Final Project – Milestone 1

Shamim Sherafati
ALY 6010 - Milestone 1

Date: 2022.11.10

Part A: Introduction

Dataset Description

The bank's dataset included some information about its costumers like; age, marital status, their job, education, the loan and housing, campaign and its duration and etc.
As I previously study and searched on financial industry and the field of bank, so based on this I choose this subject as my dataset and analyzed on it.

Kind of data

This dataset consists of both textual
and numeric fields, and we must convert it to factor format in order to use it for analysis.
We have data on their age, which we will use to analyze their job and its connection with their education so that we can get information about the amount of loan if they get and compare it with their housing.
Furthermore, we can obtain a trend analysis of the costumer's age
based on various factors such as campaign, duration, balance, contact types which they use from this dataset.

Purpose of the dataset

This analysis will help to get the majority of the answers to these questions by
predicting and identifying the right prospects at the right time, building customer loyalty, promoting efficiency across different departments and getting detailed information on different sectors like loan amount. The dataset has a total of 11162 rows and 17 fields which after cleaning decreased to 1118 rows and 16 fields.

Data source

I get this dataset from Kaggle website and used my previous; ALY 6010, work R practice- week 1 and used my previous r code for analyzing.

(Source : file:///Users/shamimsherafati/Documents/ALY%206010/w1/Sherafati_M1-Week1-R-Practice.html)

Part B: Data Analysis

Packages and library installation

```
> #install.packages("plyr")
> #install.packages("dplyr")
> #install.packages("tidyr")
> #install.packages("tidyverse")
> #install.packages("psych")
> #install.packages("ggpubr")
> #install.packages("ggplot2")
> #install.packages("plotly")
> #install.packages("moments")
> #install.packages('gmodels')
>
> library(plyr)
> library(dplyr)
> library(tidyr)
> library(tidyverse)
> library(psych)
> library(ggpubr)
> library(ggplot2)
> library(plotly)
> library(moments)
> library(gmodels)
>
```

Figure 1: library and packages

Dimension of the dataset

```
> dim(bank)
[1] 1118   16
>
```

Figure 2: dimension of the dataset

Data Cleaning

I first imported libraries, and after that I imported the "bank.csv" dataset and then proceeded to prepare the data for analysis. (reference: based on my previous work R practice- week 1)

age <int>	job <chr>	marital <chr>	education <chr>	default <chr>	balance <int>	housing <chr>	loan <chr>	contact <chr>	day <int>
59	admin.	married	secondary	no	2343	yes	no	unknown	5
56	admin.	married	secondary	no	45	no	no	unknown	5
41	technician	married	secondary	no	1270	yes	no	unknown	5
55	services	married	secondary	no	2476	yes	no	unknown	5
54	admin.	married	tertiary	no	184	no	no	unknown	5
42	management	single	tertiary	no	0	yes	yes	unknown	5
56	management	married	tertiary	no	830	yes	yes	unknown	6
60	retired	divorced	secondary	no	545	yes	no	unknown	6
37	technician	married	secondary	no	1	yes	no	unknown	6
28	services	single	secondary	no	5090	yes	no	unknown	6

1-10 of 10,000 rows | 1-10 of 17 columns Previous 1 2 3 4 5 6 ... 1000 Next

Figure 3: dataset of bank

	age <int>	job <chr>	marital <chr>	education <chr>	balance <int>	housing <chr>	loan <chr>	Contact_Info <chr>	day <int>
4738	37	management	SINGLE	tertiary	102	yes	no	cellular	6
4736	37	technician	SINGLE	tertiary	0	yes	no	cellular	23
4733	37	management	MARRIED	tertiary	156	no	no	cellular	19
4731	37	management	MARRIED	tertiary	0	no	no	cellular	15
4680	37	technician	MARRIED	secondary	480	no	no	cellular	22
4639	37	unemployed	SINGLE	secondary	443	no	no	cellular	29

6 rows | 1-10 of 17 columns

Figure 4: data set's variable and their data types.

My data was first sorted by the decreasing age column, and the default column was subsequently removed because it made no difference in the comparison of the data. Later, to be more precise, the column 'contact' was renamed to 'Contact Info'.
Rows are reduced from 11,162 rows to 1,118 rows when the top 50% and bottom 45% of the data are calculated.
Finally, the column variables for "married" were changed from tiny to capital letters.

```
> str(bank)
'data.frame':   1118 obs. of  16 variables:
 $ age         : int  37 37 37 37 37 37 37 37 37 37 ...
 $ job         : chr  "management" "technician" "management" "management" ...
 $ marital     : chr  "SINGLE" "SINGLE" "MARRIED" "MARRIED" ...
 $ education   : chr  "tertiary" "tertiary" "tertiary" "tertiary" ...
 $ balance     : int  102 0 156 0 480 443 0 4017 1113 4151 ...
 $ housing     : chr  "yes" "yes" "no" "no" ...
 $ loan        : chr  "no" "no" "no" "no" ...
 $ Contact_Info: chr  "cellular" "cellular" "cellular" "cellular" ...
 $ day         : int  6 23 19 15 22 29 8 30 2 30 ...
 $ month       : chr  "may" "jul" "nov" "jan" ...
 $ duration    : int  445 366 366 426 344 1600 257 665 229 543 ...
 $ campaign    : int  1 6 3 2 2 1 2 2 1 4 ...
 $ pdays       : int  258 -1 -1 196 182 -1 97 196 182 -1 ...
 $ previous    : int  2 0 0 1 8 0 1 1 1 0 ...
 $ poutcome    : chr  "failure" "unknown" "unknown" "other" ...
 $ deposit     : chr  "yes" "yes" "yes" "yes" ...
```

Figure 5: Bank Data Set Structure

Subsets of data

In this step, I got the subset of balance, age and campaign ; As we can see from the tables below, in figure 2, I put the balance as a discrete data equal to 1. So, 5 rows created which shows the age between 37-39 years old that consist of either married or divorced are included in this subset of data.
With using this subset function , we can compare variables in different columns more easily as the data has been filtered out.

```r
subset(bank, balance == 1) #Find subset of balance
subset(bank, age == 20:40) #subset of ages between 20-40 years old
subset(bank, campaign == 10) #subset of campaign between only counted 10
```

Description: df [5 × 16]

	age <int>	job <chr>	marital <chr>	education <chr>	balance <int>	housing <chr>	loan <chr>	Contact_Info <chr>	day <int>
9	37	technician	MARRIED	secondary	1	yes	no	unknown	6
9873	38	technician	MARRIED	secondary	1	yes	no	cellular	21
8997	39	entrepreneur	MARRIED	tertiary	1	yes	yes	unknown	29
7903	39	admin.	DIVORCED	secondary	1	no	yes	cellular	7
5414	39	technician	MARRIED	secondary	1	yes	no	unknown	26

5 rows | 1-10 of 16 columns

Figure 6: subset of balance=1

In this I created a subset for age as a continuous data; There are 53 rows constructed based on this subset that display various variables in this filtered domain in Figure 3, where I have restricted the data to show only the ages between 20 and 40.

Description: df [53 × 16]

	age <int>	job <chr>	marital <chr>	education <chr>	balance <int>	housing <chr>	loan <chr>	Contact_Info <chr>	day <int>
4007	37	blue-collar	SINGLE	unknown	217	no	no	cellular	23
3498	37	management	MARRIED	tertiary	2283	no	no	cellular	6
2591	37	technician	MARRIED	tertiary	127	yes	no	cellular	25
2031	37	entrepreneur	SINGLE	secondary	-22	yes	no	cellular	15
1355	37	services	SINGLE	secondary	1045	no	no	cellular	17
830	37	technician	MARRIED	tertiary	247	no	no	cellular	20
123	37	technician	SINGLE	secondary	3326	yes	no	unknown	21
10809	38	management	MARRIED	tertiary	2278	yes	yes	cellular	13
10202	38	blue-collar	MARRIED	primary	714	yes	no	unknown	29
9636	38	services	MARRIED	secondary	-314	yes	no	unknown	28

1-10 of 53 rows | 1-10 of 16 columns Previous 1 2 3 4 5 6 Next

Figure 7: subset of age between 20 to 40

Additionally, I made a subset of the campaign to count 10 and only 4 rows were produced, each of which included 2 age ranges in a distinct variable in a different column.

Description: df [4 × 16]

	age <int>	job <chr>	marital <chr>	education <chr>	balance <int>	housing <chr>	loan <chr>	Contact_Info <chr>	day <int>
4840	38	self-employed	SINGLE	tertiary	605	no	no	telephone	18
4780	38	blue-collar	SINGLE	tertiary	2885	yes	no	unknown	20
9105	39	admin.	DIVORCED	secondary	32	no	yes	unknown	20
8741	39	services	MARRIED	secondary	2	yes	no	cellular	25

4 rows | 1-10 of 16 columns

Figure 8: subset of campaign equal to 10

Create Table

Looking at the below data, specifically the frequency of yes and no for the Deposit and Age variables in Figure 9, we can see more customers I the age group of 38-40 have not open deposit accounts and the customers in the age group of 37 have the most deposit accounts.

```
> ftable(bank_Ftable2)
      37  38  39  40
no     0 209 200 210
yes  150 144 143  62
>
```

Figure 9: Frequency Tables for Deposit and Age

```
> ftable(bank_Ftable3)
          DIVORCED MARRIED SINGLE
cellular        90     487    197
telephone        3      31     14
unknown         49     177     70
>
```

Figure 10: Frequency Tables for Contact info and Marital Status

Looking at the below data for the Contact info and Marital Status variables in Figure 10, number of people who are using cellular phones are 774 out of which 487 are married and there are only few people

who are using telephone irrespective of their marital status. Although, we do have a total of 296 peoples with unknown status about their contact information.

```
> #create CrossTable for 'age' and 'Housing'
> banl_ct2 <- CrossTable(bank$age, bank$housing,
+                        dnn = c("Age", "Housing"))


   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|
```

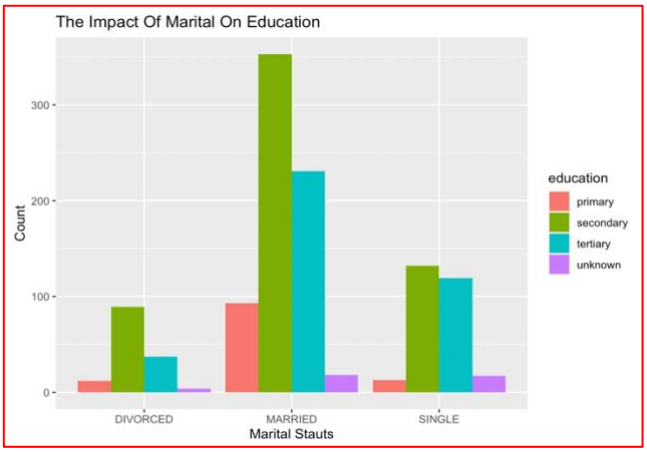
```
Total Observations in Table:  1118

             | Housing
         Age |        no |       yes | Row Total |
-------------|-----------|-----------|-----------|
          37 |        84 |        66 |       150 |
             |     6.547 |     4.818 |           |
             |     0.560 |     0.440 |     0.134 |
             |     0.177 |     0.102 |           |
             |     0.075 |     0.059 |           |
-------------|-----------|-----------|-----------|
          38 |       147 |       206 |       353 |
             |     0.047 |     0.035 |           |
             |     0.416 |     0.584 |     0.316 |
             |     0.310 |     0.320 |           |
             |     0.131 |     0.184 |           |
-------------|-----------|-----------|-----------|
          39 |       139 |       204 |       343 |
             |     0.284 |     0.209 |           |
             |     0.405 |     0.595 |     0.307 |
             |     0.293 |     0.317 |           |
             |     0.124 |     0.182 |           |
-------------|-----------|-----------|-----------|
          40 |       104 |       168 |       272 |
             |     1.111 |     0.818 |           |
             |     0.382 |     0.618 |     0.243 |
             |     0.219 |     0.261 |           |
             |     0.093 |     0.150 |           |
-------------|-----------|-----------|-----------|
Column Total |       474 |       644 |      1118 |
             |     0.424 |     0.576 |           |
-------------|-----------|-----------|-----------|
```

As we can see in Figure 11 about the cross-table representation about age and housing, there are 150 people in the age group of 37, out of which only 66 people have housing. Out of 1118 peoples, people in the age group of 38 has the highest count of 353 for both with and without housing among which 206 people have housing and 147 people with no housing as compared to other age groups.

Figure 11: Cross Tables for Age and Housing

'Age', 'Job', 'Education' and 'Campaign' are the most usage variables within the variables in dataset. By getting Mean, Standard Deviation from age and campaign I wanted to get the basic information about them to gain better analyse from them.

Diagrams



The bar chart below, the proportion of tertiary education is highest Married people which is counted more than 300 and it is followed by Single persons which is halved. (Figure 12)

The level of secondary education is also highest in married people and followed by single persons which is close to the level of tertiary education.
In divorced people, all the levels in education are the lowest one. So, it can be analysed that the sense of educating in married persons are the highest one in compare with the singles and those who divorced.

Figure 12: The impact of marital on education

As we can see from this graph, 200 of 38- and 39-years old people voted 'yes' to housing and 150 of them don't own any house. This proportion is the same in 40 years old people with around 170 and 100 in having house and not respectively. (Figure 13)
However, the 37 years people, it's a bit different which the number of them who don't own any house is more than owners with about 90 to 60 respectively.

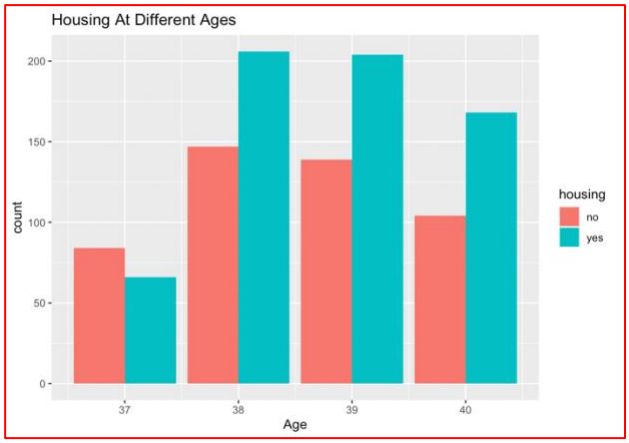
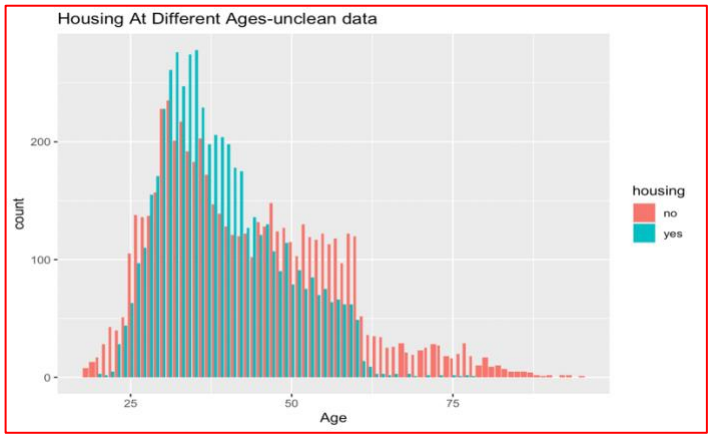


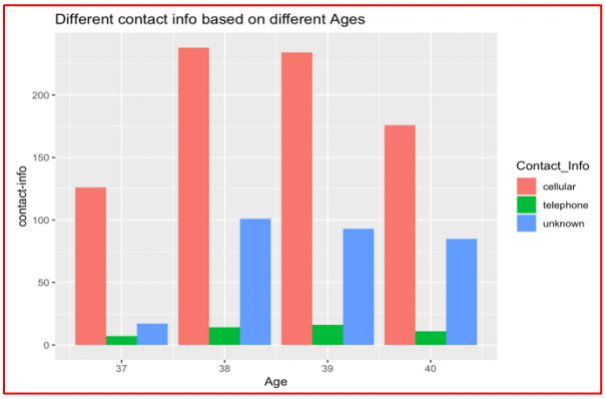
Figure 13: Housing at different ages



With using uncleaned data and compare the housing ownership based on age, it is clear, and we can see in the range of 25-45 years old the number of owners is the highest and it reached to its peak in 35 years old. On the other hand, 45 years old upward, the 'no' voted from these range age depicts that the number of ownerships has experienced a decrease trend. (Figure 14)

Figure 14: Housing at different ages- unclean data

Here, I created two graphs; a box plot and a ggplot of age and contact-info to gain information about three type of contacts which each age range use. As we can see majority of people use cellular and the usage of telephone is the least once. In the age of 38, this proportion peaked at above 200 which counted. While half of this age range is unknown their contact info but in total cellular seems the most popular type in compare with telephone. (Figure 15 - Figure 16).



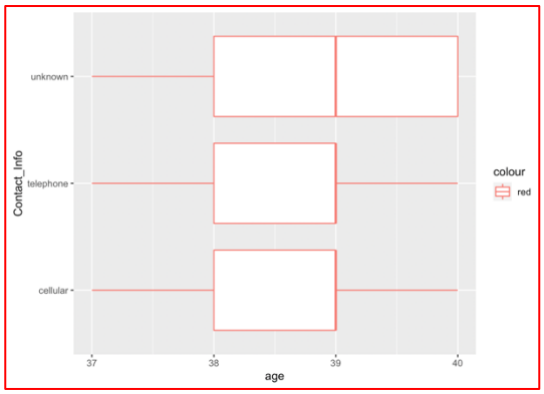
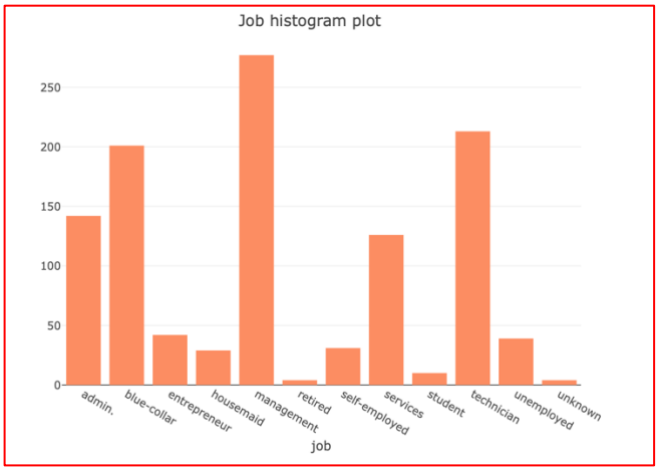


Figure 15 – Figure 16: Different contact info based on different ages



The bar plot below which created with plotly depicts the job proportion. Which 'management' has the highest proportion which is more than 250 counted. 'Blue-collar' and 'technician' has approximately the same amount which is almost 200. This is followed by 'admin' and 'services' that are counted 140 and 135 respectively. Meanwhile, student' and 'housemaid' has the lowest proportion and the number of unemployed is counted below 50. So, we can gain this that most jobs which are counted more than others are technically jobs which nowadays are became more popular and more useful one. (Figure 17)

Figure 17: Job histogram plot

Now, comparing the job with different age. First, I created a chart for comparing them (Figure 18) and then I created its box plot for be more understandable. (Figure 19)

As we can see; admin, entrepreneur, housemaid, management, and technician are being more occupancy in the range of 38 to 40 years old. While other jobs like services is only be specified in one age which is depicts in 38 years old. However, some other jobs which is depicts with unknown has a range of 37 years old occupancy.

```
#Top 6 jobs based on different ages
bankrank <- bank %>%

select(education,age, job)

Mostjob <- aggregate(age ~ job, bankrank, mean)
slice( Mostjob[order(-Mostjob$age),],
1:10)
```

job <chr>	age <dbl>
retired	40.00000
blue-collar	38.92040
housemaid	38.86207
unemployed	38.79487
admin.	38.64789
services	38.64286
self-employed	38.58065
technician	38.57746
entrepreneur	38.57143
management	38.53069

1-10 of 10 rows

Figure 18: The top 6 jobs based on different ages table

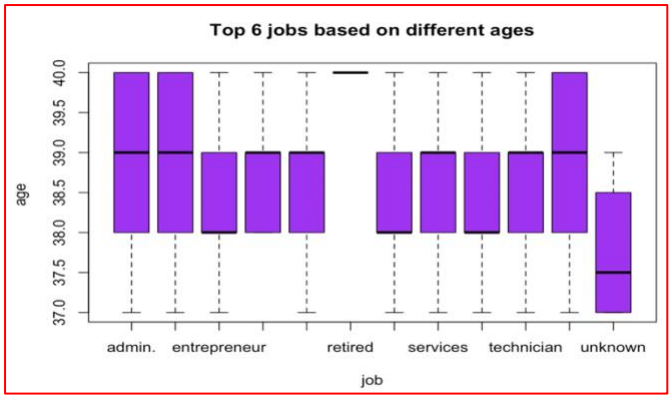


Figure 19: The top 6 jobs based on different ages box plot

Now, I created two different box plot which shows housing and loan based on different ages.
As it can be seen, the housing and loan has the direct connection with each other as from 38 years old till 40 years old are owner house and this range of age are also voted 'yes' for loan, so we can realize that this two variables has the direct connection with each other. (Figure 20 - Figure 21)

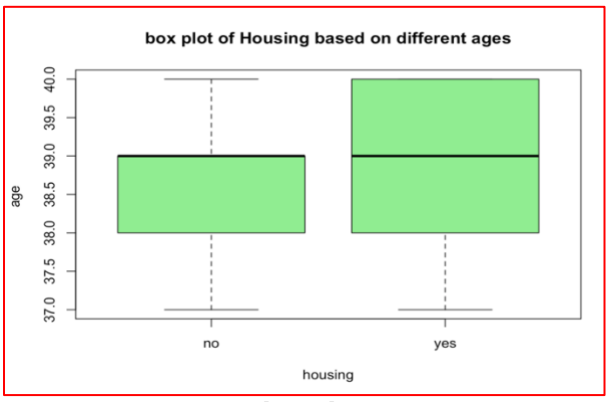


Figure 20: Housing box plot.

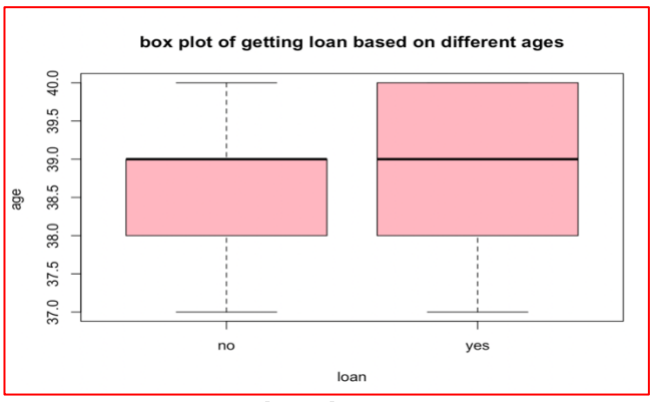


Figure 21: Loan box plot

The plotly histogram below depicts the loan which given in each day based on the ages from 37 to 40 years old. As it can be seen, the third day of month the loan proportion is at its peak and its more based on 28-29 years old people. While, in the first days with compared to the third days, has significantly lowest proportion of loan given. (Figure 22).

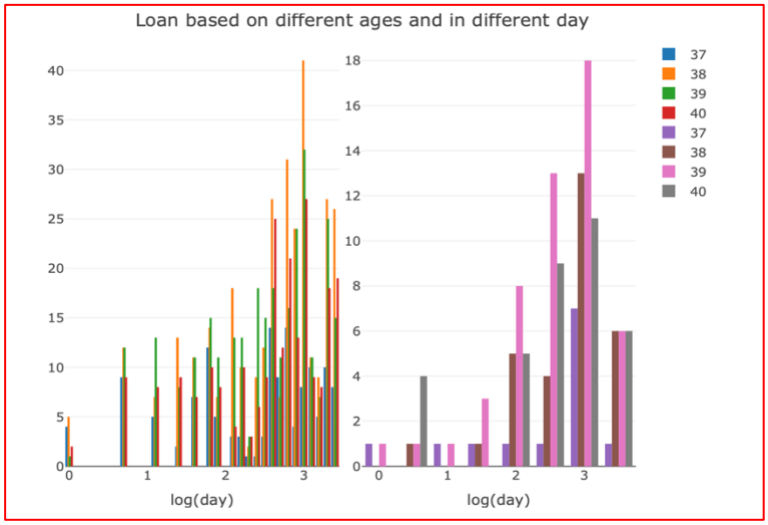
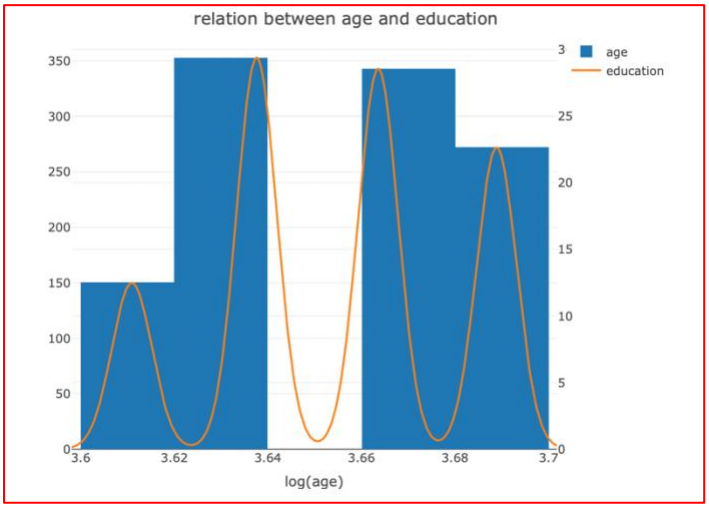


Figure 22: Loan based on different ages and in different day



Here, I created a plotly of age and education factors and added density axis to make it more accurate and understandable. (Figure 23)
As it can be seen from the chart, the education trend is fluctuating in different age range and its not affecting by increasing age as it means education can be all the time be something to consider about it.

Figure 23: relation between age and education

The plotly below shows that accumulation of campaign counted was high in the beginning up to 500 duration and after that it became scattered after this duration. Also, from density axis we can see that its mostly in the range of -1 to 2 up to 500 duration. (Figure 24).

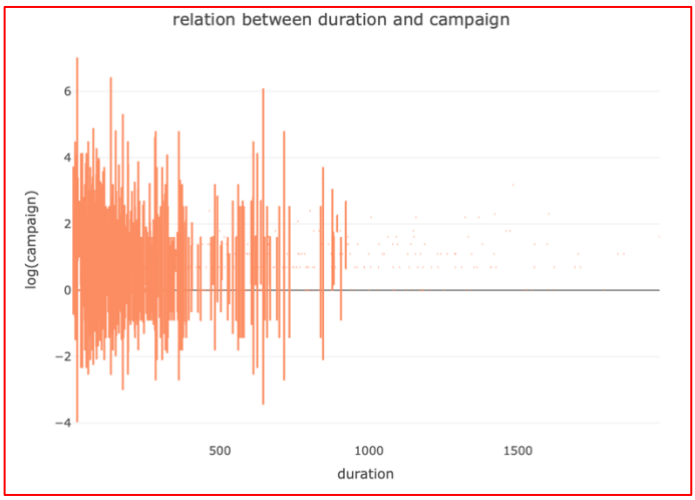
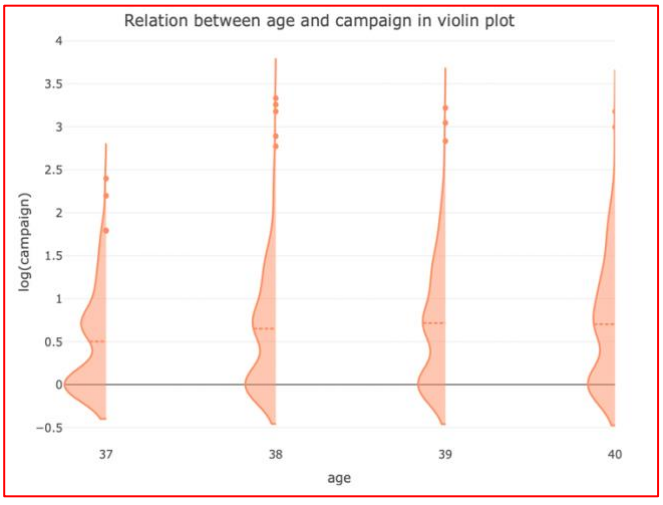


Figure 24: relation between duration and campaign



The last plotly of age and campaign shows the mean, median and upper face of each age range based on campaign variable. From their mean, we can see that 39 and 40 years old have approximately the same mean amount and its more than the previous ages. (Figure 25)

The shape of the distribution indicates the campaign for different ages are highly concentrated around the median.

Figure 25: Relation between age and campaign in violin plot

Part C: Summary

Summary

In this project, I used the bank dataset which I got from Kaggle website and imported my R practice code and added some other functions like 'subset' to find more information about my variables.

First of all, I research on bank system and its different variables like age, loan, housing, different contact info and etc. After cleaning my dataset, I started analysing them by providing different tables that gave me information about the proportion of different ages in compare with other factors like housing, loan and etc.

After that, creating different diagrams like job which gave information about different jobs which demonstrate that management has the highest proportion within the others. Then I compare these jobs with age factor in plotly that shows management, technician, admin, and entrepreneur are being more occupancy in the range of 38 to 40 years old.

Moreover, with another model of diagram, boxplot, which I also created for job, I be able to compare its top 6 job with different ages in a different and easier to understand way.

At the end, I created a theory for myself at first to compare some factors like age, different usage based on contact, education and its relation with job, housing and getting loan and etc. which are the nowadays useful factors that all of us are familiar with it, so based on these factors, I created some different diagram and tables based my need to be able to compare them and gain accurate information about them to analyse.

References

1. Kabacoff, R. I. (2015). R in Action SECOND EDITION Data analysis and graphics with R (2nd ed.) by Manning Publications Co.
2. Ploty Graphics Library (2021). Bar Charts in R. Https://Plotly.Com/. https://plotly.com/r/barcharts/
3. Quick-R by Datacamp (2017). Subsetting Data. Www.Statmethods.Net. https://www.statmethods.net/management/subset.html
4. Stackoverflow (2016) .R "Error: unexpected '}' in"} "[duplicate]. Https://Stackoverflow.Com/. https://stackoverflow.com/questions/40291675/r-error-unexpected-in
5. *Source of dataset: https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset *
6. *Reference R code : file:///Users/shamimsherafati/Documents/ALY%206010/w1/Sherafati_M1-Week1-R-Practice.html *