UNIVERSITY OF BRITISH COLUMBIA - OKANAGAN CAMPUS

Faculty of Science

Department of Computer Science

Master of Data Science Capstone Project

Modeling and Visualization of the COVID-19 Outbreak in Ontario

For
Bruno St-Aubin and Marian Radulescu
Written by

Sofia Bahmutsky
KT Hobbs
Ngan Lyle
Shreeram Murali

June 2020

Contents

List of Figures

List of Tables

1 Executive Summary

Coronavirus disease, or COVID-19 is one of the most significant infectious diseases in the last 100 years. In Canada, the burden of disease is greatest in Quebec and Ontario. As of June 26, 2020, there have been greater than 34,000 cases and 2,600 deaths in Ontario alone. Individuals at high risk for poor outcomes include seniors and people with underlying health conditions and it has been estimated that up to 80% of deaths from COVID-19 in Ontario have occurred in residents of long term care (LTC) homes. Therefore, the purpose of this project was twofold; first, we analyzed health and proximity factors that are associated with the burden of disease among different Public Health Unit (PHU) regions of Ontario and second, we explored and modeled LTC home characteristics and quality indicators that are associated with COVID-19 outbreaks.

Data and software tools used in this project are exclusively open source. Data were taken from Statistics Canada and Government of Ontario websites. Data scraping, wrangling and cleaning were performed using python, statistical analysis was performed using R, and map generation and manipulation was performed using QGIS. Visualizations were implemented using JavaScript and D3 and are hosted on Github Pages.

Using clustering methods, principal component analysis and principal component regression, we showed that Ontario PHU regions with higher measures of proximity to various amenities (e.g. transit, healthcare and employment) had higher proportions of COVID-19 cases. On the other hand, regions that have a greater proportion of individuals with underlying health problems had a lower proportion of cases. When assessed individually using LASSO and beta regression, "amenity-richness" appears to be as influential as health factors with respect to the proportion of COVID-19 cases. LASSO also highlighted the importance of several health conditions that were associated with COVID-19 proportions within a PHU, such as the effect of physical activity and having a regular healthcare provider.

The probability of a COVID-19 outbreak at Ontario LTC homes was modeled with a binary logistic regression. The factors that were positively associated with an outbreak are number of beds and total number of complaints. The factors that were negatively associated with an outbreak are the total number non-complaint inspections and municipal home type.

Ontario PHU regions with greater "amenity-richness" and a greater proportion of healthy individuals according to measures used in this study, have had greater proportions of COVID-19 cases. LTC homes with a greater number of beds and complaints filed were more likely to have had a COVID-19 outbreak. Interestingly, municipal type LTC homes and those with a greater number of non-complaint inspections were less likely to have experienced an outbreak. Altogether, this project showcased how leveraging various open data sources can produce comprehensive and meaningful results.

2 Introduction

2.1 Areas of Investigation

Since its initial appearance in China on December 31, 2019, Coronavirus disease, or COVID-19, has evolved to a global pandemic - a novel disease to which there is little pre-existing immunity, that becomes epidemic in many countries. As of June 26, 2020, there have been greater than 9 million confirmed cases and greater than 480,000 deaths worldwide (World Health Organization, 2020). In Canada there have been greater than 100,000 cases and 8500 deaths. Roughly one third of these cases have been in Ontario, the Canadian province with the second greatest number of cases and fatalities (Government of Canada, 2020). In this Capstone project we analyze the COVID-19 outbreak in Ontario on two levels. First we explore the factors affecting the rates of disease in the different Public Health Unit (PHU) regions and second we examine the factors associated with the probability of an outbreak in different long term care homes.

The province of Ontario is divided into 34 PHU regions for administration of health promotion and disease prevention programs. These regions have experienced different levels of COVID-19 activity, which may be related to characteristics of individual regions, such as urban connectedness and the prevalence of various health conditions. For example, Northern Ontario is much less populated and urbanized than Central Ontario, which may be a factor in determining the extent of COVID-19 spread. Also, given that individuals at the highest risk for severe illness are seniors over the age of 65 as well as those with underlying health conditions such as, chronic obstructive pulmonary disease (COPD), asthma and diabetes (Centers for Disease Control and Prevention, 2020), the prevalence of these conditions in a health region may be determinants of disease activity and impact. In view of this, we conducted an exploratory statistical analysis to better understand the proximity and health factors associated with the distribution COVID-19 in Ontario PHU regions.

As the pandemic has progressed in Canada, it has become clear that seniors who live in long term care (LTC) homes shoulder a disproportionate burden of COVID-19 disease. Currently almost half of the 625 LTC homes operating in Ontario have experienced an outbreak. Moreover, it has been estimated that up to 80% of COVID-19 fatalities in Canada have occurred among seniors living in long term care homes (Willms & Montgomery, 2020). Thus, another aim of this Capstone project is to investigate the LTC home characteristics and quality indicators that are associated with the probability of an outbreak.

2.2 Research Questions

1. Which measures of health and proximity to amenities best characterize a Public Health Unit in Ontario? Of these, is there an association with COVID-19 prevalence in a particular region?

2. Are any of the publicly available LTC home characteristics and quality markers associated with an increased probability of a COVID-19 outbreak?

2.3 Objectives and Deliverables

The following objectives were adapted from our original proposal to Statistics Canada and had expanded in scope throughout the duration of the project:

1. To produce an inferential statistical model of factors that may be associated with the probability of COVID-19 outbreaks in different LTC homes in Ontario.

2. To produce an inferential statistical model of the proximity and health factors that may be associated with COVID-19 disease activity at the level of the PHU regions in Ontario.

3. To produce an interactive web page using QGIS and D3 to visualize the results from both PHU region analysis and LTC homes analyses.

4. To leverage open data sources in meaningful, exploratory analyses.

Similarly, our deliverables expanded to produce the following:

1. A database of LTC homes in Ontario.

2. A method to aggregate proximity data to the PHU region level and an aggregated database containing joined data from multiple sources for Ontario.

3. Several inferential models that assess the importance of predictor variables in relation to COVID-19 disease activity in different PHU regions and to the probability of COVID-19 outbreaks in LTC homes.

4. An interactive web page with visualizations integrating information about PHU regions, LTC homes, significant factors and the COVID-19 outbreak in Ontario.

5. A report including exploratory analyses of patterns and/or connections between the PHU regional data, LTC homes data, and the COVID-19 data.

3 Background

Although COVID-19 is a novel virus, pandemics of the past have led to research about modelling the spread of diseases dating back to the 18th century with Bernoulli's probability theory (Mansnerus, 2014). Throughout the 19th and 20th centuries, more research was conducted from mathematical, statistical, medical, and molecular perspectives on disease spread. From the early 1990's to early 2000's, computational research emerged allowing scientists to better model disease transmission. Such models included demographics, contact sites, classroom sizes, and age groups as predictor variables (Mansnerus, 2014). Research from the United Kingdom has had significant contributions to spatial modelling of disease outbreaks, the 100 year anniversary of the 1918 flu being the motivation of the project (Klepac et al., 2018). The researchers used data from a smartphone tracking app to record data of human movement for development of their model which showed the impact of human movement, contact patterns, and the type of location on disease transmission.

In regard to the current pandemic, medical research and development of a vaccine or therapy has been an imminent focus. Other areas involve methods to develop detection and prevention of COVID-19 transmission. Although an overwhelming majority of COVID-19 research has been in the medical sphere, there has also been a good deal of statistical and mathematical model development. This has also been studied at the provincial level of Ontario, conducted by numerous research teams and universities including, Queen's University, University of Guelph, and University of Toronto (Tuite et al., 2020). One of the most commonly implemented epidemiological compartmental models, a model that observes disease transmission, is Susceptible Infected Recovered (SIR), which makes use of differential equations to estimate spread based on susceptible, infected, and recovered individuals in a population (Bacaër, 2011).

Currently, there is no record of peer-reviewed publications regarding COVID-19 in Ontario; however, available research largely focuses on modeling spread of the disease in Ontario with various intervention strategies (Tuite et al., 2020). Of those, few investigate outbreaks in long

term care homes and none were found that explore the incidence of COVID-19 on the PHU level. Interestingly, multiple publications leveraged open data from the Government of Ontario's Ministry of Long-term Care (Government of Ontario, 2020), which is one of the sources used in this paper. One study conducted at the University of Toronto models temporal trends of mortality counts in long term care homes to assess an LTC home's change in risk over time; nonetheless, their scope was limited to counts of positive tests and fatalities over a short time frame of 10-days and did not assess other attributes of the home, such as quality indicators, home type (for-profit, non for-profit, or municipal), or home size (Fisman et al., 2020). Another study performed at Mount Sinai Hospital using the same data source modeled the size of outbreak and number of deaths in an LTC home according to home type and reported that for-profit homes are associated with a greater outbreak size and number of fatalities (Stall et al., 2020). With limited research dedicated to understanding profiles of LTC homes and PHU regions with COVID-19 outbreaks, we offer a novel investigation.

4 Data, Tools and Resources

4.1 Data

Table 1: **Summary of LTC homes data used in the project methods.** The table references the source of the data as well as the link to the data web page if applicable.

Data	Source
COVID-19 in LTCs	Government of Ontario
LTC profiles	Public Reporting
Health Quality	Health Quality Ontario

Table 2: **Summary of PHU regions data used in the project methods.** The table references the source of the data as well as the link to the data web page if applicable.

Data (name in project repository)	Source
Open Database for Health Facilities (ODHF)	Statistics Canada - LODE - Open Databases - Open Databases of Healthcare Facilities
Dissemination Block ArcInfo shapefiles (used only in QGIS for methods, not included in repository due to large file size)	Statistics Canada - Census Program - Geography - Boundary Files
Proximity Measures (PMD-en, amenity-score)	Statistics Canada - Data Visualization Products - Proximity Measures Data Viewer
Health Indicators by Health Region (file was too large to be uploaded to Github)	Statistics Canada
Ontario COVID-19 cases (ON_cases.csv)	Government of Ontario

4.2 Software Tools

The software tools used for this project are exclusively open source. Python and QGIS were employed for data retrieval, wrangling, aggregation and cleaning. Web scraping was performed in python using the packages 'selenium', 'requests' and 'beautifulsoup'. Statistical analysis was performed in R and visualizations were implemented using QGIS, HTML, CSS, and JavaScript/D3.

4.3 Terminology

This project involves some familiarity with public health terms, Statistics Canada census definitions as well as other domain specific terms. Further information about the relevant abbreviations and terminology can be found in Appendix A, Table 7.

4.4 Data Sources

All data used in this project are open-source. Data were retrieved from the open data portal of Statistics Canada or scraped from public Government of Ontario websites.

5 Methodology

Our group members have various academic backgrounds, allowing each to bring unique strengths to the project. Kaitlyn and Ngan performed majority of web scraping, data wrangling and cleaning. Shreeram assisted in modifying web scraping code to suit Windows machines. Sofia had experience and expertise in mapping/GIS and manipulated PHU data in QGIS. Kaitlyn has previous experience with HTML and CSS and was involved in developing the visualizations and web page with the guidance of Bruno St-Aubin. Statistical analysis, interpretation of results, research, and writing was completed by Ngan, Kaitlyn, and Sofia. This project was completed over a period of 10 weeks, including proposal development and final reporting. See Appendix A for a formal schedule break-down (Figure 12).

5.1 Web Scraping

5.1.1 LTC Homes Data

Three online data sources were web scraped and joined to make up the LTC homes database.

1. General LTC homes data including home characteristics and the number and type of inspection reports for each home was scraped from the Reports on Long-Term Care Homes website. While there are 651 homes in the database, information was retrieved for only the 625 LTC homes that appeared to be fully operational. Inconsistent or incomplete entries were resolved by manual review of the flagged homes. Most of the homes that were excluded are closed. A full list of the variables that were web scraped is shown in Appendix A, Table 10.

2. Quality indicators for each LTC home was downloaded from the Health Quality Ontario website. Complete quality data were available for 615 of the 625 homes from the general database. A full list of the available quality measures is also shown in Appendix A, Table 10.

3. COVID-19 outbreak data were web scraped directly from the Government of Ontario 'How Ontario is responding to COVID-19' website. Homes that have not had an outbreak were not represented. Homes with an outbreak were classified as being either active or no longer in an outbreak. Homes classified as having an active outbreak include numerical data on the number of beds, of confirmed staff cases, of confirmed resident cases and of resident deaths. However, homes classified as no longer in an outbreak only include the number of beds and resident deaths rendering a loss of data when the outbreak is considered resolved.

All of the above data sources are maintained by or associated with the Government of Ontario Ministry of Health and Long-Term Care.

5.1.2 PHU Regions Data

Data on the PHU level was primarily retrieved from Statistics Canada with the exception of the Ontario COVID-19 case data, which was obtained from the Government of Ontario. Ontario cases were reported by a PHU and included patient gender, case acquisition information, outcome of disease, patient age group and the name of the reporting PHU. Case data was aggregated to achieve summaries for each PHU. Proximity measures were determined by Statistics Canada and calculated for each dissemination block. There are 10 proximity measures in total and one amenity density score, which evaluates overall "amenity-richness" of a dissemination block. Health data was reported at the PHU region level by Statistics Canada for 2016. Discrepancies existed between health regions and reporting PHU names, which were later resolved during data wrangling.

5.2 Statistical Analysis

In agreement with our two research questions and objectives, statistical methods were partitioned into the analysis of COVID-19 outbreak at the LTC homes level, and the analysis of COVID-19 rates at the PHU region level. As per our discussions with our Capstone partner, statistical analyses were designed to be inferential rather than predictive.

5.2.1 LTC Homes Statistical Analysis

Table 3: **Clustering of LHIN regions into 5 geographic regions for statistical analysis.**

Geographic Regions	LHIN
West	Erie-St. Clair, South West, Hamilton Niagara Haldimand Brant, Waterloo Wellington
Central	Mississauga Halton, Central West, Central, North Simcoe Muskoka
Toronto	Toronto Central
East	Central East, South East, Champlain
North	North West, North East

Count data and numeric data that were right skewed were either log or square root transformed. Following the initial binary logistic regression, variable selection was performed using backwards selection.

5.2.2 PHU Statistical Analysis

Statistical analysis consisted of two orthogonal approaches: unsupervised methods, and supervised methods.

Unsupervised Methods

Amenity score and comorbidities for each PHU were scaled and the euclidean distance was computed between each PHU. Using the 'hclust()' function in R, complete, agglomerative clusters were formed on the basis of least dissimilarity between PHU regions. This method was unsupervised because COVID-19 case proportions were excluded in the distance calculations. Resulting PHU clusters were visualized in a bar plot that was ordered by COVID-19 case proportion.

Supervised Methods

LASSO was utilized as a method of dimensionality reduction and feature selection for the dataset. The library 'glmnet' was used in R for this task. LASSO stands for Least Absolute Shrinkage and Selection operator. It is a type of regression method that simultaneously performs variable selection and regularization. LASSO models "shrink" coefficient values to zero if they are not sufficient at explaining the response variable. LASSO is also a robust method for dealing with multi-collinearity between predictor variables.

Results from the LASSO were subsequently integrated into a generalized linear model, namely a beta regression model. The library 'betareg' in R was used for this task. Resulting output of the beta regression was visualized as a scatterplot of COVID-19 case proportions explained by the most significantly influential predictors. Beta regression is appropriate for modelling when the response variable is a measure which varies from 0 to 1, but when no observations can equal exactly 0 or 1. It is also appropriate for data that is a proportion of discrete counts. The response variable for our model is COVID-19 proportion of cases or fatalities, which is constrained between 0 and 1.

5.3 Web Page

JavaScript, HTML, CSS and D3 were used to create a web page with interactive choropleth maps showing the geospatial distribution of COVID-19 disease activity for PHU regions and LTC homes in Ontario. We were assisted by one of our Capstone project contacts at Statistics Canada, Bruno St-Aubin, who is an expert in web GIS development. Features include the ability to zoom, a linked chart, and a tooltip with additional metadata on hover. The web page is hosted on Github Pages.

6 Results

6.1 LTC Homes Analysis

The results of the binary logistic regression followed by backwards selection for variable selection is shown in Table 4.

Table 4: **Logistic regression results.**

Variable	Estimate	Standard Error	P-value
Intercept	-1.608	0.257	4×10^{-10}
Number of beds	0.012	0.002	1.41×10^{-10}***
Total number of complaints	0.045	0.013	0.00065***
Total number of non-complaints	-0.025	0.257	0.034*
Municipal Home Type	-0.566	0.011	0.025*
Non-profit home type	0.167	0.267	0.437
For-profit home type (ref)			

McFadden $R^2 = 0.14$, **P-value** $= 0$

The variables 'number of beds' and 'total number of complaints' were positively associated with the likelihood of a COVID-19 outbreak. On the other hand, 'municipal home type' and the 'total number of non-complaint inspections' was negatively associated with the likelihood of a COVID-19 outbreak. The relationship between the variables and the log odds of an outbreak

is shown by the following equation:

$$\text{Log odds of an outbreak} = -1.608 + 0.012 * X1 + 0.045 * X2 - 0.025 * X3 - 0.566 * X4 \quad (1)$$

Where

- X1 = Number of beds
- X2 = Total complaints
- X3 = Total non-complaints
- X4 = Municipal home type

A visualization of the model results with LTC homes ordered on the x-axis by the probability of an outbreak is shown in Figure 1.

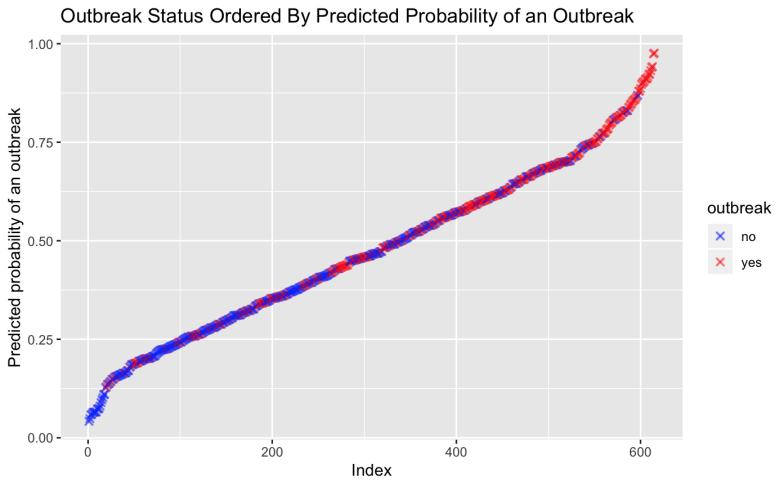

Figure 1: **Outbreak status ordered by predicted probability of an outbreak.**

6.2 PHU Regions Analysis

6.2.1 Hierarchical clustering

Agglomerative hierarchical clustering reported a coefficient of 0.67, indicating a somewhat sufficient capture of true clustering structure in the data (Rousseeuw et al., 1985). The resulting dendrogram suggested two distinct clusters, which seemed to represent Public Health Units whose COVID-19 case proportions were in the upper half of the distribution (Figures 2, 3). This observation implies that it may be possible to characterize PHU regions with higher case proportions on the basis of health and proximity measures.

Figure 2: **Agglomerative complete hierarchical linkage forming 2 clusters.**

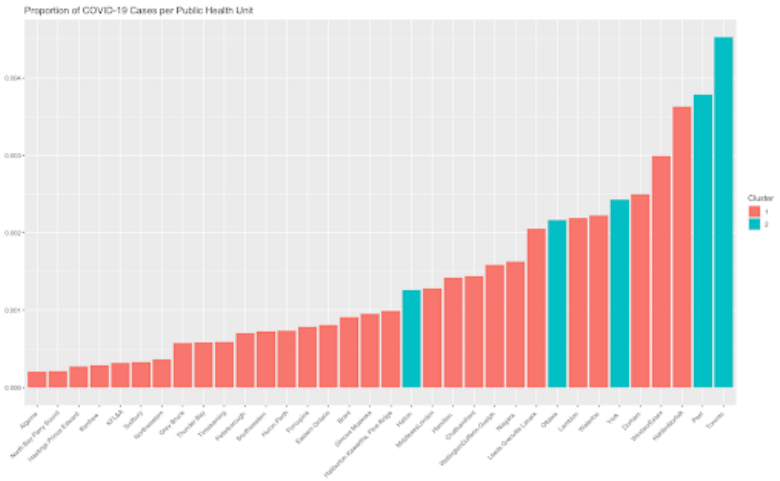

Figure 3: **Clusters ordered by PHU proportion of COVID-19 cases.**

6.2.2 Principal Component Analysis and Regression

Principal component analysis was performed on the predictor variables as a method of dimensionality reduction and to assess for underlying factors that might be associated with COVID-19 activity. Selected principal components (PCs) were then used as independent predictors in two multiple linear regression models with the proportion of COVID-19 cases and fatalities as the outcomes respectively. The top 6 PCs were retained for the regression based on the result of the scree plot shown in Figure 4.

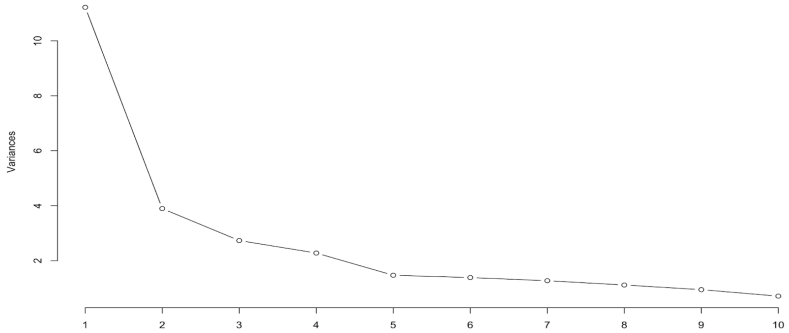

Figure 4: **Clusters ordered by PHU proportion of COVID-19 cases.**

The top 6 PCs cumulatively explain more than 70% of the variation in the data. The results of the linear regression with the proportion of cases in PHU regions as the response variable are shown below in Table 5. The results of the linear regression with the proportion of fatalities in the PHU regions as the response variable is shown in Table 6.

Table 5: **Results of principal component linear regression on the proportion of cases.**

Principal Component	Estimate	Standard Error	P-value
Intercept	1.378×10^{-3}	1.351×10^{-4}	8.62×10^{-11}***
PC1	2.258×10^{-4}	4.095×10^{-5}	7.71×10^{-6}***
PC2	-1.340×10^{-4}	6.950×10^{-5}	0.0644
PC3	2.028×10^{-4}	8.299×10^{-5}	0.0213*
PC4	-1.319×10^{-5}	9.090×10^{-5}	0.8857
PC5	7.295×10^{-5}	1.132×10^{-4}	0.5247
PC6	9.832×10^{-5}	1.166×10^{-5}	0.403115

Adjusted R2 = 0.5163, **P-value** = 0.0001648

Table 6: **Results of principal component linear regression on the proportion of fatalities.**

Principal Component	Estimate	Standard Error	P-value
Intercept	1.034×10^{-4}	1.489×10^{-5}	1.84×10^{-7} * **
PC1	1.907×10^{-5}	4.513×10^{-6}	0.000243 ***
PC2	1.536×10^{-5}	7.659×10^{-6}	0.055089
PC3	5.215×10^{-6}	9.146×10^{-6}	0.573286
PC4	6.112×10^{-6}	1.002×10^{-5}	0.546887
PC5	3.847×10^{-6}	1.247×10^{-5}	0.760142
PC6	1.091×10^{-5}	1.247×10^{-5}	0.403115

Adjusted R2 = 0.3451, **P-value** = 0.006226

As shown above, PC1 was significantly associated with both the proportion of COVID-19 cases and the proportion of COVID-19 fatalities among different PHU health regions. Interestingly, PC1 loaded positively for most of the proximity measures as well as for most of the health measures of "good health". PC1 loaded negatively for all of the health measures of "poor health". PC3 was also significantly associated (to a lesser degree) with the proportion of

COVID-19 cases, but how the loadings on PC3 should be interpreted is less clear. The loadings for each variable on PC1, PC2 and PC3 are shown in Appendix B, Table 11.

The PCA biplot below (Figure 5) shows the variables appear to cluster into three groups:

1. Variables indicating greater proximity scores or "amenity richness"
2. Variables indicating "good health"
3. Variables indicating "poor health"

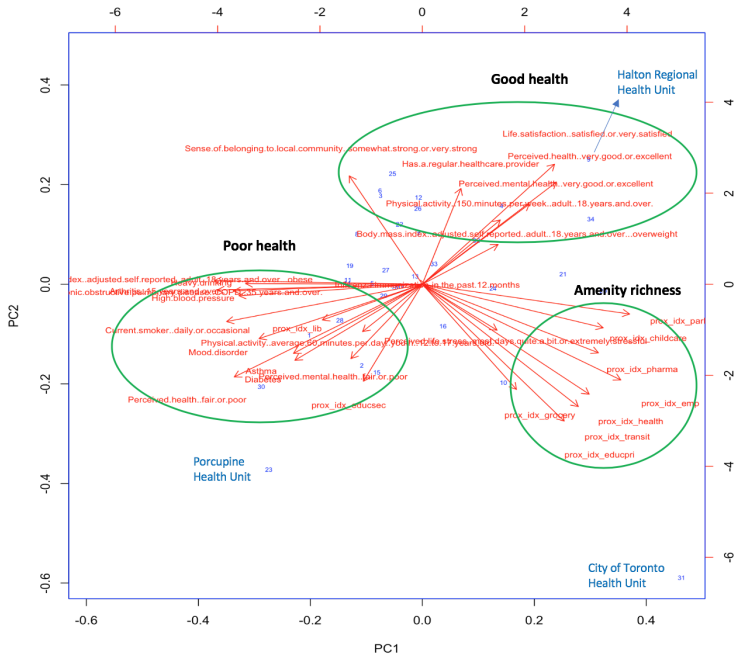

Figure 5: **Principal components analysis biplot.**
Each of the blue numbers represent a PHU. The location of the PHUs on the plot relative to each other indicates how similar/different the PHUs are with respect to the factors in the analysis. Selected PHUs that are outliers have been labeled. The red arrows each represent a predictor variable. The length and direction of the arrow indicates how strongly a predictor influences PC1 and PC2, respectively. Predictors with arrows that cluster together with small angles between them indicate that those predictors are correlated with each other.

6.2.3 LASSO and Beta Regression

Supervised analysis of the data at the PHU level involved LASSO and beta regression. Since there were a large number of predictor variables, the LASSO model was used to assist in feature selection. Based on the graphical plot from LASSO and in combination with cross-validation measurements, it was determined that 7 predictor variables were sufficient for explaining the variation of the response variable. Graphical plots are displayed in Appendix B, Figures 16, 17. Numerical output of the LASSO shows the importance of the various predictor variables, see Figure 6, below. LASSO relies on lambda, a tuning parameter that influences how many predictors will be retained by the LASSO. When lambda is equal to zero, the LASSO simply performs an ordinary least squares regression. The appropriate value for lambda, found by cross-validation, was 9.184336e-05. Using this as the lambda value results in the model with the lowest mean-squared error (MSE). The retained predictors are: arthritis, BMI - obese, diabetes, heavy drinking, physical activity (150 mins/week), strong sense of belonging to the community, and amenity density score.

```
23 x 1 sparse Matrix of class "dgCMatrix"
                                                                                  1
(Intercept)                                                          5.398123e-03
Arthritis..15.years.and.over.                                       -9.799490e-04
Asthma                                                               .
Body.mass.index..adjusted.self.reported..adult..18.years.and.over...obese      -2.498881e-03
Body.mass.index..adjusted.self.reported..adult..18.years.and.over...overweight  .
Chronic.obstructive.pulmonary.disease..COPD..35.years.and.over.                  .
Current.smoker..daily                                                            .
Current.smoker..daily.or.occasional                                              .
Diabetes                                                             3.516130e-03
Has.a.regular.healthcare.provider                                                .
Heavy.drinking                                                      -1.531532e-02
Influenza.immunization.in.the.past.12.months                                     .
Life.satisfaction..satisfied.or.very.satisfied                                   .
Mood.disorder                                                                    .
Perceived.health..fair.or.poor                                                   .
Perceived.health..very.good.or.excellent                                         .
Perceived.life.stress..most.days.quite.a.bit.or.extremely.stressful              .
Perceived.mental.health..fair.or.poor                                            .
Perceived.mental.health..very.good.or.excellent                                  .
Physical.activity..150.minutes.per.week..adult..18.years.and.over.  -4.452048e-04
Physical.activity..average.60.minutes.per.day..youth..12.to.17.years.old.        .
Sense.of.belonging.to.local.community..somewhat.strong.or.very.strong -3.515305e-05
amenity                                                              1.971229e-03
```

Figure 6: **Output from LASSO using 'glmnet' library in R.**
Predictor variables are the left-hand side column, and corresponding coefficients are shown on the right-hand column.

A beta regression was fit to the seven predictor variables, and the response variable of COVID-19 proportion of cases for the varying PHU's. Using the "loglog" as the link function increased the model accuracy. The beta regression model is shown below (Figure 7).

```
Coefficients (mean model with loglog link):
                                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                        -1.25653    0.32368  -3.882 0.000104 ***
Arthritis..15.years.and.over.                                      -0.74702    0.50685  -1.474 0.140523
Body.mass.index..adjusted.self.reported..adult..18.years.and.over...obese -0.13837    0.37517  -0.369 0.712254
Diabetes                                                            0.72335    0.86424   0.837 0.402604
Heavy.drinking                                                     -1.17542    0.57746  -2.036 0.041800 *
Physical.activity..150.minutes.per.week..adult..18.years.and.over. -0.07494    0.35674  -0.210 0.833616
Sense.of.belonging.to.local.community..somewhat.strong.or.very.strong -0.29973    0.37015  -0.810 0.418079
amenity                                                             0.12383    0.10299   1.202 0.229250

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)     3264        815   4.005  6.2e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 208.7 on 9 Df
Pseudo R-squared: 0.6251
Number of iterations: 1417 (BFGS) + 11 (Fisher scoring)
```

Figure 7: **Beta regression model output, modelling the significant LASSO predictor variables versus the response variable COVID-19 case proportions across the PHU's.**

Based on the regression model output, only the 'Heavy Drinking' variable is significantly associated with the COVID-19 case proportions. The model has a pseudo-R squared value of 0.625, which suggests that a good deal of the variation in the data is explained by the model.

The regression can be modelled by the following equation:

$$loglog(Y) = -1.256 - 1.18 * (X1) \tag{2}$$

where Y represents COVID-19 total case proportion, and $X1$ represents the proportion of the population in the PHU which participates in heavy drinking regularly. The $loglog(Y)$ can be interpreted as: $loglog(Y) = log(log(Y))$.

Displayed below is a visual representation of the relationship between the proportion of a PHU that participates in heavy drinking on a regular basis with respect to the proportion of COVID-19 cases within a PHU (Figure 8). By observing the scatter plot, there are some obvious outliers, Toronto and Hamilton stand out as PHU with high amenity scores. Toronto has the

highest proportion of COVID-19 cases, and the highest amenity score of any PHU in Ontario, and is a noticeable outlier. On the other hand, Hamilton has a relatively high amenity score, second to Toronto but it does appear as such an obvious outlier: it falls within the general scatter of points surrounding it. Overall, there seems to be a general negative trend between the variables, PHU's which have a higher proportion of heavy drinkers, tend to also be areas with lower COVID-19 case proportion.

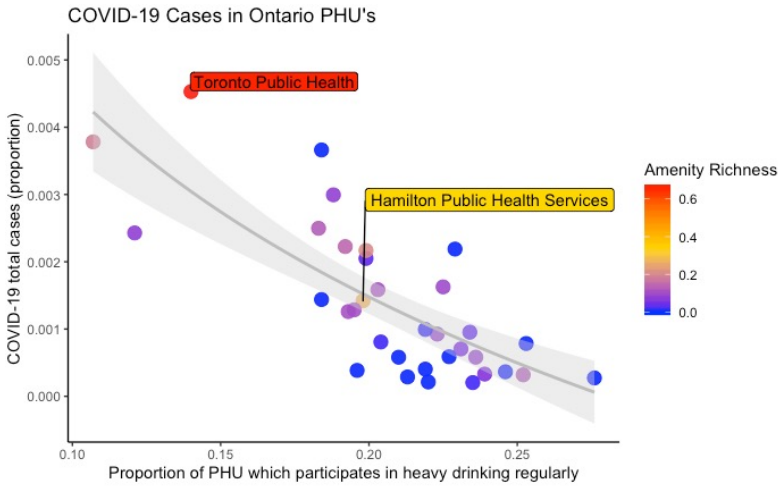

Figure 8: **Scatter plot showing the relationship between PHU and heavy drinking.** The proportion of a PHU that participates in heavy drinking on a regular basis with respect to the proportion of COVID-19 cases within a PHU. The points are coloured according to their amenity richness score. Gray fitted line is fit by default 'glm' function in ggplot. The gray shaded area represents the 95% confidence region.

In addition to investigating the COVID-19 proportion of cases within the PHU's, the number of fatalities from COVID-19 was also obtained in the data set, this was converted into a proportion in a similar manner to the COVID-19 cases proportion. The same steps were repeated using the fatalities proportion as the response variable. Based on the graphical plot from LASSO and in combination with cross-validation measurements (see Appendix B, 18, 19), it was determined that four predictor variables were sufficiently important and influential for the model. Numerical output of the LASSO shows the importance of the various predictor variables, see Figure 9, below. The appropriate value for lambda, found by cross-validation, was 1.782782×10^{-5}. Using this as the lambda value results in the model with the lowest mean-squared error (MSE). The corresponding number of predictors that are retained when using this lambda value is four. Interestingly, retained predictor variables for the COVID-19 fatalities proportion response coincide with some of the retained predictor variables for the previous analysis of COVID-19 proportions of cases.The retained predictors are: BMI - obese, heavy drinking, strong sense of community belonging, and amenity density score.

```
23 x 1 sparse Matrix of class "dgCMatrix"
                                                                                      1
(Intercept)                                                                0.0007364359
Arthritis..15.years.and.over.                                              .
Asthma                                                                     .
Body.mass.index..adjusted.self.reported..adult..18.years.and.over...obese -0.0003737596
Body.mass.index..adjusted.self.reported..adult..18.years.and.over...overweight .
Chronic.obstructive.pulmonary.disease..COPD..35.years.and.over.            .
Current.smoker..daily                                                      .
Current.smoker..daily.or.occasional                                        .
Diabetes                                                                   .
Has.a.regular.healthcare.provider                                          .
Heavy.drinking                                                             -0.0006730223
Influenza.immunization.in.the.past.12.months                               .
Life.satisfaction..satisfied.or.very.satisfied                             .
Mood.disorder                                                              .
Perceived.health..fair.or.poor                                             .
Perceived.health..very.good.or.excellent                                   .
Perceived.life.stress..most.days.quite.a.bit.or.extremely.stressful        .
Perceived.mental.health..fair.or.poor                                      .
Perceived.mental.health..very.good.or.excellent                            .
Physical.activity..150.minutes.per.week..adult..18.years.and.over.         .
Physical.activity..average.60.minutes.per.day..youth..12.to.17.years.old.  .
Sense.of.belonging.to.local.community..somewhat.strong.or.very.strong      -0.0005205899
amenity                                                                    0.0000444202
```

Figure 9: **Output from LASSO using 'glmnet' library in R.**
Predictor variables are the left-hand side column, and corresponding coefficients are shown on the right-hand column.

A slight difference within the fatalities from COVID-19 and the cases of COVID-19 is that not all PHU's had any recorded fatalities due to COVID-19. Since beta regression assumes that all response variable measurements are confined to the interval (0,1), a transformation was applied to the data (a small, near zero value was added to all response variable numbers). A beta regression was fit to the data, creating a model which included 1 predictor variable, and the response variable of COVID-19 fatalities proportion for the varying PHU's. The included predictor variables were determined to be influential from the LASSO output: BMI - obese, heavy drinking, sense of community belonging, and amenity density. Using the "loglog" as the link function increased the model accuracy. The beta regression model output is shown below (Figure 10).

```
Coefficients (mean model with loglog link):
                                                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                     -1.230843   0.251221  -4.899 9.61e-07 ***
Body.mass.index..adjusted.self.reported..adult..18.years.and.over...obese -0.258279   0.324577  -0.796  0.42618
Chronic.obstructive.pulmonary.disease..COPD..35.years.and.over. -1.199511   0.888778  -1.350  0.17714
Heavy.drinking                                                  -0.446035   0.465326  -0.959  0.33779
Sense.of.belonging.to.local.community..somewhat.strong.or.very.strong -1.038029   0.345623  -3.003  0.00267 **
amenity                                                          0.009807   0.094368   0.104  0.91723

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)    20824       5413   3.847  0.00012 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 288.2 on 7 Df
Pseudo R-squared: 0.5583
Number of iterations: 5000 (BFGS) + 16 (Fisher scoring)
```

Figure 10: **Beta regression model output, modelling the significant LASSO predictor variables versus the response variable COVID-19 fatality proportions across the PHU's.**

Based on the regression model output, only 'strong sense of community belonging' is significantly associated with the COVID-19 fatalities proportions. The model has a pseudo-R squared value of 0.558, which suggests that a good deal of the variation in the data is explained by the model. The regression can be modelled by the following equation:

$$loglog(Y) = -1.231 - 1.038 * (X1) \tag{3}$$

where Y represents COVID-19 fatalities proportion, and $X1$ represents the proportion of the

population in the PHU which reports having a very strong sense of belonging within their local community. The $loglog(Y)$ can be interpreted as: $loglog(Y) = log(log(Y))$.

Displayed below is a visual representation of the relationship between the proportion of a PHU which reports a strong sense of belonging within their community with respect to the proportion of COVID-19 fatalities within a PHU (Figure 11). By observing the scatter plot, the points representing Toronto and Hamilton are obviously visible (red and yellow, the points with highest amenity score, as discussed in previous scatter plot), however they do not appear as outliers in this graph. They both generally fall within the scatter of points surrounding them. Overall, there seems to be a general negative trend between the variables, PHU's which have a higher proportion of individuals reporting a strong sense of belonging within their community, tend to also be areas with lower COVID-19 fatalities.

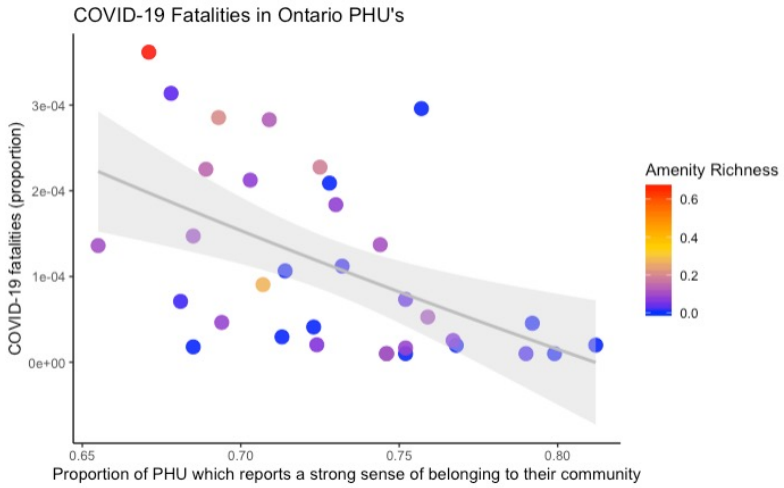

Figure 11: **Scatter plot showing the relationship PHU and a strong sense of belonging to community.**
The proportion of a PHU that reports a strong sense of belonging within their community with respect to the proportion of COVID-19 fatalities within a PHU. The points are coloured according to their amenity richness score. Gray fitted line is fit by default glm function in ggplot. The gray shaded area represents the 95% confidence region.

7 Interpretation and Discussion

7.1 LTC Homes Analysis

It has been suggested that the current COVID-19 crisis in LTC homes has been many years in the making, a consequence of chronic under funding, neglect, and ineffective models of care both in Canada and elsewhere (Werner et al., 2020). However the individual factors and mix of factors that beget quality, or that put a home at risk of a devastating COVID-19 outbreak in the medium term, are unclear. In the long term, smaller scale family style homes with small numbers of residents have been put forth as being a higher quality alternative to the current model of institutional care. In the medium term, with respect to the current pandemic, a study examining the COVID-19 outbreak in LTC homes in Ontario showed that a greater number of beds and location in a high prevalence region increases the risk of an outbreak. For-profit status was not associated with likelihood of an outbreak but was associated with the size of an outbreak and the number of resident deaths (Stall et al., 2020). Our analysis adds new information to the existing knowledge, showing that in addition to the number of beds, the total number complaint reports is also associated with the likelihood of an outbreak. Moreover, municipal non-profit home-type and the total number of non-complaint reports are protective.

7.2 PHU Analysis

In this study, open data from several Statistics Canada and Government of Ontario resources were successfully integrated and utilized in a meaningful, macro-level, exploratory analysis of COVID-19. Evaluating high proportions of COVID-19 using health indicators and degree of "connectedness" seem to suggest a profile of less healthy, highly connected Public Health Units, yet these associations are neither definitive or comprehensive. In fact, through LASSO and PCA, indicators of poor health, such as heavy drinking, were determined to be protective as they were prevalent in regions with low case proportions (Figure 8). Toronto Public Health may have low proportions of "heavy drinking" yet is the most connected health unit (Figure 11). As a result of the associative nature of this research, the results may misleadingly suggest that health indicators are protective. Heavy drinking may also be confounding to a latent variable that was not captured in this study. Further investigation with additional data may be more telling.

Similarly, investigation of COVID-19 fatality proportion showed a relationship between regions with people reporting a strong sense of belonging and lower COVID-19 fatality proportion. This also suggests that having strong community sense is protective; however, as stated previously, this may be related to a confounding or latent variable which we are not aware of, or had no way of capturing with our data.

8 Limitations

As research unfolded and more avenues of investigation came to light, the main hindrance for this project was lack of time. Moreover, some personal limitations with software slowed our progression; specifically, when working with QGIS and JavaScript/D3, which were both new tools and languages to our team with steep learning curves. Fortunately, we had guidance from our capstone partners to help resolve technical difficulties and better our abilities.

Another time-limiting factor we faced was wrangling data from various sources. Most data was only available as recently as 2016 while COVID-19 case data was recorded in real-time, throughout 2020. Since the 36 Public Health Units that Ontario was partitioned into in 2016, Public Health Units have evolved to merge two sets of regions, ultimately rendering 34 PHUs and including minor name changes (Marshall, 2019). Similar issues occurred with scraping of LTC information as, over time, some homes closed, renamed, or merged with others.

In the Public Health Unit analysis, COVID-19 case information in Ontario was limited and did not include a holistic patient profile with supplementary health or contact tracing information that would allow for statistical probabilities to be calculated. In effort to preserve privacy, cases were also reported on the PHU level, thus generalizing subsequent analysis. Some PHU regions are incredibly diverse and population dense. Such generalizations include assumptions and limit statistical analysis to interpretations of correlation or association. As a result, we turned to health indicators and amenity scores to merely characterize each unit and observe their associations with COVID-19 proportions.

Finally, it was originally of interest to investigate various response variables to determine the extent of outbreaks for long term care homes such as, extent of fatalities, or total number of cases in a home. Since data was lost when a home's outbreak status evolved from active to inactive, we were confined to investigating a binary response variable; whether a home had an outbreak or not.

9 Directions for Future Work

With the goal of providing a more holistic profile of Public Health Units in Ontario, additional census data from Statistics Canada had already been retrieved, wrangled and merged in preparation for statistical analysis. Due to uncertainty and time constraints, observations regarding these metrics were excluded from this report; however, a more comprehensive picture of each PHU might offer better insight to potential circumstances surrounding COVID-19 outbreaks. Additionally, generalized profiles may provide a template for other areas of investigation, such as, health-related policies.

Accordingly, supplementary visualizations to better demonstrate case-fatality ratios and additional important predictors could be incorporated to the web page.

Finally, replication of this study for other provinces could offer more robust interpretations of the associations between proximity, health indicators, census data and COVID-19 spread.

10 References

Bacaër, N. (2011). McKendrick and Kermack on epidemic modelling (1926–1927). In N. Bacaër (Ed.), A Short History of Mathematical Population Dynamics (pp. 89–96). Springer London. https://doi.org/10.1007/978-0-85729-115-8_16

Center for Disease Control and Prevention. (2020, June 25). People Who Are at Increased Risk for Severe Illness. Coronavirus Disease 2019 (COVID-19). https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-at-increased-risk.html ?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fneed-extra-precautions%2Fpeople-at-higher-risk.html

Fisman, D., Lapointe-Shaw, L., Bogoch, I., McCready, J., Tuite, A. (2020). Failing our Most Vulnerable: COVID-19 and Long-Term Care Facilities in Ontario [Preprint]. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2020.04.14.20065557

Government of Canada. (2020, June 25). Coronavirus disease (COVID-19): Outbreak update. https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.htmla1

Government of Ontario. (2020, June 26). How Ontario is responding to COVID-19. https://www.ontario.ca/page/how-ontario-is-responding-covid-19 Health Quality Ontario. (2020, June 26). Long-Term Care Home Performance in Ontario.

Klepac, P., Kissler, S., Gog, J. (2018). Contagion! The BBC Four Pandemic – The model behind the documentary. Epidemics, 24, 49–59. https://doi.org/10.1016/j.epidem.2018.03.003

Mansnerus, E. (2014). Modelling in Public Health Research: How Mathematical Techniques Keep Us Healthy. Palgrave Macmillan UK. https://books.google.ca/books?id=4lucBQAAQBAJ

Marshall, R. (2019, December 19). Hello Huron Perth Public Health! The Merger of Huron and Perth Health Units Takes Effect January 1, 2020. https://www.huroncounty.ca/news/hello-huron-perth-public-health-the-merger-of-huron-and-perth-health-units-takes-effect-january-1-2020/

Ontario Ministry of Health and Long-Term Care. (2020, June 26). Reports on Long-Term Care Homes. http://publicreporting.ltchomes.net/en-ca/Search_Selection.aspx

Rousseeuw, P. J., Mathematics, D. U. of T. D. of, Informatics. (1985). A Visual Display for Hierarchical Classification. Delft University of Technology. https://books.google.ca/books?id=e8bDnQEACAAJ

Stall, N. M., Jones, A., Brown, K. A., Rochon, P. A., Costa, A. P. (2020). For-profit nursing homes and the risk of COVID-19 outbreaks and resident deaths in Ontario, Canada [Preprint]. Geriatric Medicine. https://doi.org/10.1101/2020.05.25.20112664

Tuite, A. R., Fisman, D. N., Greer, A. L. (2020). Mathematical modelling of COVID-19 transmission and mitigation strategies in the population of Ontario, Canada. Canadian Medical Association Journal, 192(19), E497–E505. https://doi.org/10.1503/cmaj.200476

Werner, R. M., Hoffman, A. K., Coe, N. B. (2020). Long-Term Care Policy after Covid-19—Solving the Nursing Home Crisis. New England Journal of Medicine, NEJMp2014811. https://doi.org/10.1056/NEJMp2014811

Willms, J., Montgomery, H. (2020, June 25). Coronavirus Update: More than 80 per cent of Canada's deaths are in long-term care homes. The Globe and Mail. https://www.theglobeandmail

.com/canada/article-coronavirus-update-more-than-80-per-cent-of-canadas-deaths-are-in/

World Health Organization. (2020). Coronavirus disease (COVID-19) Situation Report – 158. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200626-covid-19-sitrep-158.pdf?sfvrsn=1d1aae8a_2

11 Appendices

11.1 Appendix A - Supplementary Data and Methods

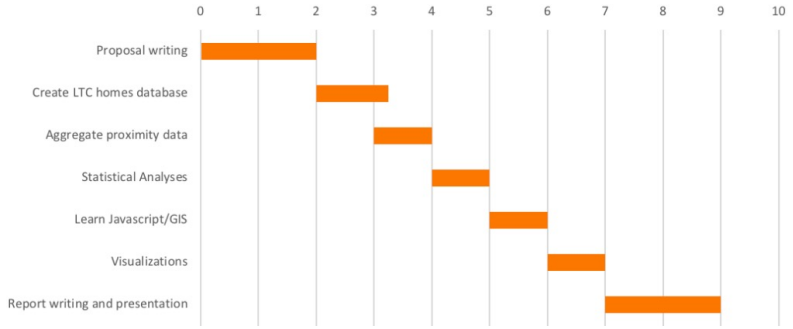

Figure 12: **Schedule of project tasks, broken down by week. April 26 to June 23 2020.**

Table 7: **Definitions and information about various terms and abbreviations which are used throughout the paper. Information is directly quoted and referenced from the source.**

Term	Meaning	Information	Source
DB	Dissemination Block	As defined by Statistics Canada: "A dissemination block (DB) is an area bounded on all sides by roads and/or boundaries of standard geographic areas. The dissemination block is the smallest geographic area for which population and dwelling counts are disseminated. Dissemination blocks cover all the territory of Canada."	Statistics Canada / Census
PHU	Public Health Unit	As defined by the Ontario government: "A Public Health Unit is an official health agency established by a group of urban and rural municipalities to provide a more efficient community health program, carried out by full-time, specially qualified staff." There are 34 PHU regions in Ontario.	Government of Ontario
LTC	Long Term Care	As defined by the Ontario government: "Long-term care homes are places where adults can live and receive help with most or all daily activities, have access to 24-hour nursing and personal care."	Government of Ontario.
LHIN	Local Health Integration Network	As defined by the Ontario government: "Local Health Integration Networks plan, integrate and fund local health care, improving access and patient experience." There are 14 LHIN regions in Ontario.	Government of Ontario
Proximity Measure		As defined by Statistics Canada: "Proximity measures are based on a simple gravity model that accounts for the distance between a reference dissemination block DB and all the DBs in which the service is located (within a given distance) and the size of the services. The measure accounts also for the presence of services within the DB of reference."	Statistics Canada
Amenity Score		As defined by Statistics Canada: "An aggregate measure was created to indicate neighbourhoods that have access to basic needs for a family with minors. A dissemination block with access to a grocery store, pharmacy, health care facility, child care facility, primary school, library, public transit stop, and source of employment.**Note:** 1 is referred to as an amenity dense neighbourhood. A high amenity density neighbourhood is defined as an amenity dense neighbourhood that has proximity measure values in the top third of the distribution for each of the eight proximity measures."	Statistics Canada

11.1.1 LTC Web Scraping

Additional Data Source Information

Table 8: **List of variables web scraped from the Reports on Long-Term Care Homes website**

A. Home profile information
1. Name
2. Address
3. LHIN
4. Licensee
5. Management
6. Home Type (For-profit, Non-profit, Municipal)
7. Number of Beds
8. Approved short stay beds
9. Residents' Council
10. Family Council
11. Accreditation
B. Home inspections information
1. Total number of inspections
- Total available
- Total in the last 5 years (since January 1, 2015)
- Total in the last 2 years (since January 1, 2018)
2. Total number of complaints inspections
- Total available
- Total in the last 5 years (since January 1, 2015)
- Total in the last 2 years (since January 1, 2018)
3. Total number of critical incident inspections
- Total available
- Total in the last 5 years (since January 1, 2015)
- Total in the last 2 years (since January 1, 2018)
4. Total number of inspections accompanied by an order(s) of the inspector
- Total available
- Total in the last 5 years (since January 1, 2015)
- Total in the last 2 years (since January 1, 2018)

Table 9: **List of quality indicators downloaded from the Health Quality Ontario website**

Quality Indicators
Antipsychotic Medication Use (%)
Pressure Ulcers (%)
Falls (%)
Physical Restraints Use (%)
Depression (%)
Pain (%)

Table 10: **Predictor variables for the logistic regression.**

Variable	Data type	Data values
Outbreak status	Binary	Yes or no
Home type	Categorical	For-profit, non-profit or municipal type
Short stay	Binary	Yes or no
Residents' council	Binary	Yes or no
Family council	Binary	Yes or no
Accreditation	Binary	Yes or no
Region	Categorical	West, Central, Toronto, East, North
Number of beds	Numeric	
Antipsychotic use percent	Numeric	
Depression percent	Numeric	
Falls percent	Numeric	
Pressure Ulcers percent	Numeric	
Pain percent	Numeric	
Total complaints	Numeric	
Complaints in the last 5 years	Numeric	
Complaints in the last 2 years	Numeric	
Total non-complaints	Numeric	
Non-complaints in the last 5 years	Numeric	
Non-complaints in the last 2 years	Numeric	
Total critical incidents	Numeric	
Critical incidents in the last 5 years	Numeric	
Critical incidents in the last 2 years	Numeric	
Total inspections with orders	Numeric	
Inspections with orders in the last 5 years	Numeric	
Inspections with orders in the last 2 years	Numeric	

Detailed Web Scraping

ltc-webscraping.ipynb

The following information applies to Linux machines. In order to use 'selenium', chromium and chromedriver must be installed. This can be done using 'brew cask install' commands in terminal. The selenium server must be running using 'java -jar /path/to/selenium-server' prior to running the python notebook.

11.1.2 PHU Wrangling

Order of Github Scripts to Run:

1. Retrieve updated COVID-19 Data for Ontario (covid-wrangling.ipynb)
2. After performing DB aggregation and labeling using QGIS (methods described below), convert health region names to current public health units (HR_to_PHU_wrangling.ipynb)

3. Prepare comorbidity data (comorbidity_data_prep.ipynb)

4. Merge proximity data, amenity score, health indicators, DB aggregations, and COVID-19 data per PHU (phu-wrangling.ipynb)

Detailed Wrangling:

covid-wrangling.ipynb

Daily updated information on COVID-19 cases was retrieved from the government of Ontario. The data was grouped according to the reporting PHU and each case attribute was counted to produce a numerical summary for each PHU (covid_wrangled.csv).

QGIS Methods

QGIS is an open source GIS application which was used to analyze, merge, and extract information from the data which were available to us in shapefile (.shp) format. QGIS version 3.10.5-A Coruña for macOS Mojave (10.14) was used in this project.

QGIS was used to layer the various shapefiles, namely the dissemination block layer (DB) underneath the public health unit layer (PHU). The main objective of doing this was to ensure that each DB belonged to exactly one single PHU, and then to classify each DB into that corresponding PHU. Fortunately, each DB was associated with only one PHU, and no DB had boundaries which extended past the boundary of the PHU. We also visualized proximity measures, LTC home locations as layers on the QGIS map. QGIS also flags problematic polygons which we encountered in the data, and using a built-in function we were able to "correct" the polygons with issues. In addition, due to discrepancies of PHU definition (Ontario had 36 PHU's at some point, but currently there are 34), some PHU regions were combined, so QGIS was used to combine the corresponding polygons (QGIS dissolve function). The main use of QGIS was to aggregate the proximity measures, so that we would be able to obtain a proximity measure from the individual DB's extrapolated to the PHU in which the DB belonged. This was necessary because all our other data which we were integrating with COVID-19 data was recorded at the PHU-level, except for the proximity measures which were recorded at the DB level. In order to avoid large outlier influence, we decided to aggregate the proximity measures by using the medina value for each proximity measure to represent the corresponding proximity measure for the PHU. For example, if the Toronto PHU had 12,000 DB's all with varying proximity scores for "proximity to secondary schools" we decided it would be more effective to use the median value for the measurement from all the DB's to become the proximity measure for at the PHU level. This is explained further and demonstrated visually in Appendix A. In addition, using the layer mapping tool allowed us to visualize our data in an effective way, in addition to provide a cross-reference to the D3 web page during development.

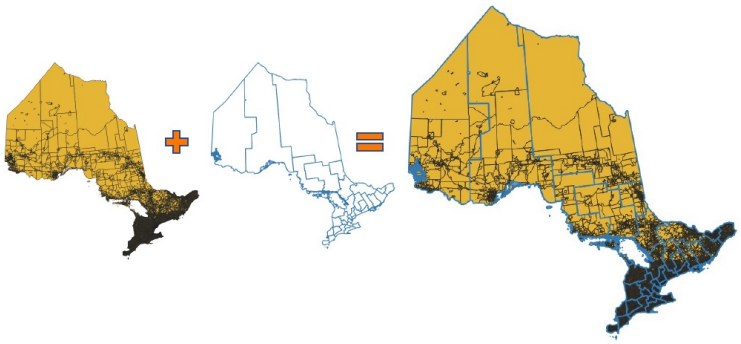

Figure 13: **Schematic showing GIS layering of two map layers to produce the map of DB's overlaid by the PHU boundaries.**
Black lines depict the boundary of DB's in Ontario, blue lines depict the boundaries of PHU's of Ontario. There are 133,214 DB's and 34 PHU's in Ontario.

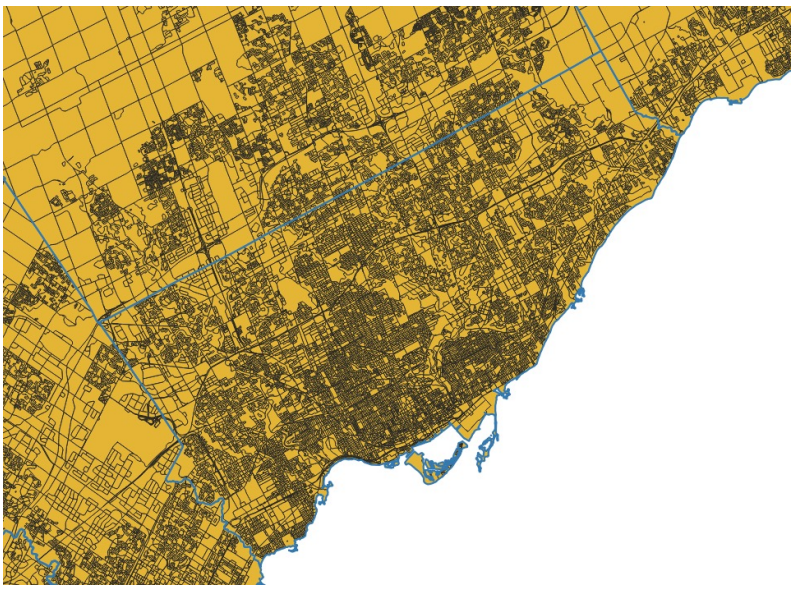

Figure 14: **Map showing how dense DB's are in some regions of Ontario.**
This region shows Toronto PHU and associated DB's. This PHU alone has over 12,000 DB's.

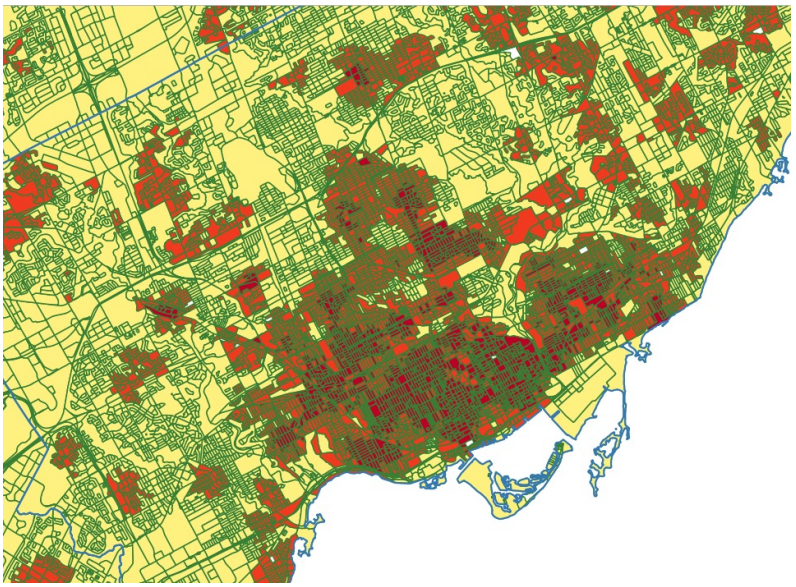

Figure 15: **Map showing the amenity scores for the Toronto PHU.**
Lighter colours depict DB's with low amenity scores, and darker colours depict DB's with high amenity scores. White areas represent DB with no data collected.

comorbidity_data_prep.ipynb
Data was filtered on Ontario and the HR_UID was adjusted to only include the last several digits. Data was subset on the following conditions: Age group: Total, 12 years and older, Sex: Both sexes, REF_DATE: 2017/2018. Data was pivoted (comorbidity_data_percent.csv).

HR_to_PHU_wrangling.ipynb
Since Ontario COVID-19 case data was not accompanied by a unique identifier, additional data must be merged on the PHU name. Previously, Ontario was divided into 36 Health Regions but convention has since evolved to merge 2 pairs of regions, ultimately rendering 34 Public Health Units, and including minor name changes. Amenity scores and Health Region-labelled dissemination blocks rendered by QGIS required labels to be updated to the current PHU nomenclature. The four merged units appeared as duplicate labels and their content was later averaged where sensible, such as for the amenity score. This issue was averted in the QGIS output as regions were merged prior to labelling.

phu-wrangling.ipynb
Proximity data was joined to the QGIS output on the unique identifier, HR_UID, primarily for the addition of the dissemination block population, which was later summed when data was grouped by PHU to provide a population. Amenity density and wrangled COVID-19 data were then merged on the PHU name and proportion of cases for each PHU were calculated by dividing the COVID-19 case total by the PHU population. Finally, wrangled proportion co-morbidity data was joined on the PHU name to render a complete PHU file (PHU_master_prop.csv).

11.2 Web Page Deployment

A web page that includes interactive visualizations and user instructions can be found on our Github repository.

To update the data, several scripts need to be run:

1. PHU_analysis/wrangling/webpage_data.ipynb

2. LTChomes_analysis/webpagedata/webpage_data.ipynb

Both scripts require the CSV files generated from the wrangling and merging steps for PHU and LTC analysis, respectively. This will automatically update the following CSV files and, subsequently, metadata on the web page:

1. docs/kt/data/ltc_points.csv
2. docs/kt/data/phu_statistics.csv

11.3 Appendix B - Results

PHU Analysis - Supervised Methods

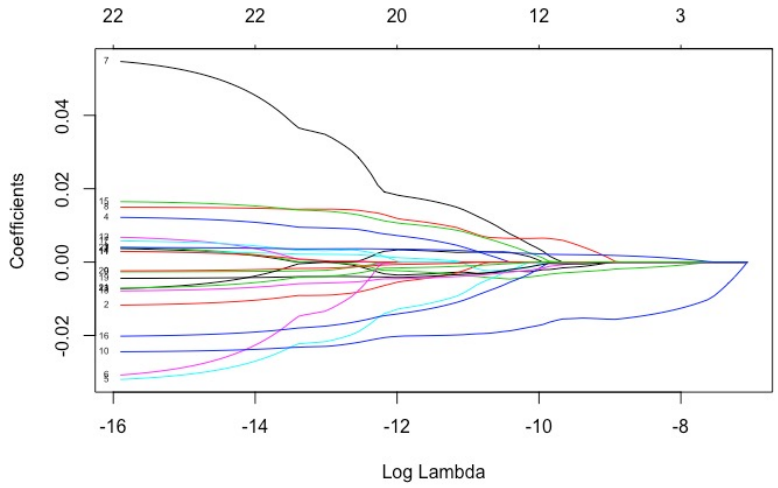

Figure 16: **LASSO output of PHU predictor variables for COVID-19 cases proportion modelling.**
Top axis shows the number of predictor variables included in the LASSO, bottom x-axis shows the Log lambda, the tuning parameter. When lambda is equal to zero, the LASSO is performing an ordinary least squares fit and includes all predictor variables. As log lambda decreases, so does the number of influential predictors retained by the LASSO, and extraneous variables' coefficients are forced to zero.

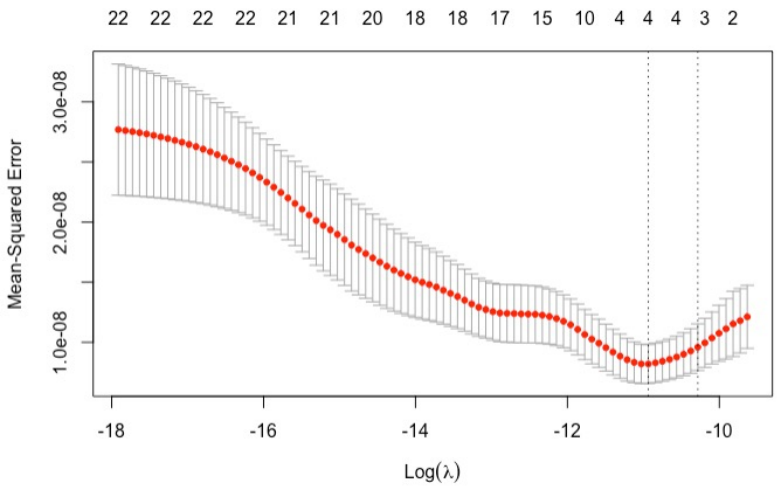

Figure 17: **Cross-validation plot output for LASSO for COVID-19 cases proportion analysis, determination of appropriate lambda value.**
The number of predictors is shown on the top axis, and log lambda (tuning parameter) on the bottom x-axis. The dashed vertical lines represent the lambda value which results in the lowest MSE, along with a secondary vertical line which represents the lambda value which is one standard error away from the minimum.

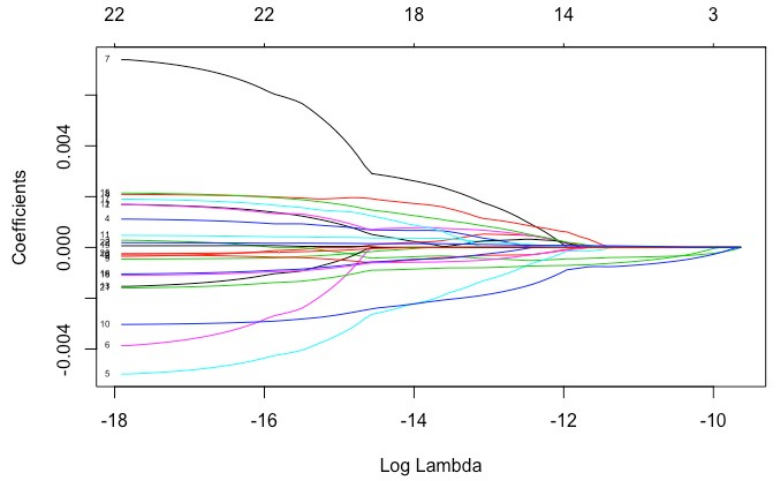

Figure 18: **LASSO output of PHU predictor variables for COVID-19 fatalities proportion modelling.**
Top axis shows the number of predictor variables included in the LASSO, bottom x-axis shows the Log lambda, the tuning parameter. When lambda is equal to zero, the LASSO is performing an ordinary least squares fit and includes all predictor variables. As log lambda decreases, so does the number of influential predictors retained by the LASSO, and extraneous variables' coefficients are forced to zero.

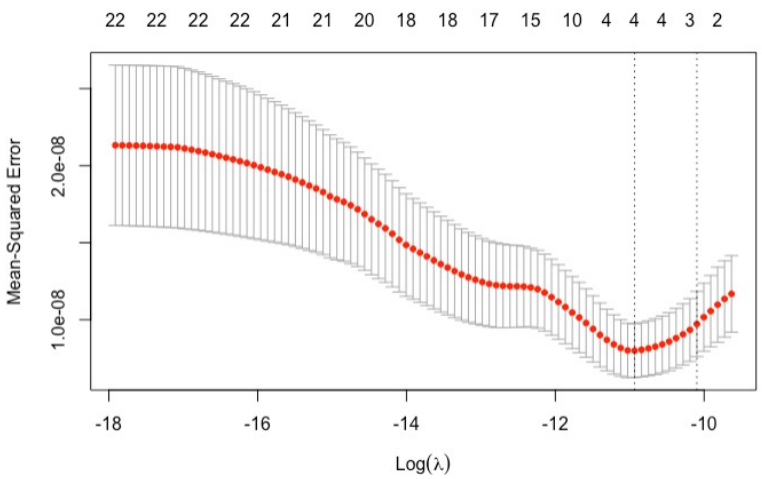

Figure 19: **Cross-validation plot output for LASSO for COVID-19 fatalities proportion analysis, determination of appropriate lambda value.**
The number of predictors is shown on the top axis, and log lambda (tuning parameter) on the bottom x-axis. The dashed vertical lines represent the lambda value that results in the lowest MSE, along with a secondary vertical line that represents the lambda value, which is one standard error away from the minimum.

Table 11: **Loadings for the variables on PC1, PC2 and PC3**

Variable	PC1	PC2	PC3
prox_idx_emp	0.248	-0.228	0.104
prox_idx_pharma	0.220	-0.164	-0.005
prox_idx_childcare	0.226	-0.104	-0.242
prox_idx_health	0.209	-0.262	0.0169
prox_idx_grocery	0.118	-0.251	-0.145
prox_idx_educpri	0.177	-0.327	-0.063
prox_idx_parks	0.259	-0.071	-0.023
prox_idx_transit	0.195	-0.292	-0.119
prox_idx_educsec	-0.074	-0.232	-0.0477
prox_idx_lib	-0.125	-0.085	-0.132
Arthritis..15.years.and.over.	-0.256	-0.014	-0.094
Body.mass.index..adjusted.self.reported..adult.. 18.years.and.over...obese	-0.260	0.009	-0.055
Chronic.obstructive.pulmonary.disease..COPD.. 35.years.and.over.	-0.236	-0.016	-0.056
Current.smoker..daily.or.occasional	-0.245	-0.089	-0.036
Diabetes	-0.159	-0.182	0.252
Heavy.drinking	-0.221	0.002	-0.267
High.blood.pressure	-0.229	-0.027	0.0788
Mood.disorder	-0.204	-0.130	-0.242
Perceived.health..fair.or.poor	-0.235	-0.221	0.0806
Perceived.mental.health..fair.or.poor	-0.089	-0.177	-0.158
Asthma	-0.161	-0.165	-0.247
Perceived.life.stress..most.days.quite.a.bit. or.extremely.stressful	0.093	-0.110	0.061
Body.mass.index..adjusted.self.reported..adult.. 18.years.and.over...overweight	0.095	0.095	-0.119
Influenza.immunization.in.the.past.12.months	0.005	-0.002	-0.475
Life.satisfaction..satisfied.or.very.satisfied	0.165	0.286	-0.147
Has.a.regular.healthcare.provider	0.048	0.228	0.076
Perceived.health..very.good.or.excellent	0.168	0.244	-0.186
Perceived.mental.health..very.good.or.excellent	0.134	0.191	-0.010
Physical.activity..150.minutes.per.week..adult.. 18.years.and.over.	0.097	0.153	-0.451
Physical.activity..average.60.minutes.per.day..youth.. 12.to.17.years.old.	-0.0752	-0.113	-0.219
Sense.of.belonging.to.local.community..somewhat. strong.or.very.strong	-0.091	0.258	-0.076