Name: Neel Adalja

Covid-19 Data Analysis

BAC66 - Data Analytics and AI Bootcamp

Date: December 24, 2024

Instructor: Ram Gupta

**1. Covid-19 Data Analysis**

**Purpose**: The purpose of this project is to visualize how covid 19 affected different countries. The main goal is to learn about the active cases, total casualties, total cases, tests done and the per million population calculations. This will help in future to prevent such outbreak and take precautions incase of an outbreak.

**Abstract**

The main goal of this project is to get meaningful insights from the covid-19 data from 2020 to 2024. The data is collected from the open source website that has covid-19 counter and table with information about the total cases, total deaths total tests, total recovered and much more, the goal is to scrape this data from the website with the help of python library and then use pandas to clean the data and then import it to csv, then the Tableau Prep Builder will come into action for further cleaning and joining additional data. Finally, the data will be used to build an interactive dashboard for visualizing the data.

**Table of Contents**

* Introduction
* Web Scraping
* Data Cleaning and Preprocessing
* Dashboard Development
* Data Analysis
* Challenges
* Learnings
* Future Work
* References

1. **Introduction**

**Background**:

it is important to collect the right Covid-19 data for this analysis because the there are many sources and the data available will be not accurate if its not from a trusted source. The website I used this for this project is <https://www.worldometers.info/coronavirus/#page-top> which is a very reliable source and they have recorded the reporting till the April 2024. The data collected is also from the ships that were in the ocean at the time of the outbreak so it has recorded minute details also.

**Objectives**: tracking the global covid-19 data and making interactive Tableau Dashboard.

1. **Web Scraping**

* **Tools Used**: Visual Studio Code, Jupyter Notebook, Github, Tablue Prep Builder, Tablue Public
* **Languages and Libraries used:** Python, (Pandas, CSV, JSON, BeautifulSoup)
* **Steps**:
  1. Identify the data source (<https://www.worldometers.info/coronavirus/#page-top>).
  2. In our case, it is in the table in the end which is inside the table tags(HTML).
  3. First the required libraries were imported and then the request to connect to the server was made. The request was successful as we used the header describing the human attempt to scrape the data. The html was then parsed, and the data was collected by identifying the table ID and the table tags inside the table body. Therefore, I created an empty list named world\_info[ ] to store the data and looped through rows of the list and stored all the values in the empty list we created (World\_info[ ]). Moving further, using pandas, created the data frame and checked for any duplicates. Cleaned the data where I performed tasks like deleting the first 2 rows because the were world statistics. Now the data was ready to be exported into a csv file for further cleaning in the Tableau Prep Builder.
* #imported all the required libraries
* import requests
* import pandas as pd
* from bs4 import BeautifulSoup
* import csv
* import json
* #url for the data source
* url = 'https://www.worldometers.info/coronavirus/#main\_table'
* #sending the request to and using headers to identify the user.
* response = requests.get(url, allow\_redirects=True)
* response.headers
* #connectin established and received the html, now parsing it to get the data.
* soup = BeautifulSoup(response.content, 'html.parser')
* #making html easy to look using prettify
* soup.prettify()
* soup
* # our data is inside the table tags and the tbody tags.
* table = soup.find('table', id='main\_table\_countries\_today').find('tbody')
* rows = table.find\_all('tr')
* #usingthe empty dictionary to store the data, looping through the dictionary we have and storing the data inside the td tags.
* world\_info = []
* for row in rows:
* columns = row.find\_all('td')
* country\_info = [column.text.strip() for column in columns]
* world\_info.append(country\_info)
* world\_info
* del world\_info[0]
* # delete the first row in the data we have
* #first row contains the world stastics.
* del world\_info[0:6]
* #this is our data
* world\_info
* #defining the column names for the dataframe.
* columns = [
* "number", "country", "total\_cases", "new\_cases", "total\_deaths", "new\_deaths", "total\_recovered",
* "active\_cases", "serious\_critical", "total\_cases\_per\_1m\_pop", "deaths\_per\_1m\_pop", "total\_tests",
* "tests\_per\_1m\_pop", "population"
* ]
* filename = 'covid\_data.csv'
* # Saving it to CSV
* with open(filename, 'w', newline='', encoding='utf-8') as csvfile:
* csvwriter = csv.writer(csvfile)
* # Writing the custom header we created
* csvwriter.writerow(columns)
* # Writing the data
* for row in world\_info:
* csvwriter.writerow(row[:len(columns)])  # Ensure it matches the column count
* print(f"Data has been saved to {filename}")
* #using pandas to convert it in dataframe and do the cleaning.
* df = pd.read\_csv('./covid\_data.csv')
* df.head(10)
* #looking at the data types and the null values
* df.info()
* #vaidating that we have same rows as the website, the 4th element was germany.
* df[df['number'] == 4]
* # droppingany duplicates
* df = df.drop\_duplicates()
* df
* df1 = df.copy()
* #copyed the dataframe, now further ceaning will be done in the tablue prep builder.

Fig: Source Code

1. **Data Cleaning**

* **Objective**: Removing null values and N/A values where data is not available.
* **Steps**:
  + Remove unwanted columns.
  + Aggregation and grouping.
  + Cleaning for N/A values
  + Correcting incorrect values and checking inconsistencies.

A screenshot of a computer

Description automatically generated

Fig: Data Flow from Tableau Prep Builder

* + Statistical analysis (e.g., growth rates, correlations).

**Steps taken:**

removed the total recovered column as it only had around 700 total recovered in the whole world. While the data for most of the countries was 0. Removed like new\_recovered, New\_deaths, etc. because they were empty columns.

Cleaned the data rows where the N/A values were in high values and filled in with 0 so that it does not affect analysis.

Added another table to have country with population. This table was required because the scraped data did not had the total population of the country. Therefore, cleaned the new table and then joined the new table with the covid-19 dataset in the Tableau prep builder. This created some conflicts in country fields, because some names were differently written in the new table. Manually corrected those mistakes and the conflicts were resolved.

Again, cleaned the new merged data and removed duplicates.

1. **Dashboard Development**

* **Tools Used**: Tableau Public.
* **Design**:
* Designed map to show total cases in the country for better visuals
* Created a bar chart for the total\_cases VS total\_deaths
* Bubble chart for the number of casualties vs the population of the country
* Heat chart for the total test per 1 million vs the country population.
* Created a KPI for total cases and total deaths.

A screenshot of a computer screen

Description automatically generated

1. **Data Analysis:**

**Key Insights:**

* The Total Cases recorded across the world were 703,215,677
* The Total Deaths were 6,996,771 across the world.
* The highest cases were recorded in USA, followed by India and Brazil, and Russia.
* The total casualties in the USA were higher compared to all other nations which was 1,219,487. And followed by India, Brazil, Russin and China.
* Region wise, the total cases were higher in Europe while the deaths total was around same for North America and Europe Region.
* The highest amount of tests were taken in Peru, Bulgaria, Bosnia and followed by USA, UK and Brazil compared to total population of those countries.

1. **Challenges:**

* The Cleaning was challenging, especially when there were many null values, and the data is important for every country so it was not easy to change it.
* While some columns were having values for few countries those columns were mostly full of null values. So, it stopped me from performing some KPIs that I was interested in performing.
* **Learnings:** Using APIs is easier in some cases than scraping the data. And the data we get is also tidier.

1. **Future work:** Fields like date will be a good addition to the dataset to see the impact and countries getting affected over the time.

**References**

Coronavirus cases:. Worldometer. (n.d.). <https://www.worldometers.info/coronavirus/#page-top>

<https://www.youtube.com/watch?v=mvpGTXRLIQc&ab_channel=AnthonySmoak>