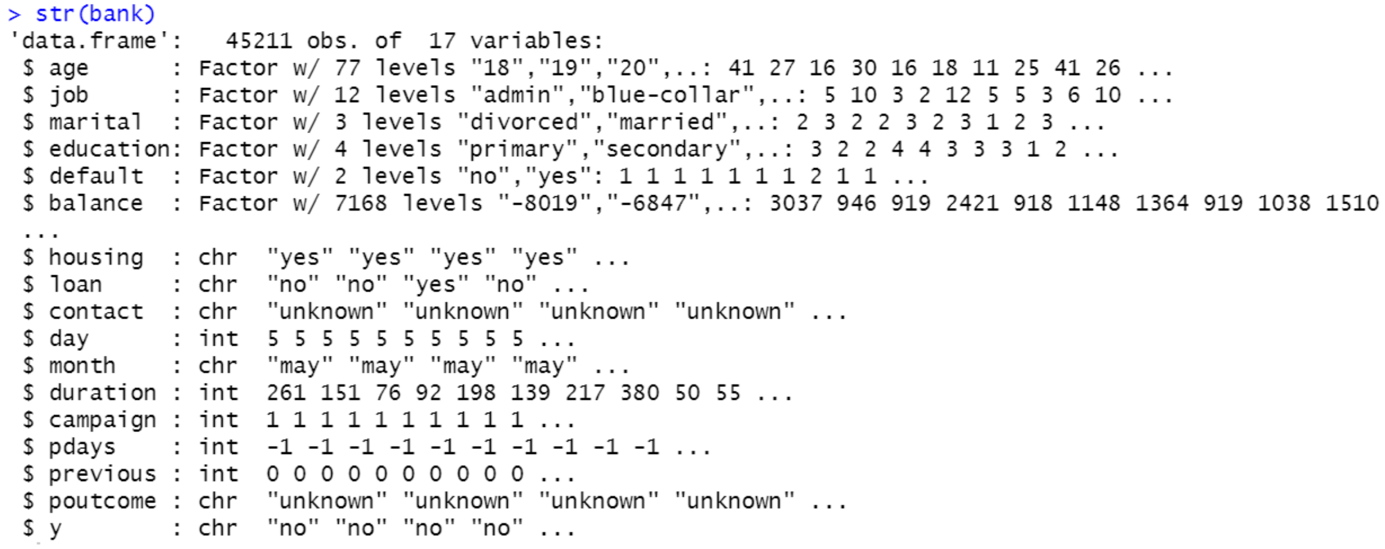
ALY 6010: Probability Theory and Introductory Statistics

Final Project

Rishabh Bansal

Introduction

As a final project, we have been working on this dataset which was assigned to us in Assignment 1 which is a banking dataset that contains 45211 observations and 17 columns:



```
> str(bank)
'data.frame':   45211 obs. of  17 variables:
 $ age      : Factor w/ 77 levels "18","19","20",..: 41 27 16 30 16 18 11 25 41 26 ...
 $ job      : Factor w/ 12 levels "admin","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ...
 $ marital  : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ...
 $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ...
 $ default  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...
 $ balance  : Factor w/ 7168 levels "-8019","-6847",..: 3037 946 919 2421 918 1148 1364 919 1038 1510
 ...
 $ housing  : chr  "yes" "yes" "yes" "yes" ...
 $ loan     : chr  "no" "no" "yes" "no" ...
 $ contact  : chr  "unknown" "unknown" "unknown" "unknown" ...
 $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
 $ month    : chr  "may" "may" "may" "may" ...
 $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
 $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome : chr  "unknown" "unknown" "unknown" "unknown" ...
 $ y        : chr  "no" "no" "no" "no" ...
```

I want to discover the answers to the following questions as part of this project:

• I'll look for a connection between those with education and balance.

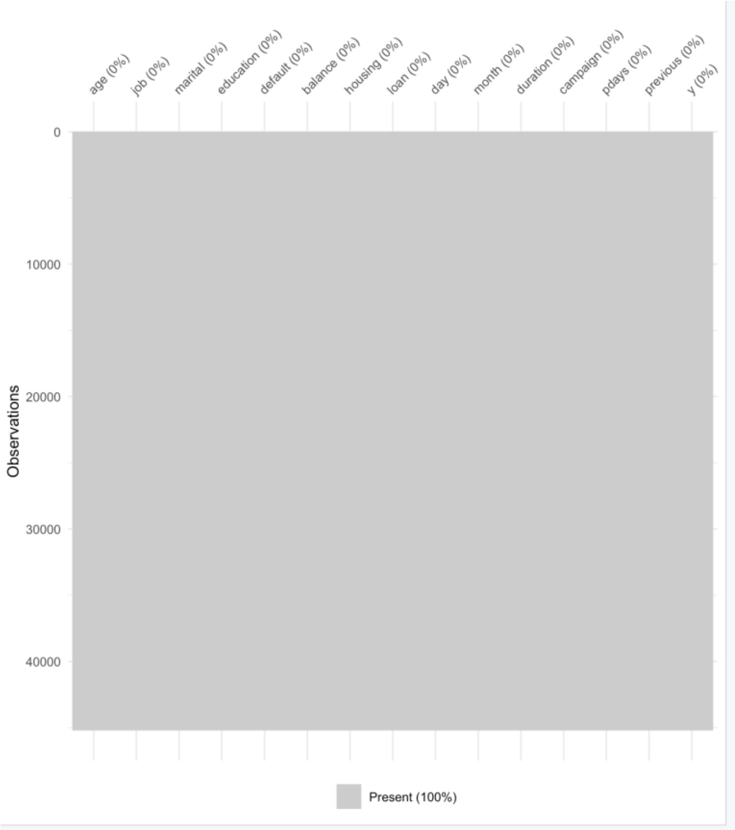
• I'll attempt to determine which profession has taken out the greatest loan by analyzing the association between profession and debt.

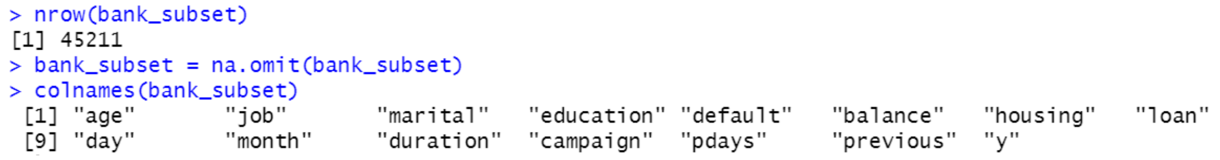
• I will attempt to produce an observation for the outliers in the balance

• I will attempt to observe if the average balance of married people with a loan of primary education is equivalent to secondary education.

• What form of education has the greatest balance in the account?

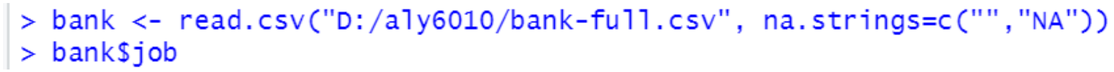
There are no null values in the data after cleaning, however, two columns ("outcome" and "contact") have unknown values that are eliminated.



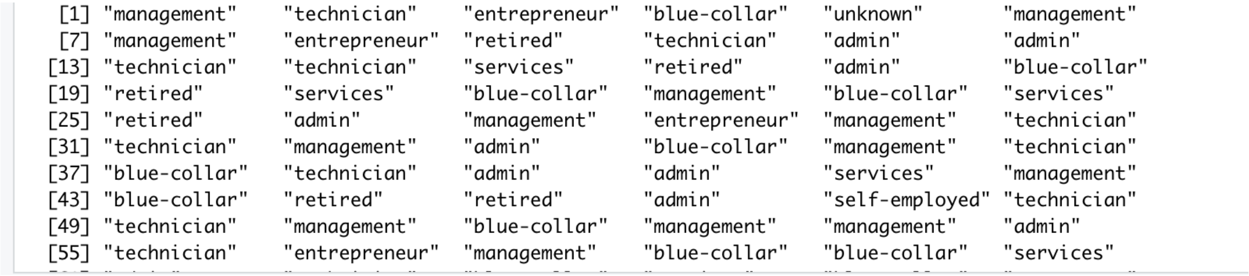


```
> nrow(bank_subset)
[1] 45211
> bank_subset = na.omit(bank_subset)
> colnames(bank_subset)
 [1] "age"       "job"       "marital"   "education" "default"   "balance"   "housing"   "loan"
 [9] "day"       "month"     "duration"  "campaign"  "pdays"     "previous"  "y"
```

The job type has the value “admin.” Which is not considered to be best practice, so replacing it with admin without dot at the end.

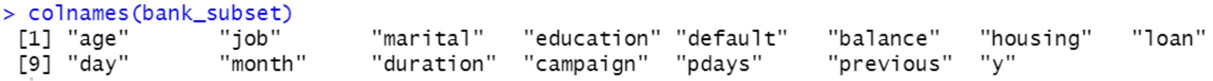


```
> bank <- read.csv("D:/aly6010/bank-full.csv", na.strings=c("","NA"))
> bank$job
```



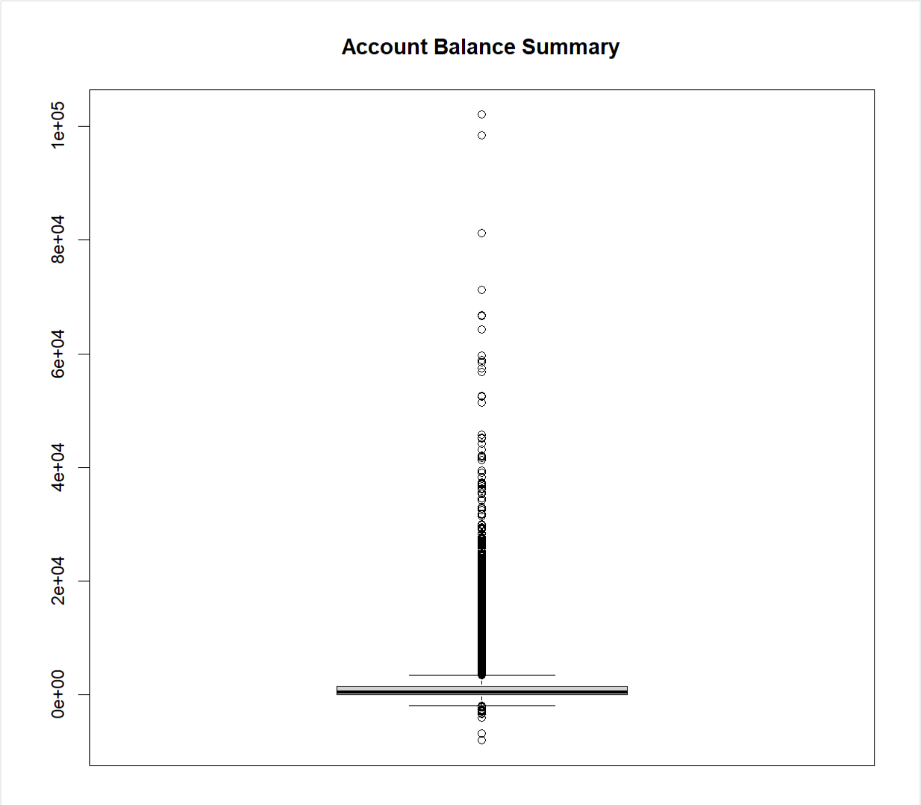
```
   [1] "management"    "technician"    "entrepreneur"  "blue-collar"   "unknown"       "management"
   [7] "management"    "entrepreneur"  "retired"       "technician"    "admin"         "admin"
  [13] "technician"    "technician"    "services"      "retired"       "admin"         "blue-collar"
  [19] "retired"       "services"      "blue-collar"   "management"    "blue-collar"   "services"
  [25] "retired"       "admin"         "management"    "entrepreneur"  "management"    "technician"
  [31] "technician"    "management"    "admin"         "blue-collar"   "management"    "technician"
  [37] "blue-collar"   "technician"    "admin"         "admin"         "services"      "management"
  [43] "blue-collar"   "retired"       "retired"       "admin"         "self-employed" "technician"
  [49] "technician"    "management"    "blue-collar"   "management"    "management"    "admin"
  [55] "technician"    "entrepreneur"  "management"    "blue-collar"   "blue-collar"   "services"
```

Following are the columns present in the dataset on which analysis will be done further.



```
> colnames(bank_subset)
 [1] "age"       "job"       "marital"   "education" "default"   "balance"   "housing"   "loan"
 [9] "day"       "month"     "duration"  "campaign"  "pdays"     "previous"  "y"
```

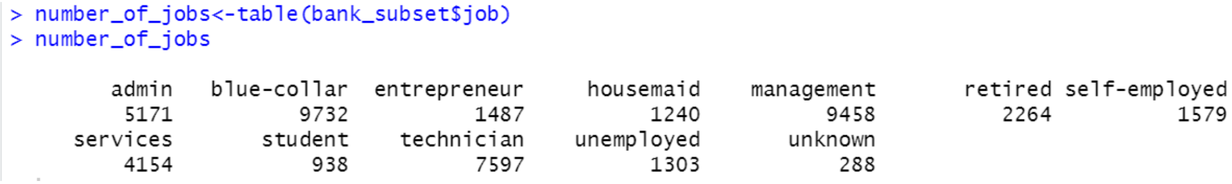
Loan Amount box plot:



The above box plot describes the balance columns in a dataset. As per the box plot, the maximum balance is 102127 and minimum – is 8019 and the median balance is 448.

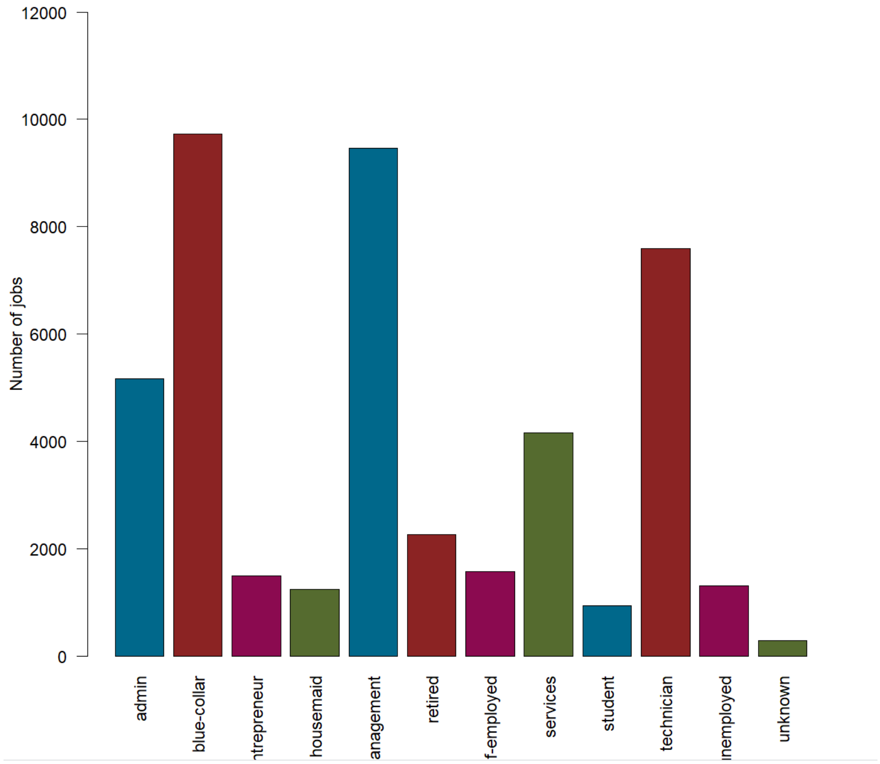
Frequency of jobs in each profile.

Below table shows the number of people in each job profile.



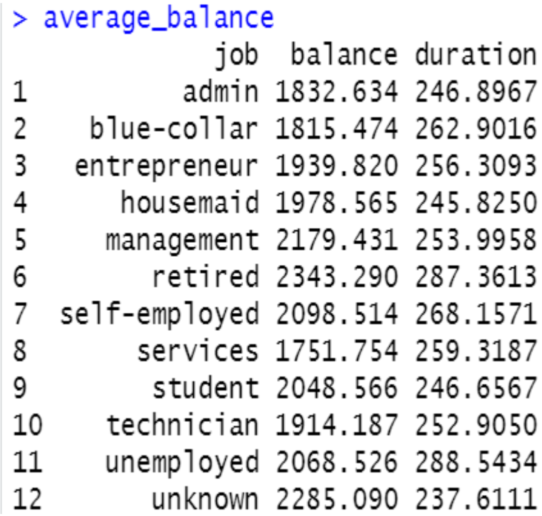
```
> number_of_jobs<-table(bank_subset$job)
> number_of_jobs

        admin   blue-collar  entrepreneur     housemaid    management       retired self-employed 
         5171          9732          1487          1240          9458          2264          1579 
     services       student    technician    unemployed       unknown 
         4154           938          7597          1303           288 
```



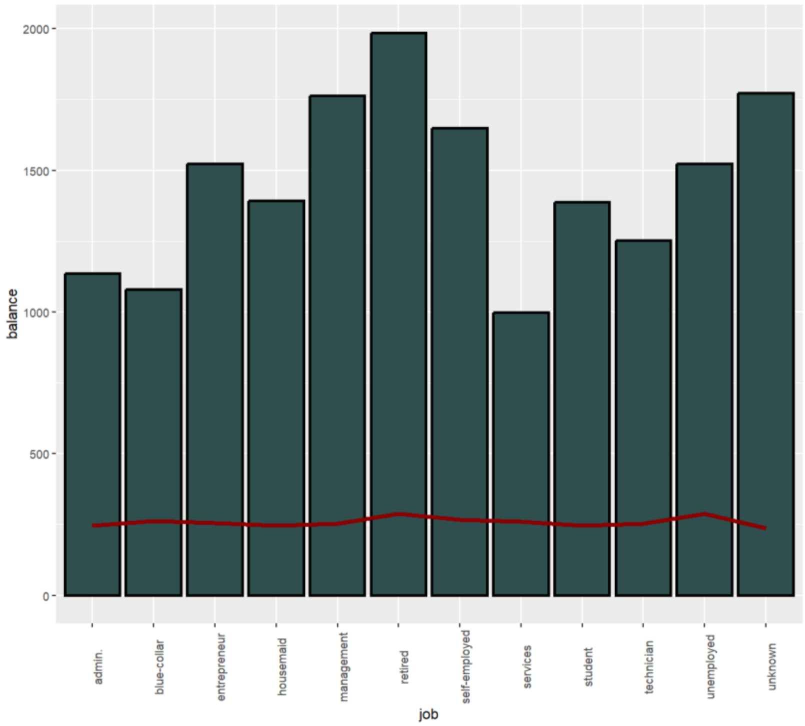
Here we can observe blue-collar, and management has the highest number of people.

Balance and duration based on the job:

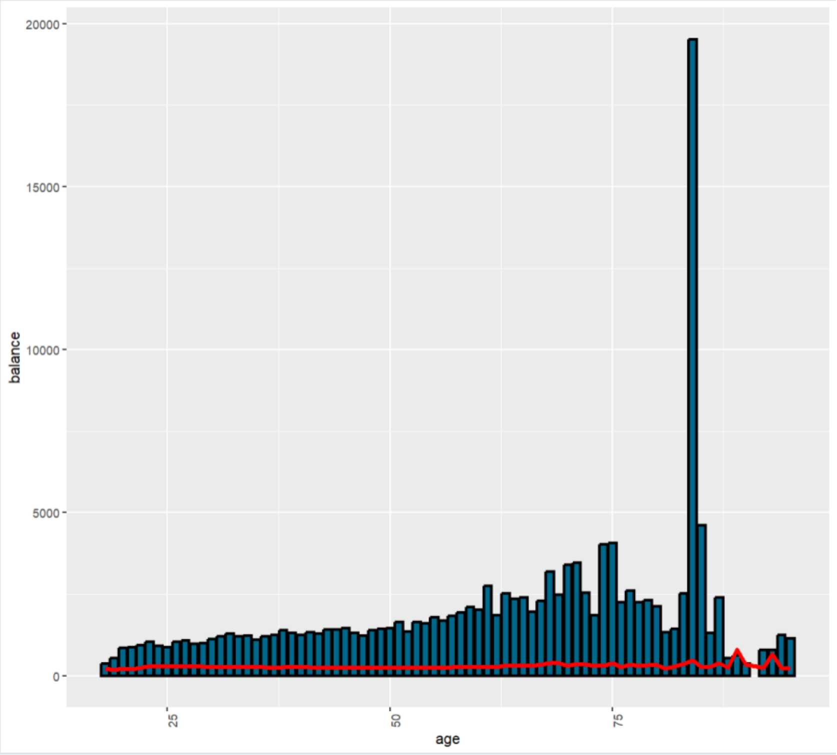


```
> average_balance
            job  balance duration
1         admin 1832.634 246.8967
2   blue-collar 1815.474 262.9016
3  entrepreneur 1939.820 256.3093
4     housemaid 1978.565 245.8250
5    management 2179.431 253.9958
6       retired 2343.290 287.3613
7 self-employed 2098.514 268.1571
8      services 1751.754 259.3187
9       student 2048.566 246.6567
10   technician 1914.187 252.9050
11   unemployed 2068.526 288.5434
12      unknown 2285.090 237.6111
```

In the above tables, I have calculated the average values of balance & duration according to education. Below is the representation of the same in the bar plot. Here the red line represents the duration of load and average balance of each job profile.



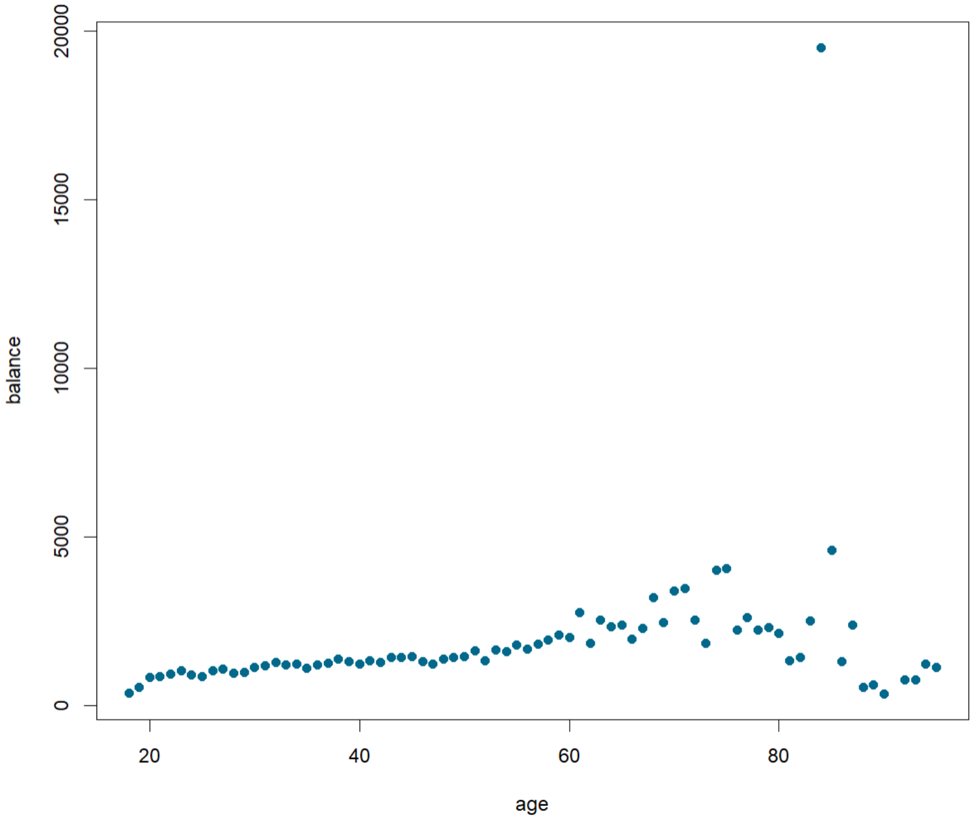
Aggregating the points and prices based on the reviewer:



The aggregation of loan balance and term regarding age is shown in the graph above. The age range here is from 18 to 95 years. As you can see, after 75 years, the length changes. The balance steadily rises from 18 to 80 years old according to trends.

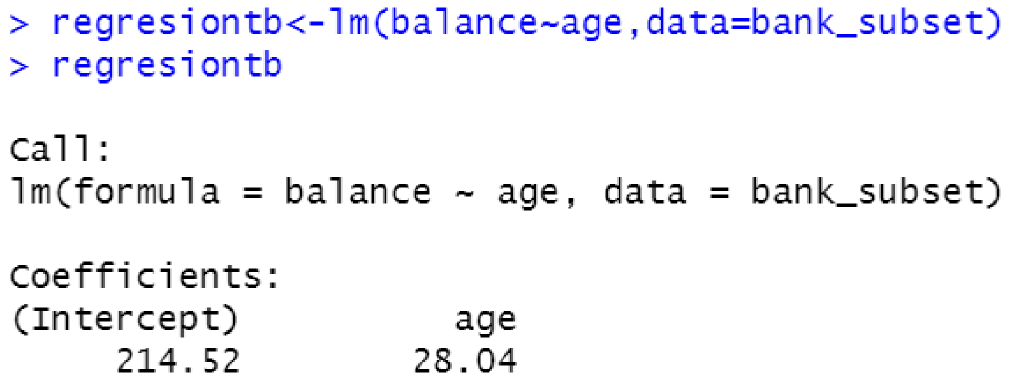
Aggregating the balance and duration of the basis of age:

As shown in the accompanying scatter plot (below) between the average balance and duration points that are dependent on the province, I have compiled and estimated the means of balance and durations for different ages.



Age and balance are directly inversely correlated, therefore we may infer that as age increases, so does balance. Therefore, age is a factor in this equilibrium. The same scattered graph is the subject of this regression model.

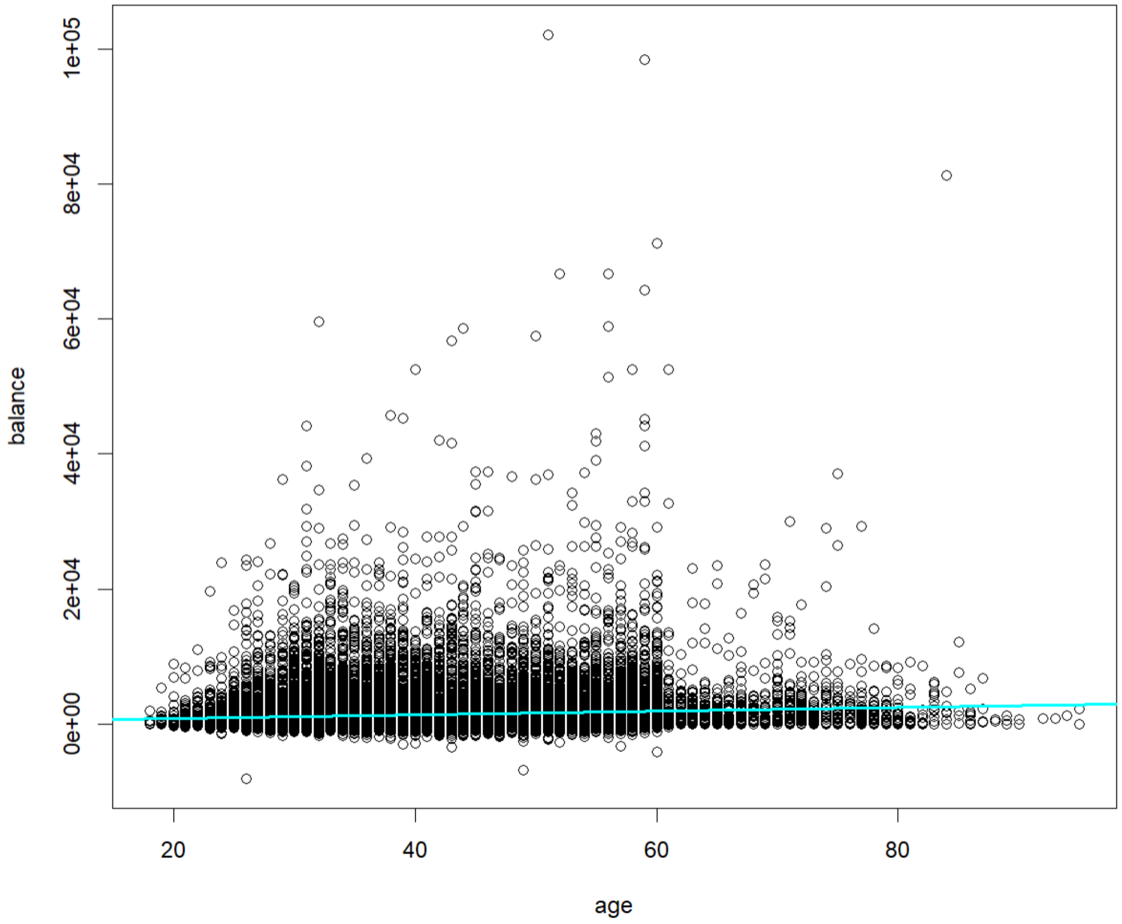
Regression for balance and age:



```
> regresiontb<-lm(balance~age,data=bank_subset)
> regresiontb

Call:
lm(formula = balance ~ age, data = bank_subset)

Coefficients:
(Intercept)          age
     214.52        28.04
```



Age is the independent variable in the graph above, whereas balance is the dependent variable. We are aware that age affects equilibrium. According to the equation used y = mx+c, we can calculate the y if we know the x value in short if we know the age, we can predict the balance.

M=28.04

y-coefficient=-214.52

x= value of point

y= dependent variable which is the price here.

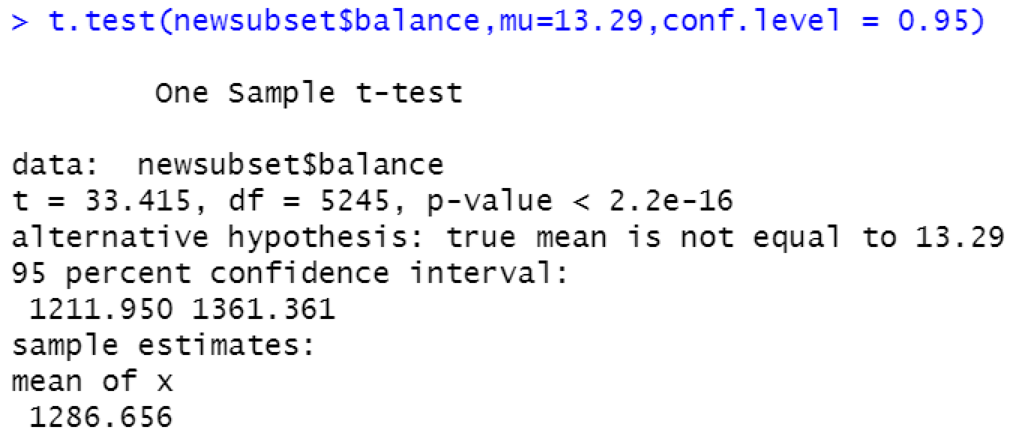
Hypothesis Testing on California and Oregon:

1 sample T-Test:

According to the observation, a married person with a primary education will have 1000. So to prove the null hypothesis I am conducting t-test to with one sample test to observe if that data has the mean value is equal to 1000 or not

H0: The average balance of a married person with primary education is 1000$.

H1: The average balance of a married person with primary education is not 1000$.



```
> t.test(newsubset$balance,mu=13.29,conf.level = 0.95)

        One Sample t-test

data:  newsubset$balance
t = 33.415, df = 5245, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 13.29
95 percent confidence interval:
 1211.950 1361.361
sample estimates:
mean of x
 1286.656
```

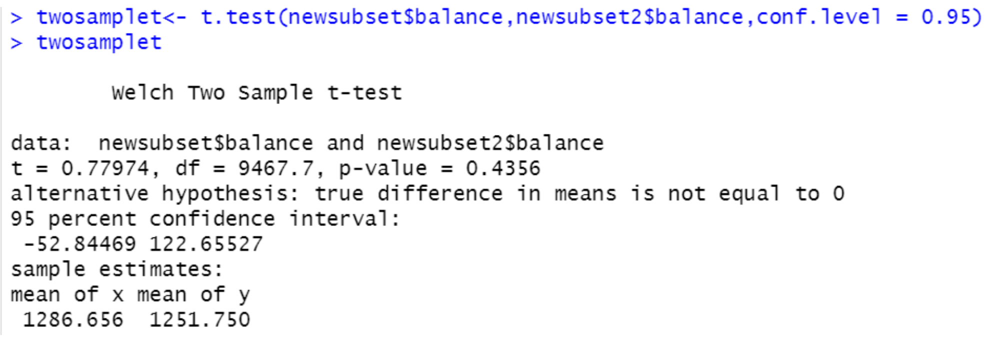
As per the test, we have a p-value of less than 2.2e-16 which is next to zero with a confidence level of 95%. We can reject the null hypothesis which is less than the significant value.

2 sample T-Test:

Now doing a two-sample test to determine whether or not the average proportion of married people with elementary and secondary schooling is equal. For this exam, my level of confidence is 95%. I am using the following null hypothesis and alternative hypothesis to do a two-sample t-test.

H0: A married guy with primary and secondary education will have the same balance

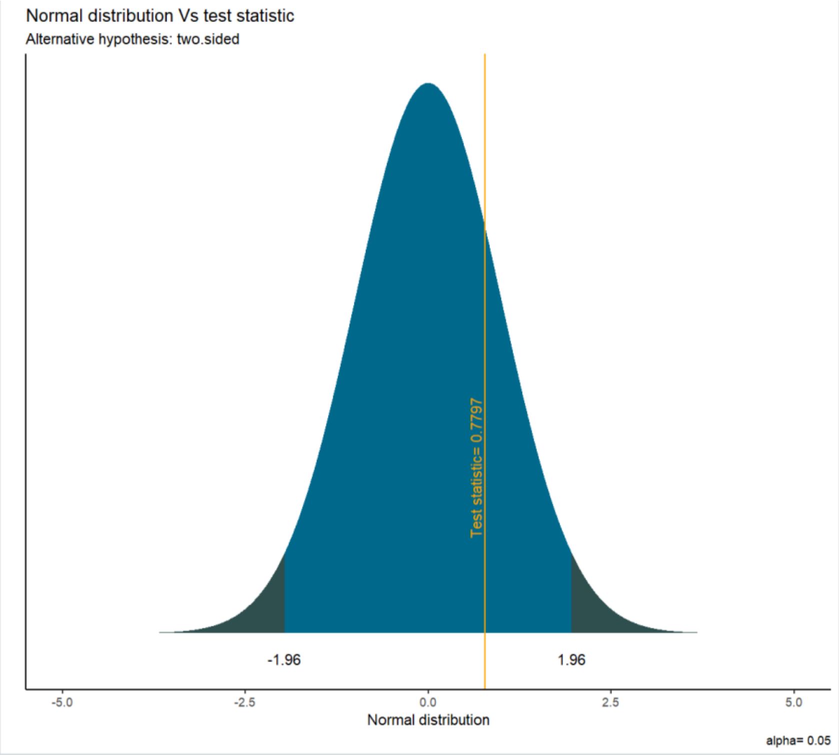
H1: A married guy with primary and secondary education will not have the same balance



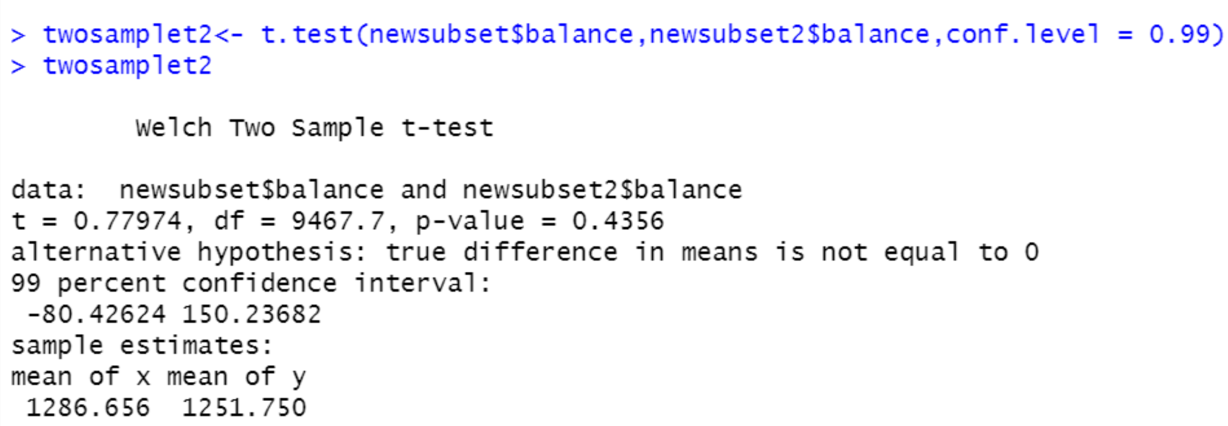
```
> twosamplet<- t.test(newsubset$balance,newsubset2$balance,conf.level = 0.95)
> twosamplet

        Welch Two Sample t-test

data:  newsubset$balance and newsubset2$balance
t = 0.77974, df = 9467.7, p-value = 0.4356
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -52.84469 122.65527
sample estimates:
mean of x mean of y
 1286.656  1251.750
```



Now, If we consider 99% confidence:



```
> twosamplet2<- t.test(newsubset$balance,newsubset2$balance,conf.level = 0.99)
> twosamplet2

        Welch Two Sample t-test

data:  newsubset$balance and newsubset2$balance
t = 0.77974, df = 9467.7, p-value = 0.4356
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -80.42624 150.23682
sample estimates:
mean of x mean of y 
 1286.656  1251.750 
```

After analyzing with a 99% confidence level, the p-value is also not sufficient enough to reject the null hypothesis. In this process, the width of the confidence interval has increased as compared to the 95% interval.

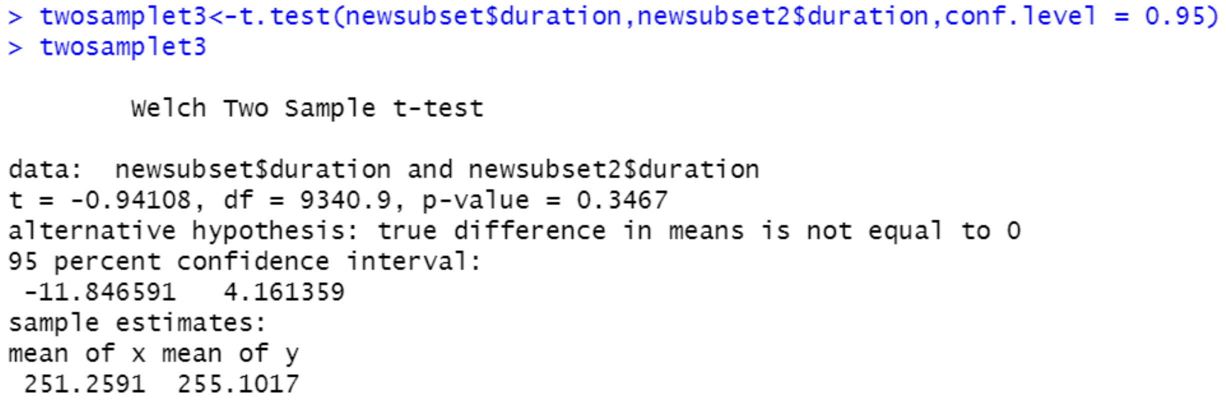
The 95 % confidence interval is -52.84469 to 122.65527.

The 99 % confidence interval is -80.42624 to 150.23682.

Second sample T-Test:

H0: The average duration of a married person with primary education is the same as that secondary

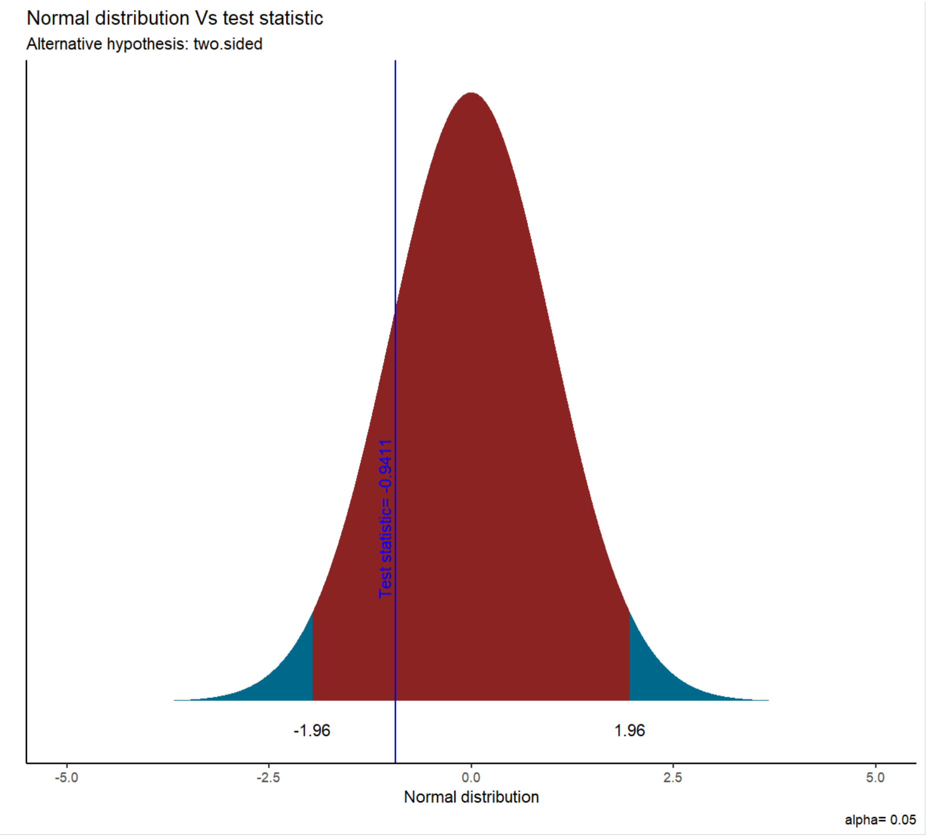
H1: The average duration of a married person with primary education is not the same as that secondary



```
> twosamplet3<-t.test(newsubset$duration,newsubset2$duration,conf.level = 0.95)
> twosamplet3

        Welch Two Sample t-test

data:  newsubset$duration and newsubset2$duration
t = -0.94108, df = 9340.9, p-value = 0.3467
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.846591   4.161359
sample estimates:
mean of x mean of y 
 251.2591  255.1017 
```



In the above test, we can observe the p-value is 0.3467 for a 95% confidence level, we will have to accept the null hypothesis if the p-value is greater than 0.05. Hence, we fail to reject the null hypothesis. Hence both means are equal.

Conclusion:

According to the study, there is a stronger positive link between balance and duration when the data are grouped by age. Following the regression model, we may calculate an individual's account balance depending on his age, which is connected to a certain sample.

References:

GGTTEST: Student's t-test plot. RDocumentation. (n.d.). Retrieved December 16, 2022, from https://www.rdocumentation.org/packages/gginference/versions/0.1.3/topics/ggttest

Holtz, Y. (n.d.). Dealing with color in GGPLOT2: The R Graph Gallery. Dealing with color in ggplot2 | the R Graph Gallery. Retrieved December 4, 2022, from https://r-graph-gallery.com/ggplot2-color.html

R. K.-. (n.d.). Subsetting Data. Quick-R: Subsetting Data. Retrieved December 4, 2022, from https://www.statmethods.net/management/subset.html