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

### Data Cleaning

Ariel Mundo

## Table of contents



## Purpose

This document focuses on the cleaning of the data from the Survey of COVID-19 related Behaviours and Attitudes which was used as the data source for the analysis in the main paper. The cleaning process encompassed exploratory analyses, covariate selection, removal of outliers, and adding geographical information to the data for analysis (Health Regions in Ontario). At the end of the data cleaning process, a dataset ready for for formal analyses was produced. This document can be rendered to re-create all the steps used in the cleaning process and to generate the same dataset that was used for analysis.

Note, however, that code chunks where intermediate csv files are generated (such as those for missing municipalities, or missing Health Regions) are commented in the current document as they were executed once to create the files. Also, to reduce the lenght of the document, the code in the code chunks does not appear in the rendered file.

## 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
```

### 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 determine which 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 | missingness |
|---|---|---|
| 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.

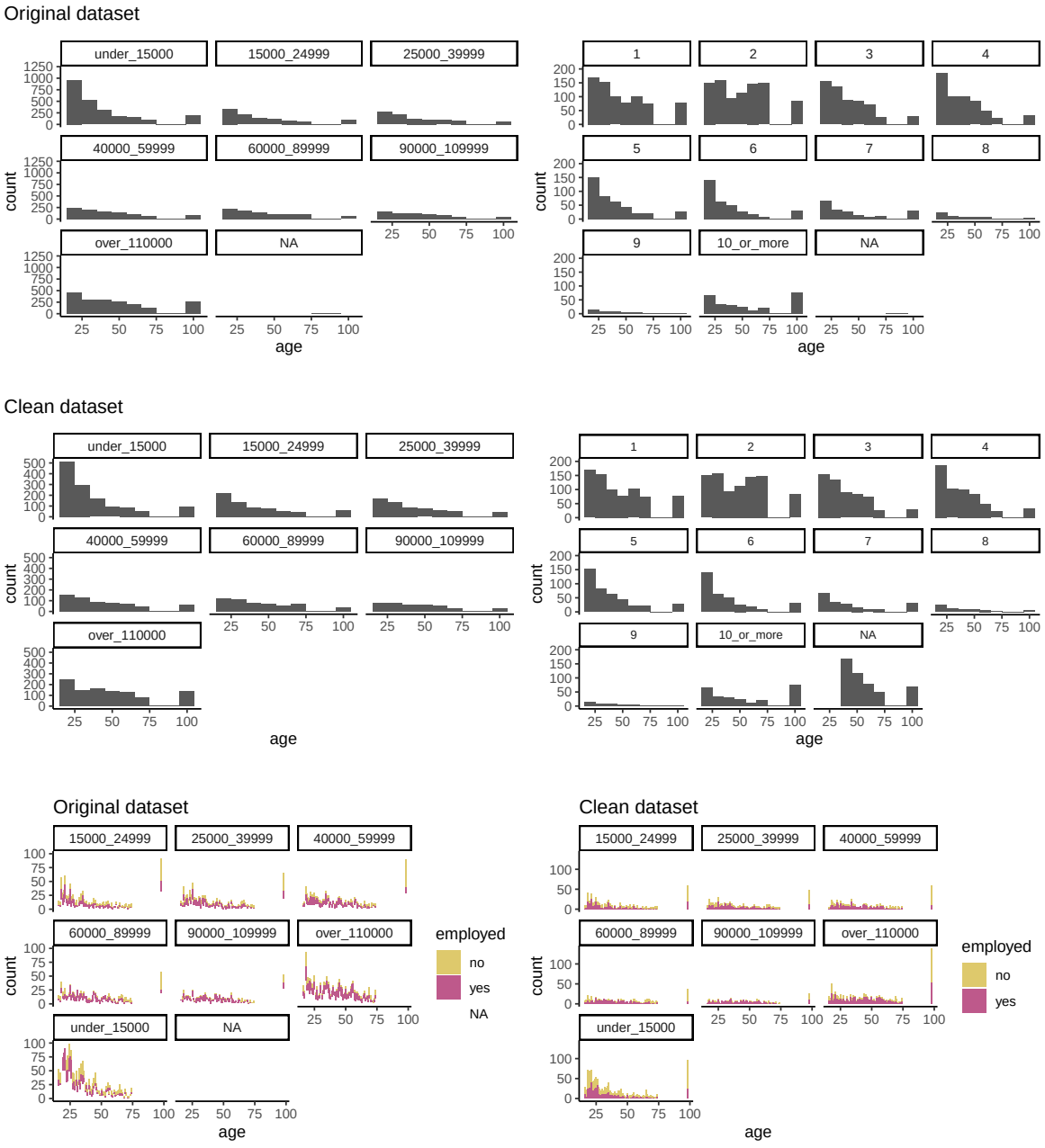
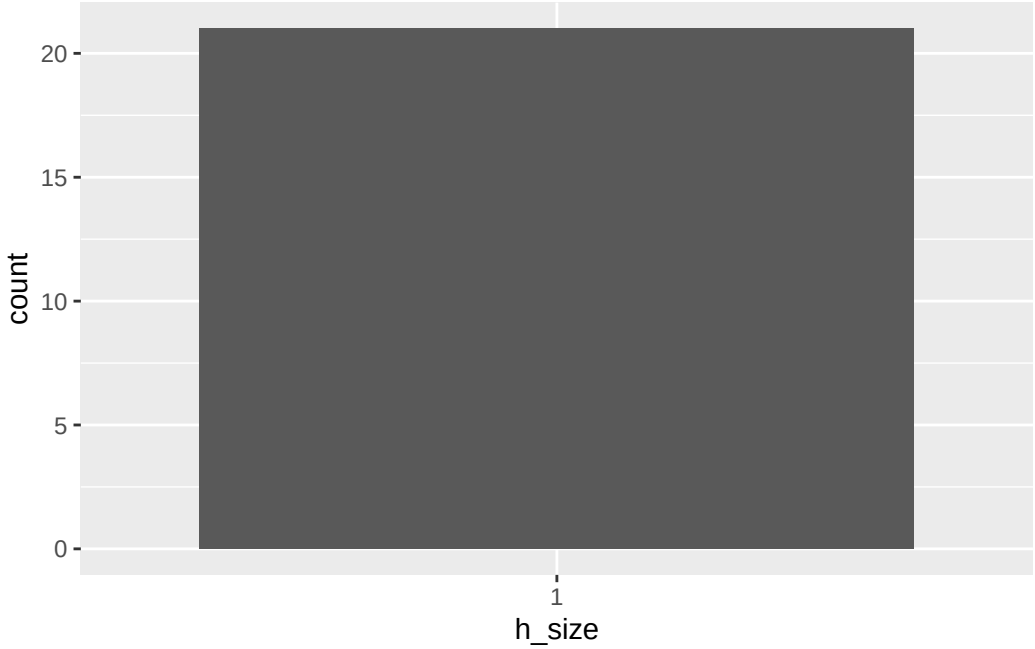



Table 2: Missing observations and histograms of the data

| 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 examined the data for outliers. We identified certain entries in the survey that corresponded to individuals that were under 25 years of age and that reported having an income >110k while living in a household 1. The next code chunk creates a plot of the number of observations that have these characteristics, which were around 20. Next, these outliers were removed from the dataset.




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 missingness

| variable | missingness |
|---|---|
| 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% |

## 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.

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

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").

### 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.

```
[1] 1
```

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).

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.

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`.