# Group 2

Final Project - Commuter Transportation Preferences

# Presentation

**Introduction and Selecting Data Sources** - Ryan

**Data Exploration Phase** - Michael

**Analysis Phase** - Robert

**Machine Learning Model** - Joey

# Selected Topic

## Vehicle and Pedestrian Traffic Volume

Source: https://wtop.com/local/2017/05/many-people-really-bike-work-around-dc-surprising-stats/

# 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

## Vehicle Traffic Data


Source: https://www.dot.state.mn.us/

- I-94 Westbound Traffic, Ramsey County, MN
- .
- MN DoT ATR Station 301
- .
- Hourly traffic data 2012 - 2018

# Data Sources (cont'd)

## Weather & Holiday Data


Source: https://openweathermap.org/api

- OpenWeathermap
- .
- Temperature
- .
- Hourly traffic data 2012 - 2018

# Data Sources (cont'd)

## Pedestrian Data


Source: https://www.dot.state.mn.us/

- The pedestrian data used also came from a MN Dept. of Transportation dataset.
- .
- Walkers and Bikers in Ramsey County from 2014 - 2020
- .
- Lots of excess 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 pedestrian traffic on a given day assuming vehicle traffic and  weather conditions?

# Data Exploration - Vehicle Traffic dataset

Some preliminary data exploration done in excel, with most of the data cleaning done in postgreSQL.

- Found duplicated datetime values in vehicle table

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)

- Not all datetimes for holidays were correctly labelled.

| | a_holiday character varying (40) | a_date_time timestamp without time zone | a_date date | b_holiday character varying (40) | b_date_time timestamp without time zone | b_date date |
|---|---|---|---|---|---|---|
| 1 | Columbus Day | 2012-10-08 00:00:00 | 2012-10-08 | None | 2012-10-08 01:00:00 | 2012-10-08 |
| 2 | Columbus Day | 2012-10-08 00:00:00 | 2012-10-08 | None | 2012-10-08 02:00:00 | 2012-10-08 |
| 3 | Columbus Day | 2012-10-08 00:00:00 | 2012-10-08 | None | 2012-10-08 03:00:00 | 2012-10-08 |
| 4 | Columbus Day | 2012-10-08 00:00:00 | 2012-10-08 | None | 2012-10-08 04:00:00 | 2012-10-08 |
| 5 | Columbus Day | 2012-10-08 00:00:00 | 2012-10-08 | None | 2012-10-08 05:00:00 | 2012-10-08 |
| 6 | Columbus Day | 2012-10-08 00:00:00 | 2012-10-08 | None | 2012-10-08 06:00:00 | 2012-10-08 |
| 7 | Columbus Day | 2012-10-08 00:00:00 | 2012-10-08 | None | 2012-10-08 07:00:00 | 2012-10-08 |

| | 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 Exploration - Pedestrian Traffic Dataset

| site<br>character varying | county<br>character varying (30) | date_day<br>date | mode<br>character varying (10) | total<br>integer | record_count<br>bigint |
|---|---|---|---|---|---|
| Metro \| St Paul \| Jacks... | Ramsey | 2016-01-16 | Pedestrian | 114 | 1 |
| Metro \| St Paul \| Jacks... | Ramsey | 2016-01-17 | Bicyclist | 6 | 1 |
| Metro \| St Paul \| Summ... | Ramsey | 2016-01-17 | Bicyclist | 175 | 1 |
| Metro \| St Paul \| Jacks... | Ramsey | 2016-01-17 | Pedestrian | 113 | 1 |
| Metro \| St Paul \| Jacks... | Ramsey | 2016-01-18 | Bicyclist | 9 | 1 |
| Metro \| St Paul \| Summ... | Ramsey | 2016-01-18 | Bicyclist | 200 | 1 |
| Metro \| St Paul \| Jacks... | Ramsey | 2016-01-18 | Pedestrian | 174 | 1 |

# Combining the Data Sets

We joined the Vehicle, Pedestrian, and Holiday tables to create a dataset for analysis and for use in the machine learning model.

| daily_non_vehicle_traffic bigint | avg_temp_f_daily double precision | total_rain_mm_daily double precision | total_snow_mm_daily double precision | avg_cloud_percent_daily numeric | total_vehicle_volume_daily numeric | date date | month double precision | day_of_week double precision | holiday character varying |
|---|---|---|---|---|---|---|---|---|---|
| 837 | 57.9847973632813 | 0 | 0 | 92.0000000000000000 | 2886 | 2015-06-11 | 6 | 4 | none |
| 1266 | 69.3290051269532 | 0 | 0 | 5.3333333333333333 | 10774 | 2015-06-12 | 6 | 5 | none |
| 935 | 69.5092907714844 | 0 | 0 | 66.0000000000000000 | 5091 | 2015-06-13 | 6 | 6 | none |
| 1328 | 73.2451916503907 | 0 | 0 | 88.0000000000000000 | 2909 | 2015-06-14 | 6 | 0 | none |
| 1260 | 78.5425793457032 | 0 | 0 | 0.0000000000000000000 | 4681 | 2015-06-19 | 6 | 5 | none |
| 1128 | 71.9941784667969 | 0 | 0 | 88.0000000000000000 | 4045 | 2015-06-20 | 6 | 6 | none |
| 729 | 70.6621960449219 | 0 | 0 | 12.0000000000000000 | 5157 | 2015-06-22 | 6 | 1 | none |
| 1240 | 72.2597008514405 | 0 | 0 | 42.0000000000000000 | 66712 | 2015-06-24 | 6 | 3 | none |

# Analysis Phase

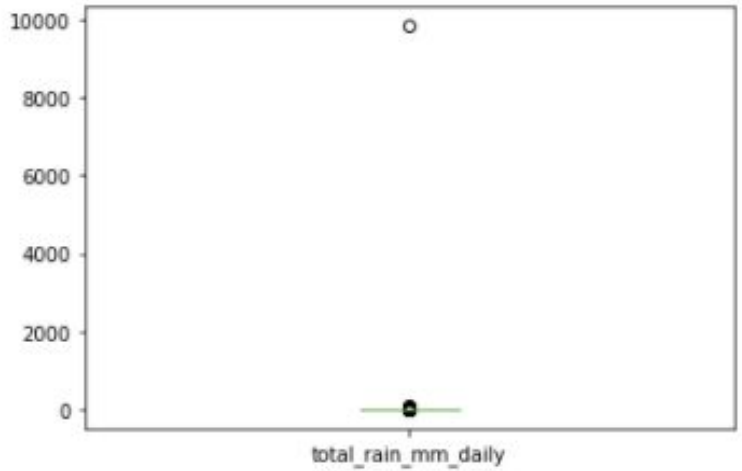


# Analysis Phase (cont'd)

## Outlier

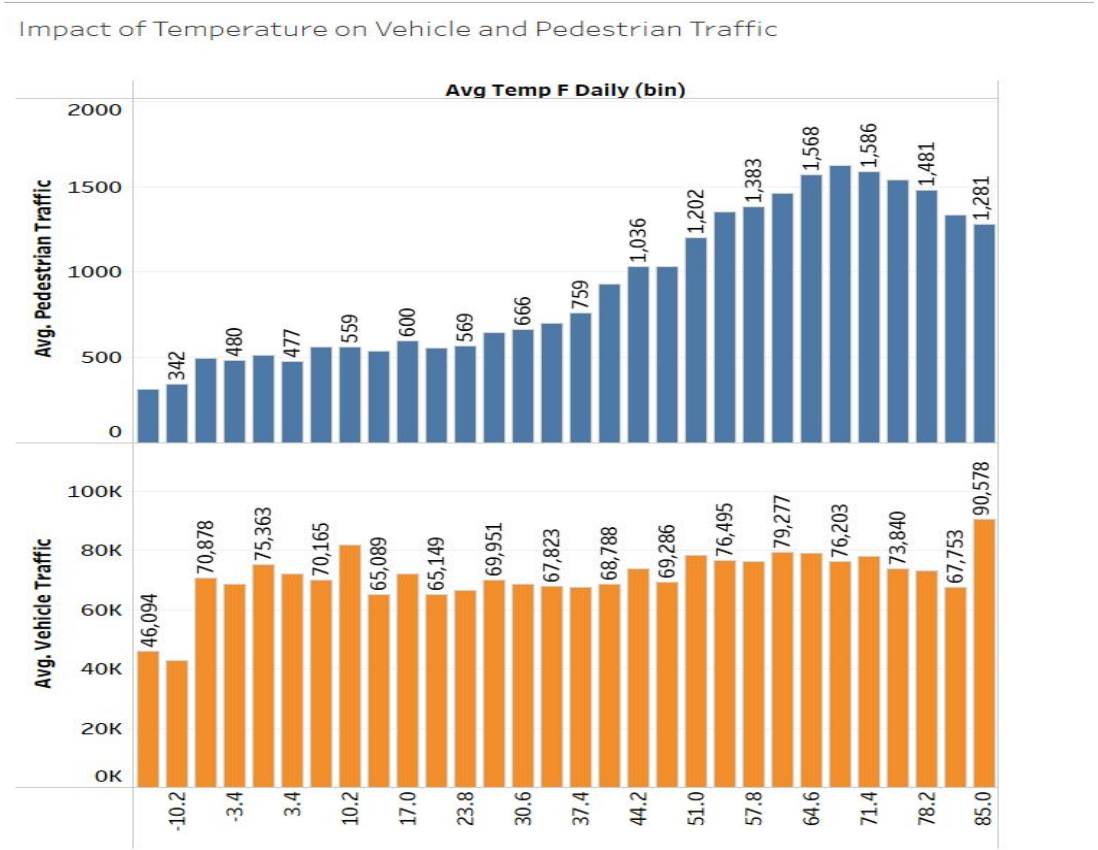
```
# Use box-and-whisker plot to idenfity any outliers
mlearning_df["total_rain_mm_daily"].plot.box()
```

```
<AxesSubplot:>
```



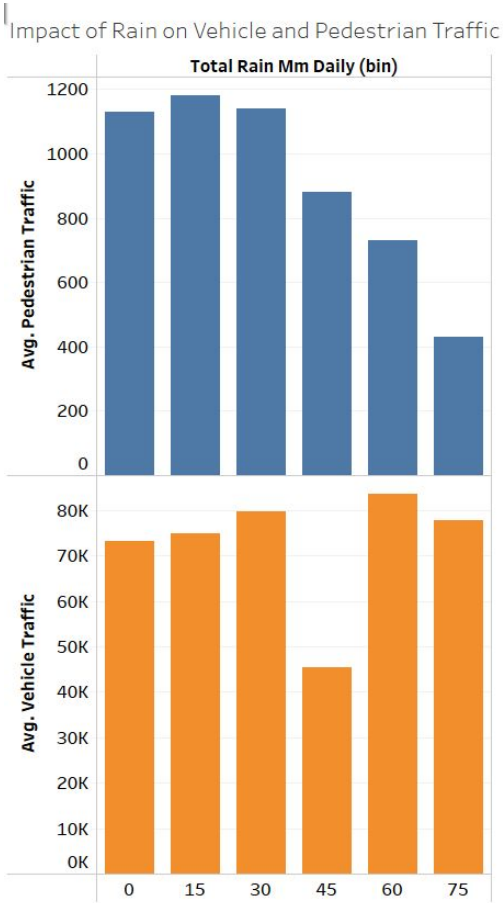
# Analysis Phase (cont'd)

## Temperature



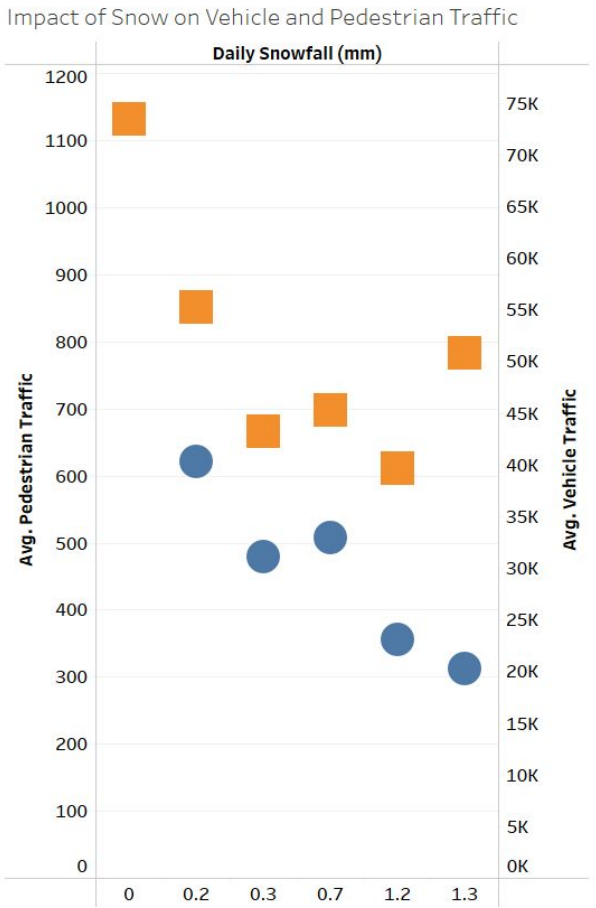
# Analysis Phase (cont'd)

## Rain



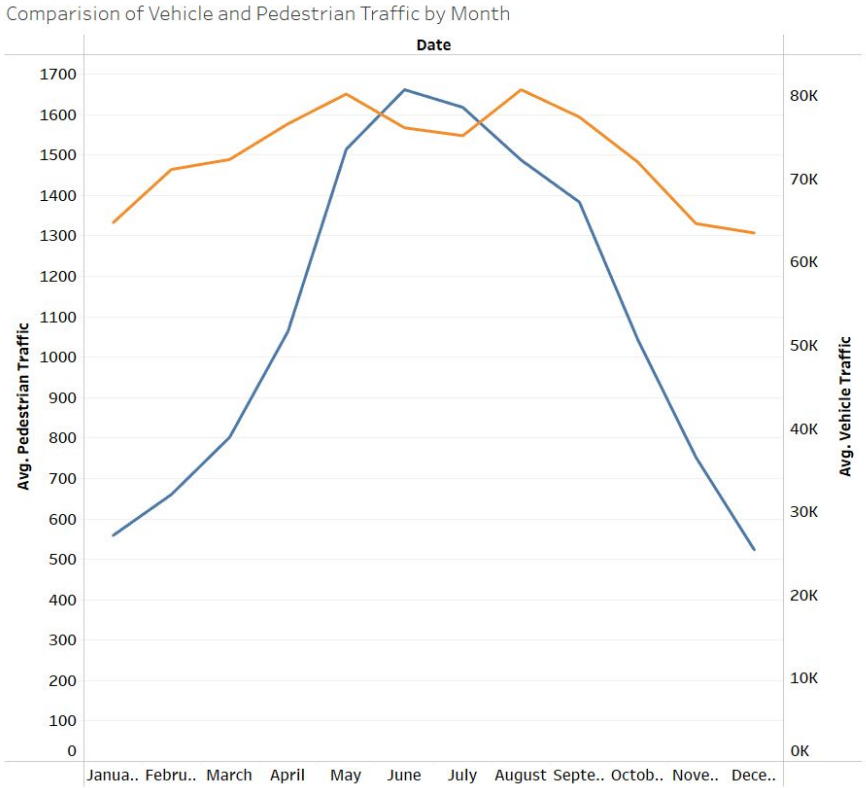
# Analysis Phase (cont'd)

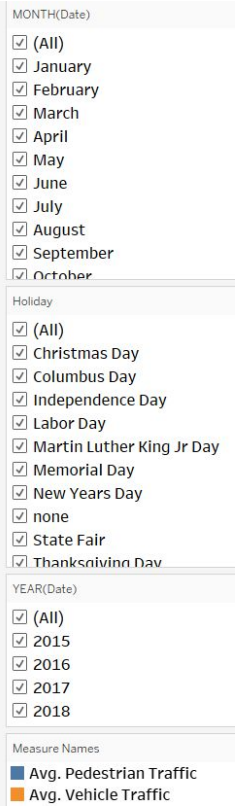
## Snow



# Analysis Phase (cont'd)

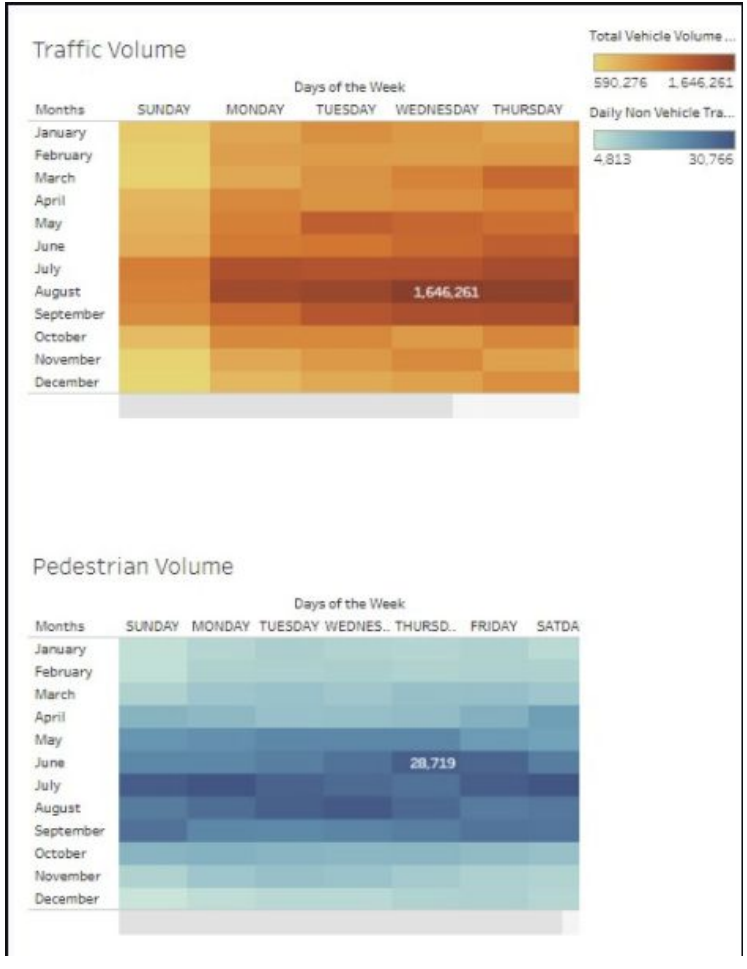
Month



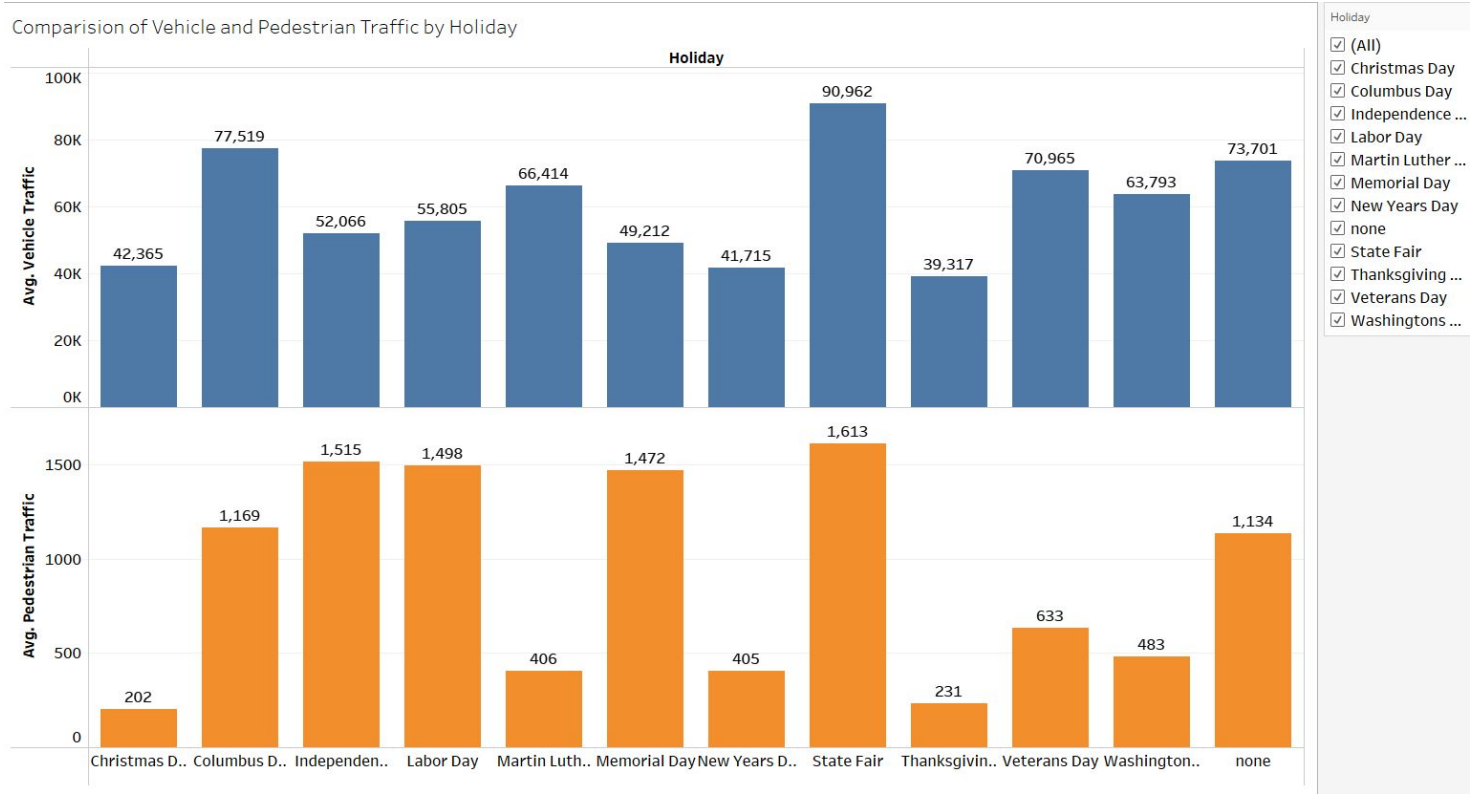

[Link to Tableau]

# Analysis Phase (cont'd)



# Analysis Phase (cont'd)

## Holiday



# Machine Learning Model

- Linear Regression
  - Understanding relationship between input and output variables
  - Input and output are numeric values

$$y= a_0+a_1x+ \varepsilon$$

**Here,**

Y= Dependent Variable (Target Variable)

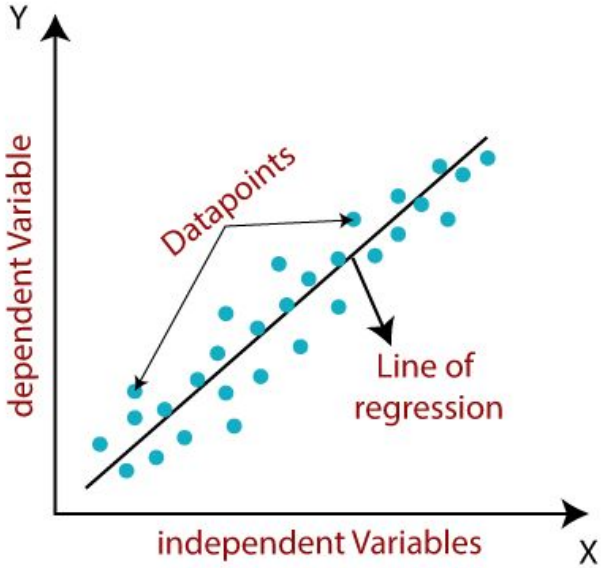
X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

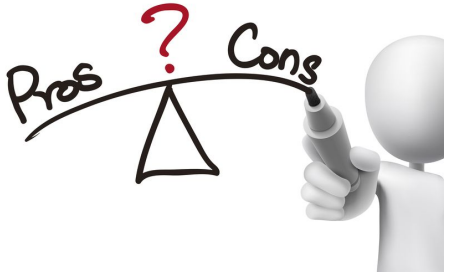
ε = random error

Source: https://www.javatpoint.com/linear-regression-in-machine-learning

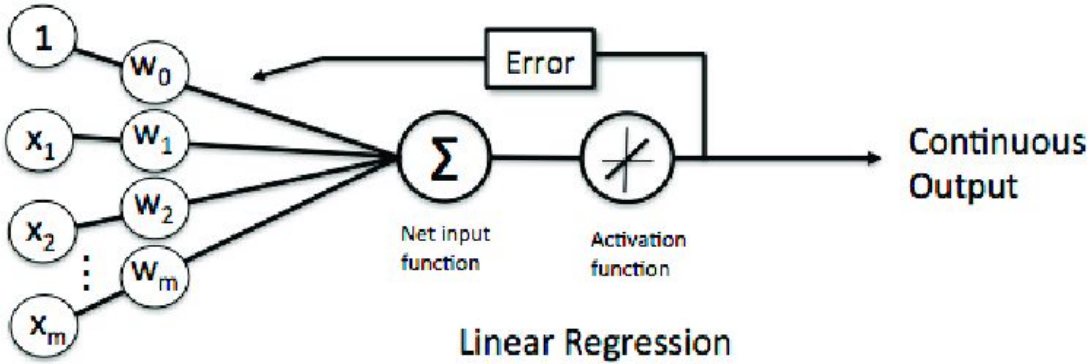



Source: https://www.javatpoint.com/linear-regression-in-machine-learning

# Machine Learning Model (cont'd)




- **Limitations**
  - Cannot predict discrete values
  - Sensitive to outliers
  - Prone to underfitting
- **Benefits**
  - Can predict continuous values
  - Linearity & simple implementation
  - Reduces overfitting by regularization



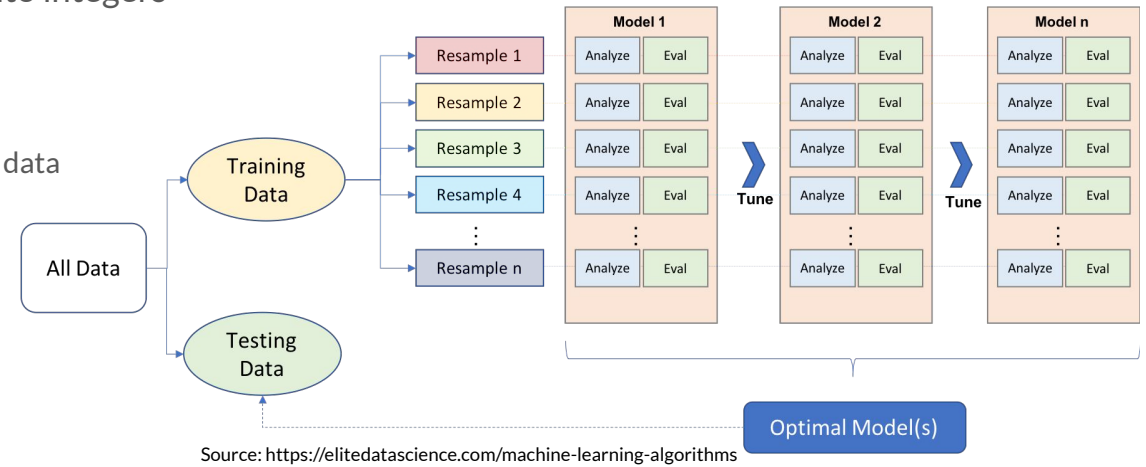

Source: https://www.researchgate.net/figure/Machine-learning-linear-regression_fig1_328948591

# Initial Machine Learning Model

- Make dataset more compatible with model
  - Dropped ‘Date’ column
  - Dropped extreme outlier
- Convert ‘Holiday’ string values into integers
  - Grouped by ‘holiday’ and ‘none’
  - Converted labels into integers
- Scale the Data
  - Compared scaled and unscaled data

**Initial Score**

```
Training Score: 0.6994128803300881
Testing Score: 0.6565277870104217
```



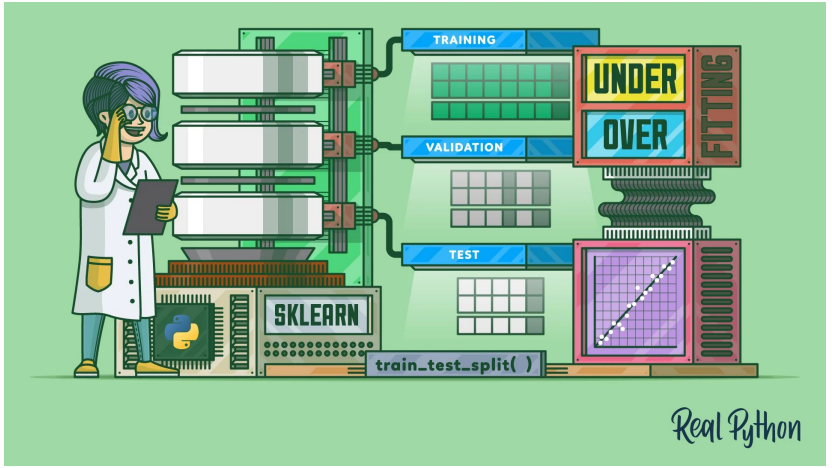

Source: https://elitedatascience.com/machine-learning-algorithms

# Optimized Machine Learning Model

- Further Feature Engineering and Selection Experimenting
  - Can model's accuracy be improved?
  - Train model on new features.
  - Keep feature change if accuracy improved.
- Optimized R-squared Score: 65.85%
  - Included all initial features
  - Left each holiday value in place
  - Grouped days into weekday or weekend

## Optimized Score

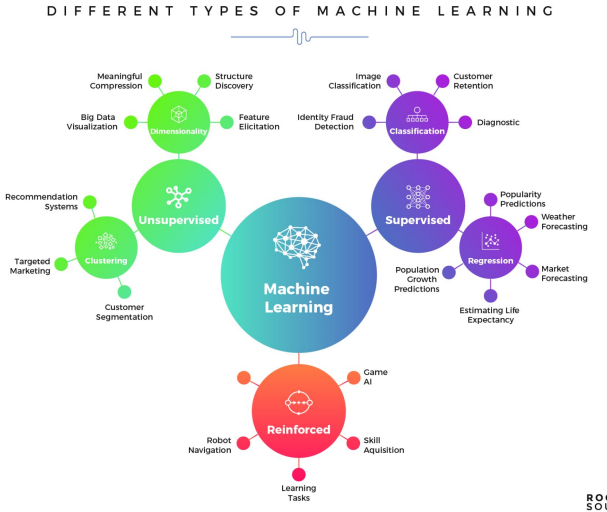
```
Training Score: 0.7007461735465028
Testing Score: 0.6585083431725288
```




Source: https://realpython.com/train-test-split-python-data/

# What we would change

- Utilize a different Machine Learning model
- Minimizing the reduction of data rows
    - Condensing 40,000+ rows of data to ~1,000
- Collect more data




ROCKET SOURCE

# Future Analysis

- Obtain additional records
  - From same dataset source
- Add additional combination features
  - New datasets
  - Comparing accessibility to public transportation
- Asking new questions
  - How does gas prices affect commuter habits?

Source: https://realpython.com/train-test-split-python-data/