Project DSC540 - Covid 19 Data Analysis and Comparison

By Kausik Chattapadhyay

Milestone – 1

Date: 12/20/2022

Context

The novel coronavirus, also known as SARS-CoV-2, is a contagious respiratory virus that first reported in Wuhan, China. On 2/11/2020, the World Health Organization designated the name COVID-19 for the disease caused by the novel coronavirus. This new strain of virus has strike fear in many countries as cities are quarantined and hospitals are overcrowded.
This project aims at exploring COVID-19 through data analysis and visualization.
.

Milestone 1 – Data Gathering

For this project, I have selected data from the following sources:

1. **CSV** – The Covid 19 data is scrapped from John Hopkins University github repo : https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series , this has
    a. Daily time series summary tables, including confirmed, deaths and recovered. All data is read in from the daily case report. The time series tables are subject to be updated if inaccuracies are identified in our historical data.
    b. Two time series tables are for the US confirmed cases and deaths, reported at the county level.
    c. Three time series tables are for the global confirmed cases, recovered cases and deaths. Australia, Canada and China are reported at the province/state level.
    d. Data is updated at a daily basis.

2. **API** – I have used the api : https://services.arcgis.com/lQySeXwbBg53XWDi/arcgis/rest/services/Map/FeatureServer/0/query?where=1%3D1&outFields=*&outSR=4326&f=json

    which will provide the demographic info (Age, employment, sex, ethnicity etc) for US at a County level for all the corona virus cases, this information is vital to understand the rate of spread across

communities in United States. I will use this data to deep dive into corona cases in US and generate some interesting facts.

3. **Web** – I am scrapping data from https://www.worldometers.info / website for more insights at a world level like population, density, area and other factors.

The datasets 1 and 2 are related by county and country code, whereas the dataset 1 and 3 are related by Country code, this will help me join the 3 datasets into one and perform any slicing dicing for visualization.

Summary of what I learned and had to do to complete the project.

First of all, I got to learn Pandas library which is new to me, I have used PySpark earlier which is little different in syntax and implementation, while working on this project, I will touch all areas of data wrangling which includes:

1. **Data Gathering from Different Sources** – While I had previous experience working with CSV for data gathering, I learned doing Web Scrapping using Beautiful Soup for processing HTML data. However, there were a few challenges as the Web has grown organically out of many sources. It combines a ton of different technologies, styles, and personalities, and it continues to grow to this day.
   Also, I learned to fetch data from Web API using requests library, Request returns a Response, a powerful object for inspecting the results of the request. Using Response, I will examine the headers and contents of the response, get a dictionary with data from JSON in the response, and also determined how successful our access to the server was by the response code from it.

2. **Data Cleaning and Transformation -** Learned how to filter, sort, merge, join two data frames in Pandas.
3. **Data Visualization -** Learned a lot of visualization libraries like matplotlib, seaborn and plotly.
4. **Importing data in Sql table from Python –** Mostly in my previous projects I never created any Db, added any table and queried, in this project I got a chance and learned sqllite3 to create database, I have also learned to use sqlalchemy library to create a database table and read into a data frame.

If data is incomplete, unreliable, or faulty, then analyses will be too—diminishing the value of any insights gleaned.

Data wrangling seeks to remove that risk by ensuring data is in a reliable state before it's analyzed and leveraged. This makes it a critical part of the analytical process. when done manually data wrangling can be time-consuming. This can be reduced by defining some policies around data—for example, requiring that data include certain information or be in a specific format before it's uploaded to a database.