

QMB 6358: Software Tools for Business Analytics
Executive Development Center
College of Business
University of Central Florida
Fall 2021

Assignment 6

Due Tuesday, November 2, 2021 at 11:59 PM
in your GitHub repo.

Instructions:

Complete this assignment within the space on your GitHub repo in a folder called `assignment_06`. In this folder, save a copy of the files called `A6Q1_data.py` and `A6Q2_data.py` that will contain all your Python code for Questions 1 and 2 in this assignment. Use the sample scripts in the `assignment_06` folder as a starting point.

When you are finished, submit your code by pushing your changes to your GitHub repo, following the instructions in Question 3. You are free to discuss your approach to each question with your classmates but you must `git push` your own work.

Question 1:

The folder `housing_data` contains 5 `.csv` files. Your job is to collect these files to form one dataset and fit a regression model. Use the file `A6Q1_data.py` as a starting point. Complete it in stages by following these steps:

- a) Write a loop that prints out the names of all the files. Use this to verify that the file names all match the names of the actual files.
- b) Extend the loop to read in each file, assign it to a data frame called `housing_single`.
- c) Extend the loop one more time by binding the files into a full dataset called `housing_full`.
- d) Verify that the statistics at the bottom of `A6Q1_data.py` indicate that the dataset has been read in correctly, so that the regression output is printed out.

Question 2:

Your research assignment is to collect and analyze a dataset that you extract from a document of sales records. Your supervisor would like to generate a model for the value of aggregate house prices by zip code as a function of several variables. Specifically, the regression model is

$$HOUSE_PRICE_i = \beta_0 + \beta_1 \times INCOME_i + \beta_2 \times IN_CALI_i + \beta_3 \times EARTHQUAKE_i + \epsilon \quad (1)$$

where:

- $HOUSE_PRICE_i$ = the average house price (in millions) in zip code i
- $INCOME_i$ = the average income (in millions) in zip code i
- IN_CALI_i = whether or not zip code i is in California
(i.e., $IN_CALI_i = 1$ if zip code i is in California, zero otherwise)
- $EARTHQUAKE_i$ = whether or not zip code i experienced an earthquake (similarly, 1 or 0)

Fortunately, you have found a source for a suitable dataset on the website of a real estate agent at https://github.com/LeeMorinUCF/QMB6358F21/tree/master/assignment_06/housing_records. The problem is that the agent's data collection efforts spanned several weeks and the data are organized into a separate page for each day of each week. The data are publicly available to clients such as yourself but you have to collect and compile the data from a `.pdf` document on the Website. Fortunately, the supporting `.tex` files are also available on the same Website. You can loop through the file names and obtain the numbers from within the `.tex` files. Your script should be automated, so it is ready to update the estimation next week, once the realtor uploads the newly obtained observations.

Create a script that reaches into the files and obtains the observations for each day within each week. Compile these observations into a single dataset and save it in your workspace. Create the script using the following stages as a guide.

- a) Write a function `is_data_row` that determines whether a particular line represents a line of data in the tables in the files such as `housing_sales_week_1_day_1.tex`.
- b) Write a function `get_obs_row` that takes in a string with the content of any of the lines like lines 9–28 in the file `housing_sales_week_1_day_1.tex` and returns a list of the numbers between the strings “&” and “\\”.
- c) Write a function `get_obs_day` that opens a file such as `housing_sales_week_1_day_1.tex`, then loops through a block of twenty consecutive lines of strings such as the ones in lines 9–28 and collects these values in a `pandas` data frame. This function should make repeated calls to `is_data_row` and `get_obs_row` and return the values in the data frame `housing_day`.
- d) Write a loop through the files for week 1 taking in the observations for each day of the week, making repeated calls to `get_obs_day`. For each day, append the values in the data frame for each day into a data frame `housing_week`.
- e) Wrap the code above within one more loop, just as was done in Question 1, that creates the data frame `housing_full` by appending data from each week.
- f) Verify that the statistics at the bottom of `A6Q2_data.py` indicate that the dataset has been read in correctly, so that the regression output is printed out.

Question 3:

Push your completed files to your GitHub repository following these steps. See the `README.md` and the `GitHub_Quick_Reference.md` in the folder `demo_04_version_control` in the QMB6358F21 course repository for more instructions.

1. Open GitBash and navigate to the folder inside your local copy of your git repo containing your assignments. Any easy way to do this is to right-click and open GitBash within the folder in Explorer. A better way is to navigate with UNIX commands.
2. Enter `git add .` to stage all of your files to commit to your repo. You can enter `git add my_filename.ext` to add files one at a time, such as `my_filename.ext`. in this example.
3. Enter `git commit -m "Describe your changes here"`, with an appropriate description, to commit the changes. This packages all the added changes into a single unit and stages them to push to your online repo.
4. Enter `git push origin main` to push the changes to the online repository. After this step, the changes should be visible on a browser, after refreshing the page.