PILGRIM CASE

NAME - ADWAIT TORO
STUDENT # - 89487135

APPROACH: -

- **Total 11 variables with 31634 records. After removing unnecessary columns, there were 7 variables with 31634 records.**
- **There are no duplicate entries for customers, otherwise there would've been a discrepancy in the profitability.**
- **Out of all the variables, 9Age and 9Inc has missing values i.e. the age bucket column and the income column.**
- **After removing the null values, the new data sheet has 22812 rows, 7 columns.**
- **Most of the customers are from geographical district of 1200 (17686).**
- **Dependent variable is Profitability (9Profit) for the bank through offline and online customers.**
- **After looking at the correlation matrix derived from the Pearson's test, Spearman's test, Kendall's test, Phik test, it is clear that there is no multicollinearity between the variables.**
- **But after taking a closer look at the data through minitab's 6-pack capability analysis, it was observed that the cronbach's alpha value was 0.017 which was < 0.5, which was unacceptable. Hence the data internally is not stable.**
- **The spread for the data is as follows –**
    - **Profit –**
        - **Range is from -221 to 2071**
        - **Standard deviation is 282.856**
        - **Mean is 127.1694**
        - **Median is 22**
        - **Thus, there is no normal distribution for the data.**

Data After cleaning (Overview): -

Overview

Dataset info

Number of variables	8
Number of observations	22812
Missing cells	0 (0.0%)
Duplicate rows	0 (0.0%)
Total size in memory	1.4 MiB
Average record size in memory	64.0 B

Variables types

Numeric	5
Categorical	1
Boolean	1
Date	0
URL	0
Text (Unique)	0
Rejected	1
Unsupported	0

Warnings

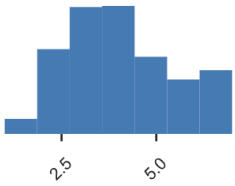
`ID` is highly correlated with `df_index` (ρ = 1) **Rejected**

Variables

9Age
Numeric

Distinct count	7
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

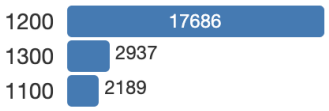
Mean	4.064658951
Minimum	1
Maximum	7
Zeros (%)	0.0%



Toggle details

9District
Categorical

Distinct count	3
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0

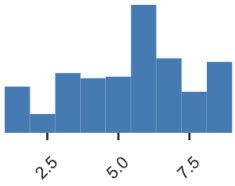


Toggle details

9Inc
Numeric

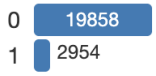
Distinct count	9
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	5.488076451
Minimum	1
Maximum	9
Zeros (%)	0.0%



9Online
Boolean

Distinct count	2
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0

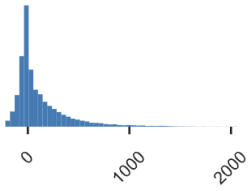


Toggle details

9Profit
Numeric

Distinct count	1548
Unique (%)	6.8%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	127.1693845
Minimum	-221
Maximum	2071
Zeros (%)	0.6%

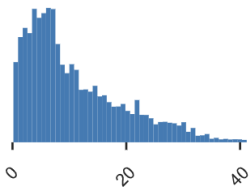


Toggle details

9Tenure
Numeric

Distinct count	487
Unique (%)	2.1%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

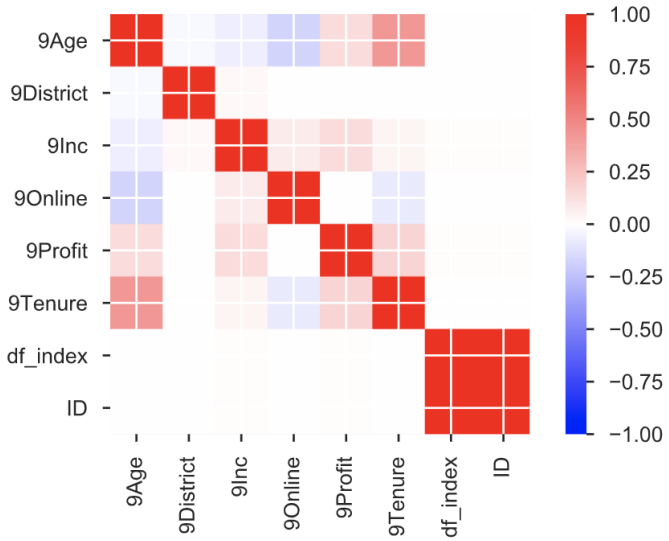
Mean	10.99631773
Minimum	0.16
Maximum	41.16
Zeros (%)	0.0%



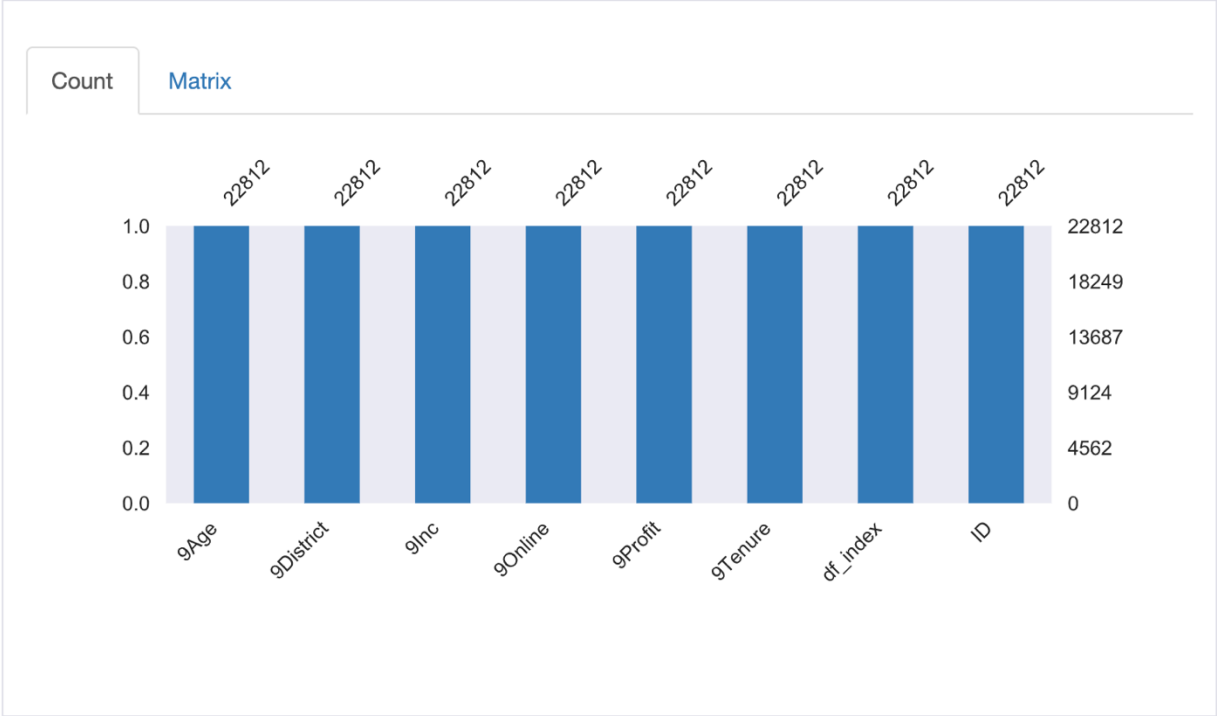
Toggle details

Correlations

Pearson's r | Spearman's ρ | Kendall's τ | Phik (ϕ_k) | Cramér's V (ϕ_c) | Recoded

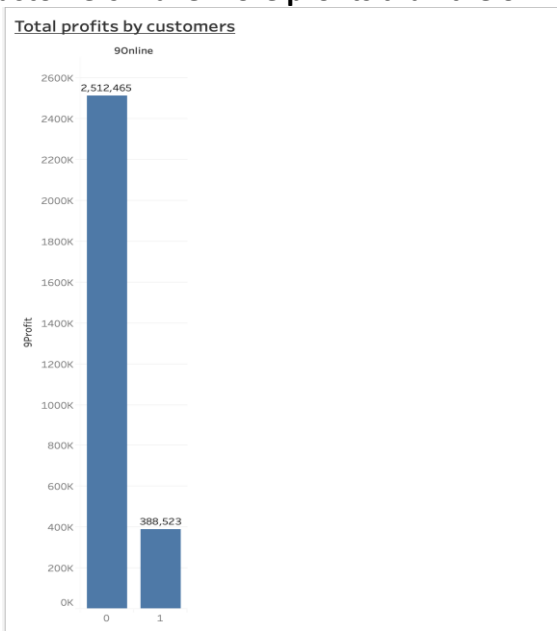


Missing values

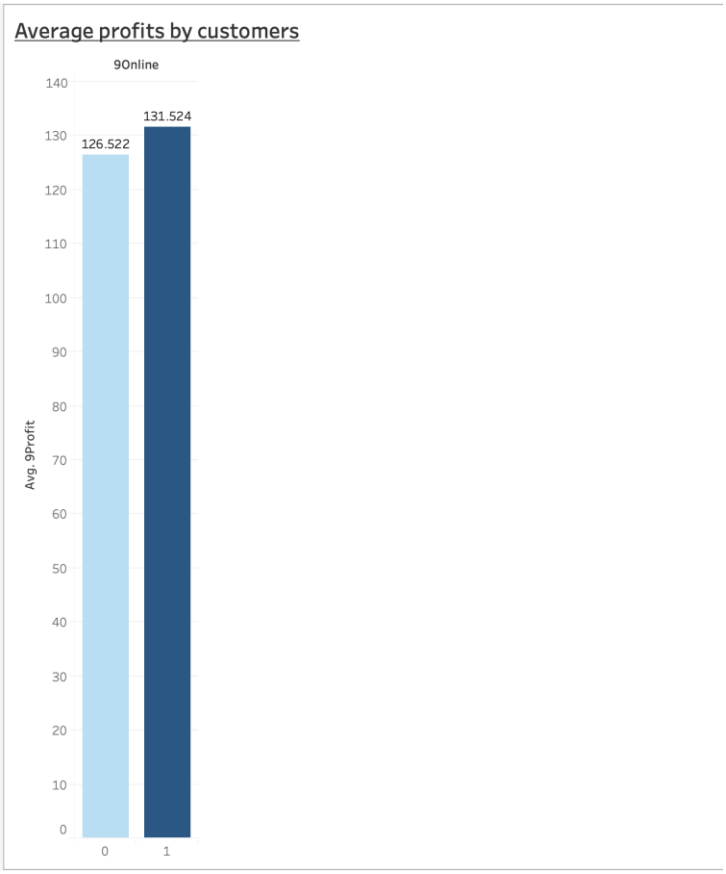


1. **Based on the customer data sample from 1999 what can Green conclude about customer profitability for Pilgrim Bank's entire customer population?**

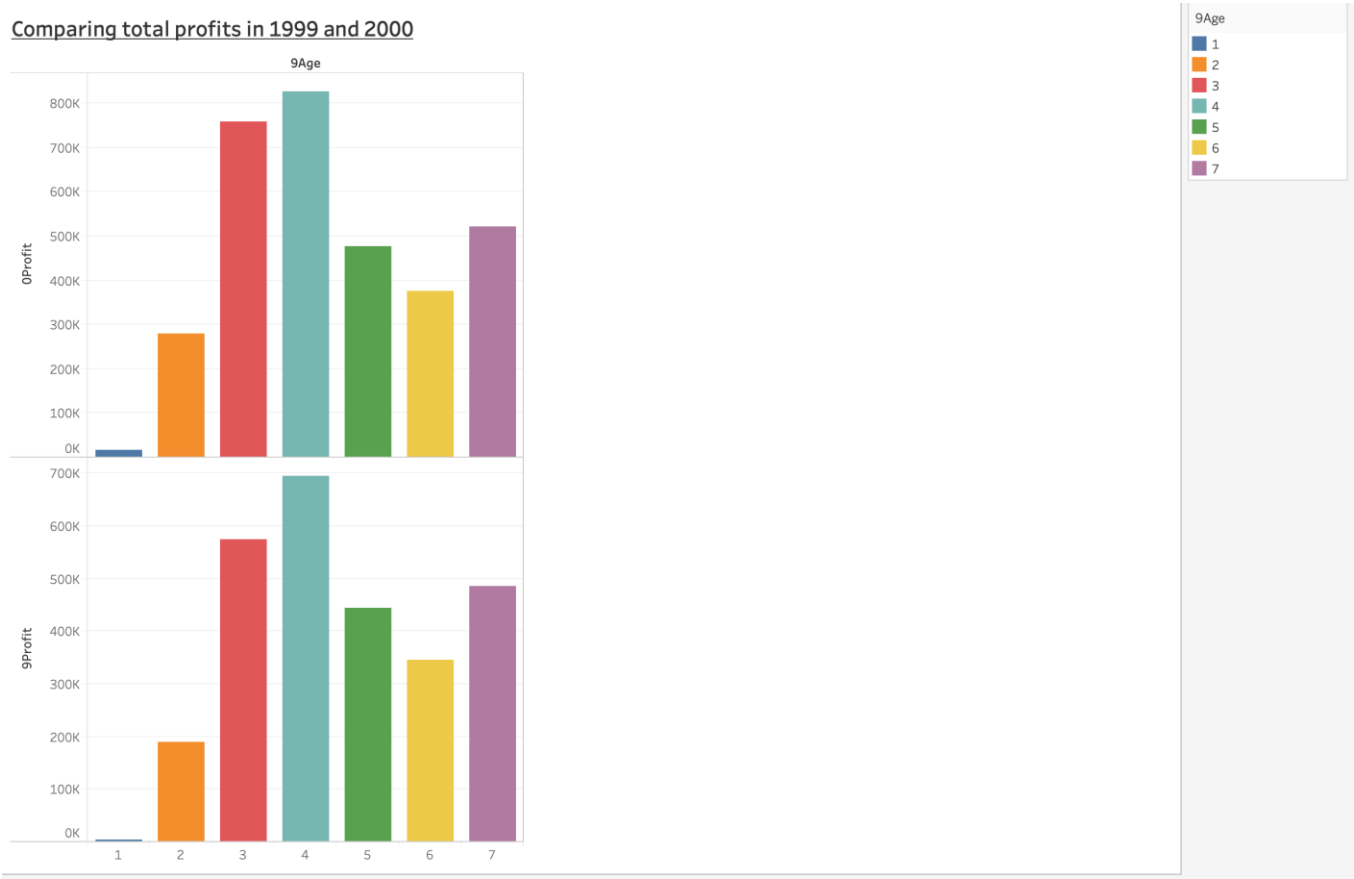
- **If we compare the total profit made by both online as well as offline mediums, it is clear that the offline customers make more profits than the online customers.**



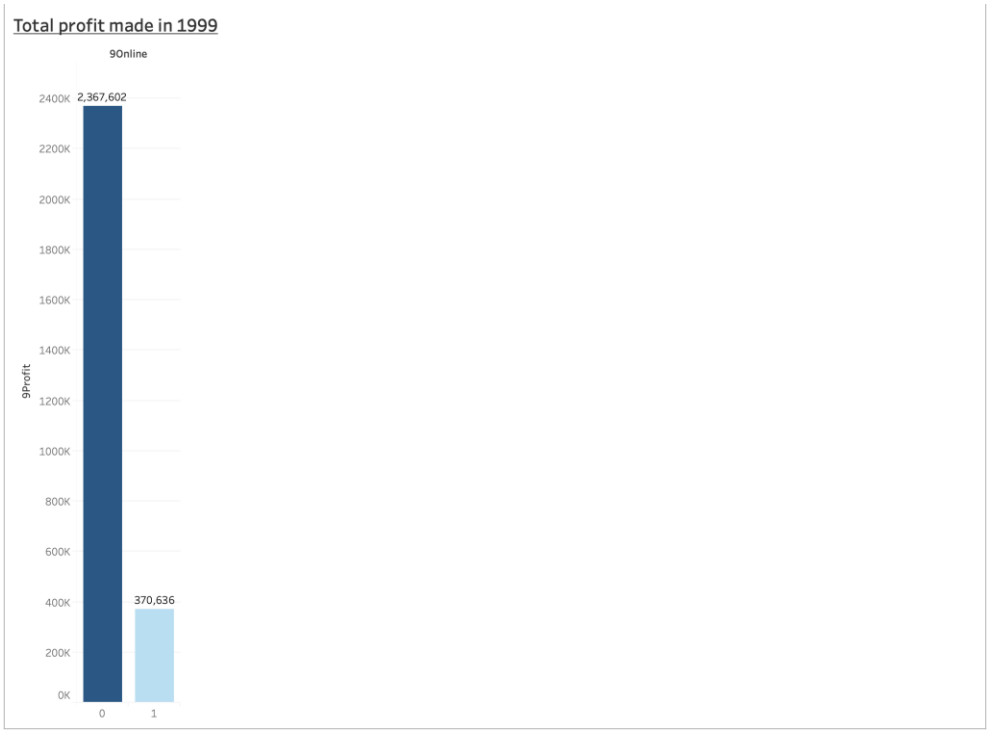
- **On the contrary, if we look at the average profit values for both the mediums, then it is clear that the average profit for both online customers and the offline customers is almost similar. This may lead us to two things-**
    - **Firstly, as the given data is imbalanced i.e. offline customers are more than the online customers, there is no sufficient data to conclude why the average price for online customers and that of offline customers is almost similar.**
    - **Secondly, there's a possibility that the negative values for profit in offline customers leads to the similarity in mean values.**

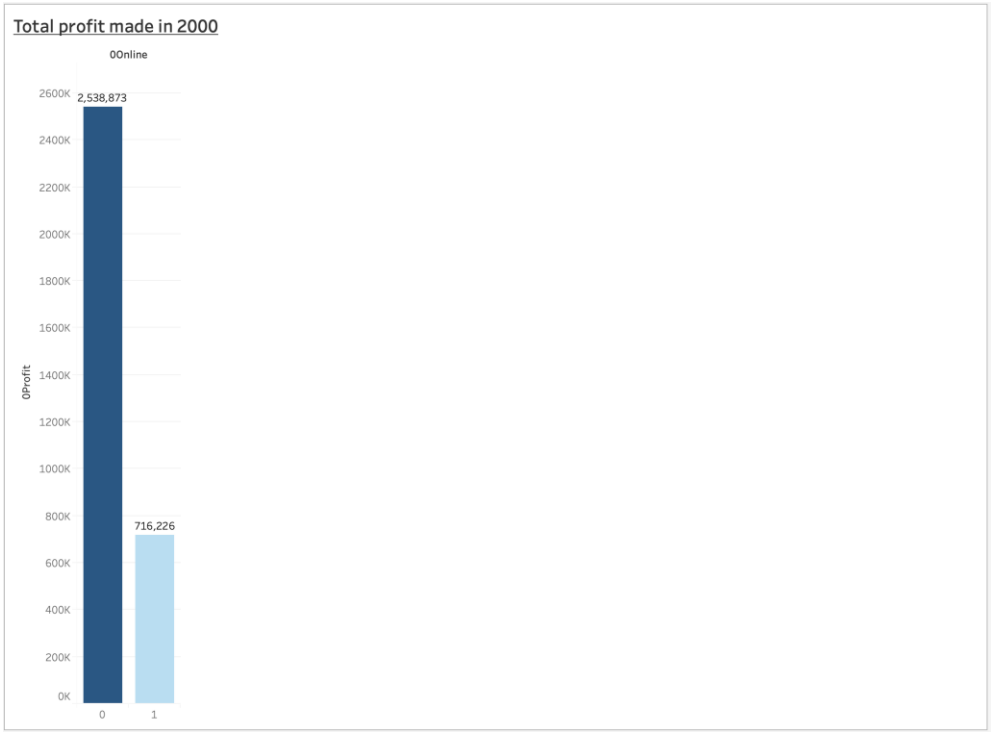


- **Next, if we compare the profits of 1999 and 2000, it is observed that the age group 4 = 35-44 years has the highest total profit in both these years. Thus, it is clear that the bank has more customers from this age group in both these years. Also, if we take a look at the graph neatly, we observe that there has been a greater increase in the total profit made by age group 0-15 years which was quite astonishing. Thus, it can be concluded that the bank has identified young people as potential targets.**



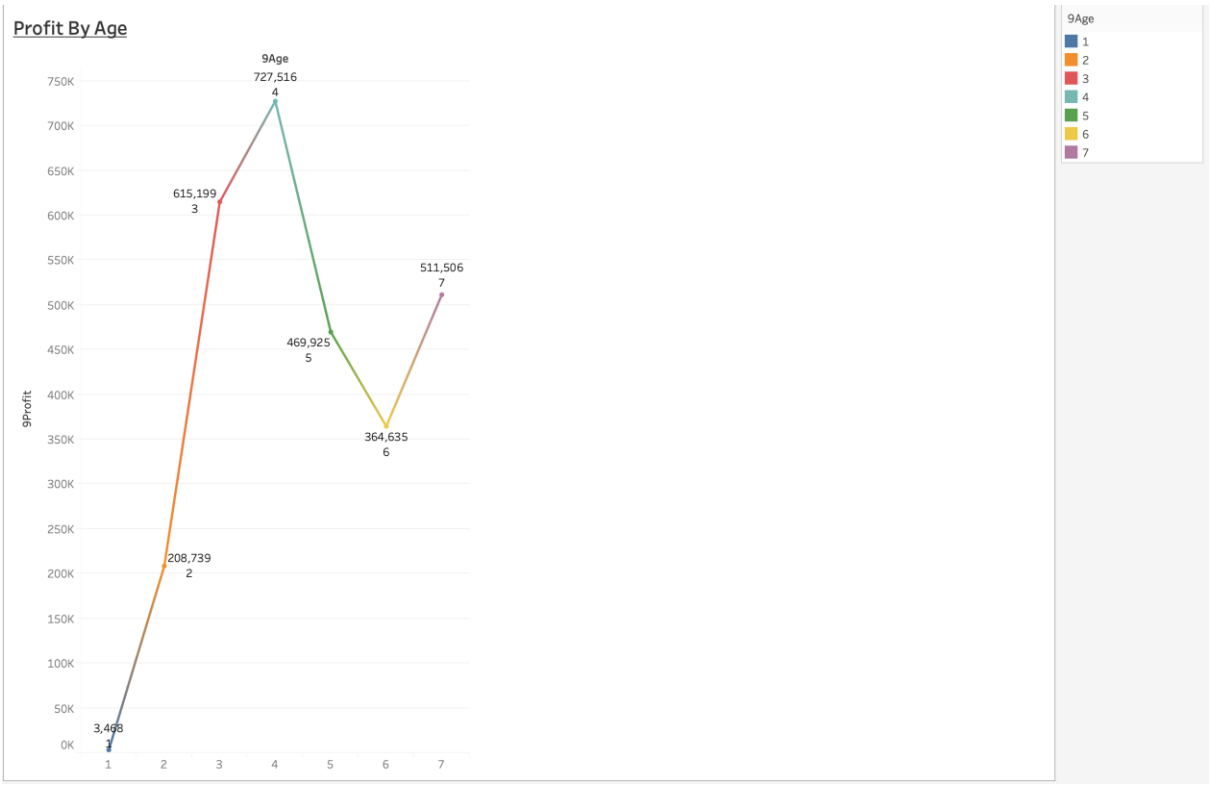
- **If we compare the total profits made in 1999 and 2000, it can be said that the number online customers increased in 2000 and hence the total profit made from online customers also increased.**



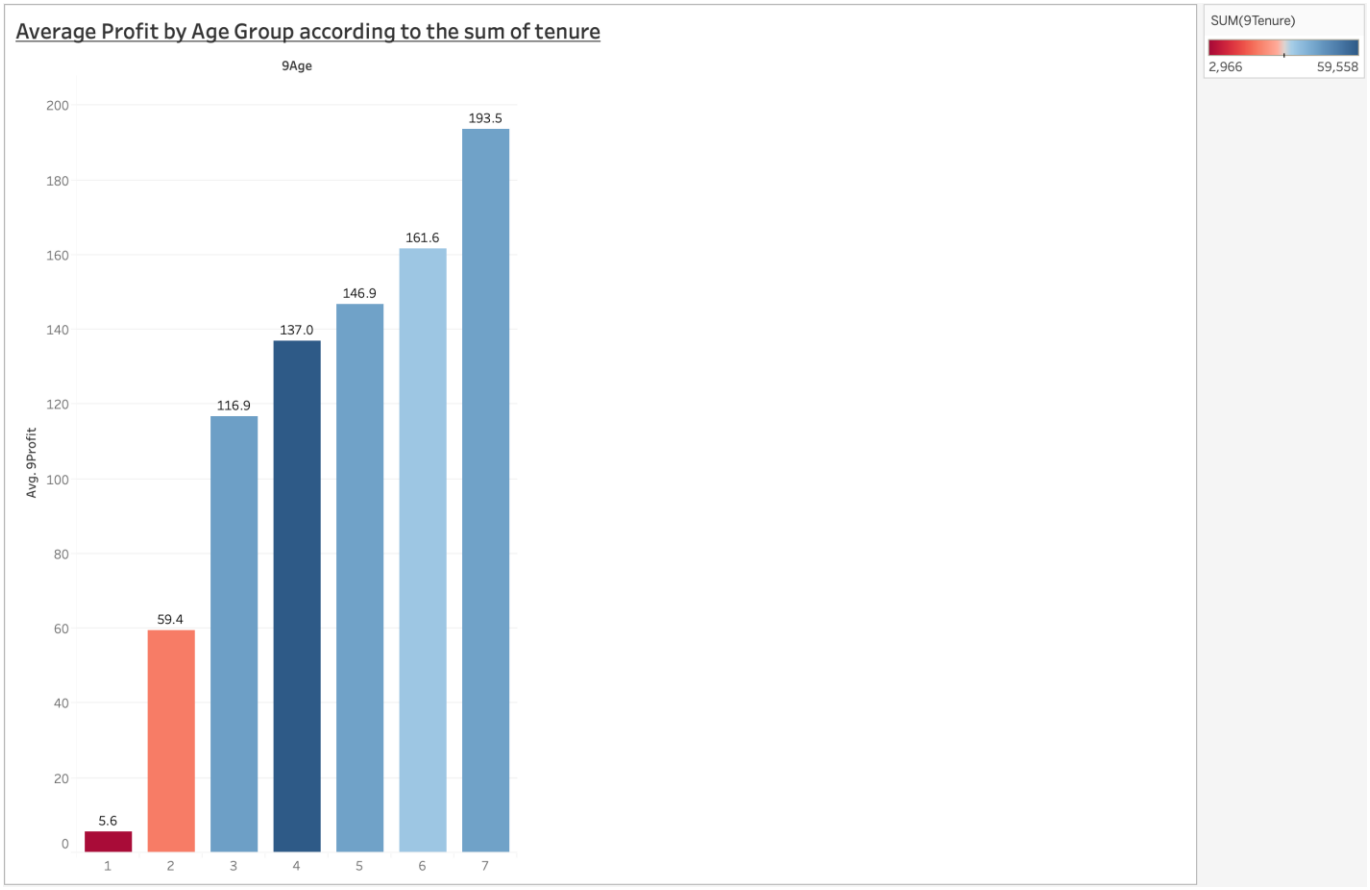


2. **What role do customer demographics play in analyzing customer profitability for online and offline customers?**

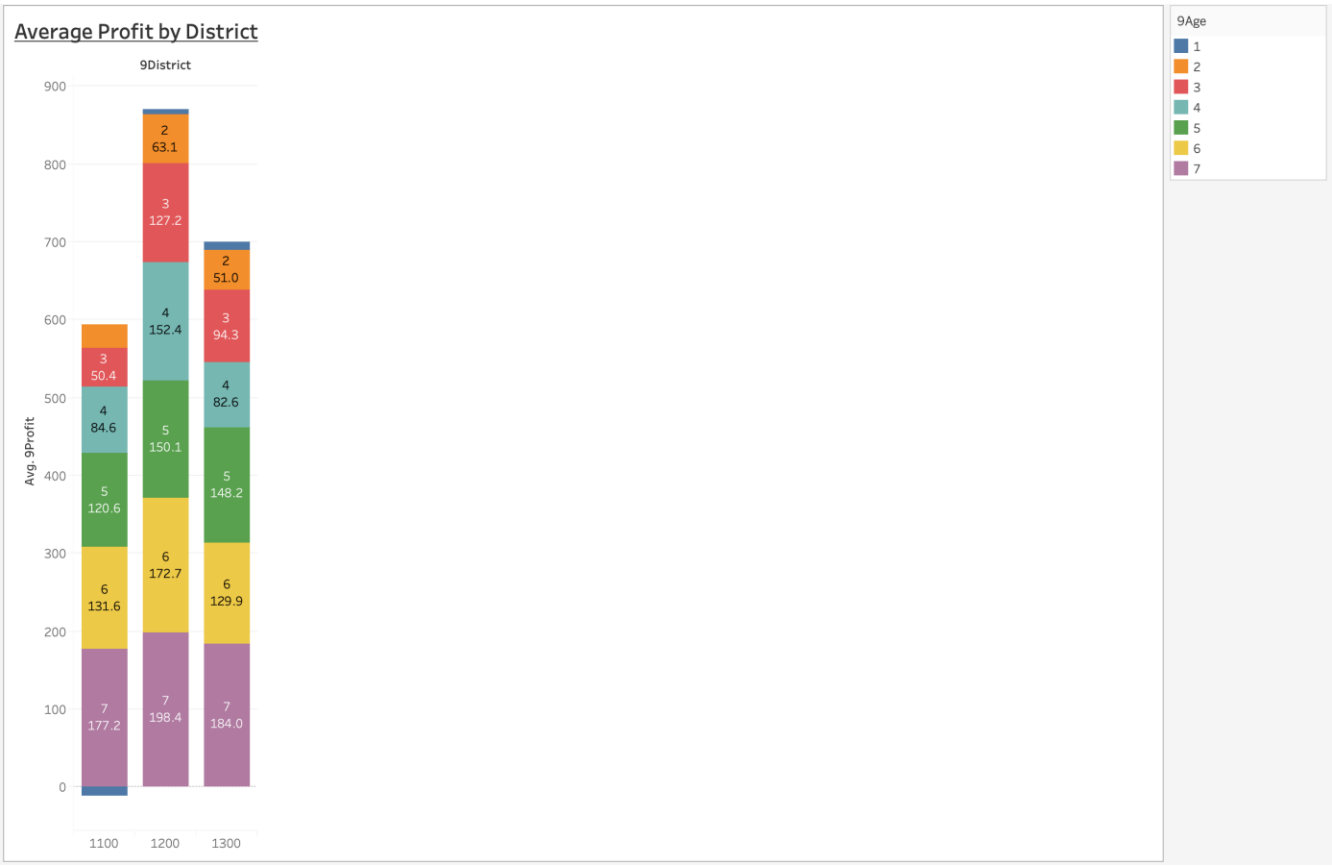
- **After observing the following graph, it is clearly observed that the total profit made by the Pilgrim bank is for the age group 4 = 35-44 years.**



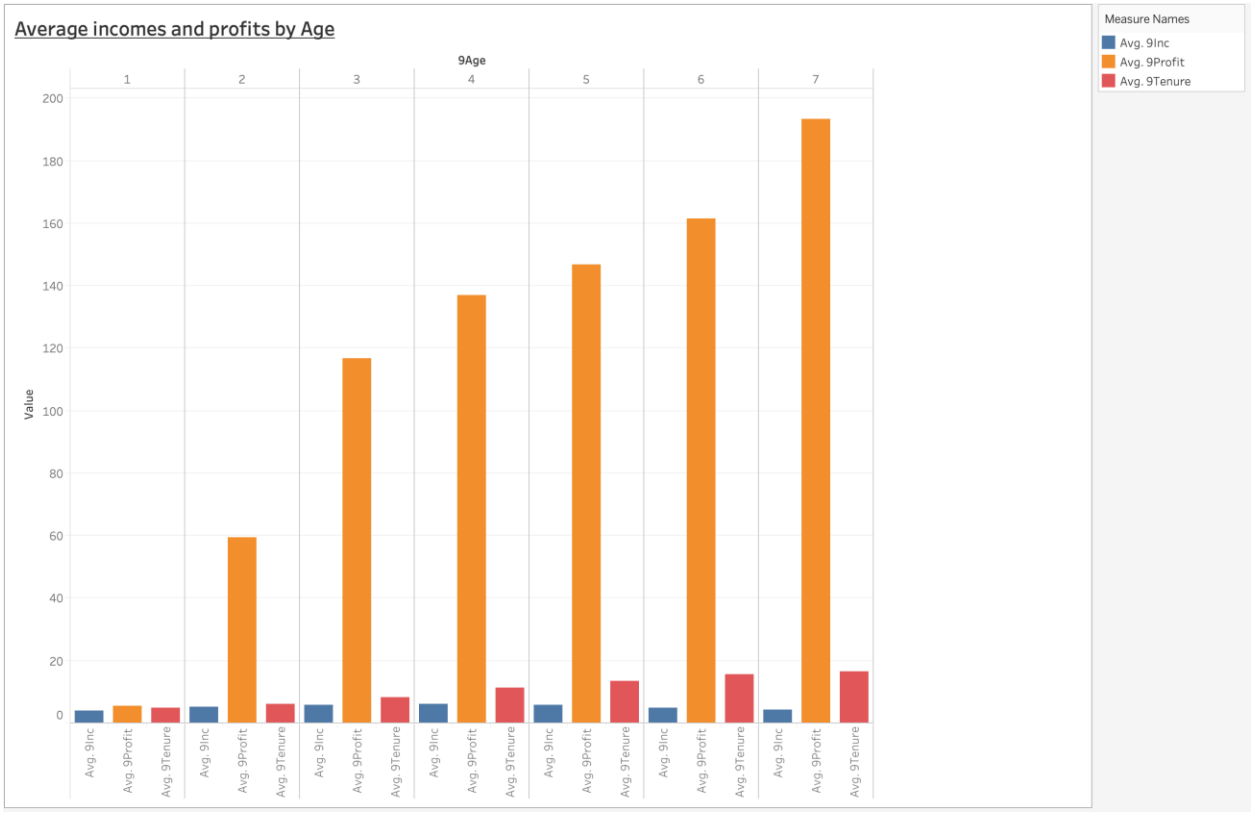
- **Furthermore, if we consider plot the average profits for each age group, we find that the age group 7 = 65 years and above has the largest profit average. To dig even deeper, if we find the average profits for each age group taking into consideration the sum of the tenures for all the clients, it is surprising that the sum of tenure is the highest for age group 4 = 35-44 years as 59,558 years but that for age group 7 = 65 years and above has lower tenure sum.**



- **Geographically, most of the customers prefer to go to branch that is present in district 1200. Furthermore, clinging on to the previous point that older people have a higher average profit value, the following graph also proves that the district number 1200 had the highest amount of older people as the district 1200 has the highest average profit value coming from people who are 65+ years old.**



- **Now, if we look at the average incomes and profit values by age, it is surprising to see that the highest average value is for people older than 65 years, with the second lowest income after young age people and highest average tenure of 16.5 years. Thus, it can be said that the average value of the profit increases with the tenure.**



- **After running a linear regression model on the data, it was found that demographics do have a significant impact on the profit numbers, though this model can be improved.**

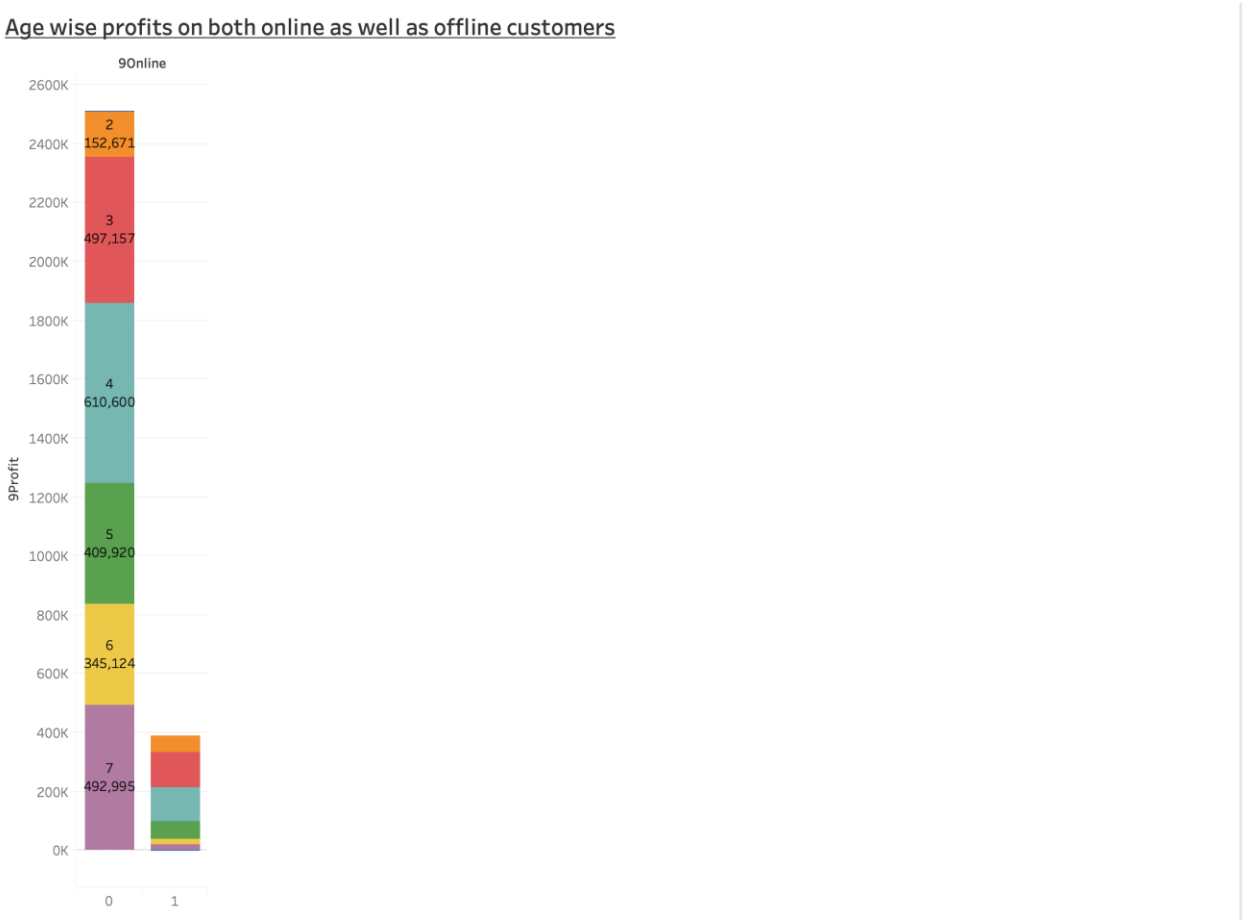
```
lm(formula = Profit ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-489.86 -164.24  -75.07   66.53 1944.71

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.018e+02  5.459e+01  -1.865  0.06223 .
...1         1.242e-04  2.322e-04   0.535  0.59284
ID                  NA         NA      NA       NA
`9Online`    2.110e+01  6.454e+00   3.269  0.00108 **
`9Age`       1.949e+01  1.448e+00  13.458  < 2e-16 ***
`9Inc`       1.899e+01  9.122e-01  20.818  < 2e-16 ***
`9Tenure`    4.305e+00  2.726e-01  15.794  < 2e-16 ***
`9District` -7.238e-04  4.470e-02  -0.016  0.98708
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 286.2 on 18348 degrees of freedom
Multiple R-squared:  0.05991,   Adjusted R-squared:  0.05961
F-statistic: 194.9 on 6 and 18348 DF,  p-value: < 2.2e-16
```

- **Clearly, age group 4 = 35-44 years have the highest amount of total profit numbers. Thus, in 1999 pilgrim bank had more number of customers from this age group.**



- **From the graph below, offline customers had the highest majority in which the highest number of people were from district 1200 in age group 35-44.**

3. **Is the difference in average profitability between online and offline customers in the sample indicative of a meaningful difference in profitability across these groups?**

T-Test

[DataSet3]

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
pro1	Online	2954	131.52	290.365	5.342
	Offline	19858	126.52	281.724	1.999

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
pro1	Equal variances assumed	8.985	.003	.897	22810	.370	5.003	5.578	-5.930	15.936
	Equal variances not assumed			.877	3826.784	.381	5.003	5.704	-6.181	16.186

- **As observed above, the Sig. value is 0.003 which is < 0.05. Hence here, we assume that we have unequal variances.**
- **Thus, due to unequal variances, we look at the Sig.(2-tailed) value which is 0.381.**
- **As a thumb of rule, the null hypotheses that there is no significant statistical difference between the mean profit values of online and offline was set.**
- **As the p-value obtained was > than 0.05, we cannot reject the null hypotheses and hence, it is true.**