Bay Area Bike Sharing Program - Who uses the bike sharing the most and when are they using them?

1. 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 different places in the city.

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.

2. Data Wrangling

For this problem set, we are working with four different tables. Given the complexity, I have chosen to clean and analyse each Dataframe individually, and then process EDA and merge data as necessary in the future.

Here I'll describe the content of each table. Afterwards I'll describe the steps taken to clean each table

- **Station.csv** - this table has information on the location for each bike sharing station in the bay area as well as when they were installed. Columns includes:
    - **Id** - this is the station id
    - **Name** - this is the station name
    - **Lat** - station latitude
    - **Long** - station longitude
    - **Dock_count** - how many bike docks each station has
    - **City** - which city the station is in
    - **Installation_date** - when each station was installed

- **Trip.csv** - this table contains information on every trip taken on the bike stations over a two year period. It includes trip locations, duration, as well as time information:
    - **Id** - this is a unique trip id
    - **Duration** - this is the trip duration in seconds
    - **Start_date** - when the trip begin. Includes date and time
    - **Start_station_name** - the name of the station where the trip begin

  - **Start_station_id** - id of the start station
  - **End_date** - when the trip ended. Includes date and time
  - **End_station_name** - the name of the station where the trip ended
  - **End_station_id** - id of the end station
  - **Bike_id** - unique id for reach bike
  - **Subscription_type** - information on what type of customer is riding the bike. This can be one of two values:
    - Subscriber
    - Customer
  - **Zip_code** - zip code of where the trip started

- **Weather.csv -** this table includes weather information with various different weather statistics, including but not limited to precipitation and temperature:
  - Date - date the weather information corresponds to
  - Zip_code

- **Status.csv** - this table includes minute by minute information on how many bikes an docks are available for every station id over a two year period:
  - Time - minute by minute breakdown on station capacity
  - Station_id - the id of the station
  - Bikes_available - how many bikes are available at that station at that time
  - Docks_available - how many docks are available at that station at that time

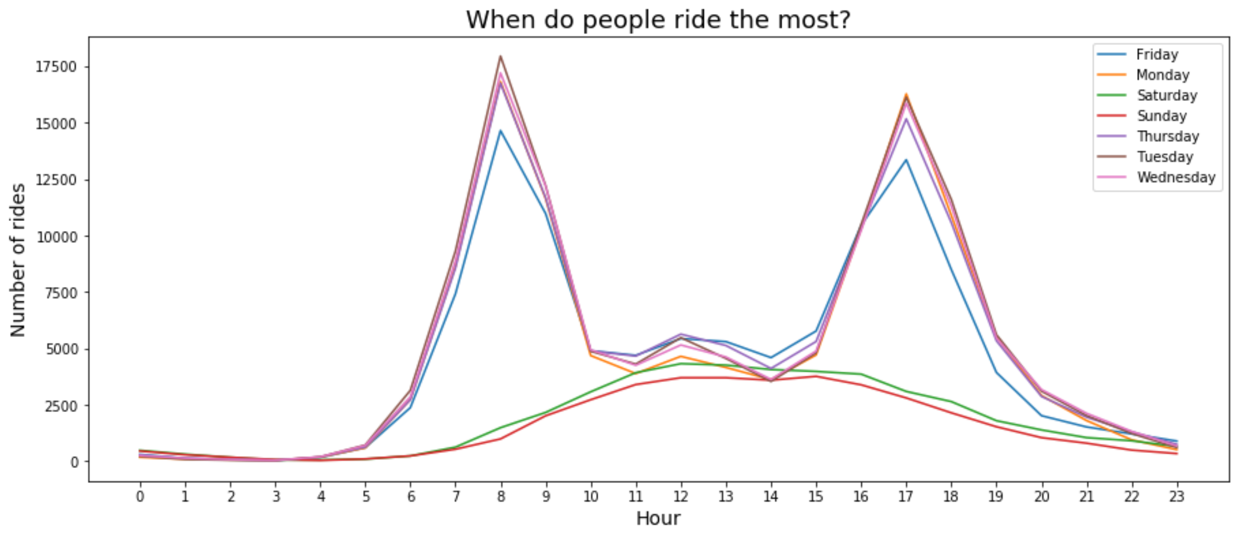
Steps to clean the data:

- **Station.csv** - this table is already clean from the outset. There are no missing values or duplicate rows. All of the dtypes are accurate with the exception of installation date, which I parsed into datetime64[ns] dtype.
- **Trip.csv** - For this table, I parsed the start_date and end_date columns to change dtype to datetime64[ns]. This table did not have duplicate values and the only column with missing values is the zip code field.
  - The duration field is provided in seconds. I have changed the values to be in minutes.
  - For the **zip_code**, there are various mistakes with the data. Some values are “nil”, other values have too many numbers (e.g. 3510011) for a zip code while others have too few (e.g. 5013). Another problem with the zip codes is that even for zip codes which have 5 digits, the zip codes clearly do not below to the bay area as they do not start with the number 9. All zip codes in the 5 cities represented in this dataset start with the digit 9.
    - Instead of trying to clean this up, I dropped the zip_code field and added the city based on start_station_id. This gives me the same information as the zip code and will allow me to join trip with weather information based on the city when doing EDA.

- **Weather.csv** - for the weather data, I dropped most of the fields in the table and kept only a few columns of interest. Date, Max, Mean, and Min temperature, precipitation_inches, and zip_code.
    - With the date field, I parsed the dates into dateimte64[ns] dtype format
    - For the zip_codes, there are only 5 values, each corresponding to an individual city. I used the zip_codes to match which city the weather information came from and added the city accordingly.
    - Precipitation_inches column was read in as an object, even though it should be a float. Taking a closer look at the data, there are strings in place of numbers and that is the reason why it was read in as an object dtype. I replaced the string ‘T’ with 0.00 for precipitation. In addition, there are four rows with NaN values for temperature and precipitation. I filled in those values with a forward fill method.
- **Status.csv** - the status dataframe is clean from the beggining. No duplicate rows or missing values in the dataframe

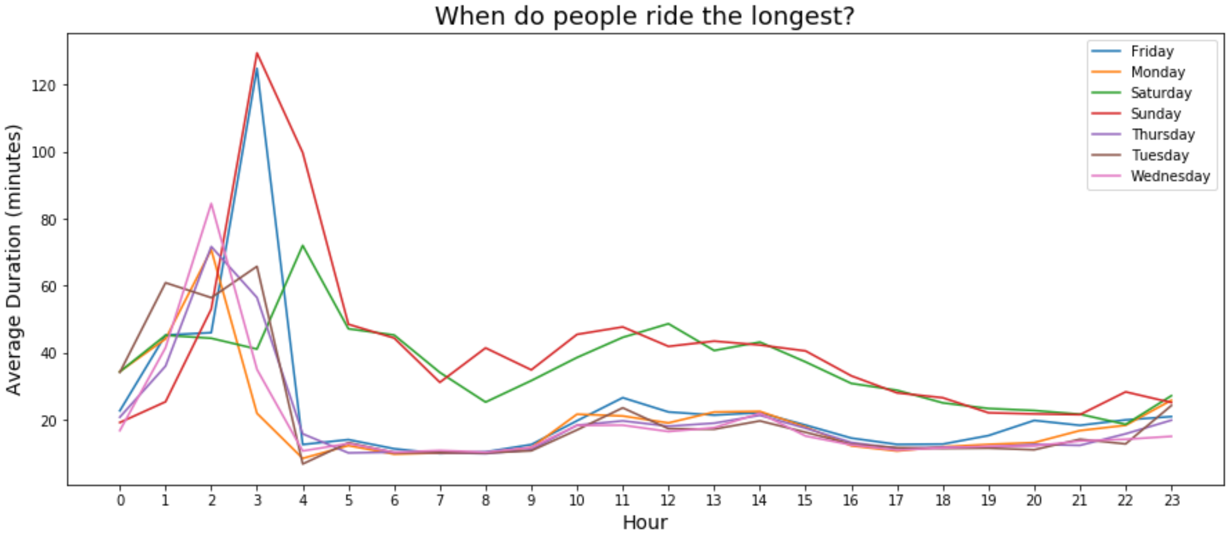
3. Exploratory Data Analysis (EDA)

- There is a lot to find out about this dataset, one of the most insightful information found through EDA is when riders are riding their bikes the most, broken down by hour in the day and day of the week:



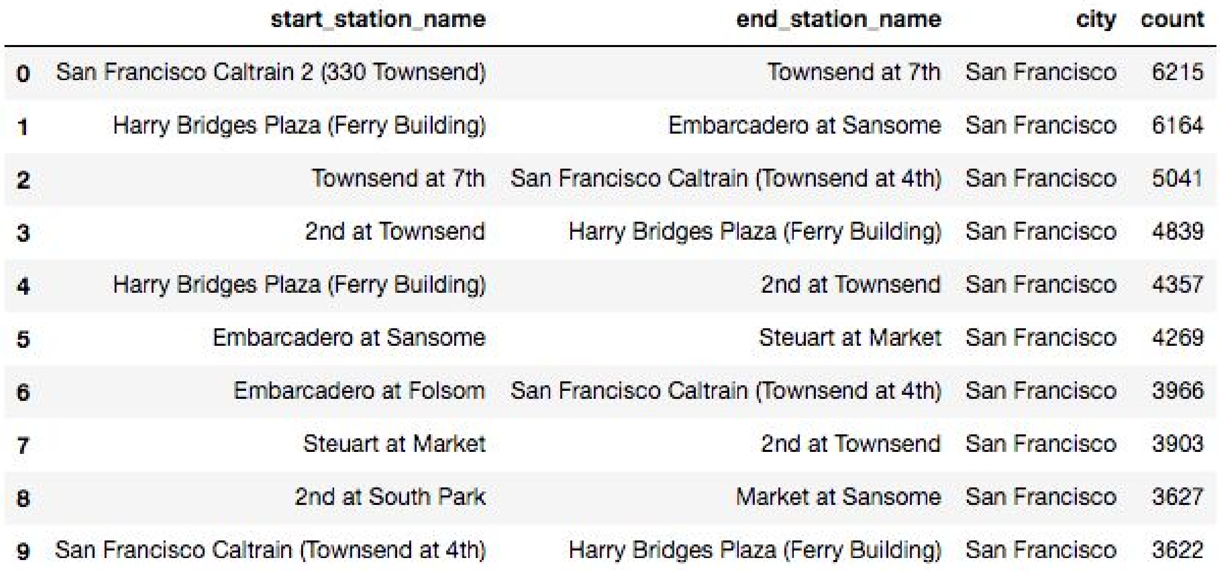
- As can be seen from the information above, riders then to ride most often during commute times, i.e. between 7:00 and 9:00 and between 16:00 and 18:00. Notice also that these times coincide with days of the week, suggesting these are commuters heading to work or home from work.
- There’s significantly less activity during the weekend, which is indicated by the green and red lines in the graph above. Also, ridership tends to peak between 12:00 and 15:00 for rides taken during the weekend.

- Another important piece of information is ride duration. Here’s another graph that is very insightful:



- There is some noise in the data for the early hours in the morning, however, for the most part, one can easily see that the rides that are being taken during the week, during commute times are shorter in duration when compared to rides being taken on weekends or outside of commute hours during the week.
- This reinforces the idea that riders who are riding their bikes at certain times of the day are heading either to work or headed home from work.

- The last point is where people are riding the most. Below is the top routes taken. We can see that riders tend to ride to commute areas, i.e. Caltrain stations, ferry building, or near bart stations:



	start_station_name	end_station_name	city	count
0	San Francisco Caltrain 2 (330 Townsend)	Townsend at 7th	San Francisco	6215
1	Harry Bridges Plaza (Ferry Building)	Embarcadero at Sansome	San Francisco	6164
2	Townsend at 7th	San Francisco Caltrain (Townsend at 4th)	San Francisco	5041
3	2nd at Townsend	Harry Bridges Plaza (Ferry Building)	San Francisco	4839
4	Harry Bridges Plaza (Ferry Building)	2nd at Townsend	San Francisco	4357
5	Embarcadero at Sansome	Steuart at Market	San Francisco	4269
6	Embarcadero at Folsom	San Francisco Caltrain (Townsend at 4th)	San Francisco	3966
7	Steuart at Market	2nd at Townsend	San Francisco	3903
8	2nd at South Park	Market at Sansome	San Francisco	3627
9	San Francisco Caltrain (Townsend at 4th)	Harry Bridges Plaza (Ferry Building)	San Francisco	3622

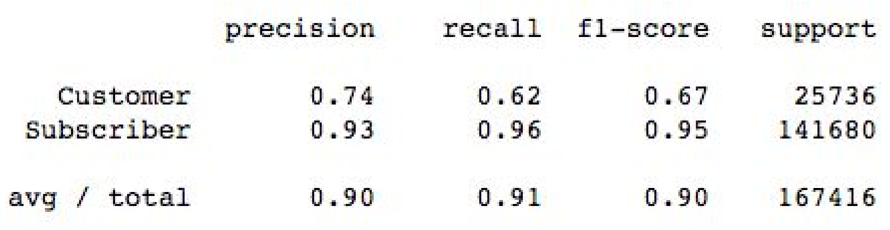
4. Machine Learning

Based on the information laid out above, I created a model to check predict if the riders are either Subscribers (people who pay a monthly fee to use the bikes) or Customers (pay as you individuals).

I tried four different models with to see which one performed better.

- Here are the models I tried:
  - LinearSVC
  - Logistic Regression
  - Naive Bayes
  - Random Forest
- I first ran the model with the following features:
  - Ride Duration
  - Start_station_id
  - End_station_id
  - Mean_temperature
  - Mean_wind_speed
  - Precipitation

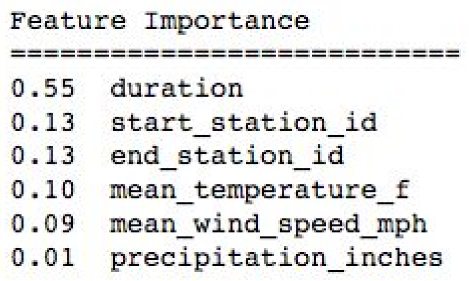
Random Forest was the best performing model out of the four. Here are the results:



```
             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
```

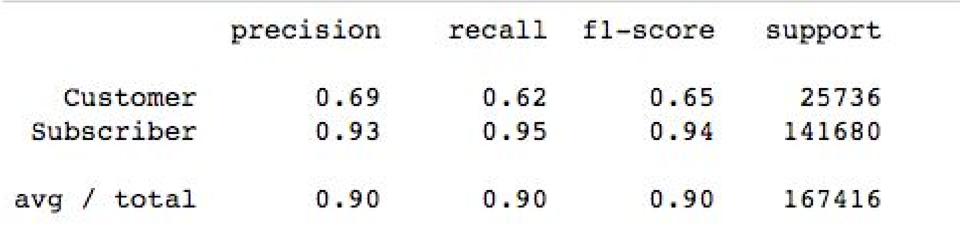


```
Feature Importance
===========================
0.55  duration
0.13  start_station_id
0.13  end_station_id
0.10  mean_temperature_f
0.09  mean_wind_speed_mph
0.01  precipitation_inches
```

After these results, I dropped the features that were below 10% importance and added another feature which indicates the Distance to Duration Ratio. Even though there's no way to know which path the rider took between two stations, we can find the distance between two stations "as the crow flies", and divide it by the trip duration.

This tells us how long a rider took between two stations. This ratio will help us to determine if the rider is a Subscriber or a Customer since Subscribers tend to take faster rides when compared to customers.

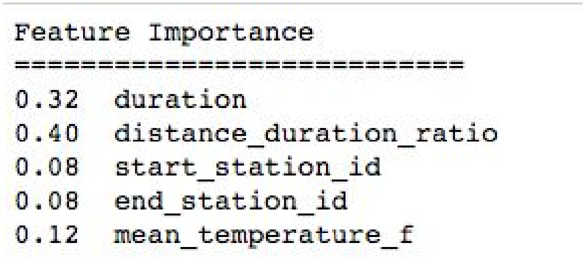
After adding the distance ratio and removing some of the less important features, here's the updated results:



```
             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
```

5. Additional Improvement for future

Next steps for this project is to work on a model that'll predict when stations do not have bikes when users are needing to ride. This is an issue as not having bikes on the bikes is not only immediately frustrating, but it can also curb future demand for the company. If riders are unable count on their station having bikes when they need them, they may be less willing to use the service in the future.