Data Ethics & Applied Analytics

Task 5.4: Intro to Data Mining

One of the data analysts for the sales team recently left Pig E. Bank, so you've agreed to take their place and help out with a customer retention project.

To increase customer retention, the sales team wants to identify the leading indicators that a customer will leave the bank. You've created a table of client attributes that you believe could indicate whether customers will leave—for example, age, estimated salary, etc. You're going to use this information to identify the top risk factors that contribute to client loss and model them in a decision tree.

Create a new text document and call it "Answers 5.4"; then, based on the data mining techniques covered in this Exercise and the course, complete the steps below:

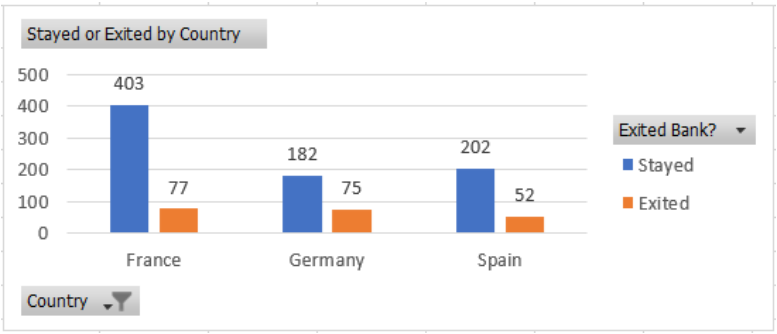
1. Download Pig E. Bank's client data set (.xlsx). Open the data set in Excel and take a moment to familiarize yourself with the data.
2. To understand the data, you'll first need to assess the quality of the data, by checking for missing values, errors, and inconsistencies.
    - You'll also need to 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
Row_Number	Column is redundant	Deleted Row_Number column	Row # not relevant to analysis
Customer_ID	No issues found	None	All #'s are unique
Last_Name	(6) font "?" & (1) Blank	None	Will not affect analysis
Credit Score	(3) Blanks	Imputed using Mean value	Mean= 649 Median= 654
Country	Format inconsistencies (FR, DE, ES)	Replaced FR = France DE = Germany ES = Spain	244=FR 23=DE 118=ES
Gender	Format inconsistencies (M, F) 1 NULL (no action)	Replaced F = Female M = Male	19=F 49=M
Age	(11) = 2 (1) = NULL	Imputed using Mean value	Mean= 39 Median= 37
Tenure	No issues found	None	Mean= 5 Median= 5
Balance	(349) = $0.00	None; Imputing would affect results inaccurately	Mean= $78,002.72 Median= $98,668.18
NumOfProducts	No issues found	None	
HasCrCard?	No issues found	None	
IsActiveMember	No issues found	None	
Estimated Salary	(1) = Blank (1) = NULL	Imputed using Mean value	Mean= $98,574.54 Median= $98,368.24
ExitedFromBank?	No issues found	None	

No duplicates were found.

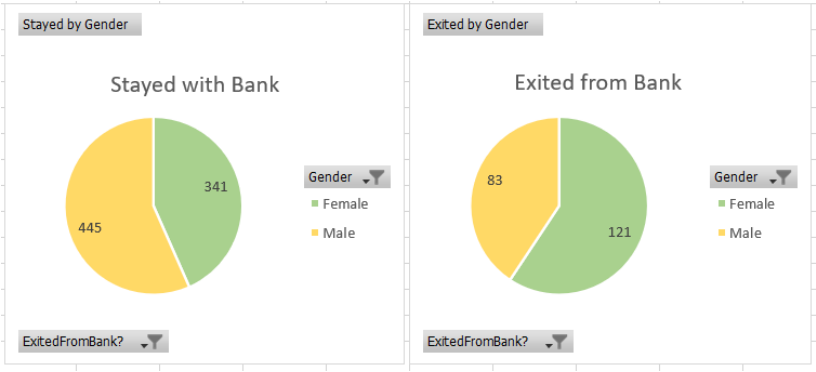
3. Now that you've cleaned the data, you're ready to calculate some basic descriptive statistics to understand the data. Remember, your goal is to identify the risk factors that have contributed to customers leaving the bank.
    - Separate the clients into 2 groups: one for those who have left the bank and a second for those who have stayed (hint: "1" in the "ExitedFromBank" column represents customers who have left).
    - Use pivot tables and other Excel functions to identify the top 3 to 4 factors that lead to clients leaving.

Country



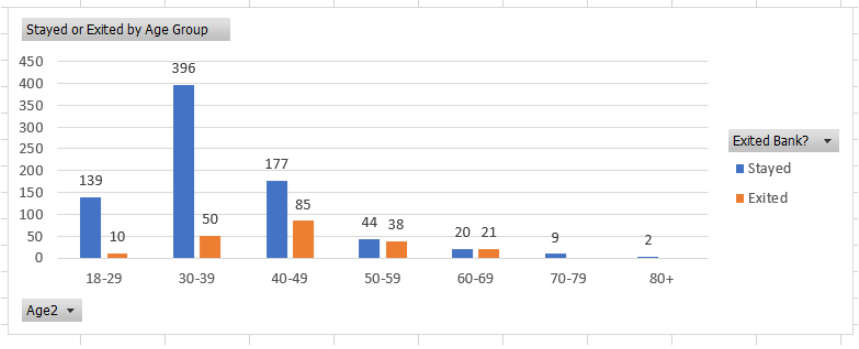
The highest ratio of Germany's total population has exited, followed by Spain and France, respectively.

Gender



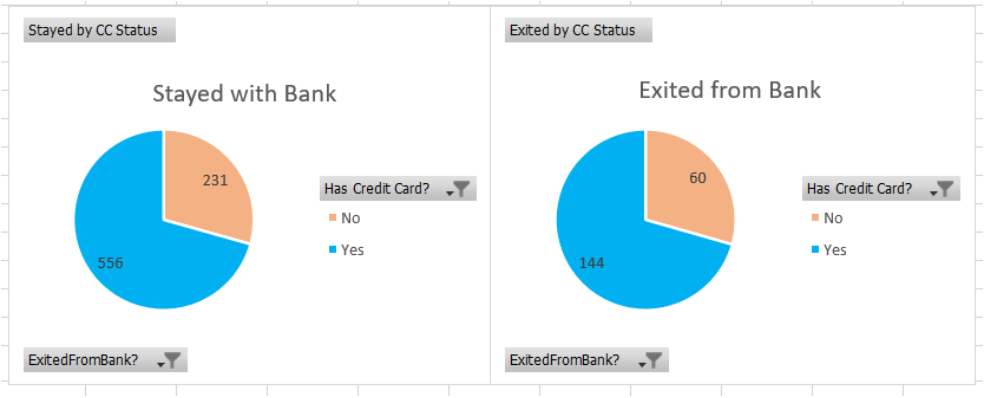
Clearly a higher percentage of exited customers are female, concurring with the higher male percentage among those who stayed.

Age



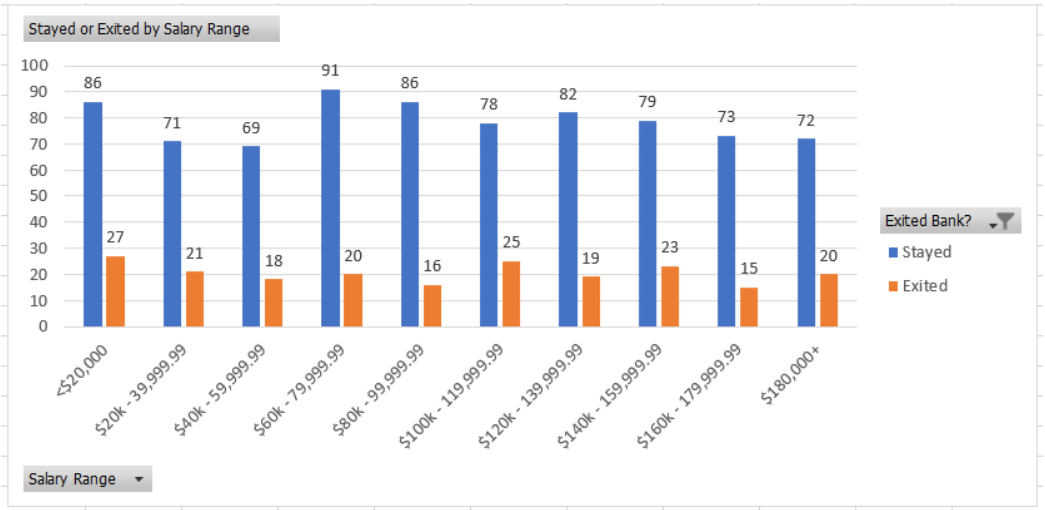
Close to half the total population of customers age 50-59 and 60-69 have exited. While the total of these age groups is notably smaller by comparison, the likelihood of exiting the bank clearly starts to increase within the 40-49 age group. This indicates a high exit risk.

Has CC



There appears to be no correlation with credit card status, among customers who stayed or exited, as the ratio of credit card holders is nearly identical among both customers who stayed or exited.

Est Salary



While most stayed to exited ratios are comparable among the various income brackets, applying the formula (value/total value)x100%, the highest exit percentages are as follows:

$100k - 24.2%; <$20k - 23.2%; $20k - 22.8%; $40k - 20.6%

We can generalize this data for the purpose of the analysis, stating the $100k bracket as the highest exit risk, followed by income less than $60k.

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

Credit Score

Exited		Stayed	
Mean	637	Mean	652
Median	644	Median	657
Max	850	Max	850
Min	376	Min	411

Slightly lower mean, median and min credit scores found with exited customers.

Age

Exited		Stayed	
Mean	45	Mean	38
Median	45	Median	36
Max	69	Max	82
Min	22	Min	18

Notably higher mean and median age indicated among those who exited.

Balance

Exited		Stayed	
Mean	$90,239.22	Mean	$74,830.87
Median	$112,433.97	Median	$93,147.00
Max	$213,146.20	Max	$197,041.80
Min	$0.00	Min	$0.00

Mean, median, and max balance are somewhat higher with exited customers.

Salary

Exited		Stayed	
Mean	$97,155.20	Mean	$98,942.45
Median	$100,375.40	Median	$98,368.24

Max	$199,725.39	Max	$199,661.50
Min	$417.41	Min	$371.05

Only slight salary variations indicated between those who stayed or exited.

- Determine the leading factors that contribute to client loss, based on your analysis of the data provided.
- Document your results and how you reached them.

As indicated by the chart findings in step 3 above, customers aged 40-69 show the overall highest likelihood of exiting, followed by customers based in Germany. All three countries are shown in the decision tree, by order of risk, in effect including the entire 40-69 population. These higher risk factors are further expanded by gender, and finally according to the two higher risk income brackets. These show higher exit by a narrow margin, although still significant.

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