Covid-19 vaccination in the province of Ontario: A geographical and socio-economical analysis

Data Cleaning

Ariel Mundo

Table of contents

Background

The original Fields Covid-19 survey contained information about Covid-19 vaccination status and other COVID measures in different cities in Ontario. Information provided by respondents included:

- Age (only above 16, if age above 75 then it appears as 98)
- Age-group (generated from age)
- Employment status
- Remote work within the last month
- If person receives paid sick leave
- Number of people in household
- Number of people from household that attend school
- Chronic illnesses within the household same time

- Race
- Three first digits of postal code
- Day, month and year the survey was accessed

Respondents provided multiple answers regarding vaccination:

- "Have you received the first dose of the COVID vaccine?" (y/n)

- (If answered "yes" above) "Have you received the second dose of the COVID vaccine?" (y/n)

- (If answer was "no" to the first question) "If a vaccine was made available to you you would:"

```
- definitely get
- definitely not get
- probably get
- probably not get
```

Preliminary Analyses

This document focuses on exploratory analyses, covariate selection, and outlier identification and removal. The goal of the exploratory analyses was to clean the dataset for formal analyses.

As way of comparison, I used the 2016 census data for Ontario.

Data Loading

The first task was to load the raw data, extract the majority of the responses from the survey that could be used for regression, and if they were categorical, make them factors to do the exploratory analysis.

Choosing covariates

The next step consisted in identify the missing rates of the covariates, and determinewhich of those could be included in the model. The following code chunk creates a table with the number of missing observations and percentages.

Table 1: Percentage of missing observations all covariates

variable	observations	missigness

age_group	39029	0.0%
income	8919	77.1%
race	6873	82.4%
employed	5247	86.6%
h_size	4129	89.4%
school	4050	89.6%
pc_1	3442	91.2%
pc_2	3319	91.5%
pc_3	3238	91.7%
remote_work	2490	93.6%
sick_leave	2441	93.7%

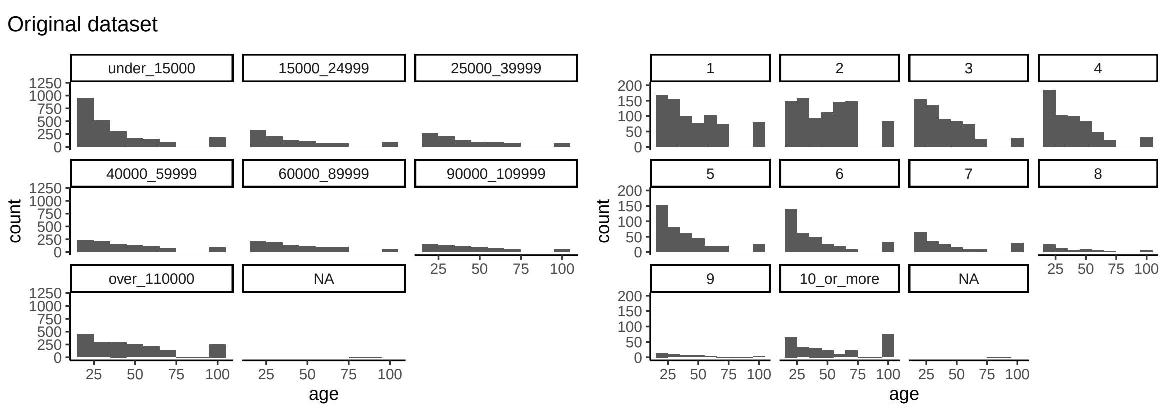
From the table, it can be seen that the covariate with the least amount of observations is "sick_leave".

However, per the dictionary in the original dataset, "sick_leave" was answered only by those that reported to be employed (the survey design made this response conditional). Therefore, those unemployed would be excluded in an analysis that considers this variable.

Therefore, we decided to select the following covariates from the original dataset:

- age group
- income
- race
- employment status

From these covariates, the one with the highest missing rate of observations is the employment status. Next, we cleaned the data in order to have complete observations about employment, and to see how the missing rates looked for the other covariates. The following code chunk creates histograms for the different covariates and a table with missing rates in the the `clean_data` data frame which was created from the raw data.



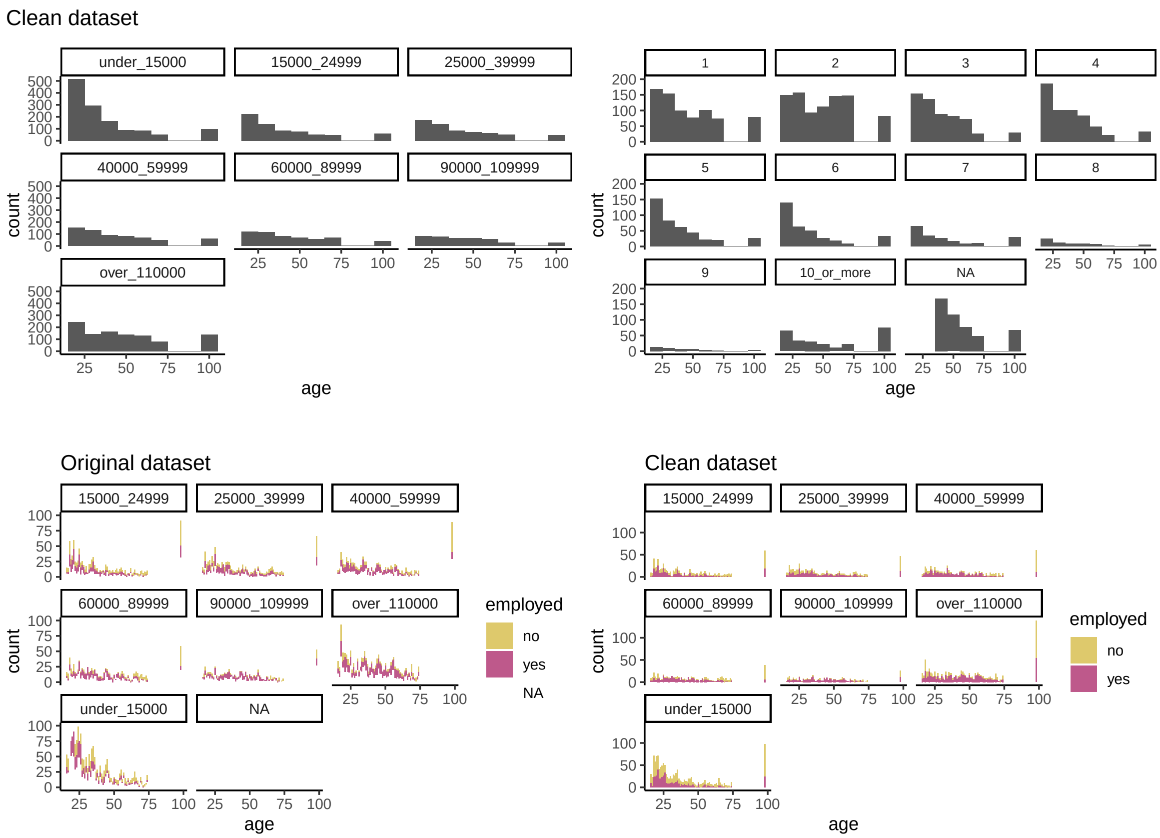


Table 2: Missing observations and histograms of the data

variable	missigness
get_vaccine	76.5%
race	27.9%
second_dose	26.4%
h_size	21.3%
age_group	0.0%
employed	0.0%
age	0.0%
income	0.0%
city	0.0%
date	0.0%
first_dose	0.0%

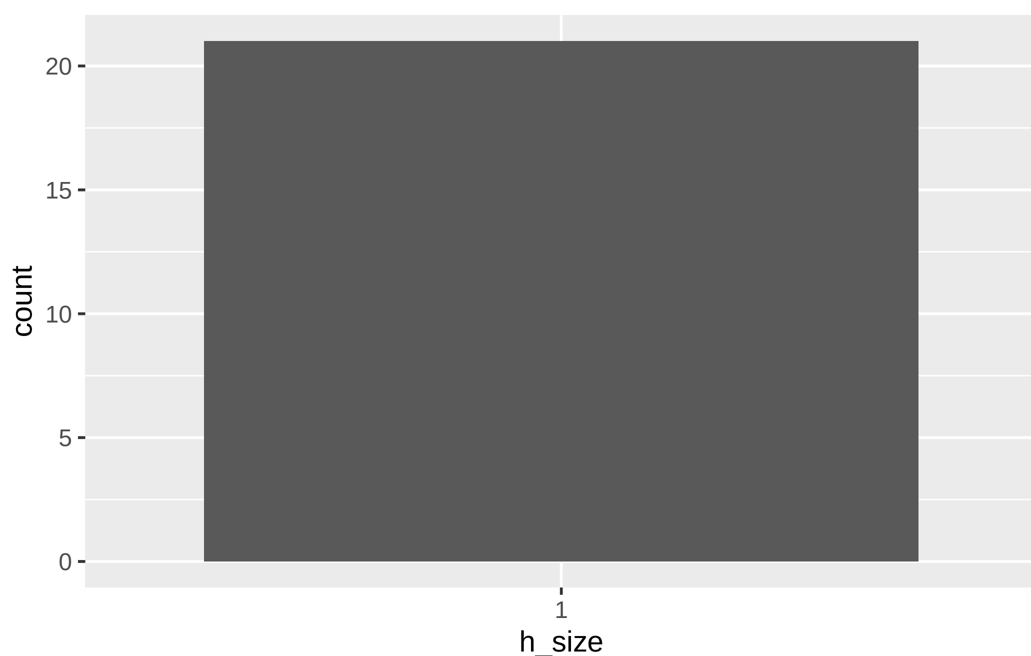
At this stage in the analysis, we identified that outliers were present in the data. Specifically, individuals that were under 25 years of age and that reported having an income >110k while living in a household of 1. We therefore, created some diagnostic plots to explore the reported household composition of those <25 with income>110k, and we removed the outliers from the dataset.

```
outliers<-clean_data %>%
  filter(age_group=="16_24" &
           h_size==1 &
           income=="over_110000"

  )

outliers%>%
  ggplot(aes(x=h_size))+
  geom_histogram(stat="count")
```

```
Warning: Ignoring unknown parameters: binwidth, bins, pad
```




```
#From the plot it is about 21 entries that are outliers. Now, remove these outliers

clean_data<-anti_join(clean_data,outliers)
```

```
Joining, by = c("age_group", "employed", "age", "income", "race", "city",
"h_size", "date", "first_dose", "second_dose", "get_vaccine")
```

The next code chunk creates another table with covariate missing rates in the clean dataset. It can be seen that at this stage the answer about the first dose of the vaccine has no missing observations, the answer about having received the second dose of the vaccine has 27% of missing observations, and the answer about if people would get a vaccine has 76% of missing observations.

Table 3: Clean dataset missigness

variable	missigness
get_vaccine	76.2%
second_dose	27.0%
age_group	0.0%
employed	0.0%
age	0.0%
income	0.0%
race	0.0%
city	0.0%
h_size	0.0%
date	0.0%
first_dose	0.0%

Race and Ethnicity

Next, we explored how the race and ethnicity information from the dataset compared to the data from the Census. The following chunk creates a summary table for the Race variable.

Table 4: Ethnic information from the clean dataset

race	observations	percentage
arab_middle_eastern	221	4.2%
black	307	5.9%
east_asian_pacific_islander	311	6.0%
indigenous	224	4.3%
latin_american	183	3.5%
mixed	328	6.3%
other	396	7.6%
south_asian	384	7.3%
white_caucasian	1410	27.0%

NA	1462	28.0%

It is important to mention that the categories for race/ethnicity provided in the survey did not match the categories used in the Census. Therefore, we used a combination of sources to obtain the information presented in Table 5, where the categories used could be matched to the Census Information. The data sources for the table were:

- Fact sheet from the Provice of Ontario for Visible Minorities link: Used to obtain percentages for Arab, Black, East Asian/Pacific Islander (adding Chinese, Korean, and Japanese percentages), Latin American, Mixed (using the percentage for "multiple visible minority"), Other (obtained by adding the Southeast Asian, Filipino, West Asian, and Minority not identified elsewhere percentages), South Asian. *Accessed on January 05, 2022*
- Census Profile for Ontario link: Used to obtain percentage of Aboriginal population. *Accessed on January 05, 2022*
- Wikipedia entry for Ontario demographics link: To corroborate that the percentage of population reported as "European" in this website matched the percentage obtained for "White Caucasian" that was independently obtained by obtaining the difference in population proportion after subtracting the sum of Visible Minorities and Aboriginal Population percentages.

Table 5: Reference Data from Race/Ethnicity

Ethnicity/Race	percentage
Arab	1.6%
Black	4.7%
East Asian/Pacific Islander	6.6%
Indigenous	2.8%
Latin American	1.5%
Mixed	1.0%
Other	5.2%
South Asian	8.7%
White Caucasian	67.8%
Total	99.9%

Visible Minorities: 29.3% Not a Visible Minority: 70.7% (Aboriginal 2.8%, White Caucasian 67.8%)

Age groups

According to the Census, the distribution of the different the age groups for the province of Ontario are as follows:

Table 6: Age distributions from the 2016 Census for Ontario

Group age	Percentage of population
15-24	12.7%
25-34	12.9%
35-44	12.8%
45-54	14.9%
55-64	13.7%
65 and over	16.7

To compare the survey data to the Census, the next chunk creates a barplot of the age groups in the survey. Here, it can be noticed that the age-group distribution from the survey data is different from the Census data.

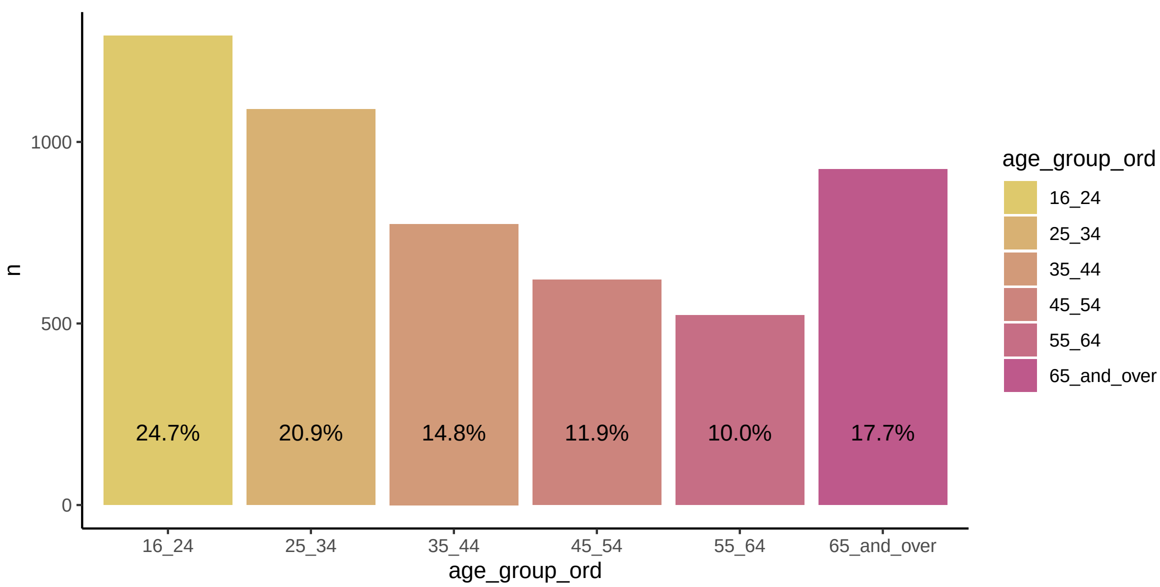


Figure 1: Age group distributions from the dataset

Income

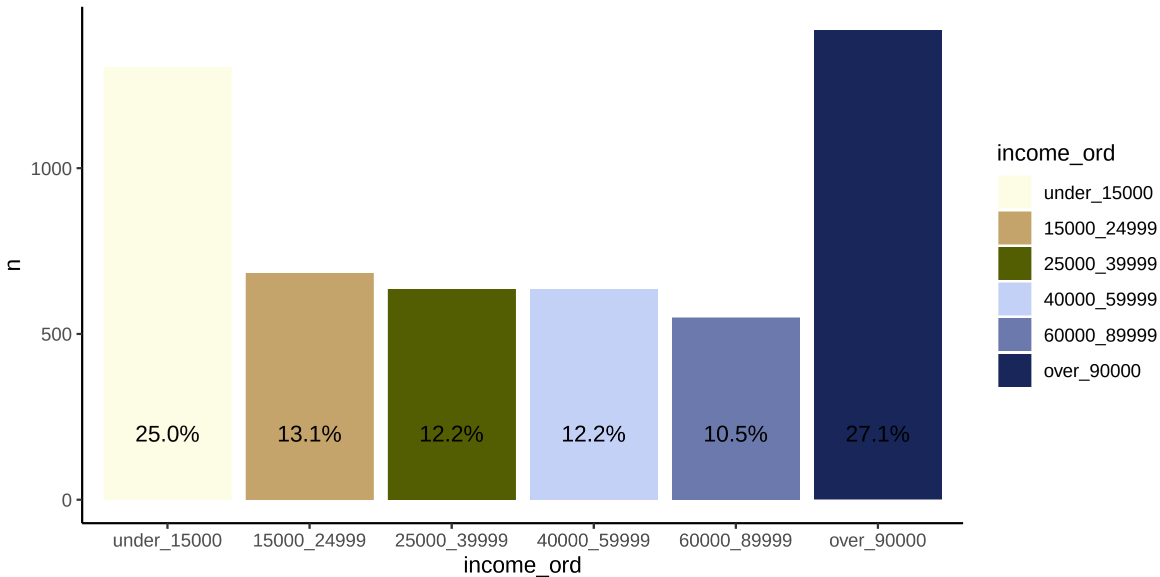
Survey respondents answered the question “What is your household annual income?”. To compare the distribution of responses to this question from the survey and the Census data, we used the “Household total income groups in 2015 for private households” from the Census data (available in the Census Data for Ontario website link).

Table 7: Income percentages from the 2016 Census for Ontario

Household income range (CAD)	Percentage
< 15,000	5.7%
15,000 - 24,999	7.5%
25,000 - 39,999	11.6%
40,000- 59,999	15.4%
60,000 - 89,999	19.5%
>90,000	40.3%

One difference that is noticeable is that the brackets for income in the census data are different than the brackets used in the survey. The census does CAD 4,999 brackets (e.g., CAD 5,000- CAD 9,9999) up to CAD 49,999, followed by CAD 9,999 brackets up to CAD 99,999. After that, the brackets increase to CAD 24,999, and therefore, it is not possible to obtain percentages for the 90,000-109,999 and >110,000 brackets from the survey.

Therefore, in the following code chunk, we created an additional category for income in the dataset that matched the information from the Census, and barplots to visualize the proportion of each income bracket in the dataset.



We identified differences again between the survey data and the Census data. For example, the <15,000 bracket accounts for about 25% of the responses, a much higher rate than the proportion in the Census. The other brackets were different as well from the Census data.

Geographical Information

Because each of the respondents of the survey was assigned a geographical location (city), we were interested in accounting for geographical location in our analysis. We used a multi-step process to assign geographical information to the entries in the dataset.

There are two parts to the geographical analysis:

1) Assign to each entry in the dataset municipalities the geographical region it belongs to using municipality and geographical region information.
2) Assign a Health Region to each survey entry using the geographical and Local Integrated Health Network (LHIN) information.

The details of each step are outlined below.

Municipality Data

We have obtained and cleaned the data of the municipalities from the province of Ontario website, and cleaned the dataset to obtain the city names and geographical locations. Further details can be found in `municipalities.qmd` in the `data_cleaning` directory. We will join and match the geographical location from the municipalities dataset to the clean dataset we have obtained in this document so far after removing the entries that do not have a corresponding geographical location.

```
clean_data<-clean_data%>%
  filter(city!="None")


municipalities<-read.csv(here("data","municipalities_clean.csv"))

municipalities$Municipal.status<-as.factor(municipalities$Municipal.status)

municipalities$Geographic.area<-as.factor(municipalities$Geographic.area)

municipalities$city<-as.factor(municipalities$city)


clean_data<-left_join(clean_data,municipalities,by="city")
```

The following chunk identified which entries were left without a geographical region:

```
test<-clean_data %>%
  filter(is.na(Geographic.area))%>%
  distinct(city)
```

This analysis identified that 2744 entries did not get a geographical region. These 2744 entries corresponded to 187 unique cities. These cities without a region were exported to a csv file in order to manually write their geographical areas following the Association of Municipalities of Ontario divisions, using Wikipedia to check the status of each municipality. The following chunk created the csv file. The code is commented now as it was run once.

After searching and manually entering the geographic area for each city, the file was saved as `missing_municipalities_updated.csv`, and this file was used for the next steps.

Note: there is one city "Kinburn", but there are two communities with such name, one in Huron County and one in Carelton County, we assigned it to Huron County. In the case of "Sydenham", which can be a ward in Kingston or a community in Frotenac, we assigned it to Frotenac.

Merging missing municipalities and geographic area names

After manually assigning the geographical regions to the municipalities that were missing (which can be found in `missing_municipalities.csv` in the `data` directory), these were merged as the dataset `geographic_areas.csv` (also in `data`), which contains the titles for each region (e.g., "County", "Region").

```
#load missing municipalities dataset

missing_municipalities<-read.csv(here("data","missing_municipalities_updated.csv"))


#combining geographical regions

clean_data<-clean_data %>%
  left_join(missing_municipalities, by = c("city")) %>%
  mutate(Geographic.area = coalesce(Geographic.area.x,Geographic.area.y)) %>%
  select(-c(Geographic.area.x,Geographic.area.y))

clean_data$Geographic.area<-as.factor(clean_data$Geographic.area)

# load the the titles for each region

geographic_areas<-read.csv(here("data","geographic_areas.csv"))

geographic_areas$Geographic_area<-as.factor(geographic_areas$Geographic_area)

geographic_areas$Geographic_area_title <-as.factor(geographic_areas$Geographic_area_title)

#merge the datasets

clean_data <-left_join(clean_data,geographic_areas,by=c("Geographic.area"="Geographic_area
   subset(select=-c(Full_title.y))
```

Health Regions

We sought to geographically analyze the information in the survey using the The Health Regions of Ontario. However, these Health Regions do not match the divisions from the census, and there is no publicly available dataset from Health Ontario that lists each municipality and its corresponding Health Region. We used therefore a multi-stage approach to incorporate the information into the dataset:

- First, we used the dataset from Paul Allen regarding long-term care homes in Ontario (https://paulallen.ca/consolidated-dataset-of-ltc-homes-in-ontario/) to obtain information about communities and the Local Health Integration Network (LHINs) where long-term care homes were located.

- Second, using the LHIN information, we added the Health Region each entry corresponded to using the information on LHIN and Health Region correspondence, which can be found here: https://www.ontariohealth.ca/about-us/our-people.
- Third, after merging the dataset, we manually added LHINs to those municipalities that did not have an entry at this stage.

The dataset from Paul Allen was downloaded and saved as `Consolidated_LTC_dataset.csv` in the `data` directory. One thing to note is that there was a missing observation (coded as "Not provided") for one of the entries (city of Napanee, located in the Lennox and Addington County). Under the LHIN divisions, Napanee was in the South East LHIN. Also, there is one entry that says "244 Main Street East" as the community entry but it should be "Stayner" (the address provided belongs to Stayner). The information was fixed before adding the Health Region.

```
#load data
ltc_data <- read.csv(here("data","Consolidated_LTC_dataset.csv"))

## checking missing observations

sum(ltc_data$X_LHIN=="Not provided")
```

```
[1] 1
```

```
ltc_data <- ltc_data %>%
  mutate(X_LHIN=replace(X_LHIN,X_LHIN=="Not provided","South East"))

ltc_data <- ltc_data %>%
  mutate(COMMUNITY=replace(COMMUNITY,COMMUNITY==" 244 Main Street East","Stayner"))

#keep relevant columns and add Health region information
ltc_data <-ltc_data %>%
 select(c("COMMUNITY","X_LHIN"))%>%
  rename(city=COMMUNITY,
         LHIN=X_LHIN)%>%
  mutate(Health_Region=
    case_when(LHIN=="Central"~ "Central",
              LHIN=="Central West"~"Central",
              LHIN=="Mississauga Halton"~"Central",
              LHIN=="North Simcoe Muskoka"~"Central",
              LHIN=="Central East"~"East",
              LHIN=="South East"~"East",
```

```
                LHIN=="Champlain"~"East",
                LHIN=="North East"~"North East",
                LHIN=="North West"~"North West",
                LHIN=="Toronto Central"~"Toronto",
                LHIN=="South West"~"West",
                LHIN=="Hamilton Niagara Haldimand Brant (Hnhb)"~"West",
                LHIN=="Waterloo Wellington"~"West",
                LHIN=="Erie St. Clair"~"West"
    )
  )%>%
  distinct(city,.keep_all = TRUE)
```

Next, we combined the datasets and wrote csv file with those cities that were not assigned a Health Region. The next code chunk does these steps (note that the line for writing the csv file has ben commented as it was run only once).

```
clean_data<-left_join(clean_data,ltc_data,by="city")


 missing_health_regions<-clean_data %>% filter(is.na(LHIN)) %>%
   distinct(city,.keep_all = TRUE)%>%
   select(city,Geographic.area,Geographic_area_title,LHIN,Health_Region)

#write.csv(missing_health_regions,here("data","missing_health_regions.csv"),row.names = FA
```

To obtain the missing LHINswe, obtained information from the LHIN websites, which listed all the municipalities within each LHIN. The websites were:

- South East link
- North Simcoe Muskoka link
- Champlain link
- Waterloo Wellington link
- North West link
- North East link
- Erie St. Clair link
- South West link
- Hamilton Niagara Haldimand Brant link
- Central West link

- Central East link
- Mississauga Halton link
- Toronto Central link

Some cities did not belong entirely to a LHIN. For example, Etobicoke was divided between the Central, Central East, and Toronto Central LHINs. We chose in these case the LHIN that covered the larger geographical region of each city (Toronto Central LHIN in the case of Etobicoke). We next assigned LHINs to the entries were they were missing using the information from the webistes, and created a csv with the updated information (the file is called `missing_health_regions_updated.csv`). The next code chunk loads this file, assigns LHINs and creates a new column in the dataset for Health Region.

```
#load health regions with missing entries fixed

mhru<-read.csv(here("data","missing_health_regions_updated.csv"))

#rename columns
mhru <- mhru %>%
  rename(Geographic_area_title=Geographic.area.title)


clean_data<-clean_data %>%
  left_join(mhru, by = c("city")) %>%
   mutate(LHIN = coalesce(LHIN.x,LHIN.y)) %>%
  select(-c(LHIN.x,LHIN.y,Full_title.x,Geographic.area.y,Geographic_area_title.y))%>%
  mutate(Health_Region=
    case_when(LHIN=="Central"~ "Central",
              LHIN=="Central West"~"Central",
              LHIN=="Mississauga Halton"~"Central",
              LHIN=="North Simcoe Muskoka"~"Central",
              LHIN=="Central East"~"East",
              LHIN=="South East"~"East",
              LHIN=="Champlain"~"East",
              LHIN=="North East"~"North East",
              LHIN=="North West"~"North West",
              LHIN=="Toronto Central"~"Toronto",
              LHIN=="South West"~"West",
              LHIN=="Hamilton Niagara Haldimand Brant (Hnhb)"~"West",
              LHIN=="Waterloo Wellington"~"West",
              LHIN=="Erie St. Clair"~"West"
    ))%>%
```

```
    rename(Geographic_area=Geographic.area.x,Geographic_area_title=Geographic_area_title.x)
```

At this point, the data was ready for formal analysis.

The clean dataset and completed dataset was saved as a *.csv file called `clean_dataset`.