New York State Statewide COVID-19 Testing Analysis

- Sahithi Dodda

ABSTRACT:

2019 Novel Coronavirus (2019-nCoV) is a virus (more specifically, a coronavirus) identified as the cause of an outbreak of respiratory illness first detected in Wuhan, China. The primary the goal of this model is to provide users with timely analysis and graphical representation about Covid 19 spread and reporting of positive cases in various regions of New York State.

INTRODUCTION AND DATA DESCRIPTION:

The dataset which we considered for this model has around 59.7K Rows and 8 Columns. This data includes the “Test Date”, “County”, “New Positives” which represents the number of new positive cases on that particular day, “Cumulative Number of Positives” which indicates the total number of positive cases as of that day in that county, “Total Number of Tests Performed” it included all kind of test results like both positive and negative, “Cumulative Number of Tests Performed”, “Test % Positive” and “Geography” which contains county or region or statewide as the value. Considering Data Provided by the New York State Department of Health which was last updated on May 6, 2022.

IMPLEMENTATION AND ANALYSIS:

Normalization:

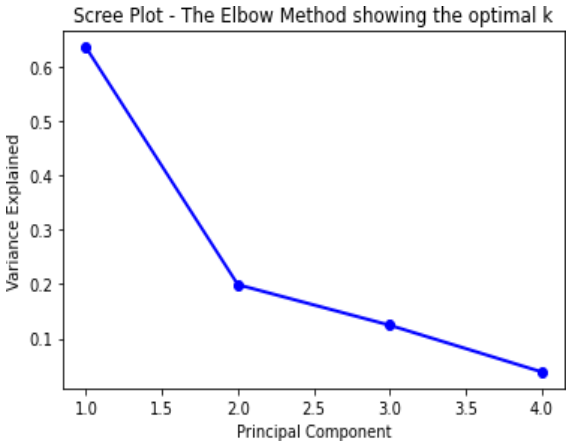
To maintain the redundancy of the data we normalized the table. After the normalization, we have 4 tables.

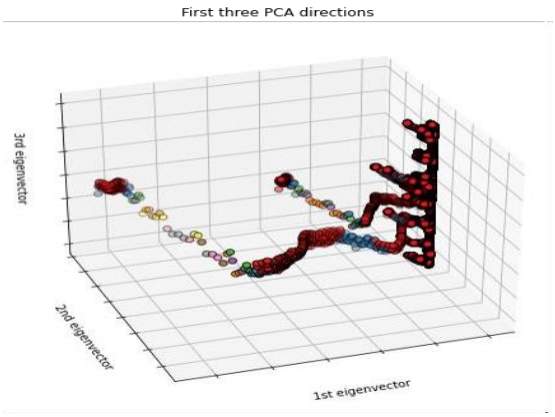
1. COUNTY table – COUNTYID, COUNTY attributes.
2. DATE_COUNTY table - Test_ID, Test_Date, County_ID attributes.
3. Geography table – GeographyID, Geography attributes.
4. COVID_RESULTS table - Test_ID, New_Positives, Cumulative_Number_of_Positives,Total_Number_of_Tests_Performed, Cumulative_Number_of_Tests_Performed, Test_Positive_Perc, GeographyID attributes.

Principal Component Analysis (PCA) and K-Means Clustering:

PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. Scree plot Gives the measure of the associated principal component’s importance with regards to how much of the total information it represents and helps in finding optimal

number of clusters to be formed. We used 6 features earlier and after the PCA, the first 3 PCA's contribute 0.95860322% (approx. 96%) and when we consider the 4 PCA's it sums to 0.99728777 (approx. 100%) The k-means clustering method is an unsupervised machine learning technique used to identify clusters of data objects in a dataset.





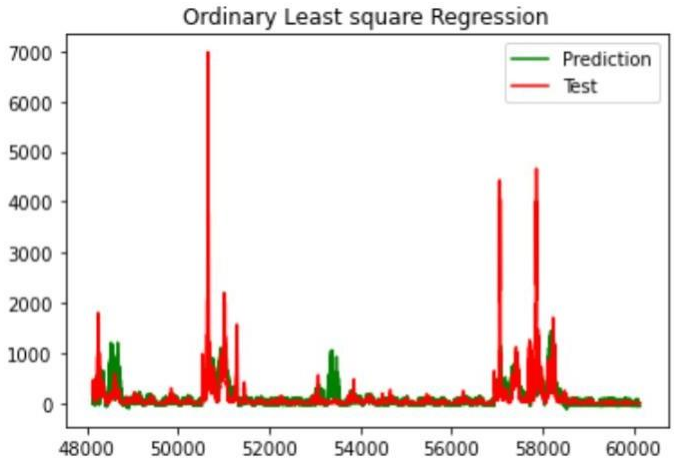
ORDINARY LEAST SQUARE REGRESSION (OLS):

Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable (simple or multiple linear regression).

In this we perform the model testing, then we can find weight vector and mean square error percentage of prediction found.

```
Weight Vector:  [-8.76424383e-01  4.98188294e-02  4.94571598e-02 -5.52797847e+01]
```

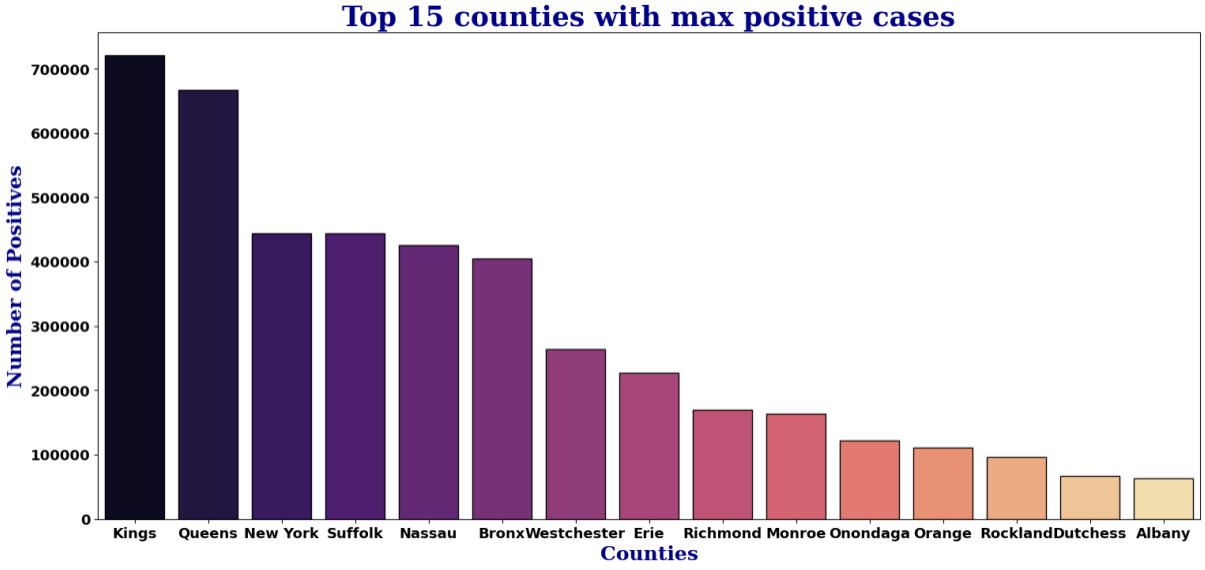
```
Mean Square Error:  2.548192
```

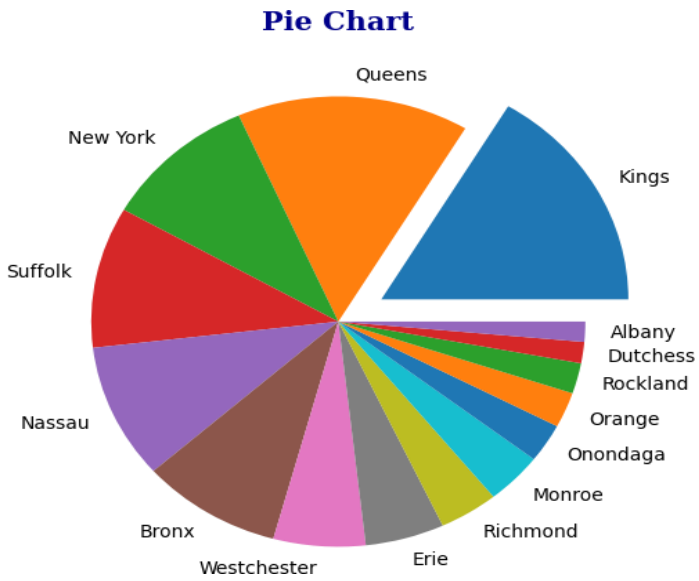


VISUALIZATIONS AND RESULTS:

Use case 1 - Top 15 counties with the highest number of positive cases

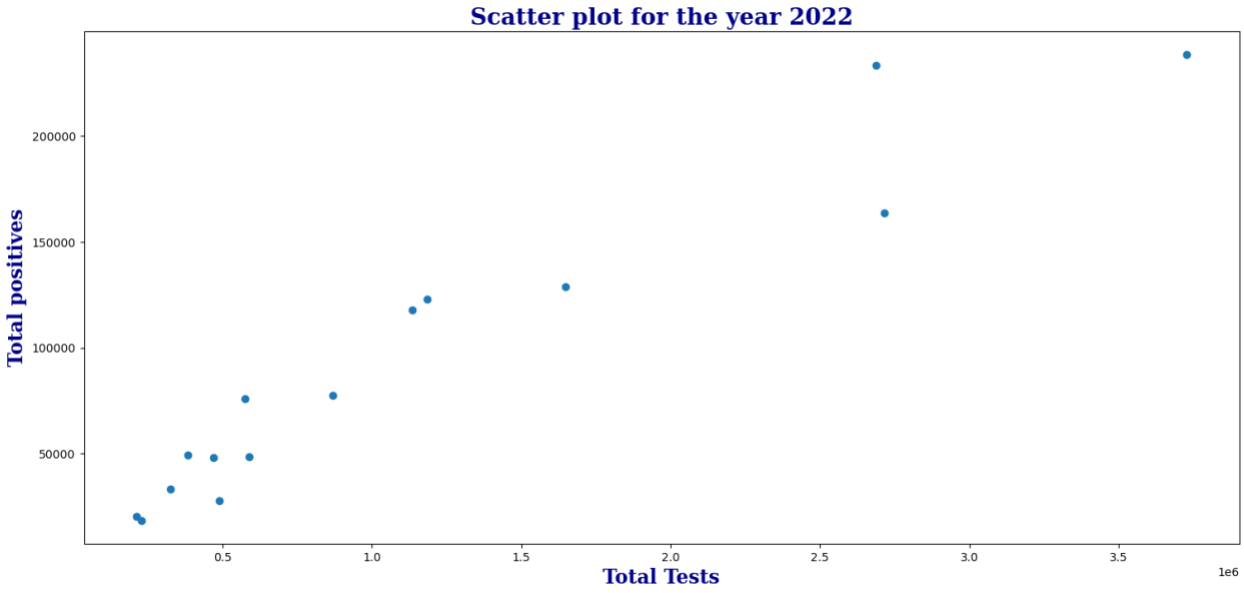
The Bar graph and the Pie chart below represents the top 15 counties in New York state with highest no. of positive cases from March 2020 to May 2022 and the highest positive cases recorded were in Kings.

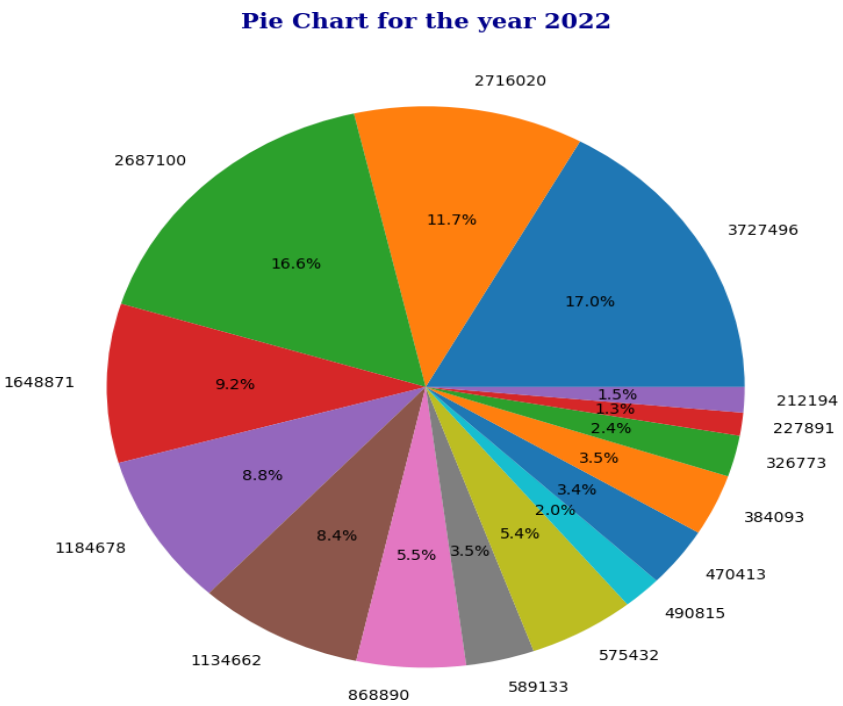




Use Case 2.1: 15 Counties with Highest total tests performed in 2022 VS positive cases in 2022

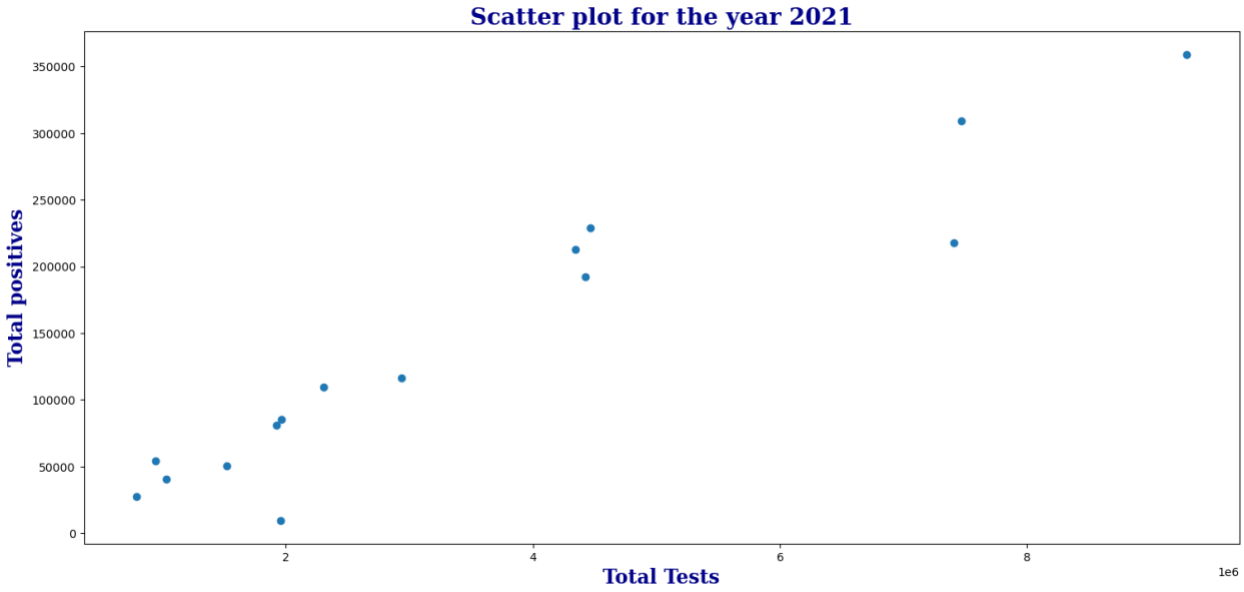
The Scatter plot and the Pie chart below represents the 15 Counties with Highest total tests performed in 2022 VS positive cases in 2022.

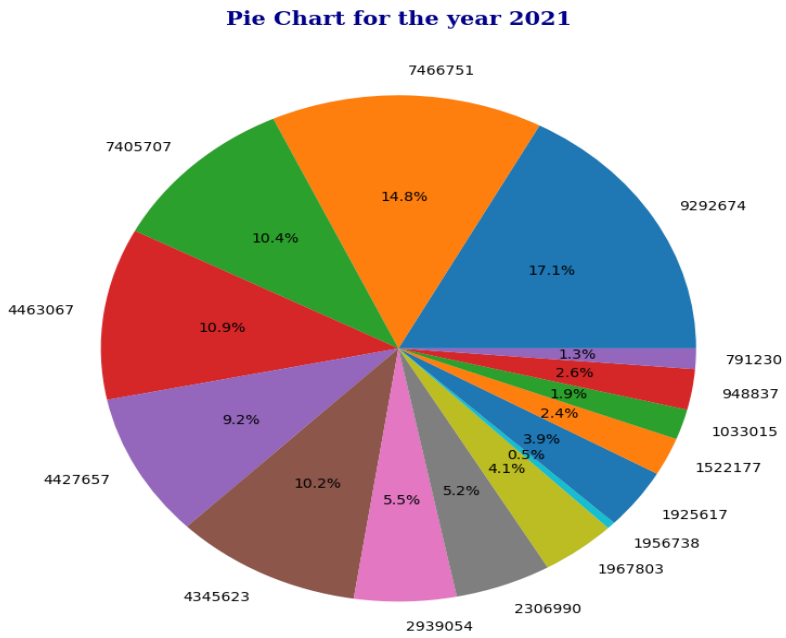




Use case 2.2: 15 Counties with Highest total tests performed in 2021 VS positive cases in 2021

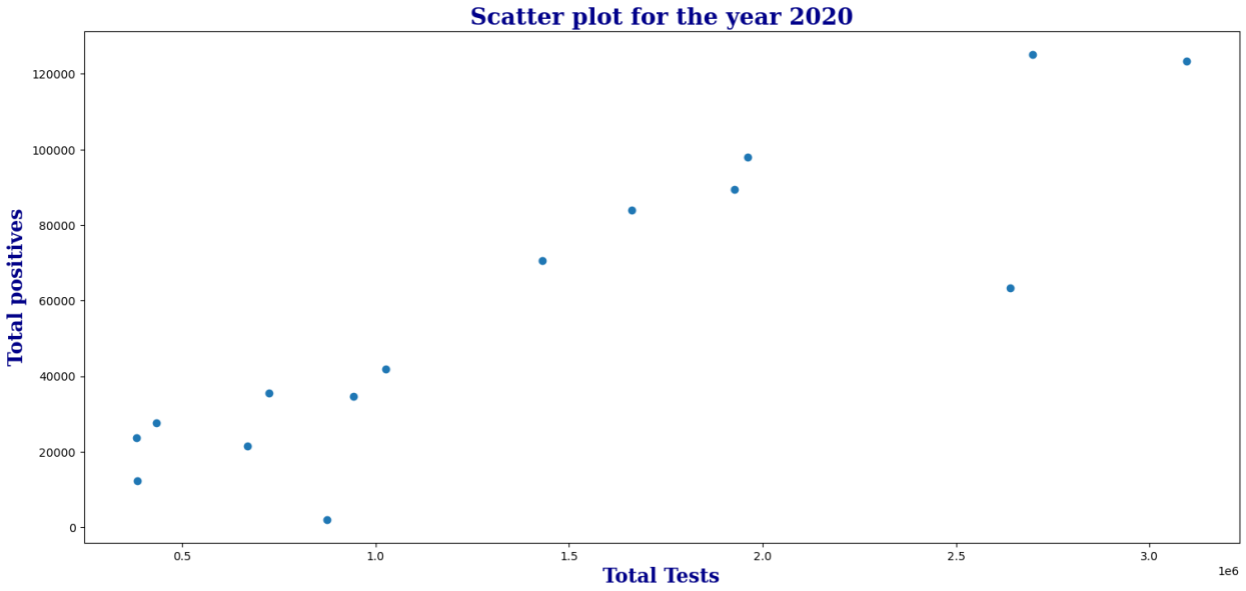
The Scatter plot and the Pie chart below represents the 15 Counties with Highest total tests performed in 2021 VS positive cases in 2021.

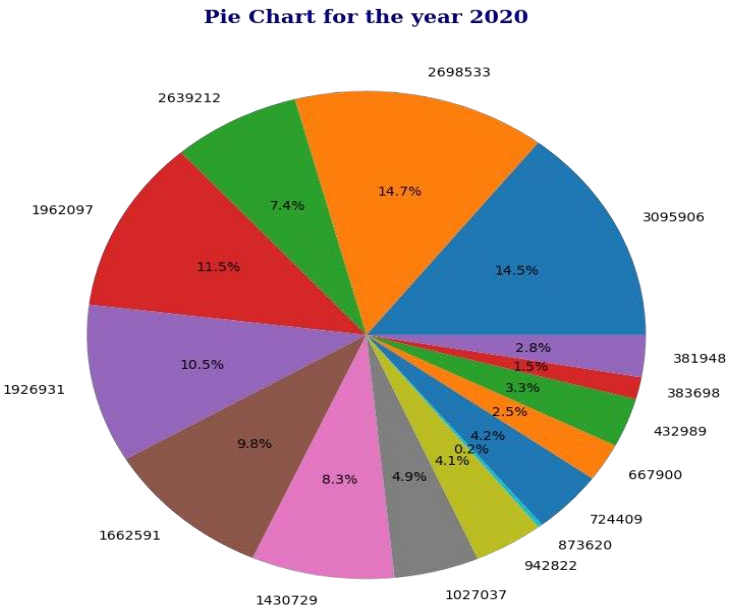




Use Case 2.3: 15 Counties with Highest total tests performed in 2020 VS positive cases in 2020

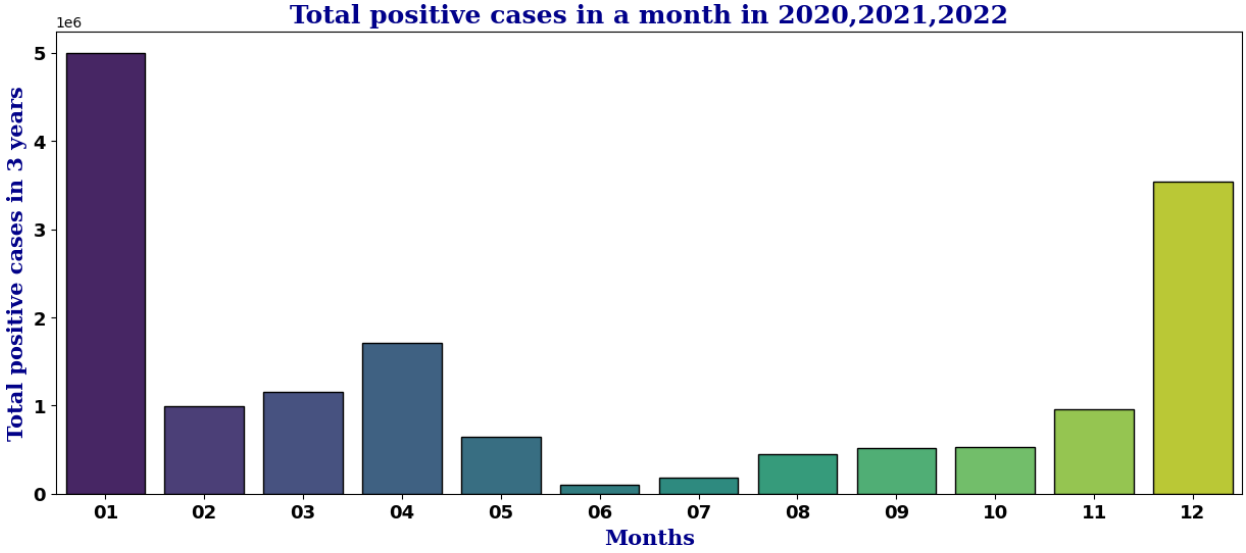
The Scatter plot and the Pie chart below represents the 15 Counties with Highest total tests performed in 2020 VS positive cases in 2020.





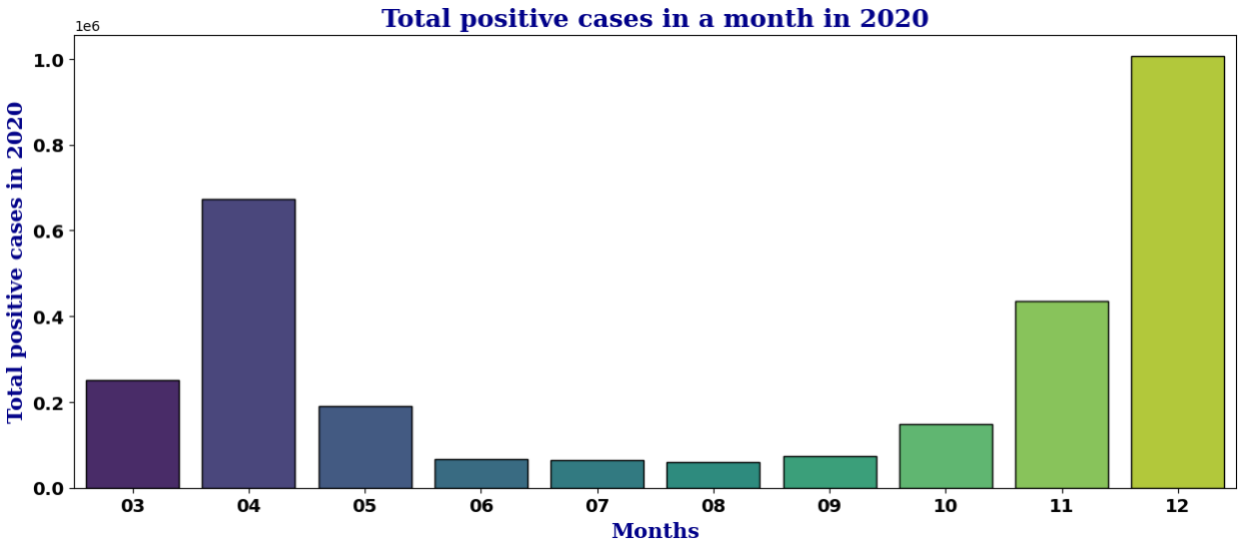
Use Case 3.1: Cumulative positive cases in a month in all the years

The Bar graph below represents the total positive cases recorded in each month in all the years. From this we can observe that positive cases rose drastically in the months of Jan, April and December in all the counties.



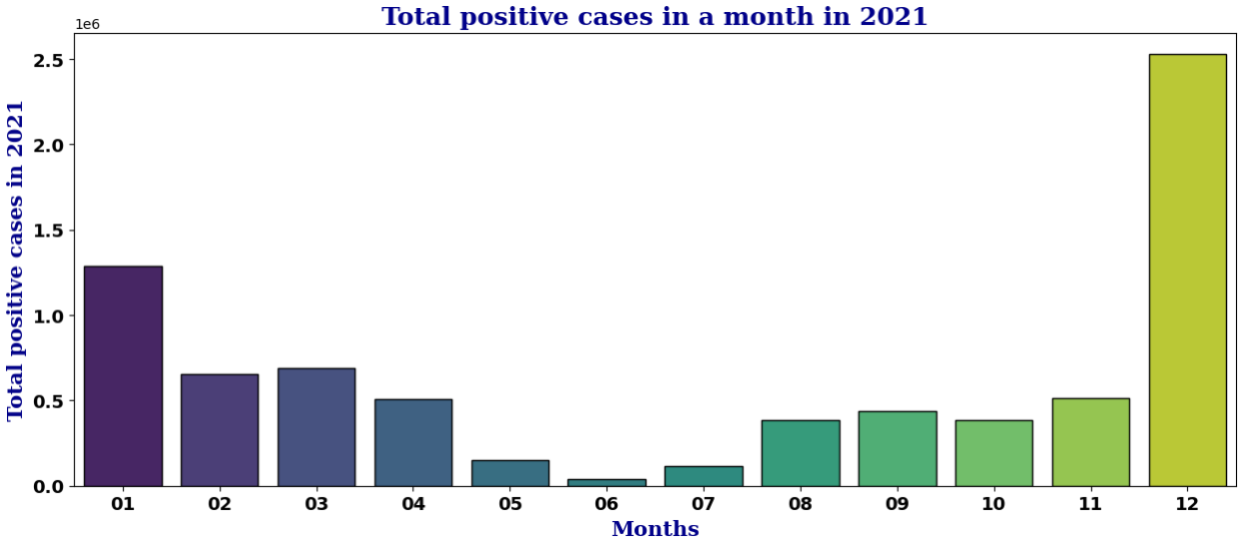
Use Case 3.2: Cumulative positive cases in a month in the year 2020

The Bar graph below represents the total positive cases recorded in each month in 2020. From this we can observe that positive cases rose drastically in the months of April, November and December in all the counties.



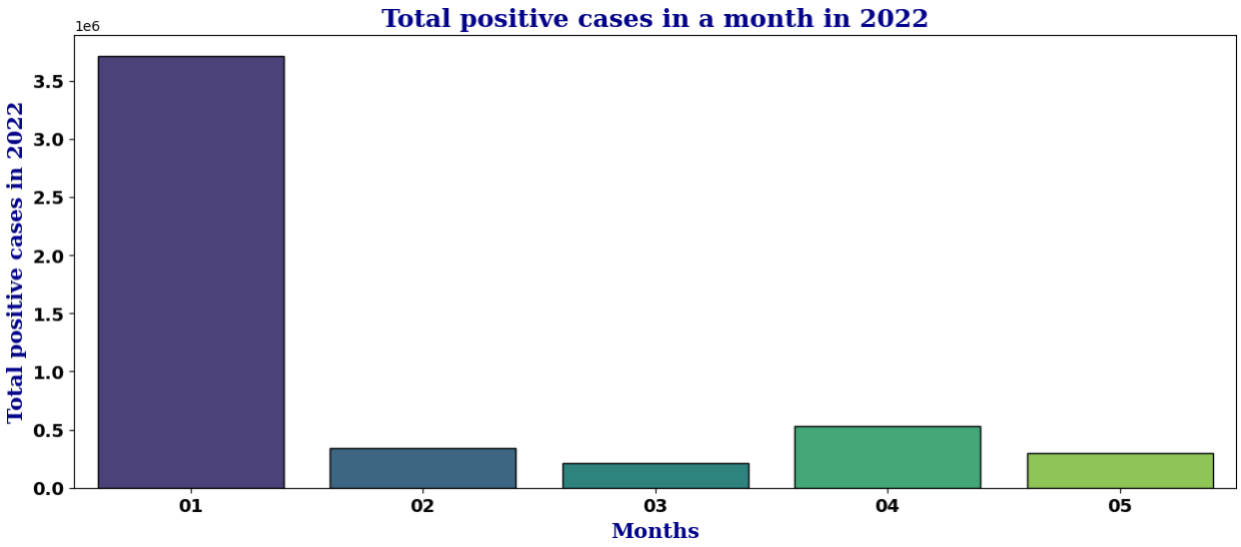
Use Case 3.3: Cumulative positive cases in a month in the year 2021

The Bar graph below represents the total positive cases recorded in each month in 2021. From this we can observe that positive cases rose drastically in the months December and January in all the counties.



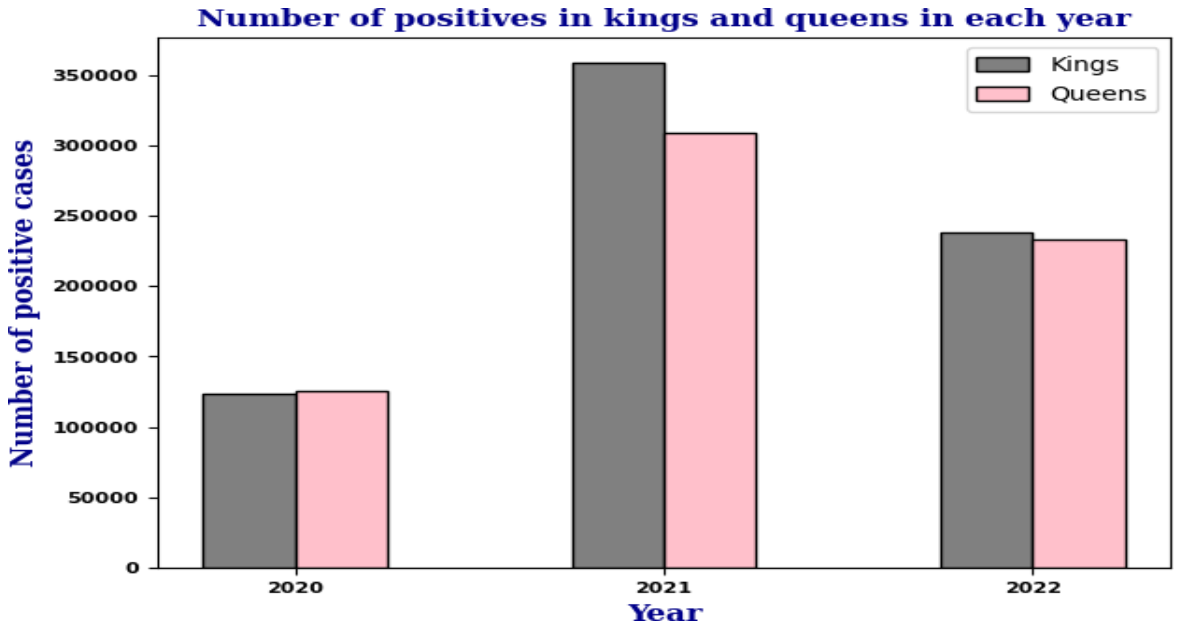
Use Case 3.4: Cumulative positive cases in a month in the year 2022

The Bar graph below represents the total positive cases recorded in each month in 2022. From this we can observe that positive cases rose drastically in January in all the counties.



USE CASE 4: Difference in Positive Cases for max affected counties in subsequent years

The Bar graph below represents the Difference in Positive Cases from 2020 – 2022 for the max affected counties i.e., Kings and Queens. From this we can infer that, in the year 2021 these counties were affected badly when compared to other years.



Use Case 5: Word Cloud

We have created word cloud for county names on basis of total number of positive cases. The counties having greater number of positive cases are viewed in higher font and the ones with least number of positive cases are viewed in lower font.

INFERENCE AND CONCLUSION:

This analysis model can be used anytime by the State Government or Health Care Facilities to check which counties are affected mostly and in which months there is a probability of maximum spike in positive cases. From this model, we can infer that overall, Kings, Queens, New York, Suffolk, Bronx counties were affected maximum across all the 3 years and needs additional support and preventive measures. Furthermore, in the months of November – January, utmost care and safety measures must be implemented for the curb of Covid 19 positive cases across all the counties. This Analysis helps them to understand the repeating patterns and implement necessary measures to prevent the spread of covid.

UB BOX LINK :

https://buffalo.box.com/s/7sph88k3wphsvkljuxkquy6i2moa1g9

References:

https://health.data.ny.gov/Health/New-York-State-Statewide-COVID-19-Testing/xdss-u53e

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

https://www.geeksforgeeks.org/generating-word-cloud-python/

https://www.w3schools.com/python/matplotlib_plotting.asp