Presentation - Group 2

Final Project - Commuter Transportation Preferences

Selected Topic

We chose to analyze vehicle, pedestrian and weather data collected in Ramsey County Minnesota from June 2015 to September 2018.

(Pedestrians = walkers + bikers)

Why we selected this topic

We are interested in how weather, day of the week, time of year and holidays affected commuters' preferred method of transportation.

Data Sources

Two datasets were used in our analysis.

The first dataset contained information on hourly traffic volume, hourly weather information, and holidays.

The second dataset contained information on pedestrian traffic (bikers and walkers) on specific days.

The original datasets can be found at the links below:

https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume#

https://www.dot.state.mn.us/bike-ped-counting/reports.html

(The following three slides further describe the contents of our datasets)

Data Sources (cont'd)

Vehicle Traffic Data

Our vehicle traffic dataset, compiled by the Minnesota Department of Transportation, contained hourly Westbound traffic volume for MN DoT ATR Station 301, roughly midway between Minneapolis and St. Paul, MN from October 2012 to September 2018.

The traffic dataset contained the date, time and the number of vehicles that passed by Station 301 every hour.

Data Sources (cont'd)

Weather & Holiday Data

The vehicle traffic data was combined with the corresponding weather data for each hour of each day.

- The weather data was obtained from OpenWeathmap, and contained information on the following: average temperature, rainfall, snowfall, cloudcover, and columns for textual descriptions of the weather.
- The holiday column contained information national holidays, plus a regional holiday (MN State Fair)

Data Sources (cont'd)

Pedestrian Data

The pedestrian data used also came from a MN Dept. of Transportation dataset.

This dataset contained information on walker and biker volume from all over the state of Minnesota from 2014 - 2020. It contained the number of pedestrians and bicyclists that went through checkpoints each day.

It also contained information we didn't need for our analysis, such as the technology used to collect the data, the type of path the walkers and bikers used, and some basic weather data.

Questions we hope to answer with the data

- How does commuter behavior change given the day of the week and time of year?
- How do weather conditions affect commuter behavior?
- Can we predict non-vehicle (bikers and walkers) traffic on a given day assuming vehicle traffic and weather conditions?

Data Exploration

Most of the data exploration and cleaning was done in SQL. We first dealt with duplicate datetime entries in the raw_vehicle_traffic table by dropping the textual description columns that did not provide any hard data, then averaging any disparate weather readings for a given datetime.

Data Output | Explain | Messages | Notifications | Scratch Pad

	holiday character varying (40)	temp real	rain_1h real	snow_1h real	clouds_all integer	weather_main character varying (15)	weather_description character varying (40)	date_time timestamp without time zone	traffic_volume integer	record_count bigint
1	None	274.79	0	0	90	Snow	snow	2013-04-18 22:00:00	1532	6
2	None	274.79	0	0	90	Snow	heavy snow	2013-04-18 22:00:00	1532	6
3	None	274.79	0	0	90	Rain	light rain	2013-04-18 22:00:00	1532	6
4	None	274.79	0	0	90	Rain	moderate rain	2013-04-18 22:00:00	1532	6
5	None	274.79	0	0	90	Drizzle	light intensity drizzle	2013-04-18 22:00:00	1532	6
6	None	274.79	0	0	90	Mist	mist	2013-04-18 22:00:00	1532	6
7	None	287.15	0	0	90	Mist	mist	2013-05-19 10:00:00	3591	6
8	None	287.15	0	0	90	Thunderstorm	thunderstorm with light rain	2013-05-19 10:00:00	3591	6
9	None	287.15	0	0	90	Thunderstorm	thunderstorm with heavy rain	2013-05-19 10:00:00	3591	6
10	None	287.15	0	0	90	Rain	moderate rain	2013-05-19 10:00:00	3591	6
11	None	287.15	0	0	90	Rain	light rain	2013-05-19 10:00:00	3591	6
12	None	287.15	0	0	90	Thunderstorm	proximity thunderstorm	2013-05-19 10:00:00	3591	6
13	None	285.8	0	0	90	Mist	mist	2012-10-25 15:00:00	5888	5
14	None	285.8	0	0	90	Drizzle	light intensity drizzle	2012-10-25 15:00:00	5888	5
15	None	285.8	0	0	90	Drizzle	drizzle	2012-10-25 15:00:00	5888	5
16	None	285.8	0	0	90	Rain	moderate rain	2012-10-25 15:00:00	5888	5
17	None	285.8	0	0	90	Rain	light rain	2012-10-25 15:00:00	5888	5
18	None	278.93	0	0	90	Drizzle	drizzle	2012-10-26 04:00:00	705	5
19	None	278.93	0	0	90	Rain	moderate rain	2012-10-26 04:00:00	705	5
20	None	278.93	0	0	90	Mist	mist	2012-10-26 04:00:00	705	5
21	None	278.93	0	0	90	Drizzle	light intensity drizzle	2012-10-26 04:00:00	705	5
22	None	278.93	0	0	90	Rain	light rain	2012-10-26 04:00:00	705	5

Data Exploration (cont'd)

We also found that not all hours for a given holiday date had the the holiday listed, so we had to assure that the holidays were correctly labeled when aggregating on date.

	holiday character varying (40)	date date
1	Columbus Day	2012-10-08
2	Veterans Day	2012-11-12
3	Thanksgiving Day	2012-11-22
4	Christmas Day	2012-12-25
5	New Years Day	2013-01-01
6	Washingtons Birthday	2013-02-18
7	Memorial Day	2013-05-27
8	Independence Day	2013-07-04
9	State Fair	2013-08-22
10	Labor Day	2013-09-02
11	Columbus Day	2013-10-14
12	Veterans Day	2013-11-11
13	Thanksgiving Day	2013-11-28
14	Christmas Day	2013-12-25
15	New Years Day	2014-01-01
16	Martin Luther King, Jr Day	2014-01-20

Data Output | Explain | Messages | Notifications | Scratch Pad

	a_holiday character varying (40)	a_date date	b_holiday character varying (40)	b_date date
1	Columbus Day	2012-10-08	None	2012-10-08
2	Columbus Day	2012-10-08	None	2012-10-08
3	Columbus Day	2012-10-08	None	2012-10-08
4	Columbus Day	2012-10-08	None	2012-10-08
5	Columbus Day	2012-10-08	None	2012-10-08
6	Columbus Day	2012-10-08	None	2012-10-08
7	Columbus Day	2012-10-08	None	2012-10-08
8	Columbus Day	2012-10-08	None	2012-10-08
9	Columbus Day	2012-10-08	None	2012-10-08
10	Columbus Day	2012-10-08	None	2012-10-08
11	Columbus Day	2012-10-08	None	2012-10-08
12	Columbus Day	2012-10-08	None	2012-10-08
13	Columbus Day	2012-10-08	None	2012-10-08

Data Exploration (cont'd)

In the raw_bike_pedestrian_traffic tabled, columns containing metadata such as type of bike path or technology used to track traffic were dropped. The dataset contained data for several counties in Minnesota, so we filtered to only Ramsey County where the vehicle traffic was recorded. The weather columns contained some null values, so we dropped those columns, since the vehicle traffic also contained weather data with no nulls. Each date in this table had an entry for pedestrian traffic and an entry for bike traffic, so we summed those values to get a total non-vehicle traffic value for each date.

Combining the Data Sets

To analyze the vehicle, pedestrian and weather data together, we combined our two datasets into a new dataset. This allowed us to have a unified set to analyze with our machine learning model.

The final combined dataset contained the following columns: daily pedestrian traffic count, average daily temperature, total daily rainfall, total daily snowfall, average daily cloud cover percentage, daily vehicle traffic count, date, month of year, day of week, and holiday.

Limitations of Our Data

The vehicle traffic dataset contained data on traffic and weather for every hour of the day. However, because our pedestrian dataset contained only information on the amount of bikers on a given day, we had to collapse our vehicle dataset into daily categories instead of hourly categories in order to join the two datasets together. This caused us to lose insights into more nuanced vehicle traffic data, such as how the time of day affected vehicle traffic.

Also, the vehicle traffic only collected data from one road in Ramsey County, whereas the pedestrian traffic data was collected from two different locations in Ramsey County. Although the data points were collected from two different locations, we believe it still conveys a useful depiction of pedestrian and vehicle traffic in the county.

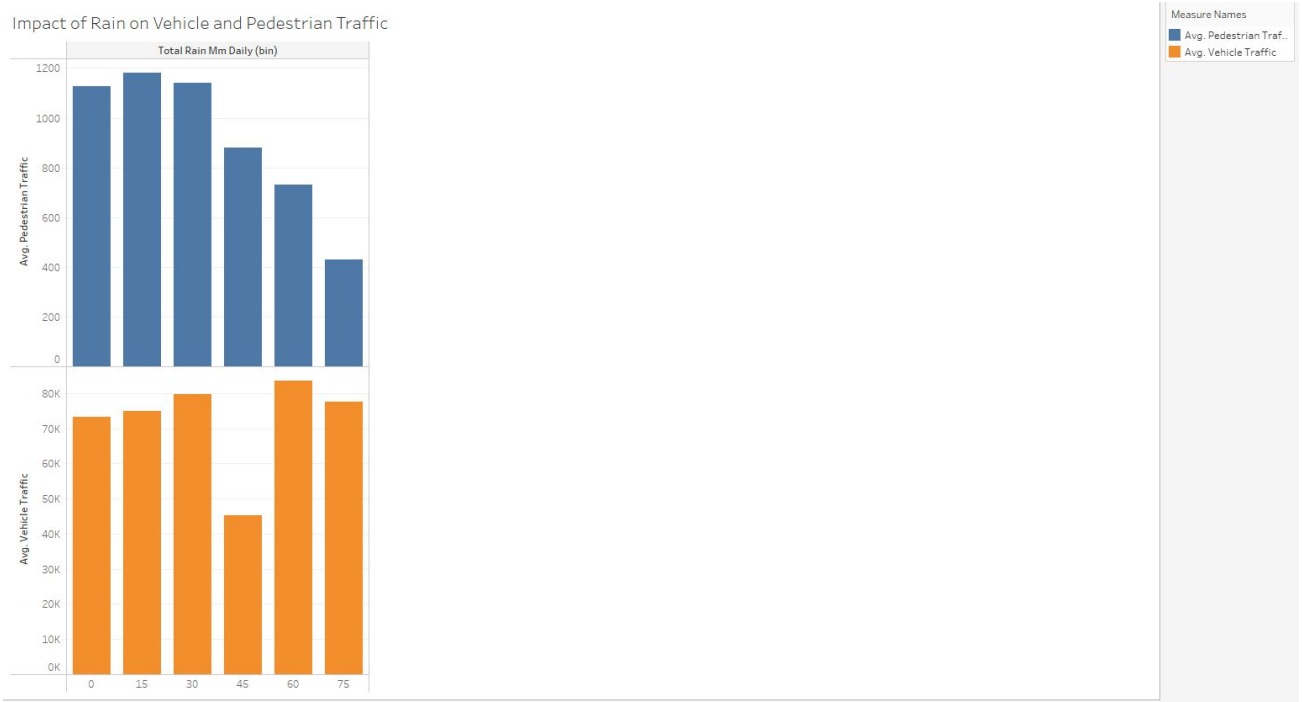
Analysis Phase

Once the datasets had been cleaned and joined our team was able to produce initial visualizations using Seaborn. A variety of visualizations including bar charts, line graphs, scatter plots, swarm graphs, and linear regression plots, allowed us to identify outliers and decide what inputs were crucial and which ones just created noise. As a second step new visualizations were created in Tableau to include in the final dashboard as well as make some of the visualizations interactive.

Link to Tableau Dashboard

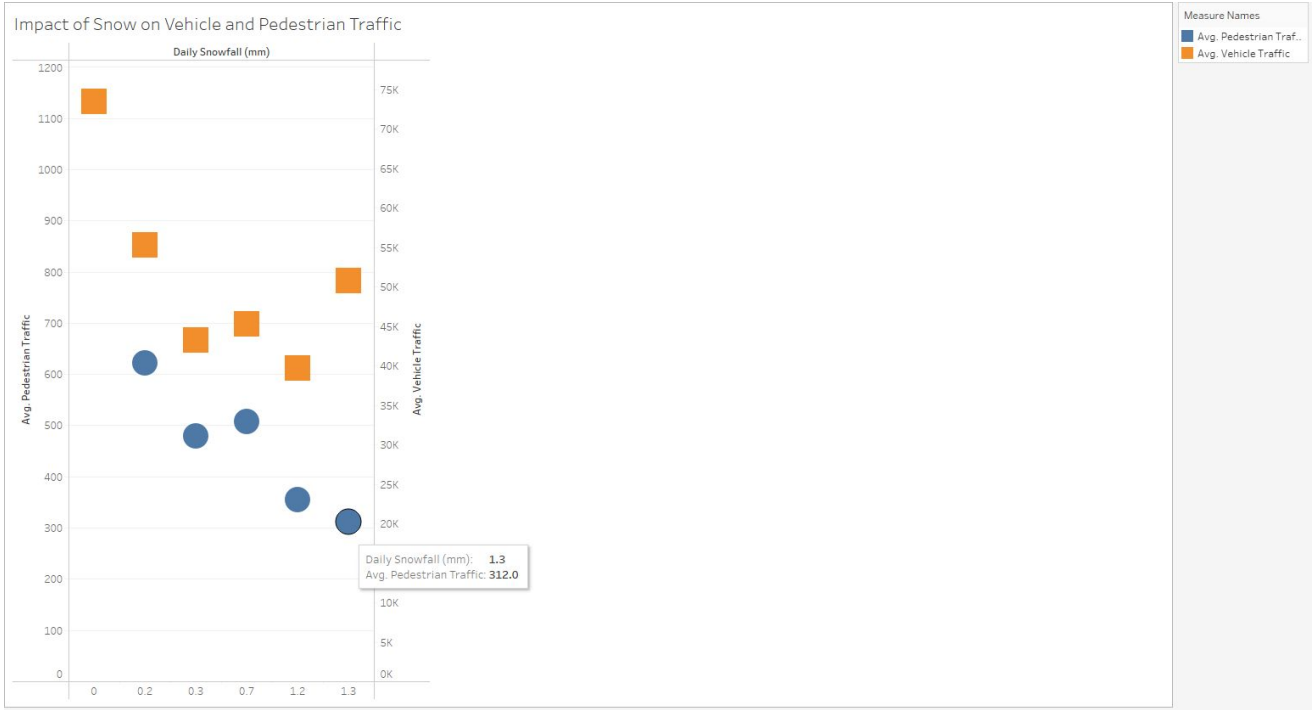
Analysis Phase (cont'd)

Rain



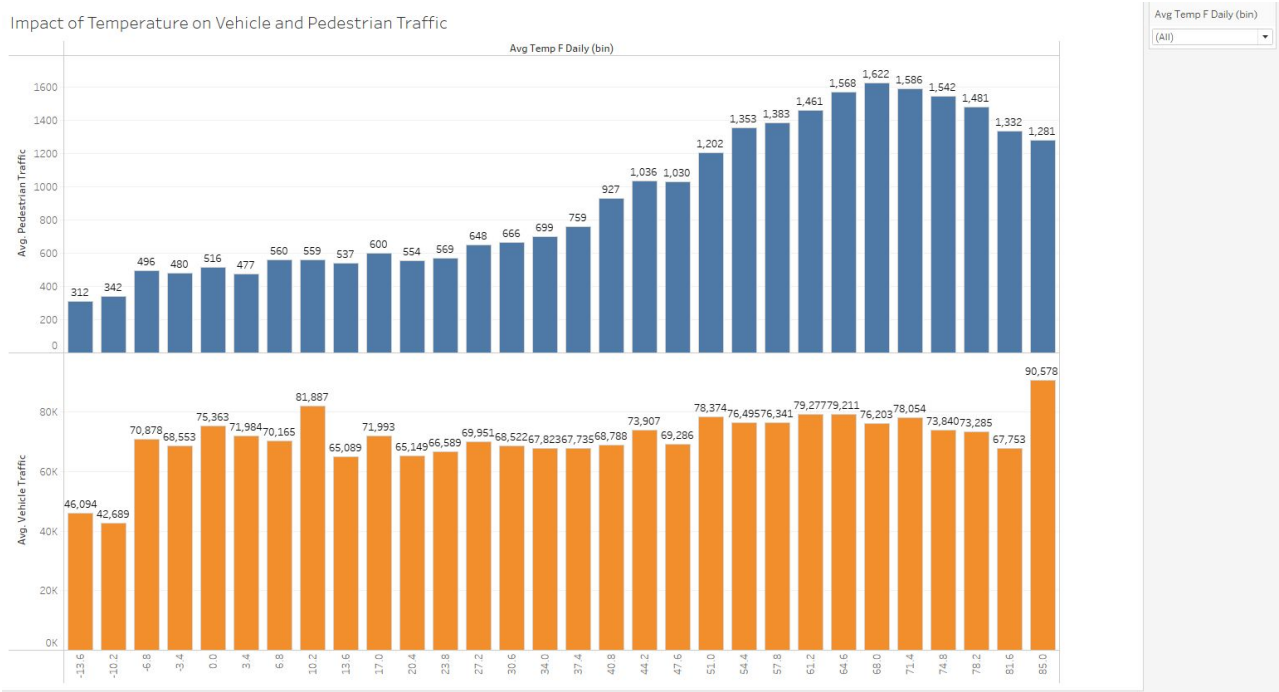
Analysis Phase (cont'd)

Snow



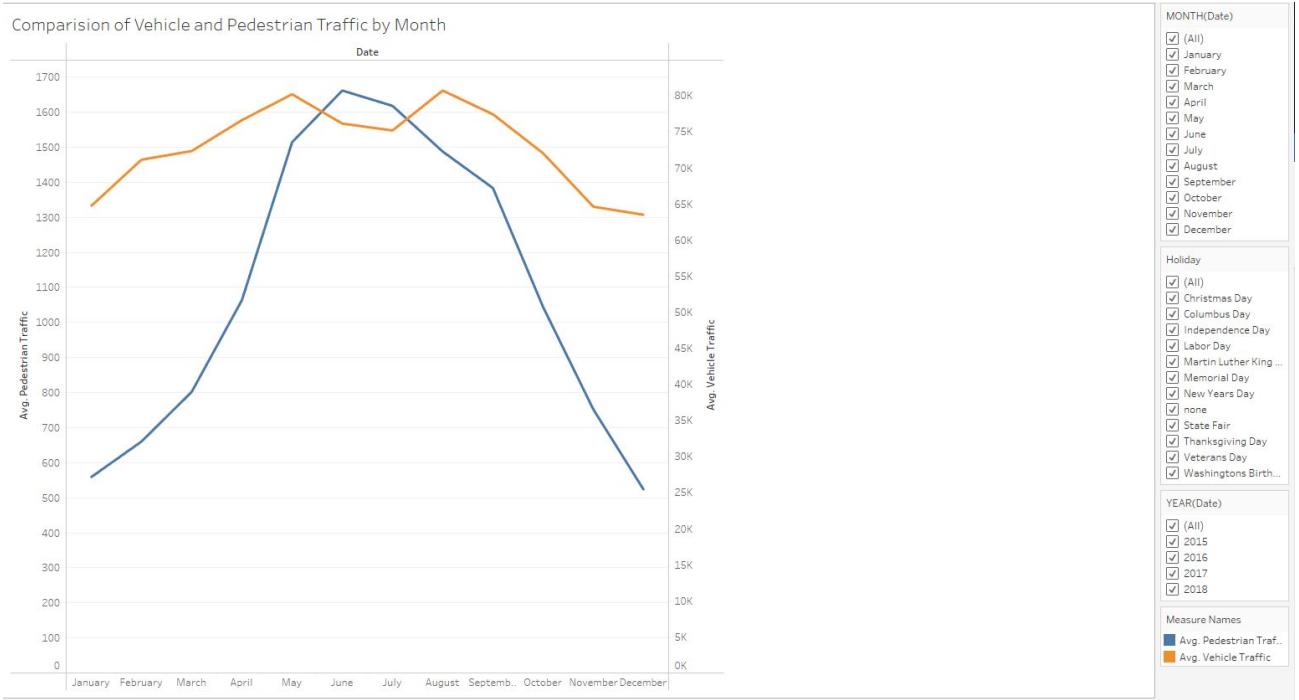
Analysis Phase (cont'd)

Temperature



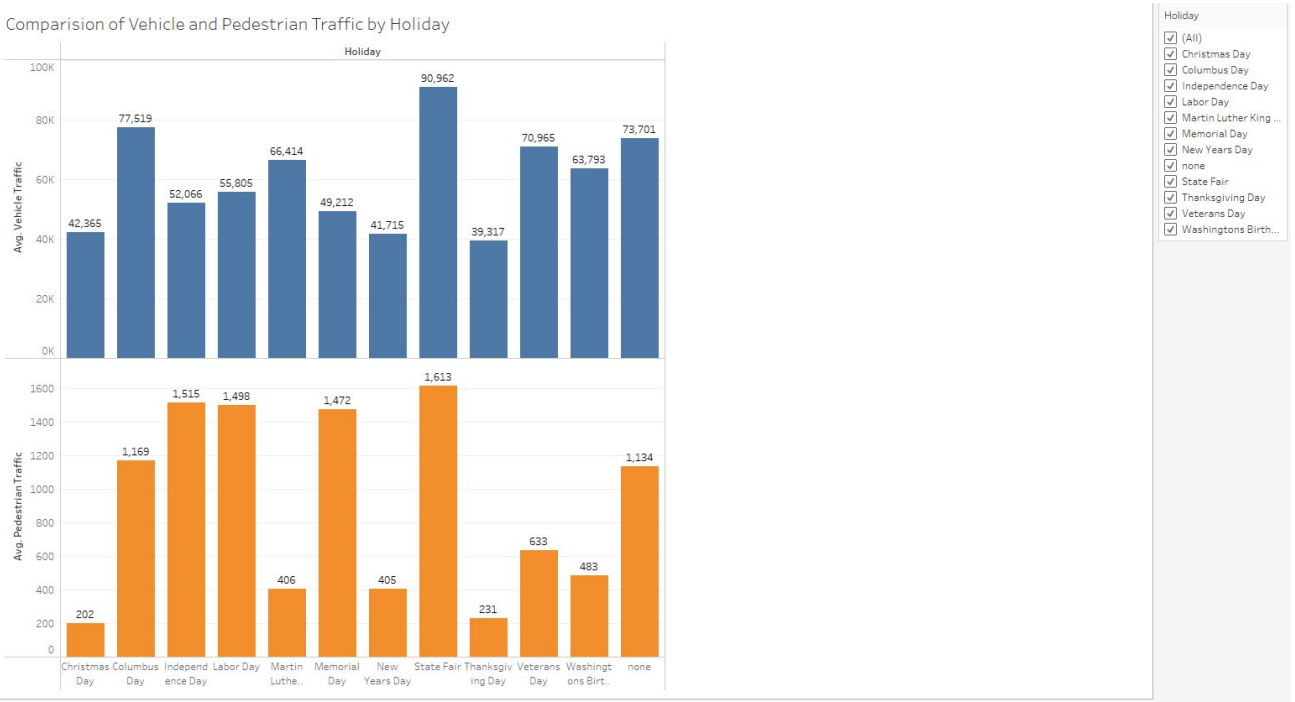
Analysis Phase (cont'd)

Month



Analysis Phase (cont'd)

Holiday

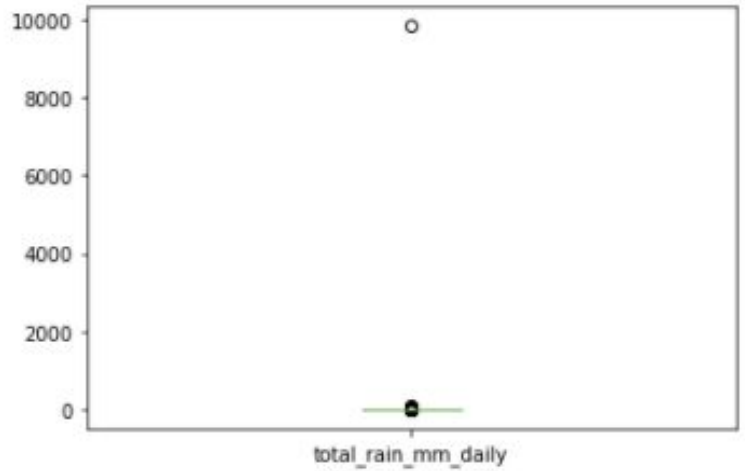


Analysis Phase (cont'd)

Outlier

```
# Use box-and-whisker plot to idenfity any outliers
mlearning_df["total_rain_mm_daily"].plot.box()
```

```
<AxesSubplot:>
```



Technologies, Languages, Tools and Algorithms

For the data exploration phase, we used SQL within pgAdmin to clean the data and join the datasets together.

For the analysis phase, we used Seaborn, Tableau, MatPlotLib and to visualize the data and derive insights.

To preprocess the data and use the machine learning model, we used Python within a Jupyter Notebook. The following dependencies were used: Pandas, pathlib, sklearn, psychopg2, sqlalchemy, configparser, matplotlib, and numpy. We used the LinearRegression model for our machine learning model.