Fields COVID-19 dataset exploration

Ariel Mundo

Table of contents

Background

This dataset is about vaccination and other COVID measures in different cities in Ontario. The covariates available are:

- 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 (*I will not analyze this one as the question is too open and asks about age and disease at the same time*)
- Race

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

There are many responses from the survey on the dataset, but there are some that I would be interesed in analyzing:

- "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
```

Note: As of Dec 30, 2022 I no longer intend to analyze this last question as after cleaning the dataset it has a missing rate of about 76% and therefore it has too few observations compared to the first two questions.

Preliminary Analyses

This document focuses on some preliminary analyses of the data I undertook to assess its representativeness. I performed five major tasks for this preliminary analysis:

- Choosing covariates to assess their missigness rates
- Cleaning the data in order to keep all the covariates with a 0% missigness rate
- Race and ethnicity representativeness of the clean data
- Trends in the survey answers by geographical location
- Trends in survey answers by group age

Overall the goal of this analysis was to see what adjustments would need to be made to formally analyze the data.

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

Choosing covariates

The next step is to see from the covariates what are the ones with the highest number of missing observations, in order to decide what can we include in the model.

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, there are a couple of things we want to consider: The answer to "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.

At the very least, it would be interesting to analyze how the reported status of vaccination changed over geographical location and time by:

- age group
- income
- race
- employment status

From these covariates, the one with the highest missing rate of observations is the employment status. I next cleaned the data in order to have complete observations about employment, and to see how the missing rates looked for the other covariates.

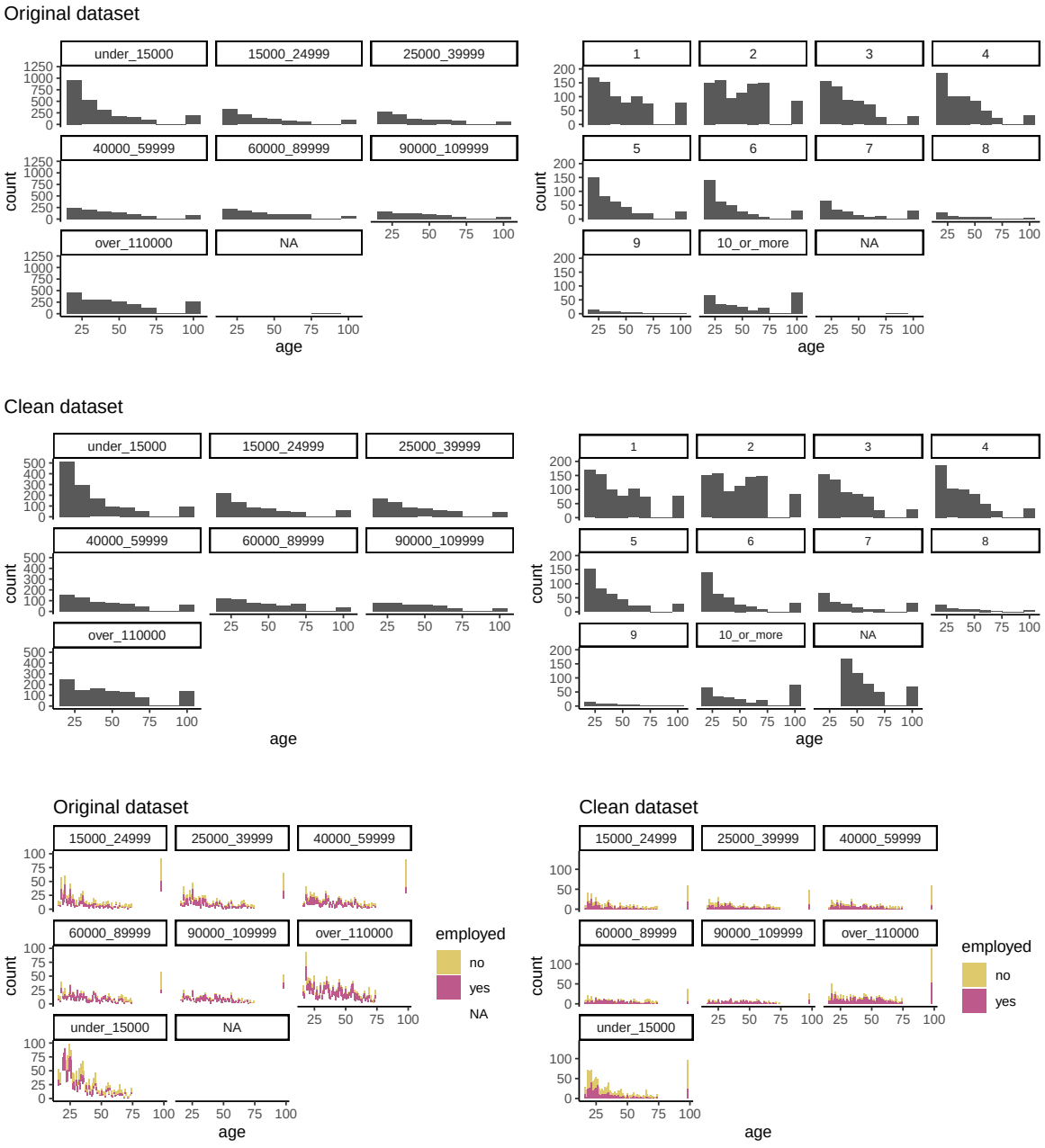


Table 2: Missing observations and histograms of the data

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%

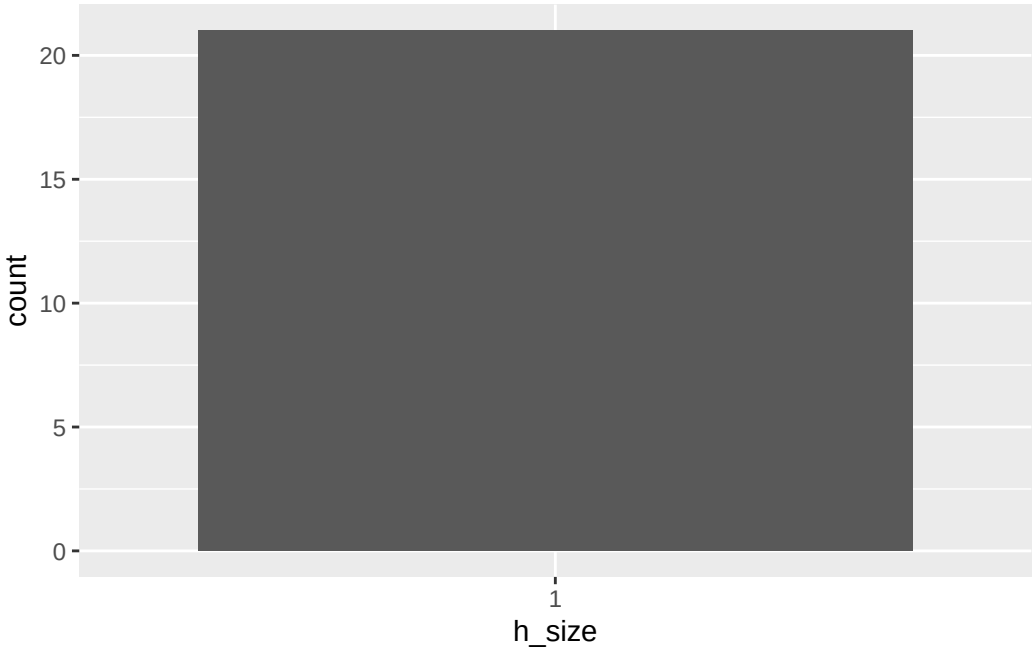
There seem to be outliers where people <25 y of age report having an income >110k while living in a household of 1. Explore the reported household composition of those <25 with income>110k.

```
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")
```

This table shows something interesting due to the nature of the dataset: the missing observations across variables are different, which is why now, although in the original dataset there were more observations of "race" than "employment", when we select complete observations for the latter we lose some observations for race. The second step of cleaning would be to remove the missing observations from race as well, and see what the proportions of covariates and responses are when compared to the census data.

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%

It can be seen that after removing all missing observations in the covariates, the only response that still has missing observations is the answers for the second dose of the vaccine, with 27% of missing observations.

Race and Ethnicity

Once the data was clean, I explored how the race and ethnicity information from the dataset compared to the data from the Census.

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%

Below is the information provided by the 2016 Canada Census for Ontario for different visible minorities and ethnic origins. Link to Census Data.

Table 5: Data from the 2016 Census for Ontario

Ethnicity/Race	percentage
Arab	1.6%
Black	4.7%
East Asian	6.6%

Ethnicity/Race	percentage
Indigenous	3.9%
Latin American	1.5%
Mixed	1.0%
Other	0.7%
South Asian	8.7%

In the table above, the following data were obtained from the "Visible Minorities" from the Census:

- Arab
- Black
- South Asian
- East Asian (computed by adding the Chinese, Korean and Japanese entries)
- Mixed ("multiple visible minorities" entry in the census)
- Other (from the "visible minority n.i.e" entry in the census)

The following data were obtained from the "ethnic origin" data from the census: - Indigenous - European Origins

However, according to the census "ethnicity" and "visible minority" seem to have different values. For example, in the Visible Minority entry of the census Latin American is 1.5%, but in the "Ethnic Origin" data the value is 2.4%.

Moreover, I would need to adjust the data for Pacific Islanders, as the survey included East Asian and Pacific Islanders in the same question but they are categorized differently in the census.

Lastly, the survey had "white caucasian" as one of the options but that category is not present in the census data. The most approximate entry on the census is "European origins" (61.6%).

Overall, we weighted the data because it is not representative of the population distribution in Ontario.

Geographical location

Next, I was interested in analyzing the geographical distribution of the participants over time, to see the percentage of answers from Toronto. This is important to see where the majority of the answers came from (is there such thing as geographical weighting?). The same will be done with the age groups of the participants, as this is data that is reported in the census.

The proportion of responses from Toronto is 30.9% of the total. The Census data indicates that Toronto has around 38% of the total population province (5,433,590/14,223,942). This would mean that a correction would be needed also for the proportion of answers.

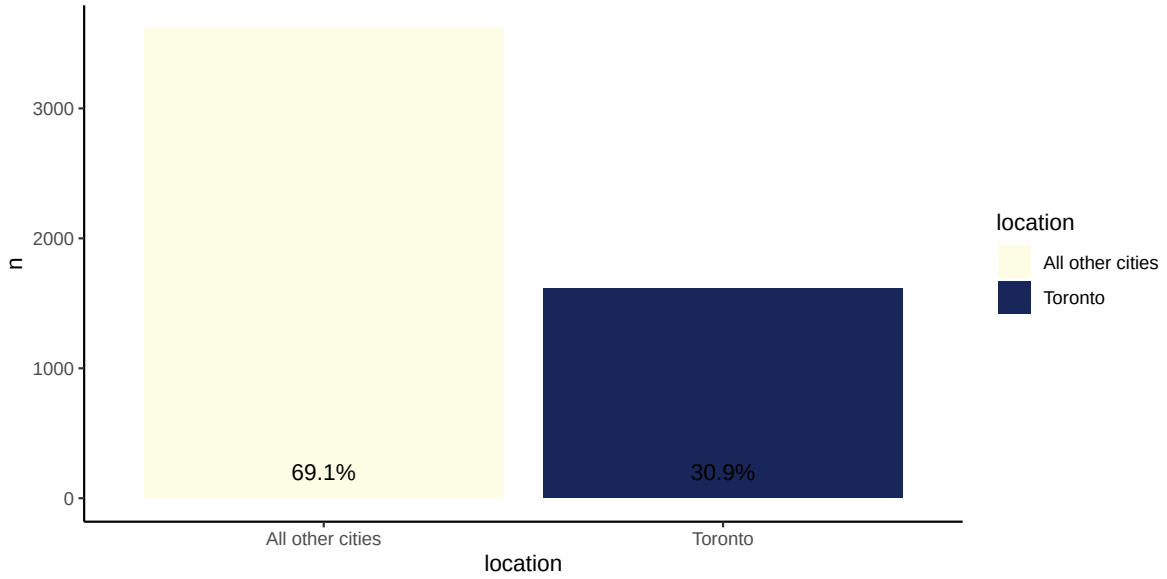


Figure 1: Percentage of survey answers by city

Additionally, the data needs to be cleaned as some of the responses do not have an assigned city, and the entry only reads "None". These entries will be removed as they lack geographical information.

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

Age groups

Another important aspect to consider is how the trends in the responses look over time by age group. This is because from the census data, the age groups for the province 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

However, when plotted the proportion of answers at each time point among groups, the distribution looks rather different.

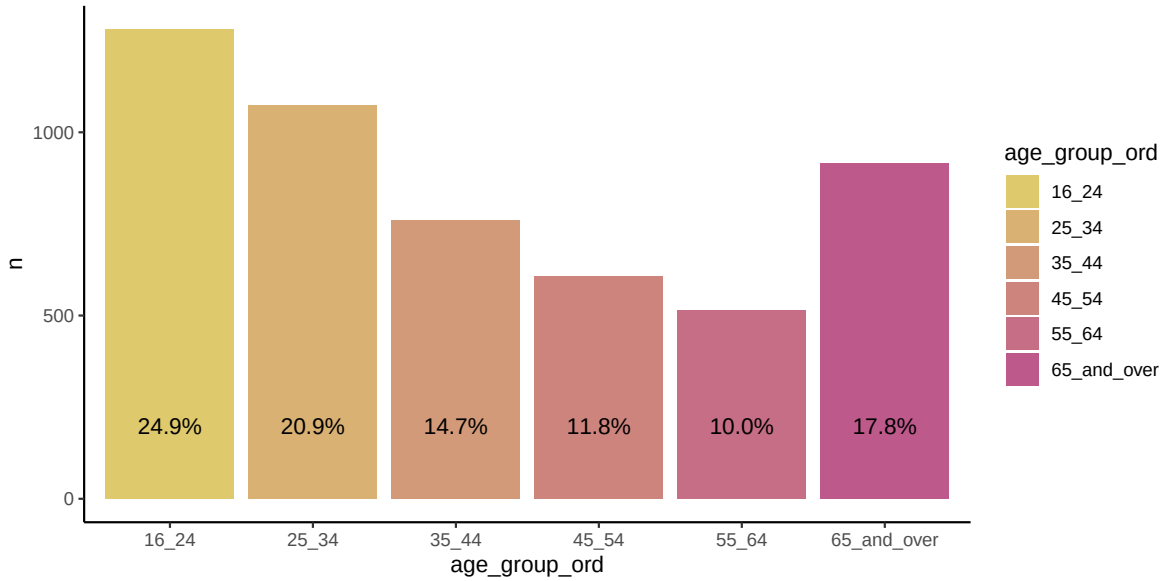


Figure 2: Age group distributions from the dataset

It can be seen from the graph that the proportion of answers by group age do not follow the overall distribution from the province.

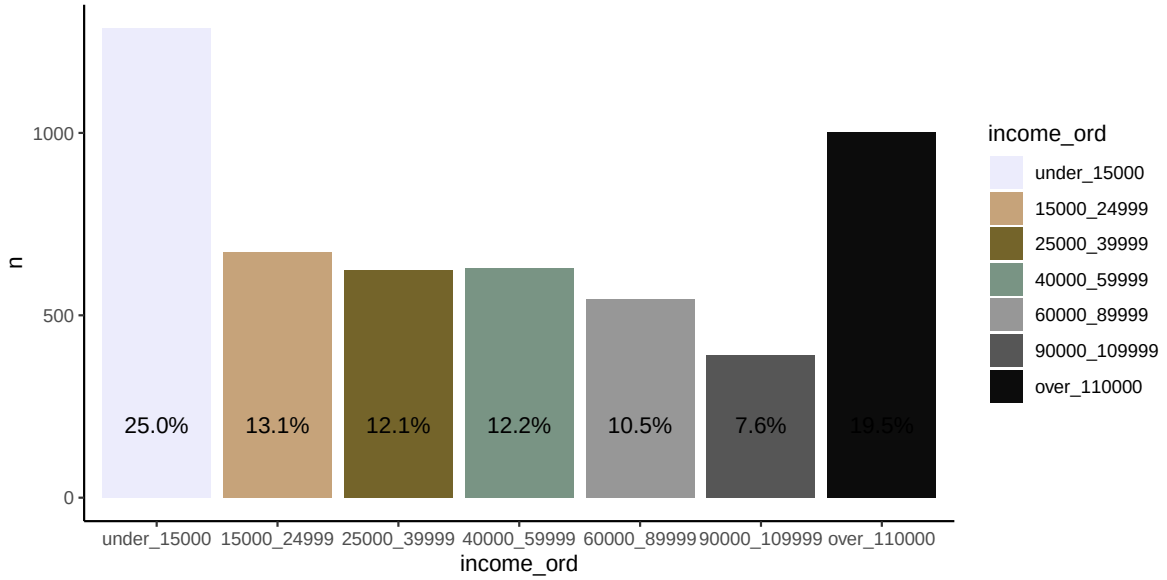
Income

The final aspect I wanted to explore was the income distribution. In this regard the question from the survey was "What is your household annual income?". To compare the results of this answer, I used the "Household total income groups in 2015 for private households" from the Census data. The information from the census appears next.

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	40.4%

One thing to keep in mind in this case 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 and then it does CAD 9,999 brackets up to CAD 99,999 but after that, the brackets increase to CAD 24,999, and therefore, one cannot obtain percentages for the 90,000-109,999 and >110,000 brackets from the survey.



Again, there are differences here in the patterns of response. The <15,000 bracket accounts for about 20% of the responses, a much higher rate than the percentage they represent from the census, and there is variation within the other brackets as well.

Joining the municipalities 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. Next, we will join and match the geographical location from the municipalities dataset to the clean dataset we have obtained in this document so far.

```
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")
```

Need to explore which entries were left without a geographical region:

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

There are 756 observations that did not get a geographical region, and they correspond to 177 unique cities. These will be exported to a csv file in order to manually write their geographical areas following the Association of Municipalities of Ontario divisions, and using Wikipedia to check the status of each municipality.

The code chunk above was ran once to get the names of the municipalities. After manually entering the geographic areas, the file was saved as `missing_municipalities_updated.csv`, and this will be the file to be 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, for the time being Huron County is assigned, will check later if is better to remove it.

Also one occurence of "Sydenham", which can be a ward in Kingston or a community in Frotenac. Will assign Frotenac for the time being.

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 will be 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="Geographic.area")
```

Raking

The next step is to adjust the variables according to the Census data (raking). I will follow the steps provided by https://sdaza.com/blog/2012/raking/