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.
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.
    - Gather and analyze statistical information on both groups (e.g., find averages, means).
    - 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.
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.
5. Combine your decision tree and answers document into one PDF and upload it here for your tutor to review.

The four factors that mean customers are most likely to leave the bank are:

- **Higher Balance**
- **Older**
- **Female**
- **German**

- **Higher Balance**

	Stayed	Left
Mean	74645.78	90030.75
Median	93012.89	112045.7
Maximum	197041.8	213146.2

All these figures are significantly higher for this that have left

To put a figure on "Higher Balance" I looked at Quartiles of balances

	1st Quartile	2nd Quartile	3rd Quartile	4th Quartile
Stayed	84%	85%	74%	75%
Left	16%	15%	26%	25%

So, the customers start to leave more frequently in the upper two quartiles i.e. above $98469.955

- **Older**

	Stayed	**Left**
Mean	37.51097	45.28856
Median	36	45
Mode	36	39
Maximum	*82*	69

The mean, media and mode age of customers that have left is higher.

NB The very oldest customers have tended to stay, but these are small numbers

To put a number on ages, I looked at age brackets

	Under 30	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65 and over
Stayed	93%	92%	87%	76%	54%	50%	58%	36%	81%
Left	7%	8%	13%	24%	46%	50%	42%	64%	19%

So, it is those aged 45-64 that re most likely to leave

- **Female**

	Female	Male
Stayed	74%	84%
Left	26%	16%

Females are considerably more likely to have left

- **German**

	France	Germany	Spain
Stayed	84%	71%	79%
Left	16%	29%	21%

Germans are considerably more likely to have left

DECISOON TREE SAVED IN EXCEL