**Project Title:** Tap Water Lead Levels Virginia vs Toronto

**Team Members:** Ciara Spencer, Alex Goldstein, Pedro Aida

**Project Description / Outline:**

Our project aims to compare lead levels in tap drinking water from the city of Toronto and the state of Virginia for the years 2014 and 2015. Measurements of lead concentration in parts per billion(ppb) for each location will be compared in terms of variance and average lead concentration dichotomized month-to-month.

The final dataset will contain lead level measurements for each region grouped by average per month and variance per month for comparison.

**Datasets:**

1. [**https://www.kaggle.com/nikperi/toronto-lead-sampling**](https://www.kaggle.com/nikperi/toronto-lead-sampling)

**Toronto Lead Sampling** – This dataset contains water samples measuring lead amounts in ppb for the City of Toronto for the years 2014 through 2019. The collection was attained through the city of Toronto’s non-regulated program allowing residents, free-of-charge, to pick up and drop off water sample kits, providing water test results to all residents.

1. [**https://data.world/albert/virginia-water-lead-levels**](https://data.world/albert/virginia-water-lead-levels)

**Virginia Water Lead Levels** – This dataset contains water samples measuring lead levels in Micrograms per liter (ug/L) for the state of Virginia from 2013 through 2015. The collection was obtained from the Virginia DOH.

**Clean-up and Analysis:**

The two datasets differ in terms of table columns. Denominating factors between the two will be correlated with dates in common grouped by Month and Year. The two datasets are also of different sizes. Therefor the smaller dataset will be referenced for value counts to determine the number of randomly selected values from the larger dataset necessary to make an accurate comparison.

Calculations will be necessary to compare lead level measurements between the two datasets. The lead levels as determined in the Virginia Water Lead Levels will be converted from micrograms per liter (ug/L) to ppb to analogize the lead level measurements between the two.

**Rough Breakdown of Tasks:**

**Ciara -** Extraction of data from web sources.

**Alex –** Transformation of Data – cleaning, joining, filtering

**Pedro –** Loading of Data into final database

**Final Report**

**Extraction:**

The two datasets used to construct the final dataset were sourced from two data catalogs.

The Toronto Lead Sampling dataset was found on Kaggle and contains the lead sampling data for years 2014 through 2019. The original source of the data was acquired from the City of Toronto Open Data Portal which contains datasets for lead sampling from 2014 to present. This data is also accessible via an API through the portal which can be accessed via it’s corresponding package ID number.

Virginia Water Lead Levels was found on data.world and contains lead level data from January of 2013 to the end of December 2015 as acquired from the Virginia Department of Health. VDH hosts this data within a portal entitled Drinking Water Watch that contains the database for lead and copper level measurements as detected in Virginia Drinking Water from 1993 up until present day.

An attempt was made to extract data directly from the database by means of a selenium script that completed field entries to reach the database but once rendered issues occurred within the extraction of the table data due to the hosting of the data across a JavaScript server.

**Transform:**

**Part 1**

Following the extraction of the data from Kaggle and DataWorld, the next step was to transform it, such that it could be loaded in to a structured database.

The first step in this process was to size the data correctly. To avoid any potential confounding due to seasonality of collection, the data was segmented by month, with datasets for both localities being divided up by month for all 24 months in the 2-year sample period. Since the governments of Virginia and Toronto seldom communicate or collaborate, and certainly didn’t regarding their respective measurements of their communities’ drinking water lead levels, the individual dates for the two datasets do not align perfectly. However, since each of the 24 months in the datasets have a certain number of water quality readings, alignment between the two localities’ data by month was feasible.

Following the segmentation of the datasets by month, it was imperative to ensure the two datasets had an equivalent number of values, so that they could then be joined in a structured data frame. To ensure the number of observations was equal for both datasets, the total number of observations was calculated for each of the 24 months in the sample period. If a given month featured a greater number of readings in Toronto than in Virginia, then a random sample equal in size to the total number of observations in Virginia for that month was collected from the set of readings in Toronto. If such was true for Virginia, then the reverse would happen.

Finally, once the segmentation and appropriate sampling for each of the 24 months in the 2-year was complete, the new data was concatenated.

For each of the 24 months, either the total number of observations for Toronto was included in the concatenation (if there were fewer readings for Toronto than Virginia for that given month in that given year), or the randomsample was included (if Toronto had more readings than Virginia). Conversely, if Virginia had fewer readings than Toronto for a given month, then the total number of Virginia readings was included in the new Virginia dataset, whereas if Virginia had more readings, the random sample was included instead. Once these two new, properly transformed datasets were assembled via concatenation, they were written to separate CSV files, to be subsequently loaded into a structured dataframe.

**Part 2**

After the rendering of data from the two CSVs into equivalent forms for each dataset, the transformed CSVs were then read into corresponding tables for each. They were then each parsed down to contain only columns necessary to make comparisons between the two tables. Both data frames, for Virginia and Toronto, contain columns for sample date, locality of collection, and Lead amount in parts per billion (ppb).

**Load:**

A connection was then made to a locally hosted Postgres database entitled lead\_water. An engine was created to check for table names within the database that were created within a query for the two tables, ‘toronto\_lead’ and ‘virginia\_lead’. The data from the two previously created data frames were then appended into the SQL database to link with their corresponding columns of each table within the database. A confirmation of load was then performed by querying each table back into a Pandas data frame from the SQL database.