Shamim Sherafati

R Practice – ALY 6010

Module 1
Week 1

First, I imported libraries and then After importing the dataset “bank.csv”, I started to clean the data for further analysis.

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

Firstly, I sorted my data with descending age column and then the column ‘default’ is dropped as it has no impact towards the comparison of data. Later, column ‘contact’ was renamed to ‘Contact_Info’ to be more specific.

After calculating the top 50% and the bottom 45% of data, the rows are trimmed down from 11,162 rows to 1,118 rows.

At the end, ‘marital’ column variables were replaced from small letters to capital letters.

	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 1 shows all the data set’s variable and their data types.

```
> 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 1: Bank Data Set Structure

Looking at the below data, specifically the frequency of yes and no for the Deposit and Age variables in Figure 2, 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 2: Frequency Tables for ***Deposit and Age***

Looking at the below data for the Contact info and Marital Status variables in Figure 3, 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.

```
> ftable(bank_Ftable3)
           DIVORCED MARRIED SINGLE

cellular         90     487    197
telephone         3      31     14
unknown          49     177     70
>
```

Figure 3: Frequency Tables for ***Contact info and Marital Status***

As we can see in Figure 4 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.

```
> #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 |           |
-------------|-----------|-----------|-----------|
```

Figure 4: 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.

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 5)*

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.

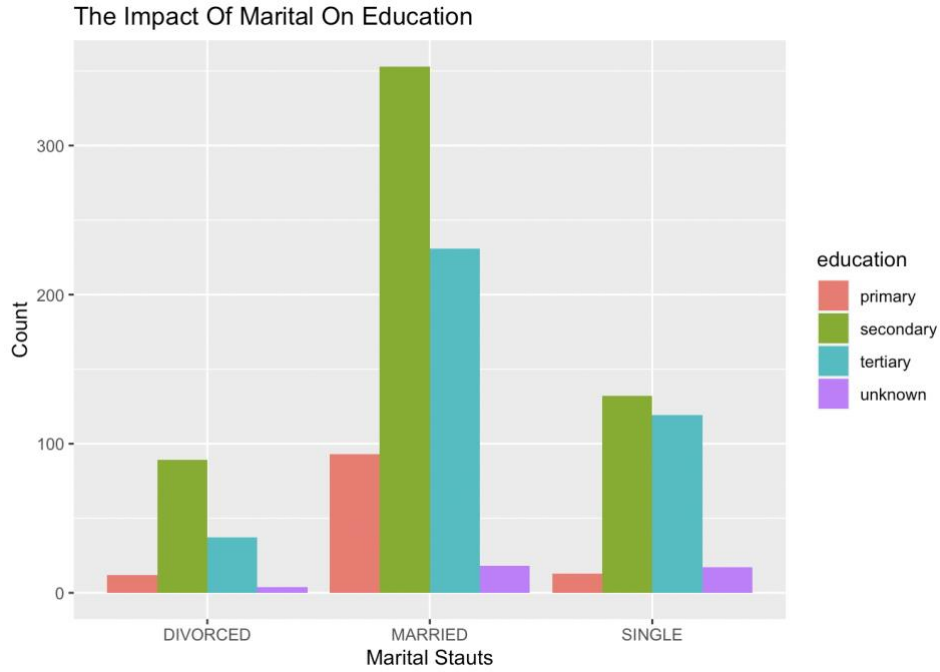

Figure 5: The impact of marital on education

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 6)*

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.

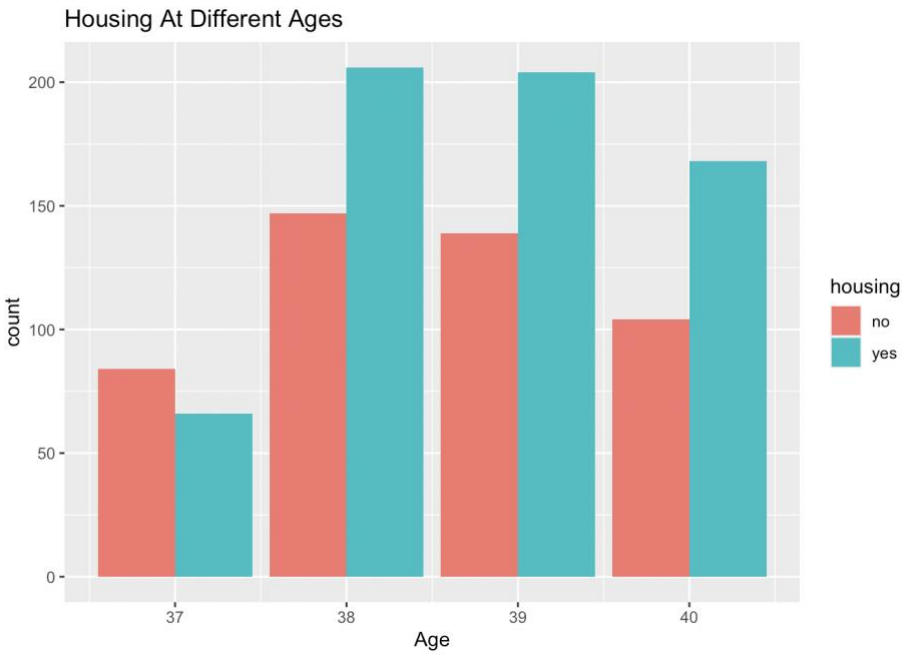

Figure 6: Housing at different ages

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 7)*

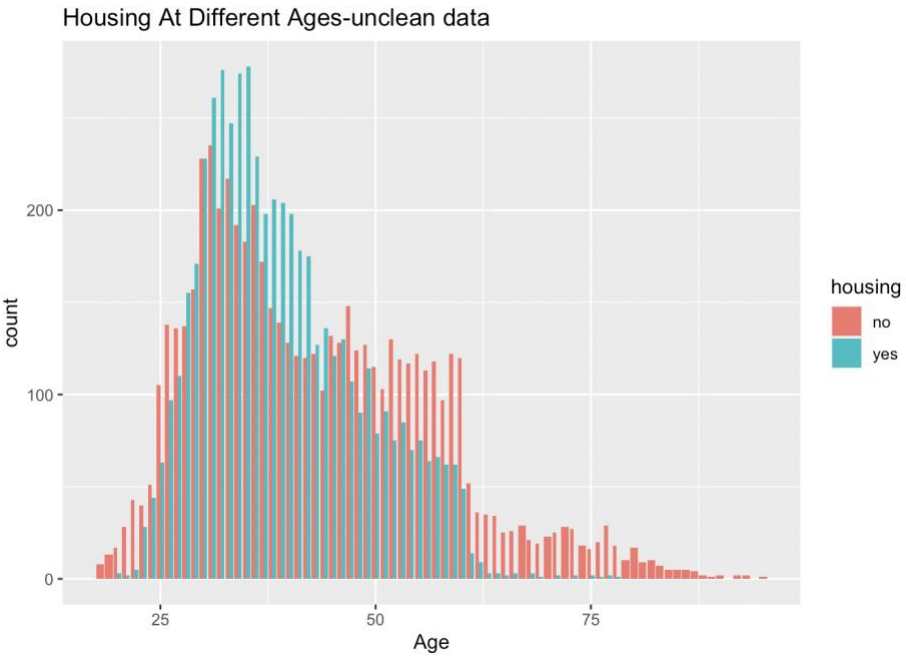

Figure 7: Housing at different ages- unclean data

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 8 - Figure 9)*

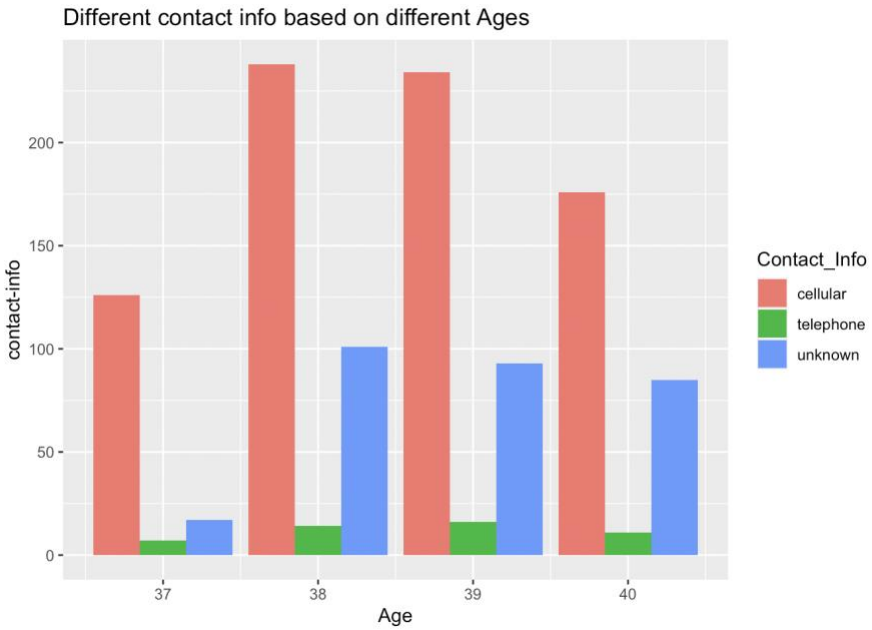

Figure 8: Different contact info based on different ages

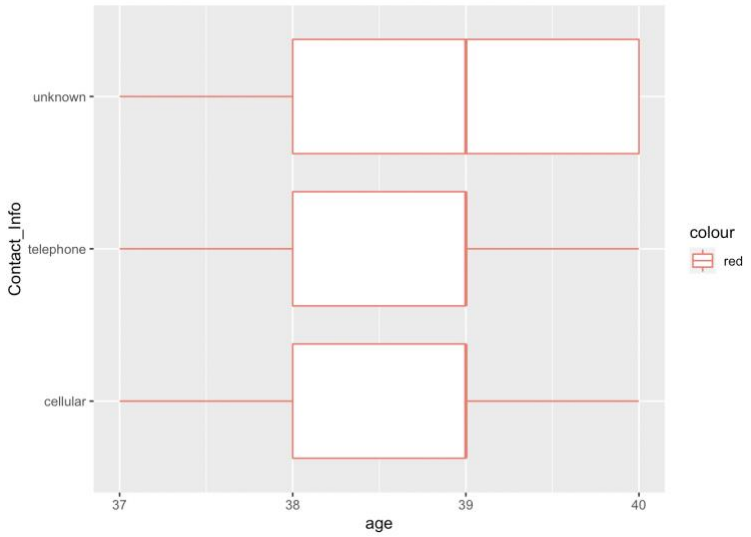

Figure 9: Different contact info based on different ages

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 10)*

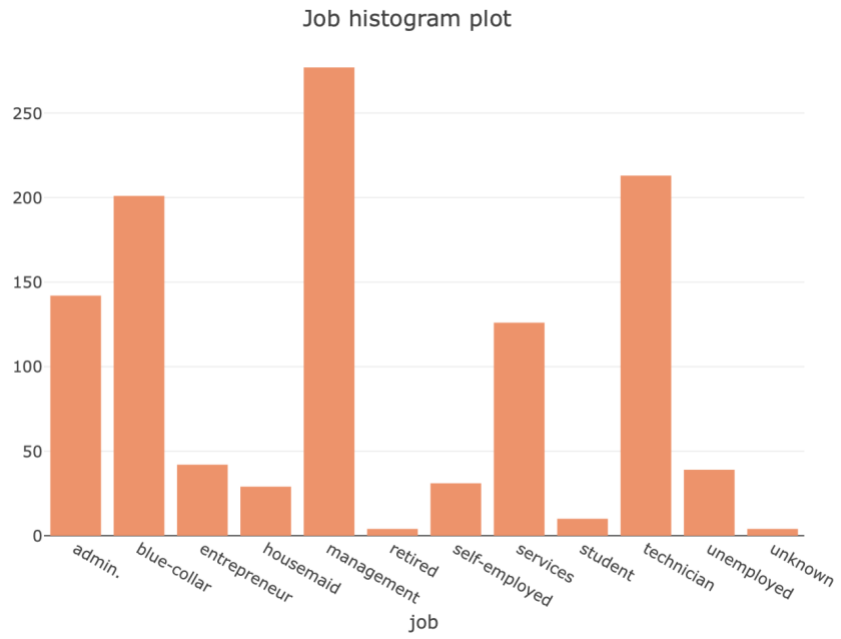

Figure 10: Job histogram plot

Now, comparing the job with different age.
First, I created a chart for comparing them (*Figure 11*) and then I created its box plot for be more understandable. (*Figure 12*)

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 11: The top 6 jobs based on different ages table

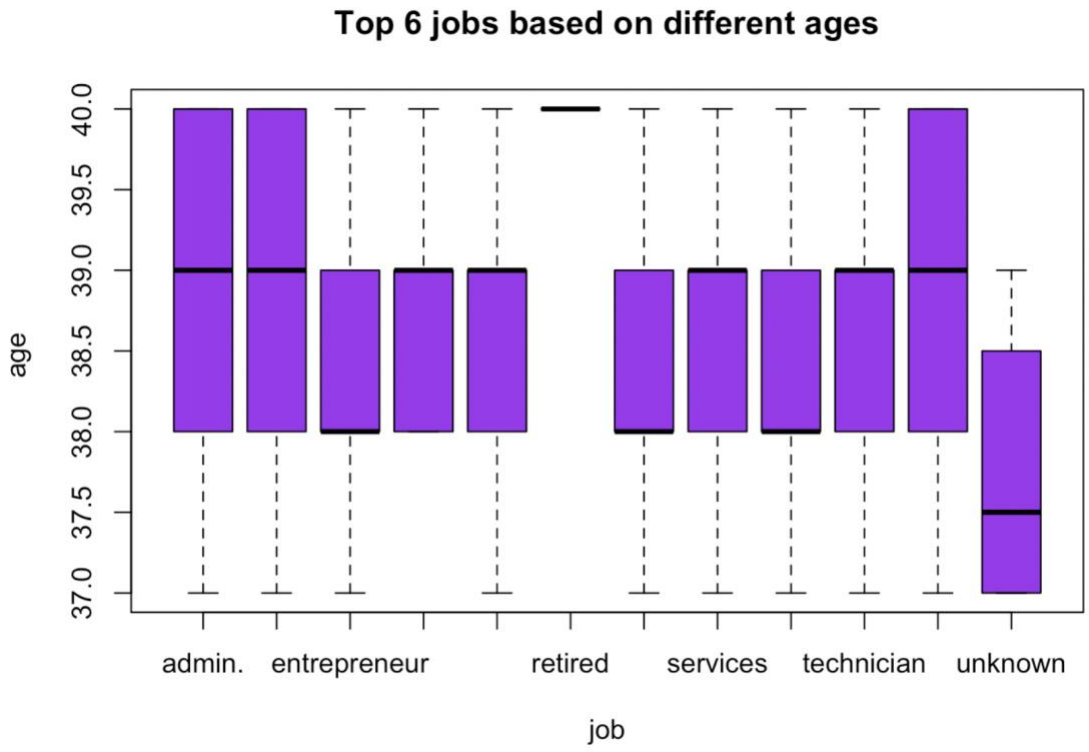

Figure 12: The top 6 jobs based on different ages box plot

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 13 - Figure 14)*

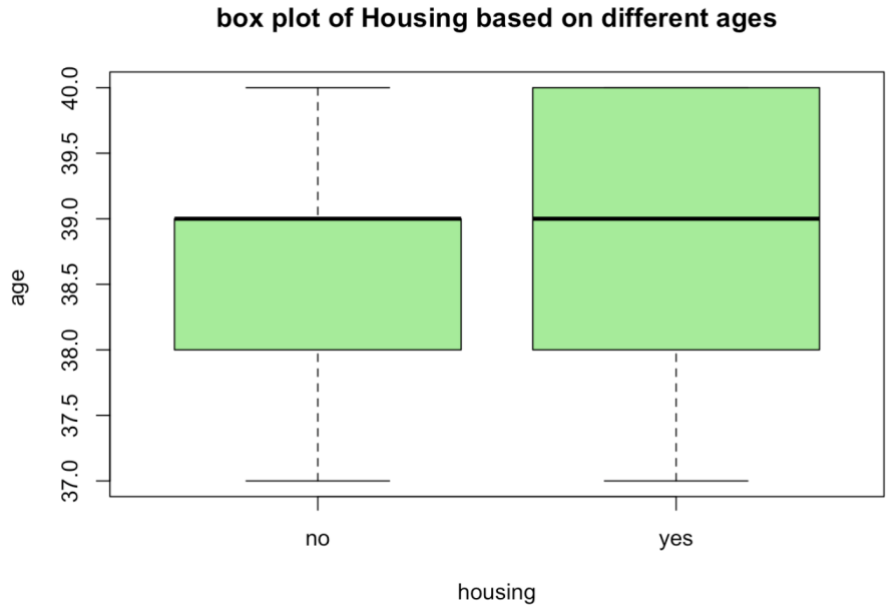

Figure 13: Housing box plot

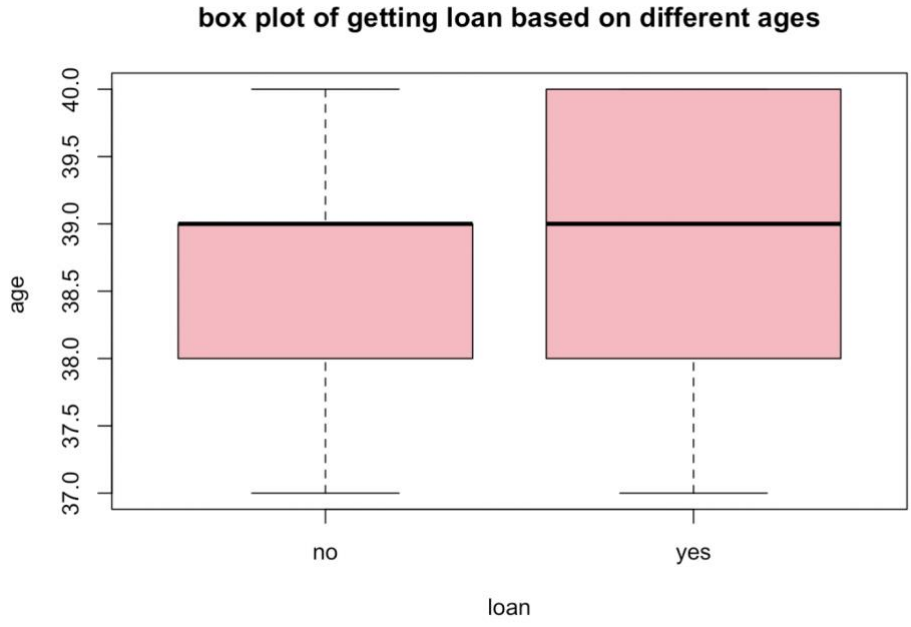

Figure 14: Loan box plot

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 15)*

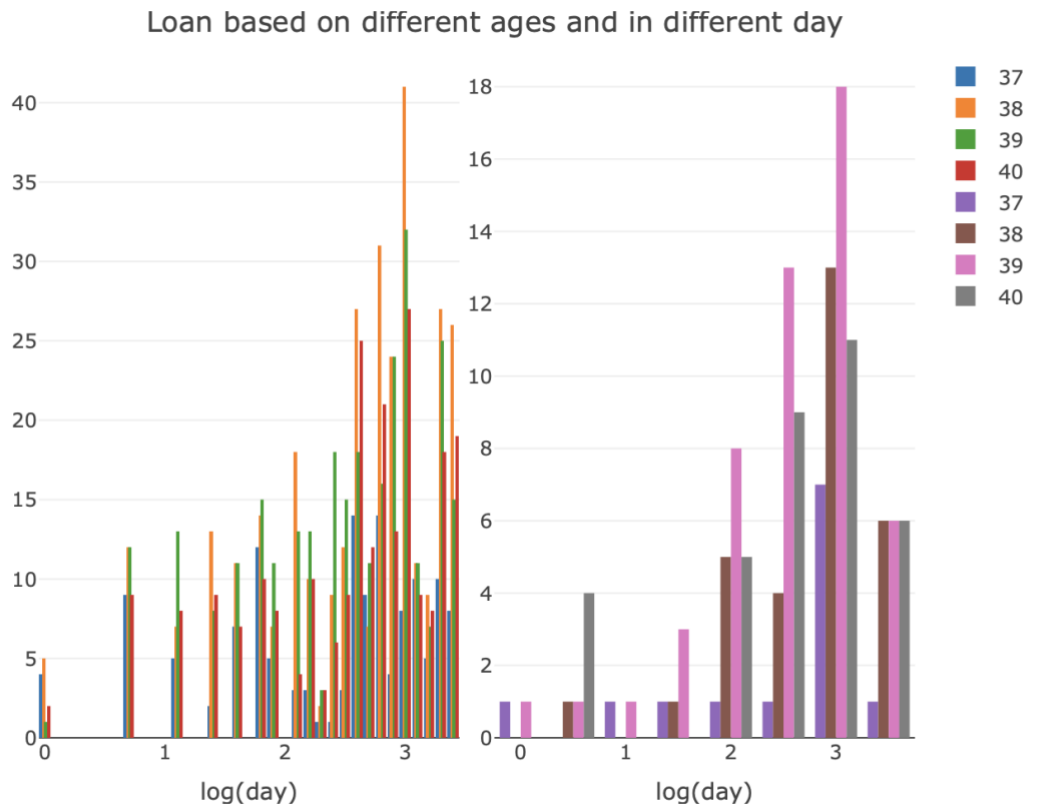

Figure 15: Loan based on different ages and in different day

Here, I created a plotly of age and education factors and added density axis to make it more accurate and understandable. (*Figure 16)*
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.

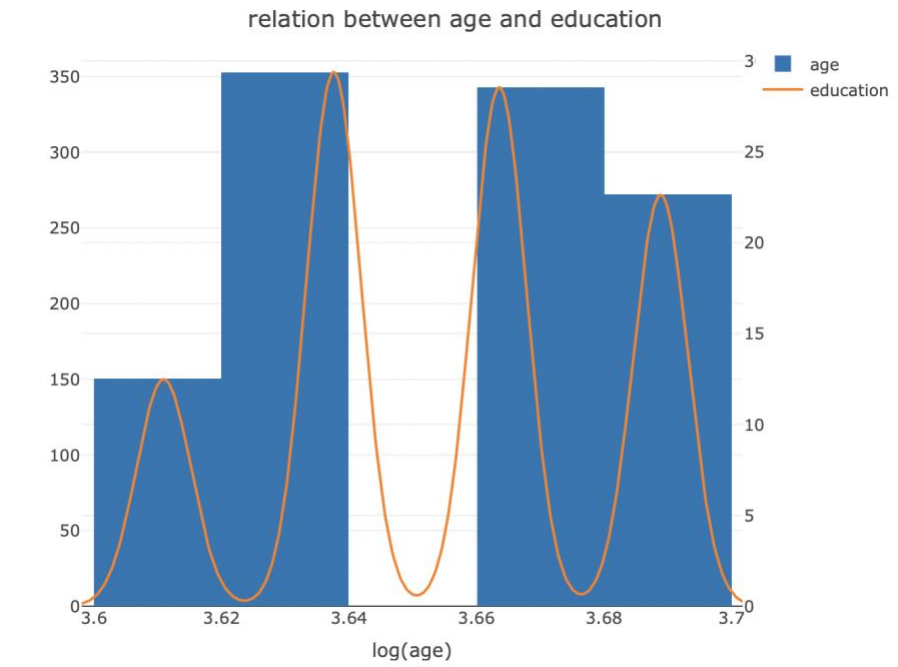

Figure 16: relation between age and education

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 17)*

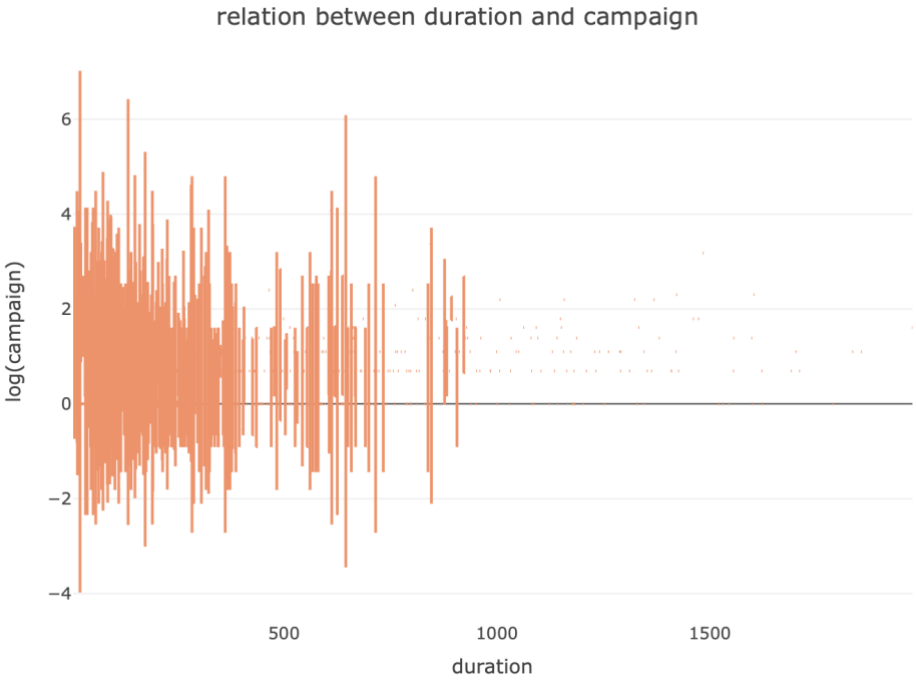

Figure 17: relation between duration and campaign

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 18)*

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

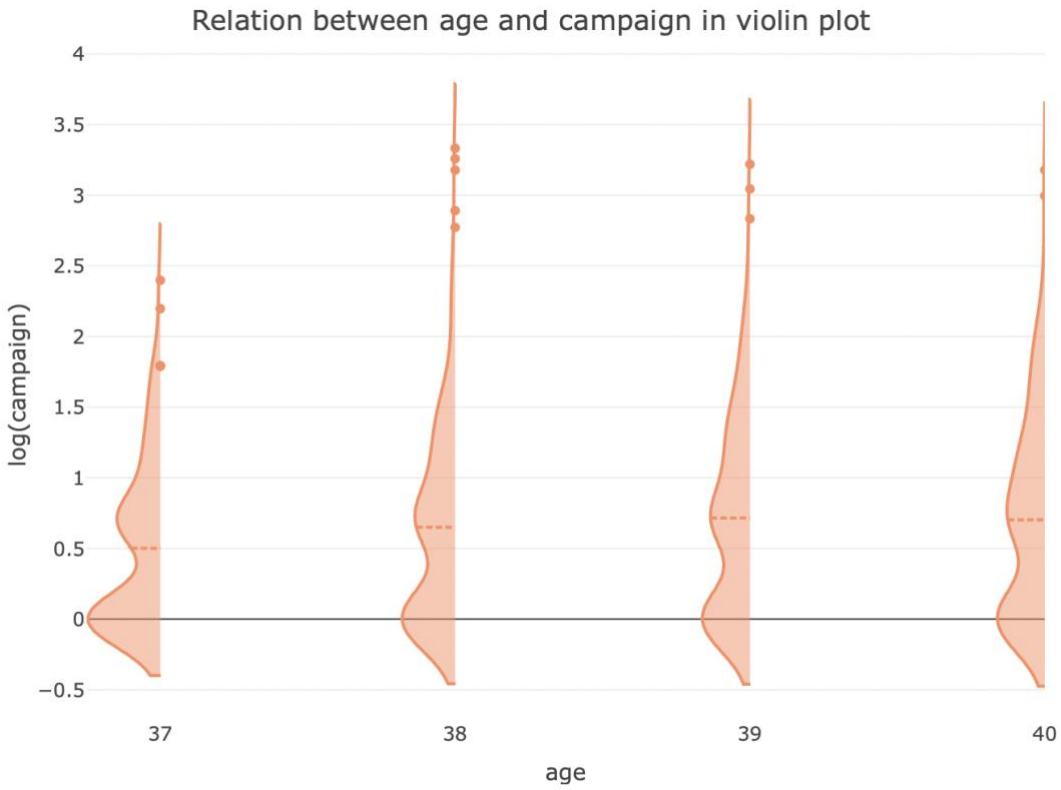

Figure 18: Relation between age and campaign in violin plot

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