

Instructions:

Complete this assignment on your private GitHub repo in a folder called `Assignment_08`. In this folder, save your answers to Questions 1 to 4 in a file called `my_A8_queries.py`, by completing the script in the file `my_A8_queries.py` in the course repository.. When you are finished, submit your files to your repository and upload the link to Webcourses.

1. The folder `Assignment_08` contains three `.csv` files: `applications.csv`, `credit_bureau.csv`, and `demographic.csv`. The first dataset `applications.csv` contains the following variables.
 1. `app_id` = a unique key for each customer who applied for credit
 2. `ssn` = the social security number
 3. `zip_code` = the the zip code in which the applicant resides
 4. `income` = the applicant's reported income
 5. `homeownership` = a categorical variable that indicates whether an applicantowns or rents a home
 6. `purchases` = the monthly value of purchases on the account
 7. `credit_limit` = the maximum amount that an applicant is approved to spend

You will use this dataset to estimate a regression model to predict the monthly amount of `purchases` for each customer.

- (a) Create a new database called `credit.db`.
 - (b) Read in the `applications.csv` dataset and store the contents in a data frame called `applications` in your workspace.
 - (c) Use the sample code in `my_A8_queries.py` to estimate a regression model to predict `purchases` as a function of the other variables in the dataset (ignoring the variables `app_id`, `ssn` and `zip_code`, which are keys for databases). Notice the value of the adjusted R-squared statistic.
 - (d) `CREATE` a `TABLE` called `Applications` with a schema that is appropriate for the variables.
 - (e) Populate the table `Applications` with the observations in the data frame `applications`.
2. Now use two files `applications.csv` and `credit_bureau.csv` in the folder `Assignment_08`. The dataset `credit_bureau.csv` contains the following variables.
 1. `ssn` = the consumers unique social security number
 2. `zip_code` = the zip code in which the consumer resides
 3. `fico` = the consumer's credit score
 4. `num_late` = the number of number of times a consumer has made a payment after the due date
 5. `past_def` = the number of number of times a consumer has defaulted on a line of credit
 6. `num_bankruptcy` = the number of number of times a consumer has filed for bankruptcy

You will use the variables from both datasets to estimate a better regression model to predict monthly purchase volume.

- (a) Read the new dataset and store it in a data frame called `credit_bureau` in your workspace.
 - (b) **CREATE** a **TABLE** called `CreditBureau` with a schema that is appropriate for the variables.
 - (c) Populate the table `CreditBureau` with the observations in the data frame `credit_bureau`.
 - (d) Join the two tables by `ssn` and `zip_code` and output the result as a `pandas` data frame called `app_bureau`.
 - (e) Use the sample code in `my_A8_queries.py` to estimate a regression model to predict purchases as a function of the other variables in the dataset. (Again, ignore the variables `app_id`, `ssn` and `zip_code`, which are keys for databases.)
3. Now use all three files `applications.csv`, `credit bureau.csv`, and `demographic.csv` in the folder `Assignment_08`. The dataset `demographic.csv` contains the following variables.
1. `zip_code` = the zip code to indicate each geographic region
 2. `avg_income` = the average income in each zip code
 3. `density` = the population density in each zip code

You will use the variables from all three datasets to estimate an even better regression model to predict monthly purchase volume.

- (a) Read the new dataset and store it in a data frame called `demographic` in your workspace.
- (b) **CREATE** a **TABLE** called `Demographic` with a schema that is appropriate for the variables.
- (c) Populate the table `Demographic` with the observations in the data frame `demographic`.
- (d) Join the new table `Demographic` to the information from the other two tables by `zip_code`. You can use your query from Question 2 as a nested query. Output the result as a `pandas` data frame called `purchase_full`.
- (e) Use the sample code in `my_A8_queries.py` to estimate a regression model to predict purchases as a function of the other variables in the dataset. As above, ignore the variables `app_id`, `ssn` and `zip_code`, which are keys for databases.