CSI4142
Winter 2023

Tracking covid exposure in different provinces in Canada

Group ID (21)
Fah, Fatimetou
Ogunfowora, Opemipo
Tiendrebeogo, Othniel

Professor:

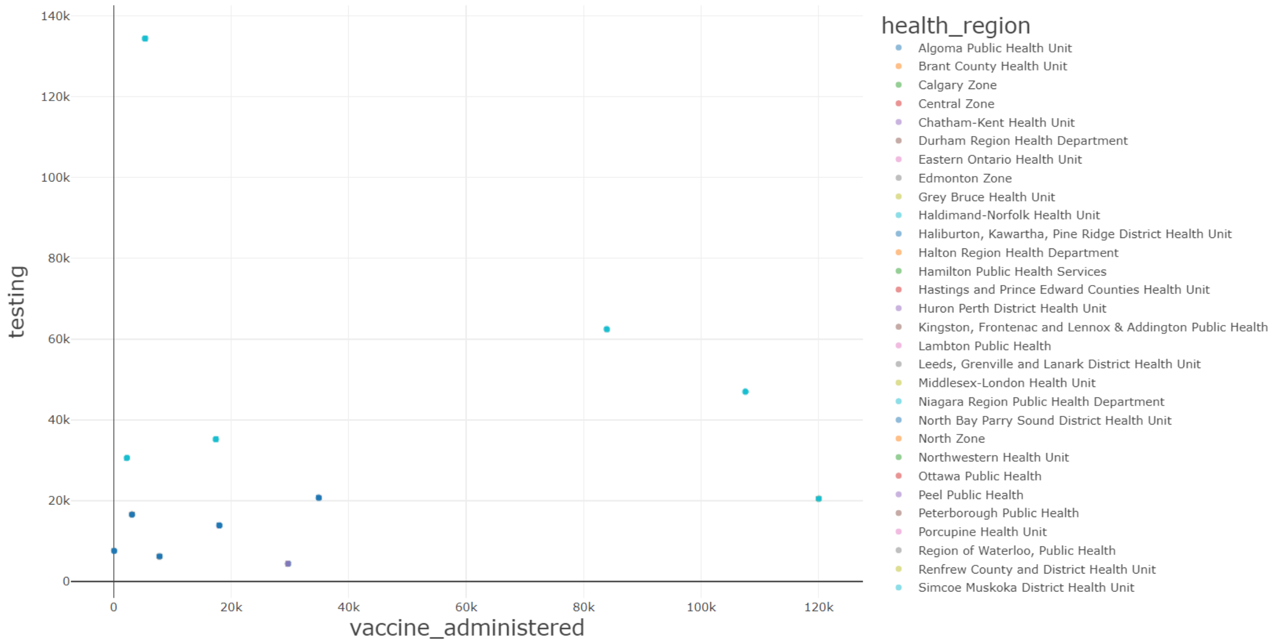
Yazan Otoum

Phase 4: Data Mining

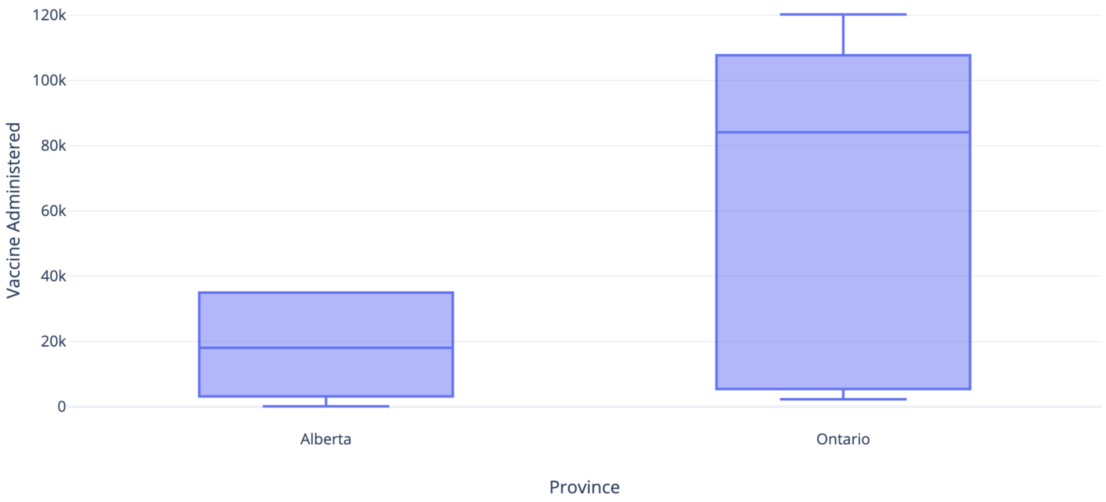
Part A. Data summarization, data preprocessing and feature selections:

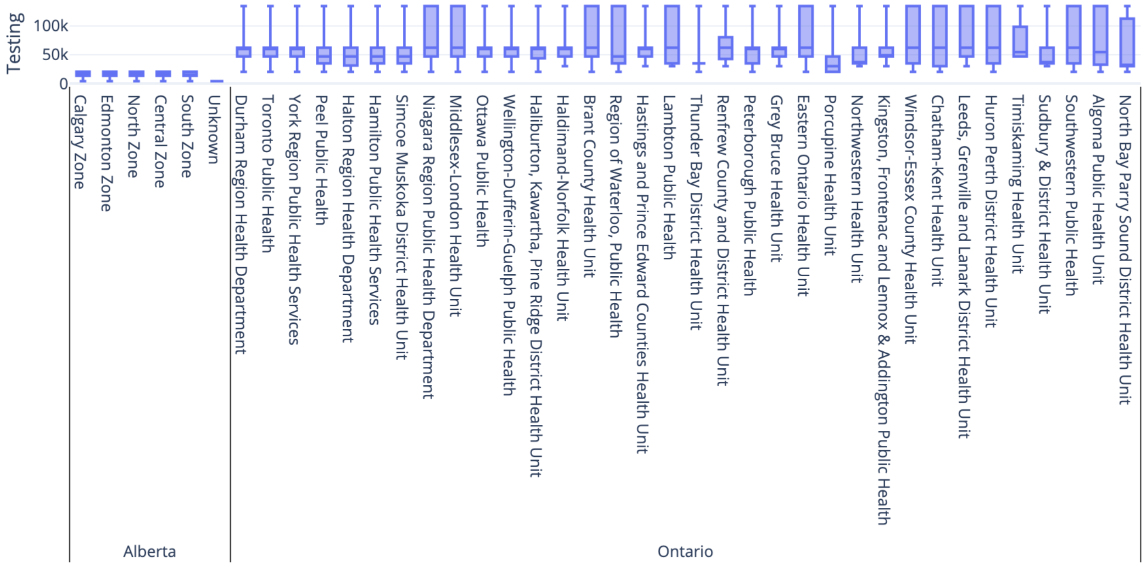
1.

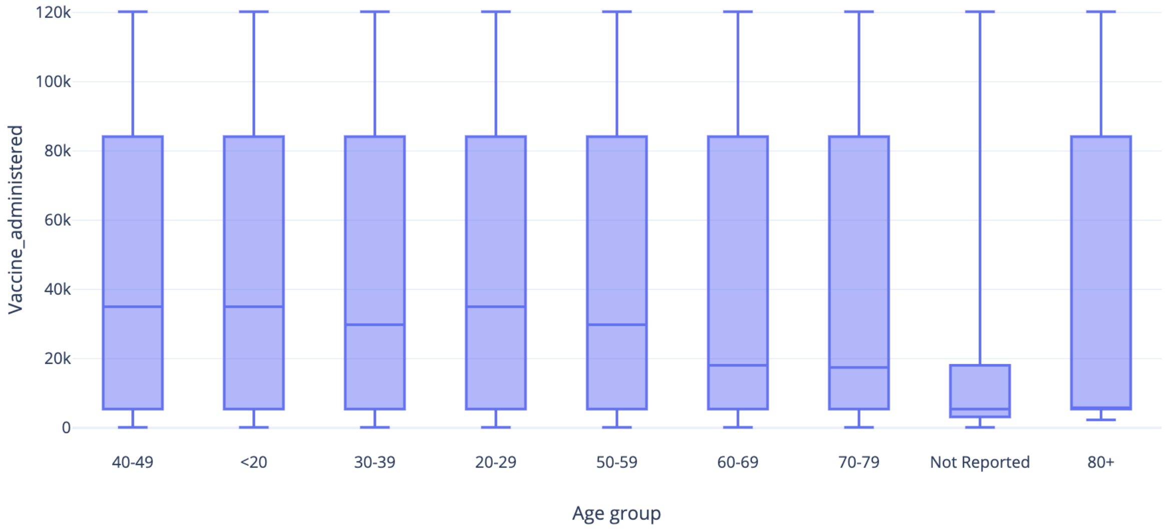
- Scatter plots:



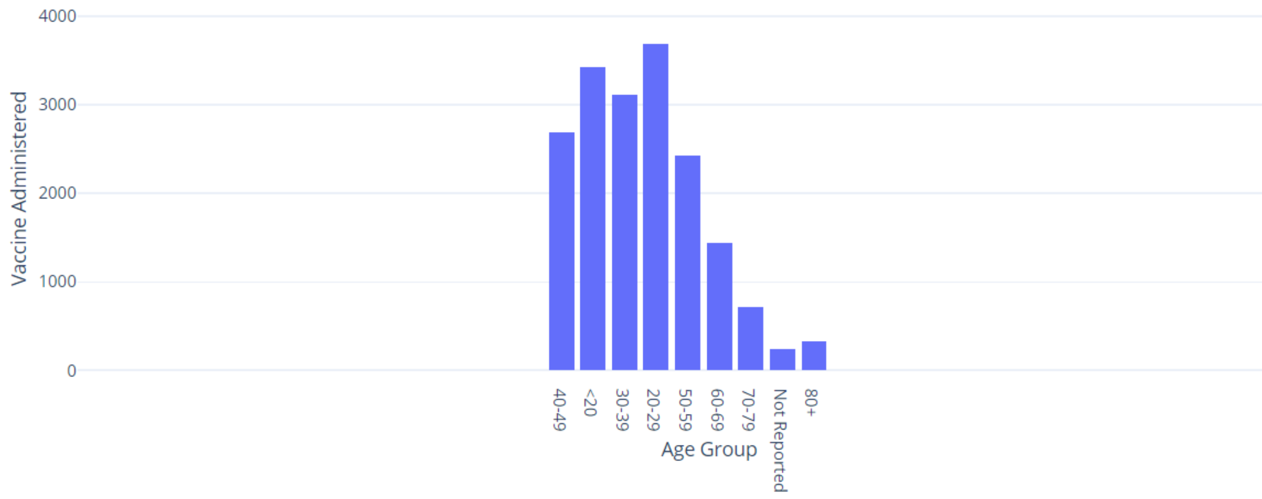
- Boxplots:




Testing
100k
50k
0
Calgary Zone
Edmonton Zone
North Zone
Central Zone
South Zone
Unknown
Durham Region Health Department
Toronto Public Health
York Region Public Health Services
Peel Public Health
Halton Region Health Department
Hamilton Public Health Services
Simcoe Muskoka District Health Unit
Niagara Region Public Health Department
Middlesex-London Health Unit
Ottawa Public Health
Wellington-Dufferin-Guelph Public Health
Haliburton, Kawartha, Pine Ridge District Health Unit
Haldimand-Norfolk Health Unit
Brant County Health Unit
Region of Waterloo, Public Health
Hastings and Prince Edward Counties Health Unit
Lambton Public Health
Thunder Bay District Health Unit
Renfrew County and District Health Unit
Peterborough Public Health
Grey Bruce Health Unit
Eastern Ontario Health Unit
Porcupine Health Unit
Northwestern Health Unit
Kingston, Frontenac and Lennox & Addington Public Health
Windsor-Essex County Health Unit
Chatham-Kent Health Unit
Leeds, Grenville and Lanark District Health Unit
Huron Perth District Health Unit
Timiskaming Health Unit
Sudbury & District Health Unit
Southwestern Public Health
Algoma Public Health Unit
North Bay Parry Sound District Health Unit
Alberta
Ontario


120k
100k
80k
60k
40k
20k
0
Vaccine_administered
40-49
<20
30-39
20-29
50-59
60-69
70-79
Not Reported
80+
Age group

- Histogram



Based on our data summarization, The data points on the scatter plot graph are distributed around the same area for health regions that are in the same range in terms of population. This is showing that regions will share characteristics such as numbers of testings and the numbers of vaccines administered when they are in the same population range. This also explains the outliers as they can be as far away as possible from other data points if the population in that health region is a lot larger than the others. This also implies that a higher population will probably result in more testing and vaccines administered which will be plotted on the scatter plot higher when compared to regions with a lower population.

2.

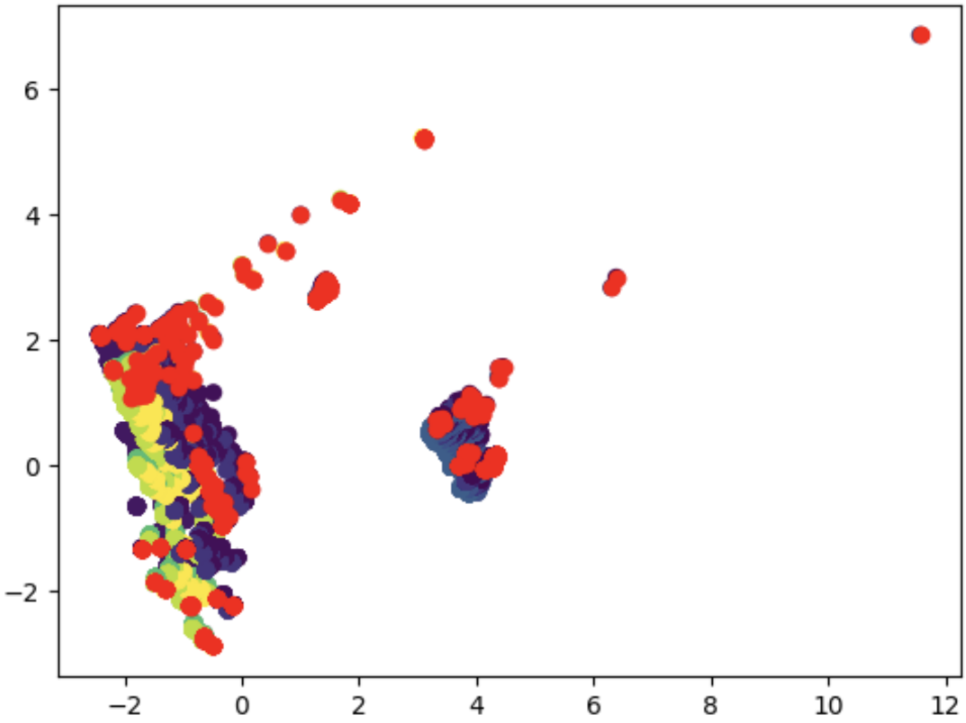
- Handling missed values: To handle the missing values in our datasets, we replaced it with “Not Reported” or “Unknown” for columns with type string. Otherwise, during the data staging phase, we handled all missing values.
- Handling categorical attributes: We handling categorical attributes by converting them to numerical data through one-hot encoding.The age_group attribute was originally a categorical attribute with values such as 20-29, 30-39, 40-49, etc. After one-hot encoding, the age_group attribute was transformed into multiple binary columns with values of either 0 or 1, indicating whether a given row belongs to a specific age group.
- Feature selection to remove potentially redundant attributes: Feature selection was performed by dropping the cumulative_avaccine column from the df_vaccine dataframe . This column was considered redundant since it provides cumulative vaccine data, which is already captured in the vaccine_administered column. Add to that we dropped cumulative_testing and testing_info attributes from the testing dataset since they were not relevant to the analysis. Additionally, the 'ObjectId' attribute was removed from the final merged dataset since it was not necessary for analysis.

- Normalization of numeric attributes to ensure all attributes are of equal, importance during learning: The vaccine_administered and testing columns were on different scales and units. To address this, they were both divided by the total population of the dataset to make them comparable. This process ensured that each attribute carried equal importance during the learning process.

Part B. Classification (Supervised Learning):

For this part, we couldn't really do it because our dataset doesn't have any Machine Learning components so no scope of implementing it.

Part C. Detecting Outliers: (Bonus)



In summary, the One-Class SVM algorithm is designed to identify outliers in a dataset, which are data points that are significantly different from the rest of the data. In the context of our dataset, one of the reasons that outliers are present in our dataset can be due to the correlation between higher population, higher testing, and higher vaccine administration. This can lead to outliers in the data because there are data points that deviate significantly from the overall pattern of the data. For example, a region with a very high population may have a much higher number of tests and vaccines administered than other regions with similar populations, which could be considered an outlier.

Moreover, the algorithm's ability to identify outliers provided an opportunity to investigate the regions with significantly higher or lower numbers of tests and vaccines administered. By investigating these regions, it is possible to identify underlying factors that may be contributing to these differences. For example, regions with lower numbers of tests and vaccines administered may require additional resources to improve their healthcare systems. Similarly, regions with significantly higher numbers of tests and vaccines administered may have unique characteristics that require further investigation to understand the underlying factors that contribute to these differences. Some potential factors that may contribute to these differences include population density, age distribution, income levels, previous disease outbreaks or cultural/religious practices.

The One-Class SVM algorithm helped identify these outliers by learning a decision boundary that separates the "normal" data points from the outliers. The support vectors, which are the data points closest to the decision boundary of the One-Class SVM, can be visualized to see which points are considered "normal" in the dataset. By identifying the outliers and removing or investigating them further, we can improve the quality and reliability of our analysis.

Tools used for data summarization:

- https://www.csvplot.com/
- https://chart-studio.plotly.com/create/box-plot/#/