*December 2022*

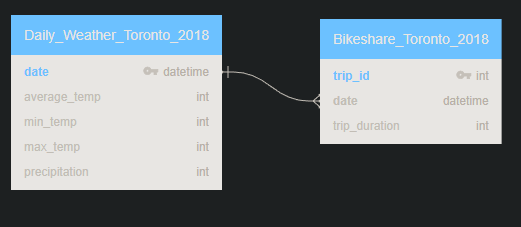
Project 2

## Group Name: Bikers

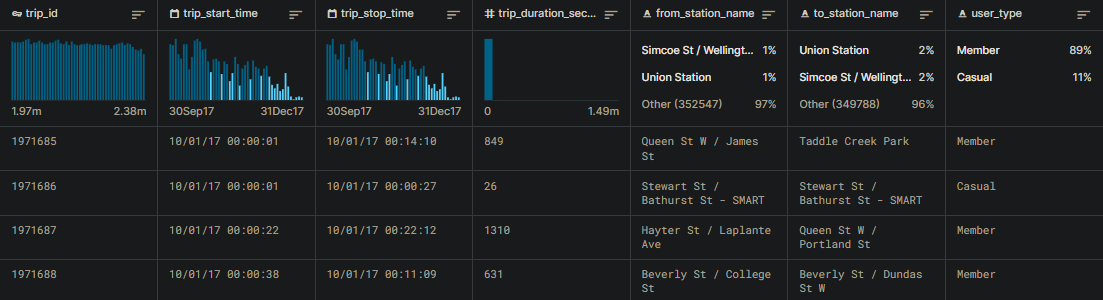
## Group Members: Bronwyn Milne, Oluwatobi Adelaja, Patrick Brennan

### Project Report:

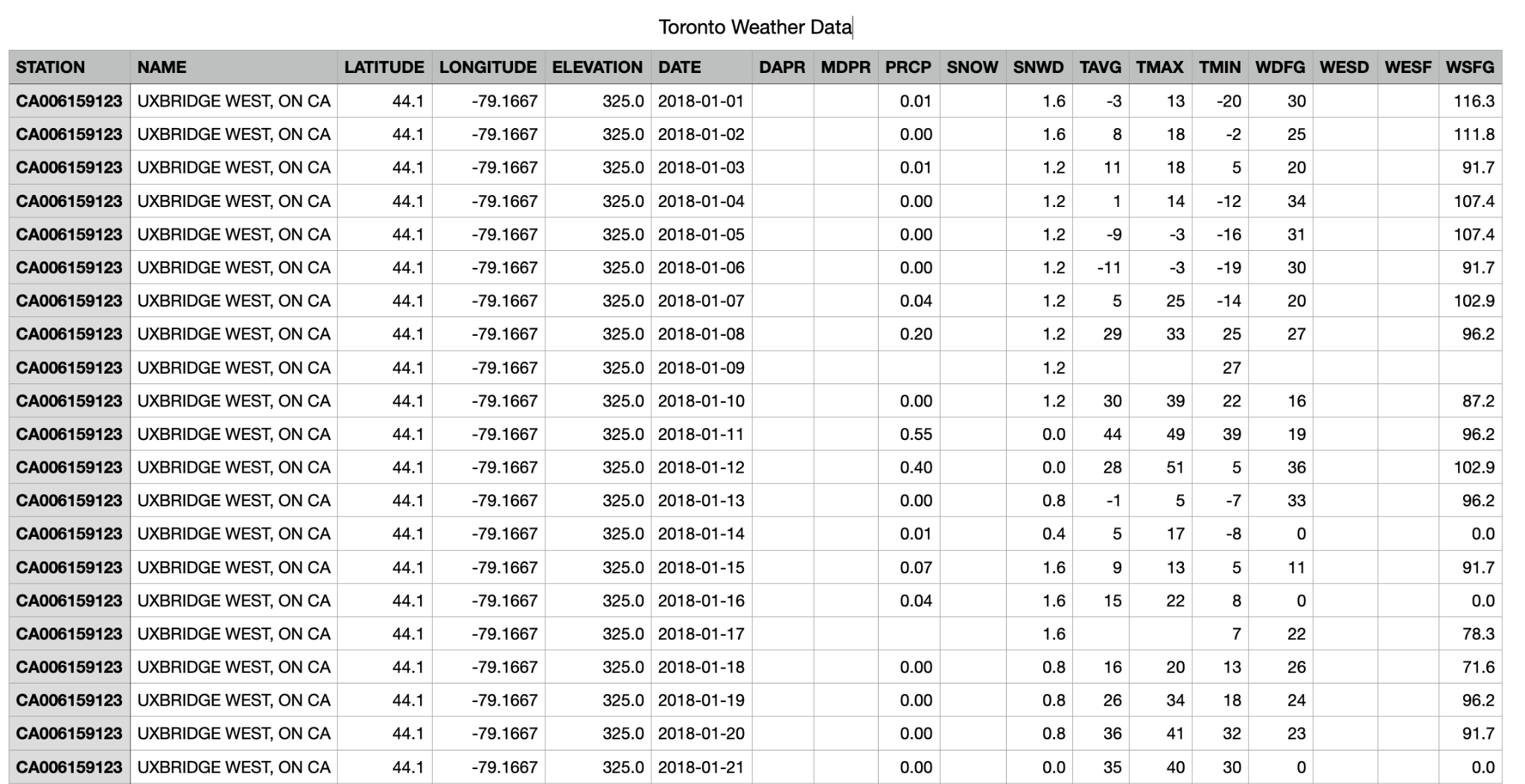
Our team will be joining a [dataset](https://www.kaggle.com/datasets/jackywang529/toronto-bikeshare-data) of bikeshare rides in Toronto from Kaggle in the year 2018 with the daily weather data for the same year pulled from the [NOAA](https://www.ncdc.noaa.gov/cdo-web/) weather database. The data will be stored in a relational database, relying on the date column to join the two datasets together, as seen below.



The bikeshare database includes a unique id number for each trip, the trip start time and end time, the total trip duration in seconds, the starting and ending bike stations, and the user’s account type. For the purposes of this project, we will keep the start time of the ride (stripping out the hour, minutes, and seconds) as our date of the ride as well as the unique trip id and the trip’s duration in seconds. The data was downloaded via CSV and uploaded into a pandas dataframe. The data was available for each quarter of the year; we used a concatenation method to merge all four CSVs into one dataframe.



The NOAA allows us to pull historical weather data for the Toronto area. By querying using the name of the city as well as the date range comprising all of 2018, we can pull a dataset in a CSV that gives us the average, minimum and maximum temperature for that day as well as various other weather measurements collected from all stations within a certain mile radius of the desired location.



We then loaded the CSV into a dataframe via pandas. By reviewing the dataset, we determined we wanted only the daily weather data for the station named ‘TORONTO CITY, ON CA’. After filtering for that station, we dropped columns such that we were only left with the station name, the date, the average, minimum and maximum temperatures, the precipitation, and the snowfall depth. Empty columns were filled with nulls to complete the dataset.

Once both tables were uploaded into pandas dataframes, we were able to upload them to our locally created database via sqlalchemy. We then were able to join the two datasets using the date (after converting each table’s ‘date’ column to datetime in pandas), creating a many to one connection between the bikeshare table and the weather table. Our final database can be seen below. Potential improvements include more accurate weather information based on the starting station and cross referencing the location of the various weather stations.

