San José State University
Department of Applied Data Science

DATA 220
Mathematical Methods for Data Analysis

Spring 2021
Instructor: Ron Mak

Assignment #1

Assigned: Thursday, January 28
Due: Thursday, February 4 at 5:30 pm
100 points max

CSV datasets and Jupyter notebooks

The purpose of this assignment is to give you and your lab partner practice finding and downloading interesting CSV (comma-separated values) datasets on the internet, and then creating Jupyter notebooks to work with the data.

Find three CSV (comma-separated values) datasets on the Internet that are interesting to you. For each dataset, create a Jupyter notebook that does the following:

- Load the dataset into a Pandas dataframe. Your Python code can access the dataset directly via a URL, or you can first download the data to a file on your local machine and then your code loads the file into a dataframe.
- Display the data as a table, up to 60 rows. The table should have suitable column headings.
- Display some simple statistics of the data.
- Create one or more histograms of the numerical values in the dataset.

You may use any Python code from the sample notebooks that were demonstrated in class:

- Passengers on the Titanic:
  http://www.cs.sjsu.edu/~mak/DATA220/notebooks/0123/TitanicCSV.ipynb
- Airline safety:
  http://www.cs.sjsu.edu/~mak/DATA220/notebooks/0123/AirlineSafetyCSV.ipynb
- Crimes in Boston:
  http://www.cs.sjsu.edu/~mak/DATA220/notebooks/0123/BostonCrimeCSV.ipynb

Running Jupyter Lab

Install Anaconda, which includes Python and Jupyter Lab. The download page is at https://www.anaconda.com/distribution/#download-section Select the download that is appropriate for your platform (Windows, macOS, or Linux). The installation may take a while.

In a terminal window, change to the directory (`cd` command) where you've stored your Jupyter notebooks. If you're on Windows, you must run the command window called "Anaconda Prompt" which you access from the start menu. On the command line, execute the command

```
jupyter lab
```

Jupyter Lab will open in your default browser. Your notebooks will appear in the left panel. Double-click a notebook to open it.

What to submit to Canvas

Submit your notebook files into Canvas: **Assignment #1**
You can submit the notebooks individually (they are text files) or you can first zip them together.

Rubric

Your notebooks will be graded according to these criteria:

Criteria	Max points
Successful Anaconda installation	10
Three CSV datasets	30
Three Jupyter notebooks: • Load the dataset into a Pandas dataframe. • Display the data as a table (no more than 60 rows) with suitable column headings. • Display some simple statistics of the data. • Create one or more histograms.	60 • 15 • 15 • 15 • 15