**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**

[**Section 1.0 - Sara Adi**](#_8ay46778ly4p)4

[Question](#_54cjtbpaazq2) 4

[Group Interest](#_xcr80o2b2y4v) 4

[Data Sources](#_hn1ng5g1a0ud) 4

[Update Frequency](#_hybln5g27o2l) 4

[Variable Values](#_9unfejdruiza) 4

[Data Organization](#_7s5104v1fd84) 4

[Relevant Data](#_t5g0ocomj83a) 5

[Preprocessing](#_ojw3govxfp4i) 5

[Parameterization](#_oy2mqfyije7h) 5

[Visualization](#_xzbfmp3lo6qh) 5

[**Section 2.0 - Jennifer Lithgow**](#_9arc8qibsbyb)6

[Question](#_caszyrjpoba7) 6

[Group Interest](#_fae557xm2mt1) 6

[Data Sources](#_g6309u987mlu) 6

[Update Frequency](#_qwhx1o1cus5i) 6

[Variable Values](#_ryu4ekvtykyj) 6

[Data Organisation](#_7fmzvciylj5z) 7

[Relevant Data](#_m244c4q4kdhz) 7

[Preprocessing](#_36qt6fd53n7j) 7

[Parameterization](#_mnaw1o474rqm) 7

[Visualization](#_g7uea9whkwvf) 7

[**Section 3.0 - Emily Kozatchiner**](#_kmj59zzdwz2g)8

[Question](#_x4miwwtocffd) 8

[Group Interest](#_huq8mwsjuzx1) 8

[Data Sources](#_g7csj549wwwj) 8

[Update Frequency](#_xr36hwfioq4i) 8

[Variable Values](#_pc904im8nc4k) 8

[Data Organization](#_ljdyurxph3ad) 8

[Relevant Data](#_pj49aymdzo7u) 8

[Preprocessing](#_tcwe3y52kon5) 9

[Parameterization](#_otg785793z0r) 9

[Visualization](#_wb7ijf98a4p7) 9

[**Section 4.0 - Sophie Mlodzik**](#_jkg957g4slfe)10

[Question](#_huvizu406wxc) 10

[Group Interest](#_qvb4e5c90ykg) 10

[Data Sources](#_qo45wpr5azm3) 10

[Update Frequency](#_vmuf2yi2g2wz) 10

[Variable Values](#_la14xnfw1bli) 10

[Data Organization](#_ourcivfhi245) 10

[Relevant Data](#_n773i6hluys4) 10

[Preprocessing](#_jyr9qt2vwk9f) 10

[Parameterization](#_58089hipo878) 11

[Visualization](#_i3sbmblue3yp) 11

# Section 1.0 - Sara Adi

## Question

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?

## Group Interest

The biggest debate between the ministry of education and the general public was, “is it safe, for younger kids, to be back in school?”. The reopening of school meant higher touchpoints, increased sanitization and protocols, and increased risk of spreading the disease to our younger population, who can go home and spread it to relatives of high risk. Interestingly, to my group, many teachers and faculty members fought back against the reopening of schools due to their safety. Although some parents and community members agree, others say that it is time children go back and get the in-person education they need.

Therefore, by comparing the data of cases found within schools in the province compared to that of the total province count for the same timeframe, we can determine just how serious the consequences are of the reopening plans.

## Data Sources

1. Government of Ontario

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

1. Government of Ontario

<https://data.ontario.ca/dataset/confirmed-positive-cases-of-covid-19-in-ontario/resource/455fd63b-603d-4608-8216-7d8647f43350>

## Update Frequency

* Daily

## Variable Values

* Integers, only real numbers above 0 (the values inputted/outputted will be case counts and dates).
  + Case total for schools for that date
  + Case total for ontario for that date
  + Specified year and month (command line arguments)
* No “special” codes

## Data Organization

* Date of the case reported
* Date of the case collected

## Relevant Data

1. Relevant: columns 2, 6 Irrelevant: columns 1, 3-5, 7-21
2. Relevant: columns 3 Irrelevant: columns 1-2, 4-18

## Preprocessing

Organizing the first csv file (school data), in terms of a new file for each month and year (i.e. filename: 2020-09.txt). Where in the given example each row will hold all the cases reported for every valid day within the month September of 2020. This will help organize the data in terms of time, making it easier to do any addition to account for the total number of cases per month. Organizing the second csv file (total ontario data as a province), would be done similarly as the first csv file.

## Parameterization

| $python school\_covid.py <Filename #1> <Filename #2> <specified year > <specified month>  $python school\_covid.py data/schoolcovidsummary.csv data/conposcovidloc.csv 2020 06 |
| --- |

argv[0] = python script

argv[1] = filename with school related data

argv[2] = filename with Ontario data

argv [3] = specified year of search

argv [4] = specified month of search

## Visualization

The x-axis consists of a specific ranged time period (in the month of a chosen year), while the y-axis consists of the total number of positive cases. Since we are looking at a numerical comparison over time, a histogram is a good choice.

In order to separate the data sets between days in the chosen month, a grouped histogram would be used with different coloured bars distinguishing between positive cases in schools vs. in Ontario (a stacked histogram).

# Section 2.0 - Jennifer Lithgow

## Question

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

## Group Interest

This is a question of interest because there is potential misinformation about the impact of Covid-19 on different age groups. For example, there are claims that the virus doesn’t affect young adults as much as other age groups. By showing the percentage of a user’s chosen age range that has been infected, hopefully a better understanding of Covid-19’s impact can be demonstrated.

## Data Sources

1. 2016 Canadian census age and gender data tables (Census):
   1. Contains age ranges as well as specific ages for Ontario (the selected section of data)
   2. <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>
2. Ontario Covid-19 infection data (Infection):
   1. Ontario positive cases of Covid-19
   2. Age groups are sorted by decades; this creates a constraint on our ability to manipulate the data
   3. <https://data.ontario.ca/dataset/confirmed-positive-cases-of-covid-19-in-ontario>

## Update Frequency

1. Census: Every 5 Years
2. Infection: Daily

## 

## Variable Values

1. Census:
   1. Age will be integers
   2. Total populations of age groups will be integers
2. Infection:
   1. Age will be integers
      1. We will have to remove the ‘s’ character that is present in the column
   2. Total infections for a given age group will be integers

## Data Organisation

The data from the census will be sorted by increasing age group, as that is the important part of the information to use in answering the question. As for the data in regards to infection rates, it would be better to sort it by date, since that is the order in which it will be displayed.

## Relevant Data

1. Census: 2 columns relevant:
   1. Col. 1 (index 0) - Age. There is a mix of integers and ‘phrases’ that will need to be ‘translated’ into the decade age-range (e.g. 20-29, 50-59)
   2. Col. 2 (index 1) - total population that are of the given age/age range
   3. Cols. 3, 4 ignored
2. Infection: 2 columns relevant:
   1. Col. 3 (index 2) - “Accurate\_Episode\_Date” - If we want to show the progress over time of total percentage of the age group infected
   2. Col. 7 (index 6) - “Age\_Group” (decade) - this will be tallied for the appropriate age range
   3. Col. 12 (index 11) - “Reporting\_PHU\_ID”
   4. Cols. 1, 2, 4,-6, 8-11, 13+ ignored

## Preprocessing

1. Census: Find decade age range, create a cumulative sum
2. Infection: Calculate sum of age range infected on specific date

## Parameterization

| $python age\_percentage.py <census filename> <covid-19 infections filename> <start decade> <end decade>  $python age\_percentage.py data/10952620210311023233.csv data/conposcovidloc.csv 20 50 |
| --- |

argv[0] = python script

argv[1] = Filename holding census data

argv[2] = Filename holding infection data

argv[3] = integer indicating start decade of age population

argv[4] = integer indicating end decade (exclusive) of age population

## Visualization

Credit for suggestions given: Emily Kozatchiner

Graph: Histogram of user-selected age group over time, across Ontario

X-axis: Time displayed in regular intervals (years, months, days, etc.)

Y-axis: Percentage of chosen population infected

Singular output (optional): total percentage infected

# Section 3.0 - Emily Kozatchiner

## Question

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

## Group Interest

This question was of interest to us because PHU status often indicates how dangerous a region could be, and comparing how PHU’s differentiate over time from one another could provide some insight into the flow of the pandemic across regions. Showing the difference in change between regions could show that some regions are more unstable than others and may need to be avoided/helped depending on the focus of the user. It could also tell the user which months between all the PHU regions were the hardest, and where the trend is headed for future labeling of regions (for better or for worse). Although it is simply displaying the data for the user to view, the user could extrapolate and connect ideas about how the PHU status has seemed to progress.

## Data Sources

Obtained from: Government of Ontario

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

## Update Frequency

Interchangeable. Usually data is updated approximately every five-to-seven days (there are exceptions - January of 2021 has no data)

## Variable Values

Integers: PHU Id, PHU status, (year, month, day) - all dates

Doubles: The rate of change

Strings: Taking in the PHU status

* No “special” codes

## Data Organization

* PHU status will be a data set organized per a certain PHU Id and and a certain date.
* The data will be organized over a specific period of time

## Relevant Data

Column 3, 4, and 5 are relevant containing the data sets:

Column 3 = PHU Id

Column 4 = PHU Status

Column 5 = Starting date of newly applied status

Columns 1, 2, 6-8 are irrelevant.

## Preprocessing

The data will be preprocessed by pulling out a PHU Id and organizing all data based on the PHU Id (PHU status & date) in separate files. Therefore the csv file will contain data specific to each PHU. When the user enters a PHU Id, the script will access the PHU-specific file for efficient data collection.

## Parameterization

| $python PHU\_status\_change.py <Filename #1> <specified PHU>  $python school\_covid.py data/response\_framwork.csv 2262 |
| --- |

argv[0] = python script

argv[1] = Filename holding data

argv[2] = PHU Id parameter

## Visualization

A line graph will show the rate of change between the changes of PHU status over time. The independent variable being time and the dependent variable being a value that corresponds with the severity of a PHU status. The elements of the line graph will consist of however many dates are found for the specific PHU in the data file.

The x-axis will be organized by an orderly time difference (months, days, etc.) rather than each time the PHU status changes.

# Section 4.0 - Sophie Mlodzik

## Question

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?

## Group Interest

This is a question of interest to the group because many people, ourselves included, are very anxious to get the Covid-19 vaccine. By creating a program that can predict the number of vaccines administered on a given date in the near future, it can give us peace of mind and help us predict when we might be able to receive the vaccine based on the number of doses being administered daily.

## Data Sources

1. Government of Ontario Covid-19 Vaccine Data

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

## Update Frequency

* Daily

## Variable Values

* Integers
  + Number of days into the future (to predict the number of vaccines)
  + Number of vaccines administered each day, number of days
* Reals
  + Rate of change in number of vaccines administered
* No “special” codes

## Data Organization

* Line graph to represent data from past days
* Trend line, in addition, to represent predictions for future days

## Relevant Data

* Columns 2 and 3, report\_date and previous\_day\_doses\_administered, are relevant in calculating the rate of growth
* All other columns (1, 4+) can be disregarded

## Preprocessing

Sort the data into two lists:

* Dates
* Vaccines administered per day

## Parameterization

| $python <program.py> <filename\_1> <number of days into the future> $python vaccine\_predictions.py data/vaccines\_administered 5 |
| --- |

argv[0] = python script

argv[1] = filename with vaccine data

argv[2] = number of days

## Visualization

A line graph is suitable for analyzing growth rate of the number of daily COVID vaccinations, as well as predicting future vaccination values. The independent value, time, will be represented through days, and the dependent value, vaccinations administered, will be represented in the number of doses. The line graph will have a trend line in order to provide the user with a projection of future vaccines that will be administered in the coming days.