CS 696 Intro to Big Data
Fall Semester, 2020
Assignment 1

San Diego State University -- This page last updated 1/28/20

Due Feb 9 23:59

Weather Data

The file US_WeatherEvents_2016-2019 contains over 5,000,000 rows and 13 columns. The column labels are:

EventId - This is the identifier of a record.

Type - The type of an event; examples are rain, fog. snow storm, cold.

Severity - The severity of an event, wherever applicable.

StartTime(UTC) - The start time of an event in UTC time zone.

EndTime(UTC -)The end time of an event in UTC time zone.

TimeZone - The US-based timezone based on the location of an event (eastern, central, mountain, and pacific).

AirportCode - The airport station that a weather event is reported from.

LocationLat - The latitude in GPS coordinate.

LocationLng - The longitude in GPS coordinate.

City - The city in address record.

County - The county in address record.

State - The state in address record.

ZipCode - The zipcode in address record.

Answer the following questions about the data set.

1. How many snow events were recorded?
2. How many moderate snow events were recorded each year? Produce a table and line plot
3. How many snow events were recorded in each state?
4. How many snow events were recorded each hour?
5. How many events of each severity did San Diego have? Produce a table and bar plot.

Trump Tweets

The file trumpttweets.csv contains all President Trump's tweets up to Jan 20, 2020. The file has 9 columns with headers given below

id - tweet id

link - link to tweet

content - tweet text

date - date of tweet

retweets - number of retweets

favorites - number of favorites

mentions - mentions (if any)

hashtags - hashtags (if any)

geo - location of tweet (if specified)

Answer the following questions about the dataset.

1. How many tweets did the President make each year? Produce a table of results and plot the values.

2. How many tweets contain "Fake News"?

3. A tweet with the location could be considered a security issue as it would give the current location of the President. How many tweets contain a location?

4. How many tweets were done each hour? Produce a table and a plot.

Instructions

You are free to use any IDE to write your code.. However you are to turn in a Jupyter Python notebook. Your jupyter notebook should be self contained. All calculations and answers to the questions are to be in one notebook. This assignment requires you to use files, some of which are provided. You notebook needs to read the unmodified files, including names. Any needed modification to the files needs to be done in the notebook.

At the beginning of your notebook you should create variable that hold the path (plus name) of any input files that you use. It is likely that for grading purposes those paths will need to change. I should be able to run your notebook using my input files by just changing the path to files at the top of your notebook.

Notebooks can contain text, code and output. Use text to indicate what problem you are solving. The code used to answer the problem need to be complete.

Grading

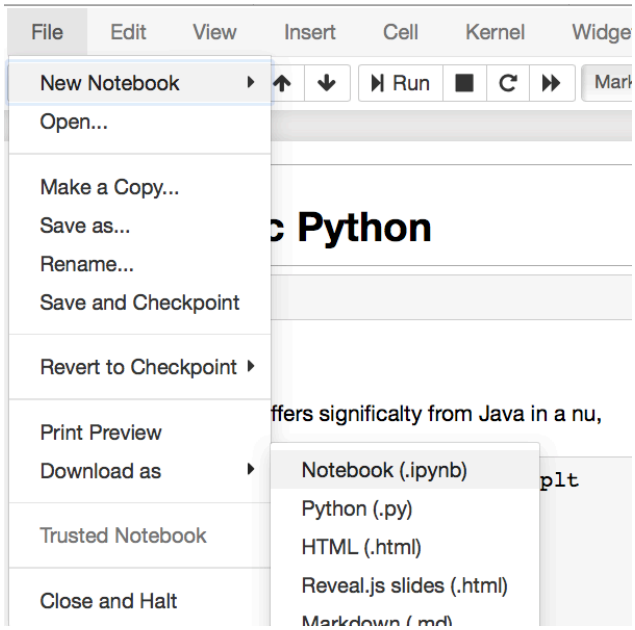
Each problem is worth 10 points.

What to turn in

To turn in your assignment download your Jupyter notebook as an IPython Notebook (.ipynb). See image below. This will allow me to run your assignment in Jupyter. Note that when you download your assignment it will create a file with the extension .ipynb.json. I will remove the .json extension.

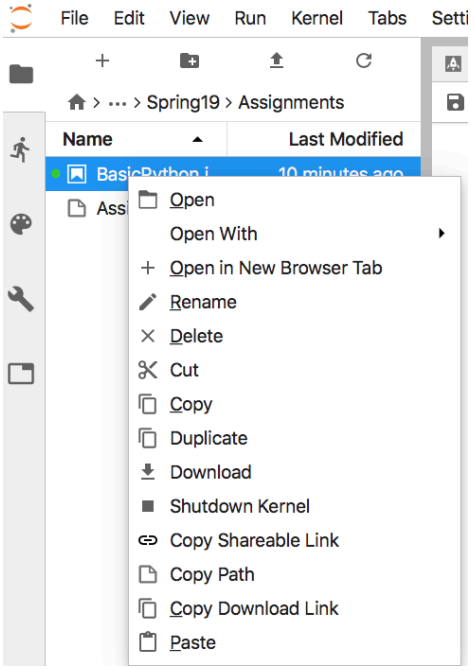
Once you have downloaded the assignment zip it up and then upload the zip file to the course portal.

Using Classic Juptyer Notebook



Using JupterLab

Right-click on the Notebook name in the file browser and select download.



Late Penalty

An assignment turned in 1-7 days late, will lose 5% of the total value of the assignment per day late. The eight day late the penalty will be 40% of the assignment, the ninth day late the penalty will be 60%, after the ninth day late the penalty will be 90%. Once a solution to an assignment has been posted or discussed in class, the assignment will no longer be accepted. Late penalties are always rounded up to the next integer value.