2018). *TMA07 - Overseas Trips to and from Ireland by Route of Travel, Year and Statistic - Overseas Trips to and from Ireland by Route of Travel, Year and Statistic - data.gov.ie*. [online] Available at: https://data.gov.ie/dataset/overseas-trips-to-and-from-ireland-by-route-of-travel-year-and-statistic/resource/f3c1fbf3-66fe-47b5-a2fb-31009cde4ed3 [Accessed 29 Mar. 2018].
4. DBD, U. (2017). *KDD Process/Overview*. [online] Www2.cs.uregina.ca. Available at: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html [Accessed 12 Nov. 2017].
5. Fischetti, T. (2018). *Package 'assertr'*. [online] Cran.r-project.org. Available at: https://cran.r-project.org/web/packages/assertr/assertr.pdf [Accessed 8 May 2018].
6. Henley, J. (2018). *Sweden leads the race to become cashless society*. [online] the Guardian. Available at: https://www.theguardian.com/business/2016/jun/04/sweden-cashless-society-cards-phone-apps-leading-europe [Accessed 7 Feb. 2018].
7. Quora. (2017). *What are the main objectives of cashless transactions?*. [online] Available at: https://www.quora.com/What-are-the-main-objectives-of-cashless-transactions [Accessed 20 Oct. 2017].
8. Sdw.ecb.europa.eu. (2018). *ECB Statistical Data Warehouse*. [online] Available at: http://sdw.ecb.europa.eu/browseTable.do?node=434881 [Accessed 7 Oct. 2017].

6 Appendix

6.1 Project Proposal

6.1.1 Objectives

The aims of Cashless Payment Analyst are to identify the trends of the public make payment without using cash. Cashless payment is used to reduce the corruption of cash medium takes place and reduce the burden of the cost of printing currency. While using Cashless Payment, the movement of the money is able to track completely how the money has been used.

For Cashless Payment Analyst, it's able to specify the population of people used cashless to payment and predict the trends of using cashless payment for next few years. It will bring a new transformation for the way of consume and improve the economic consumption.

6.1.2 Background

This analyst is to identify the trends of using cashless payment. This is idea is come with when I am in a long queue at grocery shop. I figure out the spending time of payment process is much different when people used cash and card. Most of the people will only take out the cash when they had been noticed by cashers the amount of the things they buy. It would be cause a long queue especially at the peak hour. Due of this issue, I want to observe and analyst people are likely to pay with cash or cashless and how people frequently used card to make payment also which area they would like to spent with card. For example, people may prefer to pay with card when purchase a high cost product such as washing machine because some of the shops are allows the consumer to make installment for few months.

Moreover, cashless payment is highly praise nowadays. Cashless payment describes not only credit/debit card payment but also used virtual cash such as Bitcoin, Paypal or mobile application to make payment. Through cashless payment, it will increase the efficiency of payment process and indirectly improve the quality of life for people to spend their time in more meaningful ways.

In Sweden nowadays is the most cash-free society in the world. The reason why Sweden government encouraged their citizen used cashless to make payment is because it can decreases the budget for printing money and reduced crime rate. Even though the homeless vendors that sold magazine in the capital of Sweden are using electronic payment to deal their business.

Cashless payment would be a tendency within this few years, so that my project is work to generate an accurate and reliable result to prove the frequency of cashless payment.

6.1.3 Technical Approach

To carry out this analyst, I do some research of population of cashless payment and whose countries are pursuing cashless payment. I also figure out European Central Bank have bulk of open dataset that are related to my analyst. It reduces the difficulty for me to form the data from various website. For the main technical approach in this project was KDD methodology.

6.1.4 Technical Details

R Studio

In this project, R Studio as an open-source integrated development environment and perform statistical computing and graphical in R Language. R Studio operates as retrieval data from Ms Excel and store as a dataset in RStudio. It will works with KDD methodology to select, process, transform and visualize the data and information.

SPSS

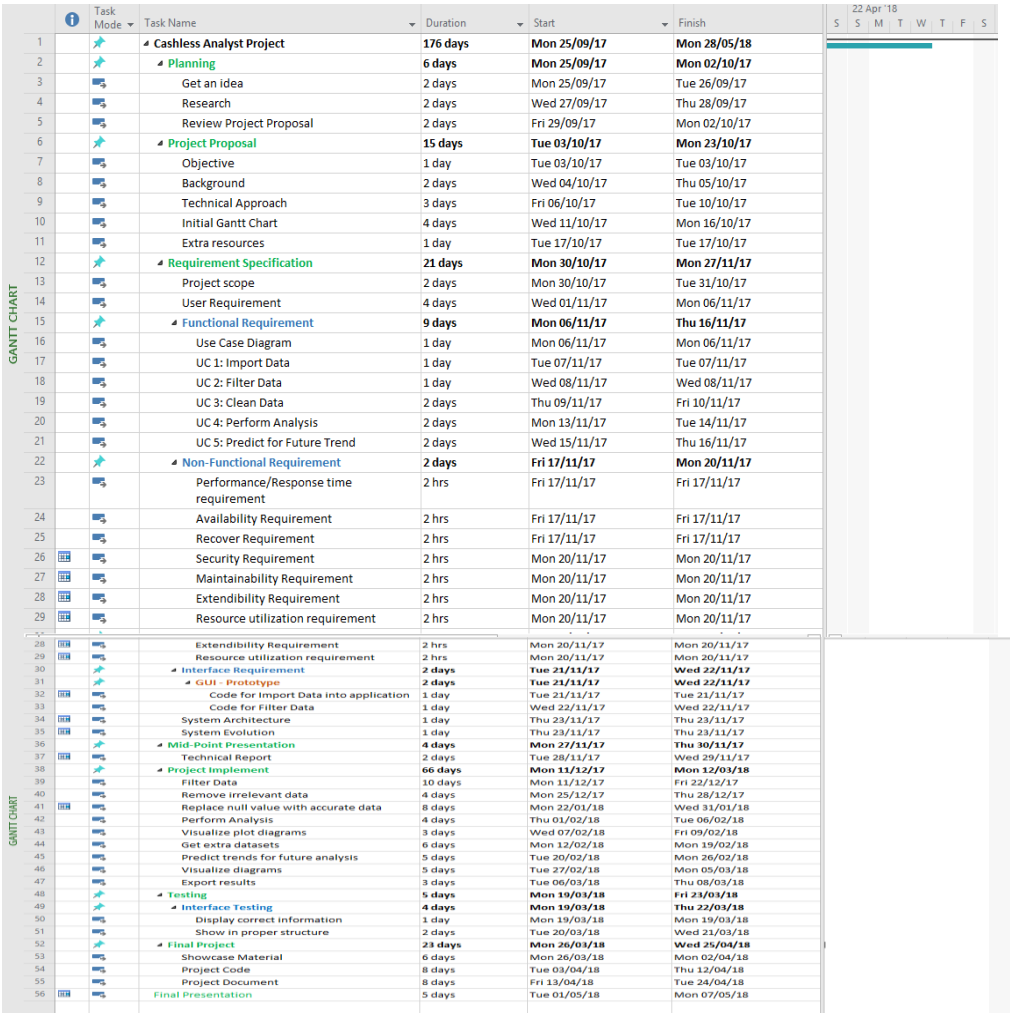
SPSS is leading statistical software used to solve a variety business and research problems. It provides a range of techniques such as ad-hoc analysis, hypothesis testing and reporting which easier to manage data, select and perform analyses. SPSS in this project is used to perform different types of diagram and ease-to-understand information.

6.1.5 Evaluation

Testing will be perform in integration tests to clarifies the system is working well after integrate all datasets into a storage. It ensures the generated results are accurate and reliable information. It also avoid there is exclusive appearance in the storage that will affect the data. In addition, system tests will be implementing to make sure the result come out is what expected for user. The results should be logical and useful for user to manage or make decision, otherwise the results would be useless.

Furthermore, the way I will evaluate the system with an end user are prove them the minor action that they will take everyday are gradually changes the economy and encourage them get more understand about the advancement of society. The results of this project may not bring impact to anyone but for whole world because it brings lots of intangible advantages.

6.1.6 Project Plan



	Task Mode	Task Name	Duration	Start	Finish
1		**Cashless Analyst Project**	**176 days**	**Mon 25/09/17**	**Mon 28/05/18**
2		**Planning**	**6 days**	**Mon 25/09/17**	**Mon 02/10/17**
3		Get an idea	2 days	Mon 25/09/17	Tue 26/09/17
4		Research	2 days	Wed 27/09/17	Thu 28/09/17
5		Review Project Proposal	2 days	Fri 29/09/17	Mon 02/10/17
6		**Project Proposal**	**15 days**	**Tue 03/10/17**	**Mon 23/10/17**
7		Objective	1 day	Tue 03/10/17	Tue 03/10/17
8		Background	2 days	Wed 04/10/17	Thu 05/10/17
9		Technical Approach	3 days	Fri 06/10/17	Tue 10/10/17
10		Initial Gantt Chart	4 days	Wed 11/10/17	Mon 16/10/17
11		Extra resources	1 day	Tue 17/10/17	Tue 17/10/17
12		**Requirement Specification**	**21 days**	**Mon 30/10/17**	**Mon 27/11/17**
13		Project scope	2 days	Mon 30/10/17	Tue 31/10/17
14		User Requirement	4 days	Wed 01/11/17	Mon 06/11/17
15		**Functional Requirement**	**9 days**	**Mon 06/11/17**	**Thu 16/11/17**
16		Use Case Diagram	1 day	Mon 06/11/17	Mon 06/11/17
17		UC 1: Import Data	1 day	Tue 07/11/17	Tue 07/11/17
18		UC 2: Filter Data	1 day	Wed 08/11/17	Wed 08/11/17
19		UC 3: Clean Data	2 days	Thu 09/11/17	Fri 10/11/17
20		UC 4: Perform Analysis	2 days	Mon 13/11/17	Tue 14/11/17
21		UC 5: Predict for Future Trend	2 days	Wed 15/11/17	Thu 16/11/17
22		**Non-Functional Requirement**	**2 days**	**Fri 17/11/17**	**Mon 20/11/17**
23		Performance/Response time requirement	2 hrs	Fri 17/11/17	Fri 17/11/17
24		Availability Requirement	2 hrs	Fri 17/11/17	Fri 17/11/17
25		Recover Requirement	2 hrs	Fri 17/11/17	Fri 17/11/17
26		Security Requirement	2 hrs	Mon 20/11/17	Mon 20/11/17
27		Maintainability Requirement	2 hrs	Mon 20/11/17	Mon 20/11/17
28		Extendibility Requirement	2 hrs	Mon 20/11/17	Mon 20/11/17
29		Resource utilization requirement	2 hrs	Mon 20/11/17	Mon 20/11/17
28		Extendibility Requirement	2 hrs	Mon 20/11/17	Mon 20/11/17
29		Resource utilization requirement	2 hrs	Mon 20/11/17	Mon 20/11/17
30		**Interface Requirement**	**2 days**	**Tue 21/11/17**	**Wed 22/11/17**
31		**GUI - Prototype**	**2 days**	**Tue 21/11/17**	**Wed 22/11/17**
32		Code for Import Data into application	1 day	Tue 21/11/17	Tue 21/11/17
33		Code for Filter Data	1 day	Wed 22/11/17	Wed 22/11/17
34		System Architecture	1 day	Thu 23/11/17	Thu 23/11/17
35		System Evolution	1 day	Thu 23/11/17	Thu 23/11/17
36		**Mid-Point Presentation**	**4 days**	**Mon 27/11/17**	**Thu 30/11/17**
37		Technical Report	2 days	Tue 28/11/17	Wed 29/11/17
38		**Project Implement**	**66 days**	**Mon 11/12/17**	**Mon 12/03/18**
39		Filter Data	10 days	Mon 11/12/17	Fri 22/12/17
40		Remove irrelevant data	4 days	Mon 25/12/17	Thu 28/12/17
41		Replace null value with accurate data	8 days	Mon 22/01/18	Wed 31/01/18
42		Perform Analysis	4 days	Thu 01/02/18	Tue 06/02/18
43		Visualize plot diagrams	3 days	Wed 07/02/18	Fri 09/02/18
44		Get extra datasets	6 days	Mon 12/02/18	Mon 19/02/18
45		Predict trends for future analysis	5 days	Tue 20/02/18	Mon 26/02/18
46		Visualize diagrams	5 days	Tue 27/02/18	Mon 05/03/18
47		Export results	3 days	Tue 06/03/18	Thu 08/03/18
48		**Testing**	**5 days**	**Mon 19/03/18**	**Fri 23/03/18**
49		**Interface Testing**	**4 days**	**Mon 19/03/18**	**Thu 22/03/18**
50		Display correct information	1 day	Mon 19/03/18	Mon 19/03/18
51		Show in proper structure	2 days	Tue 20/03/18	Wed 21/03/18
52		**Final Project**	**23 days**	**Mon 26/03/18**	**Wed 25/04/18**
53		Showcase Material	6 days	Mon 26/03/18	Mon 02/04/18
54		Project Code	8 days	Tue 03/04/18	Thu 12/04/18
55		Project Document	8 days	Fri 13/04/18	Tue 24/04/18
56		Final Presentation	5 days	Tue 01/05/18	Mon 07/05/18

Figure 27: Gantt chart with details on planning and implementation steps

7 Monthly Journals

7.1 September

Student name: Siuk Ching Tan

Programme: BSc in Computing

Specialization: Data Analytics

Month: September 2017

My Achievements

This month was the first month for final year and this subject could be my most important subject in this year. The first week I go back college I been noticed that this year was the tough year and everything must be done on time instead of working in the last minutes. The schedule of Software Project has been listed out for which document need to be submitted on which period. I struggled about what I should do for my project and need to relate with my specialization. Moreover, CA for the other subjects were come continuously and I felt bewildered with the subjects that totally no idea with, especially with Introduction to Artificial Intelligent and Web Service and API Development.

The second week was step forward on few subjects' assignment. Try to catch up everything that I miss last week as procrastination may delay every task and brings stress to me. I had done the first assignment in this semester which is research document for Introduction of AI. In addition, I am still confused with the idea of final year project.

My Reflection

I felt, it worked well in understanding what had been taught within the first two weeks. Did some of the research about each subject and get to know the theory of AI.

However, I was not successful come to a conclusion about the idea of final year project. I got dispirited because the ideas I was thinking that were made in past semester. I did not improve myself by provided challenges to myself.

Intended Changes

Next month, I will try to think out of the box for the new idea. I realized that I need to encourage myself to be more confidence to accept every challenges and difficulty. Challenges means new things, I hope I can step forward in every week.

Supervisor Meetings

On September I was not been scheduled to meet the supervisor yet so that there is no supervisor meeting.

7.2 October

Student name: Siuk Ching Tan
Programme: BSc in Computing
Specialization: Data Analytics
Month: October 2017

My Achievements

At the beginning of October, I had presented my project pitch to lecturers but unfortunately my idea had been rejected due to privacy issues. After this I was thinking few ideas to discuss with my supervisor while waiting the list of supervisor release. The first assignment of Web Service and API Development has to be submitted on end of the first week. Moreover, in Business Data Analyst class, I get to know more knowledge about various types of analyst can be used to test the hypothesis. This subject is my favorite subject because it made a little step forward every week throughout weekly lecture and tutorial class.

For the continued weeks, there were many project assignments had to submit successively such as practical assignment of Artificial Intelligent. I had a group meeting with my group mates of API Development to make progress on our assignment. Other than that, project proposal of Software Project has due on the last week of October which mean I need to put more effort on this document because I did nothing for past few week as my idea has been rejected and waiting to meet with supervisor. Furthermore,

assignment of Strategies Management was due at the same week with project proposal and CA test of Data Application Development as well.

My Reflection

I felt I did not manage the time well for every subject to finish each CAs so that I got hurry and confusion in this month. While I get stuck on some task, I would spend more time to solve it. It takes me longer time to finish my works.

Intended Changes

Next month, I will try to focus on my final year project in order to complete everything on time. This is a big project and fall in two semesters, I would like to finish as many as I can in the first semester rather than postpone until second semester.

Supervisor Meetings

I had the first meeting with my supervisor - Michael Bradford to discuss about my project idea and get some guide from him. We figure out there was difficult to get datasets from public because payment details was privacy information. Michael gave me a hand where I can get the datasets from open source website-European Central Bank. I had enquired some issues for Project Proposal document and get good advice from my supervisor.

7.3 November

Student name: Siuk Ching Tan

Programme: BSc in Computing

Specialization: Data Analytics

Month: November 2017

My Achievements

November would be a productive month as most of the assignments were due in this month. For examples, second document assignment of Artificial Intelligent and second CA of Web service and API Development are due in the same day. I worked in a team

with for API assignment but I and my group mates got some problems with this assignment. It lets me felt how tough for the final year.

The week after submission of second CA test of API was the week of its test. Lecturer had told the test could be similar with the assignment so If we did well in assignment there not a problem to attend the test. I get little bit stress on this subject because I did not do more in the past assignment due to misunderstand issues. Discussion with friends and lecturer had carry out to overcome the problem of lack of knowledge.

Requirement Specification document for Software Project has to be completed in this month and compile with project proposal to generate Technical Report document. Functional requirement, non-functional requirement and prototype have been required in this document. The knowledge I learnt in Data Application Development was suitable to use in Software Project.

My Reflection

When there are two subjects due in the same time I might not do well in other subjects. Especially CAs of the assignments are quite hard and I need to spend more time to practice and do revision on that. I might leave out the other subjects.

Intended Changes

Most of the works in this semester had almost completed and I might have the same problems with last month as procrastinate issues. I will improve myself to overcome delay of works.

Supervisor Meetings

I meet my supervisor twice in this month to enquire some problems that I face in my final year project and show him what I did in the project proposal and requirement specification document. After completed these two documents, I got some advice from Michael and start to prepare my prototype for Mid-Point Presentation.

7.4 December

Student name: Siuk Ching Tan

Programme: BSc in Computing

Specialization: Data Analytics

Month: December 2017

My Achievements

During December there is nothing much can do except for mid-point presentation. All the related document has been submitted last month, therefore this month is only focus on the prepare presentation slide. Summary, aim and future work have to be included in the presentation slide as well as prototype to demonstrate in the presentation.

My Reflection

There was plenty time to prepare the presentation since I have more than 1 week time between the submission date of Technical Report and presentation date. However, i was facing some problem in coding for prototype demonstrate so that i assume may not be high marks in the presentation. In overall, i only received a grading of 47. More effort to involve to getting a higher marks.

Intended Changes

I am depressed on the grading of the mid-point presentation, therefore I will keep improve my presentation and writing skill.

Supervisor Meetings

I meet my supervisor three times in this month before and after presentation to get some feedback to enhance my weakness. Some discussion had made for the next phase of the project.

7.5 January

Student name: Siuk Ching Tan

Programme: BSc in Computing

Specialization: Data Analytics

Month: January 2018

My Achievements

At the beginning of January, I have to sitting for three exam it is good to leave my project behind and get back after exam. Some extra learning for R programming language has carry through in the week off. Before I start to implement analysis for my project, get familiar with all the datasets is the priority work to do.

My Reflection

Small step go forward for my knowledge about R programming language but long way to go. As timetable for semester two is out, there is more time for me to fulfilling my project requirements.

Intended Changes

As my knowledge about R programming language improved, more function and analysis can be implement in this project.

Supervisor Meetings

I has no meeting with supervisor this month due to exam period and week off. Appointment a meeting for February has been made.

7.6 February

Student name: Siuk Ching Tan

Programme: BSc in Computing

Specialization: Data Analytics

Month: February 2018

My Achievements

Coming back from semester break and focus to my project in connecting SQL database with RStudio. After I did some research through online, I get an idea for writing those code. Some packages such as “DBI”, “RMySQL” and “dbConnect” was installed to execute the code. In SQL Workbench, few sentences of codes was written to create and insert data into table. Moreover, few dataset has also included in the database to perform correlation analysis.

My Reflection

Time flies while my project keep making progress so I assume I able to complete my project within the time limit.

Intended Changes

Most of the time I stuck in some problems, my supervisor will give me a good solution to overcome the problems. The way to do a project might have some changes after getting a new idea from my supervisor.

Supervisor Meetings

Weekly meeting is carry out to drive my project to next stages. The idea of involve merge function and regression analysis has be suggested from my supervisor. Good advice to enhance my project.

7.7 March

Student name: Siuk Ching Tan

Programme: BSc in Computing

Specialization: Data Analytics

Month: March 2018

My Achievements

In March, I was focus on perform some visualization by using R orTableau and generate some prediction by using Machine Learning Algorithm. I using decision tree technique in this project and it did not perform good at the beginning. Taking longer time to understand and figure out the problems and solution of this technique.

My Reflection

Feel like I have done most of the things but still have more works to go. Testing report may cause me longer time to complete it because I have not much idea to do that.

Intended Changes

The way to perform decision tree is not good enough therefore i will concentrate on this issue.

Supervisor Meetings

Discussion of Machine Learning Algorithm in depth with my supervisor and enhance my decision tree to improve the accuracy of the data mining techniques.