Phase 2 - Physical Design and Data Staging

CSI4142[A] Fundamentals of Data Science

Group 27

Julien Martinet, 300115646
Dalia Sawaya, 300111681
Zahra Manochehri, 300131603

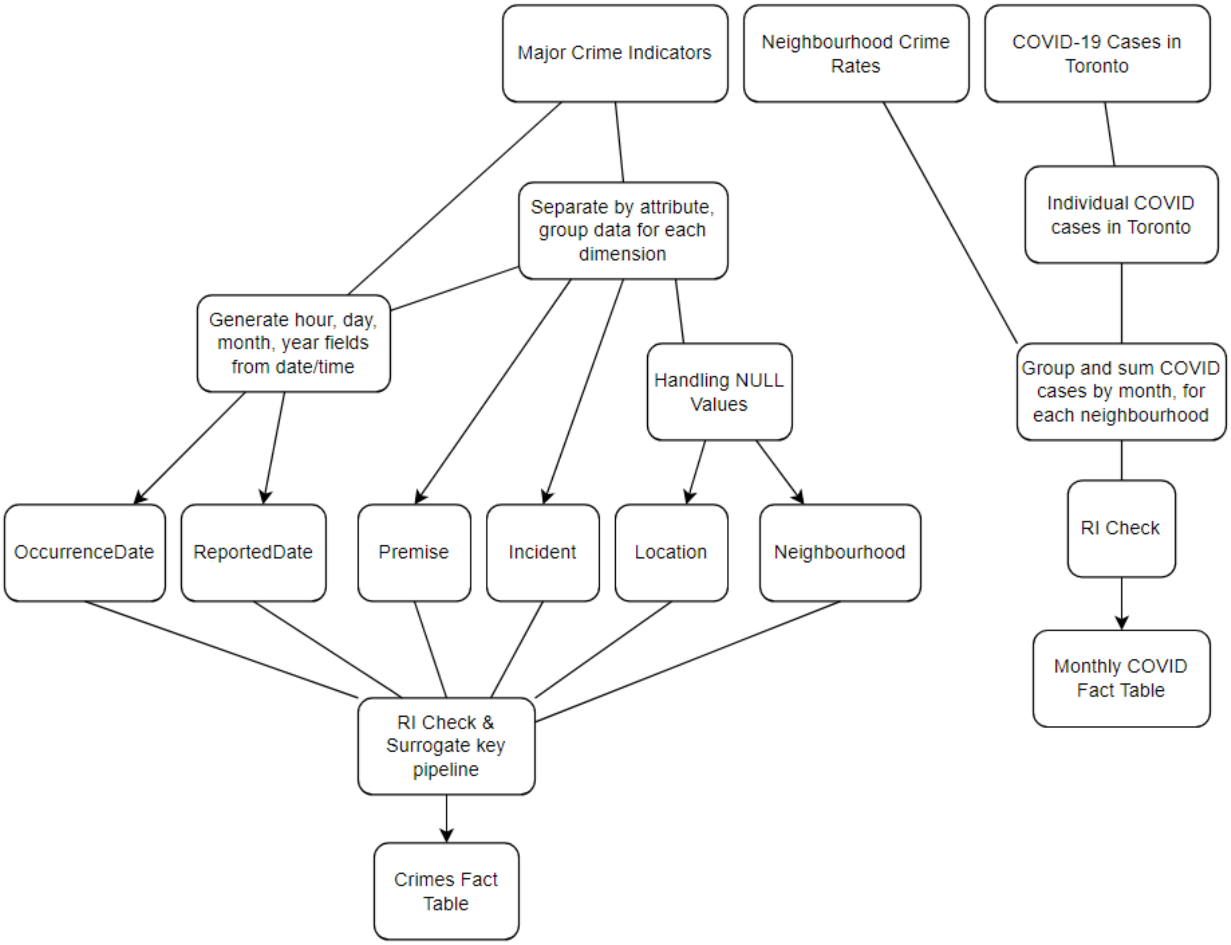
Table of Contents

Source Code

The source code for this deliverable can be found at our GitHub repository:

https://github.com/zhrmnch/CSI4142_Phase3

High-Level Data Staging Plan


Major Crime Indicators
Neighbourhood Crime Rates
COVID-19 Cases in Toronto
Separate by attribute, group data for each dimension
Individual COVID cases in Toronto
Generate hour, day, month, year fields from date/time
Handling NULL Values
Group and sum COVID cases by month, for each neighbourhood
OccurrenceDate
ReportedDate
Premise
Incident
Location
Neighbourhood
RI Check
Monthly COVID Fact Table
RI Check & Surrogate key pipeline
Crimes Fact Table

Data Quality Issues

Handling Duplicates

For every table, we removed duplicates to make sure we did not have the same values more than once in the tables. This would cause data quality issues, so we made sure to remove duplicate entries.

Handling Null Values

1. Location and Neighbourhood

   Problem:
   Some values have one, some or all:
   - Division = NSA (null)
   - Latitude = 0
   - Longitude = 0
   - Neighbourhood = NSA (null)

   3880 crimes have no longitude/latitude where 3588 of which do not have an associated neighbourhood.

   Solution:
   We spoke to the TA about these and since they were a small number compared to the amount of values we had in the entire dataset, we decided to remove them.

2. OccurrenceDate and ReportedDate

   Problem:
   OccurenceYear, OccurenceDay, OccurenceDayOfYear, OccurenceDayOfWeek, ReportedYear, ReportedDay, ReportedDayOfYear, ReportedDayOfWeek contain null values.

   Solution:
   Since the ‘Occurencedate’ and the ‘ReportedDate’ columns can be found in every row, we decided to remove all columns mentioned above that had some missing values. We decided to recreate the deleted columns by using the values from the ‘Occurencedate’ and the ‘ReportedDate’ columns.

Handling Inaccurate Fields

1. Covid Fact table fields

   Problem:
   This is a warning we found from the Toronto Open Data Website, where we got our COVID dataset from (link in references below):

   “As of February 2023, the fields ‘currently hospitalized’, ‘currently in ICU’ and ‘currently intubated’ have been removed from the Open Data set. Due to current provincial guidelines on COVID-19 case management, discharge information is not always available. This makes it difficult to report accurately on these fields. The time period for the inaccuracy is not known therefore data in these fields from previous downloads of the open data set should be interpreted with caution.”

   Solution:
   We decided to remove these measures that we planned on having as part of our Monthly COVID fact table. We made this decision based on the warning above put on the website where we got the dataset from.

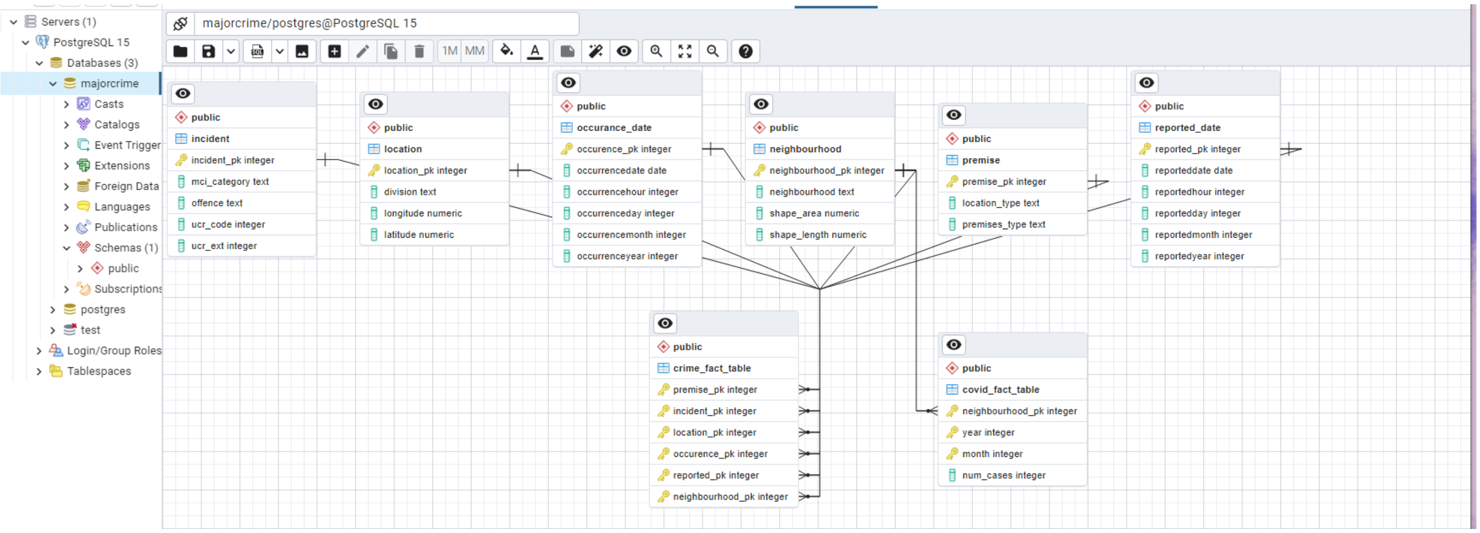
2. Neighbourhood

   Problem:
   The “Neighbourhood Name” fields were not consistent across all datasets. For example, in our Major Crime Dataset, there exists fields with Neighbourhood Name values as “LAmoreux”. In contrast, in our Neighbourhood Crime Rates dataset, that same neighbourhood was labelled as “L’Amoreux”. This is only one example of this type of data quality issue, there were in fact multiple neighbourhood values that were skewed across our 3 tables.

   Solution:
   To ensure referential integrity, we manually modified all occurrences of inconsistent neighbourhood names using the pandas.replace() function on our dataframes. Once finished, we additionally verified that all rows were maintained when merging, and no null values were generated to ensure referential integrity.

Creating Data Mart

In order to build our data mart, we opted to use PostgreSQL as the database management system, along with pgAdmin as the graphical user interface. The figure provided below offers a comprehensive view of all tables present within the primary crime/covid database. For access to the SQL code utilised in creating these tables, please refer to the file in the Git repository.



public
incident
incident_pk integer
mci_category text
offence text
ucr_code integer
ucr_ext integer

public
location
location_pk integer
division text
longitude numeric
latitude numeric

public
occurance_date
occurence_pk integer
occurrencedate date
occurrencehour integer
occurrenceday integer
occurrencemonth integer
occurrenceyear integer

public
neighbourhood
neighbourhood_pk integer
neighbourhood text
shape_area numeric
shape_length numeric

public
premise
premise_pk integer
location_type text
premises_type text

public
reported_date
reported_pk integer
reporteddate date
reportedhour integer
reportedday integer
reportedmonth integer
reportedyear integer

public
crime_fact_table
premise_pk integer
incident_pk integer
location_pk integer
occurence_pk integer
reported_pk integer
neighbourhood_pk integer

public
covid_fact_table
neighbourhood_pk integer
year integer
month integer
num_cases integer

Team Planning Spreadsheet

Deliverable Checklist	Responsible Team Member(s)	Expected Completion Date	Actual Completion Date	Estimated time to complete	Actual time to complete	Notes (if any)
Create database instance	Julien, Zahra, Dalia	March 14	March 12	30 mins	30 mins	–
Create OccurenceDate dimension	Julien, Zahra, Dalia	March 14	March 12	15 mins	15 mins	–
Create ReportedDate dimension	Julien, Zahra, Dalia	March 14	March 12	15 mins	10 mins	–
Create Premise dimension	Julien, Zahra, Dalia	March 14	March 12	15 mins	10 mins	–
Create Incident dimension	Julien, Zahra, Dalia	March 14	March 12	15 mins	10 mins	–
Create Location dimension	Julien, Zahra, Dalia	March 14	March 12	15 mins	10 mins	–
Create Neighbourhood dimension	Julien, Zahra, Dalia	March 14	March 12	15 mins	10 mins	–
Staging OccurenceDate dimension	Julien, Zahra, Dalia	March 20	March 20	30 mins	30 mins	–
Staging ReportedDate dimension	Julien, Zahra, Dalia	March 20	March 20	30 mins	30 mins	–
Staging Premise dimension	Julien, Zahra, Dalia	March 20	March 20	30 mins	20 mins	–
Staging Incident dimension	Julien, Zahra, Dalia	March 20	March 20	30 mins	20 mins	–
Staging Location dimension	Julien, Zahra, Dalia	March 20	March 20	30 mins	20 mins	–
Staging Neighbourhood dimension	Julien, Zahra, Dalia	March 20	March 20	30 mins	20 mins	–
Surrogate key pipeline	Julien, Zahra, Dalia	March 21	March 20	45 mins	40 mins	–
Staging of Crime fact table – including FKs and measures	Julien, Zahra, Dalia	March 21	March 20	45 mins	45 mins	–
Staging of Monthly COVID fact table – including FKs and measures	Julien, Zahra, Dalia	March 21	March 20	30 mins	30 mins	–
Data quality handling and reporting	Julien, Zahra, Dalia	March 22	March 22	1 hour	1.5 hours	–
Writing Report	Julien, Zahra, Dalia	March 24	March 23	3 hours	4 hours	–

Meeting Times

Dates and Times we met as a group:

1. Friday March 10, at 10:00AM-11:30AM (1.5 hour)
2. Wednesday March 15, at 11:30AM-1:00PM (1.5 hour)
3. Monday March 20, at 12:30PM-4:30PM (4 hours)
4. Tuesday March 21, at 2:00PM-3:00PM (1 hour)
5. Wednesday March 22, at 11:30AM-1:00PM (1.5 hours)

Dates and Times we met with the TA:

1. Tuesday March 21, at 1:30PM-2:00PM

References

1. https://data.torontopolice.on.ca/datasets/TorontoPS::major-crime-indicators-1/explore?location=20.627203%2C-40.021433%2C4.75&showTable=true
2. https://open.toronto.ca/dataset/covid-19-cases-in-toronto/
3. https://data.torontopolice.on.ca/datasets/TorontoPS::neighbourhood-crime-rates/about