Scientific Programming – Spring 2022 – Final Assignment Guide

Overview: The final assignment is an opportunity to practice applying skills from throughout the semester to the exploratory analysis of a dataset. Your task is to choose a dataset of interest (details below) and follow the general guidance to produce some analyses and figures.

You will be looking at the data, thinking about what the columns are, and deciding on some exploratory analyses to help understand the data. You will be generating some figures and doing some statistics to answer some questions that you choose.

Format: Your submitted assignment should be a notebook that carries out the tasks described below. Your assignment should be written so that all we have to do is change the directory where the files are located and then we can replicate your analysis. This means that your code should be written to run sequentially from the top of your notebook to the bottom.

Where indicated (see below), you should also include markdown cells that describe your analysis or answer specific questions.

A note about APA format: you are not required to follow strict APA format for any stats reporting or interpretation that you do. You are of course welcome to follow APA format, but overall your goal should be to focus on conveying information in a clear fashion that highlights what you were trying to do and what you found. Do not worry too much about whether you need two decimals versus three or anything like that. If in doubt, get in touch with Shannon or Marishka.

Due date: Your final assignment is due on Friday May 6th. You should submit your assignment by email to *nyu.sci.programming@gmail.com*. Your submitted notebook should have a title with the format:

lastname_firstname_netID_final.ipynb

Datasets

You will be provided with some datasets and their descriptions (see below). You will choose the dataset you wish to work on.

The data files (csv files and accompanying documentation in a readme.txt file) are located on Brightspace under Content / data / final_project_data /

Inside of that directory is a folder for each of the datasets containing the individual csv files and the readme.

IBM attrition data:

This is a dataset we have already seen this semester in HW3. It includes data on a number of variables for employees working at IBM and whether they stayed with or left the company. A description of the variables (columns in the data) can be found here:

https://inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2018FBL/IBM_Attrition_VSS.html

For the final assignment the IBM data can be found in the IBM_data folder with contents:

IBM-Attrition-Human Resources.csv
IBM-Attrition-Research & Development.csv
IBM-Attrition-Sales.csv

These data are on Brightspace in Content / final_project_data / ibm_data

OMNI data:

The OMNI dataset has data from individual people participating in a memory experiment. Each person studied 45 pairs of words and gave a rating of how well they thought they learned the pairs. This is called a "judgment of learning", or JOL.

After a delay, each person engaged in a retrieval test where they were presented with a Lithuanian word and attempted to recall the associated English word. Retrieval trials were scored as correct if the person recalled the English word and incorrect otherwise.

Each person was assigned to one of four groups, and the group assignment determined the approximate study-to-test delay (i.e., how long they waited between studying the words and trying to recall).

In addition, some of the participants did the learning task while in an MRI scanner and some did not.

The OMNI data is located in the omni_data folder and has the following files (one for each of the study-to-test delay groups).

omni_data_group-H.csv
omni_data_group-A.csv
omni_data_group-C.csv

omni_data_group-F.csv

In addition, that folder has a readme.txt file in it that gives more information on the data collection and the columns in the csv files.

These data are on Brightspace in Content / final_project_data / omni_data

Gender Differences in Moral Foundations dataset:

This dataset includes data from a survey of respondents in different countries. The survey was the Moral Foundations Questionnaire. The result of this Questionnaire is a score for each person on five core "moral foundations" that are held to underlie Moral Foundations Theory. This theory posits that moral intuitions derive from innate psychological mechanisms that coevolved with cultural institutions.

These foundations concern dislike for the suffering of others (Care), proportional versus egalitarian fairness (Fairness), ingroup loyalty (Loyalty), deference to authority and tradition (Authority), and concerns with physical/spiritual purity and contamination (Purity). Care and Fairness are entailed in individuals' well-being and have been referred to as the 'individualizing' foundations. On the other hand, Loyalty, Authority, and Purity encompass concerns about community, social order, and the maintenance of group bonds and, therefore, have been referred to as the 'binding' foundations.

The dataset has data from individual people and gives a score on each of the five foundations as well as gender and country of each person. A fuller description of the dataset can be found in the readme.txt included in the MFQ data folder.

MFQ_Hungary.csv
MFQ_Korea.csv
MFQ_Netherlands.csv
MFQ_UK.csv

These data are on Brightspace in Content / final_project_data / mfq_data

Your own data:

If you have a dataset of your own that you would like to explore that is OK. This could be a dataset you are aware of that you are interested in or data that you are working on for a lab, thesis, or other project. The only restrictions are that your data has at least two categorical variables and at least two interval or ratio variables.

IF YOU ARE CHOOSING OR CONSIDERING THIS OPTION: please get in touch with me soon to verify that your data are suitable for the assignment

YOUR TASK:

1. Choose a dataset to explore.
2. Download the data files and put them into a directory somewhere you know how to access
3. **Load and concatenate the csv files into a dataframe:**
    - Each dataset has a set of csv files containing a subset of the data
    - Import the `os` library and use the `listdir()` function to list get a list of all of the files in your data directory. This screenshot shows how to use it

```
In [1]: # import the os libary
        import os

In [4]: # use os.listdir() to get the contents of a directory
        # the input to listdir() is path (absolute or relative)
        # to a directory

        # relative path
        my_dir = '../../data/'

        files = os.listdir(my_dir)
        print(files)

['vibes.wav', 'chico.csv', 'WHR2018.csv', 'zeppo.csv', '.DS_Store', 'mlg-
airhorn.mp3', 'Programming SU21 intro survey.csv', 'nyu_admission_fake.cs
v', 'revised_salary.csv', 'demo_and_data.csv', 'student', 'meditation_wit
hin_person.csv', 'IBM-Attrition.csv', 'mental_rotation.csv', 'student.zi
p', 'sleep_mood.csv', 'salary2.csv', 'ibm', 'salary.csv', 'meditation.cs
v', 'pandas_data_1.zip', 'student-mat.csv', 'harpo.csv', 'covid_data', 's
tats_data.zip', 'awesome.csv', 'pyimages.zip', 'stats_data', 'parenthood.
csv']

In [6]: # absolute path
        my_dir = '/Users/shannon/science/courses/SciProgramming_SP22/data/'


        files = os.listdir(my_dir)
        print(files)

['vibes.wav', 'chico.csv', 'WHR2018.csv', 'zeppo.csv', '.DS_Store', 'mlg-
airhorn.mp3', 'Programming SU21 intro survey.csv', 'nyu_admission_fake.cs
```

    - Once you have a list of the csv files for your chosen dataset, **write a for loop that iterates through the files in your list and loads them into a dataframe one at a time and then concatenate that list into a single dataframe with all the data**
        i. Hint 1: when you list the file contents there will be a readme.txt file in addition to the csv data files. In your for loop you should check that the current file in your list is a csv file. You can do this using string checks by either seeing if the filename has an expected string in it or check to see if the filename does *not* have some undesired text in it.

ii. Hint 2: we have done individual file loading and concatenating in the pandas_part2 notebook

4. **Get rid of missing data**
    - Each row of your data corresponds to a single record and columns are measurements for that row. If there is missing data for some row you should drop that record from your analysis
    - You have seen a dropna() function in the pandas notebooks that will get rid of any row in the dataframe that is missing data for some column
    - **Use dropna() to clean up your dataset**
        i. **How many rows were in your data before dropna? How many after?**
        ii. **Include a markdown cell that reports how many records were dropped from your analysis due to missing data. If your dataset is the same size before and after it simply means you had no missing data.**

5. **Generate descriptive statistics and save dataframe to csv**
    - Choose one categorical variable and one numeric variable from your data
    - Use the groupby() function to get descriptive statistics for your numeric variable column broken down by the groupings in your categorical variable
    - The result of the previous steps will be a dataframe. Save this dataframe to a csv file.

6. **Explore the relationship between two interval or ratio variables**
    - Choose two numeric interval or ratio variables that you think might have a relationship to each other
    - **Make a scatterplot showing the values of these two variables in relation to each other**
    - **Compute pearson's correlation between these two variables**
    - **Make a markdown cell that describes why you hypothesized that these two variables might have a relationship and whether the Pearson's correlation reveals a significant positive relationship, negative relationship, or non-significant result**

7. **Run a t-test (at least one)**
    - You should run at least one of the following t-tests (or more if you think it will be interesting):
        i. One-sample t-test: are the values in some numeric column likely to have been drawn from a population with some particular average value?
        ii. Two-sample t-test: are the values in some numeric column different, on average, when we compare one group to another?
    - **Make a markdown cell that briefly says what your analysis choice was and whether there was a significant difference**
    - **Make a bar graph that shows the average value for your numeric variable (broken down by group if you are doing a two-sample test)**

8. **Carry out a regression or ANOVA analysis**
    - At minimum you should execute either an ANOVA or a regression analysis on your data
    - If you would like to carry out both kinds of analyses in your dataset you are free to do so
    - **REGRESSION**
        i. Conduct a single *and* multiple variable regression to see whether an outcome variables values can be accounted for based on the value of one or more predictor variables
        ii. For the single regression, choose an outcome variable (interval or ratio) and a predictor variable (interval or ratio) and run a regression
        iii. Use seaborn to make a regplot() showing your two chosen variables in a scatterplot along with the best fit line
        iv. Take a look at the coefficient for your predictor: is it different from zero? How do you know?
        v. For the multiple regression, add at least one additional predictor variable to your single variable regression
        vi. Using the results of either the single or multiple regression (i.e., the coefficients and the intercept) report what the expected value of the outcome variable would be for some arbitrary value of the predictor variable(s). In other words, use your regression results to make a prediction
        vii. Use a markdown cell to report: your motivation for assessing the relationship between these particular variables; whether the coefficient in your single variable regression was significant; whether the additional coefficients in your multiple regression was significant; how the R^2 value changes (or not) between the single and multiple regression
    - **ANOVA**
        i. In this analysis you are asking whether the average level of some outcome variable differs based on different groupings created by one or more categorical grouping variables
        ii. Conduct a one-way ANOVA (single grouping variable with > 2 levels and a single numeric interval or ratio outcome variable) or two-way ANOVA (two grouping variables with at least 2 levels each and a single numeric interval or ratio outcome variable)
        iii. Make a plot that shows the value of your chosen outcome variable within each of the groupings from your independent grouping variable(s)
        iv. If your ANOVA analysis is significant, conduct follow-up tests to try to understand what is driving the effect
        v. Include a markdown cell that describes: why you hypothesized that the grouping variable(s) you choose might have an impact on the

outcome variable; whether the ANOVA was significant; which groups were different from each other

Quick overview of tasks and expected outputs (details above):

1. Use OS tools to get list of files
2. Loop over list of files loading into dataframes and then concatenate into a single dataframe with all of the data in it
3. Use dropna() to clean up the data
    a. Report how many rows were dropped from the data
4. Descriptive statistics for one column after grouping by another
    a. groupby() and descriptive stats
    b. Save result to csv file
5. Scatterplot and pearson's correlation between two numeric variables
    a. Markdown with description of analysis and results
    b. scatterplot
6. T-test and bar plot
    a. Markdown with description of analysis and results
    b. Bar plot
7. Choose one (or do both): Regression or ANOVA
    a. Regression: carry out single and multiple variable regression
        i. Regplot()
        ii. Report coefficients and R^2 values
        iii. Make a prediction using coefficients and intercept
    b. ANOVA: carry out one-way or two-way ANOVA
        i. Report whether ANOVA was significant
        ii. Do follow-up tests as appropriate
        iii. Make plot showing values of outcome variable based on groupings of the independent variable(s)

Notes on different kinds of measurement scales (ordinal, interval, etc):
https://statsthinking21.github.io/statsthinking21-core-site/working-with-data.html#appendix