CS 696 Intro to Big Data
Fall Semester, 2020
Exam

San Diego State University -- This page last updated 4/6/20

Due Apr 21 23:59

Version 1.1

Covid-19

The data is from the website:

https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

There are three files:

- covid_deaths_usafacts.csv
- covid_confirmed_usafacts.csv
- covid_county_population_usafacts.csv

You can down load the files from the assignment page.

covid_deaths_usafacts.csv

This contains the known deaths in each county in the USA each day from January 22 to April 6. The column headers besides the dates are:

- countyFIPS - Federally assigned number identifying each county in the country
- County Name - obvious
- State - The state the county is in
- stateFIPS - Federally assigned number identifying each state in the country. This number starts the countyFIPS

covid_confirmed_usafacts.csv

This file contains the know covid cases in each county in the USA each day from January 22 to April 5. The column headers are the same as above.

covid_county_population_usafacts.csv

This file contains the population of each county in the USA. The column headers are:

- countyFIPS - Federally assigned number identifying each county in the country
- County Name
- State - The state the county is in
- population - Population of the county

Questions

Use Pandas to answer the following questions.

1. Compute the total confirmed cases per day in the country. Produce a table and a line plot.

2. Compute the total confirmed cases per week in the country. Produce a table of results. Plot the results using a log line plot. That is the log of the number of cases. A log plot of an exponential process will produce a straight line. Does the resulting plot look like an exponential?

3. Compute the number of new confirmed cases per week in the the country. Produce a table of results.

4. Repeat #2 with the number of known deaths.

5. Find the 10 counties with the most known number of covid-19 cases. Plot the number of cases each week. What are the differences or similarities?

6. Compute the number of known cases per population in each county that have cases of covid-19. Produce a bar graph and table of the results for the 10 counties with the highest values and the 10 counties with the lowest values.

7. Compute the death rate (deaths/ known cases) in each county that have both deaths and known cases. Compute the mean and standard deviation of the results.

8. Produce a violin plot of the death rates found in #7

9. There is some hope that covid-19 may have peaked in NYC (counties Bronx, Kings, New York, Queens and Richmond). Is there any evidence of that in the data?

10. Seattle (King county in Washington) also hopes that they have peaked. Is there any evidence of that in the data?

Names

The names files contains the names of babies born in the USA from 1880 and 2018. See the read me file for file format. You will create two models to predict the sex of a person. Download the file names.zip from the assignment web page.

1. Combine the files into one dataframe that contains year, name, sex and number of people born that year. Create a training set and a test set from this dataframe.

2. For the first model determine for each name which sex used the name most in the training data. How good is the model on the test set? How to measure good?

3. For the second model create a decision tree using the training data. How good is the model on the test set?

4. How do the two models compare?

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

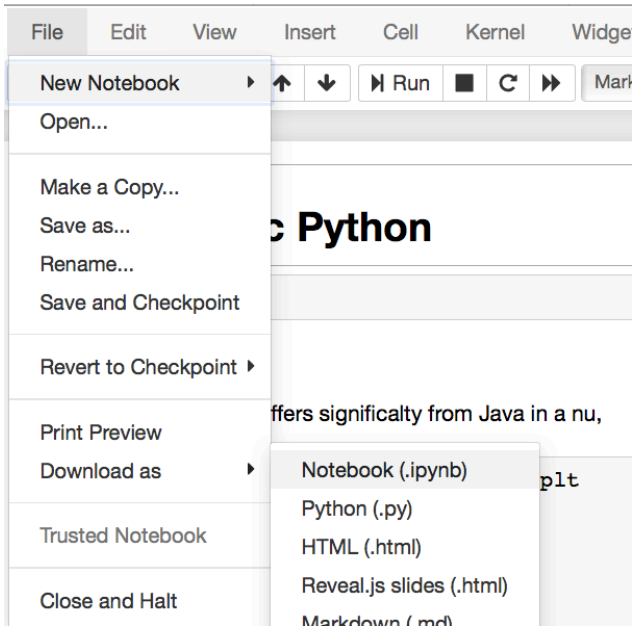
Each problem is worth 10 points.

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.

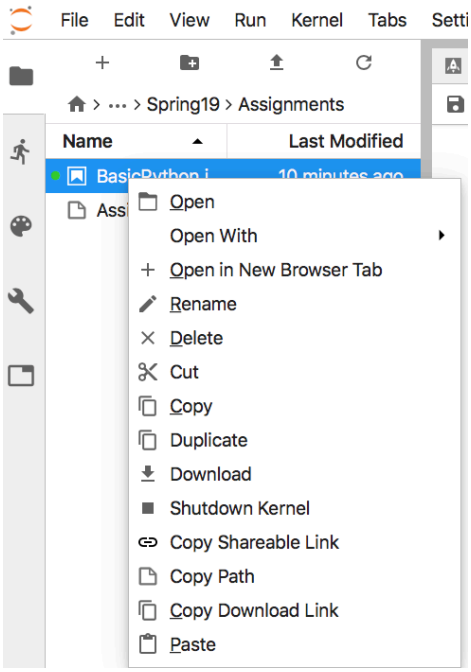
Once you have downloaded the assignment zip it up and then upload the zip file to the course portal.

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.

Version History

1.1 Clarified covid question 6. 3/9/2020