**Team Name:**

Delhi

**Group Members & Roles:**

Sara Adi (1129361),

Jennifer Lithgow (1134108), Sophie Mlodzik (1130462), Emily Kozatchiner (1149665)

| **Position** | **Primary** | **Secondary** |
| --- | --- | --- |
| **Product Owner** | Jen/Emily | Sara |
| **Team Leader/Project Manager**  **& Process Coordinator/Agile Lead** | Sara | Sophie |
| **Architect/Conceptual Interaction Lead** | Emily | Jen |
| **Communications Officer/Documentation Lead & Archivist/Revision Control Lead** | Sophie | Jen |
| **Quality Assurance/Testability Lead** | Jen | Emily |

Table of Contents

[Data Set 1 (Jennifer Lithgow)](#_kii8dn25fcee) 1

[Data Set 2 (Sara Adi, Jennifer Lithgow)](#_5bdtt6vti3sj)3

[Data Set 3 (Sara Adi)](#_x9uvo69wim19) 4

[Data Set 4 (Sophie Mlodzik)](#_c1db9d9akf03) 5

[Data Set 5 (Emily Kozatchiner)](#_jekj22lkbqkx) 6

[Questions](#_rma44mmqqg6t) 8

[Question 1 (Sara Adi)](#_q765upmtu21v) 8

[Question 2 (Jennifer Lithgow)](#_ic9aqyb1hhf8) 8

[Question 3 (Sophie Mlodzik)](#_he74asqahzhf) 8

[Question 4 (Emily Kozatchiner)](#_egrgwmx41c9j) 9

Each dataset section will contain the following information:

[Data Description](#_rsuj1nn4r17)

[Format / Organization](#_87imnwqef1b9)

[Encodings](#_arovsrtixxmc)

[Useful Fields](#_416k5h999ppc)

[Data Reorganization](#_3ca5zyf0gd7e)

[Update Frequency](#_s719pc6x8oks)

[Incorporating Updated Data](#_du1ekmmvz7vx)

# Data Set 1 (Jennifer Lithgow)

<https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dt-td/Rp-eng.cfm?TABID=1&LANG=E&A=R&APATH=3&DETAIL=0&DIM=0&FL=A&FREE=0&GC=35&GL=-1&GID=1161871&GK=1&GRP=1&O=D&PID=109526&PRID=0&PTYPE=109445&S=0&SHOWALL=0&SUB=0&Temporal=2016&THEME=115&VID=0&VNAMEE=&VNAMEF=&D1=0&D2=0&D3=0&D4=0&D5=0&D6=0>

Data Description

The census data describes the population living in Ontario divided by age (in single years and in ranges).

Format / Organization

It appears that the first three and last three rows contain some general information about the population. The file was small enough to remove these manually, as they were not needed to answer the relevant question. The remaining rows have four fields: Age or age range, total population, and two for the population divided as male/female. The age ranges include five-year ranges (e.g. 0-4 years), as well as larger, inconsistently divided age ranges. However, ages 100 years and older are not further divided.

Encodings

* **Col. 1**: “0 to 14 years” “0 to 4 years” “1” - Age, either in ranges or individual years
* **Col. 2-4**: Population number (integer), by total or male/female

Useful Fields

* **Col. 1 (index 0)** - Age. There is a mix of integers and ‘phrases’ in the column. The single years/integers are translated into the relevant age-ranges (e.g. 20-29, 50-59)
* **Col. 2 (index 1)** - The total population that are of the given age/age range

Data Reorganization

**Grouping:** Isolating so that only individual ages are listed to use (removing unnecessary rows, ignoring the third and fourth columns)

**Aggregate:** Sum of ages within age ranges that match the second data set (Ontario infections)

**Encodings:** Age\_Group, Population

Update Frequency

None; census may update in 2021, but not within the timeframe of the course/project.

Incorporating Updated Data

Not applicable to this dataset.

# Data Set 2 (Sara Adi, Jennifer Lithgow)

<https://data.ontario.ca/dataset/confirmed-positive-cases-of-covid-19-in-ontario>

## 

Data Description

This dataset consists of the daily snapshots of confirmed, reported cases of Covid-19 reported in Ontario by person.

Format / Organization

The data is formatted in a CSV data file, consisting of 330573 rows, each representing a unique case, which will only increase as new cases are confirmed and the file is updated. The file is also formatted with 18 columns. The columns consist of information on data, age group, gender, Acquisition information, Outcome, PHU, and location of the patient.

Encodings

* **Col. 2-5:** “year-month-day”
* **Col. 8:** “OB” “CC” “Travel” “NO KNOWN EPI LINK” - How the outbreak occurred
* **Col. 11:** “Public Health Unit Number”

Useful Fields

For Jennifer:

* **Col. 3 (index 2**) - “Accurate\_Episode\_Date” - Show the progress over time of the total percentage of the age group infected
* **Col. 7 (index 6)** - “Age\_Group” (decade) - this will be tallied appropriately

For Sara:

* **Col.6:** “Specimen Date”

Groupings - Aggregations - Encodings

**Group:** Sort by accurate episode date

**Aggregate:** None, as this dataset is used for two questions that require different aggregations

**Encodings:** Accurate\_Episode\_Date, Specimen\_Date, Age\_Group

Update Frequency

The source dataset updates daily, but as this is not a publicly used analysis, it is not a priority to be up-to-date. Therefore, weekly updates will be done before demonstrations.

Incorporating Updated Data

Manually re-uploading the CSV file from the Government of Ontario website to the repl and running the preprocessing program.

# Data Set 3 (Sara Adi)

<https://data.ontario.ca/dataset/b1fef838-8784-4338-8ef9-ae7cfd405b41/resource/7fbdbb48-d074-45d9-93cb-f7de58950418/download/schoolcovidsummary.csv>

Data Description

This dataset provides a summary of the confirmed COVID-19 cases within Ontario schools. This set includes faculty and students, and the total cases per day in Ontario schools.

Format / Organization

The data is formatted in a CSV data file. The file consists of 21 columns, and 121 rows, which increase as each update occurs. Each row represents a new report/collected date.

* **Col. 1-2:** Dates of reported/collected
* **Col. 3-5:** Current school information (schools with cases, shutdown, the current total number of schools)
* **Col. 6-9:** Newly, reported cases (student, faculty, unspecified, total)
* **Col. 10-13**: Recently, reported cases (student, faculty, unspecified, total)
* **Col. 14-17:** Past, reported cases (student, faculty, unspecified, total)
* **Col. 18-21:** Cumulative, reported cases (student, faculty, unspecified, total)

Encodings

* **Col. 1 - 2:** “year-month-day” (collected and reported date, respectively)
* **Col. 9, 13, 17, 21:** “unspecified\_case”: Cases from non-staff /students that have a connection to the school

Useful fields

* **Col. 1:** Collected date
* **Col. 6:** New total school-related cases

Data Reorganization

**Grouping**: The necessary 2 columns will be grouped and exported to a separate CSV file. As the original CSV data file is already organized by date, no sorting is required. Preprocessed form:

[collected\_date, new\_total\_school\_related\_cases]

**Encoding**: Col.1 “year-month-day”

Update Frequency

The data is updated daily on weekdays, excluding any provincial holidays. The update occurs from the previous day, data is current as of 2:00 pm the previous day.

Incorporating Updated Data

Manually re-uploading the CSV file from the Government of Ontario website.

# Data Set 4 (Sophie Mlodzik)

<https://data.ontario.ca/dataset/covid-19-vaccine-data-in-ontario/resource/8a89caa9-511c-4568-af89-7f2174b4378c>

Data Description

This data set provides a summary of Covid-19 vaccines in Ontario, including the number of vaccines administered daily, the total number of doses administered, the total number of doses in fully vaccinated individuals, and the total number of fully vaccinated individuals.

Format / Organization

The data is formatted in a CSV file. The file consists of six columns and currently has 85 rows, although a new row is added daily. The data is organized in the following way:

* **Col. 1:** Column ID
* **Col. 2:** Date of reported cases
* **Col. 3:** Number of doses of the vaccine administered the previous day
* **Col. 4-5:** Total number of doses administered (all time, in fully vaccinated individuals)
* **Col. 6:** Total number of individuals that are fully vaccinated

Encodings

* **Col. 2:** “year-month-dayThours:minutes:seconds”
* **Col. 3:** number of vaccine doses administered the previous day (integer)

Useful Fields

* **Col. 2:** “report\_date”
* **Col. 3:** “previous\_day\_doses\_administered”

Data Reorganization

The necessary 2 of the 6 columns have been grouped and exported to a separate CSV file in Repl.it named “data/preprocessed\_vaccines\_administered.csv”. The original CSV file is organized by date, therefore no further organization is required. The new preprocessed file is formatted in the following way:

[report\_date, previous\_day\_doses\_administered]

Column 1 Encoding: “year-month-dayThours:minutes:seconds”

Update Frequency

The data is updated daily at 12am EST. Data from the previous day is appended.

Incorporating Updated Data

Manually re-uploading the CSV file from the Government of Ontario website to the Repl repository and running the preprocessing program.

# Data Set 5 (Emily Kozatchiner)

<https://data.ontario.ca/dataset/cbb4d08c-4e56-4b07-9db6-48335241b88a/resource/ce9f043d-f0d4-40f0-9b96-4c8a83ded3f6/download/response_framework.csv>

Data Description

The data set keeps information on each zone’s different public health and workplace safety measures. Specifically, it includes every update on different severity statuses in each PHU zone. The start date and end date of the zone severity are tied to each PHU status, while multiple statuses are tied to a specific PHU location, both by id and name. There are five different status severities that the Government of Ontario website specified, and are stored in string format for each PHU and date. Created by the Government of Canada.

Format / Organization

The data is organized by 7 columns, currently consisting of 182 records all stored and formatted within a CSV data file. The specifics of column and data organization can be found below:

* **Col. 1:** The current row number representing the record number (amount of inputs)
* **Col. 2:** The reporting PHU name as a string
* **Col. 3:** The id of the PHU (enumerating the reporting PHU string into an integer)
* **Col. 4**: The severity of the PHU status in a specific zone (represented as one of six strings)
* **Col. 5-6:** The start and end date for when the status was enforced onto the PHU zone (strings)
* **Col. 7:** The URL to each PHU (string format)

Encodings

* **Col. 4:** Enumerated strings of PHU status into integers based on severity (“Prevent” = 1, “Protect” = 2, “Restrict” = 3, “Control” = 4, “Stay-at-home” and “Lockdown” = 5)
* **Col. 5:** Changed starting date string into “year-month-day” by removing “T00:00:00”
* **Col. 1, 2, 3, 6, 7:** Dropped to clear up data

Useful Fields

* **Col. 3:** The id of the PHU in order to keep track of information
* **Col. 4:** Severity of PHU zone to keep track of positive/negative change
* **Col. 5:** The start date to measure the changing point of the PHU status

Data Reorganization

All remaining columns will come together and be grouped into a modified CSV file (data/prep\_phu\_reponse.csv). The original CSV data file is organized from least to greatest PHU id, as well as earliest to latest start date for each PHU, so no sorting was done. The modified CSV file will be in form:

[Reporting\_PHU\_id,Status\_PHU,start\_date]

\*Refer to the encodings in order to see what data was modified\*

Update Frequency

The data is updated interchangeably during the week. The start dates between the changes usually have a five-to-seven day difference (there are exceptions - January of 2021 has no data). Therefore I presume the data is updated weekly at the most.

Incorporating Updated Data

* Acquiring and manually updating the CSV file weekly from the Government of Ontario website

# 

# Questions

## Question 1 (Sara Adi)

What is the relation between the number of confirmed positive Covid-19 cases within Ontario schools, vs. the total number of cases for Ontario as a province over time as a function of year and month?

From the Ontario school dataset, the data to be used to answer the question can be found in columns one and six, collected data, and new total school-related cases respectively. The data will be reorganized by date first, then the respective total related cases on each line.

From the Covid-19 dataset, column 6, accurate episode data, is to be used to answer the question. The data will be sorted by specimen date, each row representing a single case

The question parameters are a specified month and year of search, and the program will make the comparison and provide the percentage of school-related cases that contribute to the provinces total.

## Question 2 (Jennifer Lithgow)

What is the percentage of a chosen Ontarian age population that has been diagnosed with Covid-19 since the beginning of the pandemic?

From the census dataset, the data to be used to answer the question are the age (in single years) and total population. The data will be reorganised by age group, the respective population in the same line.

From the Covid-19 infection dataset, the accurate episode date and the age group of the infected person will be used to answer the question. The data will be sorted by accurate episode date, the corresponding age in the same line.

The question parameters are the age range the user wants to know the percentage infected, and the program will find the cumulative sum (and graph it) from the earliest recorded date to the latest update.

## Question 3 (Sophie Mlodzik)

Given a growth rate of vaccines administered daily, calculated through the number of vaccines previously administered per day over a period of time, how many vaccines will be administered a specified number of days into the future?

From the Covid-19 vaccine dataset, the report date and number of doses administered the previous day will be used. The data is sorted by report date.

The question parameters are the number of days into the future that the user would like to predict the number of vaccines administered, and the program will calculate a prediction based on the data from previous days.

## Question 4 (Emily Kozatchiner)

What is the rate of change for a PHU status throughout a period of time in the COVID pandemic? Is it a rise or decline in extremity?

By reading and inputting the values from the chosen dataset, we can calculate the rate of change using the start dates and change in status per PHU id. The start dates will be used in order to indicate when the change in status occurred. The data will be organized by date, so it should flow linearly.

The question parameters require the user input, and so the user’s choice of PHU id is the one the rate of change is measured by. The program will then output the rate of change and display the results via line graph.