COVID-19 Vaccine Distribution Analysis

Foundations of Data Science

Objective

Analyze the "***COVID-19 case details dataset of Ontario"*** to determine how vaccine distribution will be prioritized for each group or category of people within the province.

Goal:

The goal of this project is to determine which group/category of individuals should receive the vaccine first amongst the different categories of people in various regions of each city in Ontario. The COVID-19 pandemic is a once-in-a-lifetime event and organizations around the world are pulling out the stops to get in front of the disease.

Questions to answer:

Some of the questions to answer are:

- *Which age groups by gender will receive the vaccine first?*
- *Which health region within Ontario will be the first to receive the vaccine?*
- *How does the case exposure type affect the number of cases amongst different ages?*
- *Which area in Ontario has the most number of cases based on the geographic locations?*

Hypothesis:

The rollout of the vaccine will be prioritized in health regions with the *highest active and total number of COVID-19 infections starting with elderly individuals over the age of 65* based on the idea that they have the highest death rate amongst other age groups. Also the health regions within the province of Ontario which has the highest number of active and total cases will be prioritized. Shipment deployments of a safe and effective vaccine will be prioritized based on where they will make the most impact and save lives. That means vaccinating our vulnerable seniors and those who care for them so that the virus will be under control.

Approach:

There are a lot of meaningful columns within the COVID-19 dataset that gives us important information into where and who should receive the vaccine first.

These columns include:

- Age Group
- Gender
- Exposure
- Case Status
- Health Region
- Longitude and Latitude

We will be analyzing the data within Ontario because most cases are known to be within this province and also many of the other provinces do not provide the necessary data to do an analysis

for the same. For example, the other provinces have a case status of "*Not Reported*". This data is critical to making a conclusion based on the hypothesis.

Our main approach for the analysis would be to focus on comparing different factors using the columns listed above and their correlation between each other. We will break down each column and analyze how the vaccine would be distributed to limit the spread of the virus and reduce mortality rate.

We will analyze the data and remove nulls values using different functions based on whether the data gives necessary importance. We will also present a map based on the geographic location of various health regions using the longitude and latitude columns to further look into the spread of COVID amongst different locations. Pivot table will be created to analyze the connections between the different columns. For example, health region vs the number of cases and the case status. Pie charts, bar graphs and plots will be created to analyze the exposure and gender to see who should receive the vaccination first.

Data Preparation

Data source:

Our dataset was downloaded from https://resources-covid19canada.hub.arcgis.com/datasets/compiled-covid-19-case-details-canada/data which is an open data source. As the coronavirus pandemic seems to be the most significant issue all over the world, we assumed that the distribution of vaccines in Ontario would be the best topic for data-analysis.

https://www150.statcan.gc.ca/t1/tbl1/en/cv.action?pid=1710000501 will also be used to back up the data of gender and death rates within different age groups.

There are over 90,000 records within the main dataset and the reported date of the cases are between January 2020 to September 2020 which is the most up to date dataset that is obtainable. Having a large number of records with dates ranging for 8 months can help with analysis of the data.

Data quality:

In terms of data quality, our dataset is considered trustworthy, reliable, accurate and consistent. This dataset is growing exponentially and it is updated on a regular basis showing province-level details about the confirmed cases including recovered and death counts. In order to ensure data quality, we assessed the contents of the data and its underlying structure to confirm that there are no questionable anomalies.

The dataset also includes all the important and necessary information, the most important being the gender, age group, health region and provinces. The quantity and diversity of the data will allow us to give a thorough analysis which can give an accurate conclusion to the hypothesis.

Tools/code used for analysis:

We utilized the following methods and functions to clean and add data to prepare it for analysis:

1. Studying the Data:

We studied data quality using the following:

a. Data overview: the head() function
b. Understanding content: the info() function
c. Missing values: isnull().sum() function
d. Understanding missing data: data_missing.nunique()

2. Cleaning data:

Removed missing data by:

a. Utilizing data.dropna() function
b. Utilizing data[~data['*Column_Name*'].isin(["*Removed_string*"])]
c. Substituting non-aligned values: pandas pd.replace function

3. Adding new columns/features:

a. Creating new columns for date data: pd.DatetimeIndex function
b. Columns were added to the health region pivot tables to further analyze the odds of survival and the severity zones. For the severity range a new function was created to calculate the different zones which includes the green, yellow and red zone. If-else conditions were used to calculate each zone based on the active cases and labeling as the different types of zones.

4. Organizing data:

Crosstab: the crosstab function was utilized to pivot data for building tables and for the creation of stacked bar charts in various analysis sections.

5. Visualizing data:

a. *Geopandas* was used in our analysis to plot the map of Ontario. This package combines geometry objects of shapely. Geometries are stored in a column called geometry that is a default column name for storing geometric information in geopandas. This allows us to easily manipulate, plot and analyze data.
b. Library, *adjustTex*t was added to the analysis so that text positions for names of various data points on matplotlib visualization plots are adjusted to remove or minimize overlaps with each other.

Challenges we faced?

Processing over 90,000 records, cleaning and refining the data set is a challenge we faced during the analysis. Since many records data have null values we had to look at the importance and the quantity of values that were null. For example, some longitude and latitude values were missing but after analyzing the missing values we concluded that it does not cause an impact on our overall data so we removed the null values within the dataset.

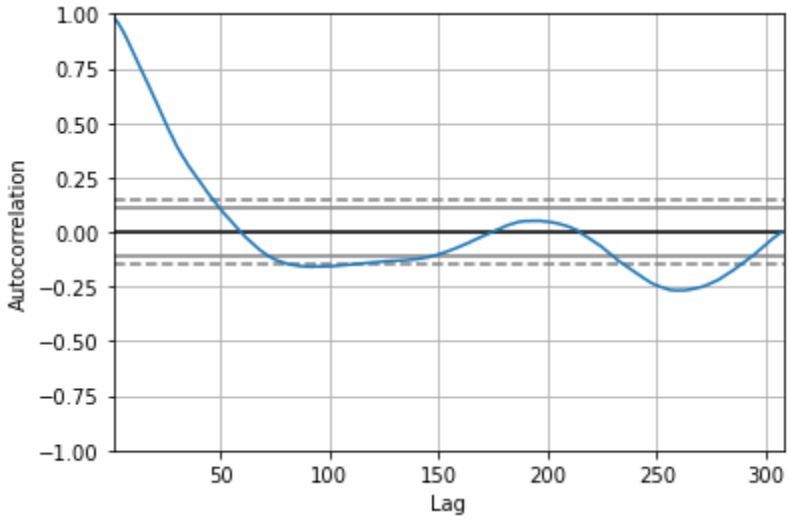
Another challenge we faced was making the necessary assumptions while analyzing the dataset. The main assumptions were:

a. Data analysed after cleaning and removing missing values is large enough and had significant sample size to represent the population of the entire dataset.
b. Data is not misrepresented and is accurately reported from the source.

Data Analysis

To understand the spreading of this pandemic we need to create a time series for our data to be able to put some guidelines for vaccine distribution priorities to control the outbreak; so we will focus on the “Active” cases along with the others.

Applying correlation for all cases and active notice we notice seasonal correlation and today we are in the second wave were we did not reach the peak as shown in the graph below:



Correlating all active cases we noticed that data is 93% correlated, this means the more cases we have the more active cases exist and vice versa. On the other hand, we have developed a forecasting model for next data and we noticed that the number of cases is still increasing and this supports the correlation graph that we have not yet reached the peak of the second wave.

a. Age Group

Analysing age groups is important to understand most vital age group to lower mortality rates; in the data set 44% of the records age groups are not reported, still analysing the 56% of available data gives us the following table:

Age Group	Percentage
80+	65.98%
70-79	19.66%
60-69	9.6%
50-59	3.38%
40-49	0.86%
30-39	0.30%
20-29	0.18%
0<20	0.03%

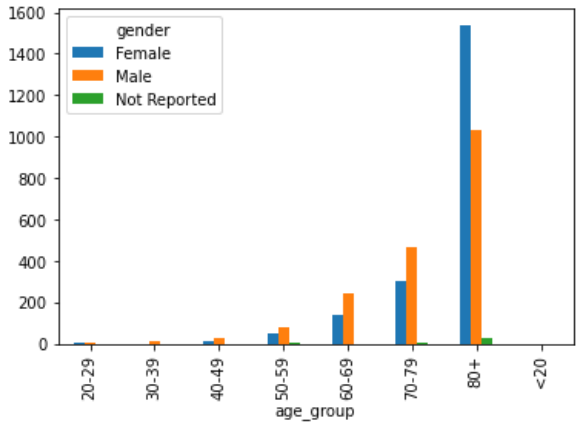
The data shows that 98.63% of mortality rates are in the age group of 60 and above, therefore vaccination should focus on people with age groups above 60 to lower the mortality rates in the province.

b. Gender

Analysing the gender column is important to understand most vital gender group to lower mortality rates; similar to age groups in the data set 44% of the records for gender are not reported, still analysing the 56% of available data shows us that *52% of deaths are females while 48% are males*. This means that being a female increases the probability of dying over a male by 8%.

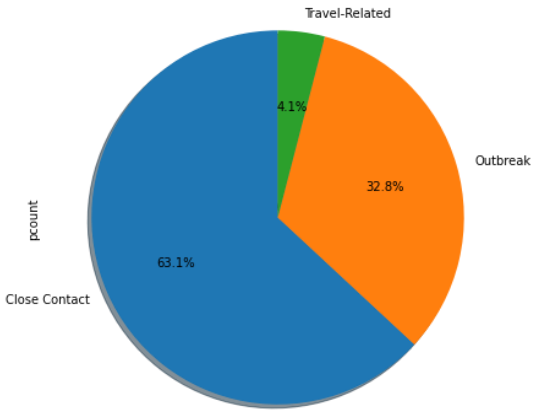
In all age groups analysed above, there are more male deaths compared to females. However, in the 80+ group, the trend is the other way around. To understand that trend, we did analysis on the male and female population in Ontario for different age groups (Table #6.2.7). Female population in the 80+ age group is significantly higher than the male population. Hence, we can conclude that male population is at a higher risk of death.

Diving further into *gender and age groups* we find that females overseed males for age group above 80 only while in all other age groups males mortalities brevales femalis.

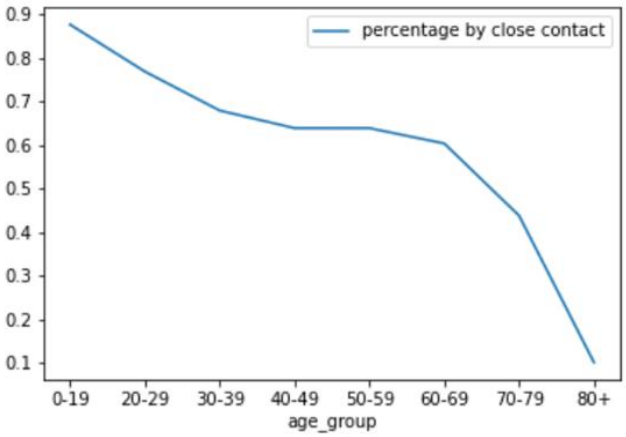


c. Exposure

The exposure data breaks down to four main categories, *1) Close Contact, 2) Travel Related, 3) Community Outbreak, 4) Not Reported*. The data has been cleaned to exclude the non-reported data from the exposure analysis. As it can be observed from the table of section 6.4.7, there is no significant impact by the exposure type of the virus and the likely fate of the patient indicating that the status of individuals impacted by the virus is independent of the exposure type. This is likely due to the independence of exposure type to the impact the virus has on the human body.



The exposure type was assessed with its relation to age group to identify if a certain age group is associated with an exposure type more than others. Based on the chart above, also shown in section 6.4.8, it can be concluded that the majority of the cases are related to close contact encounters with around 63% belonging to that category. It has been also determined that the exposure type, mainly focusing on the close contact category (as it is the main driver of the exposure), varies with the age group. As shown in the graph below, also shown in table of section 6.4.9 it can be concluded that the younger the age, the higher is the exposure percentage of total by close contact.



The exposure by close contact in the age group below 19 is at around 86% and it declines with age group members of the population with age 80+, showing a significantly lower percentage of exposure by close contact at around 15%. This could indicate that the younger population is having higher mobility in the community and thus, the younger age

groups could potentially be the main drivers of the virus via the exposure type of close contact. The age group of 20-29 shows the highest level of exposure overall by all types, this is due to the young age factor combined with being within the working class which increases likelihood of exposure.

The exposure type was assessed further by looking at the gender interaction with the likelihood of being exposed to the virus by one method or another. As shown in the graphs of sections 6.4.12-14, it can be concluded that the exposure type is not impacted and is independent of gender where female or male. The graphs also further indicate that the age group classification within the different exposure types is also independent of gender.

The exposure analysis indicates that age group has a significant impact not only on survival rate as indicated by analysis of other sections of the report, but by also exposure and likelihood to carry the exposure to others within the community. The approach for vaccination of citizens should take into consideration the younger population's impact in spreading the virus via close contact with others.

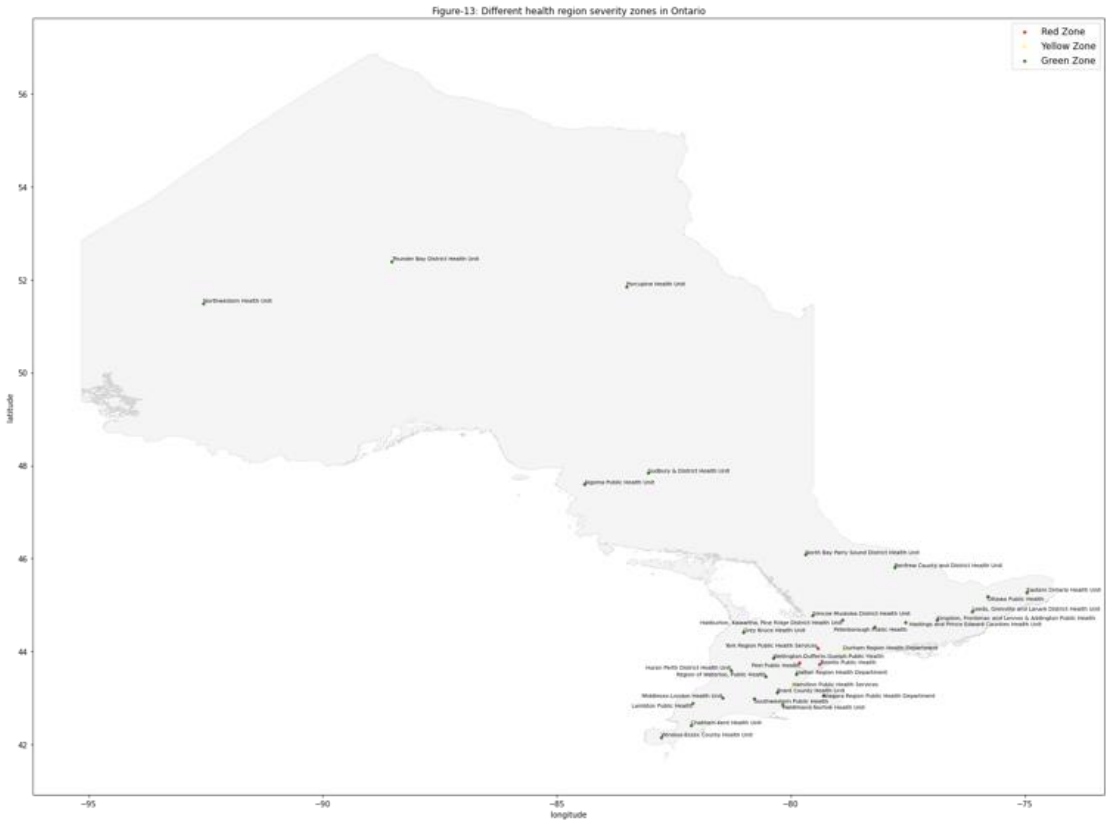
d. Health Region and Case Status

health_region	Active	Deceased	Recovered	Total Number of Cases	Odds of Survival	Severity Zone
Toronto Public Health	4935.0	1640.0	36388.0	42963.0	95.687388	Red Zone
Peel Public Health	4526.0	383.0	23318.0	28227.0	98.384034	Red Zone
York Region Public Health Services	1330.0	324.0	10273.0	11927.0	96.942531	Red Zone
Durham Region Health Department	639.0	192.0	4101.0	4932.0	95.527603	Yellow Zone
Hamilton Public Health Services	585.0	94.0	2891.0	3570.0	96.850921	Yellow Zone

Analyzing the health region is important to determine who should receive the vaccine first. Distributing the vaccine to the region with the highest total number of cases and active cases can limit the spread of the virus and reduce the mortality rates.

In the dataset there are many health regions and the approach used in analyzing the data is to create a pivot table with the health region as the index, case status as the columns and the count of the case status as the values.

Looking at the pivot table under the *Health Region section* in the Jupyter Notebook that contains the health region and case status it can be seen that the region with the highest number of active cases is the **Toronto Public Health Region** at 4935 cases. The Toronto Public Health Region also has the highest number of total cases at 42963 cases. The total number of cases is created by adding the number of active, deceased and recovered cases.



The **Severity Zone** column was created as an additional feature column to analyze and rank the regions based on the number of active cases with over 1000 active cases being in the red zone, 500 to 1000 active cases being in the yellow zone, and finally under 500 active cases being in the green zone. The map below shows us all health regions categorized as Red, Yellow and Green zones. As you can see, since the Toronto Public Health Region has 4935 cases, it is within the ***Red Zone***.

case_status health_region	Active	Deceased	Recovered	Total Number of Cases	Odds of Survival	Severity Zone
Leeds, Grenville and Lanark District Health Unit	34.0	53.0	434.0	521.0	89.117043	Green Zone
Porcupine Health Unit	8.0	9.0	99.0	116.0	91.666667	Green Zone
Haliburton, Kawartha, Pine Ridge District Health Unit	38.0	21.0	284.0	343.0	93.114754	Green Zone
Lambton Public Health	21.0	25.0	381.0	427.0	93.842365	Green Zone
Haldimand-Norfolk Health Unit	34.0	37.0	585.0	656.0	94.051447	Green Zone

The Odds of Survival column shows that the "***Leeds, Grenville and Lanark District Health Unit***" has the lowest odds of survival at 89.11%. With only 434 total cases within the "Leeds, Grenville and Lanark District Health Unit", it should not be used to come to any sort of conclusions within the health region analysis. Also, the five regions with lowest odds of survival have 29 or lower active cases and in the green zone.

Overall, ***Toronto Public Health Region*** is the first region that should receive the vaccine as it is in the Red Zone having 4043 active cases and has the highest number of 31,994 total cases in the province of Ontario.

Conclusion

Based on the analysis, the age group above 60 should receive the COVID-19 vaccination and should have top priority compared to the various age groups. People above 80 should receive the vaccine first based on death rates being the highest in that category. Male elders with ages between 60 and 80 should receive the vaccine after that as it has the second highest death rate. We believe age groups 20-50 should receive the vaccine after the people of age group between 60 and 80. This is based on the assumption that the age groups being 20-50 who represent most of the students/working class are the most active which increases their total number of cases. Finally children should receive the vaccination last with the least number of cases and death rates. The lower death rates could be because of their lower exposure when staying indoors.

Based on exposure type analysis, to limit the spread of the virus, the younger population, particularly <29 should be targeted. But since mortality rates for the younger generation are negligible compared to the older generation, the vaccination should be focused on the older generation to limit mortalities by the virus.

Looking at the health region where the elderly people should receive their vaccinations first, it would be the ***Toronto Health Region***. Within health regions, Toronto Health Region has the highest active cases as well as the highest total number of cases. Though it does not have the highest death rate we believe with the higher total number of cases, it should be the first location in which the elderly people should receive their vaccination first.

Overall, our hypothesis was accurate where we stated that elderly people should receive the vaccination first within the region with the highest number of active and total cases. The in-depth analysis showed a thorough understanding of how the vaccinate should be distributed amongst different age groups, genders, health regions and locations.