CS 649 Intro to Big Data
Spring Semester, 2021
Assignment 1

San Diego State University -- This page last updated 2/7/21

Due Feb 23 23:59

Using the datasets described below answer the following questions. Turn in your assignment in Jupiter notebook format. Include the questions in your notebook. Make sure that your answers are clearly marked.

1. Plot the weekly number of new cases of covid in the following states: California, Oregon, Washington, and Nevada. A week starts on Monday and ends on Sunday.

2. Do those states follow the the same trends?

3. Plot the weekly number of covid deaths in the following states: California, Oregon, Washington, and Nevada.

4. How do the deaths and the new cases compare?

5. Compute the number of new covid cases per 100,000 population in each county per week. What are the 20 highest rates achieved. Show the date, county and the rate.

6. Compute the number of covid deaths per 100,000 population in each county per week. What are the 20 highest rates achieved. Show the date, county and the rate.

7. What is the relation between the results in #5 and #6.

8. Find the week that each county achieved their highest rate of new covid cases per 100,000 population. If a county reach the peak multiple ties pick the earliest one. Plot via a bar chart the number of countries that reached their peak each week. What does the plot indicate about the pandemic?

9. Compute the total number of covid cases in each county and the number of covid deaths. Compute the percent of the population in each county that is still alive and has had covid. Show the top 20 rates. Is any county getting close to herd immunity, that is 70% or more people immune?

10. Produce a Violin and box plots of all the values computed in #8. What do the plots show about the pandemic in the USA?

Data

The data used is from USAFacts (https://usafacts.org). The data sets can be downloaded on the assignment page of the course website.

covid_confirmed_usafacts

This data set from USA Facts contains the new covid cases in each county in the country each day from the beginning of the pandemic.

Column Labels

State

2 letter abbreviate for the State

County Name

Name of the County

stateFIPS

Federal ID number for the state

countyFIPS

Federal ID number for the county

2020-01-22

Column contains the number of new cases.

covid_deaths_usafacts

This data set from USA Facts contains the covid deaths in each county in the country each day from the beginning of the pandemic.

Column Labels

State

2 letter abbreviate for the State

County Name

Name of the County

stateFIPS

Federal ID number for the state

countyFIPS

Federal ID number for the county

2020-01-22

Column contains the number of deaths.

covid_county_population_usafacts

Column Labels

countyFIPS

Federal ID number for the county

County Name

State

population

Population of the indicated county.

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.

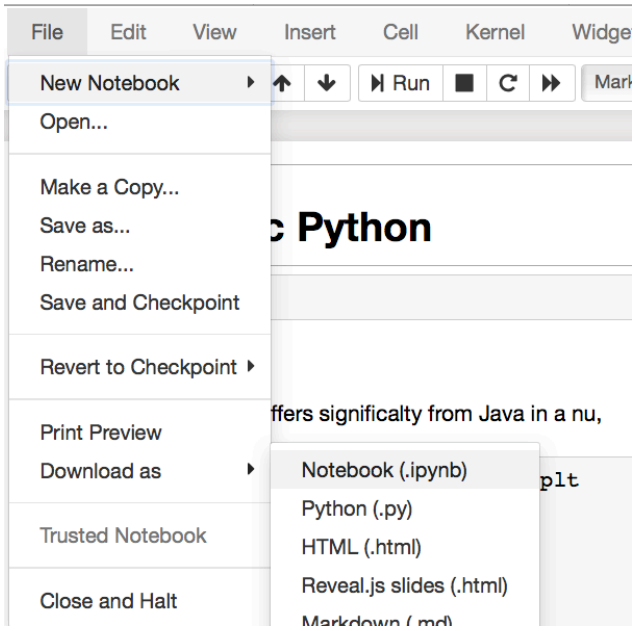
Grading

Each problem is worth 10 points.

What to turn in

You need to turn in a zipped version of your notebook file. There are several ways to do this. Once is just to zip up your notebook file. Another is to download your Jupyter notebook as an IPython Notebook (.ipynb). See image below. 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



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.