# Bike Sharing Stations - San Francisco Bay Area

By: Douglas Malfacini

# Overview

- Introduction
- About the data
- Exploratory Analysis
- Machine Learning
    - Classification & Results

# Introduction

# Introduction

The San Francisco Bay Area is a popular tourist destination, as well as the heart of the technology sector in the United States. This fact brings in a lot of people to the Bay Area every day.

Bike sharing stations have recently come up in cities like San Francisco and San Jose as a quick and affordable way to get around in the Bay Area.

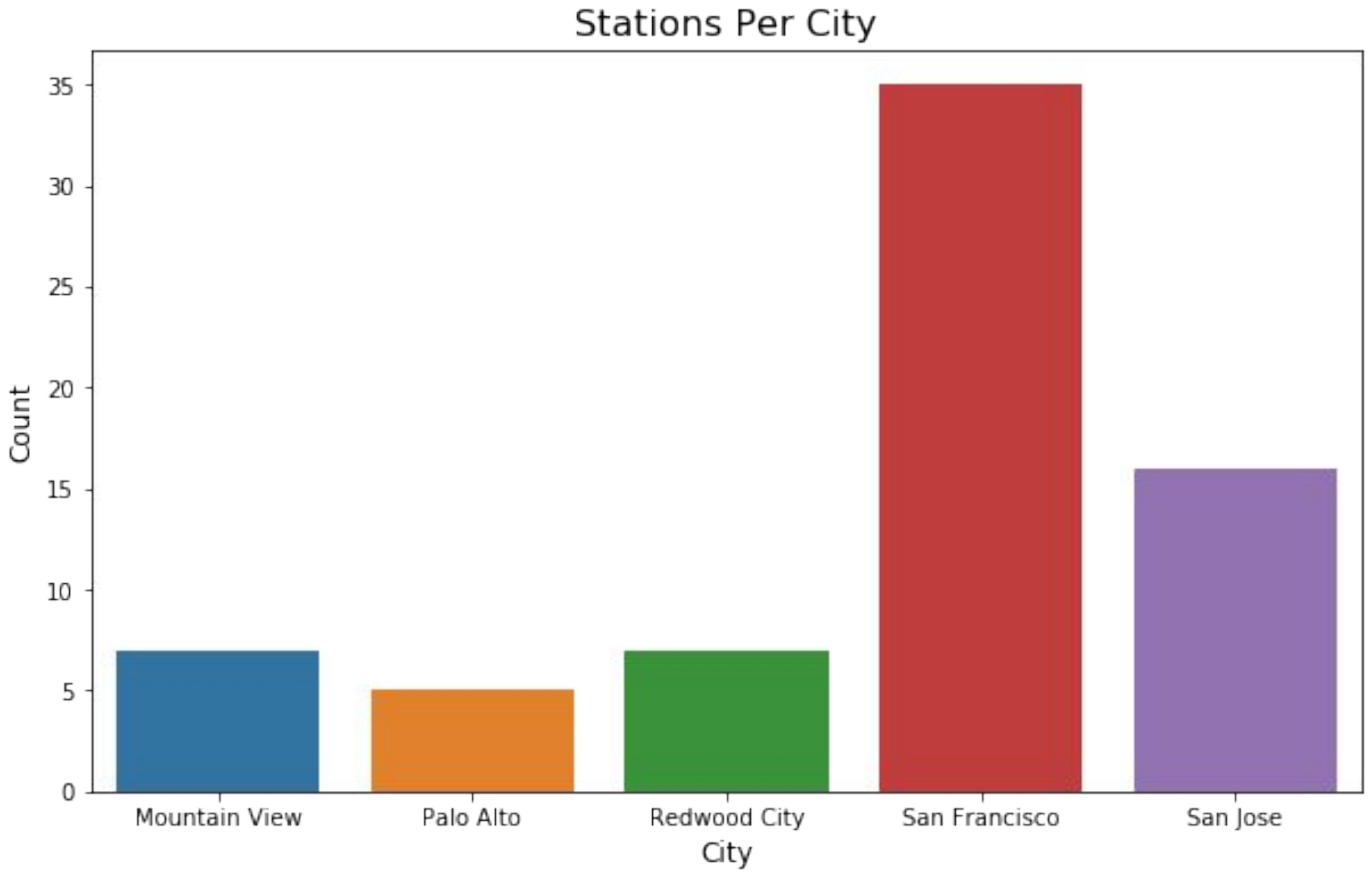
In this analysis, we'll explore who is using bike sharing stations, what times and how the weather impacts ridership. We'll also take a look at which stations are low on capacity (running out of bikes) and could pose a risk of running out of bikes for riders to use.

# About the Data

# About the data

- The data set has a some trips that have taken longer than 24 hours. Since we're only interested in daily trips, these trips have been filtered out.
- The data contains trips taken from Aug-2013 to Aug-2015.
- Five cities are included in this dataset.

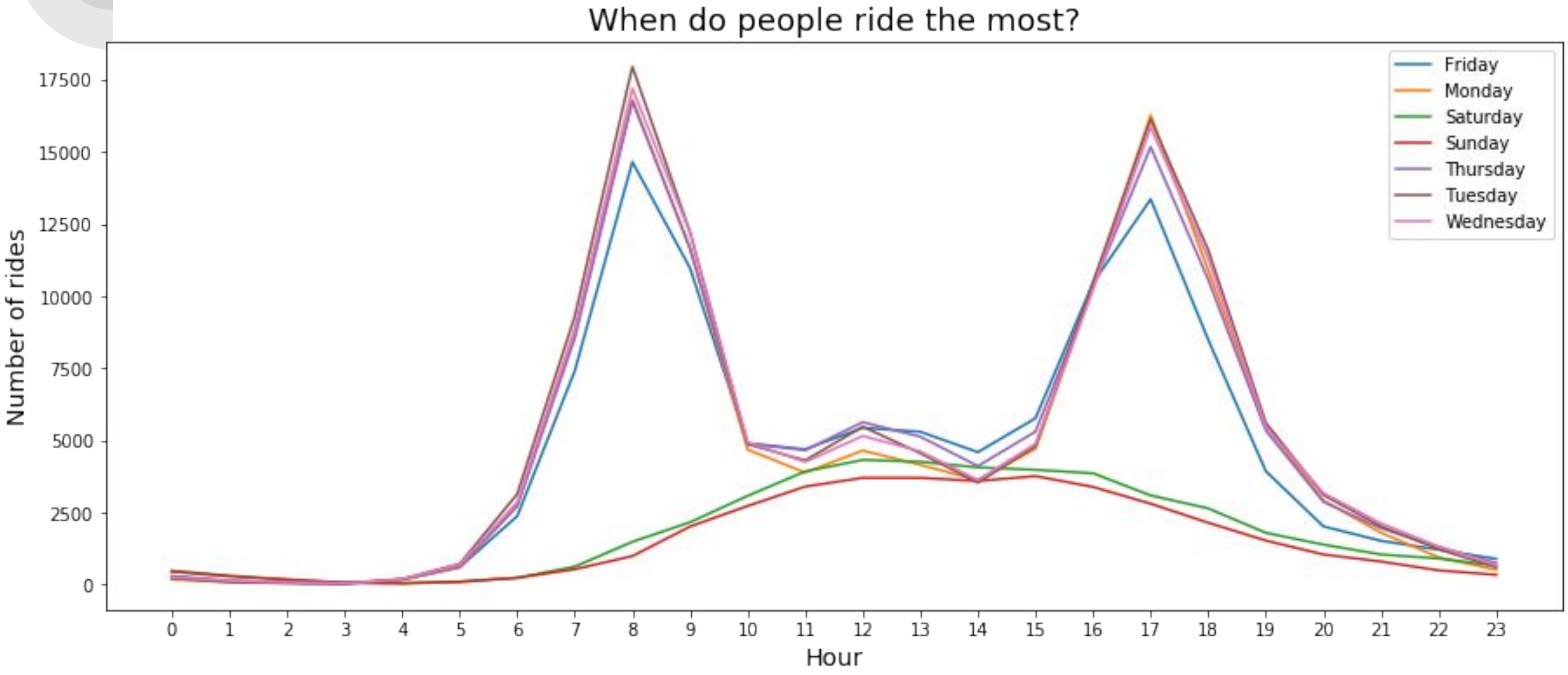
# Exploratory Analysis

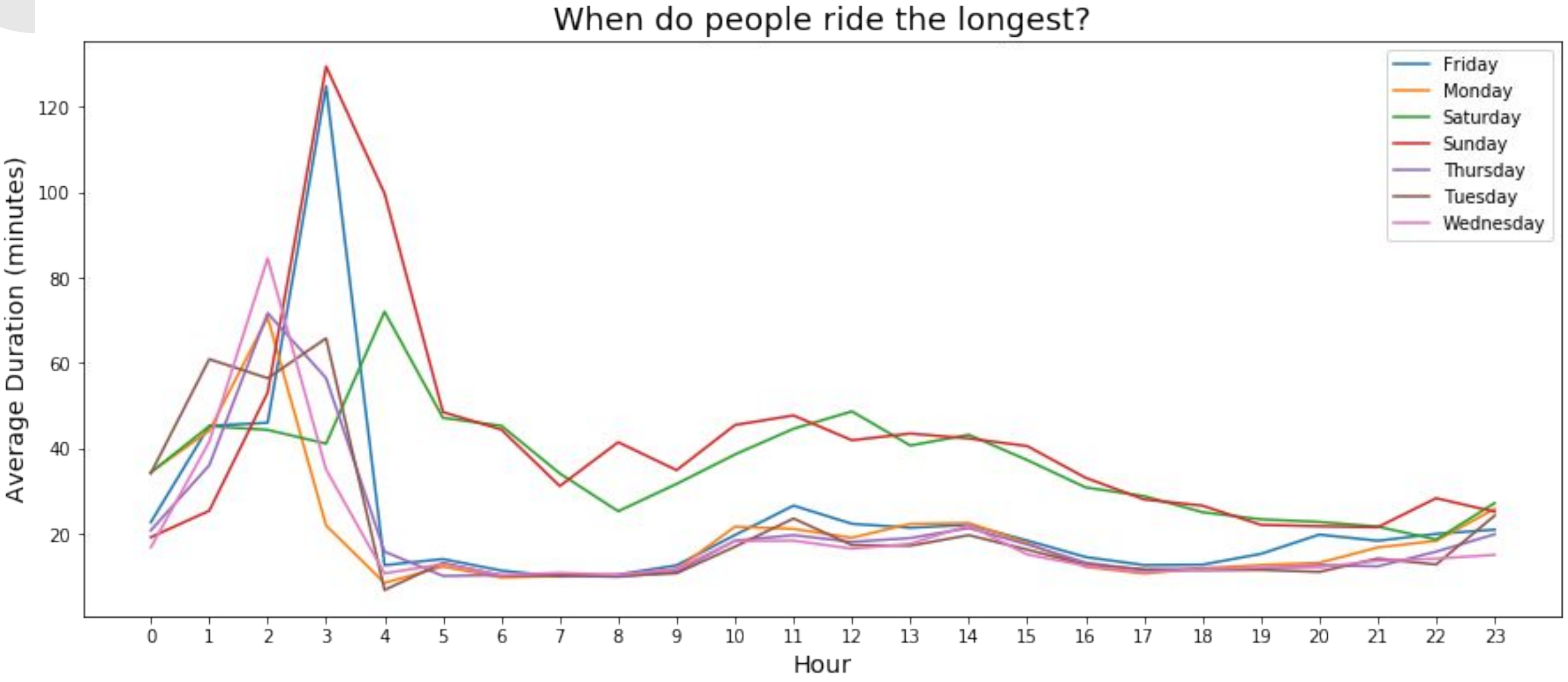

Stations Per City
Count
35
30
25
20
15
10
5
0
Mountain View
Palo Alto
Redwood City
San Francisco
San Jose
City

Bike Sharing Stations - San Francisco

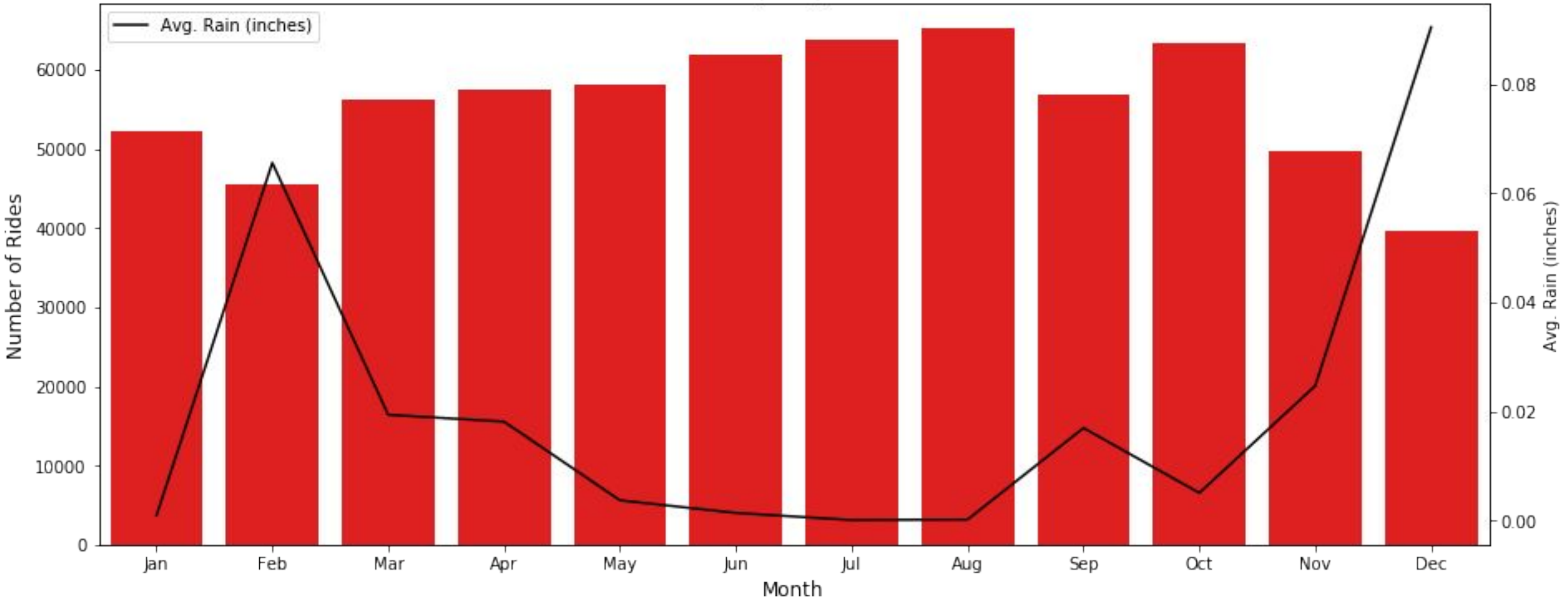
# Top ten routes

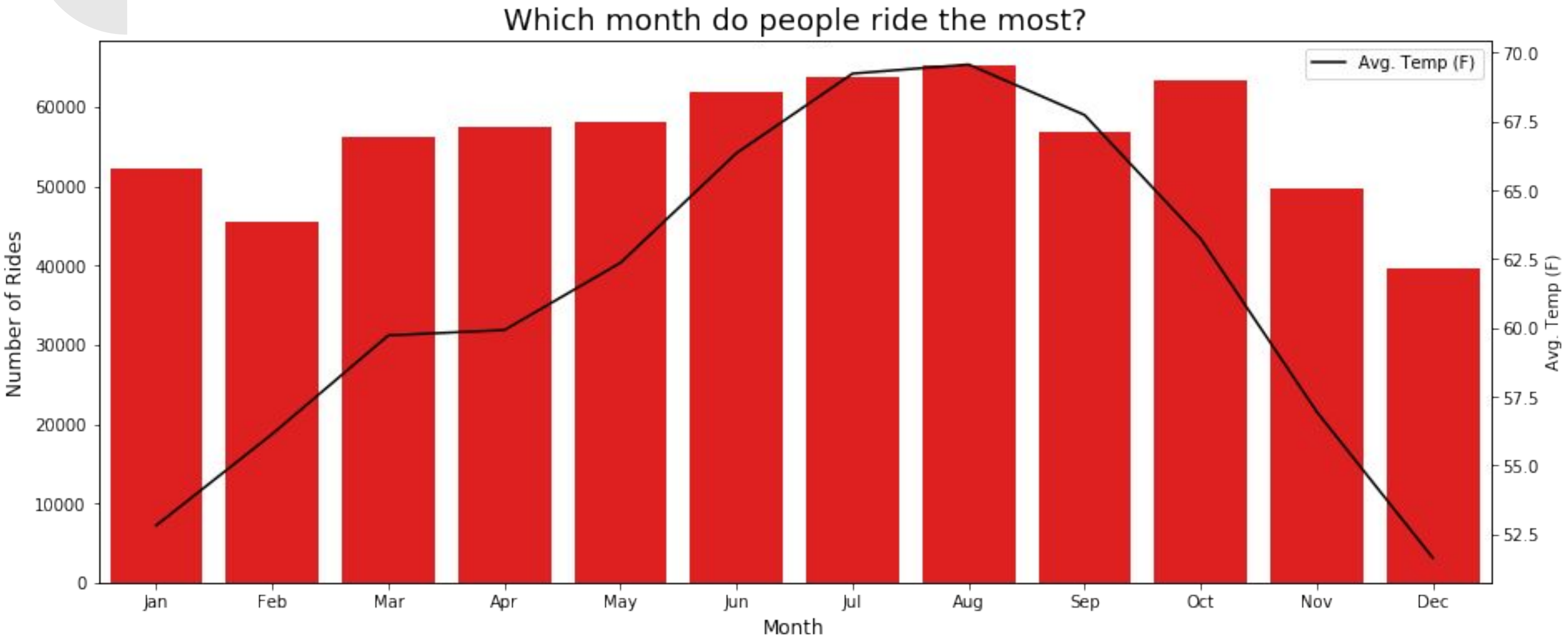
| | start_station_name | end_station_name | city | count |
|---|---|---|---|---|
| 1394 | San Francisco Caltrain 2 (330 Townsend) | Townsend at 7th | San Francisco | 6215 |
| 710 | Harry Bridges Plaza (Ferry Building) | Embarcadero at Sansome | San Francisco | 6164 |
| 1778 | Townsend at 7th | San Francisco Caltrain (Townsend at 4th) | San Francisco | 5041 |
| 90 | 2nd at Townsend | Harry Bridges Plaza (Ferry Building) | San Francisco | 4839 |
| 700 | Harry Bridges Plaza (Ferry Building) | 2nd at Townsend | San Francisco | 4357 |
| 562 | Embarcadero at Sansome | Steuart at Market | San Francisco | 4269 |
| 520 | Embarcadero at Folsom | San Francisco Caltrain (Townsend at 4th) | San Francisco | 3966 |
| 1680 | Steuart at Market | 2nd at Townsend | San Francisco | 3903 |
| 57 | 2nd at South Park | Market at Sansome | San Francisco | 3627 |
| 1338 | San Francisco Caltrain (Townsend at 4th) | Harry Bridges Plaza (Ferry Building) | San Francisco | 3622 |


When do people ride the most?
Number of rides
Hour
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
17500
15000
12500
10000
7500
5000
2500
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23


When do people ride the longest?
Average Duration (minutes)
Hour
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
20 40 60 80 100 120

# Which month do people ride the most?




Which month do people ride the most?
Avg. Temp (F)
Number of Rides
0
10000
20000
30000
40000
50000
60000
Avg. Temp (F)
52.5
55.0
57.5
60.0
62.5
65.0
67.5
70.0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

# Take away

As we can see from graphs, there is a clear distinction between when people ride the bikes vs. how long people are riding for.

Trips on workdays (Mon-Fri), during traditional commuting hours (7-9am and 4-6pm), are much shorter and more frequent when compared to trips taken on the weekends and outside of commuting hours.

In addition, the top 10 routes either starts, or ends around public transit stations like Caltrain and the Ferry Building. This strongly suggests the bikes are mostly being used by commuters in the bay.

# Classification

# Predictive Model

Based on the information discovered under the EDA steps, I created a model to predict if the riders are either Subscribers (people who pay a monthly fee to use the bikes) or Customers (pay as you go customers).

The model labels are skewed towards Subscribers with a 21:4 ratio. This puts the model baseline at 85% as that is the percentage of Subscribers in the tager variables.

# Predictive Model

I tried four different models initially to measure which one performed the best:

- LinearSVC
- Logistic Regression
- Naive Bayes
- Random Forest

I also ran the model with the following features:

- Ride Duration
- Start_station_id
- End_station_id
- Mean_temperature
- Mean Wind Speed
- Precipitation

# Predictive Model

Random Forest performed the best out of the four models. Here are the results:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Customer | 0.74 | 0.62 | 0.67 | 25736 |
| Subscriber | 0.93 | 0.96 | 0.95 | 141680 |
| avg / total | 0.90 | 0.91 | 0.90 | 167416 |

```
Feature Importance
================================
0.55  duration
0.13  start_station_id
0.13  end_station_id
0.10  mean_temperature_f
0.09  mean_wind_speed_mph
0.01  precipitation_inches
```

# Predictive Model

The model performed well, however, there were some features that had a less than 10% impact on the model performance:

- Wind Speed
- Precipitation

I dropped these two features from the model and added another feature:

- Distance/Duration Ratio

# Predictive Model

It is impossible to know which route a rider took to get between two stations, however, we can find out the distance between two stations “as the crow flies”. Once I found the distance, I was able to get a ratio. The intent is that riders who are Subscribers are getting from point A to point B as quickly as they can (they are commuters) while Customers will take a longer time to ride the same route

Here are the results:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Customer | 0.69 | 0.62 | 0.65 | 25736 |
| Subscriber | 0.93 | 0.95 | 0.94 | 141680 |
| avg / total | 0.90 | 0.90 | 0.90 | 167416 |

| Feature | Importance |
|---|---|
| 0.32 | duration |
| 0.40 | distance_duration_ratio |
| 0.08 | start_station_id |
| 0.08 | end_station_id |
| 0.12 | mean_temperature_f |

# Predictive Model

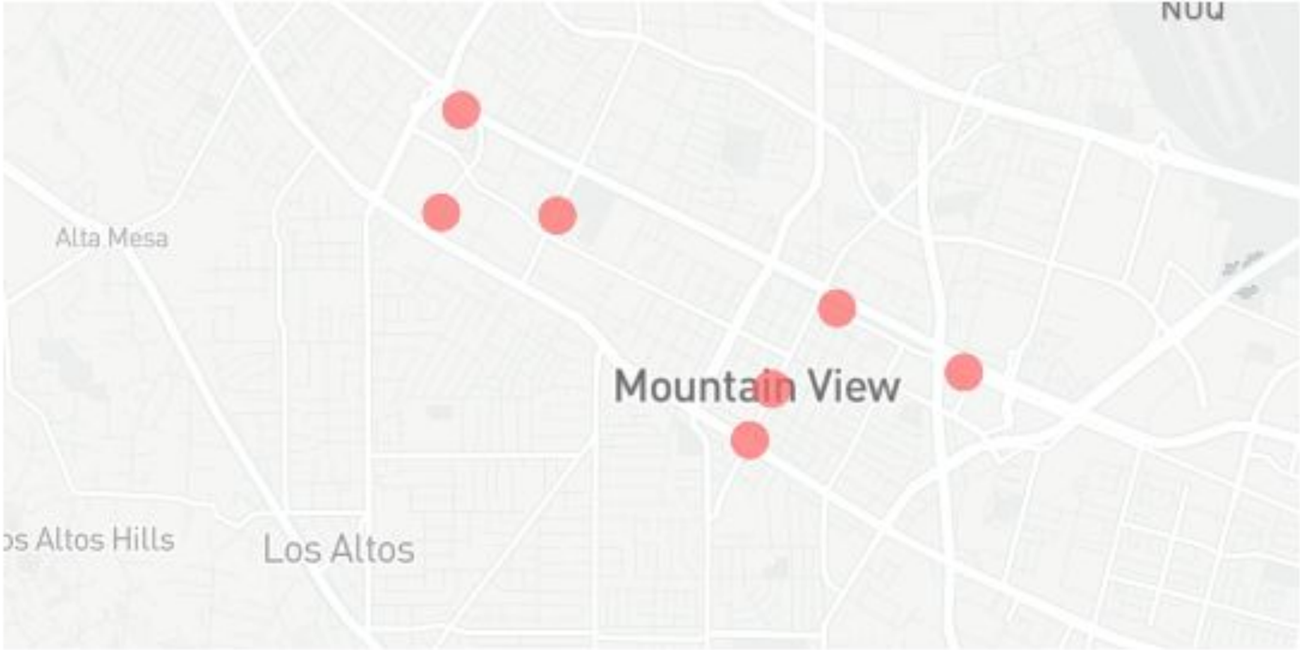
We got mixed results from the addition of the Distance to Duration Ratio. Model performance is similar, however, we got lower precision, recall, and f1-score for customers. The interest part is that the distance_duration_ratio actually turned out to be a higher importance than just the duration, which was over 50% importance in the first model.

For future iterations, it would be interested to add weather data back in and check the results.

# Appendix

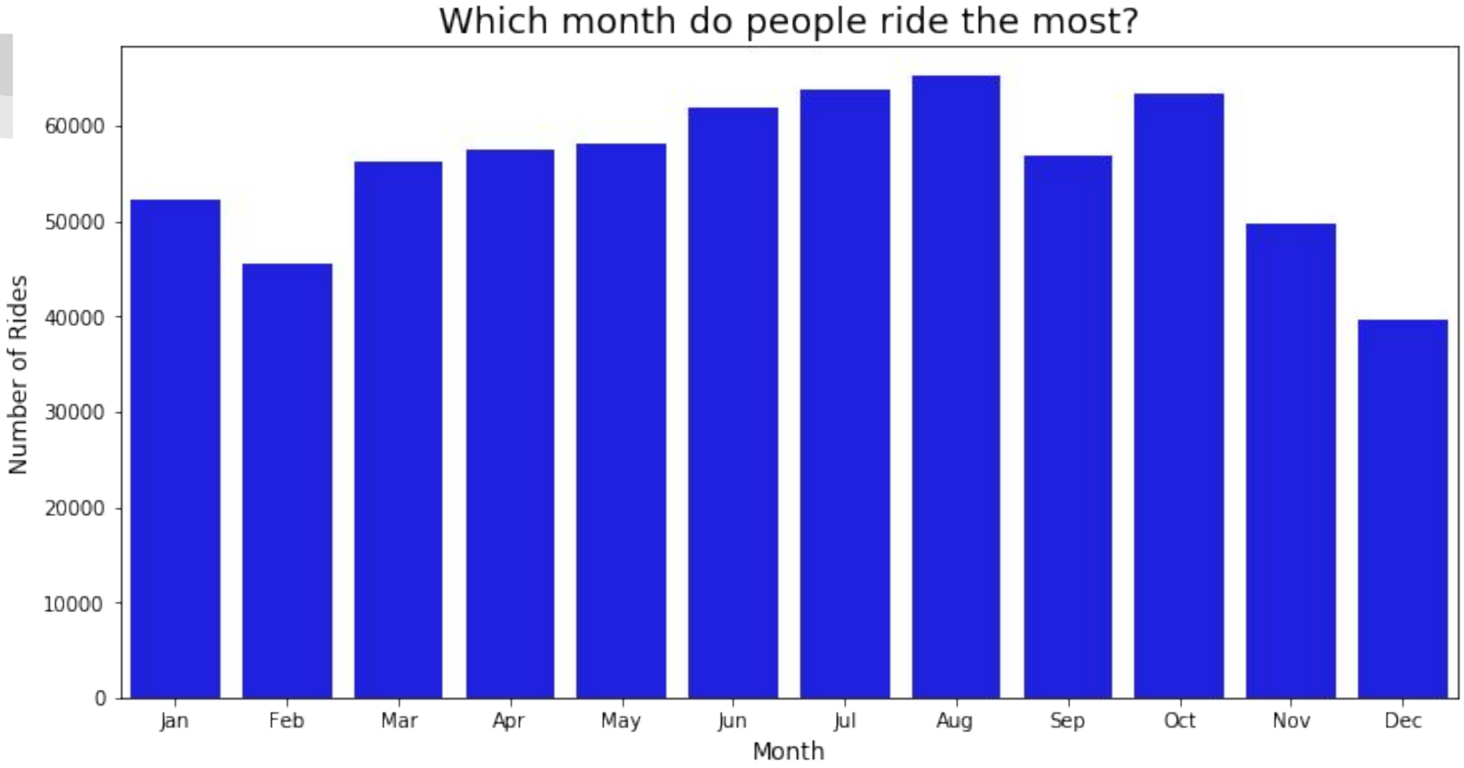
# Bike Sharing Stations - Redwood City

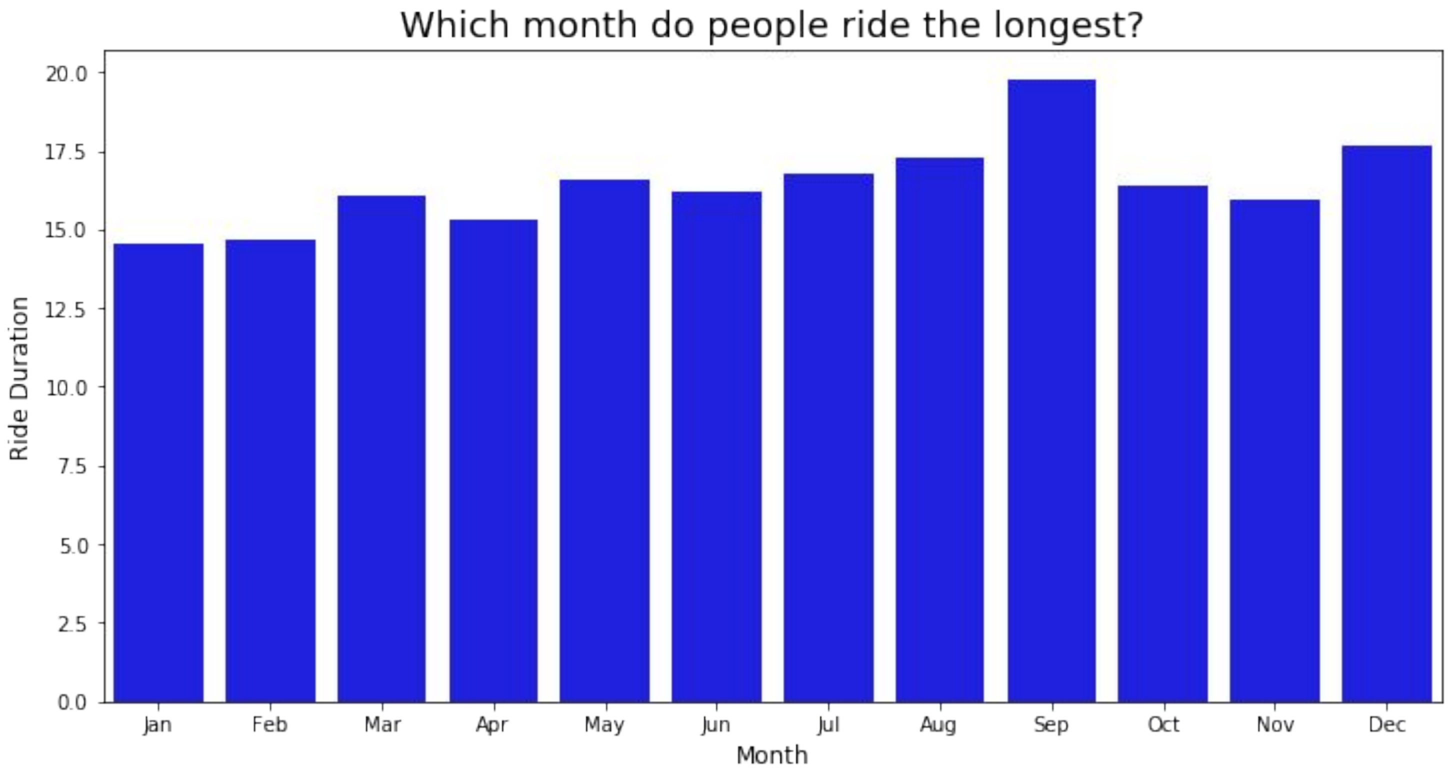
Bike Sharing Stations - Mountain View



# Bike Sharing Stations - Palo Alto

Bike Sharing Stations - San Jose


Which month do people ride the most?
Number of Rides
0
10000
20000
30000
40000
50000
60000
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month


Which month do people ride the longest?
Ride Duration
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month