INDIVIDUAL TASK COVER SHEET

JAMES COOK UNIVERSITY
AUSTRALIA

[MA1580 Foundations of Data Science]

Assessment Task	Assessment 3
College	College of Science and Engineering

Student: Please sign, date, and attach this cover sheet to the front of your assessment task for all hard copy submissions.

Student Family Name	**Student Given Name**	**JCU Student Number**							
LI	YANGYANG	1	3	7	6	3	3	3	0

Assessment Title	Data Wrangling and Data Tiding Prior to Instrument Calibration
Due Date	03/11/2020
Lecturer Name	Sourav Das
Tutor Name	

Student Declaration

1. This assignment is my original work and no part has been copied/ reproduced from any other person's work or from any other source, except where acknowledgement has been made (see *Learning, Teaching and Assessment Policy 5.1*).
2. This work has not been submitted previously for assessment and received a grade OR concurrently for assessment, either in whole or part, for this subject (unless part of integrated assessment design/approved by the Subject Coordinator), any other subject or any other course (see *Learning, Teaching and Assessment Policy 5.9*).
3. This assignment has not been written for me.
4. We hold a copy of this assignment and can produce a copy if requested.
5. This work may be used for the purposes of moderation and identifying plagiarism.
6. We give permission for a copy of this marked assignment to be retained by the College for benchmarking and course review and accreditation purposes.

Learning, Teaching and Assessment Policy 5.1. A student who submits work containing plagiarised material for assessment will be subject to the provisions of the Student Academic Misconduct Requirements Policy.

Note the definition of plagiarism and self plagiarism in the Learning, Teaching and Assessment Policy:

Plagiarism: reproduction without acknowledgement of another person's words, work or expressed thoughts from any source. The definition of words, works and thoughts includes such representations as diagrams, drawings, sketches, pictures, objects, text, lecture hand-outs, artistic works and other such expressions of ideas, but hereafter the term 'work' is used to embrace all of these. Plagiarism comprises not only direct copying of aspects of another person's work but also the reproduction, even if slightly rewritten or adapted, of someone else's ideas. In both cases, someone else's work is presented as the student's own. Under the Australian *Copyright Act 1968* a copyright owner can take legal action in the courts against a party who has infringed their copyright.

Self Plagiarism: the use of one's own previously assessed material being resubmitted without acknowledgement or citing of the original.

Student Signature		Submission Date: 03/11/2020

An investigation of what kind of customers are more likely to respond to marketing strategies by the bank.

Author: YangYang Li

Abstract:

i. *Introductory statement:* Marketing in today's World Bank is an enormous part of our life investment. This analysis covers the possibility of the various types of customers responding to marketing strategies by the bank.

ii. *Purpose of the report:* The purpose of this report is to identify form a raw Bank data that what type of customers that would have a higher chance to subscribe for a term deposit and concentrate marketing efforts on such customers.

iii. *Methodological approach:* The data is related with direct marketing campaigns of a Portuguese banking institution. The Bank data consists of 45211 observations of 17 variables. Methods used in this report includes Data Cleaning, Gathering, Selection/Sampling, Type Conversion, Variable Transformation, Group Based Summarization and Exploratory Visualization of the final data.

iv. *Findings or Achievements:* The finings from this investigation, is the duration of the call has a great affect of the customer's choice.

v. *Conclusions and Implications:* This information will assist banks to choose a better strategy toward the target customers, which allows the bank to invest in higher gain financial products to make profit.

Introduction:

Banking goods are problematic and less competitive for the needs of customers in the sense of modern unstable business relationships. A data related to direct marketing campaigns of a Portuguese banking institution will be used to address this investigation. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The motivation for the report is to gain insights whether the clients are interested to the marketing campaigns that the bank has provided.

The objective of the report is to identify the type of customers who would have a greater chance to subscribe to a term deposit and concentrate marketing effort on such customers in a raw bank information data set.

The position of this study is that when subscribed to the marketing campaigns, different categories of customers will have different options.

The outcomes of this report should provide informative and persuasive proof that helps us to evaluate the form of customers that will make a bank term deposit and focuses on that.

Data:

The data used in this report is related with direct marketing campaigns of a Portuguese banking institution.

i. *Data Source:* The source of the data (bank-full.csv) is from UCI Machine Learning Repository [1].

ii. *Data Collection:* The data was originally collected form a direct marketing campaigns of a Portuguese banking institution. (from May 2008 to November 2010).

iii. *Sample Size:* The data size is with 45211 observations and 17 variables.

iv. *Number and Types of Variables:* The data includes 17 different variables and are listed below in Table 1：

Table 1.

Variable	**Variable Type**	**Description**
Age	Numeric	Client’s age
Job	Categorical	Type of job
Marital	Categorical	Marital status
Education	Categorical	Client’s education level
Default	Categorical	Has credit in default?
Housing	Categorical	Has housing loan?
Loan	Categorical	Has personal loan?
Contact	Categorical	Contact communication type
Month	Categorical	Last contact month of year
Day	Categorical	Last contact day of the week
Duration	Numeric	Last contact duration, in seconds
Campaign	Numeric	Number of contacts performed during this campaign and for this client
Pdays	Numeric	Number of days that passed by after the client was last contacted from a previous campaign
Previous	Numeric	Number of contacts performed before this campaign and for this client
Poutcome	Categorical	Outcome of the previous marketing campaign
y(output)	Binary	Has the client subscribed a term deposit?

v. *Known interventions or pre-processing:* There are no known interventions or pre-processing.

vi. *Additional Information:* There are not any missing values in the data collection.

Methods:

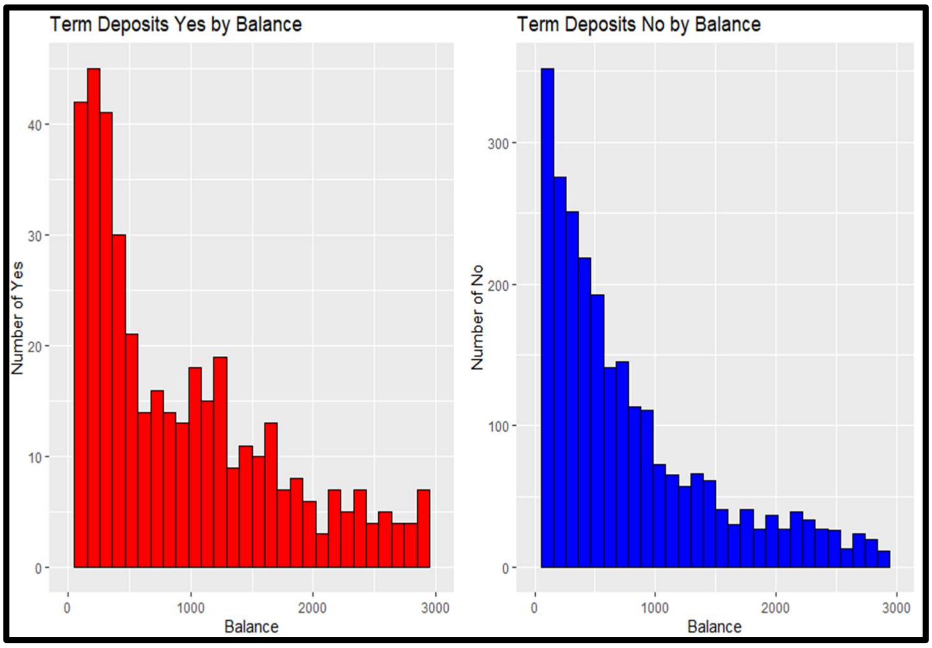
To pre-process and explore the Bank data, several data science methods were used. All methods were performed using tools from RStudio [2, 3, 4]. The targeted key topics as follows: ***1***-Data Representation, ***2***-Unstructured to Structured, ***3***-Cleaning, ***4***-Type Conversion, ***5***-Missing Value Imputation, ***6***-Gathering/Spreading, ***7***-Subsampling, ***8***-Group Based Summarisation, ***9***-Variable selection/Transformation, ***10***-Exploratory Visualization using ggplot2.

1. **Setup RStudio environment:** Setting up the right working directory by *setwd()*, then install the packages that will be used throughout the project by *install.packages(),* such as "*dplyr*", and "*ggplot2*" then call these libraries in RStudio by *library().*
2. **Import Data:** Using the *read.csv()* function to import the Bank data including Headers and the field separator character, to allow us to process the data. (1,6)
3. **Preparation:** By use *summary()* and *str()* function the data was inspected to allows decision making. To ensure the correct data was loaded and the unnecessary cells would be ignored later in the stage. *is.na()* to check is there any missing values in the Bank data so it could be remove later. (1,6,9)
4. **Remove Unwanted Data:** As mentioned above, the unnecessary fields in the data would be removed by *"dplyr", filter()* function to result in a structured and tidy data set. Which allows further progress to data. Such as balance observation, by using *filter()* to remove any values that are less than 0. The columns that are less effective or irrelevant observations 'day', 'month', 'pdays' and 'poutcome' will be removed by *select()* function. (2,3,5,9)
5. **Subsampling:** By using *group_by()* and *sample_frac()* from "*dplyr*" package, the size of the data has been reduced to 10% of each of the education categorise in the original set. Then use *summarise()* function to check the sample data. As in this way the data is more spreading which would be a great representation of the complete data. (2,3,7,8,9)
6. **Transform variables:** The balance variable has been transformed by rscl_01() function and *mutate()* function to allows great representation of the data. (3,9)
7. **Explore and Visualise the Data:** By using *filter()* function creates 2 tables with the variable y named as 'bank_yes' and 'bank_no'. Then use *ggplot(), geom_histogram()* and *grid.arrange()* to plot 2 histogram (***Plot 1***) base on the 2 table previously created in order to identify the relationship between the balance (x) and number of yes or no that the customer would subscribed (y). After that, use *ggplot()* and *geom_boxplot()* to plot (***Plot 2***) duration (x) and number of yes or no that the customer would subscribed (y), to find the relationship. Also, a bar plot (***Plot 3***) of duration (x) and jobs (y), to find the relationship between type of customers and call duration. Last, using *ggplot()* and *geom_jitter()* to plot a distribution plot (***Plot 4***) education (x) and balance (y). (2,3,6,9,10)

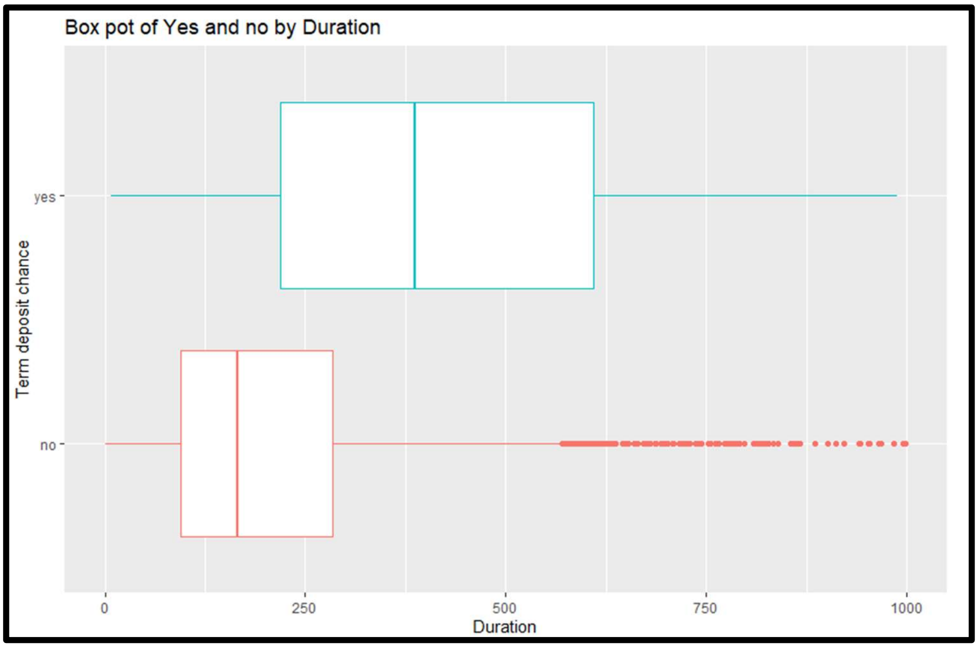
Results and Discussion:

Plot 1 is 2 histogram that illustrates if the product (bank term deposit) would be ('yes') or not ('no') subscribed base on customer's balance.

This outcome is beneficial since it indicates that there is some correlation between whether or not the client is subscribed on the basis of their balance. But more should be studied as it cannot be inferred on this plot basis. The outcome is consistent with the goal of knowing client and product relationships. By comparing the 2 histograms between 1000 to 2000, the more balance the consumer has, the stronger the shift that the product will be subscribed to.



Plot 1: Histograms of number of Yes or No vs Balance



Plot 2: Box plot of change of Yes or No vs Duration

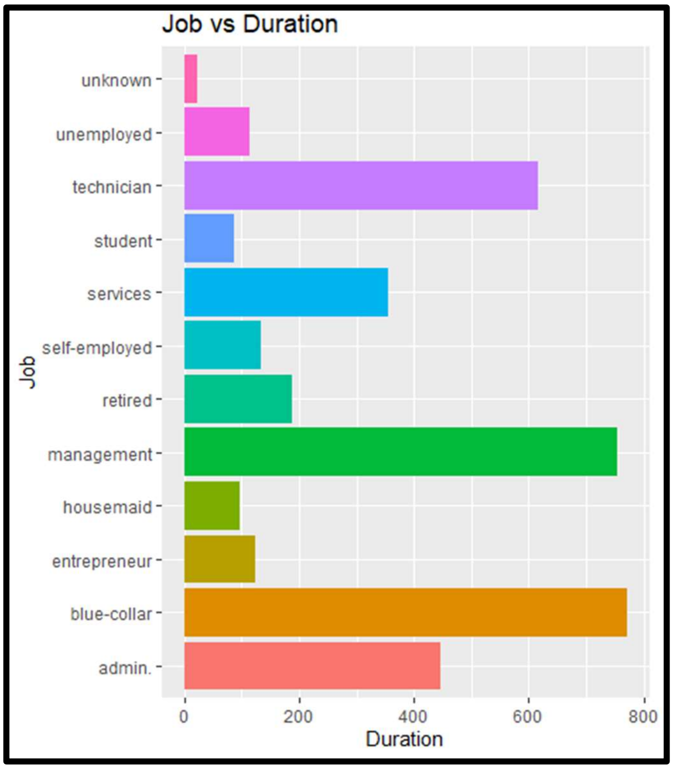
Plot 2 is a box plot, that represents the distribution of duration between the customer that would subscribe or not.

The outcome of this plot is very useful, as it reveals the length of the call that would significantly affect the option of the consumer. The result aligns with the goal that helps us to determine the form of customer preference based on call durations. The period will usually range from 270 seconds to 600 seconds from an analysis point of view, as this is the time the consumer will finish asking questions about the product and subscribing it. It can be assumed on the basis of the plot that the customers will have a greater probability of selecting the product between the previously stated time period.

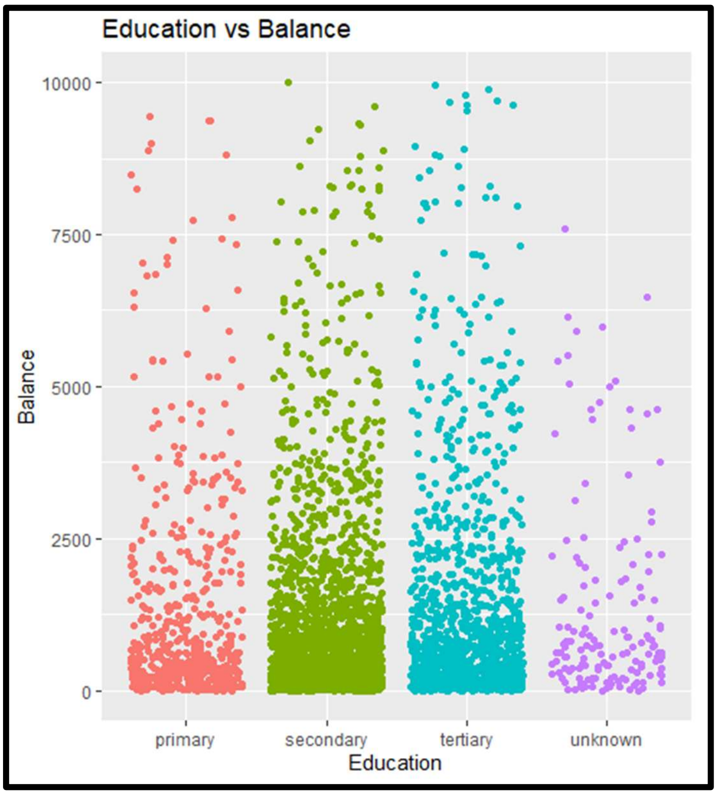
Plot 3 is a bar plot, that showing the relationship between the job and duration of a customer.

The outcome of this plot is crucial, as it shows what job of the customer would do that has a better probability to have bank term deposit. The result associates with the goal, which supports us to determine the form of customer preference based on jobs. It reveals from the plot that it will have less impact to subscribe to the bank product, such as unemployed, student's housemaid.

It can be concluded that if the customer is does jobs such as technician and admin etc. that they may choose a bank product from the bank.



Plot 3: Bar plot Job vs Duration



Plot 4: Distribution plot of Education vs Balance

Plot 4 is a distribution plot, that illustrates the balance of a customer would be depends on their education levels.

The outcome of this plot is valuable as it helps us to assess whether the customer is capable of choosing bank product form their level of education. The truth is that almost any level of education would be capable to have a bank term deposit. But from an empirical point of view, customer who has complete their secondary school, would have a better chance to have a higher balance, hence may choose a bank product. Still the data need to be analysed further in order to gain more information form the plots. This aligns with the purpose of knowing what type of customer the bank should concentrate on.

Conclusions:

A valuable initial overview of the data collected was provided by the research carried out and areas for further improvement were highlighted This investigation has accomplished the objective as it identified the type of customers who would have a greater chance to subscribe to a term deposit and concentrate marketing effort on such customers in a raw bank information data set.

In summary, it can be concluded that, based on the findings, the if the customer does jobs such as technician and admin etc. also that the customer who has complete their secondary school, they will have a higher chance of choosing a bank product from the bank. As a recommendation for future the bank should focus on these types of customers. Unavoidable limitation of this investigation, as this data was only based on one nation, so maybe inaccurate predictions towards other country or even other suburban. Therefore, other common segmentation factors like demographics, religion, region/country, and the credit score of the customer should be further analysed.

References.

- [1] UCI Machine Learning Repository,(2012,02,14), Bank Marketing Data Set, retrieved October 2020, URL: https://archive.ics.uci.edu/ml/datasets/bank+marketing

- [2] RStudio Team (2016). RStudio:Integrated Development for R. RStudio,Inc., Boston, MA Mode: Desktop. Version 1.1.1093, used October 2020, URL: https://rstudio.com/products/rstudio/download/

- [3] R(2020), R 4.0.2 for Windows, used October 2020, URL: https://cran.r-project.org/

- [4] R for Data Science, Hadley Wickham (2017), used October 2020, URL: https://r4ds.had.co.nz/

Appendices (R code):

```
# Setup RStudio environment:
setwd("C:\\Users\\liyan\\Desktop\\MA1580\\Assessment 3\\YangYang_Li_Assessment_3")

# Libraries
#~~~~~~~~~~~
library(dplyr)
library(ggplot2)
library(gridExtra)


# Import data
bank = read.csv(file = "Data 8/bank-full.csv",header = T, sep = ";" )


# Preparation
summary(bank)
str(bank)
dim(bank)
is.na(bank)


# Remove unwanted data
new_bank = bank %>% select(-c(day,month,pdays,poutcome))
sum(bank$balance<0)
filtered_bank = new_bank %>% filter(balance > 0) %>% filter(balance < 10000)


# Subsampling:
grouped_bank = filtered_bank %>% group_by(education)
summarise(grouped_bank, n.fren = n())
sub_bank = grouped_bank %>% sample_frac(size = 0.1, replace = F)


# Transform variables
rscl_01 = function(x)(x-min(x))/(max(x)-min(x))
rscl_bank = sub_bank %>% mutate(rscl_balance = rscl_01(balance))


# Explore and Visualise the data
bank_yes = rscl_bank %>% filter(y == "yes")
bank_no = rscl_bank %>% filter(y == "no")

# Balance histogram
ggyesBalance = ggplot(bank_yes, aes(balance)) + geom_histogram(col = "black",fill ="red") +
  labs(title = "Term Deposits Yes by Balance", x="Balance", y="Number of Yes")  +  xlim(0,3000)
ggnoBalance = ggplot(bank_no, aes(balance)) + geom_histogram(col = "black",fill = "blue") +
  labs(title = "Term Deposits No by Balance", x="Balance", y="Number of No")+  xlim(0,3000)
grid.arrange(ggyesBalance, ggnoBalance, ncol = 2)
```

```
# Duration Boxplot
ggplot(rscl_bank, aes(duration,y)) +geom_boxplot(aes(col = y)) + xlim(0,1000) +
  labs(title = "Box pot of Yes and no by Duration", x="Duration", y="Term deposit chance")

# Job vs Duration
ggplot(rscl_bank, aes(job)) + geom_bar(aes(size = duration, fill = job )) +
  labs(title = "Job vs Duration", y="Duration", x="Job") + coord_flip()

# education vs balance
ggplot(rscl_bank, aes(education, balance)) + geom_jitter(aes(col = education)) +
  labs(title = "Education vs Balance", y="Balance", x="Education")
```