Exercise 5.4

Intro to Data Mining

Stefan Rieß

2. To understand the data, you'll first need to assess the quality of the data, by checking for missing values, errors, and inconsistencies.
    - You'll also need to clean your data, using the techniques that you learned in previous Achievements. Fix any inconsistencies in the table and/or any errors, as far as it is possible.
    - Document your processes for assessing the data quality and cleaning the data, and note down any missing values or errors.

Column	Name	Issue	Action	Comment
A	Row_Number	No Issue found	No Action taken	
B	Customer_ID	No Issue found	No Action taken	
C	Last_Name	missing value L? H? Hs? Y? an	No Action taken	Last_name is not relevant for analysis and should in fact be removed for Data Protection PII
D	Credit Score	3 missing values found	Imputed Average	Average of 649 used for imputing the missing 3 values

E	Country	Columns containing shortcut and written name of Country	Changed DE to Germany ES to Spain And FR to France	Common issue and replaced with correct name so that the Colum is consistent again
F	Gender	Inconsistency between M for Male and F for Female 1 value Null	Changed M to Male and F to Female Ignored Value Null	One value Null wont impact our Analysis
G	Age	11 rows with Age 2 found	Replaced Age 2 with Average age of 39 as two years old probably wont have a credit card	Average might be wrong, but would impact the Analysis less than the obvious error of Age 2
M	Estimated Salary	2 rows found with Blank , Null values	Imputed with Average 98,574 USD	

No other Issues or Duplicates have been identified in the other Columns

3. Now that you've cleaned the data, you're ready to calculate some basic descriptive statistics to understand the data. Remember, your goal is to identify the risk factors that have contributed to customers leaving the bank.

    a. Separate the clients into 2 groups: one for those who have left the bank and a second for those who have stayed (hint: “1” in the “ExitedFromBank” column represents customers who have left).

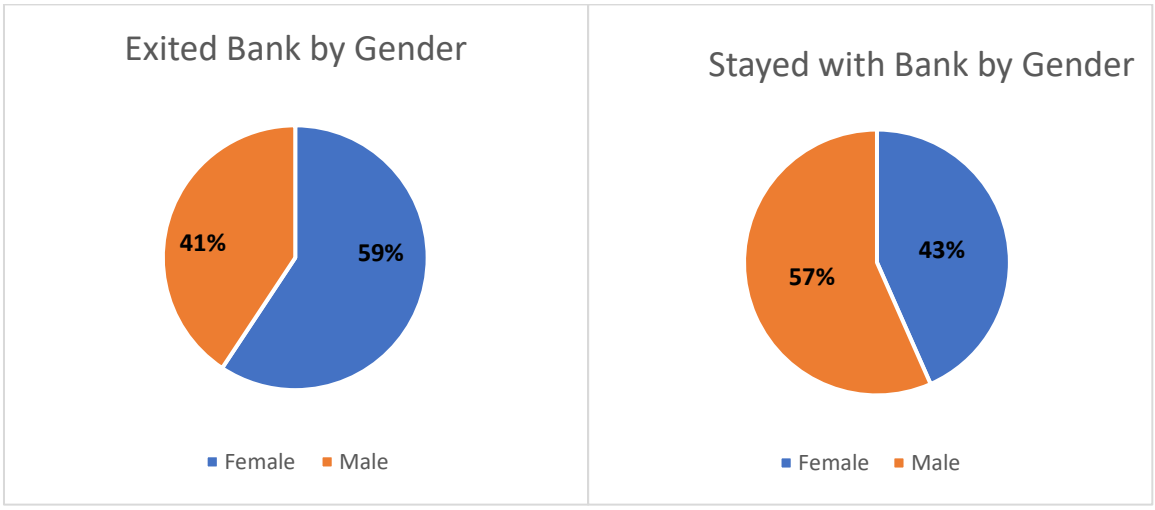
    b. Use pivot tables and other Excel functions to identify the top 3 to 4 factors that lead to clients leaving.

    c. Gather and analyze statistical information on both groups (e.g., find averages, means).

    d. Determine the leading factors that contribute to client loss, based on your analysis of the data provided.

    e. Document your results and how you reached them.

Comparison by Gender



Here we can see that the majority of customers leaving the bank are female with 59 Percent who left are Woman compared to Man with 41 Percent.

Comparison by Age Group

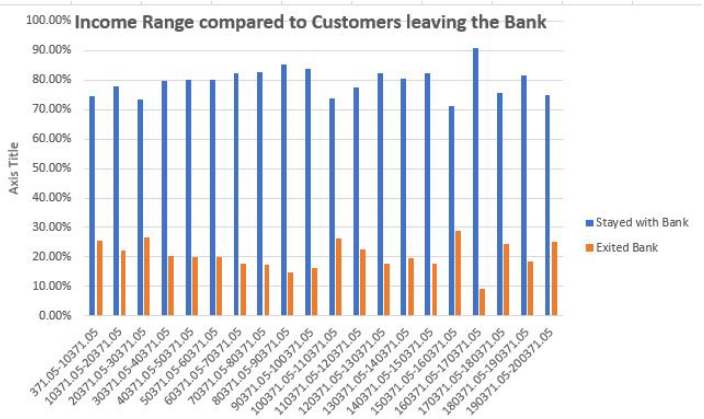
Count of ExitedFromBank?	Column Labels		
Row Labels	Stayed with Bank	Exited from Bank	Grand Total
18-27	96.88%	3.13%	100.00%
28-37	91.34%	8.66%	100.00%
38-47	72.73%	27.27%	100.00%
48-57	52.38%	47.62%	100.00%
58-67	48.08%	51.92%	100.00%
68-77	90.91%	9.09%	100.00%
78-87	100.00%	0.00%	100.00%
Grand Total	**79.49%**	**20.51%**	**100.00%**

In order to identify the most common reasons customers are leaving Pig.E Bank, i have compared the Age of their customers with the status of leaving or staying with the Bank.

I used a pivot Table where I created 10 Year Age Groups to display the Data. Here we can see that most of the customers who are leaving the Bank are between 48 to 67 of Age.

Comparison by Income Range

Count of ExitedFromBank?	Column Labels		
Row Labels	Stayed with Bank	Exited Bank	Grand Total
371.05-10371.05	74.58%	25.42%	100.00%
10371.05-20371.05	77.78%	22.22%	100.00%
20371.05-30371.05	73.33%	26.67%	100.00%
30371.05-40371.05	79.59%	20.41%	100.00%
40371.05-50371.05	80.00%	20.00%	100.00%
50371.05-60371.05	80.00%	20.00%	100.00%
60371.05-70371.05	82.22%	17.78%	100.00%
70371.05-80371.05	82.61%	17.39%	100.00%
80371.05-90371.05	85.11%	14.89%	100.00%
90371.05-100371.05	83.64%	16.36%	100.00%
100371.05-110371.05	73.81%	26.19%	100.00%
110371.05-120371.05	77.59%	22.41%	100.00%
120371.05-130371.05	82.14%	17.86%	100.00%
130371.05-140371.05	80.43%	19.57%	100.00%
140371.05-150371.05	82.14%	17.86%	100.00%
150371.05-160371.05	71.11%	28.89%	100.00%
160371.05-170371.05	90.70%	9.30%	100.00%
170371.05-180371.05	75.56%	24.44%	100.00%
180371.05-190371.05	81.40%	18.60%	100.00%
190371.05-200371.05	75.00%	25.00%	100.00%
Grand Total	**79.49%**	**20.51%**	**100.00%**

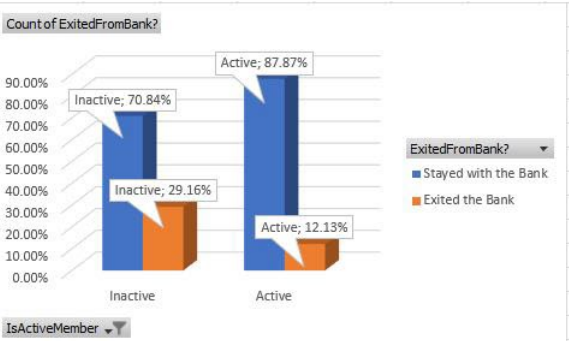


I have created another Comparison with the Data of the Income Range and Customers who stayed/ left the Bank. I used for the bracket of 10000 for the Income Range as it fits, especially the lower Income Range better than using a higher Bracket [such as 20 or 30k] .

However in this Data we can see that there is no connection between the Income and the customers who left the bank as the amount of people who leave, stay relatively stable throughout all income brackets.

Comparison by Activity Status

Count of ExitedFromBank?	Column Labels		
Row Labels	Stayed with the Bank	Exited the Bank	Grand Total
Inactive	70.84%	29.16%	100.00%
Active	87.87%	12.13%	100.00%
Grand Total	**79.49%**	**20.51%**	**100.00%**



I also compared the Activity Status of Pig.E's Customers in another Pivot Table.

There it was clearly visible that Customers who are considered inactive, are more likely to leave Pig.E Bank .

Comparison of descriptive Statistics of Customers who left and stayed within Pig E. Bank

Customers who exited Pig E. Bank

	Credit Score	Age	Tenure	Balance	Estimated Salary
Minimum	376	22	0	$0.00	$417.41
Maximum	850	69	10	$213,146.20	$199,725.39
Mean	637	45	5	$90,342.28	$97,051.39

Customers who stayed with Pig E. Bank

	Credit Score	Age	Tenure	Balance	Estimated Salary
Minimum	411	19	0	$0.00	$600.36
Maximum	850	79	10	$190,479.48	$199,638.56
Mean	648	37	5	$71,171.97	$104,230.22

As a summary and mentioned before we can see that there are several mainfactors that are playing a Role if and why customers are leaving Pig E. Bank. Mainly I would recommend to focus on Gender and Activity and Age.

4. Using the information you've uncovered so far, create a decision tree to determine the probability of customers leaving the bank.

    a. Pick which tool you'll use to create your decision tree. You can either create your own template using Excel or Powerpoint, for example, or download a decision-tree template.

    b. Determine which decision node will have the greatest impact and place it at top of the tree. For example, if you decide that an estimated salary below 15,000 USD is the biggest risk factor, then you would put this at the top and build your tree from there. Make sure that your decision tree includes the top 3 to 4 risk factors you identified in step 3.