Harsh Bhatia

CPS 3320

Dr. Robert J Domanski

Final Project Proposal

County-wise COVID19 Tracker

This project plans to be a continued model of the little research I started with in the 2nd project part wherein I used the Pandas library to first extract open sourced data and then work on it to find out the trend in the number of cases and deaths for Middlesex county in NJ. I want to continue the work on it and extend the whole project to cover all counties in NJ and show the data by representative charts of different kinds.

On working with a dataset imported straight from Kaggle. The goal will be to complete the data analysis cycle. Thus, the project will include:

1. **Finding the dataset**: researching and finding the dataset appropriate for the use here
2. **Data wrangling:** cleaning the data we have found and filtering out the unnecessary and irrelevant fields, like fields that don’t represent the correct data, for example: outliers, or null records that take up space. Also non-numerical entries in numerical fields can result in incorrect results.
3. **Data segregation:** Now that we have a clean dataset what we need to do is to separate the data we need. For example, filtering out the data to only contain entries with NJ state and removing all the other data entries from the dataset.
4. **Data study:** Studying what the data represents, the structure, what each column in trying to represent, and how we can use each column to add value to our research. For example, during project 2, the dataset did not have a column for new cases and new deaths per day, so I had to calculate that and insert it into the dataset to be used.
5. **Data analysis:** Now that we have a dataset that we can use to showcase COVID19 information across NJ, we have to show that to someone that would not understand this code and the best way to do that is by drawing charts and graphs, visualizing the data. I plan on using a combination of matplotlib and seaborn libraries for this purpose.
6. **Explaining visualized data:** Now we have to explain this visualized data to someone to make sense out of it. For example, you want a comparison between Union and Hudson county, we cannot leave you with 2 charts, but I will try to go a step further and explain these numbers.

This is my purpose plan for this project and I will be using Jupyter notebooks to showcase this information just because I feel it is the best way of showcasing data, as using terminal may sometimes have complexities involved. However this will still be under consideration throughout the coding phase.

Thanks for going through this let me know if there is anything you would like to change.