The Impact of COVID-19 on Canada: A Comprehensive Data Analysis

Mr.Siju Philip , Professor at School of Advance Technology , At Algonquin College ,Ontario, Canada - philips@algonquincollege.com

Saket Mahar, Alis Findoliya, Ankit Gamechi, Sai Akhil

School of Advacnce Technology , Algonquin College, Ontario, Canada

Abstract—The COVID-19 pandemic has had far-reaching impacts worldwide, presenting substantial challenges to public health, economies, and societal norms. Tracking critical indicators like mortality, recovery, and morbidity rates is essential for gauging the pandemic's progression and guiding targeted responses. Using publicly accessible data, this study conducts extensive data analysis to uncover factors influencing these rates, empowering decision-makers in healthcare and policy to enact informed strategies.

Keywords— Covid-19, Canada, Python, Data Analysis, Data Exploration.

I. INTRODUCTION

The primary objective of this paper is to conduct an in-dept Exploratory Data Analysis (EDA) of Covid-19 metrics, focusing on key indicators including morality, recovery, and morbidity rates. By examining publicly available data, we aim to uncover underlying trends, patterns, and factors influencing these metrics across different regions of Canada. Though we would investigate all the provinces, but on the other hand we will try to focus more on the provinces which were drastically affected by the incident in compared to other, these provinces which were drastically affected by Covid-19 generally consist of densely populated regions like Ontario, Quebec, Manitoba, British Columbia. etc. Through this analysis, we seek to provide valuable insights for policymakers, healthcare professionals, and researchers to better understand the dynamics of the pandemic and guide evidence-based decision making.

A. Methodology

For doing the Exploratory Data Analysis (EDA) on Covid-19 data with the mentioned attributes, we will follow these steps:

1. Data Collection: Obtaining COVID-19 dataset with attributes including province, Daily totals, Summary Date, Total Cases, Total Recovered/Daily Recovered. etc. These data set are openly available on the internet but to show high integrity we downloaded the data set from a reputable government-based site called Statistics of Canada. Also collected similar dataset from organizations like WHO (World health organization), Kaggle, and Google Dataset, but the most appropriate dataset that we selected among those was from Statistics of Canada.

2. Data Cleaning: This step involves handing missing values, correcting data formats, removing duplicates, and dealing with outliers in the selected data set. These steps are taken to ensure consistency in the format of dares and other variables for better analysis of the raw data.
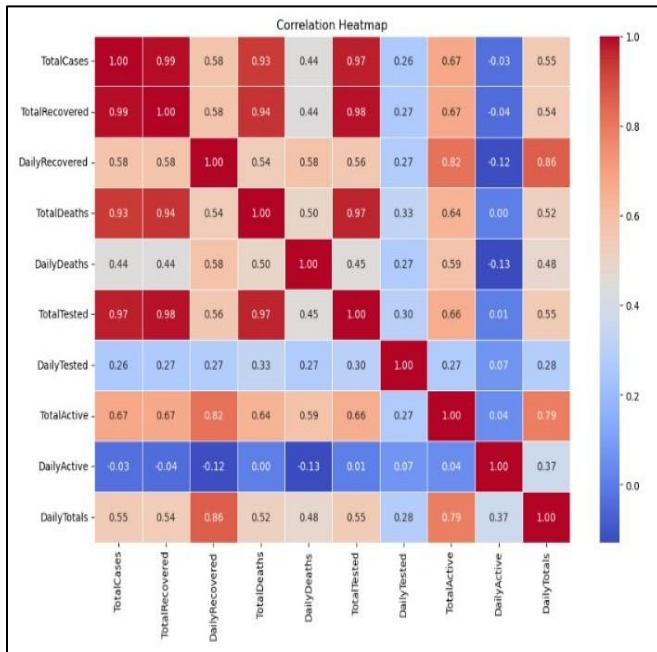
3. Data Exploration: In this part we work with the cleaned data and try to find valuable insights by performing various standards operations like:

   a. Univariate Analysis: This step includes to analyses each variable individually for this analysis we use summary statistics like mean, median, mode, standard deviation.

   b. Bivariate Analysis: This step includes exploring relationships between pairs of variables. For this we use scatter plots for continuous variables and box plots or violin plots for categorical variables against continuous variable. Here we also calculate correlation to measure the strength and direction of relationship.

   c. Multivariate Analysis: This step includes interaction between multiple variables. This uses techniques like heatmaps to visualize correlations between variables. Dimension reduction techniques could also be used like PCA (Principal Component Analysis) if dealing with high-dimensional data.

4. Data Visualization: Here we use appropriate visualization such as line plot, bar charts, heatmaps, etc. to represent the trends and patterns in the data effectively. By this, we can compare trends in different provinces or regions to get the data trend of COVID-19 cases over time across different provinces.

5. Hypothesis Testing: We formulate hypotheses about the data cases we collected and based on the observations during the exploration we test those hypotheses and validate them. For example, we created a hypothesis that more population regions generally would consist of more covid cases and found that hypothesis true during EDA.

6. Feature Engineering: Sometimes we need to create a new feature if necessary, such as calculating daily changes from cumulative data, deriving new metrics, or encoding categorical variables.

7. Insights and conclusion: Finally, we summarize the findings using the EDA process, identifying key insights, trends, correlations, and patterns in data and documenting them.

II. EDA

For EDA we will start with Multivariate Analysis where we compare multiple variables from the data set using the heatmap as shown in the picture below

The Heatmap is an effective way of depicting the co-relation between different variables, here the co-relation factors are scored between -1 to 1, where -1 refers to the factors that are less likely to co-relate with each other whereas 1 refers to the variables which are more likely to co-relate, and the decimals between them shows this strength of co-relation. This co-relation is also colored schemed to make it easier for us to judge the co-relation.

Then for the required columns having strong co-relation factors we create scatter plot charts for each of them against daily totals to create a better understanding of the dataset. The list of such columns is:

1. Province
2. Daily Totals
3. Summary Date
4. Total Cases
5. Total Recovered / Daily Recovered
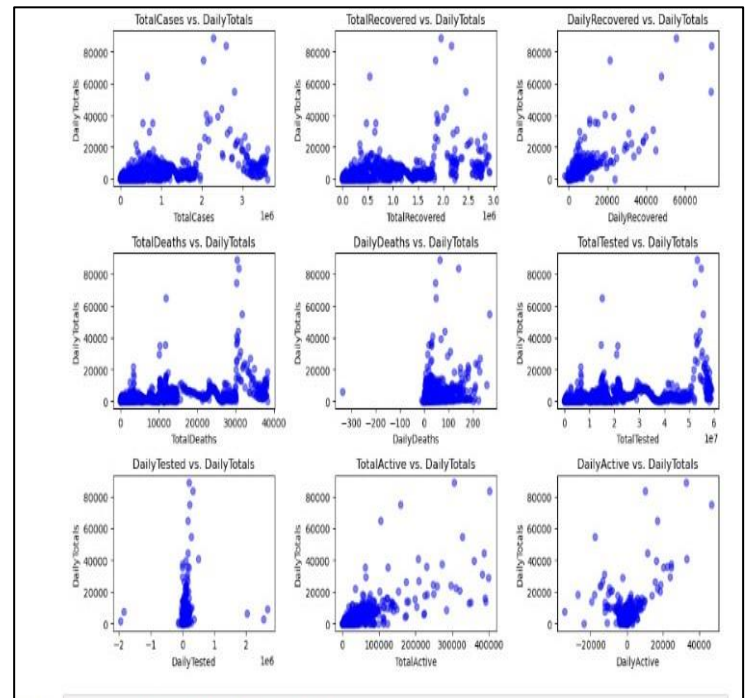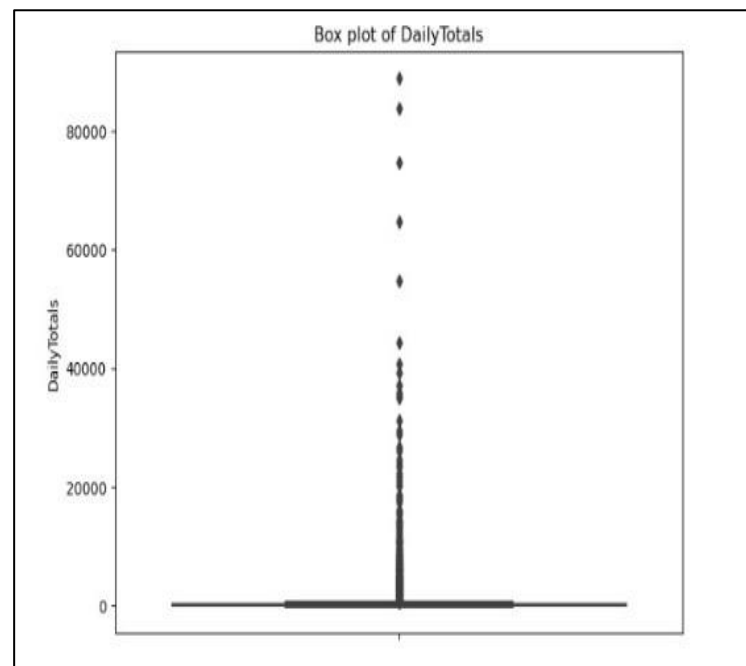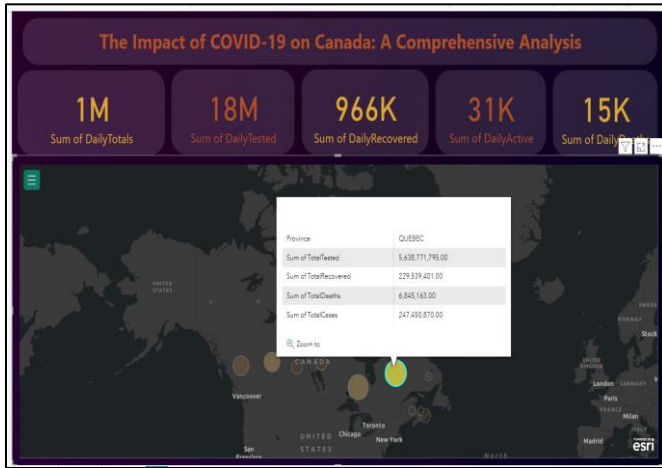6. Total Deaths / Daily Deaths
7. Total Tested / Daily Tested
8. Total Active / Daily Active

The above box plot shows the quarter of 2020 which has recorded the highest hit peak of recorded cases in Canada, estimated around 8.000 new cases per day.

This graph here, which is generated in Power BI with an interactive dashboard shows a provincial demographics of the cases occurred for the time period, where we can individually select the province and analyze the factors for each province.

This dashboard analysis gives us the names of the top 5 provinces which were affected most by the covid-19 incident which are:

1. Ontario
2. Quebec
3. Alberta
4. British Columbia
5. Manitoba.

   This observation here proves one of the hypotheses we have made earlier that the provinces with more population density have experienced more covid cases.

III. CONCLUSION

Some of the EDA conclusion we have drawn from the analysis of COVID-19 in Canada are:

1. Temporal Trends: Covid-19 cases in Canada exhibited distinct waves of infection, with peaks occurring at different points in time. Peaks often coincided with periods of relaxed restrictions and increased social interactions, highlighting the importance of ongoing vigilance and adherence to public health guidelines.

2. Regional disparities: There were significant regional disparities in the distribution of COVID-19 cases across provinces and territories. While some regions experienced higher case counts and more severe outbreaks, others managed to control more effectively through proactive public health measures and community engagement.

3. Age Distribution: Young adults and the working-age population were disproportionately affected by COVID-19, with higher case rates observed in these demographic groups. This underscores the importance of targeted outreach and messaging to address risk factors and promote adherence to preventive measures among younger populations.

4. Testing and Positivity Rates: Variations in testing and positivity rates were observed over time across regions, reflecting differences in testing strategies, access to testing, and the prevalence of COVID-19 in the population.

5. Vaccinated Progress: The rollout of COVID-19 vaccines led to significant progress in reducing case counts and severe outcomes.

REFERENCES

[1] X. Li, Y. Wu, Y. Yu, H. Guan, T. Zhou and Y. Fan, "Impact of population movement on the spread of 2019-nCoV in China," IEEE Transactions on Network Science and Engineering, vol. 7, no. 2, pp. 1068-1079, 2020.

[2] J. H. Chen, S. Lerman and R. Ferrara, "Covid-19: The First Public Coronavirus Twitter Dataset," arXiv:2003.07372 [cs], Mar. 2020.

[3] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[4] S. Seabold and J. Perktold, "Statsmodels: Econometric and Statistical

[5] Modeling with Python," in Proceedings of the 9th Python in Science Conference, 2010.