INF412 Data Analytics

Beyond the Infection Rate: Unveiling Key Factors Influencing Public Response to Health Events: Taking COVID-19 Pandemic in Toronto as An Example

Student name & ID

Julie Nguyen - 1009596475
Linrong Li - 1006697969
Missy Zhang - 1008620261
Myra Li - 1008439101

Table of Contents

1 Introduction

The COVID-19 pandemic, which emerged in late 2019, has led to a dramatic loss of human life worldwide and presents an unprecedented challenge to public health, food systems, and the world of work (Chriscaden, 2020). The sheer scale of this pandemic has compelled various divisions of healthcare institutions to take immediate and decisive action. The impact of the pandemic extends beyond the immediate health crisis, affecting economies, social structures, and the mental well-being of populations globally.

Predicting the next major outbreak and its underlying causes is a complex and daunting task. However, analyzing past pandemics and their key factors can provide valuable insights into managing and intervening in diseases more effectively. This historical analysis can aid health centers in developing more efficient and targeted plans when faced with the next pandemic, potentially saving countless lives and mitigating widespread disruption.

Central to all containment policies is the public's response to the crisis. Public behavior and attitudes play a crucial role in the success or failure of public health strategies. Therefore, understanding how the public reacts to a pandemic, including factors such as compliance with health guidelines, acceptance of vaccination, and adherence to social distancing measures, is crucial for devising effective management strategies.

This leads to our research question: What are the key factors influencing the public's response to health events, beyond the infection rate? This question aims to uncover the underlying dynamics that shape public behavior during health crises, which is essential for crafting policies that are not only scientifically sound but also socially acceptable and effective.

In this paper, we will explore two dimensions of public response, medical and social support in response to COVID-19, and how they interact to shape the overall reaction to pandemics. By understanding these factors, policymakers can design more nuanced and effective strategies that resonate with the public, thereby enhancing the efficacy of public health interventions and ultimately leading to better outcomes in managing future pandemics.

2 Review of Similar Research

Our secondary research aims to identify the key factors that influence the public's response to health events, specifically on the COVID-19 pandemic. We reviewed existing literature to understand the dynamics of public health response and the social determinants of health that play a crucial role in shaping these responses.

Mohammadpour et al. (2021) conducted a comprehensive study on the readiness and responsiveness of healthcare systems during the COVID-19 pandemic. They identified five main factors that significantly impact the effectiveness of healthcare systems in managing the crisis: community-related interventions, managerial interventions, socioeconomic factors, the readiness of hospitals and health centers, and environmental factors.

Nwakasi, Esiaka, Uchendu, & Bosun-Arije (2021) explored the multifaceted impact of the COVID-19 pandemic on the African continent, shedding light on the challenges and responses across various sectors. Their findings highlighted that non-compliance with public health recommendations often stemmed from economic and religious reasons, underscoring the complex interplay of factors influencing public behavior during a health crisis.

In the context of Canada, we aligned our analysis with the social determinants of health outlined by the Government of Canada (2024). These determinants include income and social status, employment and working conditions, education and literacy, childhood experiences, physical environments, social supports and coping skills, healthy behaviors, access to health services, biology and genetic endowment, gender, culture, and race/racism.

Based on the previous research, we decided to include the following factors in our analysis to fill in the gap of understanding the public's response to the COVID-19 pandemic: median income, education level, median age, population density, employment rate, average household size, infection rate and the presence of a COVID-19 response shelter and immunization clinics (IFS). These factors will be crucial in understanding the nuanced interactions between social determinants and public health responses during the COVID-19 pandemic.

3 Data

3.1 Data Sources

https://open.toronto.ca/dataset/covid-19-cases-in-toronto/

- This dataset includes the identified COVID-19 cases in Toronto from 2020 to 2024
- It covers variables such as age groups, living neighborhood, infection source, and reported date

https://open.toronto.ca/dataset/neighbourhood-profiles/

- This dataset is the toronto census data by neighborhood
- It covers socio-economic factors such as income level, education level, and marital status.

https://open.toronto.ca/dataset/daily-shelter-overnight-service-occupancy-capacity/

- This dataset contains the location of COVID response shelter available in 2021, which can be viewed as a source of social support.

https://open.toronto.ca/dataset/covid-19-immunization-clinics/

- This dataset contains the location of COVID immunization clinics built in 2021, which can be viewed as a source of medical support.

3.2 Strengths and weaknesses of the dataset

The datasets provided by the Toronto open data are both reliable and authoritative, ensuring a high level of credibility and trustworthiness that is essential for informed decision-making and precise analysis.

However, they do exhibit certain weaknesses. One notable limitation is the restricted range of socio-economic indicators included. In addition, there are concerns regarding the completeness and accuracy of the data, especially pertaining to the COVID-19 case data sourced from Toronto Public Health. The reliability of this data is potentially undermined by the common use of home testing kits, as a significant number of individuals fail to report positive results. This leads to a likely underreporting of the actual number of cases in the official records, which is a challenge to the integrity of the data. Furthermore, the open datasets on COVID response shelters and immunization clinics we selected for this analysis provide valuable information about certain aspects of social and medical support available in Toronto during 2021. However, it's important to recognize that these datasets represent only a portion of the vast landscape of services, resources, and factors that contribute to the overall well-being and support systems within the city. While these datasets offer insights into specific areas of social and medical assistance, they may not fully capture the complete picture or represent the intricate network of support available to individuals and communities.

3.3 Ethical considerations

From an ethical standpoint, the use of the data posed questions about representation. When using neighborhood-level measures for correlation analysis, one ethical concern is the risk of ecological fallacy, which is the assumption that characteristics at the aggregate level can be directly applied to individuals within that group, leading to potential inaccuracies and misinterpretations.

In our study, correlations between socio-economic factors and COVID-19 response rate at the neighborhood level cannot be automatically assumed that these relationships hold true for every individual within that neighborhood. Individuals may not conform to the average characteristics of their neighborhood and that specific circumstances and variations exist within any given group. Thus, it is crucial not to draw definitive conclusions about individuals based solely on aggregated data, which may lead to oversimplification and reinforce stereotypes.

3.4 Descriptive analysis

3.4.1 Dataset description

The whole dataset we built contains data from 140 neighborhoods in Toronto. Each observation has 7 predictor variables (Population density, Median income per person, Employment rate, Proportion of people with a Bachelor's degree or higher, Median age, Average household size, Infection rate) and 3 target variables (COVID-19 response shelter, Immunization Clinics, IFS).

After performing initial data cleaning and exploration, we did not identify any null or duplicate values in the dataset. Regarding outliers, we have decided to retain them in our analysis rather than removing them. This decision is based on two key considerations. Firstly, we are working with a relatively small dataset, and removing outliers could potentially lead to a significant loss of valuable information. Secondly, socio-economic factors often exhibit a wide range of values across different neighborhoods, and what may appear as outliers could, in fact, represent the inherent diversity and disparities present in the data. By including these outliers, we aim to capture and analyze the full spectrum of socio-economic conditions, ensuring a comprehensive understanding of the relationships and patterns within the dataset.

3.4.2 Target variable

COVID response shelter

This variable comes from shelter occupancy data. We filtered the data in 2021 and removed duplicate addresses. Based on whether there is a response shelter in the neighborhood, we coded the variable as a binary variable. We used this variable to represent the **social support** in response to the COVID outbreak.

Frequency table shows that out of the 140 observations, 122 (87.1%) have a value of 0, indicating no COVID-19 response shelter, while 18 (12.9%) have a value of 1, indicating the presence of a COVID-19 response shelter.

	Count	Percentage
0	122	0.871429
1	18	0.128571

Immunization clinics

Based on the GeoJSON data of immunization clinics, we counted the number of points in each neighborhood in QGIS. Then, we coded it as a binary variable based on whether there is an immunization clinic in the neighborhood. We used this variable to represent the **medical support** and response after the COVID outbreak.

Frequency table shows that out of the 140 observations, 12 (8.6%) have a value of 0, indicating no COVID-19 immunization clinics, while 128 (91.4%) have a value of 1, indicating the presence of at least one COVID-19 immunization clinic.

	Count	Percentage
0	12	0.085714
1	128	0.914286

IFS (The presence of COVID-19 response shelters and immunization clinics)

Based on the previous two variables, we created a new variable that combines the presence of a COVID-19 response shelter and immunization clinics called IFS. It assigns each neighborhood a value of 0, 1, or 2 depending on the availability of these facilities.

A value of 0 indicates that a neighborhood has neither a COVID-19 response shelter nor any immunization clinics. A value of 1 is assigned to neighborhoods that have either a COVID-19 response shelter or immunization clinics, but not both. A value of 2 represents neighborhoods that have both a COVID-19 response shelter and immunization clinics.

The frequency table shows that the majority of neighborhoods (79.29%) have either a COVID-19 response shelter or immunization clinics, while a smaller proportion (12.86%) have both facilities. Only a small percentage of neighborhoods (7.86%) lack both types of facilities.

	Count	Percentage
0	11	0.078571
1	111	0.792857
2	18	0.128571

3.4.3 Predictor variable

Table 1 Descriptive Summary of numerical variables

	Education Level	Population Density	Median Income per Person	Employment rate	Median age of the population	Average household size	Infection Rate
Count	140	140	140	140	140	140	140
Mean	0.34716	0.62677	41840.47619	0.545047	40.920952	2.478929	0.061881
std	0.132721	0.521928	10160.32218	0.063016	3.870069	0.375321	0.017925
Min	0.096713	0.10111	28400	0.402	30.4	1.55	0.031885
25%	0.241785	0.344397	34000	0.502	38.8	2.2375	0.048562

50%	0.337578	0.501785	38600	0.537	40.6	2.5	0.059507
75%	0.446291	0.742614	46400	0.5925	43.2	2.7	0.0746
Max	0.673323	4.367895	74500	0.775	50.8	3.35	0.114881

Education Level (Proportion of people with a Bachelor's degree or higher)

The data reveals an average educational attainment of 34.72%, which suggests that on average, about one-third of each neighborhood's population has obtained higher education. The median, at 33.76%, indicates a moderate left skew in the data, where more neighborhoods fall just below the average in terms of higher education attainment.

The histogram (see Figure 1) demonstrates a normal-like distribution with a slight concentration in the 20-40% range, while the box plot shows that the middle 50% of neighborhoods have an education level ranging from approximately 24.18% to 44.63%. There are some neighborhoods with exceptionally high education levels, with the highest being 67.33%. This factor's near-normal distribution, albeit slightly left-skewed, indicates that while some neighborhoods have high levels of education attainment, there are areas with significantly lower education levels.

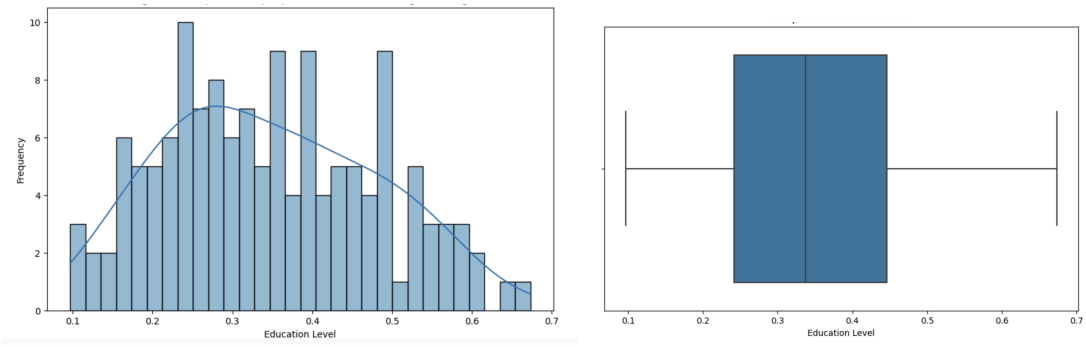

Figure 1 Distribution of people with a Bachelor's degree or higher

Population Density

The investigation into population density shows an average of 0.6268 individuals per 0.01 hectare across the neighborhoods. This factor exhibits a significant right skew, with the histogram displaying most neighborhoods having a lower density and a tail extending towards higher densities (see Figure 2).

The box plot (see Figure 2) accentuates that the middle 50% of the neighborhoods have densities that range from 0.3444 to 0.7426, with a few neighborhoods showing considerably higher densities. These dense neighborhoods, reaching up to a density of 4.3679, are marked as outliers, but won't be excluded from our research.

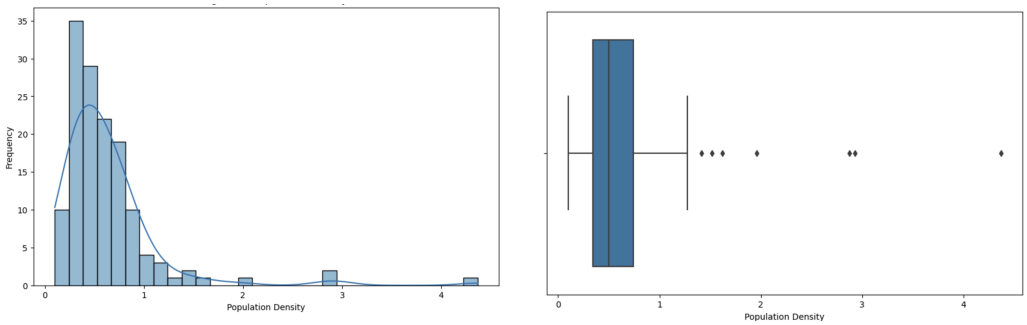

Figure 2 Distributions of Population Density

Median Income

The dataset shows a mean median income of $41,840.48 and a median income of $38,600, which indicates that half of the neighborhoods have lower incomes, pointing to a distribution that is skewed to the right.

This skewness is more evident in the histogram (see Figure 3), where income is predominantly spread between $30,000 and $50,000. The box plot further illustrates that the interquartile range extends from $34,000 to $46,400, encapsulating the middle 50% of the data. Notably, there are outliers with incomes at $67,500.00, $72,000.00, $74,500.00, $68,000.00, and $67,500.00, significantly surpassing the median income levels and suggesting a degree of economic disparity within the sample.

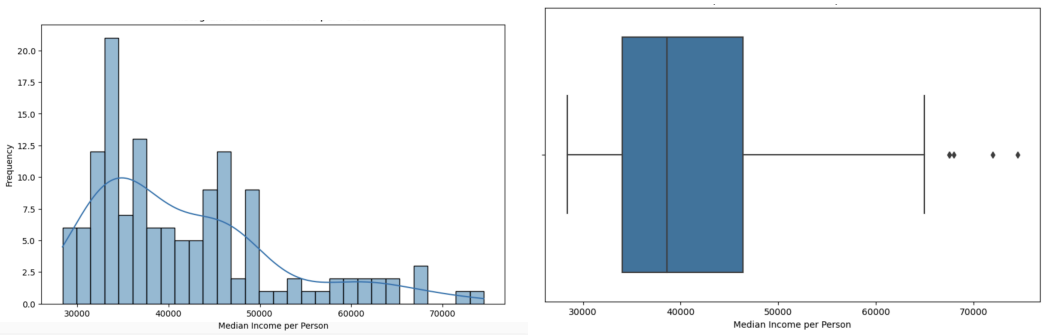

Figure 3 Distributions of Median Income Per Person

Employment Rate

The employment rate in the City of Toronto recorded a mean of 0.55. The median employment rate among 140 neighborhoods is 0.53. Since the mean and the median are not significantly

different, we may assume that this portrays a fairly normal distribution. Looking further into the histogram and boxplot (see Figure 4), the data of employment rate shows small numbers of outliers.

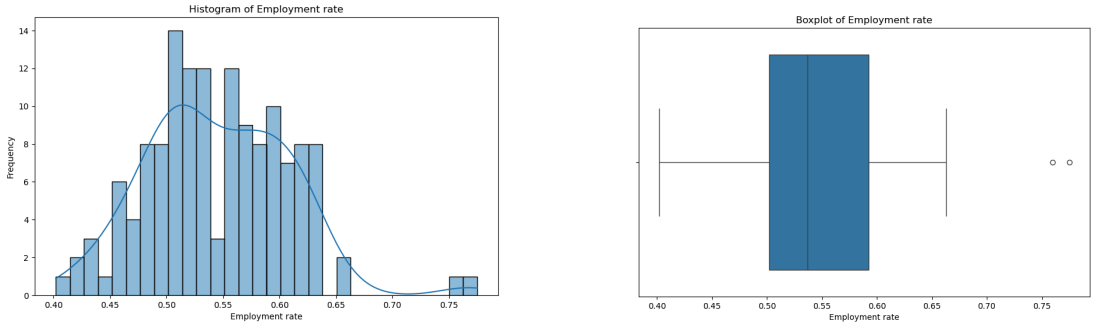

Figure 4 Distributions of Employment Rate

Median Age of the Population

In reviewing the median age of the population, the dataset presents a mean age of approximately 40.92 years, suggesting that the Toronto neighborhoods have a relatively mature population. The median age is close to the mean, at 40.60 years, hinting at a symmetric distribution in the population's age.

This symmetry is confirmed in the histogram (see Figure 5), which shows a balanced spread around the median value. The box plot discloses that the middle 50% of the neighborhoods' median ages are clustered between 38.80 and 43.20 years. The age distribution does not indicate significant outliers.

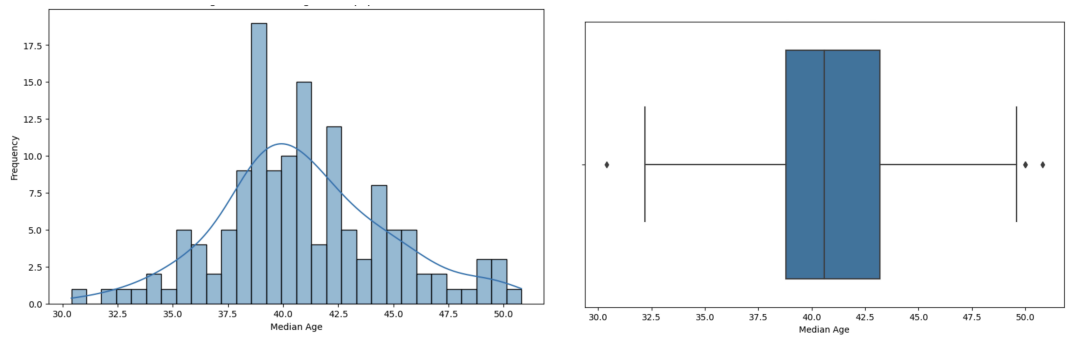

Figure 5 Distributions of Median age of the population

Average Household Size

The average household size is reported to be 2.48 people per household, which is close to the median of 2.5. The histogram of average household size shows a fairly normal distribution (see Figure 6), meanwhile the boxplot shows no outliers.

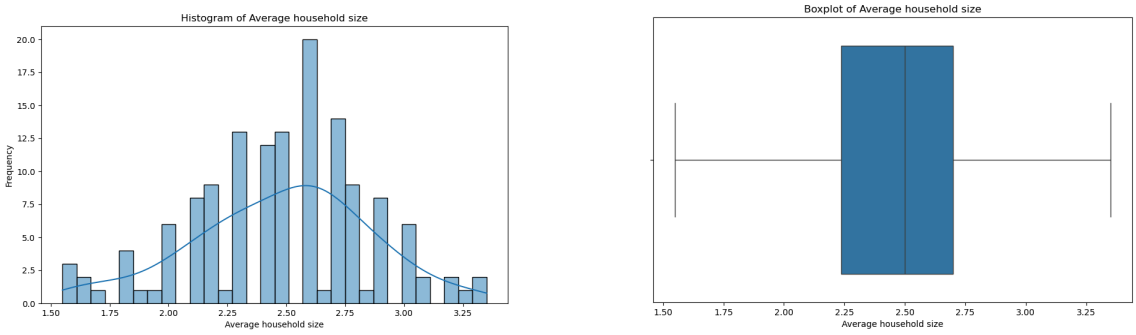

Figure 6 Distributions of Average Household Size

Infection Rate

The COVID-19 infection rate in Toronto neighborhoods in 2020 saw an average of 0.062 in 140 neighborhoods, alongside a median of 0.059. The histogram confirms the skewness of the distribution to the left side, while displaying only one outlier in the boxplot (see Figure 7).

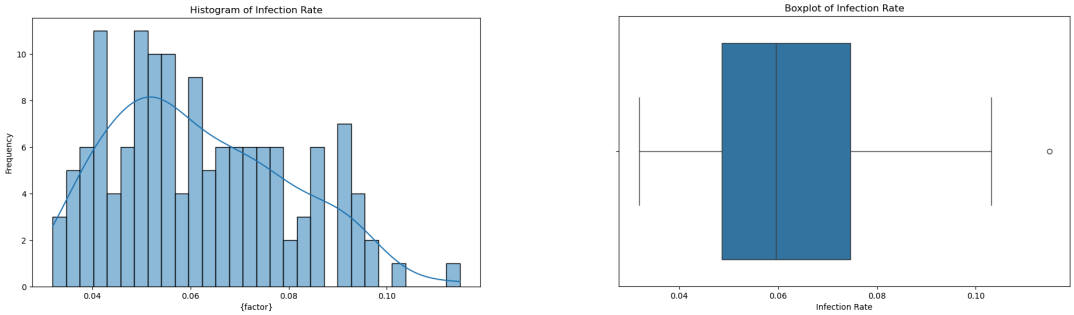

Figure 7 Distributions of Infection Rate

3.4.4 Correlations

Figure 8 shows the distributions of variables. It can be seen that age and household sizes are normally distributed in the data set, while infection rate and the rate of having a Bachelor's or above degree are right-skewed. Population density, median income per person, and employment rate have shown a right-skewed shape with a weak tendency of multimodal.

Regarding the infection rate, median income per person and the rate of people with a Bachelor's degree or higher have shown a negative correlation with it. Household size, mean age and employment rate seem to have a weak relationship with infection rate. In the scatterplot of

population and infection rate, it illustrates an almost vertical straight line, suggesting that a similar population has a broad range of possible infection rates.

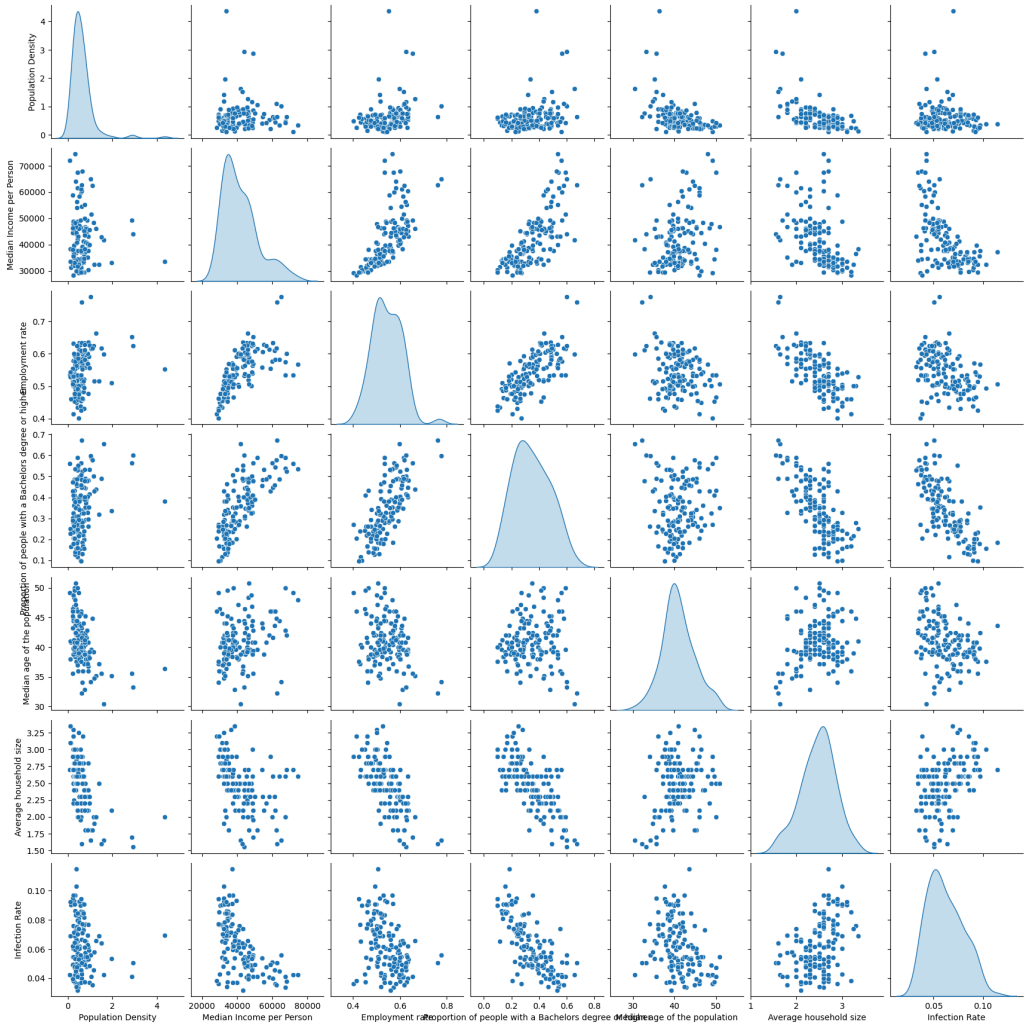

Figure 8 Pairplots of Predictor Variables

From the correlation heatmap (Figure 9), it is noted that median income per person and the proportion of people with a Bachelor's degree or higher have shown a strong negative correlation with the infection rate, with correlation values of -0.58 and -0.73 respectively. Particularly, average household size and the infection rate have shown a moderate positive correlation with a value of 0.42.

Figure 9 Correlation Heatmap of Predictor Variables

4 Model/Methodology

4.1 Model set-up

The dataset was split into features and target variables for classification analysis. The features included education level, population density, median income, employment rate, median age, average household size, and infection rate. The target variable was 'IFS' (if there was medical/social support). The data was then split into training and testing sets using a 70-30 ratio, with 70% of the data used for training the models and 30% for testing their performance. To ensure fair comparison and improve model performance, the features were standardized using Z-score normalization (StandardScaler) to have zero mean and unit variance.

4.2 Model Selection

In the modeling process, a pipeline was constructed to test the accuracy of several machine learning algorithms, including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Gaussian Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machines (SVM) with linear and RBF kernels, Random Forest, Bagging, and Decision Tree. And the accuracy scores of each model are shown below.

```
          LDA: 0.7619
          QDA: 0.7619
  Naive Bayes: 0.7143
          KNN: 0.8333
          SVC: 0.7619
          SVM: 0.8333
Random Forest: 0.7143
     Bagging: 0.7143
Decision Tree: 0.6905
```

Based on the output, among the tested models, KNN and SVM (RBF) demonstrated the highest accuracy on the test set, both achieving 83.33%. The optimal value of k was determined through hyperparameter tuning, and k=4 yielded the best results.

4.3 Cross-validation

To obtain a more robust estimate of model performance and assess their generalization ability, 10-fold cross-validation was performed on the training data for all the models. The mean accuracy score across the 10 folds was calculated for each model to provide a more reliable estimate of their performance, which is shown below.

```
          LDA: 0.7967
          QDA: 0.8178
  Naive Bayes: 0.7878
          KNN: 0.7778
          SVC: 0.7967
          SVM: 0.7778
Random Forest: 0.7767
     Bagging: 0.7456
Decision Tree: 0.6822
```

QDA achieved the highest mean accuracy of 81.78% across the 10 folds, while KNN and SVM (RBF) had slightly lower cross-validation scores of 77.78% each.

Based on the results, KNN and SVM (RBF) are the top-performing models for this classification task. Their high accuracy on the test set and competitive cross-validation scores demonstrate their ability to effectively predict the IFS based on the given socio-economic factors. However, given the limited size of the dataset, with fewer than 150 observations, we have decided to proceed with KNN as our preferred model.

5 Result and Interpretation

After selecting KNN as our preferred algorithm for this classification task, we proceeded to train the model using the optimal value of k=4 determined from the previous step. Once the model was trained, we evaluated its performance by making predictions on the test set and generating a confusion table. Here is the resulting confusion table:

Truth	0	1	2
Predicted			
0	0	0	0
1	5	35	2
2	0	0	0

Upon analyzing the confusion table, we observe that the KNN classifier exhibits a strong bias towards predicting class 1. All the instances in the test set, regardless of their true labels, are predicted as belonging to class 1. This indicates that the model struggles to distinguish between the different classes based on the given socio-economic factors.

6 Discussion & Next Steps

The confusion table highlights a significant issue with the KNN classifier's performance – it is heavily biased towards predicting class 1, resulting in a complete failure to predict instances of class 0 and class 2 correctly. This bias could stem from class imbalance in the raw data, where class 1 has a substantially higher representation compared to the other classes. Moreover, the model's inability to distinguish between the classes suggests that the selected socio-economic factors may not be sufficiently informative or discriminative for accurate classification. The features used may not capture the underlying patterns or relationships that differentiate the classes effectively.

To improve the model's performance and address the class imbalance issue, one approach is to expand the scope of the analysis by incorporating additional socio-economic and response factors such as financial support and psychological support. This will modify the coding method for the IFS variable to better capture the nuances and variations in the response levels. Moreover, introducing temporal data, such as the dates of the responses, could offer valuable insights into

the timeliness and effectiveness of the COVID-19 response measures deployed in different neighborhoods over time.

By refining the approach and incorporating these suggestions, the KNN model could possibly provide more accurate and meaningful predictions of the relationship between socio-economic factors and COVID-19 response in Toronto neighborhoods.

7 Reference

Chriscaden, K. (2020, October 13). Impact of COVID-19 on people's livelihoods, their health and our food systems. World Health Organization. https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people%27s-livelihoods-their-health-and-our-food-systems

Mohammadpour, M., Zarifinezhad, E., Ghanbarzadegan, A., Naderimanesh, K., Shaarbafchizadeh, N., & Bastani, P. (2021). Main factors affecting the readiness and responsiveness of healthcare systems during epidemic crises: A scoping review on cases of SARS, MERS, and COVID-19. Iranian Journal of Medical Sciences, 46(2), 81–92. https://doi.org/10.30476/ijms.2020.87608.1801

Nwakasi, C., Esiaka, D., Uchendu, I., & Bosun-Arije, S. (2021). Factors influencing compliance with public health directives and support for government's actions against COVID-19: Nigerian case study. Scientific African. https://doi.org/10.1016/j.sciaf.2021.e01089

Public Health Agency of Canada. (2024, February 20). Social determinants of health and health inequalities. Government of Canada. https://www.canada.ca/en/public-health/services/health-promotion/population-health/what-determines-health.html

8 Appendix

A. Link to Google Colaboratory Notebook

https://colab.research.google.com/drive/1qJx20b7TyUqqSWxL-cCaWLHK5mbmM4q0?usp=sharing