Pilgrim Bank

Jahnavi Chavali – ID: 86169170

Introduction

Pilgrim Bank would like to determine how to deal with online customers in order to maximise profits for the company. If the bank were to discover that online customers brought in more profit, then they would be offered incentives to continue online banking with Pilgrim. However, if the online customers were not any better than offline customers in terms of the amount of profit they generated, then the bank would need to consider charging the customers for using their online services. The data for customers' profit, whether they are online or offline customers and other demographic information is available. While there is a difference in the average profits of online and offline customers, it is required to determine whether this difference is statistically significant and determine which of the above two options is better for the company's profitability.

Hypothesis

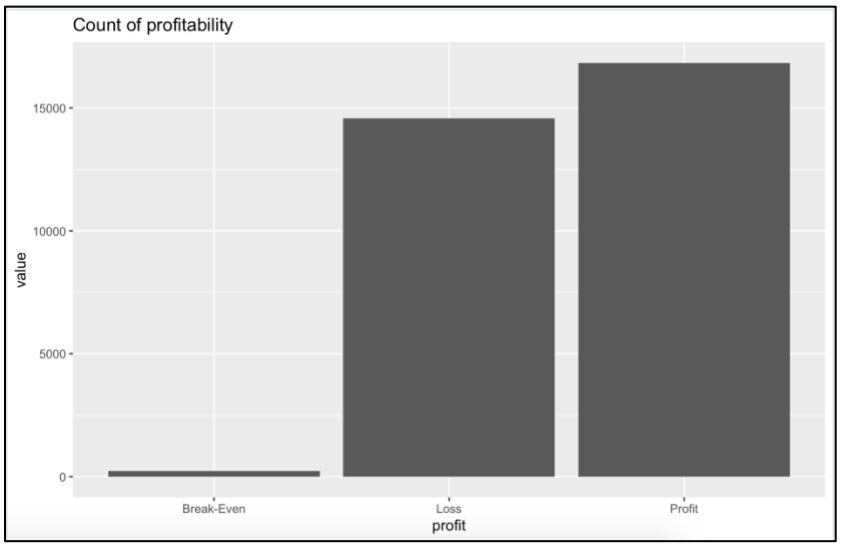
Based on the data, the mean profit for online customers is 116.7 and that of offline customers is 110.8. There is a 6-unit difference in the means, therefore it appears that the average profit of online customers is greater. Therefore, the hypothesis for this analysis is that the difference in the mean profit between online and offline customers is statistically significant.

Data Analysis

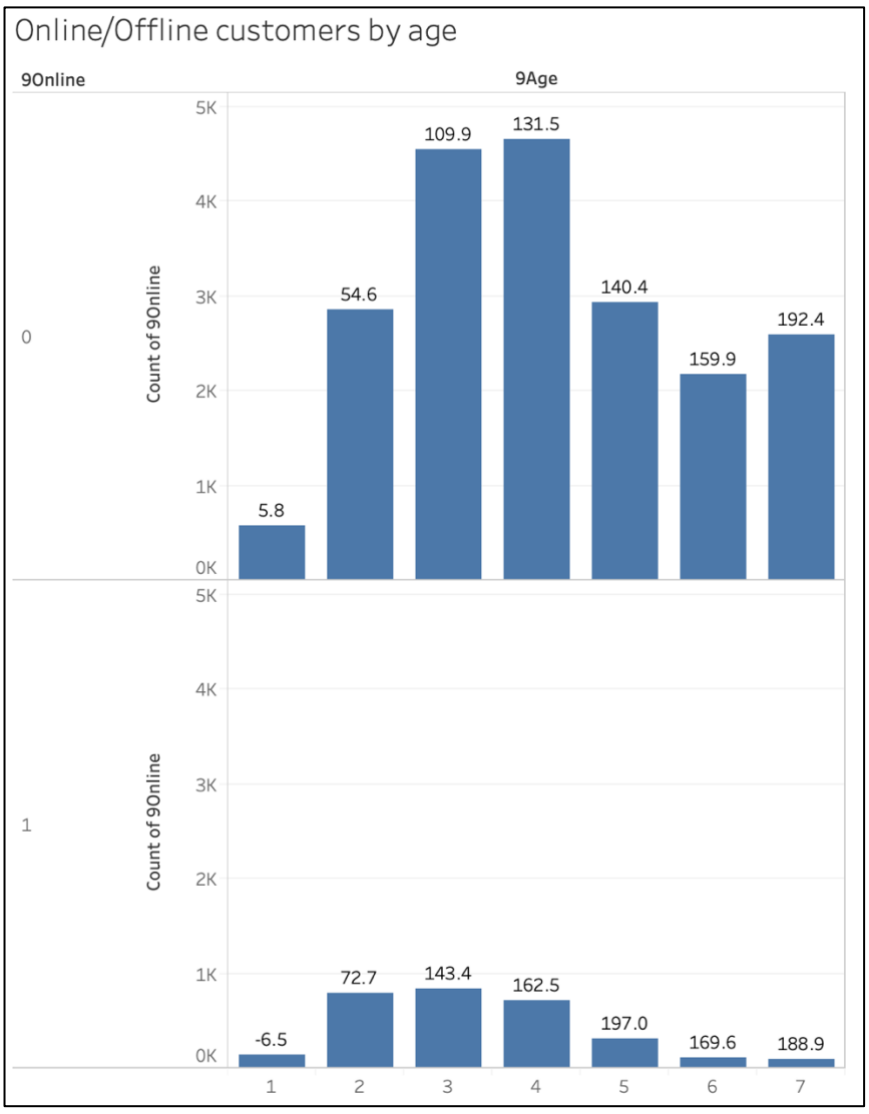
Part 1: Conclusion about customer profitability over the entire population

Data Type

Pilgrim Bank's customer population consists customers of a wide age range, income range and belonging to 3 geographic locations. The dependent variable, 'Profit', is a numeric variable, as is 'Tenure'. 'Age', 'Income', 'District' and 'Online' are categorical variables. By looking over the profitability data, it can be concluded that there is very little difference in the number of customers that have recorded a profit, and those that have recorded a loss. There are 2252 more people who are profitable compared to those reporting a loss. Only 222 people broke even.

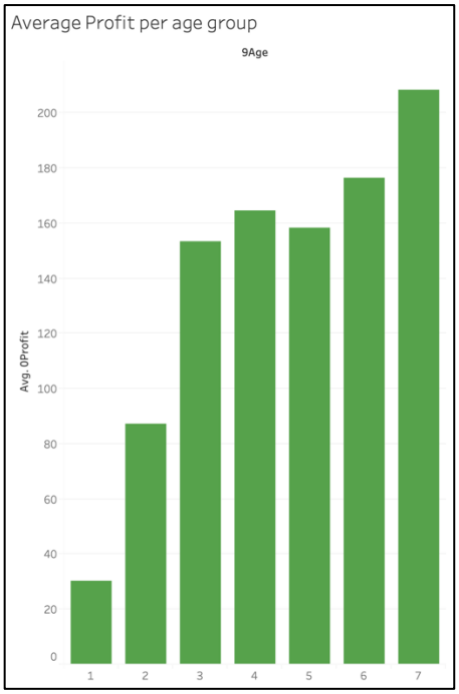


Graph 1

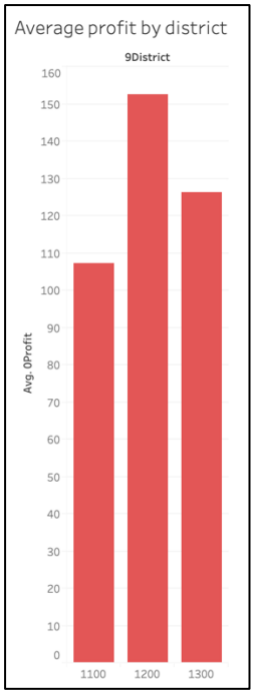


Graph 2

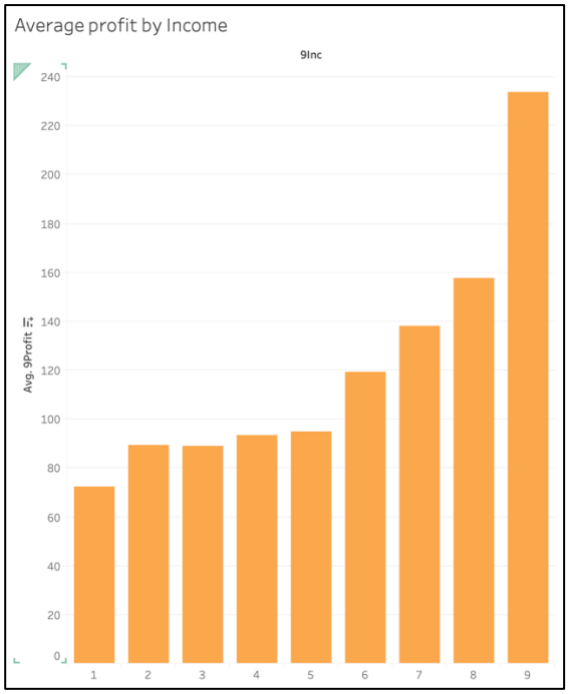
In Graph 2, the top plot shows the number of offline customers by age, and the bottom plot shows the number of online customers. There is a significant difference between the number of customers in the second age range and the third age range. There are significant;y more offline customers in the age range of 25-34 compared to the 15-24 age range.
The same cannot be said for online customers, where the number of people in the 2nd, 3rd and 4th age groups are similar.



Graph 3



Graph 4



Graph 5

The average profit increases as age range increases (Graph 3). The customers who are below 15 years make the least amount of profit, while the customers who are 65 and older make the most amount of profit. There is a small dip in average profit in the 5th age group (45-54 years).

Those in district “1200” have the highest average profit (Graph 4), and, in general, the average profit increases with increase in Salary (Graph 5) – higher income groups tend to have higher profits.

Stability

Split-Half Test

In order to determine the stability of the profit in the required time period, a split half test is run. The dataset was randomly divided into two parts, and the correlation was determined in multiple trials. The results of the test are as shown below. The correlation between the two halves is fairly low, indicating that the profit data is not very stable.

-0.0118651

-0.0171399

-0.0136209

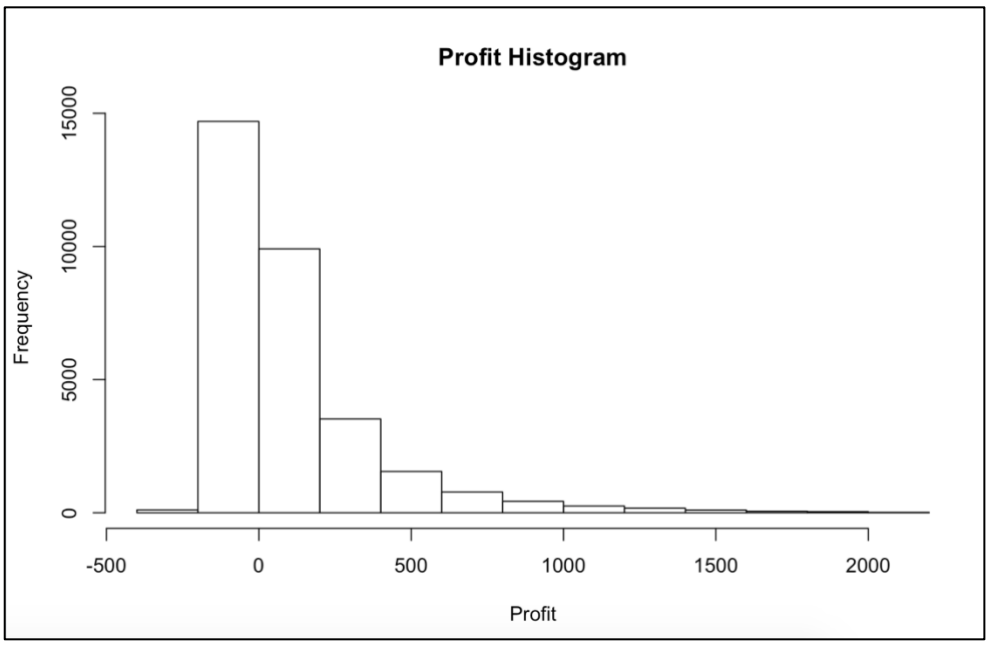
The customer data includes profits over two years – 1999 and 2000. In order to determine whether the profits of the year 2000 are representative of the profits in year 1999, correlation was run between the two years' profits. The results are as shown below:

0.599337

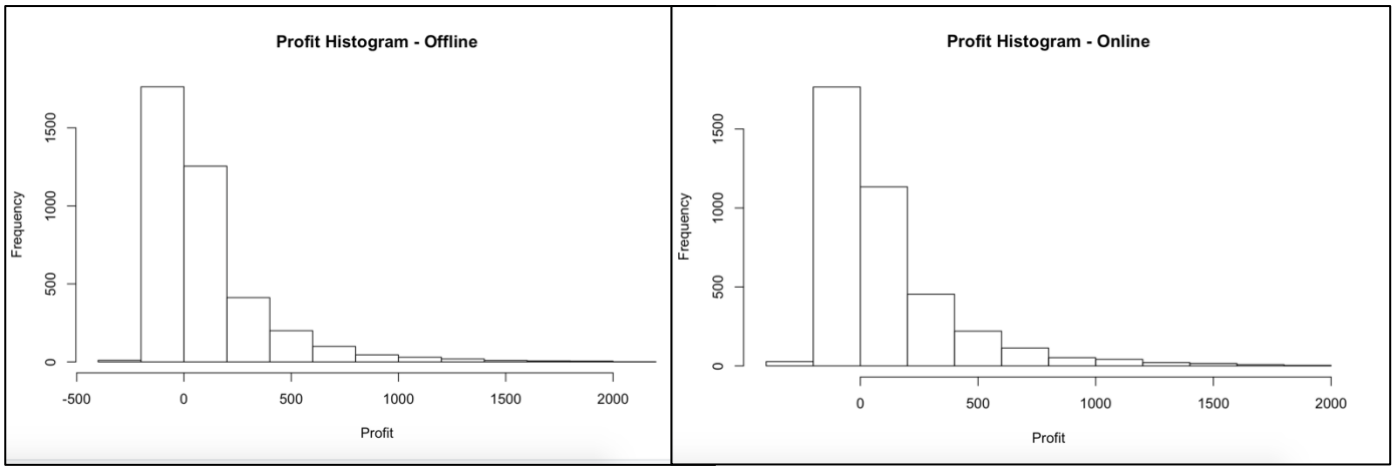
The correlation in this case is much higher at 0.6, indicating that the correlation between the two profits is strong.

Shape

The profit distribution of the entire sample is right skewed, i.e. most of the values are small, and very few that go past + 500 and -500.



After the data is split between 'Online' and 'Offline' customers, the profits remained right-skewed. This indicates that the distribution of the profits amongst 'Online' and 'Offline' customers are similar.



Spread

```
> summary(customer_type_online)
 profit....customer_type..1...9Profit.
 Min.   :-221.0
 1st Qu.: -43.0
 Median :  12.0
 Mean   : 116.7
 3rd Qu.: 186.0
 Max.   :1979.0
```

```
> summary(customer_type_offline)
 profit....customer_type..0...9Profit.
 Min.   :-221.0
 1st Qu.: -33.0
 Median :   9.0
 Mean   : 110.8
 3rd Qu.: 161.0
 Max.   :2071.0
```

Both online and offline customers have the same minimum value, but the offline data has a higher maximum value of profit. Despite this, it's median is lower – indicating that online customers have more positive values for 'Profit'. The averages of both are similar, with online customers having a slightly higher average.

Levene's Test

The Levene's test determines if the variances of the two samples are homogeneous. Levene's test is used instead of a regular f-test because the 'profit' is not normally distributed.

Null Hypothesis (H_0):

$\boldsymbol{H_0}$: $There\ is\ no\ significant\ difference\ in\ the\ variance\ in\ the\ profit\ between\ online\ and\ offline\ customers.\ i.e.(\sigma^2_{online} = \sigma^2_{offline})$

Alternate Hypothesis (H_a):

$\boldsymbol{H_a}$: $There\ is\ a\ significant\ difference\ in\ the\ variance\ in\ the\ profit\ between\ online\ and\ offline\ customers.\ i.e.(\sigma^2_{online} \neq \sigma^2_{offline})$

```
Levene's Test for Homogeneity of Variance (center = mean)
         Df F value   Pr(>F)
group     1  8.9851 0.002725 **
      22810
```

As the p-value is very small, we can reject the null hypothesis and conclude that there is a significant difference in the variances of profit amongst online and offline customers. Because there is a significant difference in variance, the data is not normalised, and the sample sizes are unequal, **Welch t-test** (parametric test) will need to be conducted to determine if there is a significant difference in the average profit between online and offline customers, as opposed to a regular t-test.

Part 2: Statistical significance of difference in means between online and offline customers

Central Tendency

Welch t-test

Student's t-test assumes equal sample sizes and equal variance between the two samples. As neither assumption is satisfied, Welch's t-test will be used to determine whether the difference in mean profits between the two groups is significant.

Null Hypothesis (H_0):

$\boldsymbol{H_0}$: $There\ is\ no\ significant\ difference\ in\ the\ average\ profitability\ between\ the\ online\ and\ offline\ customers\ at\ Pilgrim\ Bank.\ \ i.e.(\mu_{online} = \mu_{offline})$

Alternate Hypothesis (H_a):

$\boldsymbol{H_a}$: $There\ is\ a\ significant\ difference\ in\ the\ average\ profitability\ between\ the\ online\ and\ offline\ customers\ at\ Pilgrim\ Bank.i.e.(\mu_{online}\ \neq\ \mu_{offline})$

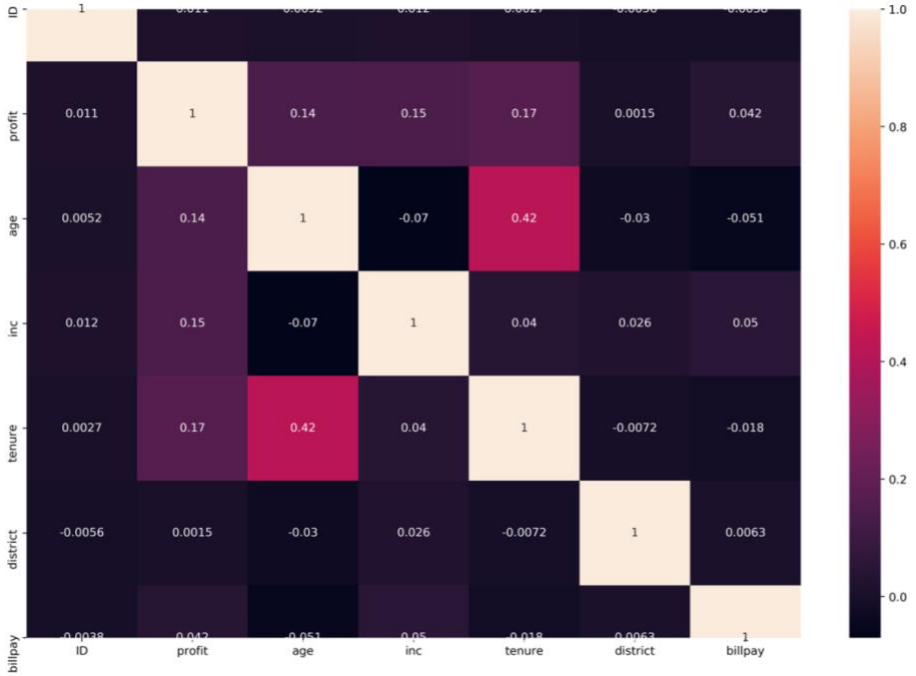
```
        Welch Two Sample t-test

data:  profit_types$profit_online and profit_types$profit_offline
t = 1.0729, df = 7662.7, p-value = 0.2834
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.530142 18.902223
sample estimates:
mean of x mean of y
 116.6668  109.9808
```

As the p-value is greater than the significance value ($\propto = 0.05$), the null hypothesis **cannot** be rejected. Therefore, it can be concluded that ***there is no significant difference in the average profits between online and offline customers***.

Part 3: Role of demographics on customer profitability for online and offline customers.

Linear Regression

In order to determine the factors that have the greatest influence on the profit, a Linear Regression Equation is generated. As seen from the Graphs 3, 4, and 5, the relationships of various dependent variables with profit is linear. To run a Linear Regression, there cannot be a high degree of multicollinearity – a high correlation between the dependent variables. By generating a correlation Matrix, it is seen that there is not a high degree of correlation between the variables. All the values are under the threshold (0.5) for multicollinearity. A more detailed matrix is given in Appendix A.



The table in Appendix B shows the coefficients for a linear equation with 'Profit' as the dependent variable. According to their p-values, the variables 'Income'=2, 'Income'=3, 'Income'=4, 'Income'=5 and 'District'=1300 do not have a significant affect on the profit of a customer. i.e. those who earn below $50,000 per year, and those who are living in district '1300'.

All other variables have p-values < 0.05. Those customers in age group '6' and '7' (55 years and above) have the greatest impact on the profit. When a customer is in on of these age groups, the profit is expected to increase by 100.1 and 135.7 units respectively, keeping all other variables constant. Additionally, those in Income group '9' (those earning more than $125,000) are expected to increase the profit by an average of 146 units, keeping all other variables constant. The age range of 25-54 increase profit by 70-80 unites on average, while the income range 75,000-125,000 also have a significant affect on the profit. The lower age groups and income classes have a lesser influence on the profit.

Running Linear Regression for online and offline customers separately generates the tables shown below. Table 1 represents online customers' profit, and Table 2, offline. The p-value for Income group 5 for online customers is 0.024, and therefore is significant for determining profit. However, it is not significant for offline customers. Therefore, those customers in the income range $40,000-$49,999 have a higher impact for profitability for online customers. The same can be observed for those in District 1200 – whose p-value is 0.376 for offline customers, indicating that it is not as important a factor in determining the profit of offline customers. However, it is significant for online customers.

For offline customers, those customers in Age groups 2, 4 and 5 impact the profit negatively – they cause the average profit to decrease. However in the case of online customers, their coefficients have a positive value – indicating that if the customer belongs to any of these age groups, the profit increases.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-49.5587	14.309	-3.463	0.001	-77.606	-21.511
C(age)[T.2]	25.1682	13.313	1.891	0.059	-0.926	51.263
C(age)[T.3]	62.3481	12.995	4.798	0.000	36.877	87.819
C(age)[T.4]	67.6783	13.060	5.182	0.000	42.080	93.277
C(age)[T.5]	69.7204	13.486	5.170	0.000	43.287	96.154
C(age)[T.6]	93.0734	13.871	6.710	0.000	65.884	120.262
C(age)[T.7]	129.7821	13.677	9.489	0.000	102.974	156.590
C(inc)[T.2]	3.3238	12.131	0.274	0.784	-20.455	27.102
C(inc)[T.3]	12.1118	8.790	1.378	0.168	-5.117	29.341
C(inc)[T.4]	13.1108	8.954	1.464	0.143	-4.440	30.662
C(inc)[T.5]	20.2775	8.989	2.256	0.024	2.657	37.898
C(inc)[T.6]	41.1052	7.826	5.252	0.000	25.765	56.445
C(inc)[T.7]	60.2396	8.607	6.999	0.000	43.368	77.111
C(inc)[T.8]	76.3740	9.900	7.715	0.000	56.970	95.778
C(inc)[T.9]	142.2881	8.885	16.014	0.000	124.872	159.704
C(district)[T.1200]	18.7133	6.657	2.811	0.005	5.664	31.762
C(district)[T.1300]	7.5464	8.113	0.930	0.352	-8.357	23.449
tenure	4.0411	0.247	16.385	0.000	3.558	4.525

Table 1 – online customers

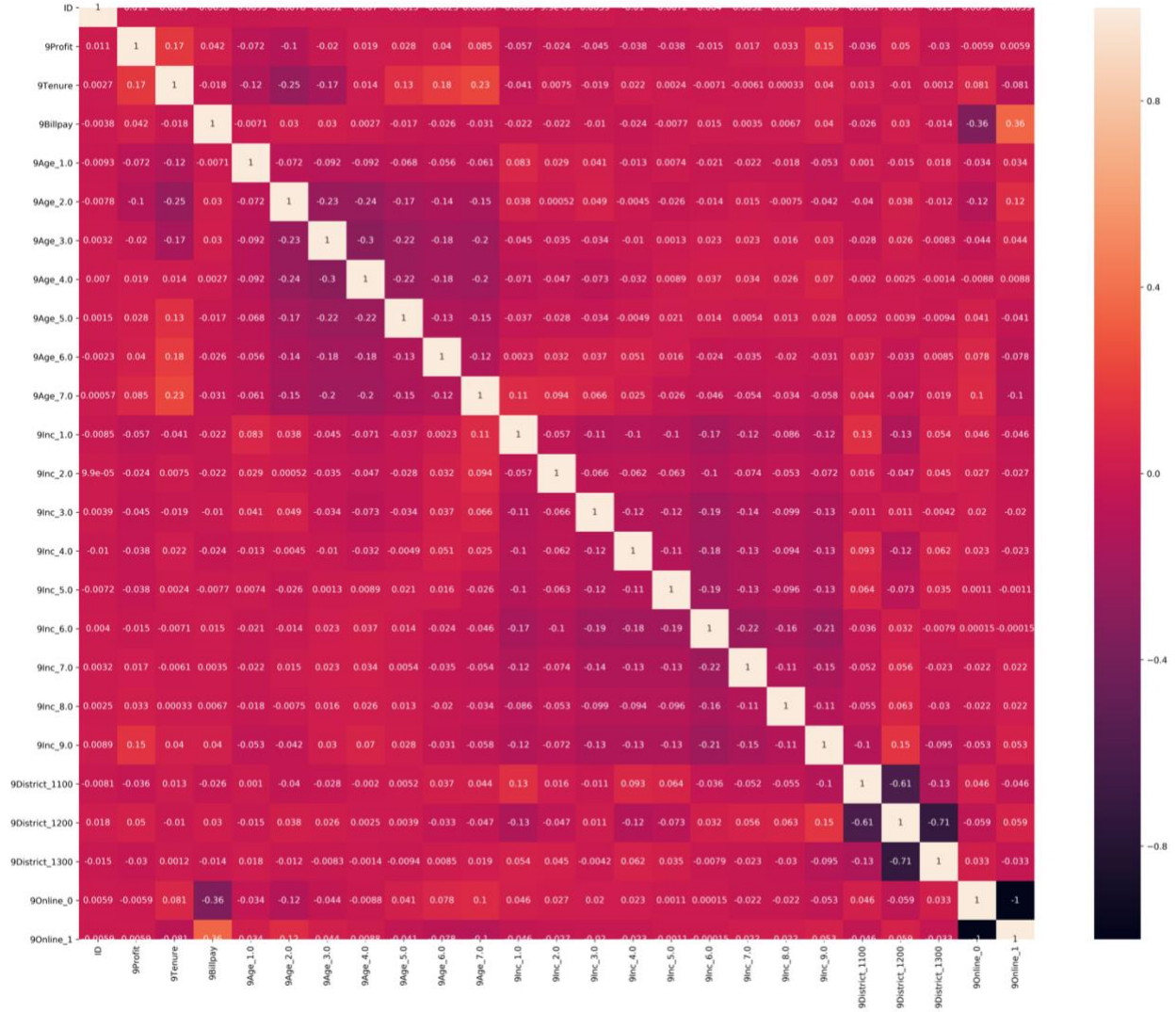
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-66.5329	37.654	-1.767	0.077	-140.363	7.297
C(age)[T.2]	48.7625	27.130	1.797	0.072	-4.434	101.959
C(age)[T.3]	96.9889	27.290	3.554	0.000	43.479	150.499
C(age)[T.4]	98.8588	27.960	3.536	0.000	44.036	153.682
C(age)[T.5]	130.4871	30.559	4.270	0.000	70.568	190.406
C(age)[T.6]	125.5742	36.799	3.412	0.001	53.420	197.729
C(age)[T.7]	134.9047	38.283	3.524	0.000	59.841	209.968
C(inc)[T.2]	-23.7901	41.464	-0.574	0.566	-105.092	57.511
C(inc)[T.3]	2.6111	28.470	0.092	0.927	-53.212	58.434
C(inc)[T.4]	-7.1185	29.009	-0.245	0.806	-63.999	49.762
C(inc)[T.5]	-11.1199	27.927	-0.398	0.691	-65.879	43.639
C(inc)[T.6]	30.5624	25.422	1.202	0.229	-19.285	80.410
C(inc)[T.7]	63.6576	26.574	2.396	0.017	11.553	115.763
C(inc)[T.8]	89.4181	28.947	3.089	0.002	32.659	146.177
C(inc)[T.9]	166.9738	26.516	6.297	0.000	114.982	218.965
C(district)[T.1200]	19.6285	22.178	0.885	0.376	-23.858	63.115
C(district)[T.1300]	3.8272	26.474	0.145	0.885	-48.082	55.736
tenure	4.5999	0.794	5.792	0.000	3.043	6.157

Table 2 – offline customers

Conclusion

The profit generated by online bank customers is not any different to the profit generated by offline customers; there is no statistical difference between their average profits. Therefore, there is no need for Pilgrim Bank to invest in any incentives in order to persuade their existing online customers to continue online banking. However, the data is not very stable, and efforts should be made to acquire data that is more representative of the customer population.

Appendix A


ID
9Profit
9Tenure
9Billpay
9Age_1.0
9Age_2.0
9Age_3.0
9Age_4.0
9Age_5.0
9Age_6.0
9Age_7.0
9Inc_1.0
9Inc_2.0
9Inc_3.0
9Inc_4.0
9Inc_5.0
9Inc_6.0
9Inc_7.0
9Inc_8.0
9Inc_9.0
9District_1100
9District_1200
9District_1300
9Online_0
9Online_1
0.8
0.4
0.0
−0.4
−0.8

Appendix B

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-56.3975	13.196	-4.274	0.000	-82.263	-30.532
C(online)[T.1]	17.0251	5.498	3.097	0.002	6.249	27.801
C(age)[T.2]	29.9474	11.932	2.510	0.012	6.559	53.335
C(age)[T.3]	69.8003	11.709	5.961	0.000	46.849	92.751
C(age)[T.4]	74.6121	11.794	6.326	0.000	51.495	97.729
C(age)[T.5]	79.4354	12.250	6.484	0.000	55.424	103.447
C(age)[T.6]	100.0856	12.705	7.877	0.000	75.182	124.989
C(age)[T.7]	135.7456	12.528	10.835	0.000	111.190	160.302
C(inc)[T.2]	0.9934	11.651	0.085	0.932	-21.844	23.831
C(inc)[T.3]	10.9358	8.395	1.303	0.193	-5.519	27.390
C(inc)[T.4]	10.8613	8.553	1.270	0.204	-5.902	27.625
C(inc)[T.5]	15.9018	8.537	1.863	0.063	-0.831	32.634
C(inc)[T.6]	39.6959	7.471	5.313	0.000	25.053	54.339
C(inc)[T.7]	60.7904	8.159	7.450	0.000	44.797	76.783
C(inc)[T.8]	78.5513	9.316	8.432	0.000	60.291	96.812
C(inc)[T.9]	146.8121	8.367	17.547	0.000	130.413	163.212
C(district)[T.1200]	18.6401	6.379	2.922	0.003	6.137	31.143
C(district)[T.1300]	7.0957	7.758	0.915	0.360	-8.110	22.302
tenure	4.0877	0.235	17.363	0.000	3.626	4.549