CSV Data ->
https://drive.google.com/drive/u/0/folders/1GWeC4LEKvO5WL5rLk19k8HtHEiSJxZMz

Complete as many of the following steps as you can. If you can't complete a step that's fine, move on to the next. The steps in bold are stretch goals for extra credit - feel free to skip them if you don't have the time. Reorder steps if it makes more sense to you to do so.

The output of the project should be the last step you can accomplish, as well as the code used throughout the process. Please leave brief comments in the code, and/or add a readme.md if you prefer.

[1] Load the attached CSV

[2] Perform as many of the following transformations as you can:

- Remove all rows where pdays is -1
- Split name into first name and second name columns (drop name)
- Replace the values in the age column with bucketed values, such that < 10 becomes 0, 10 <= x < 20 becomes 1, etc.
- Replace yes/no values with booleans
- Replace day and month with a single date column, of the form dd/MM
- Rename the y column "outcome"

[3] Add a column which categorizes geographical features in the address, where present. Note the dirtiness of the address data and that the exact categories :

  - "water", where the address contains e.g. lake, creek
  - "relief", where the address contains e.g. hill, canyon
  - "flat", where the address contains e.g. plain

[3] Group by the feature (if you created it, or by some other field if not) and filter out any empty values, sort by the age bucket (or age if you didn't do the bucketing), and return a row count.

[4] Write the row level data from step [2], and aggregated data from step[3] to both CSV and parquet formats.