FINAL PROJECT REPORT

EAS 503 – PROGRAMMING AND DATABASE FUNDAMENTALS FOR DATA SCIENCE MAJOR PROJECT

Fall 2023

Group 94

22nd December 2023

Chinmaya Lakshmi Rupa Devi Seelam - 50538411

Indra Kiran Reddy Bonthu - 50518805

Mounika Male - 50541143

Mallikarjun Reddy Reddy - 50537829

Importance of the Problem:

The selection of this specific problem was motivated by the increase in the covid cases from the recent most variants (JN1 which is a recent variant i.e.; DEC. 14, 2023) of covid here and there, the primary goal is to derive meaningful insights and actionable information to support public health decision-making and crisis management. The main objective is to perform a comprehensive analysis of the COVID-19 Outcomes by Testing Cohorts: Cases, Hospitalizations, and Death's dataset, which includes essential attributes such as 'extract_date,' 'specimen_date,' 'Number_tested,' 'Number_confirmed,' 'Number_hospitalized,' and 'Number_deaths.' The primary goal is to derive meaningful insights and actionable information to support public health decision-making and crisis management. The analysis will encompass temporal trends, providing a holistic understanding of the COVID-19 data. The implementation of interactive features, including buttons and user-friendly interface, will enhance the analytical capabilities, to dynamically explore the data and extract valuable information for evidence-based decision-making and resource allocation.

Data Description

The data is taken from: https://data.cityofnewyork.us/Health/COVID-19-Outcomes-by-Testing-Cohorts-Cases-Hospita/cwmx-mvra It contains 176k rows and 6 columns.

Columns in this Dataset:

1. extract_date: Date on which a biological specimen (such as a sample of blood, tissue, or other bodily fluid) was collected for testing or analysis.

2. specimen_date: Date of specimen collection, equivalent to diagnosis date.

3. Number_tested: Count of NYC residents newly tested for SARS-CoV-2.

4. Number_confirmed: Count of patients tested who were confirmed to be COVID-19 cases.

5. Number_hospitalized: Count of confirmed COVID-19 cases among patients ever hospitalized.

6. Number_deaths: Count of confirmed COVID-19 cases among patients who died.

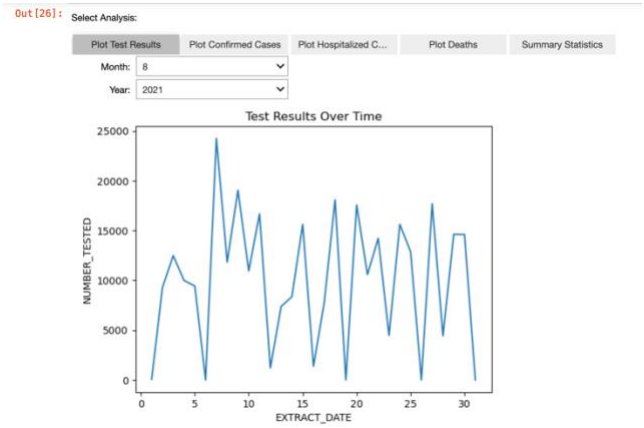
Approach to Problem Analysis:

- Parsed the raw data in the .csv file using file handling techniques and created a list of data for each Normalized table.
- Created a Connection with SQL database and Normalized tables i.e., DATEINFO, TESTINFO, and CASESINFO, Loaded the parsed data into the normalized tables. These tables are designed to store information related to COVID-19 testing, including dates, test information, and case outcomes.

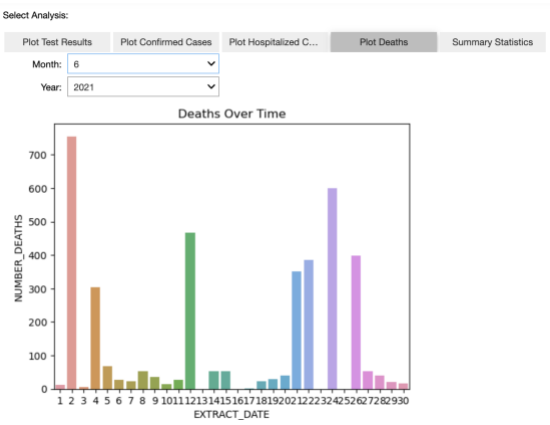
- Executed a SQL query using joins and retrieved data using Pandas
- Conducted Exploratory Data Analysis (EDA) to comprehend data distribution, drawing meaningful conclusions through visualizations and statistics of Number of tests, Number of confirmed cases, number of patients hospitalized and number of deaths for a particular date in a particular month of a particular year using ipywidgets to develop a user interface.
- Visualization using pie chart and Box plot to understand the proportion and distribution of confirmed cases, hospitalized cases, and deaths to distribution of confirmed cases, hospitalized cases, and deaths

Conclusion From the plots:

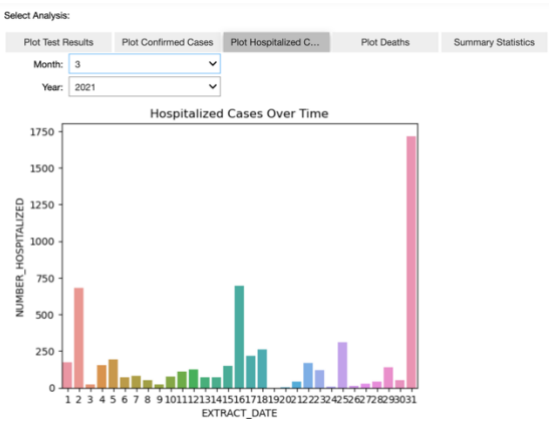
- Line Plot identify periods of increase or decrease in testing activity Note any spikes or drops in the number of tests, which might indicate specific events or changes.
- Bar Plot examine the daily variations in the number of confirmed cases, no.of deaths and no. of people hospitalized and identify any sudden increases in confirmed cases, which might indicate outbreaks or changes in testing strategies. Look for patterns in the data that might reveal insights into the spread of the virus.
- Summary Statistics examine key statistics such as mean, median, minimum, maximum, and quartiles. Understand the central tendency and variability in the data.
- Pie Chart understand the proportion of confirmed cases, hospitalized cases, and deaths. Identify the percentage of cases in each category relative to the total.
- Box Plot identify the distribution of cases by examining the box plot.Identify the median, quartiles, and any outliers. Understand the spread and variability in the number of confirmed, hospitalized, and death cases.



Out[26]:



Out[26]:



Out[26]: Select Analysis:

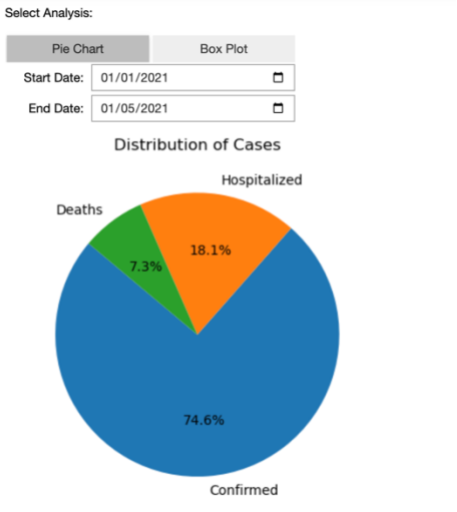
Plot Test Results | Plot Confirmed Cases | Plot Hospitalized C... | Plot Deaths | Summary Statistics

Month: 6

Year: 2021

	TEST_ID	NUMBER_TESTED	NUMBER_CONFIRMED	NUMBER_HOSPITALIZED	NUMBER_DEATHS	EXTRACT_MONTH	EXTRACT_YEAR	SPECIMEN_I
count	30.000000	30.000000	30.000000	30.000000	30.000000	30.0	30.0	30
mean	241.033333	10905.866667	1948.833333	385.500000	128.633333	6.0	2021.0	6
std	207.581224	7518.061378	1614.975672	509.438755	203.512515	0.0	0.0	3
min	2.000000	1.000000	0.000000	0.000000	0.000000	6.0	2021.0	1
25%	83.750000	5689.750000	551.750000	86.000000	15.500000	6.0	2021.0	3
50%	199.500000	10186.500000	1463.000000	149.000000	33.000000	6.0	2021.0	5
75%	354.750000	16616.750000	3461.500000	277.750000	64.250000	6.0	2021.0	8
max	742.000000	23350.000000	5890.000000	1852.000000	755.000000	6.0	2021.0	12

Out[30]:



Out[30]: