Data Analyst Take Home Exam

You were given files related to our users credit performance.   
We have no clue what is going on and we need your help to tell us a story that you find in the data.

**Requirements:**

1. You have 72h to return the assignment.
2. Use python for all Assignments and (when asked for) python-sql[[1]](#footnote-0) for data transformations (you can use pandas or any other package if the problem does not ask for python-sql)
3. For visualizations, please use python (common libraries), excel, or whatever visualization tool you are familiar with
4. For presentation please use Microsoft PowerPoint or Google Slides

**Deliverables:**

1. Python Jupyter notebook with your code. Add comments to the code !!!! Comments let us follow your thought process.
2. Core presentation no more than 15 slides (excludes the Appendix). This presentation is to the upper management so keep the technical jargon to a minimum (put technical stuff in the Appendix).
   1. 2 Slides about you (past projects, hobbies, favourite colour, etc.)
   2. 10-13 Slides about the assignments. What story do you see in the data? (do not include code in the slides, move it to Appendix)
   3. You are allowed to put as many slides into the Appendix as you want
3. Any additional documents that you would like to share with us

**What are we looking for:**

1. Tell the story with data (especially Assignment 3)
2. There is no right solution to this assignment (even SQL can be written in many different ways)
3. Be precise and to the point (don’t be long winded). Provide justifications to your recommendations
4. Use comments in the code !!!
5. Completeness of the analysis/story
6. The rationale behind the decisions you made

# Assignments

**Assignment 1:**

[SQL] Please use Python-sql

1. Combine all datasets into one
2. Provide summary statistics by:
   1. “Default\_flag”
   2. “Branch Code”
3. Select top 5 users with the most number of “transactions” from each “branch code”
4. Arrange (ascending) users by “outstanding” and provide cumulative default rates by equal size deciles. Each decile should have:
   1. Number of users
   2. Number of default users
   3. The default rate for each decile
   4. Cumulative default rate
   5. Min, max, avg “outstanding”

**Assignment 2:**

[Cleaning] Please examine features and tell us if there are any outliers or “anomalies”. If yes, then please fix them. Please provide the justification to why you marked given observations as anomalous (outliers) and which method you decided to use.

**Assignment 3:**

[Visualisations] Use visualizations (graphs and tables) to tell a story about the users who default. Do we see any significant relationships between variables? Any recommendations on who we should (or shouldn’t give the loans to)?

**Assignment 4:**

[Freestyle] Show us something cool with the data. We all have different skills, maybe you are good at modeling (build a predictive model). Maybe you are good at stats and know of a good way to look (examine) data. Maybe you can create great interactive charts. Whatever your special skill is, this is a place to show it (it does not have to be related to the credit defaulters, but please use this dataset).

1. https://pypi.org/project/python-sql/ [↑](#footnote-ref-0)