ETL Project

By Ravi Selva, Aron Rotman, Nicky Pant

Sources: Kaggle

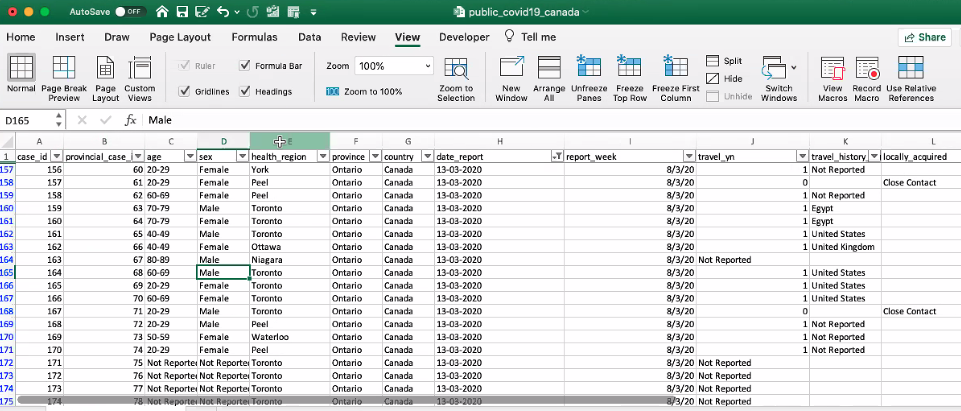
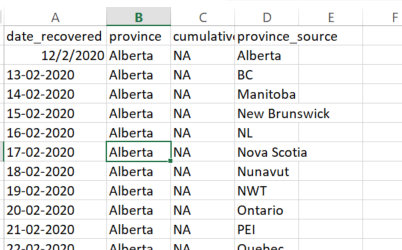
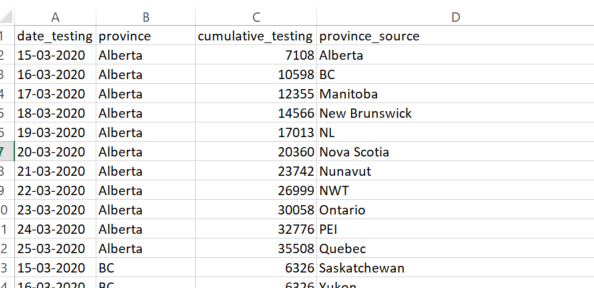
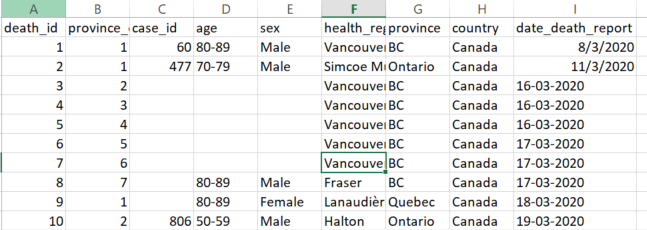
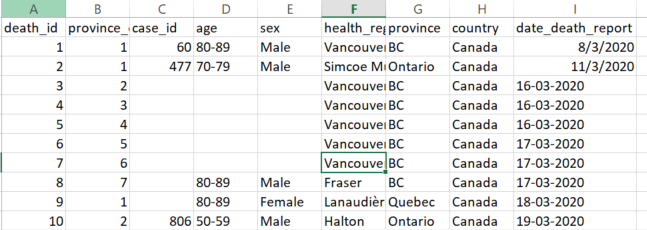
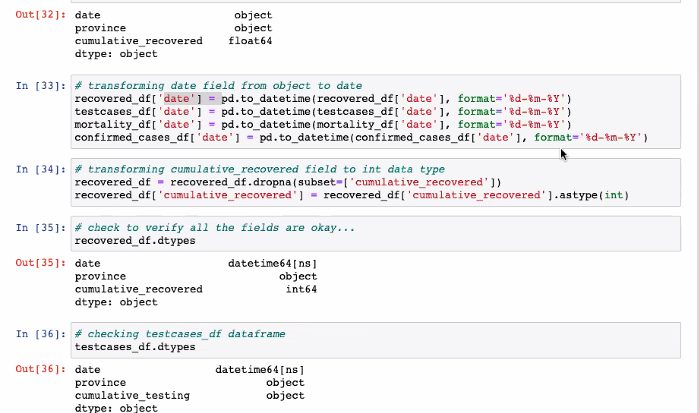
# **Topic:**

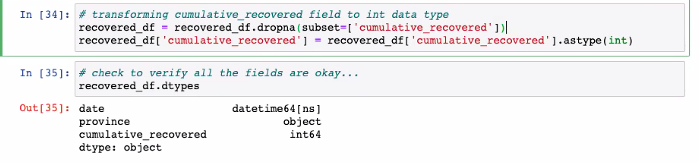
* **Corona virus** 
  + Objective: Aggregate, clean and load data for the Novel Corona Virus from multiple sources for different metrics such as mortality, testing and recovered

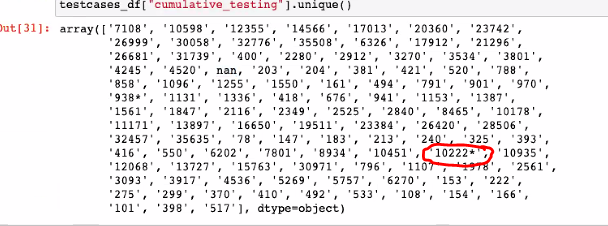
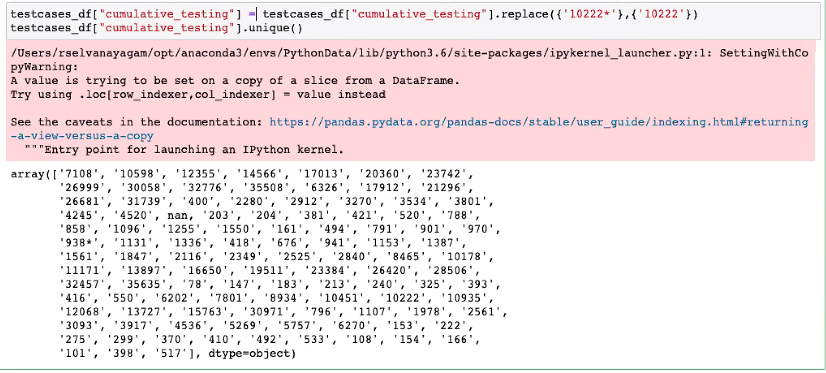
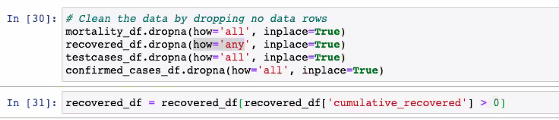
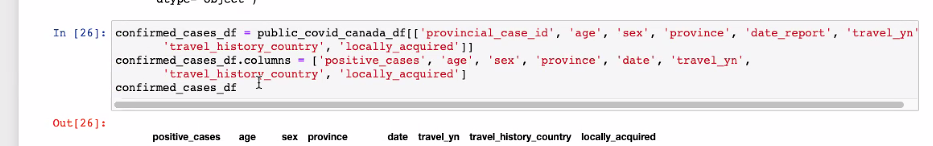
# **Extraction:**

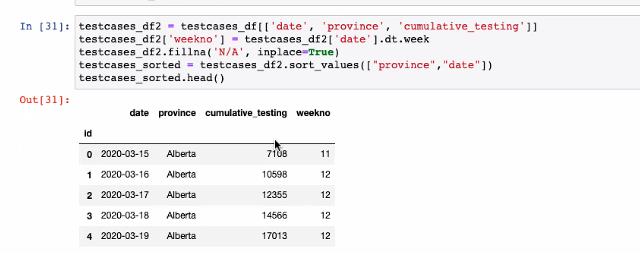
* + We used 3 different datasets from the public platform Kaggle. The data in the three included information for Canada only
  + The data in the three files included the following information
    - Positive Cases + general information (Age/Gender/Travel info)
    - Number of Deaths
    - Number of Test Cases
    - Number of Recovered
  + Source:
    - Canada
      * <https://www.kaggle.com/zinx1991/covid19-in-canada?select=Testing_Canada+.csv>

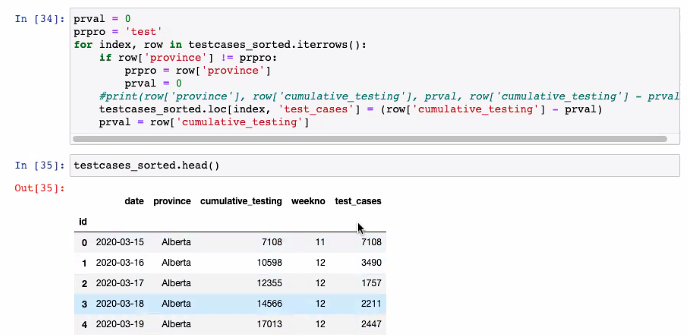
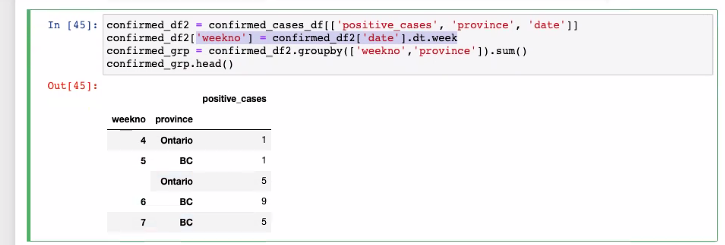
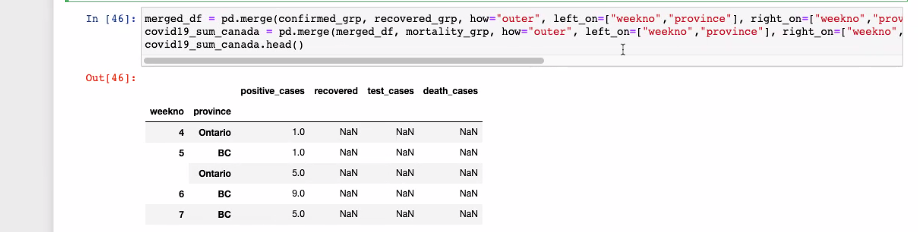
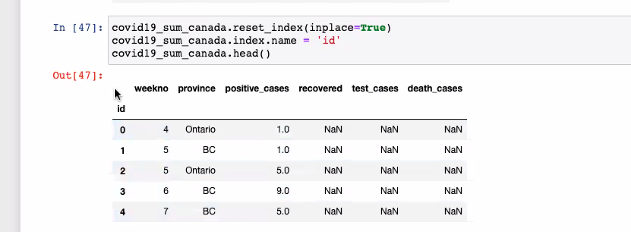
# **Transformation:**

* In order to transform the public data and use it in our study we performed the following actions:
  + Used Pandas functions in Jupyter Notebook to load all four CSV files
  + Removed a few columns
    - **Public Covid\_19 Canada**
      * 
      * Columns we dropped: case\_id, health region, country, report\_week, travel\_yn, travel history, locally acquired etc.
    - **Recovered Data:** 
      * 
      * Columns we dropped: province\_ source
    - **Testing Data:**
      * 
      * Columns we dropped: province\_ source
    - **Mortality:**
      * 
      * Columns we dropped: health region etc.
  + Formatting the data:
    - Cleaned up date formats that were inconsistent across csv files
    - 
    - We are changing the date field to one format, and then instead of using the VarChar we are able to use the Date in SQL
  + 
  + Changed multiple records as integer from other object types, taking out NA’s non integer values

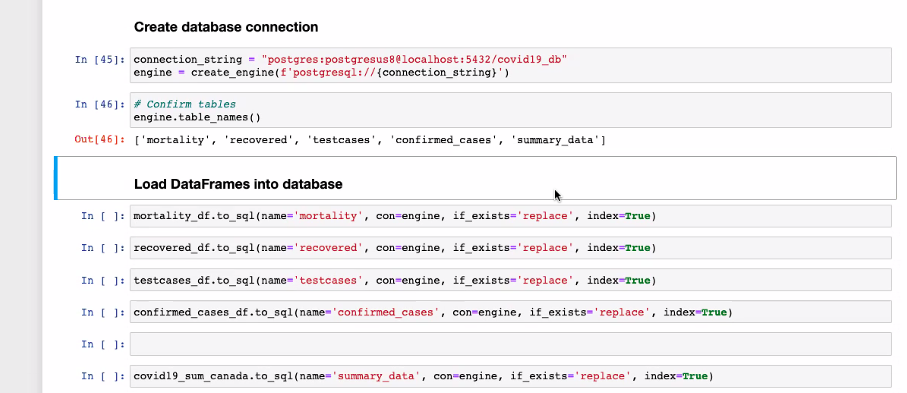
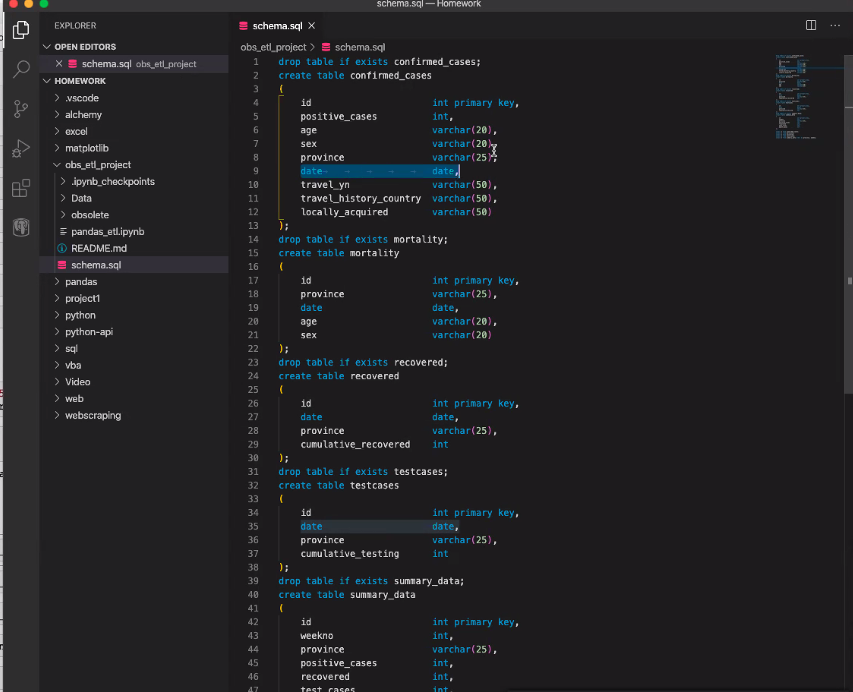
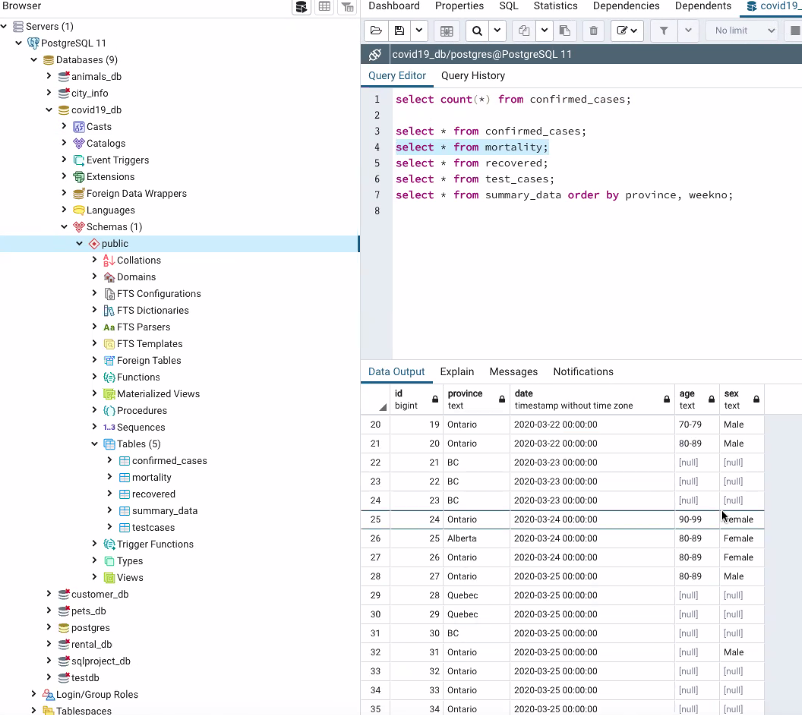
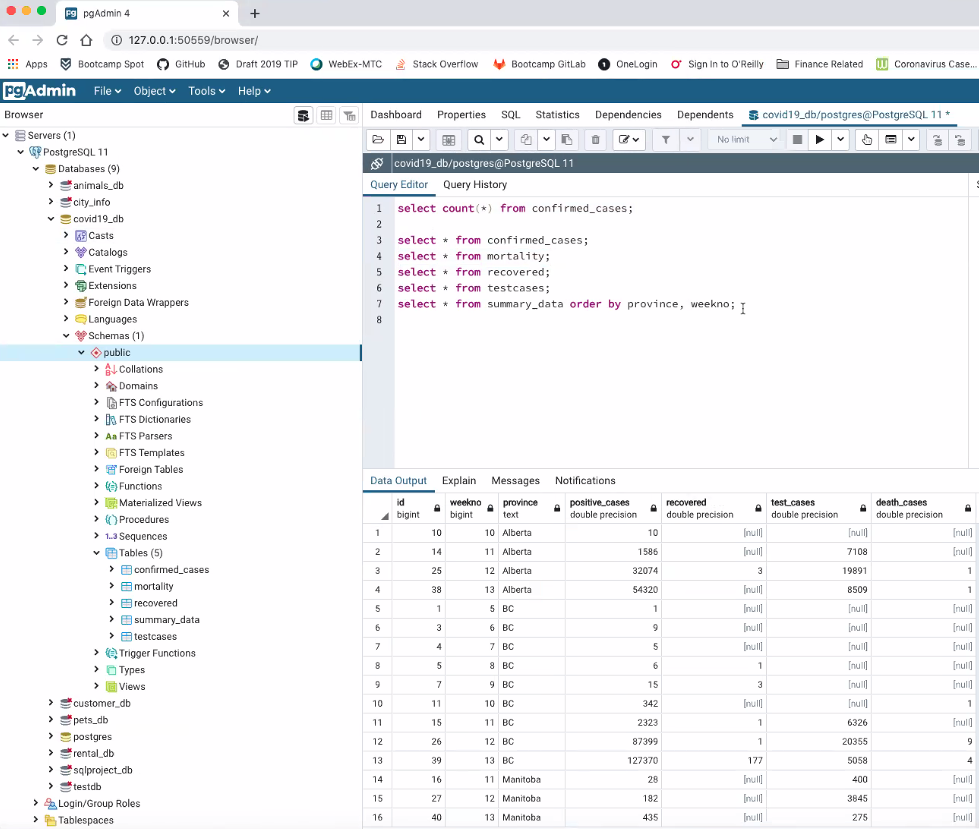
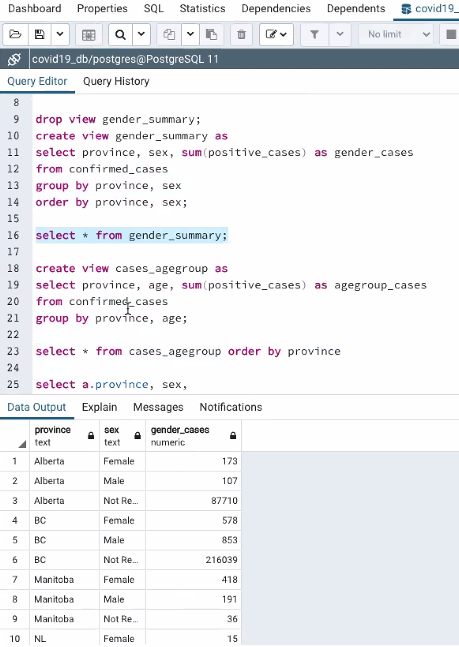
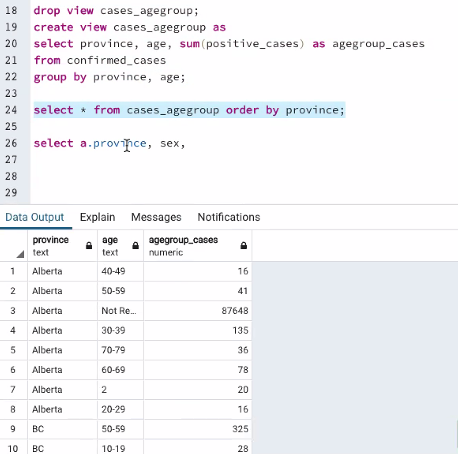


* Cleaned up data by removing “value\*” from tables:
  + 
  + Update code::
  + 
* **Dropping** **rows:** with NAN value across columns:
  + 
* **Clean-up:** Removed ‘0’ or blank values
  + 
* **Clean-up:** Renamed columns
  + 
* **Grouping** 
  + Week number field by grouping date & province



* + Discovered that adding a new field, also added that field to the original source file. Fixed issue by adding [field].
* **Adding new field:**
  + Delta form previous day of test cases and recovered cases
  + 
* **Adding new field:** count of death cases and sum of positive cases
  + 
* **Merging** **four data-frames:** using outer join
  + 
* **Setting index and primary key**
  + 

# **Loading:**

* After we pulled in the CSV files and loaded them into the data frames, we did an initial connection to the Postgres database using PG admin to store our original clean data sets.
  + 
* Created a SQL script to create the table schema within Postgres database that generated the first set of tables.
  + 
* After created the new tables we reconnected to the database and generated additional tables for the data frames.
* **Postgres Database:**
  + 
  + **Mortality Data table:**
  + 
  + **Summary Data** that aggregates all the tables:
  + 
* **Bonus:** Create additional views
  + 
  + 
  + 