CSI 4142 – Fundamentals of Data Science

Winter 2021

School of Electrical Engineering and Computer Science

University of Ottawa

Professor: Dr. Herna L. Viktor (hviktor@uottawa.ca)

Teaching Assistants: Nicolas (nflee092@uottawa.ca) and Parsa (pvafa014@uottawa.ca)

Project Deliverable C: Physical Design and Data Staging

Group #5

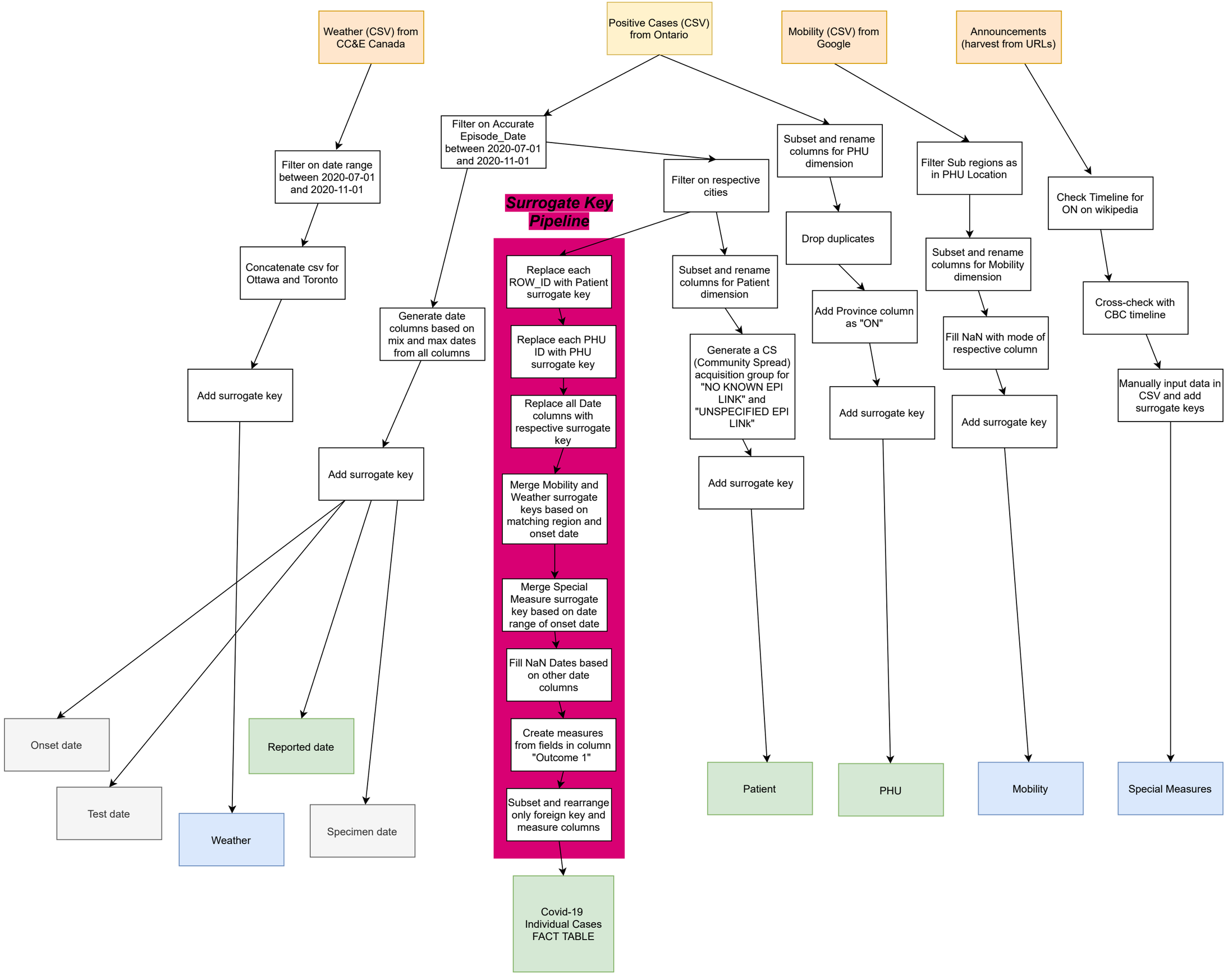
Vekshan Bundhoo (vbund062@uottawa.ca), 300035157

Le Nguyen (lnguy042@uottawa.ca), 300013304

Sukhsimranpreet Sekhon (ssekh066@uottawa.ca), 300018861

Due Date: March 10th, 2021 by 9:30 PM

Submission Date: March 10th, 2021


Weather (CSV) from CC&E Canada
Positive Cases (CSV) from Ontario
Mobility (CSV) from Google
Announcements (harvest from URLs)
Filter on date range between 2020-07-01 and 2020-11-01
Filter on Accurate Episode_Date between 2020-07-01 and 2020-11-01
Filter on respective cities
Subset and rename columns for PHU dimension
Filter Sub regions as in PHU Location
Check Timeline for ON on wikipedia
Concatenate csv for Ottawa and Toronto
Generate date columns based on mix and max dates from all columns
Surrogate Key Pipeline
Subset and rename columns for Patient dimension
Drop duplicates
Subset and rename columns for Mobility dimension
Cross-check with CBC timeline
Add surrogate key
Add surrogate key
Replace each ROW_ID with Patient surrogate key
Replace each PHU ID with PHU surrogate key
Replace all Date columns with respective surrogate key
Merge Mobility and Weather surrogate keys based on matching region and onset date
Merge Special Measure surrogate key based on date range of onset date
Fill NaN Dates based on other date columns
Create measures from fields in column "Outcome 1"
Subset and rearrange only foreign key and measure columns
Covid-19 Individual Cases FACT TABLE
Generate a CS (Community Spread) acquisition group for "NO KNOWN EPI LINK" and "UNSPECIFIED EPI LINk"
Add surrogate key
Add Province column as "ON"
Add surrogate key
Fill NaN with mode of respective column
Add surrogate key
Manually input data in CSV and add surrogate keys
Add surrogate key
Onset date
Test date
Weather
Reported date
Specimen date
Patient
PHU
Mobility
Special Measures

Tool Stack

- ➔ Python version 3.8
- ➔ Python packages: numpy, pandas, holidays and pandasql
- ➔ Microsoft Excel
- ➔ Jupyter Notebook
- ➔ PostgreSQL and pgAdmin 4 v4

Data Quality Issues

- Overview / Fact Table
    - The data was filtered on the onset_date (accurate episode date) from 1st of July 2020 to 31st of October 2020. However, the other date fields had values ranging from 31st March 2020 to 21st February 2021.
    - Columns Test Date and Specimen Date had missing values. We noticed a trend in the dataset that Test Reported date in general can be taken as the same date that the case was reported and specimen date can be taken as 2 days. We estimate it takes on average 2 days to get results/report based on the data.
    - Cities naming was inconsistent across data sources and several transformations were made for joining on city/region during the surrogate key pipeline step.
    - Most special measures had overlapping dates so only the main measures (specific stages) were joined to each row as they would not overlap ('Stage 3', 'Stage 2', 'Stage 3 Modified', 'Stage 2 Modified'). Other events happening in parallel can still be queried.
- Patient
    - Individual cases were recorded and the surrogate key is preventing duplicates. Aggregation can still be done via queries.
    - In the data dictionary for the source, *'NO KNOWN EPI LINK'* and *'UNSPECIFIED EPI LINK'* were present. By definition and for simplicity we can save both as Community Spread which we labelled as CS upon staging. There was also a naming change in October 2020 which is not in our timeline but could be accounted for in a similar manner.

        - "A 'case with an epidemiological link' is a case that has either been exposed to a confirmed case, or has had the same exposure as a confirmed case (e.g. eaten the same food, stayed in the same hotel, etc)." https://deputyprimeminister.gov.mt/en/health-promotion/idpcu/Pages/case definition.aspx
    - The Outbreak Related column was changed to proper boolean format replacing all yeses with 1 and NaN with 0.
- Mobility
    - When dealing with any null values in any of the baseline columns in the mobility table, we determined that all the missing values should be replaced by the mode value based on their respected column. This is so we can show changes in the value of baseline rather than setting all missing values to 0 that can mislead viewers to think that zero changes occur in the baseline. In addition, the mode value was used since it contains the least amount of outliers that can greatly alter the replacement values unlike the values resulted from the mean or median.
- Weather
    - Since the comparison is between the 2 big cities in general, we took a design decision of using weather data from 1 station in each location. We should also note that weather would be very similar for Municipalities around Toronto and the Main City compared to *Mobility*. Mapping was made from a combination of City/ Subregion and Onset Date. See generate_fact_table.ipynb for more details.
- Special Measures
    - When gathering and recording special measures, we had to cross check between the timeline of Covid-19 in Ontario from CBC and Wikipedia manually which sometimes resulted in dates differing by 2 to 3 days for province wide announcements and measures such "Stage 2", "Stage 3", etc. We decided to take the minimum of the dates to ensure a smooth merging.
- Date
    - See Overview / Fact Table section

Deliverable Checklist & Project Planning

Deliverable	Team member(s) responsible	Expected completion date	Actual completion date	Estimate time (hours) to complete	Actual time (hours) to complete	Notes (if any)
Create database instance and tables	Le Nguyen and Sukhsimranpreet Sekhon	03/02/2021	03/09/2021	1	2.5	This also includes importing csv on pgsql.
Staging of Date dimension	Sukhsimranpreet Sekhon	02/26/2021	03/09/2021	1	2.5	We filtered 4 months (1st July 2020 to 31st October 2020) based on the onset date but other date fields had values ranging from 31st March 2020 to 21st February 2021.
Staging of Patient dimension	Vekshan Bundhoo	02/24/2021	02/24/2021	1	1	
Staging of PHU dimension	Vekshan Bundhoo	02/24/2021	03/03/2021	1	0.5	
Staging of Mobility dimension	Le Nguyen	02/26/2021	03/06/2021	1	2	When dealing with any null values in any of the baseline columns in the mobility table, we determine that all the missing values should be replaced by the mode value based on their respected column.
Staging of Weather dimension	Le Nguyen	02/23/2021	03/03/2021	1	1.5	
Map PHU, Mobility and Weather dimensions	Vekshan Bundhoo	02/28/2021	03/05/2021	1	1.5	Since we are comparing the 2 big cities in general, we took a design decision of using weather data from 1 station in each location. Mapping was made from a combination of City and Onset Date. See generate_fact_table. ipynb for more details.
Staging of Special Measures dimension	Sukhsimranpreet Sekhon	02/28/2021	03/08/2021	1	2	Cross checked between the timeline of Covid-19 in Ontario from CBC and Wikipedia manually.
Surrogate key pipeline – including role-playing dates	Vekshan Bundhoo and Sukhsimranpreet Sekhon (Dates)	03/02/2021	03/09/2021	1.5	2.5	We tried as far as possible (for dimensions as well) to avoid the use of loops and took advantage of pandas available functions. In case of loops, pandas has to create a Series for each row and it is very time-consuming.
Staging of fact table – including FKs and measures	Vekshan Bundhoo	03/02/2021	03/09/2021	1	1.5	This was done in parallel with surrogate key pipeline.
Data quality handling and reporting	All team members	03/03/2021	03/09/2021	1	1.5	
Others - if any						