Shamim Sherafati

ALY 6010

Instructor: Mohsen Soltanifar

Module 4 – Final project Milestone 2

Date: 2022/12/02

Part A: Introduction

A data analyst should be able to analyze any dataset by understanding a few essential components. Does the dataset attempt to explanations, illustrate a relationship, or showcase something? Once we have the answers to any of these queries, we may attempt to analyze and determine any connections between the dataset something to the user?

Data Analysis

Once we are aware of the data's requirements and have a look at the dataset's various columns. By dividing the entire data set into smaller data sets, we were able to start making connections and meaning of the data as a few things started to become evident. A data analyst may create graphical representations using these smaller datasets. The dataset has the following features:

- The dataset has a total of 11162 rows and 17 fields which after cleaning decreased to 1118 rows and 16 fields.
- This dataset consists of both textual and numeric fields, and we must convert it to factor format in order to use it for analysis.
- 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.

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.

Data source

I get this dataset from Kaggle website and used my previous; ALY 6010, Final project milestone 1 week 1 and referred to my previous r code for analyzing. (Source: file:///Users/shamimsherafati/Documents/Northeastern%20University/ALY%206010/week%201/milestone%201/Sherafati_M1-Final-Project--Milestone1.html)

In this Project, I created 3 different question from the bank dataset and did hypothesis testing and inferential statistics on them.
The dataset is cleaned based on the cleaning process which I done in 'Final project Milestone 1'.

Part B: Data Analysis

Section 1: One-sample t-test

Question 1: Do Balance of account of customers greater than 1000 or not?

In this question, I wonder of the customer's account balance as if greater than specific amount or not. Based on this purpose, first I examine the data.

The histogram below, depicts the account balance of customers. Although most of them has the balance below 2000 but as it is clear from the chart, the highest amount of their balance is up to 8000.

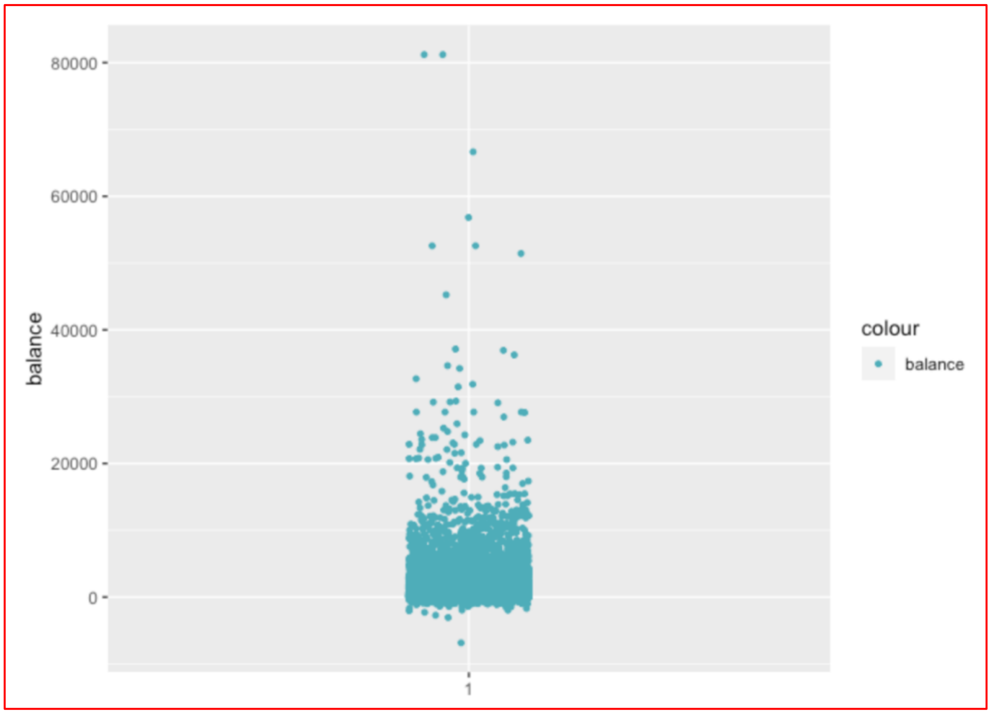
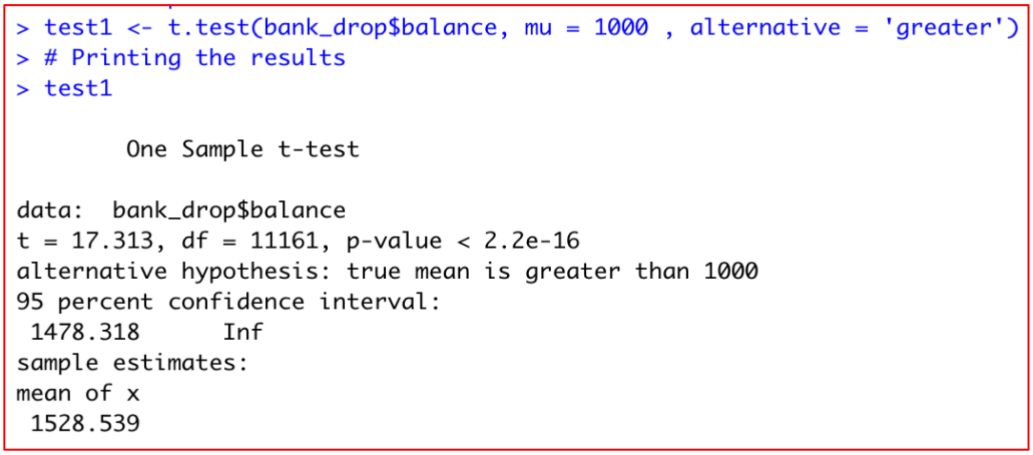


Figure 1: Plot of account balance

Now, I created a one sample t-test for this question. First, I get the average of account balance which is 1528.539. Then, as I want to know whether it is greater than 1000 or not, I used right tail in my hypothesis test. Our null and alternate hypothesis are:

H0 (claim)= The account balance is equal or lower than 1000
H1= The account balance is more than 1000

Figure 2: Null and Alternate hypothesis



```
> test1 <- t.test(bank_drop$balance, mu = 1000 , alternative = 'greater')
> # Printing the results
> test1

	One Sample t-test

data:  bank_drop$balance
t = 17.313, df = 11161, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 1000
95 percent confidence interval:
 1478.318      Inf
sample estimates:
mean of x
 1528.539
```

Figure 3: One sample t-test

Here in this test, p value is less than 2.2e-16 with 95% confidence interval running from 1478.318. When compared to the significance level of 0.05, the P-value of 2.2e-16 is extremely low. As a result, I reject the null hypothesis H0. Therefore, I can say that the account balance of customers is more than 1000.

After that, I do statistical measures for this question. First get the mean and median of balance:

Mean of balance: 1528.539
Median of balance: 550

As we can see, the mean is greater than the median and it is obvious in the above diagram as its skewed to the right.

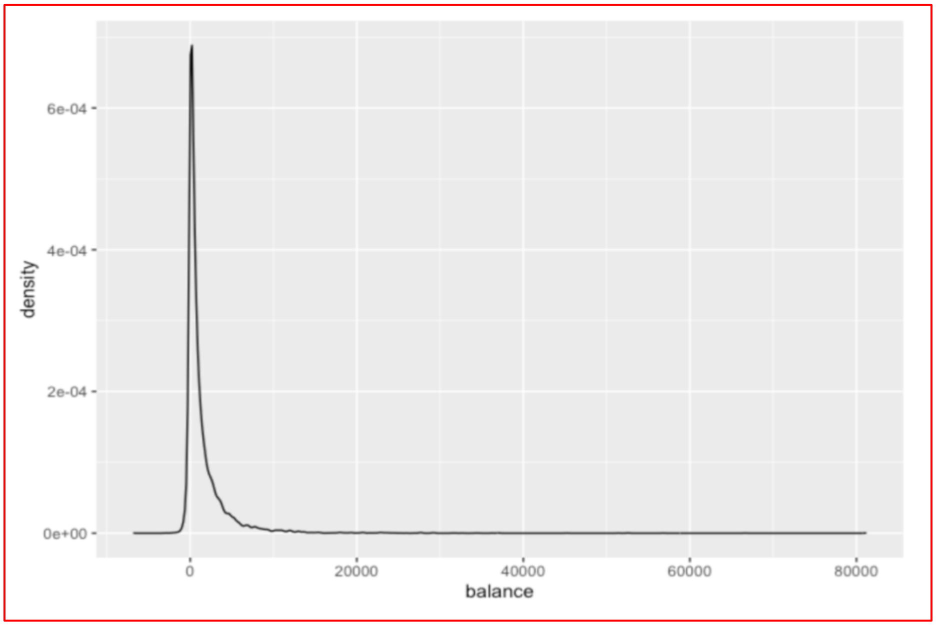
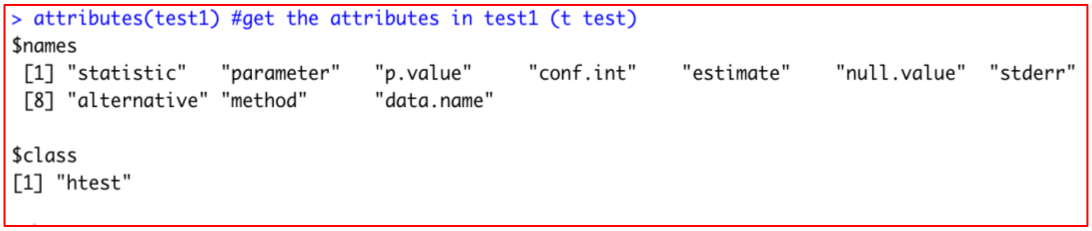


Figure 4: Right skewed diagram

Next, I get the attributes that stored in the t-test of balance and extract some of them:

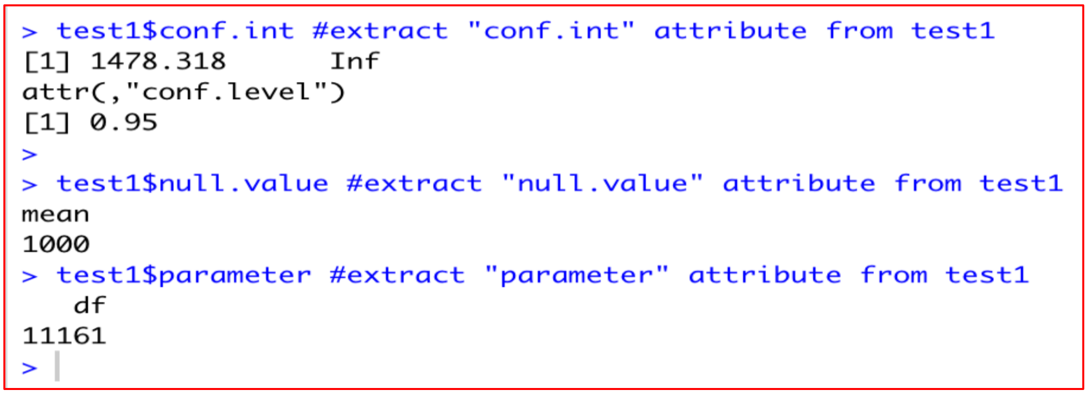


```
> attributes(test1) #get the attributes in test1 (t test)
$names
 [1] "statistic"   "parameter"   "p.value"     "conf.int"    "estimate"    "null.value"  "stderr"
 [8] "alternative" "method"      "data.name"

$class
[1] "htest"
```

Figure 5: Attributes of data

We have 10 attributes from balance t-test and from these, I extract the conf.int, null.value and parameter as follows which shows the mean, the confident level and the parameter from the test:



```
> test1$conf.int #extract "conf.int" attribute from test1
[1] 1478.318      Inf
attr(,"conf.level")
[1] 0.95
>
> test1$null.value #extract "null.value" attribute from test1
mean
1000
> test1$parameter #extract "parameter" attribute from test1
   df
11161
>
```

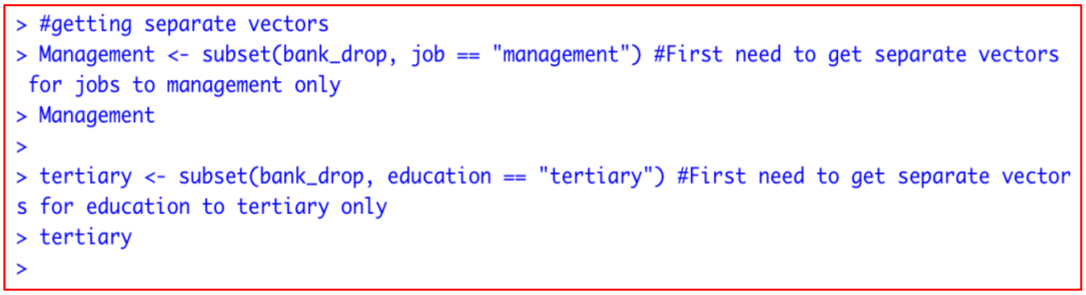
Figure 6: Extract attributes from data

Section 2: Two-sample t-test:

Question 2: Working as a Management, must need a graduate degree from university. Is that true?

In this question, I select one job, management, and took hypothesis test to see if those who work as a manager, must need university degree or not.

For this purpose, I got the separate vectors from the job and education column as a management and tertiary respectively.



```
> #getting separate vectors
> Management <- subset(bank_drop, job == "management") #First need to get separate vectors
 for jobs to management only
> Management
>
> tertiary <- subset(bank_drop, education == "tertiary") #First need to get separate vector
s for education to tertiary only
> tertiary
>
```

Figure 7: Subset of vectors

Based on this purpose, first I examine the data.

This bar graph, depicts the all of these different education level can participate in management job but the proportion of those with university degree is significantly higher.

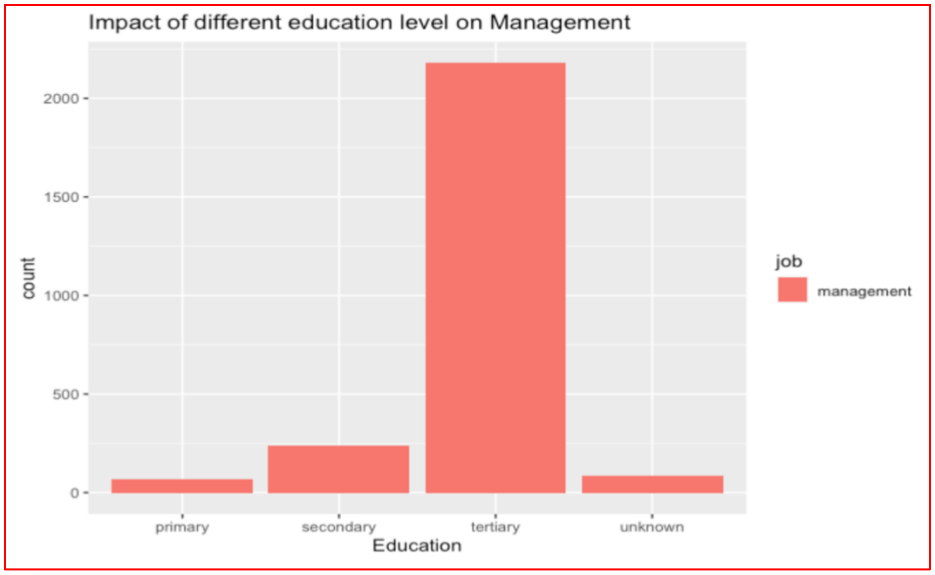
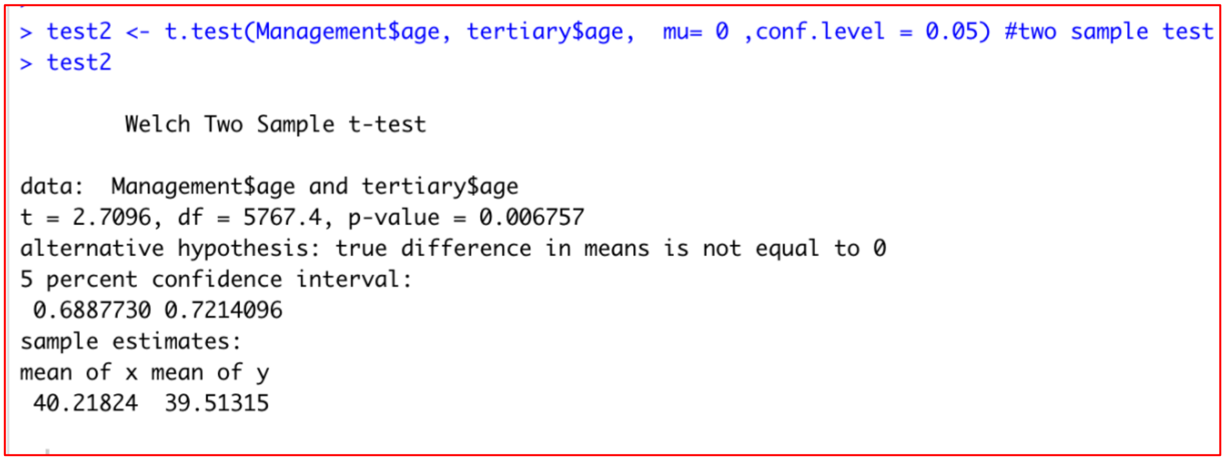


Figure 8: Impact of different education level on 'management'

Now, I created a two sample t-test for this question.

H0 (claim)= Management job need university degree.
H1= Management job may not need university degree.

Figure 9: Null and Alternate hypothesis



```
> test2 <- t.test(Management$age, tertiary$age,  mu= 0 ,conf.level = 0.05) #two sample test
> test2

	Welch Two Sample t-test

data:  Management$age and tertiary$age
t = 2.7096, df = 5767.4, p-value = 0.006757
alternative hypothesis: true difference in means is not equal to 0
5 percent confidence interval:
 0.6887730 0.7214096
sample estimates:
mean of x mean of y 
 40.21824  39.51315 
```

Figure 10: Two sample t-test

Here in this test, p value is 0.006757 with 5% confidence interval running from 0.6887730 to 0.7214096. When compared to the significance level of 0.05, the P-value of 0.006757 is lower. As a result, I reject the null hypothesis H0 and it means that having a university degree is not essential to work in management sector.

After that, I do statistical measures for this question and create a chart as below:

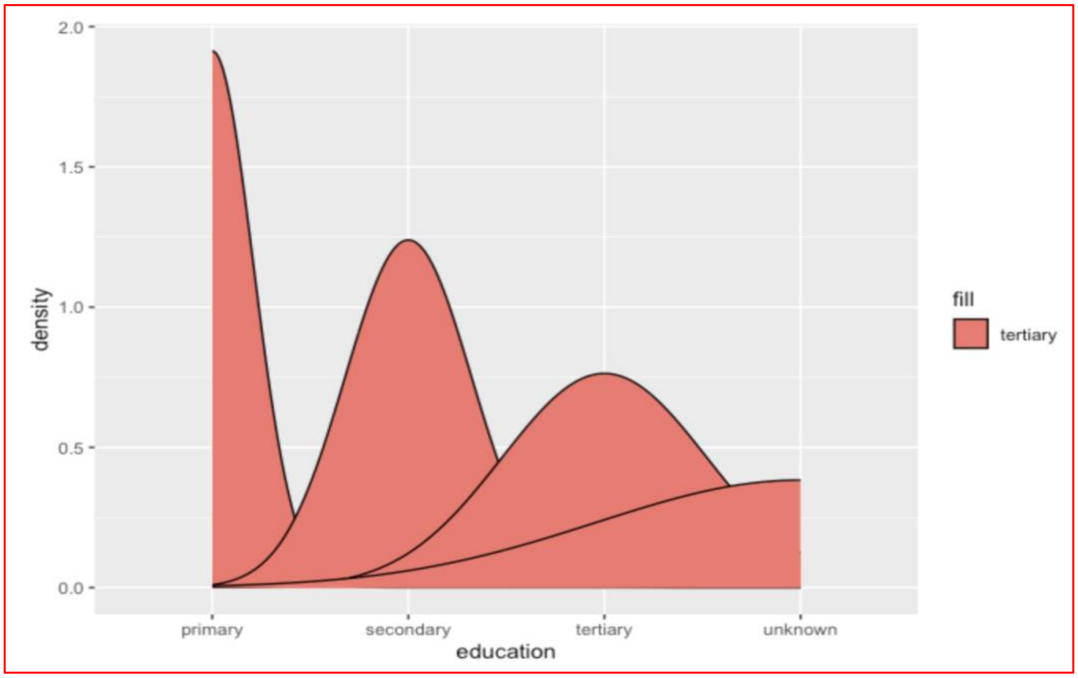
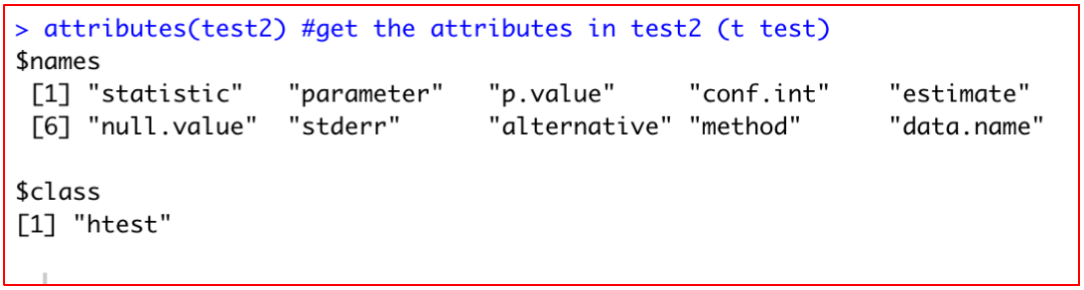


Figure 11: Skewed shape diagram

Next, I get the attributes that stored in the two sample t-test in question 2 and extract some of them.

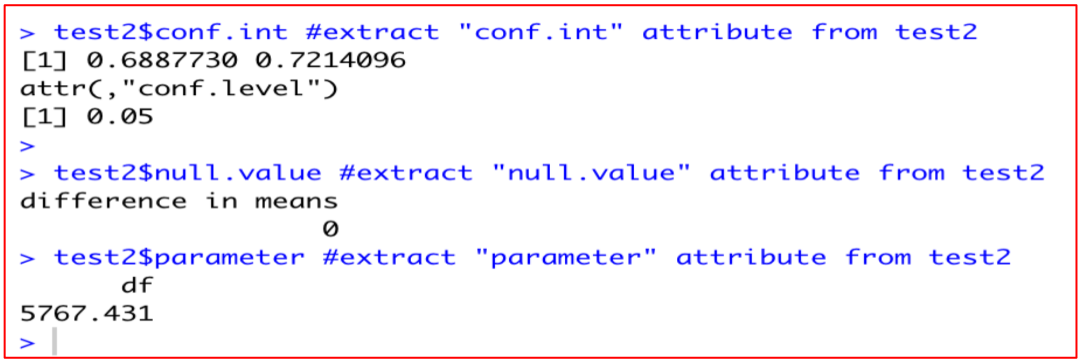


```
> attributes(test2) #get the attributes in test2 (t test)
$names
 [1] "statistic"   "parameter"   "p.value"     "conf.int"    "estimate"
 [6] "null.value"  "stderr"      "alternative" "method"      "data.name"

$class
[1] "htest"
```

Figure 12: Attributes of data

We have 10 attributes from balance t-test and from these, I extract the conf.int, null.value and parameter as follows which shows the difference in mean which is zero, the confident level of 0.05 and the parameter from the test:



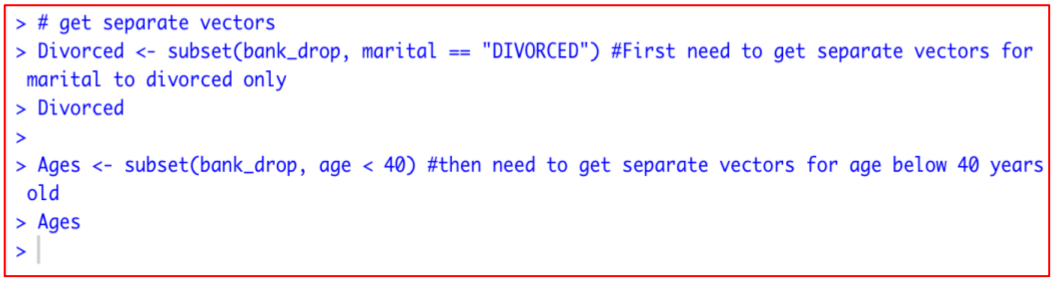
```
> test2$conf.int #extract "conf.int" attribute from test2
[1] 0.6887730 0.7214096
attr(,"conf.level")
[1] 0.05
>
> test2$null.value #extract "null.value" attribute from test2
difference in means
                  0
> test2$parameter #extract "parameter" attribute from test2
      df
5767.431
>
```

Figure 13: Extract attributes from data

Question 3: Most of divorced happened at early age. Is it true or not?

In created this question to gain information that at which the most of the divorced happened, if it mostly happen in early age means below 40 years old or not.

For this purpose, I got the separate vectors from the marital status and age column as a divorced and below 40 years old respectively.



```
> # get separate vectors
> Divorced <- subset(bank_drop, marital == "DIVORCED") #First need to get separate vectors for
 marital to divorced only
> Divorced
>
> Ages <- subset(bank_drop, age < 40) #then need to get separate vectors for age below 40 years
 old
> Ages
>
```

Figure 14: Subset of vectors

Based on this purpose, I examine the data before creating t-test:

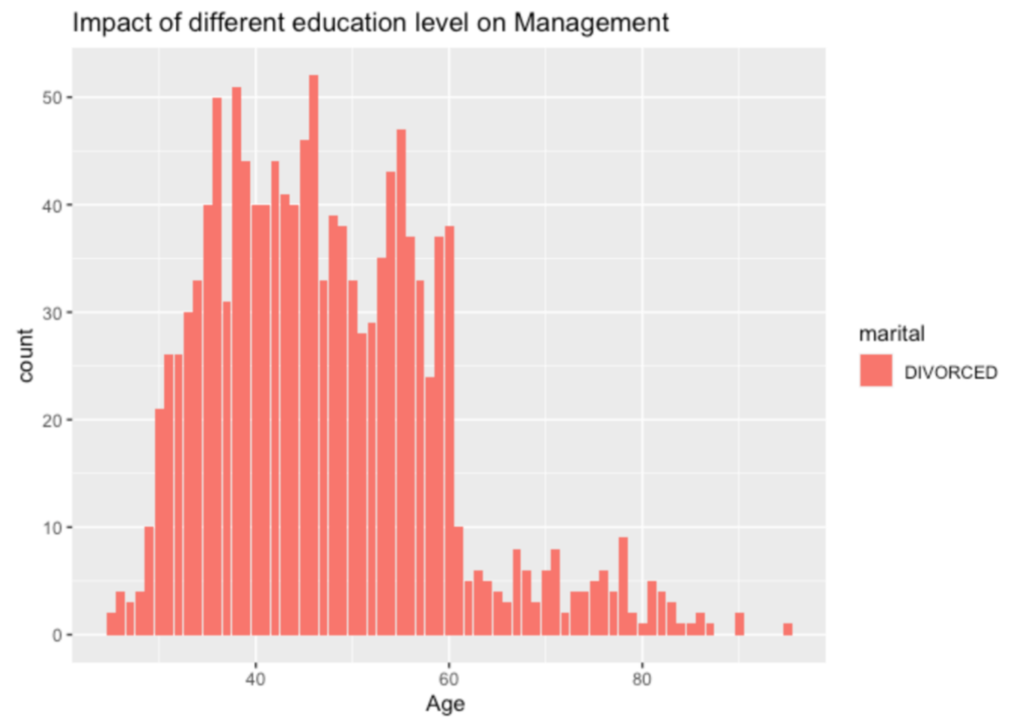
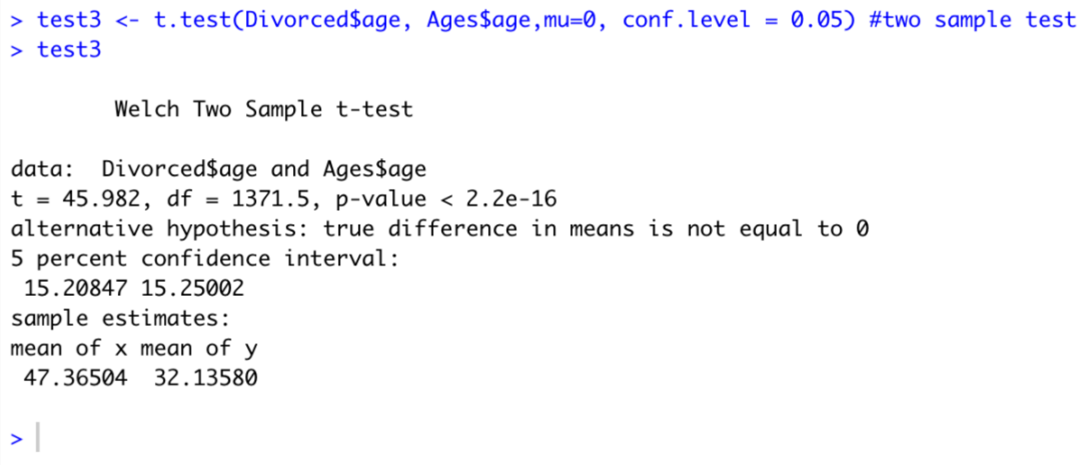


Figure 15: Histogram of marital status on ages

The above histogram, depicts divorced proportion in different ages. As we can see, between 40-50 years old, most of the divorced happened. Now, I created a two sample t-test for this question.

H0 (claim)= Most of the divorce happened at 40 years old or below.
H1= Most of the divorce happened after 40 years old.

Figure 16: Null and Alternate hypothesis



```
> test3 <- t.test(Divorced$age, Ages$age,mu=0, conf.level = 0.05) #two sample test
> test3

        Welch Two Sample t-test

data:  Divorced$age and Ages$age
t = 45.982, df = 1371.5, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
5 percent confidence interval:
 15.20847 15.25002
sample estimates:
mean of x mean of y
 47.36504  32.13580

> 
```

Figure 17: Two sample t-test

Here in this test, p value is less than 2.2e-16 with 5% confidence interval running from 15.20847 to 15.25002. When compared to the significance level of 0.05, the P-value is extremely lower than sig.level. As a result, I reject the null hypothesis H0 and I can say that most of the divorced happened after 40 years old.

After that, I do statistical measures for this question and create a chart as below:

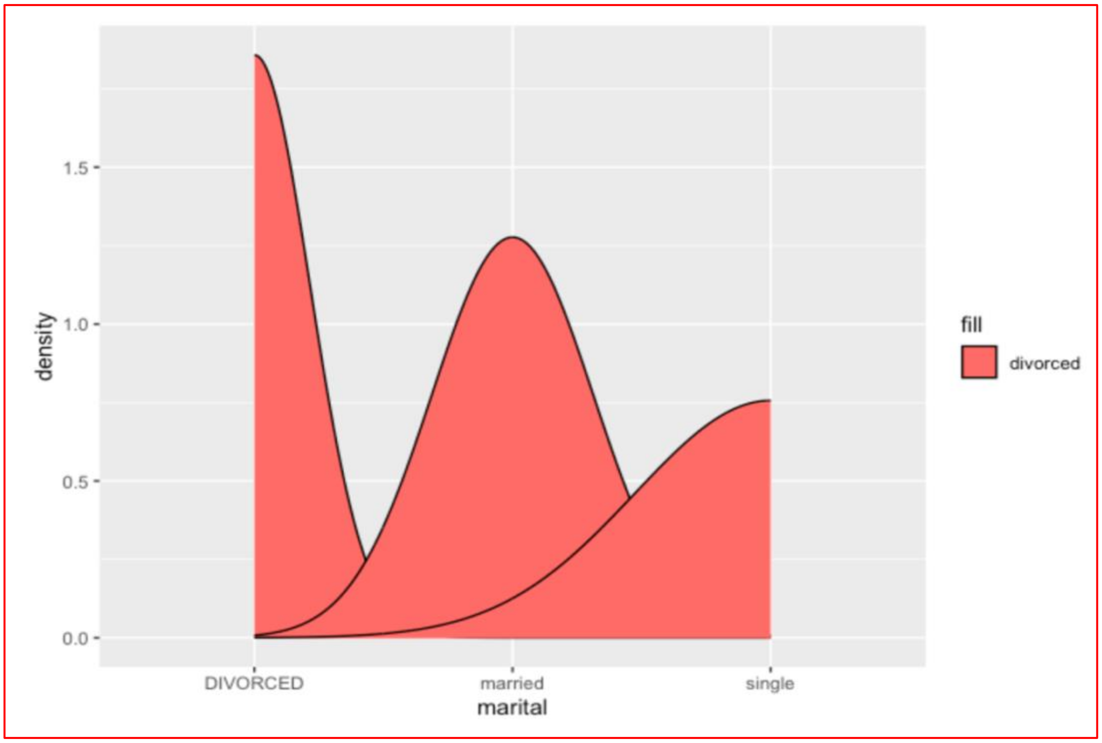
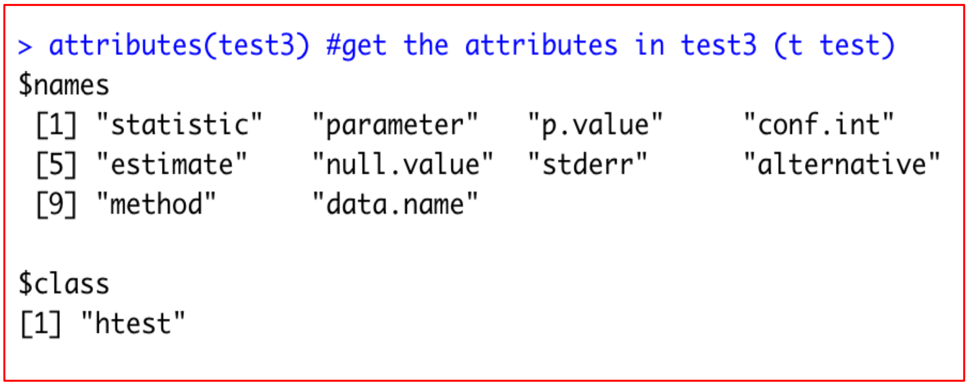


Figure 18: Skewed shape diagram

Next, I get the attributes that stored in the two sample t-test in question 3 and extract some of them:

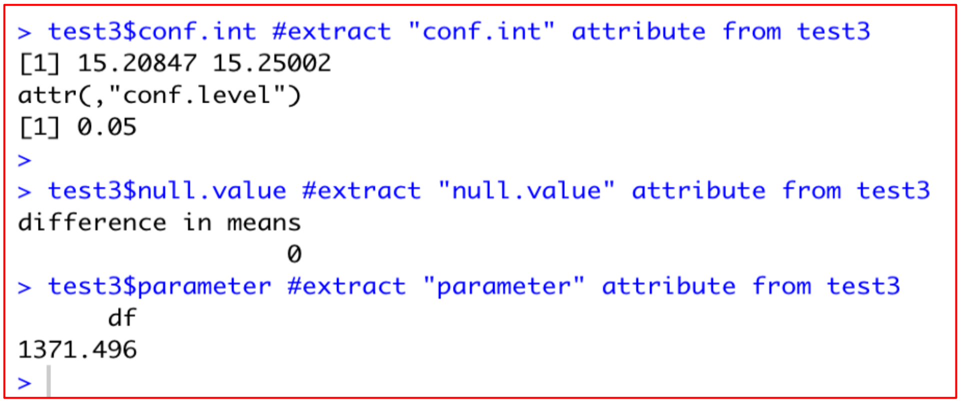


```
> attributes(test3) #get the attributes in test3 (t test)
$names
 [1] "statistic"   "parameter"   "p.value"     "conf.int"
 [5] "estimate"    "null.value"  "stderr"      "alternative"
 [9] "method"      "data.name"

$class
[1] "htest"
```

Figure 19: Attributes of data

We have 10 attributes from balance t-test and from these, I extract the conf.int, null.value and parameter as follows which shows the difference in mean which is zero, the confident level of 0.05 and the parameter from the test:



```
> test3$conf.int #extract "conf.int" attribute from test3
[1] 15.20847 15.25002
attr(,"conf.level")
[1] 0.05
>
> test3$null.value #extract "null.value" attribute from test3
difference in means
                  0
> test3$parameter #extract "parameter" attribute from test3
      df
1371.496
>
```

Figure 20: Extract attributes from data

Part C: Summary

Summary

In this project, I used the bank dataset which I got from Kaggle website and imported my Milestone 1 code and do some test on it.

First of all after cleaning my dataset, I provide three different questions

References

1. Kabacoff, R. I. (2015). R in Action SECOND EDITION Data analysis and graphics with R (2nd ed.) by Manning Publications Co.
2. I. Miller and M. Miller, John E. Freund's Mathematical Statistics with Applications; 7th edition; Pearson Prentice-Hall; Upper Saddle River, NJ; 2004.
3. Y. M. M. Bishop and S. E. Fienberg, Incomplete Two-Dimensional Contingency Tables, Biometrics 25 (1969), no. 1, 119–128.
4. R. Johnson and G. Bhattacharyya, Statistics: Principles and Methods; 3rd edition; John Wiley and Sons, Inc.; 1996.
5. *Source of dataset: https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset *
6. *Reference R code: file:///Users/shamimsherafati/Documents/Northeastern%20University/ALY%206010/week%201/milestone%201/Sherafati_M1-Final-Project--Milestone1.html *