2022

FIN1002 – FINANCIAL STATISTICS

MAJOR ASSIGNMENT – BANK CUSTOMER ANALYSIS

KHAI NGUYEN – 103844625

Contents

Executive Summary

With an aim to providing insights into the bank's customer profiles, a random sample of 80 customers was chosen and data analysis was undertaken across the 7 categories: Credit score, Geography, Gender, Age, Tenure, Account balance, Number of Products and Estimated income.

Critical statistical parameters were computed and their data was displayed using graphs and tables.

A dashboard was generated using Excel to provide a visual illustration regarding different information of the customers: Average Salary by Number of Products, Average Balance of Gender vs Age, Average Credit Score by Gender and Estimated income by Country. The dashboard is interactive with slicers showcasing different filters to enable the user to explore the relationships between the different variables.

Statements of 95% confidence were made regarding the average balance for female customers and the average age for customers from Spain. The results were consistent with the true mean of the population.

Claims regarding French customers being more loyal than German ones and that there is a difference in the estimated salary between male and female customers were also tested. These claims, however, were both debunked.

Lastly, the relationship between customers' credit score and their estimated salary was investigated. Results showed that there is a lack of correlation between these 2 variables and the linear model was a poor attempt to fit the data points.

Introduction

This report will identify the key findings extracted from the analysis of customer details of a banking system. The population consists of 10,000 customers and their details span across 9 variables as follows:

1. Customer Id
2. Credit Score
3. Geography
4. Gender
5. Age in years
6. Tenure
7. Balance
8. Number of Products
9. Estimated income

In order to generate insights from the given data set, a randomized sample of 80 customers is selected (refer to Appendix 1), on which the data analysis is performed. A range of statistical techniques are employed using Microsoft Excel and Python, which include descriptive statistics, confidence intervals, testing of hypothesis and the determination of correlation between variables and their regression models. Additionally, tables and graphs of each variable are produced to provide a visual illustration for the data and a dashboard is created to facilitate the identification of trends and relationships between different variables.

Analysis

1. Descriptive statistics

In order to perform descriptive statistics analysis, Microsoft Excel was utilized to produce key statistical parameters.

1.1 Credit Score

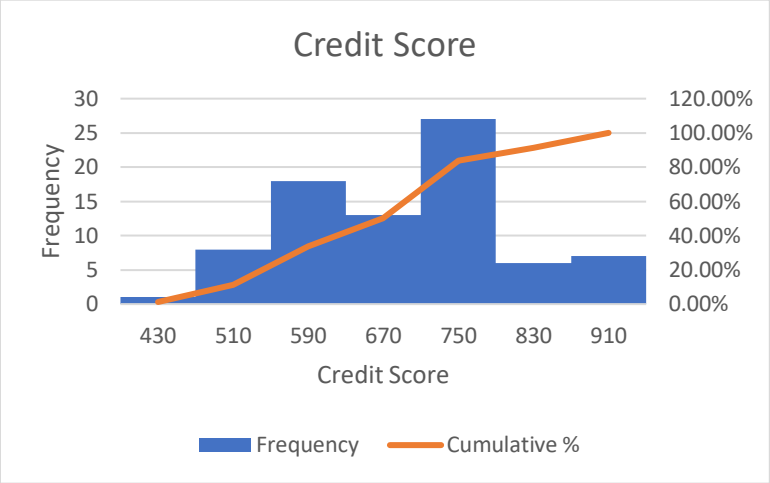


Figure 1 displays the histogram of customers' credit score

Credit Score	
Mean	649.5625
Median	667.5
Standard Deviation	110.50126
Range	500
Minimum	350
Maximum	850
Q1	573.75
Q3	720.75

Table 1 shows the key parameters for the customers' credit score

It can be seen from Table 1 that the mean and median credit score of the sample is approximately 650 and 668 respectively. The standard deviation is calculated to be 110.5, meaning that the data is quite spread out and this can be reflected by a considerably large range of 500 (with max credit score = 850 and min credit score = 350)

1.2 Geography

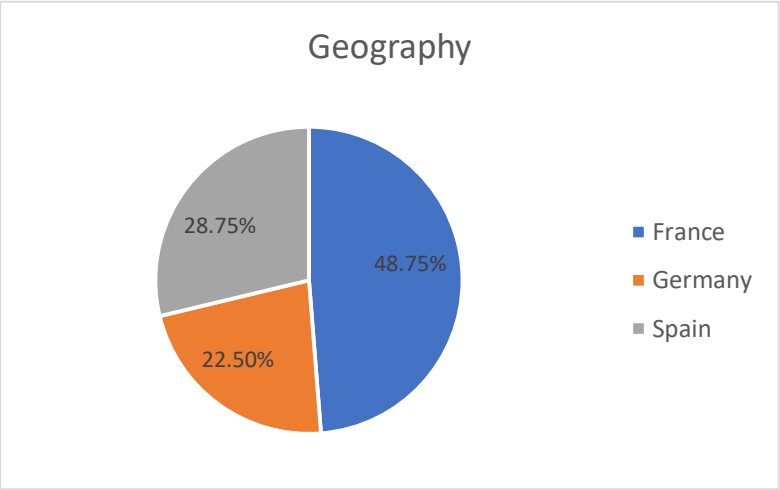


Figure 2 shows the percentage of customers from different countries

Table 2

Geography	Percentage	Count of Geography
France	48.75%	39
Germany	22.50%	18
Spain	28.75%	23

It can be observed from Figure 2 that the percentages of customers from France, Germany and Spain are 48.75%, 22.50% and 28.75% respectively, with France accounting for the majority.

1.3 Gender

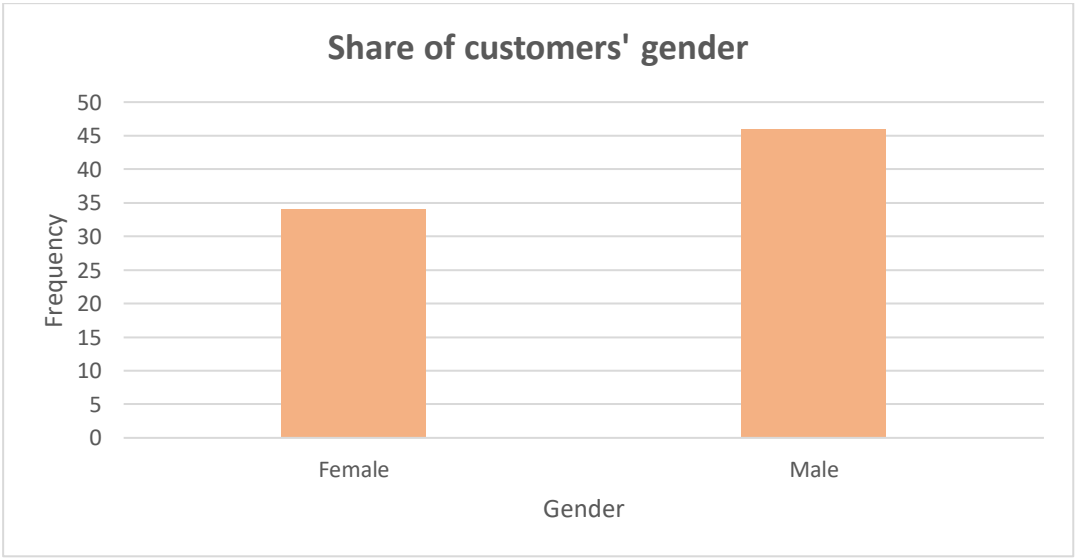


Figure 3 displays the number of males and females amongst the customer sample

Table 3

Gender	Count of Gender
Female	34
Male	46

From Figure 3 and Table 3, it can be concluded that there are more males than females in the gender pool, with 46 customers being males in comparison to 34 females.

1.4 Age

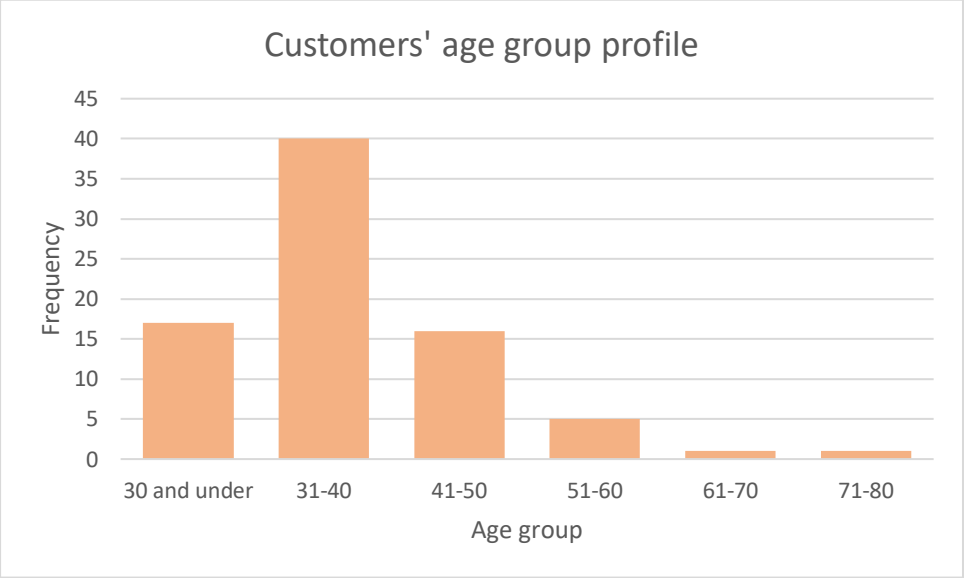


Figure 4 displays the age group profile of the customer sample

Table 4

Bin	*Frequency*
30 and under	17
31-40	40
41-50	16
51-60	5
61-70	1
71-80	1

It can be observed from Figure 4 and Table 4 that the majority of the customers fall into the 31-40 age group, accounting for 50% of the sample's profile. The least common age groups are 61-70 and 71-80, with only 1 customer in each age group, accounting each for 1% of the sample.

1.5 Tenure

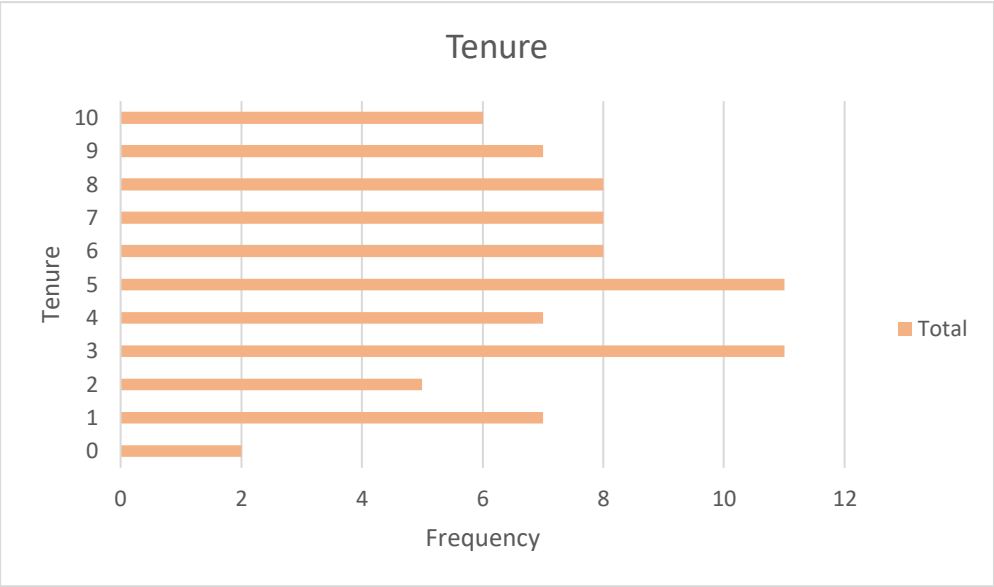


Figure 5 displays customers' tenure and their frequency

Table 5 shows customers' tenure and their frequency

Tenure	Count of Tenure
0	2
1	7
2	5
3	11
4	7
5	11
6	8
7	8
8	8
9	7
10	6
Grand Total	**80**

It can be observed from Figure 5 and Table 5 that the most common tenure is 3 years and 5 years. The least common tenure is 0 year with only 2 customers. It could also be observed from Figure 5 that the distribution for customers' tenure is approximately normal as it models a bell-shaped curve.

1.6 Account balance

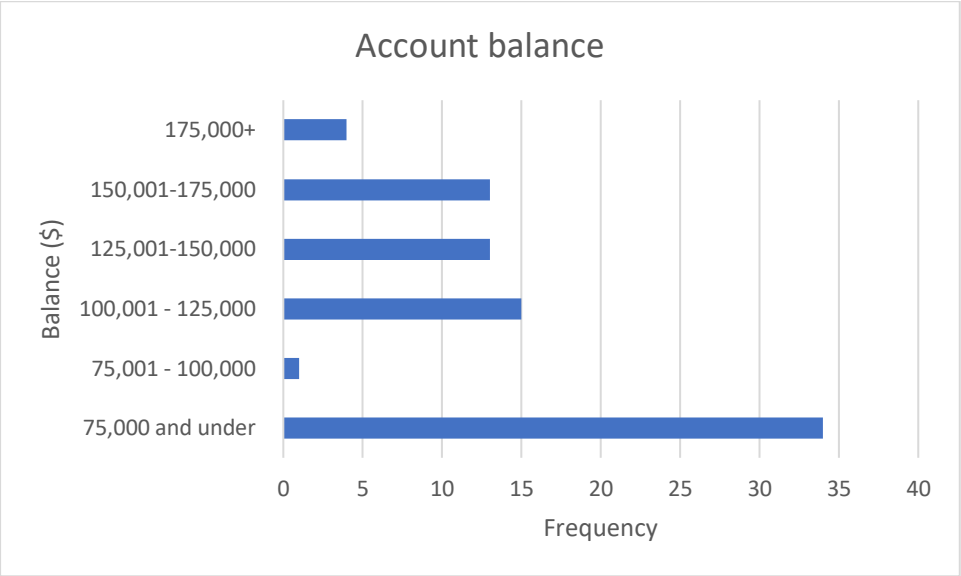


Figure 6 displays the customers' account balance

Table 6 shows the key parameters for the customers' account balance

Balance	
Mean	65554.66
Median	89013.73
Range	169025.8
Standard Deviation	59779.78457
Minimum	0
Maximum	169025.8
Q1	0
Q3	114641.5

It can be seen from Table 6 that the mean and median account balance of the customer are $65,555 and $89,013 respectively. Such a difference between the two central values can be explained by the considerably large standard deviation of $59,779, meaning that the customers' balances vary quite significantly.

1.7 Number of products

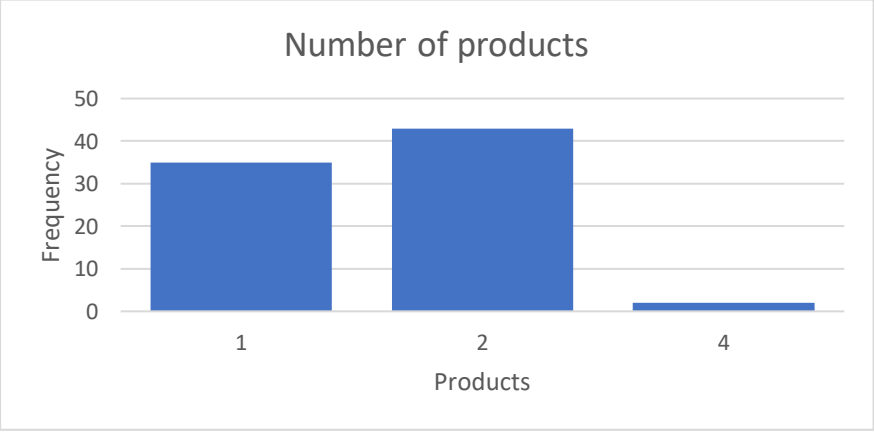


Figure 7 displays the number of products of the customers

Table 7

Number of products	Frequency
1	35
2	43
4	2

Most customers use 2 products from the bank, with very little using 4 products. There is no customer using 3 products from the bank from the sample. This suggests that the sample might have not been representative enough to include those who use 3 products from the population.

1.8 Estimated Salary

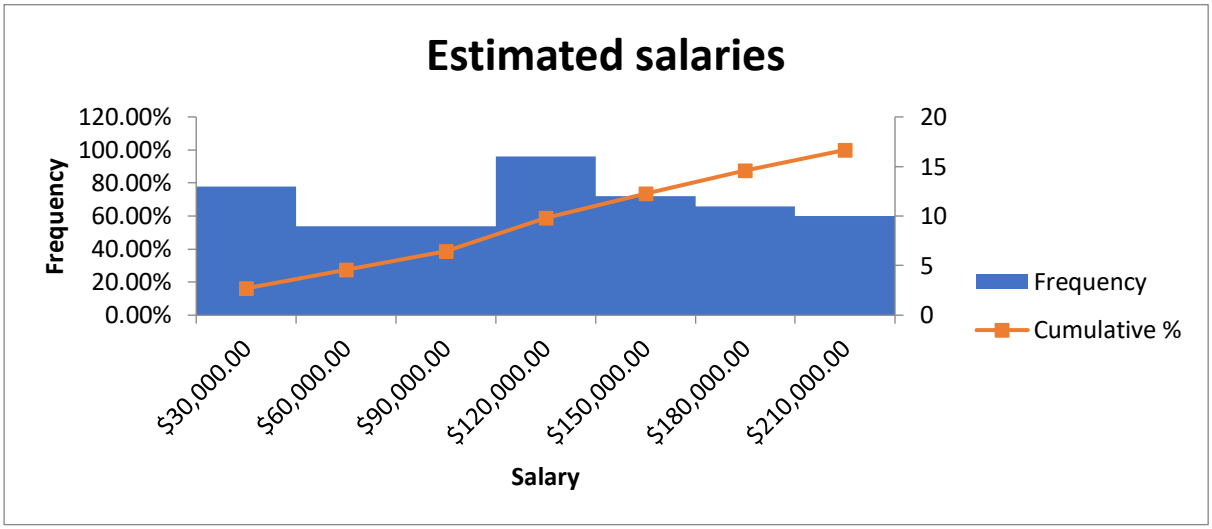


Figure 8 displays the estimated salaries of customers

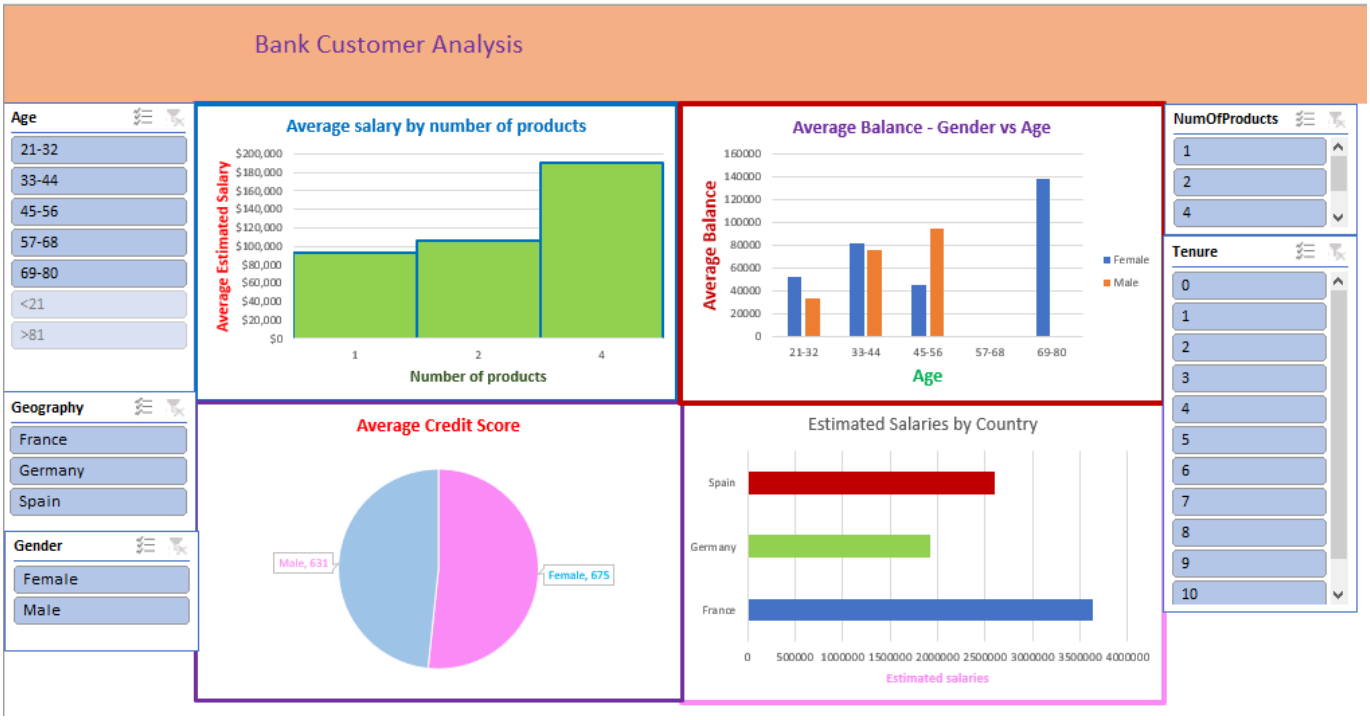
Table 8

Estimated Salary	
Mean	101779
Median	105858.5
Range	196048.5
Minimum	468.94
Maximum	196517.4
Q1	48859.99
Q3	151733.5

The customers' mean estimated salary is determined to be $101,779, ranging from $469.84 up to $196,517. As there are outliers, the median estimated salary of $105,858 would be a better parameter for the central value of customers' estimated salary.

2. Dashboard

To enable the stakeholders to analyse the relationships between different data, a dashboard was created using Excel. The slicers on the sides can be utilized to filter the data for the required parameter, making the dashboard interactive. Find attached Excel file to use the dashboard.



3. Confidence intervals

a. The average balance for females

In order to estimate the average balance for female customers, a list of female customers from the sample were filtered out. Using this sample, data analysis was undertaken. Results showed that we are 95% confident that the population mean of the balance for female customers lie between $44,790 and $87,815. True population mean was determined to be $66,307.11, suggesting that the sample represented the population well.

b. The average age for customers from Spain

The same process was applied to filter out only the customers who are from Spain. Results showed that we are 95% confident that the true population mean of the age for Spanish customers lie between 33 and 39 (years old). As the actual population mean was calculated to be 35.9 (years old), this suggests that the sample was a good representation of the whole population.

4. Hypothesis testing

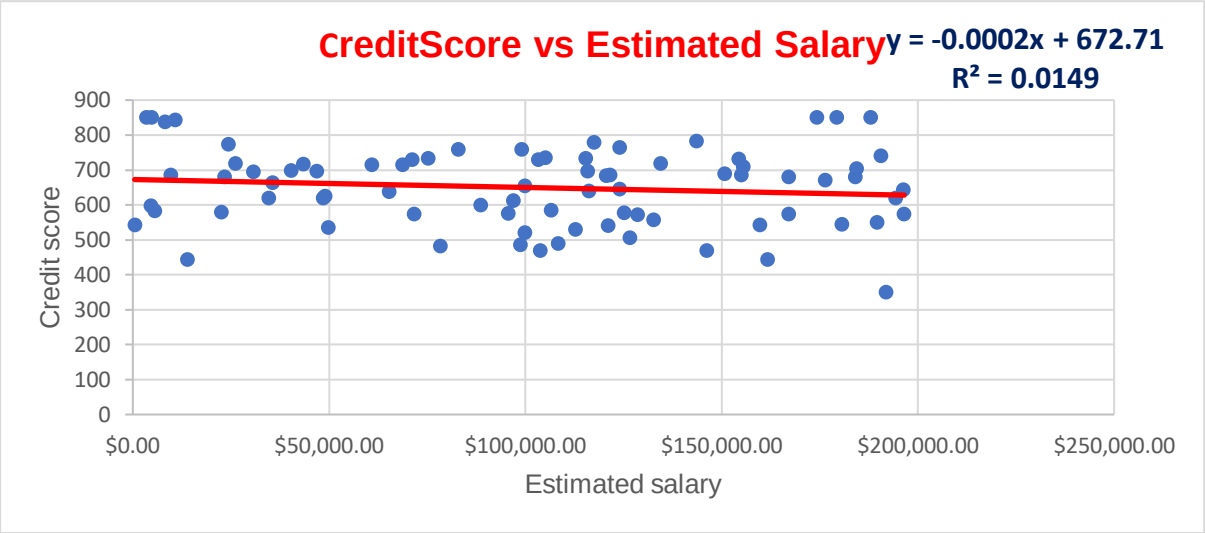
As it seems that French customers tend to be more loyal and stay with the bank for longer than their German counterparts, a test was carried out to see if there is enough evidence to support this claim. From the sample, it was deduced that there is not sufficient evidence to reject that the French

customers would stay less than the German customers. Hence, the claim that French customers would stay longer with the bank is not supported.

Additionally, claims have been made about the difference in the average estimated salary between the two genders and experiment was conducted to test the claim. However, results showed that there is not sufficient evidence to reject that the average salaries for males and females are similar. Therefore, it cannot be concluded that there is a difference in the average salary between males and females.

5. Correlation and regression

To investigate the relationship between customers' credit score and their estimated salary, a scatter plot of the two variables was graphed. As can be seen from Figure, the dots are randomly spread, suggesting little correlation between the 2 variables.



A linear model was plotted in an attempt to model the relationship between the 2 variables and was determined to have the equation **y = -0.0002x + 672.71.** This means that with an estimated salary of $0, a customer is expected to have a credit score of 672.21 and as their estimated salary increases, their credit score will experience a slight decrease of -0.0002 times their estimated salary.

To validate these observations, regression analysis was used in Excel to evaluate the correlation between the 2 variables and the goodness of fit of the linear model. The multiple R, which is the correlation coefficient, was calculated to be -0.122. This means that there is a slight negative correlation between credit score and estimated salary. However, as this value is very close to 0, it indicates that there is hardly any linear relationship between the two variables. Additionally, the R^2, which is the coefficient of determination, was calculated to be 0.0149. This means that the linear model only explains 1.49% of the data points, indicating a poor fit to the data.

In order to further verify these conclusions, a hypothesis test was conducted. From the results obtained, it can be concluded that there was not sufficient evidence to reject the hypothesis that there is no linear relationship between customers' credit score and their estimated salary. This further validates the interpretation of the data.

Conclusion

In conclusion, several insights were extracted from the analysis of different data using a sample of 80 customers. The customers' mean credit score is 650, ranging from 350 to 850. The majority of the bank customers are from France, accounting for almost half of the customers' the geography profile (48.75%). Most customers fall into the age group of 31-40, whereas those in the 61-70 and 71-80 age groups account for only 2% of the sample. The length of stay of customers is approximately normally distributed, with most customers having stayed with the bank for 3 or 5 years. There is also a large variance in the customers' account balance, ranging from $0 up to $169,025. Most customers use 2 products from the bank, with very few using 4. Additionally, the customers' mean estimated salary is determined to be $101,779, ranging from $469 to $196,517.

Additionally, the relationships between different variables were also discovered through using the dashboard.

The analysis of these data, however, is still limited to a certain extent. A sample size of 80 might not be large enough to reflect the population size of 10,000. Sample variances of certain variables remain quite large, indicating that

Appendices

Appendix 1 – Sample

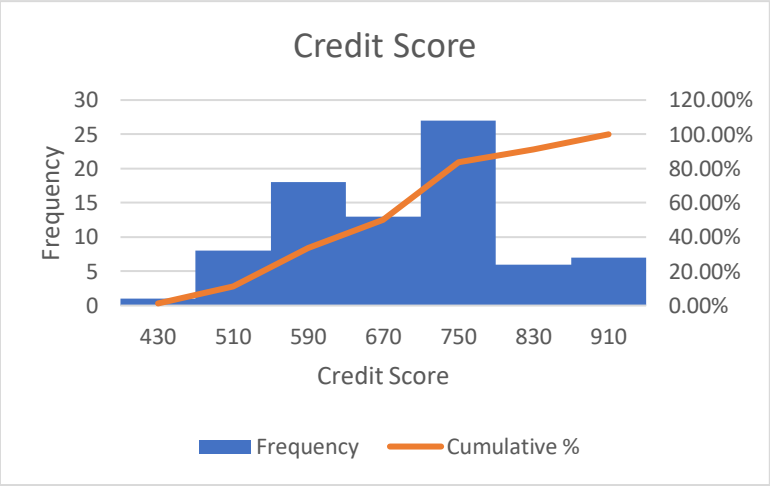
CustomerId	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	EstimatedSalary
15784124	645	Germany	Male	41	2	$93,925.30	1	$123,982.14
15616172	838	France	Male	31	2	$0.00	2	$8,222.96
15571193	579	Germany	Male	42	0	$144,386.32	1	$22,497.10
15763029	612	Germany	Male	46	9	$161,450.03	1	$96,961.00
15571843	486	Spain	Male	24	1	$0.00	1	$98,802.76
15613292	715	France	Male	21	8	$0.00	2	$68,666.63
15617648	584	France	Female	44	5	$95,671.75	2	$106,564.88
15725818	583	Germany	Male	40	4	$107,041.30	1	$5,635.63
15607598	575	Spain	Female	31	6	$0.00	2	$95,686.42
15751137	850	Germany	Female	36	3	$169,025.83	1	$174,235.06
15575694	729	Spain	Female	45	7	$91,091.06	2	$71,133.12
15732293	759	Spain	Male	31	8	$0.00	2	$99,086.74
15779529	620	France	Male	32	7	$0.00	2	$34,665.79
15599182	597	Spain	Female	33	2	$0.00	2	$4,700.66
15579787	686	France	Male	39	4	$0.00	2	$155,023.93
15678720	741	France	Female	44	7	$0.00	2	$190,534.76
15649379	850	France	Female	46	3	$0.00	2	$187,980.21
15633260	600	France	Male	37	1	$142,663.46	1	$88,669.89
15622494	718	France	Male	27	2	$0.00	2	$26,229.24
15635388	640	Spain	Male	47	6	$89,047.14	1	$116,286.25
15685372	350	Spain	Male	54	1	$152,677.48	1	$191,973.49
15743456	715	France	Female	32	10	$0.00	2	$60,907.49
15792064	545	Germany	Male	53	5	$114,421.55	1	$180,598.28
15718406	540	France	Male	41	3	$0.00	2	$121,098.65
15611947	557	France	Male	34	3	$83,074.00	1	$132,673.22
15749328	697	France	Female	45	1	$0.00	2	$46,807.62

15717898	542	Spain	Male	32	2	$131,945.94	1	$159,737.56
15566292	574	Spain	Male	36	1	$0.00	2	$71,709.12
15761340	521	France	Male	22	5	$0.00	2	$99,828.45
15714789	664	France	Male	24	7	$0.00	1	$35,611.35
15705113	685	Spain	Male	34	6	$83,264.28	1	$9,663.28
15815316	644	France	Male	50	9	$76,817.00	4	$196,371.13
15789413	733	France	Male	64	3	$0.00	2	$75,272.63
15726310	782	Spain	Female	27	3	$0.00	2	$143,614.01
15654489	843	France	Female	38	8	$134,887.53	1	$10,804.04
15604588	850	Spain	Female	38	3	$0.00	2	$179,360.76
15793343	549	France	Female	29	8	$0.00	2	$189,558.44
15570629	655	Germany	Female	72	5	$138,089.97	2	$99,920.41
15607629	679	France	Male	48	8	$0.00	2	$23,344.94
15688612	850	France	Male	33	7	$140,956.99	1	$3,510.18
15684461	469	Spain	Female	31	6	$0.00	1	$146,213.75
15718465	671	Germany	Male	51	3	$96,891.46	1	$176,403.33
15665222	625	Spain	Male	52	8	$121,161.57	1	$48,988.28
15566708	444	France	Female	45	4	$0.00	2	$161,653.50
15650889	710	Germany	Female	30	10	$133,537.10	2	$155,593.74
15625716	637	France	Female	33	9	$113,913.53	1	$65,316.50
15790235	778	Spain	Male	40	8	$104,291.41	2	$117,507.11
15632521	689	Germany	Male	45	0	$130,170.82	2	$150,856.38
15630167	684	Spain	Female	39	4	$139,723.90	1	$120,612.11
15640855	729	Germany	Male	40	5	$113,574.61	2	$103,396.08
15767891	619	Germany	Female	28	6	$99,152.73	2	$48,475.12
15763063	685	Spain	Female	25	10	$128,509.63	1	$121,562.33
15584580	443	France	Male	35	6	$161,111.45	1	$13,946.66

15671148	490	Germany	Male	33	5	$96,341.00	2	$108,313.34
15753837	573	Spain	Male	38	4	$0.00	2	$196,517.43
15579548	735	Spain	Male	36	5	$0.00	2	$105,152.17
15781678	470	Spain	Male	31	4	$55,732.92	2	$103,792.53
15622442	619	France	Male	29	5	$0.00	2	$194,310.10
15655007	758	France	Female	33	7	$0.00	2	$82,996.47
15742358	696	Germany	Male	32	8	$101,160.99	1	$115,916.55
15606849	698	France	Female	27	1	$94,920.71	1	$40,339.90
15807923	716	Germany	Female	39	10	$115,301.31	1	$43,527.40
15619616	571	France	Female	33	9	$102,017.25	2	$128,600.49
15780804	482	France	Male	55	5	$97,318.25	1	$78,416.14
15683618	774	France	Female	35	3	$121,418.62	1	$24,400.37
15772632	680	France	Female	34	1	$0.00	2	$167,035.07
15631170	695	France	Male	45	3	$0.00	2	$30,793.61
15752488	733	Spain	Female	31	9	$102,289.85	1	$115,441.66
15689526	542	Germany	Female	35	9	$127,543.11	2	$468.94
15613463	679	Germany	Female	50	6	$132,598.38	2	$184,017.98
15590268	529	Spain	Male	35	5	$95,772.97	1	$112,781.50
15637366	505	Germany	Female	25	5	$114,268.85	2	$126,728.27
15615176	732	France	Male	26	7	$0.00	2	$154,364.66
15790594	535	France	Female	27	6	$0.00	2	$49,775.58
15603851	704	France	Male	32	7	$127,785.17	4	$184,464.70
15732943	574	Spain	Male	36	4	$77,967.50	1	$167,066.95
15657085	578	France	Male	23	10	$88,980.32	1	$125,222.36
15598331	764	France	Female	40	9	$100,480.53	1	$124,095.69
15640866	718	France	Female	29	3	$0.00	1	$134,462.29
15733247	850	France	Male	33	10	$0.00	1	$4,861.72

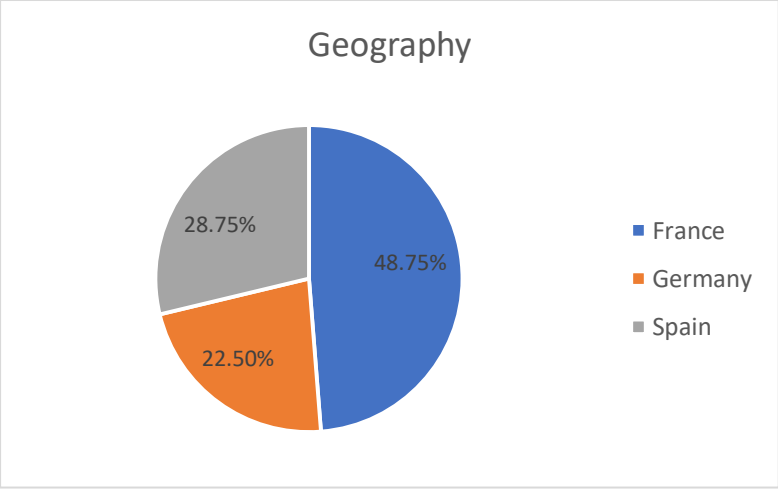
Appendix 2 – Descriptive statistics

Credit Score



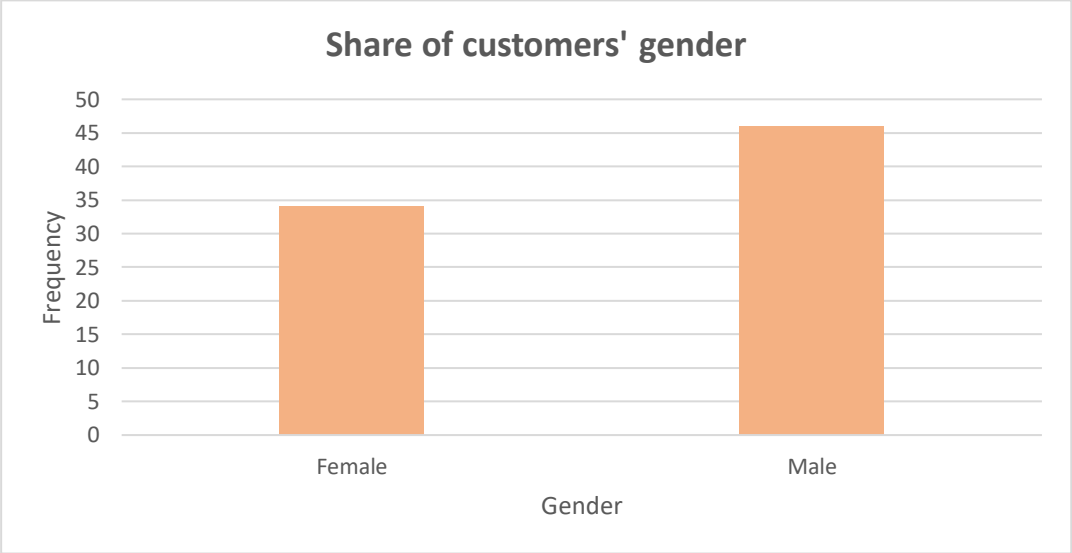
Credit Score	
Mean	649.5625
Median	667.5
Standard Deviation	110.50126
Range	500
Minimum	350
Maximum	850
Q1	573.75
Q3	720.75

Geography



Geography	Percentage	Count of Geography
France	48.75%	39
Germany	22.50%	18
Spain	28.75%	23

Gender



Gender	Count of Gender
Female	34
Male	46

Age group

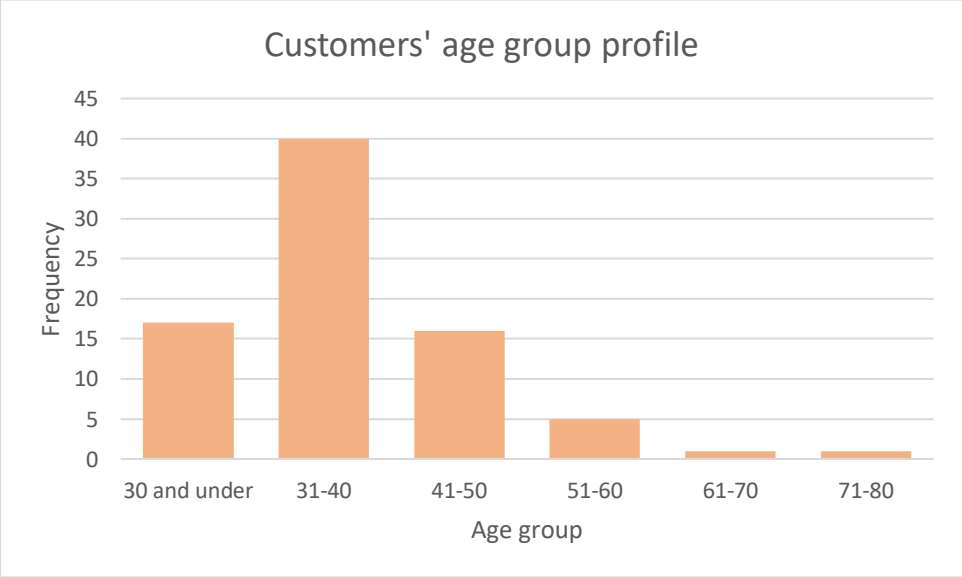


Figure 9 displays the age group profile of the customer sample

Table 9

Bin	*Frequency*
30 and under	17
31-40	40
41-50	16
51-60	5
61-70	1
71-80	1

Tenure

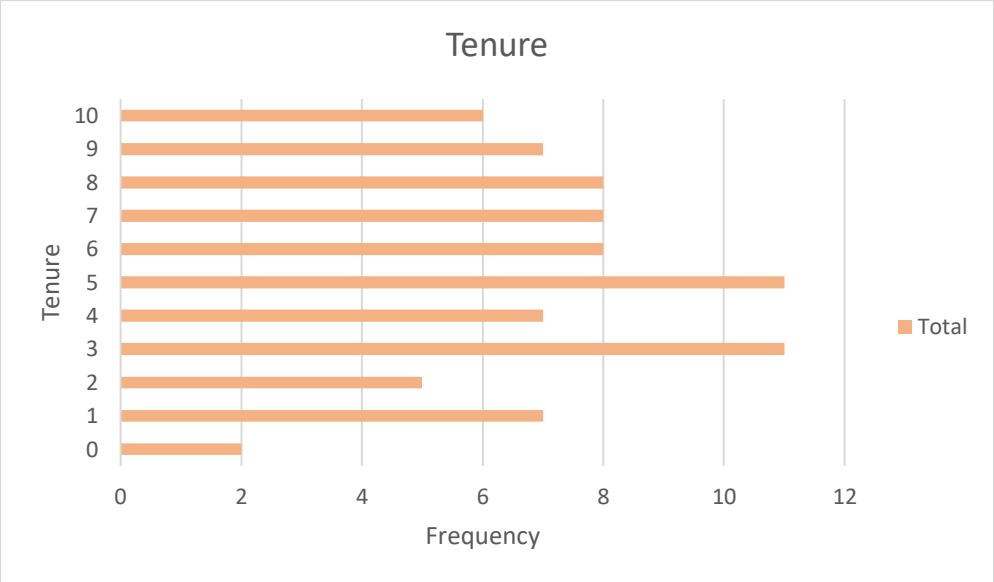


Figure 10 displays customers' tenure and their frequency

Table 10 shows customers' tenure and their frequency

Tenure	**Count of Tenure**
0	2
1	7
2	5
3	11
4	7
5	11
6	8
7	8
8	8
9	7
10	6
Grand Total	**80**

Account balance

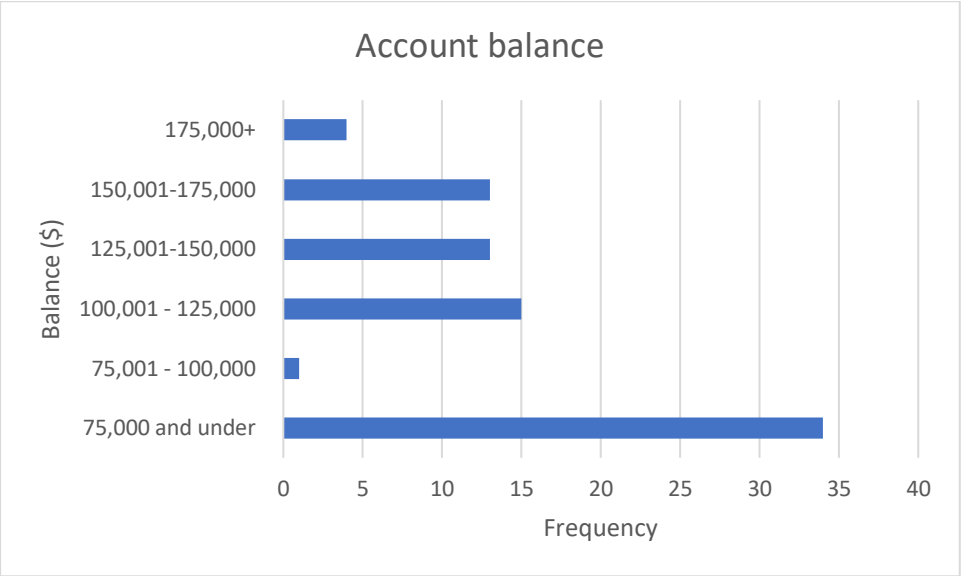


Figure 11 displays the customers' account balance

Table 11 shows the key parameters for the customers' account balance

Balance	
Mean	65554.66
Median	89013.73
Range	169025.8
Standard Deviation	59779.78457
Minimum	0
Maximum	169025.8
Q1	0
Q3	114641.5

Number of products

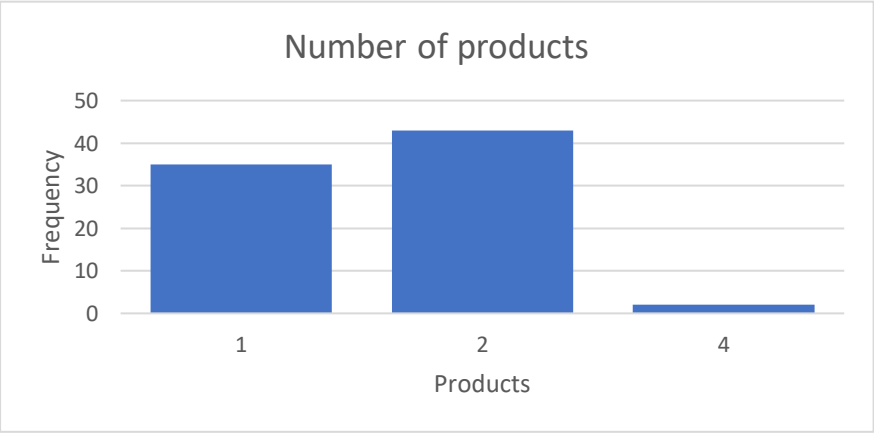


Figure 12 displays the number of products of the customers

Table 12

Number of products	Frequency
1	35
2	43
4	2

Estimated Salary

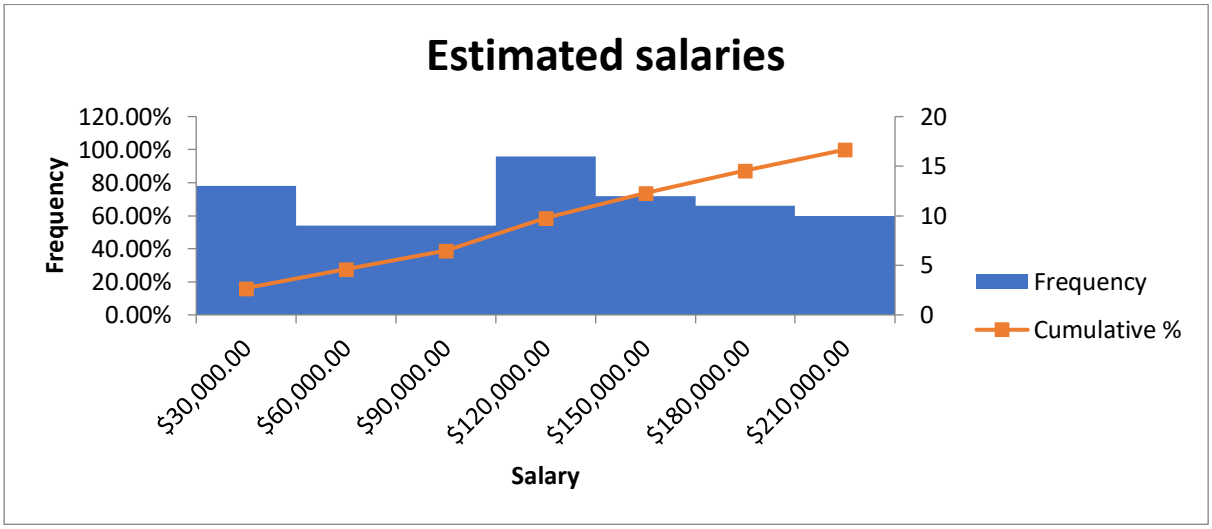


Figure 13 displays the estimated salaries of customers

Table 13

Estimated Salary	
Mean	101779
Median	105858.5
Range	196048.5
Minimum	468.94
Maximum	196517.4
Q1	48859.99
Q3	151733.5

Appendix 3 – Confidence Intervals

1. Average balance for females

Balance	
Mean	66307.107
Standard Error	10571.732
Median	95296.23
Mode	0
Standard Deviation	61643.259
Sample Variance	3.8E+09
Kurtosis	-1.831111
Skewness	-0.055095
Range	169025.83
Minimum	0
Maximum	169025.83
Sum	2254441.6
Count	34
Confidence Level(95.0%)	21508.35

Confidence intervals:	
Upper value	87815.457
Lower value	44798.757

True population mean:

Row Labels	**Average of Balance**
Female	$66,307.11
Grand Total	**$66,307.11**

2. Average age for customers from Spain

Age	
Mean	35.913043
Standard Error	1.6125901
Median	35
Mode	31
Standard Deviation	7.7337104
Sample Variance	59.810277
Kurtosis	0.5541366
Skewness	0.8529753
Range	30
Minimum	24
Maximum	54
Sum	826
Count	23
Confidence Level(95.0%)	3.3443072

Confidence interval:

Upper value	39.257351
Lower value	32.568736

True population mean:

Row Labels	**Average of Age**
Spain	35.91304348
Grand Total	**35.91304348**

Appendix 4 – Hypothesis testing

1. French customers more loyal than German customers?

Excel:

Ho: mean F - mean G <=0
Ha: mean F - mean G > 0
1 tailed test
Alpha = 0.05
Equal variance

t-Test: Two-Sample Assuming Equal Variances

	Tenure F	*Tenure G*
Mean	5.538461538	5.277778
Variance	8.044534413	9.388889
Observations	39	18
Pooled Variance	8.46006216	
Hypothesized Mean Difference	0	
df	55	
t Stat	0.314527287	
P(T<=t) one-tail	0.377155336	
t Critical one-tail	1.673033965	
P(T<=t) two-tail	0.754310672	
t Critical two-tail	2.004044783	

Since the t stat < t critical, do not reject

Python:

In [3]:
```
import pandas as pd
import scipy.stats as st
df = pd.read_excel("Assignment.xlsx", "Sample")
#filtered data
fembal = df[df["Gender"] == "Female"]
#print(fembal.shape)

n = fembal [["Balance"]].count()
df_fembal = n - 1
meanfb = fembal [["Balance"]].mean()
stdev = fembal [["Balance"]].std()
stderrfb = stdev/(n**0.5)
print("The 95% confidence interval for the mean balance is: ", st.t.interval(0.95, df_fembal, meanfb, stderrfb))
```

```
The 95% confidence interval for the mean balance is:  (array([44798.75727245]), array([87815.4568452]))
```

In [6]:
```
import pandas as pd
import scipy.stats as st
dfp = pd.read_excel("Assignment.xlsx", "Sample")
#filtered data
fembalp = dfp[dfp["Gender"] == "Female"]
print(fembalp.shape)
meanfbp = fembalp [["Balance"]].mean()
print ("The true mean is", meanfbp)
```

```
(34, 9)
The true mean is Balance    66307.107059
dtype: float64
```

2. Different in salary between male and female?

Excel:

Ho: meab f - mean m = 0
Ha: mean f - mean m not equal to 0
2 tailed test
t-Test: Two-Sample Assuming Equal Variances

	Female Salary	*Male salary*
Mean	105974.2659	98678.12913
Variance	3418020293	3626657325
Observations	34	46
Pooled Variance	3538387811	
Hypothesized Mean Difference	0	
df	78	
t Stat	0.542330003	
P(T<=t) one-tail	0.294568669	
t Critical one-tail	1.664624645	
P(T<=t) two-tail	0.589137337	
t Critical two-tail	1.990847069	

Since t stat is between the two critical values, do not reject

Python:

In [9]:
```
import pandas as pd
import scipy.stats as st
df = pd.read_excel("Assignment.xlsx", "Sample")
#filtered data
spain = df[df["Geography"] == "Spain"]
print(spain.shape)
n = spain [["Age"]].count()
df_spain = n - 1
meanspain = spain [["Age"]].mean()
stdev = spain [["Age"]].std()
stderrspain = stdev/(n**0.5)
print("The 95% confidence interval for the mean age is: ", st.t.interval(0.95, df_spain, meanspain, stderrspain))
```

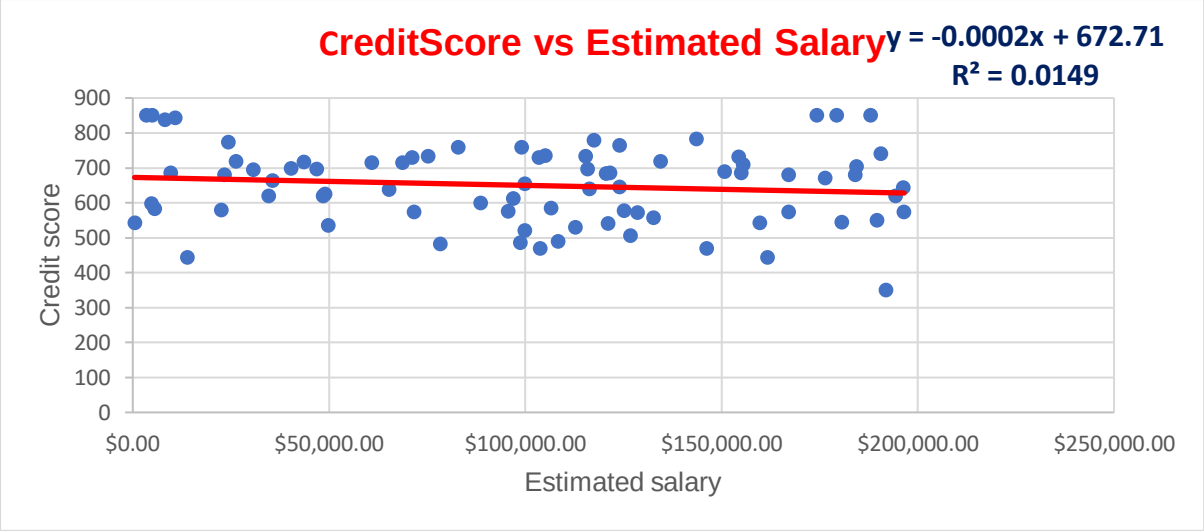
```
(23, 9)
The 95% confidence interval for the mean age is:  (array([32.56873631]), array([39.25735064]))
```

In [10]:
```
dfp = pd.read_excel("Assignment.xlsx", "Sample")
#filtered data
spainp = dfp[dfp["Geography"] == "Spain"]
meanspainp = spainp [["Age"]].mean()
print ("The true mean age is", meanspainp)
```

```
The true mean age is Age    35.913043
dtype: float64
```

In [ ]:

Appendix 4 – Correlation and regression



Regression Statistics		
Multiple R	-0.12188573	Correlation coefficient
R Square	0.014856132	coefficient of determination
Adjusted R Square	0.002226082	
Standard Error	110.3781955	
Observations	80	

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	
Intercept	672.711086	24.65470774	27.2853	1.19E-41	623.6273	7
EstimatedSalary	-0.00022744	0.000209708	-1.08455	0.281461	-0.00064	

Credit score = -0.000227 estimated Salary + 672.711

Ho: beta1 = 0
HA: beta 1 not equal
to 0
Alpha = 0.05
2 tailed test
df = 78
t value -1.08455192

t critical value 1.990847069

Since t is in between the two critical values, do not reject