Sociodemographic factors that affect the risk of death from COVID-19 contraction in Toronto, Canada

Written by: Ragavie Manoragavan

Ryerson University

INTRODUCTION

In January 2020, the world was just starting a new decade. However, the world was entirely unaware of what it was going to endure. The World Health Organization (WHO) was made aware in January 2020 that an outbreak in Wuhan, China was caused by a novel coronavirus. Fast forward to two months later, the world was impacted by this coronavirus, later identified as COVID-19. The coronavirus made its way around the world, heavily impacting people's health. In March 2020, WHO declared Coronavirus COVID-19 a pandemic. In Toronto, Canada, even with an abundant number of resources, COVID-19 remains a persistent challenge that is not going to be easy to bring down. The people of Toronto had to quickly adapt to fight and learn about this new, ambiguous virus. There have been uncontrollable daily cases, increase in the need for health services, and pressure on the health system and public health services. Amongst the fight to find balance, there have been an unfortunate influx in deaths from COVID-19 contraction in Toronto from the start of the pandemic. As the city experiences subsequent waves of this virus outbreak, researchers are recognizing that there are factors that affect the probability of dying from the virus. It is known that sociodemographic factors can explain the nature of disease contraction, but there is inadequate information with regards to COVID-19 and death. The purpose of this project is to look into sociodemographic factors of Toronto citizens and the risk of dying from COVID-19 contraction. The hope is to be able to predict who may be at higher risk of dying from COVID-19 contraction, based on certain sociodemographic factors.

LITERATURE REVIEWS

Analysis of the spatial distribution of cases of Zika virus infection and congenital Zika virus syndrome in a state in the southeastern region of Brazil: Sociodemographic factors and implications for public health

The purpose of this ecological study was to understand how sociodemographic factors may have had implications of the Zika virus infection and congenital Zika syndrome (CZS) in the state of Espirito Santo, Brazil, by neighbourhood. The study concluded that people with specific socioeconomic and demographic factors did impact the cases of Zika virus infection and CZS, which included class, social group, or gender.

The Impact of Sociodemographic Factors, Comorbidities, and Physiologic Responses on 30-Day Mortality in Coronavirus Disease 2019 (COVID-19) Patients in Metropolitan Detroit

This retrospective cohort study assessed the risk factors that may affect the mortality of patients with the COVID-19 infection in Metropolitan Detroit. Patients who presented to the emergency departments between March and April 2020 were included in the study. The primary outcomes were hospitalization and 30-day mortality. The study concluded that in regard to sociodemographic factors, disparities in income or source of health insurance did not affect mortality. It also found that black women had a lower risk of dying.

A model of disparities: risk factors associated with COVID-19 infection

This is a multivariable statistical model study where the objective was to understand clinical, sociodemographic, and environmental variables that may cause a risk with the initial infection of COVID-19. The study found that the risk of COVID-19 contraction across the US is higher among groups already affected by the following sociodemographic factors: race, ethnicity, language, income, and living conditions.

Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the COVID-19 epidemic in France under lockdown: a population-based study

This population-based study looked at assessing mobility reductions of individual residents after the lockdown in France, while evaluation the effect of sociodemographic factors. The data was broken down by analyzing trip distance, user age and residency, and time of day. The study was able to show that there were links between geography, demography and the timing of the response to mobility restrictions.

Emotional distress and associated sociodemographic risk factors during the COVID-19 outbreak in Spain

This was a cross-sectional survey where sociodemographic variables were assessed to see how they would influence the emotional disorders affected by COVID-19 in Spain. The study sample suggests that mental health interventions are needed in Spain due to the emotional stress from the impact of COVID-19.

The AGE Effect on Protective Behaviours During the COVID-19 Outbreak: Sociodemographic, Perceptions and Psychological Accounts

The purpose of this study was to focus on age-related differences of the different protective behaviours during the first wave of the pandemic. It considered the sociodemographic, COVID-related, perceived risk and psychosocial variables. The study found that the ones that reported higher traumatic experiences happened to have retired and have lower educational levels. It was found individuals with lower educational levels needed accessible messages to better communicate the health education.

Socio-demographic factors associated with self-protecting behaviour during the COVID-19 pandemic

This paper looks at how sociodemographic factors could affect the self-protective measures one takes during the COVID-19 pandemic in the USA. The data included factors such as income, gender, race, work arrangements due to COVID-19, and housing quality. The study was able to find that people with higher income is inked to larger changes in self-protective behaviour. It also partially explained that people with less income are more likely to report the difficulties of the self-protective measures. To conclude, the study recognizes that policies that assume universal compliance with self-protective measures or don’t take into consideration of sociodemographic or economic differences are unlikely to be effective.

Sociodemographic, clinical and laboratory factors on admission associated with COVID-19 mortality in hospitalized patients: A retrospective observational study

This was a retrospective observational study that looked at finding any association between baseline characteristics (sociodemographic, clinical and laboratory factors) and hospital admission and mortality with COVID-19 patients at a Spain hospital. This study was able to find patients of older age and those who lived in a retirement facility experienced a higher mortality rate.

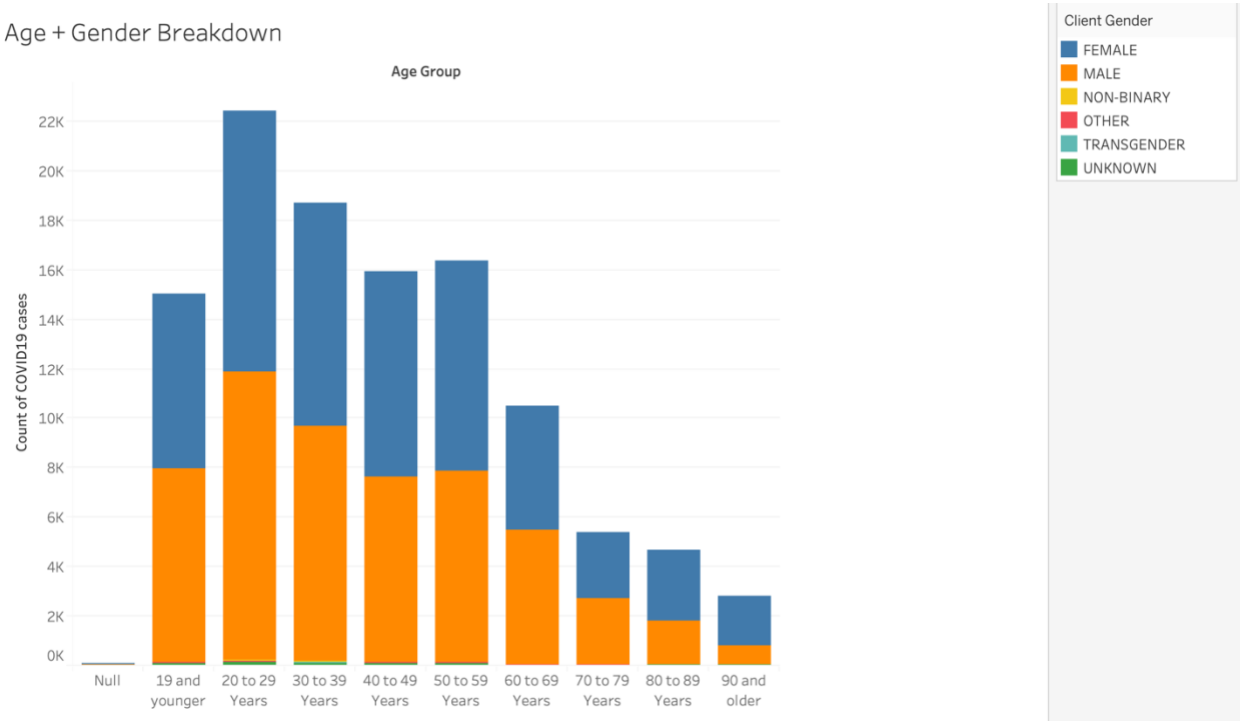
DATASETS

The project will look at two datasets from the City of Toronto's Open Data Portal: COVID-19 Cases in Toronto (https://open.toronto.ca/dataset/covid-19-cases-in-toronto/) and Neighbourhood Profiles 2016 (https://open.toronto.ca/dataset/neighbourhood-profiles/).

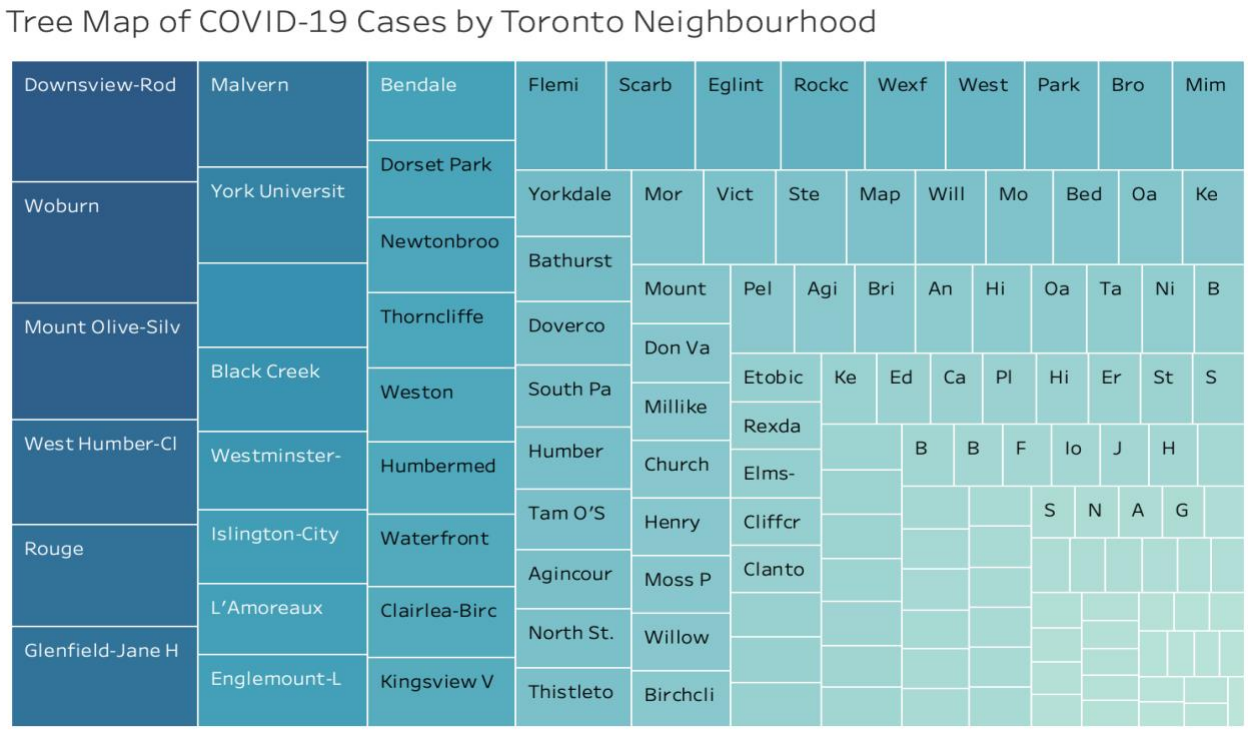
COVID-19 Cases in Toronto

This dataset is updated weekly on the data portal. The project will look at COVID-19 cases between the date range of January 23rd, 2020 to March 29th, 2021. The dataset comes with 18 columns which include two unique IDs, demographics, data pertaining to the COVID-19 contraction of the person. For the study, the following columns will be looked at: Age Group, Neighbourhood Name, Reported Date, Client Gender, and Outcome. 'Age Group', 'Neighbourhood Name' and 'Client Gender' are categorical, demographic factors that were provided. 'Age Group' were categorized as age ranges for every 10 years. 'Neighbourhood Name' covered the 140 neighbourhoods within Toronto and indicated the area of living for each person. 'Client Gender' looked at if person was male, female, transgender, non-binary, other or unknown. Column 'Outcome' told us if the case was active, resolved, or fatal. The project will be using this column as what would need to be predicted.

To better understand the dataset that is being worked with, exploratory data analysis was performed using Microsoft Excel and Tableau. The main purpose of the EDA is to visualize the various factors and try to understand any correlations.

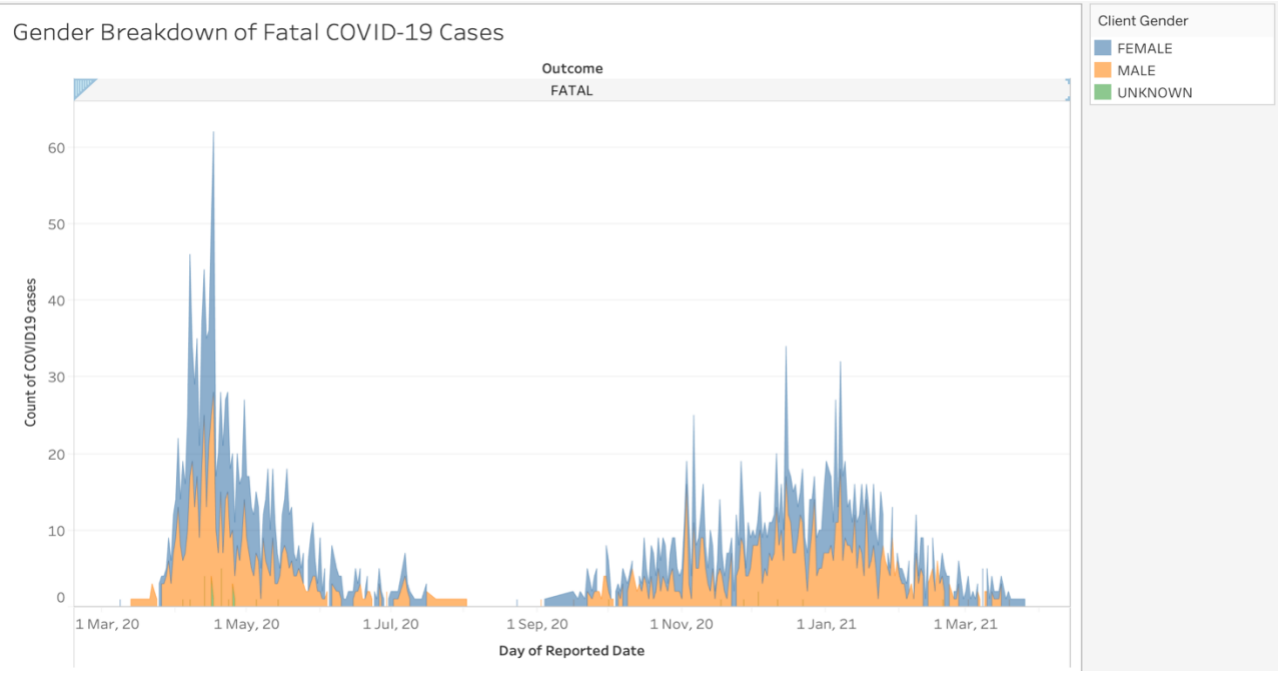


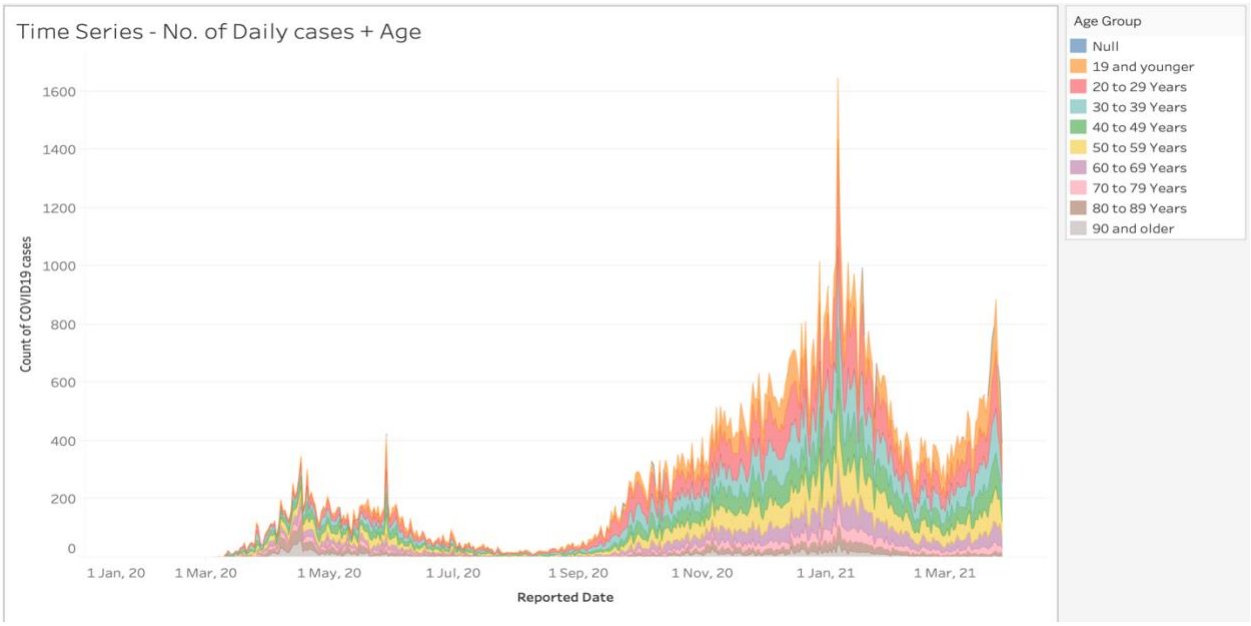
This graph looks at the total COVID-19 case count up to March 2021. This is the age and gender breakdown of the total cases. Overall, it can be easily seen that between all the genders, most COVID-19 contractions have been by females. Furthermore, more younger age



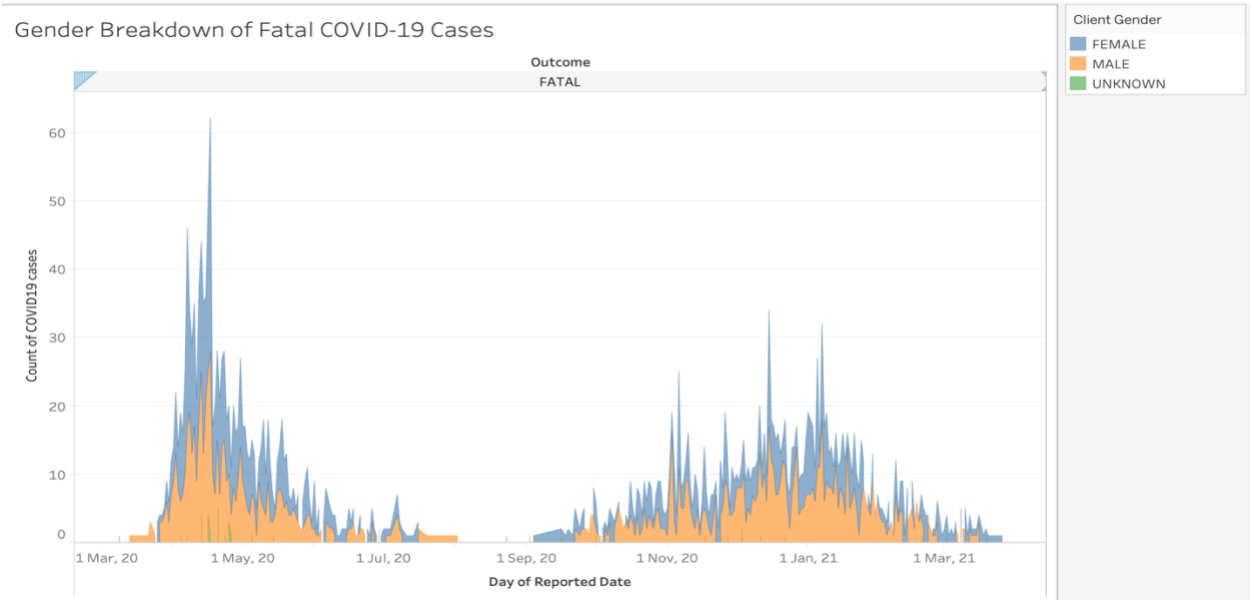
people (<40 years) are contracting COVID-19, compared to older age. With this, it can be said younger aged females are at higher risk of contracting COVID-19.

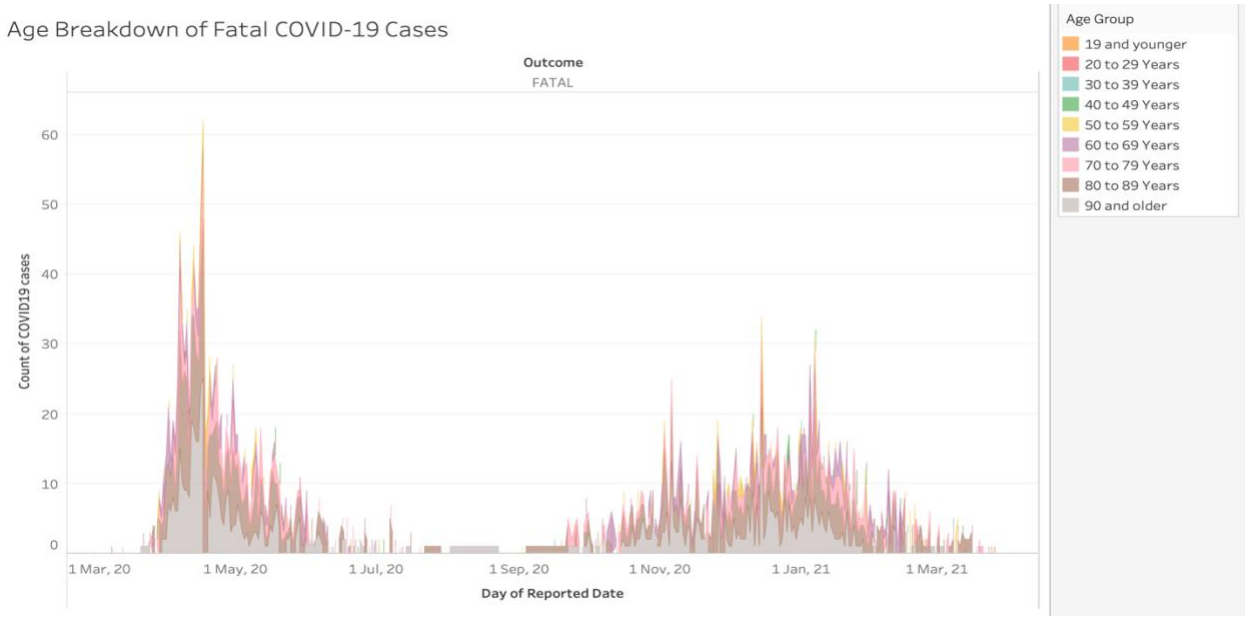
This tree map looks at the COVID-19 cases within each neighbourhood of Toronto, with the densest neighbourhood being the darkest and least dense neighbourhood being the lightest. Looking at the neighbourhood breakdown would help understand hot spot areas in Toronto with a higher risk of contraction. As seen in the map, the most COVID-19 dense neighbourhoods are from the west and east end of Toronto. With demographic breakdown (income) and comparison with the fatal COVID-19 cases, it may be more apparent if there are any specific correlations.





These time series graphs are plotting the COVID-19 cases from March 2020 to March 2021. These graphs are great for visualization and better understanding the pandemic. The peaks in the graphs indicate the pandemic waves that were experienced and experiencing. With the two breakdowns (age and gender), it continues to corroborate the original graph, where it was seen that females and younger aged people were heavily affected, in terms of COVID-19 contraction.

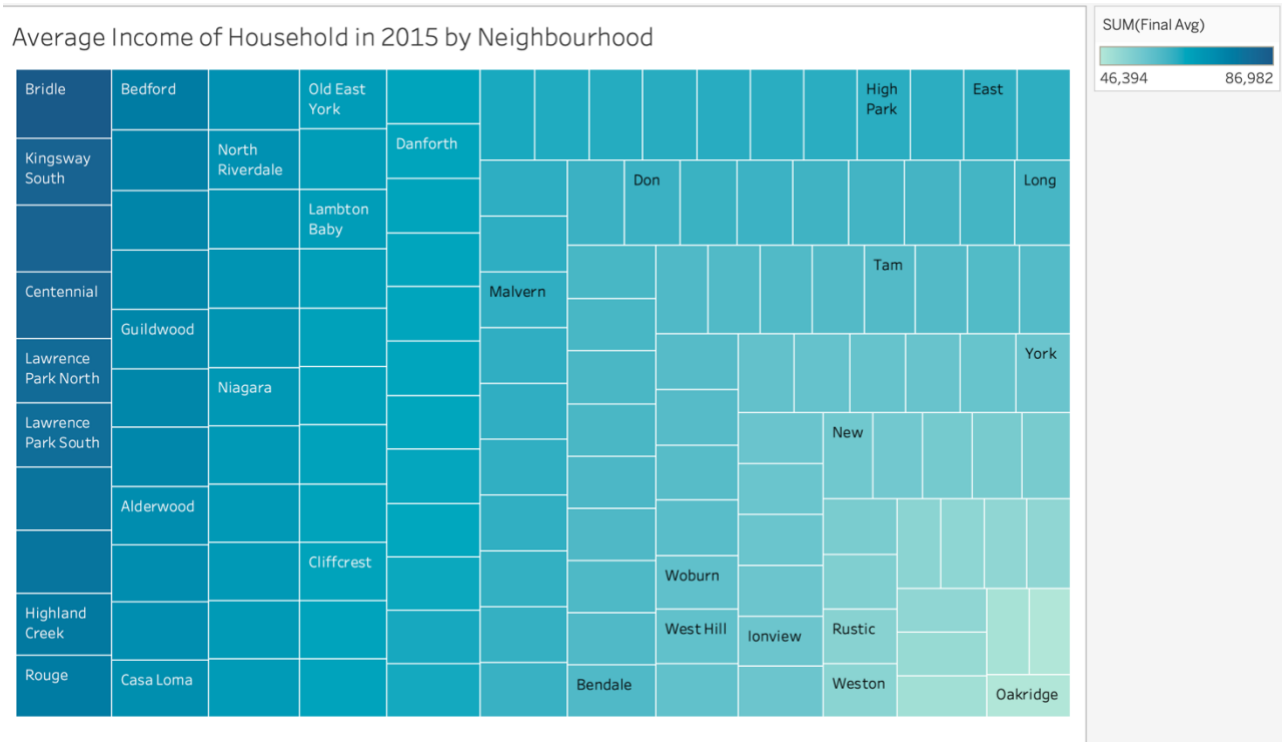




On the other hand, these time series look at the fatal COVID-19 cases between March 2020 and 2021. Just like the contraction, the graph shows that there are more females dying due to COVID-19 than males. However, unlike the contraction trends, the older age groups are dying due to COVID-19, compared to the younger aged people. With this general visualization, it would be interesting to do further classification to better understand this observation.

Neighbourhood Profiles 2016

This is a demographic dataset of Toronto in 2015. It breaks down the demographics of each neighbourhood. It looks at population characteristics, income, education, housing, ethnicity, citizenship and work industry. For the purpose of this project, the only demographic that will be extracted will be income for each neighbourhood. With the income, the data will be analyzed to visualize the average income of each neighbourhood and how it would correlate with the contracted cases of COVID-19. Out of the entire demographic set, the only columns that were extracted were “Income of households in 2015”, looking at the various income ranges. After doing some analysis through Python, the average income was calculated for each neighbourhood.



The analysis through Python shows that the highest income neighbourhood was Bridle Path, with Oakridge as the lowest income.

Fatal COVID-19 Cases by Neighbourhood

Neighbourhood Name	Outcome FATAL
Steeles	96
Rouge	89
Birchcliffe-Cliffside	82
Glenfield-Jane Heights	81
Islington-City Centre West	79
Dorset Park	77
York University Heights	76
South Parkdale	62
Weston	58
Thistletown-Beaumond H..	57
Mount Pleasant West	57
Clairlea-Birchmount	53
Yorkdale-Glen Park	52
Woburn	52
Morningside	51
West Humber-Clairville	50
Humber Heights-Westmo..	50
Guildwood	50
Bendale	43
Kensington-Chinatown	41
Annex	40
Mount Dennis	39
High Park-Swansea	38
Bathurst Manor	37
L'Amoreaux	33

Neighbourhood Count - Highlight Tables

Neighbourhood Name	
Downsview-Roding-CFB	3,093
Woburn	3,070
Mount Olive-Silverstone-J..	2,941
West Humber-Clairville	2,689
Rouge	2,576
Glenfield-Jane Heights	2,551
Malvern	2,465
York University Heights	2,221
Black Creek	1,918
Westminster-Branson	1,843
Islington-City Centre West	1,688
L'Amoreaux	1,663
Englemount-Lawrence	1,662
Bendale	1,607
Dorset Park	1,549
Newtonbrook West	1,520
Thorncliffe Park	1,512
Weston	1,480
Humbermede	1,469
Waterfront Communities-..	1,442
Clairlea-Birchmount	1,422
Kingsview Village-The We..	1,402
Flemingdon Park	1,359
Scarborough Village	1,328
Eglinton East	1,271
Rockcliffe-Smythe	1,258
Wexford/Maryvale	1,200

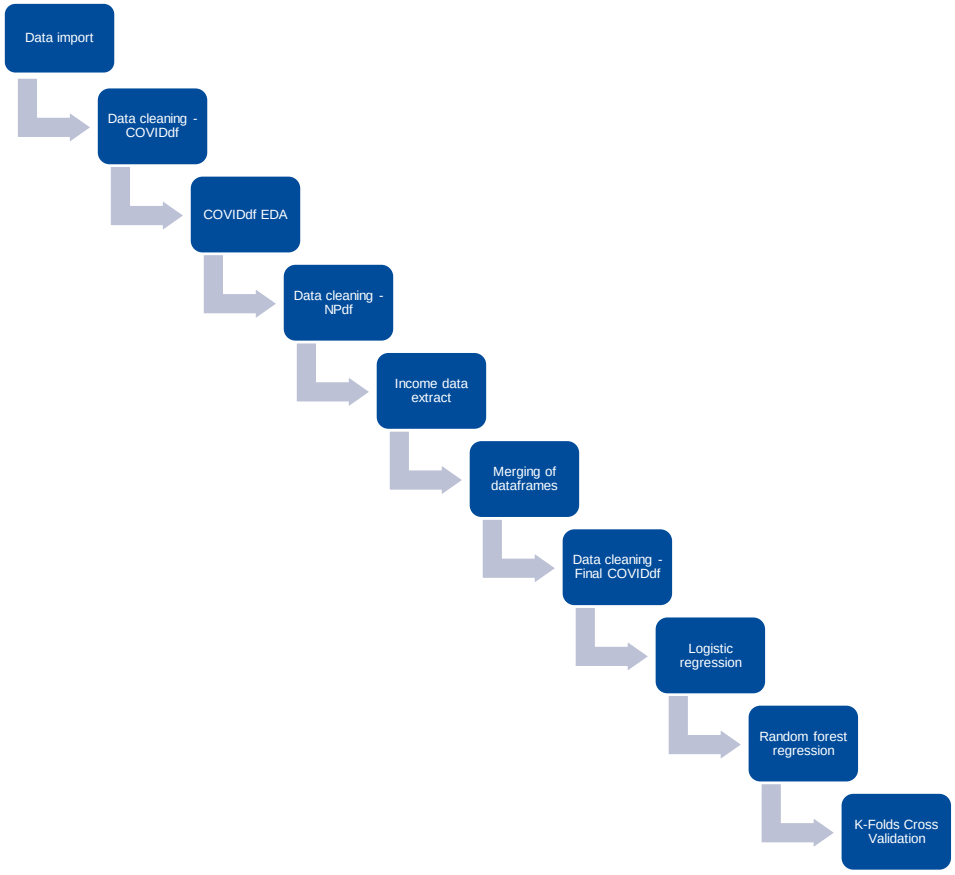
When observing the fatal COVID-19 cases by neighbourhood, there doesn't seem to be any immediate correlation with COVID-19 death and the neighbourhood income. This could be an indication that a household's income may not have a relationship with a person's chance of dying after COVID-19 contraction. When looking at COVID-19 contraction and average income by neighbourhood, on the other hand, there seems to be a correlation between a neighbourhood's average income and the number of COVID-19 cases in each neighbourhood. If further evaluated, it may be possible to see a relationship between income and contraction of the virus.

Preparation for analysis

From the dataset EDA, there were some evident factors that showed there is a relationship with COVID-19 fatal cases. On the contrary, there is a lack of clarity if factors such as neighbourhood and the neighbourhood average income would play a strong role in a person's fatality from COVID-19. With that said, with further analysis, this can be clarified.

Because this is a classification problem, the datasets will be set up for logistic regression and random forest regression. Furthermore, K-Folds cross validation will be used on the more accurate model to train and test. With these modelling, it should become apparent which specific factors have a stronger importance on COVID-19 mortality in Toronto.

Approach



GitHub link: https://github.com/rmanoragavan/CIND820-Capstone-Project/blob/main/FinalCode.ipynb

Jupyter Notebook link: https://cind820-xj0-w21.cejupyter.ryerson.ca/hub/user-redirect/lab/tree/Capstone%20CIND820

Step 1: Data import

The two data sets were imported onto the Jupyter Notebook and data frames were created to further analyze the datasets.

Step 2: Data cleaning – COVIDdf

For the COVIDdf data frame, dataset needed to be cleaned prior to any EDA or statistical analysis. This was done by removing unnecessary columns and removing any missing values.

Step 3: COVIDdf EDA

The data frame was ready for EDA. To start, each column was explored to see the different values, value counts and data types. After reviewing each variable, they were grouped to see the breakdown within each category.

Step 4: Data cleaning – NPdf

The NPdf data frame needed to be cleaned as the project only needs to extract the income profile of Toronto neighbourhoods. Unnecessary columns and null values were removed from the dataset.

Step 5: Income data extract

This was a critical step in the project as this was a prefix to the merging of the two data frames. This extract needed to get done to ensure all the interested variables were under one data frame.

In the NPdf data frame, the dataset gives the value count of each income range for each neighbourhood. For the purpose of the project, these values needed to be used to find the average income of each neighbourhood. This was calculated through the method of weighted average. The product of each value count and income range was calculated. A new data frame was created to store this information. The product was then divided by the sum of the entire column, giving the average income of each neighbourhood. Once this was calculated, it was

merged onto the COVIDdf data frame, to ensure all the needed variables and values were in one data frame.

Step 6: Merging of data frames

Once the average incomes for each neighbourhood was calculated, this data frame was merged onto the COVIDdf data frame to have unanimity. They were merged by matching the neighbourhood names in both data frames and matching it with the appropriate average income. This way, each case had the average income of the neighbourhood of the person.

Step 7: Data cleaning – Final COVIDdf

To prepare for the statistical analysis, the final COVIDdf data frame needed some data cleaning. In the column 'Outcome', there were three values: 'ACTIVE', 'RESOLVED', and 'FATAL'. For the purpose of the project and analysis, the value 'ACTIVE' needed to be removed for accuracy. Furthermore, the values 'RESOLVED' and 'FATAL' were changed into binary values of 1 and 0. This change was made for modelling purposes. It was also recognized that there was an unbalanced number of cases between 'RESOLVED' and 'FATAL'. There were 2759 'FATAL' cases and 100346 'RESOLVED' cases. To ensure no bias, 2759 'RESOLVED' cases were randomly selected to remain in the data frame, while 97587 'RESOLVED' cases were dropped.

Afterwards, any empty cell values were filled with the value '0'. To prepare for the modelling, the categorical variables were one-hot encoded.

Step 8 – Logistic Regression

With the data frame, the input variables and predict variable were determined. Using the SMOTE algorithm, the training data was over-sampled. Afterwards, Recursive Feature Elimination (RFE) was performed on the data to choose the best features. With that, the process is repeated with the rest of the features, to construct the best model possible. After that, with the best features, the logistic regression model was implemented. After that run, variables with a p-value higher than 0.05 was removed, and the model was run again. After the models were created, precision, recall, F-measure and support were calculated. Finally, the ROC Curve was plotted.

Step 9 – Random Forest Regression

With the data frame, the features and targets were identified, and converted into arrays. With that, training and testing sets were made for the RF model. Before predictions can be evaluated, a baseline needed to be established. The model was trained with the training data. Afterwards, the model was set to make predictions on the test set. After the model was interpreted, the

decision tree was visualized. Furthermore, variables' importance was calculated to quantify the usefulness.

Step 10 – K-Folds Cross Validation

After the two regression models, it was deemed the logistic regression gave a more accurate and better-fitting model, K-Folds cross validation was done, between 1 and 16 folds. The mean and standard error were plotted as a box plot for visualization purposes.

Results

LOGISTIC REGRESSION

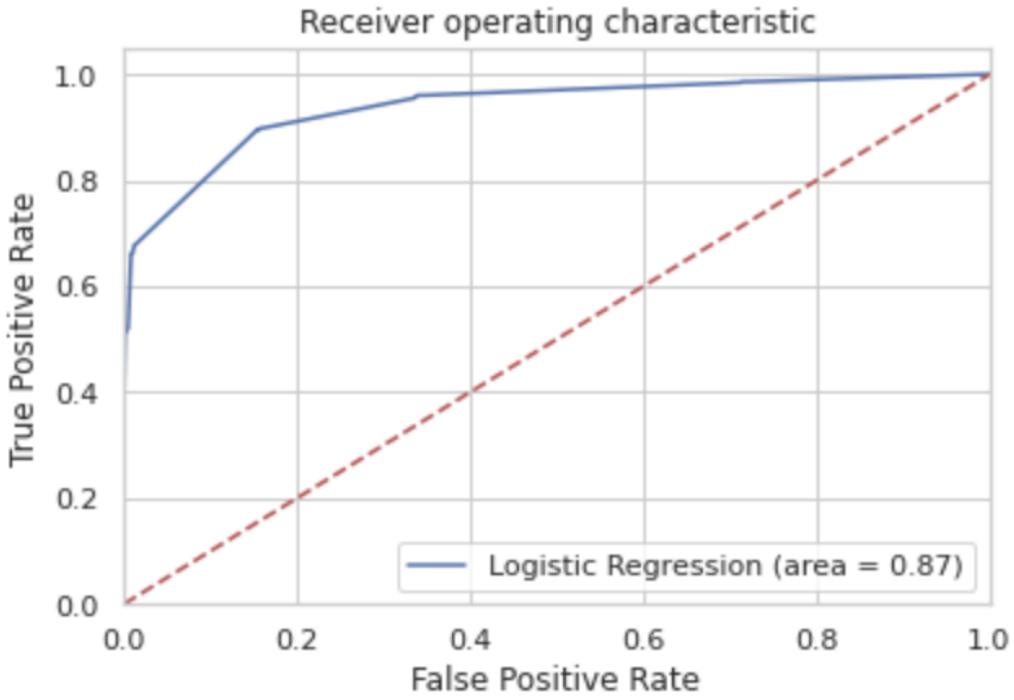
```
Optimization terminated successfully.
         Current function value: 0.287772
         Iterations 10
                                 Results: Logit
================================================================================
Model:                      Logit                  Pseudo R-squared:       0.585
Dependent Variable:         Outcome                AIC:                    2249.9023
Date:                       2021-04-02 23:08       BIC:                    2331.2752
No. Observations:           3864                   Log-Likelihood:         -1112.0
Df Model:                   12                     LL-Null:                -2678.3
Df Residuals:               3851                   LLR p-value:            0.0000
Converged:                  1.0000                 Scale:                  1.0000
No. Iterations:             10.0000
--------------------------------------------------------------------------------
                                      Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
--------------------------------------------------------------------------------
Age Group_19 and younger               5.6195   1.0022   5.6069 0.0000  3.6551  7.5839
Age Group_20 to 29 Years               6.0367   1.0052   6.0053 0.0000  4.0665  8.0068
Age Group_30 to 39 Years               4.4588   0.5075   8.7851 0.0000  3.4641  5.4536
Age Group_40 to 49 Years               2.7850   0.2507  11.1109 0.0000  2.2938  3.2763
Age Group_70 to 79 Years              -1.4112   0.1185 -11.9082 0.0000 -1.6435 -1.1790
Age Group_80 to 89 Years              -2.5049   0.1380 -18.1539 0.0000 -2.7753 -2.2345
Age Group_90 and older                -3.2702   0.2089 -15.6546 0.0000 -3.6796 -2.8608
Neighbourhood Name_Casa Loma           1.4879   0.6856   2.1702 0.0300  0.1441  2.8317
Neighbourhood Name_Downsview-Roding-CFB 1.4893  0.4231   3.5201 0.0004  0.6601  2.3186
Neighbourhood Name_Englemount-Lawrence 1.3792   0.5340   2.5827 0.0098  0.3326  2.4259
Neighbourhood Name_Forest Hill North   2.8683   0.8706   3.2945 0.0010  1.1619  4.5747
Neighbourhood Name_Trinity-Bellwoods   2.7500   1.2881   2.1350 0.0328  0.2254  5.2746
Neighbourhood Name_Yorkdale-Glen Park -1.8671   0.9207  -2.0280 0.0426 -3.6716 -0.0627
================================================================================
```

This was the final logistic regression model that was implemented with the best features out of the data set. As it can be seen, there are 13 variables, all of which have a P-value less than 0.05. In regard to the age groups, based on the coefficient, it can be seen that people that are 70 and older are less likely to have resolved cases. Furthermore, living in the neighbourhood Yorkdale-Glen Park, there is a higher chance of fatality. The logistic regression model shows

that out of all the selected sociodemographic factors, age has an effect on the chance of dying after COVID-19 contraction.

	precision	recall	f1-score	support
0	0.89	0.85	0.87	577
1	0.85	0.90	0.87	583
accuracy			0.87	1160
macro avg	0.87	0.87	0.87	1160
weighted avg	0.87	0.87	0.87	1160

This was the calculation of precision, recall, F-measure and support of the final model. As seen, there is a high precision rate, which is indicative of a low positive rate. The recall rate is very



high, which indicates the correctly predicted positive observations to all the observations in the actual class. The F1-score is 0.87, which is very close to 1. This is indicative of the model reaching its best value.

This is the ROC curve produced by the model. As seen in the graph, the curve is close to the top-left corner. This is a strong indication that this a great performance of the model.

All in all, it can be seen that the logistic regression was able to make accurate predictions. As for now, this is a very strong model.

RANDOM FOREST REGRESSION

The random forest regression was run with various parameters. Out of all, this run produced the highest accuracy (73.55%). With one estimator and one random state, this gave the highest accuracy out of all the models. Compared to the logistic regression model, this has a lower accuracy.

The variable importance was calculated to help build a stronger and more accurate model. As seen above, the top four important variables are the four age groups, specifically age ranges from 60 and older. This is parallel to what was seen in the logistic regression, where it was evident that the older age groups had a bigger impact on the predicting model on the COVID-19 outcome.

[140]:
```
rf = RandomForestRegressor(n_estimators = 1, random_state = 1)
rf.fit(train_COVIDdf, train_labels);


predictions = rf.predict(test_COVIDdf)
errors = abs(predictions - test_labels)
print('Mean Absolute Error:', round(np.mean(errors), 2), 'degrees.')
Accuracycount = 0

for i in range(0, len(predictions)):
    if predictions[i] == test_labels[i]:
          Accuracycount +=1

Accuracy = 100* Accuracycount/ len(predictions)
Accuracy

# Instantiate model with 1 decision trees
# Train the model on training data
# Use the forest's predict method on the test data
# Calculate the absolute errors
# Print out the mean absolute error (mae)
# Print out accuracy
# This model has highest accuracy out of all models
```

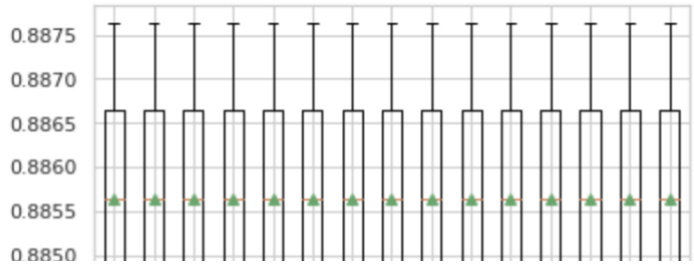
```
Mean Absolute Error: 0.16 degrees.
```

[140]:
```
73.55072463768116
```

```
Variable: Age Group_90 and older Importance: 0.29
Variable: Age Group_80 to 89 Years Importance: 0.22
Variable: Age Group_70 to 79 Years Importance: 0.19
Variable: Age Group_60 to 69 Years Importance: 0.07
Variable: NeighbourhoodAvgIncome Importance: 0.04
Variable: Client Gender_FEMALE Importance: 0.02
Variable: Age Group_50 to 59 Years Importance: 0.01
Variable: Client Gender_MALE   Importance: 0.01
Variable: Age Group_19 and younger Importance: 0.0
Variable: Age Group_20 to 29 Years Importance: 0.0
Variable: Age Group_30 to 39 Years Importance: 0.0
Variable: Age Group_40 to 49 Years Importance: 0.0
Variable: Neighbourhood Name_Agincourt North Importance: 0.0
Variable: Neighbourhood Name_Agincourt South-Malvern West Importance: 0.0
Variable: Neighbourhood Name_Alderwood Importance: 0.0
Variable: Neighbourhood Name_Annex Importance: 0.0
Variable: Neighbourhood Name_Banbury-Don Mills Importance: 0.0
Variable: Neighbourhood Name_Bathurst Manor Importance: 0.0
Variable: Neighbourhood Name_Bay Street Corridor Importance: 0.0
Variable: Neighbourhood Name_Bayview Village Importance: 0.0
Variable: Neighbourhood Name_Bayview Woods-Steeles Importance: 0.0
Variable: Neighbourhood Name_Bedford Park-Nortown Importance: 0.0
Variable: Neighbourhood Name_Beechborough-Greenbrook Importance: 0.0
Variable: Neighbourhood Name_Bendale Importance: 0.0
Variable: Neighbourhood Name_Birchcliffe-Cliffside Importance: 0.0
Variable: Neighbourhood Name_Black Creek Importance: 0.0
Variable: Neighbourhood Name_Blake-Jones Importance: 0.0
Variable: Neighbourhood Name_Briar Hill-Belgravia Importance: 0.0
Variable: Neighbourhood Name_Bridle Path-Sunnybrook-York Mills Importance: 0.0
Variable: Neighbourhood Name_Broadview North Importance: 0.0
Variable: Neighbourhood Name_Brookhaven-Amesbury Importance: 0.0
Variable: Neighbourhood Name_Cabbagetown-South St. James Town Importance: 0.0
```

K-FOLDS CROSS VALIDATION

```
>1 mean=0.8856 se=0.002
>2 mean=0.8856 se=0.002
>3 mean=0.8856 se=0.002
>4 mean=0.8856 se=0.002
>5 mean=0.8856 se=0.002
>6 mean=0.8856 se=0.002
>7 mean=0.8856 se=0.002
>8 mean=0.8856 se=0.002
>9 mean=0.8856 se=0.002
>10 mean=0.8856 se=0.002
>11 mean=0.8856 se=0.002
>12 mean=0.8856 se=0.002
>13 mean=0.8856 se=0.002
>14 mean=0.8856 se=0.002
>15 mean=0.8856 se=0.002
```



[148]:
```
kfold = model_selection.KFold(n_splits=10, random_state=100, shuffle = True)
model_kfold = LogisticRegression()
results_kfold = model_selection.cross_val_score(model_kfold, X, Y, cv=kfold)
print("Accuracy: %.2f%%" % (results_kfold.mean()*100.0))

# evaluate a logistic regression model using repeated k-fold cross-validation, 10-fold
```

```
Accuracy: 88.93%
```

After seeing that between the two types of regression models, logistic was more successful, it was decided to perform K-Folds cross validation on the logistic model. The validation was run with between 1 and 16 folds. It was found that ten folds had the highest accuracy, of 88.93%. This average shows that this model has a very high performance.

DISCUSSION

After running both regression models, it can be seen that the logistical regression model was a better fit than the random forest regression. As a classification problem, the whole purpose is to predict the outcome of a person who had contracted COVID-19. Both models were able to do that, to some satisfaction.

From both models, it was made clear that the age group has a significance on the prediction, more specifically the older age groups. In the final model of the logistic regression, the only sociodemographic factors were age and neighbourhood. However, on the other hand, in the random forest regression model, the variables with importance were age group and gender.

On a logical point of view, some sociodemographic factors would be indicative of contraction but not necessarily fatality. The chance of someone dying can have to do with their age and gender, as it is associated with frailty and possible comorbidities. However, factors such as location and income, may not be so indicative. Regardless of income or residential area, in Toronto, Canada, a person has access to free healthcare, equal services, and access to help. Given that, a person's economic status may not directly affect their chance of dying if they contract COVID-19.

For future improvements, it would be interesting to see how these very same sociodemographic factors affect COVID-19 contraction instead. When doing the initial EDA, there was correlation between location, a person's income, age and gender, and their COVID-19 contraction. As for further investigating COVID-19 mortality, it would be interesting to look at a person's ethnicity, past medical history, and course of journey if hospitalized. Such factors could possibly have a more direct effect in predicting mortality, compared to just sociodemographic factors.

CONCLUSION

The purpose of this project was to be able to classify and predict COVID-19 mortality in Toronto, Canada, based off of specific sociodemographic factors. To create models and predict, two models were used: logistic regression and random forest regression. Between these models, logistic regression gave a higher performing model, with an accuracy of 87%. Froom this model, it was understood that age groups of people over the age of 60 had a higher importance, compared to other sociodemographic factors.

From these classifications, it can be concluded that not all the considered sociodemographic factors can affect the risk of mortality after COVID-19 contraction. Based on the models, it can be said that people of older age (>=60) and of females have a higher chance of dying if contracted with COVID-19, compared to others with COVID-19.

REFERENCES

Miller, J., Fadel, R. A., Tang, A., Perrotta, G., Herc, E., Soman, S., . . . Suleyman, G. (2020). The impact of Sociodemographic Factors, comorbidities, and Physiologic responses on 30-day mortality in Coronavirus DISEASE 2019 (COVID-19) patients in Metropolitan Detroit. *Clinical Infectious Diseases*. doi:10.1093/cid/ciaa1420

Mocelin, H. J., Catão, R. C., Freitas, P. S., Prado, T. N., Bertolde, A. I., Castro, M. C., & Maciel, E. L. (2020). Analysis of the spatial distribution of cases of Zika virus infection and CONGENITAL Zika VIRUS syndrome in a state in the southeastern region of Brazil: Sociodemographic factors and implications for public health. *International Journal of Gynecology & Obstetrics, 148*(S2), 61-69. doi:10.1002/ijgo.13049

Muñoz-Navarro, R., Cano Vindel, A., Schmitz, F., Cabello, R., & Fernández-Berrocal, P. (2020). Emotional distress and ASSOCIATED Sociodemographic risk factors during the COVID-19 outbreak in Spain. doi:10.1101/2020.05.30.20117457

Papageorge, N. W., Zahn, M. V., Belot, M., Van den Broek-Altenburg, E., Choi, S., Jamison, J. C., & Tripodi, E. (2021). Socio-demographic factors associated With Self-protecting behavior during the Covid-19 pandemic. *Journal of Population Economics, 34*(2), 691-738. doi:10.1007/s00148-020-00818-x

Pasion, R., Paiva, T. O., Fernandes, C., & Barbosa, F. (2020). The age effect on protective behaviors during the covid-19 outbreak: Sociodemographic, perceptions and psychological accounts. *Frontiers in Psychology, 11*. doi:10.3389/fpsyg.2020.561785

Pullano, G., Valdano, E., Scarpa, N., Rubrichi, S., & Colizza, V. (2020). Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the COVID-19 epidemic in France Under Lockdown: A population-based study. *The Lancet Digital Health, 2*(12). doi:10.1016/s2589-7500(20)30243-0

Rivera-Izquierdo, M., Del Carmen Valero-Ubierna, M., R-delAmo, J. L., Fernández-García, M. Á, Martínez-Diz, S., Tahery-Mahmoud, A., . . . Jiménez-Mejías, E. (2020). Sociodemographic, clinical and LABORATORY factors on ADMISSION associated With COVID-19 mortality in hospitalized patients: A RETROSPECTIVE observational study. *PLOS ONE, 15*(6). doi:10.1371/journal.pone.0235107

Rozenfeld, Y., Beam, J., Maier, H., Haggerson, W., Boudreau, K., Carlson, J., & Medows, R. (2020). A model of disparities: Risk factors associated with covid-19 infection. doi:10.21203/rs.3.rs-31918/v2