Name: **Muley, Tushar**
Assignment: **DSC 680 - Week 3 Milestone 2 White Paper**
Date: **January 30, 2022**

# Metro Bike Share System

**Draft White Paper – Milestone 2**

## Table of Contents:

## Introduction:

Bike sharing has come a long way since it was first seen in 1965 in Amsterdam. The real draw to bike sharing is getting around in highly populated urban city. Cities like Los Angeles with increasing population and cars that are trapped in traffic jams, bikes are becoming viable options to getting the city population around. You have many other reasons like healthier lifestyle, climate change, and increasing pollution forcing a change in lifestyles. As automobiles are contributor to increase health risks and global warming concerns many of the world's largest cities are looking to limit the usage cars or band the usage of cars from entering city centers. Enter a possible solution to getting around. Bike sharing is a simple and elegant answer to the gridlock, air pollution and healthier lifestyle.

Now with technology bike station can be anywhere and don't require an attended to watch over them to make sure the bikes get returned. The bike sharing business steadily pulled in $1.5 billion in 2020. That number is expected to grow to $4.4 billion by the year 2027. [2] In the city of Los Angeles Metro Bike Share charges about $1.75 for a single 30-minute ride, $5.00 for 24-hours, $17.00 for a monthly pass and $150 for the annual pass. All reasonable amount if you need transportation in a bind or cover a larger distance than walking.

These sharing networks are showing up in various cities around the world. The bikes themselves have little tech on them. But they allow user from the existing city and visitors to the city to use them. The technology makes the sharing network work. The bikes are tracked via GPS as is the phone used to purchase the ride. The same technology helps riders find available bikes for use. The rider's credit card is kept on file in the event the bikes is not returned, lost or

parts are missing, so the last user can be charged (this is similar model to car rental company). Overall, the tech makes the ride sharing network simple.

In this case study I will perform Exploratory Data Analysis, answer research questions about bike sharing. Then I will split the data into test and training to perform a predictive analysis on the number of trips from a station using various models. From there will provide the best model for the analysis with resulting data.

## Data Set/Research Questions:

I obtained the data set from Metro Bike Share. Metro partnered with the city of Los Angeles to provide bikes 24/7, 365 days a year in Downtown Los Angeles, Central Los Angeles and North Hollywood. I pulled four quarters of data for 2021, which is the most recent data available. The appendix section contains column names and description of the attributes for two tables. One table contains bike usage data (titled Bike Usages Table) and the second table contain station information (titled Station Information Table). I also combined the data with a calendar table that provide day of week to determine, which days of the week are heavily utilized by riders. The total data set is 220,997 rows with 17 columns. I did remove data points and columns contain nulls values. A data dictionary is provided in the appendix section of this paper for all the tables used in the analysis.

Data obtained from: https://bikeshare.metro.net/about/data/

Below are the research questions I will be answering with the bike sharing data. Hopefully this assist in making the bike sharing networks a little more efficient, easier to use and maybe a little more profitable.

1. Does ridership change over the course of a year? Does it go up or does it go down?
2. How is usage of the bikes? Does weekend or weekdays impact usage?
3. How is the ride duration change between membership riders and casual riders (one-time users)?
4. Identify the busiest bike stations
5. What features (station location, electric bites) are influencing trip count?
6. Building a model to predict the number of trips from a station?

## Method:

Once the tables were loaded into Python I combined all the quarterly data into a single dataframe. I brought in other pieces of data like station names and calendar information to further help the analysis. I had to do some data cleansing as there where some null values. I removed ‘test’ rides and transformed some of the dates and times to be more analytic friendly. From there I stated my EDA process.

## Exploratory Analysis:

The dataset was divided in four different quarters for the 2021 year. Below is a sample of the bike usage data.

| | trip_id | duration | start_time | end_time | start_station | start_lat | start_lon | end_station | end_lat | end_lon | bike_id | plan_duration | trip_route_c |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 151713183 | 17 | 1/1/2021 1:45 | 1/1/2021 2:02 | 3005 | 34.048500 | -118.258537 | 4304 | 34.062580 | -118.290092 | 5894 | 1 | ( |
| 1 | 151713983 | 7 | 1/1/2021 2:35 | 1/1/2021 2:42 | 4390 | 34.069271 | -118.296593 | 4456 | 34.052429 | -118.302017 | 16901 | 365 | ( |
| 2 | 151716483 | 8 | 1/1/2021 4:28 | 1/1/2021 4:36 | 3052 | 34.051102 | -118.264557 | 4314 | 34.057709 | -118.279762 | 6005 | 30 | ( |
| 3 | 151721185 | 208 | 1/1/2021 4:43 | 1/1/2021 8:11 | 3034 | 34.042061 | -118.263382 | 3031 | 34.044701 | -118.252441 | 5852 | 1 | ( |
| 4 | 151720984 | 129 | 1/1/2021 5:53 | 1/1/2021 8:02 | 4446 | 34.053230 | -118.278419 | 4446 | 34.053230 | -118.278419 | 12075 | 1 | Ro |

**Table 1: Sample of the bike usage data**

Table 2 contains the column names, number of rows and datatypes.

```
RangeIndex: 220997 entries, 0 to 220996
Data columns (total 17 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   trip_id              220997 non-null  int64
 1   duration             220997 non-null  int64
 2   start_time           220997 non-null  object
 3   end_time             220997 non-null  object
 4   start_station        220997 non-null  int64
 5   start_lat            220770 non-null  float64
 6   start_lon            220770 non-null  float64
 7   end_station          220997 non-null  int64
 8   end_lat              215391 non-null  float64
 9   end_lon              215391 non-null  float64
 10  bike_id              220997 non-null  object
 11  plan_duration        220997 non-null  int64
 12  trip_route_category  220997 non-null  object
 13  passholder_type      220996 non-null  object
 14  bike_type            220997 non-null  object
 15  start station name   59081 non-null   object
 16  end station name     59081 non-null   object
dtypes: float64(4), int64(5), object(8)
memory usage: 28.7+ MB
```

**Table 2: Bike usage dataset with column names and data types**

The next table contain the dataset with station information which I will be combining with bike usage dataset. This will help make the analysis easier since there will only be one dataframe to work with.

| | Station_ID | Station_Name | Go_live_date | Region | Status |
|---|---|---|---|---|---|
| 0 | 3000 | Virtual Station | 7/7/2016 | NaN | Active |
| 1 | 3005 | 7th & Flower | 7/7/2016 | DTLA | Active |
| 2 | 3006 | Olive & 8th | 7/7/2016 | DTLA | Active |
| 3 | 3007 | 5th & Grand | 7/7/2016 | DTLA | Active |
| 4 | 3008 | Figueroa & 9th | 7/7/2016 | DTLA | Active |

**Table 3: Sample of Station information data**

The next table show the descriptive statistic of the data in the bike usage dataset.

| | trip_id | duration | start_station | start_lat | start_lon | end_station | end_lat | end_lon | plan_duration |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 2.209970e+05 | 220997.000000 | 220997.000000 | 220770.000000 | 220770.000000 | 220997.000000 | 215391.000000 | 215391.000000 | 220997.000000 |
| **mean** | 1.654887e+08 | 47.007543 | 3967.623352 | 34.030383 | -118.342844 | 3942.645217 | 34.030388 | -118.343207 | 50.520423 |
| **std** | 7.768117e+06 | 136.665633 | 636.370070 | 0.037749 | 0.095357 | 644.900711 | 0.037234 | 0.095916 | 105.045401 |
| **min** | 1.517132e+08 | 1.000000 | 3000.000000 | 33.928459 | -118.491341 | 3000.000000 | 33.928459 | -118.491341 | 1.000000 |
| **25%** | 1.587778e+08 | 9.000000 | 3064.000000 | 33.998341 | -118.451248 | 3062.000000 | 34.000309 | -118.451248 | 1.000000 |
| **50%** | 1.665581e+08 | 19.000000 | 4215.000000 | 34.039188 | -118.290092 | 4215.000000 | 34.038609 | -118.291313 | 30.000000 |
| **75%** | 1.714749e+08 | 36.000000 | 4491.000000 | 34.050880 | -118.258537 | 4490.000000 | 34.050480 | -118.258537 | 30.000000 |
| **max** | 1.794925e+08 | 1440.000000 | 4594.000000 | 34.186569 | -118.225410 | 4594.000000 | 34.186569 | -118.225410 | 999.000000 |

**Table 4: Descriptive Statistic of the Bike Usage Dataset**

Not a lot going on with the data as it has been cleaned and very well prepared by Metro Bike Share before making it public.

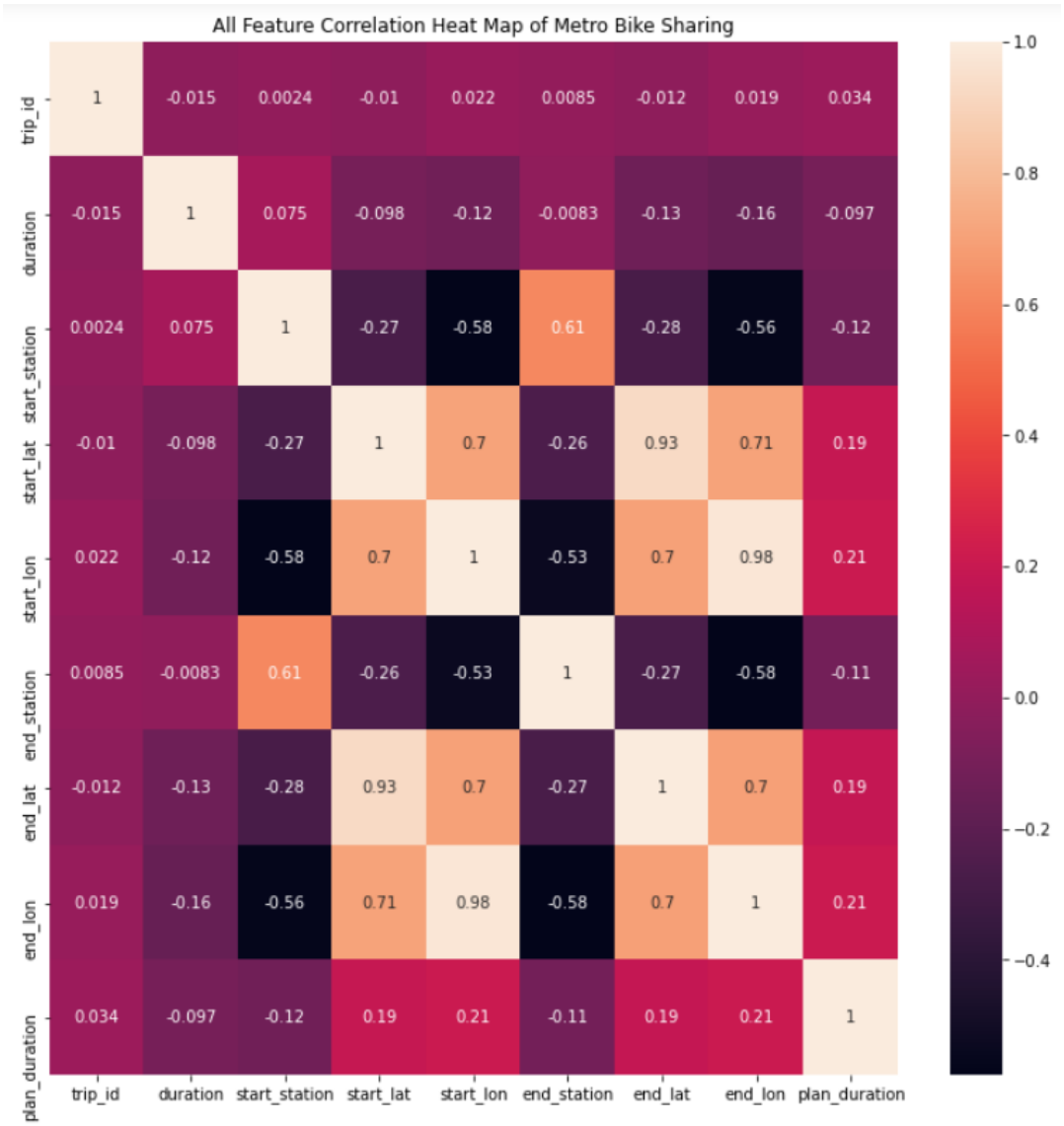


**Figure 1: Correlation Matrix for all feature in Metro Bike Sharing Analysis**

Figure 1 is a correlation matrix of the Metro Bike Sharing. The latitude and longitude part of the data have great correlation as that is expected. There is good correlation between duration and location, which might be a good indicator of areas that have larger amount of tourist or have good sightseeing opportunities or have well-built and safe biking infrastructure.

As part of the research, I was going to ask question of the data. Two of those questions were does ridership change over the course of a year? Does it go up or does it go down? Below is a line chart showing what ridership looks like throughout 2021.

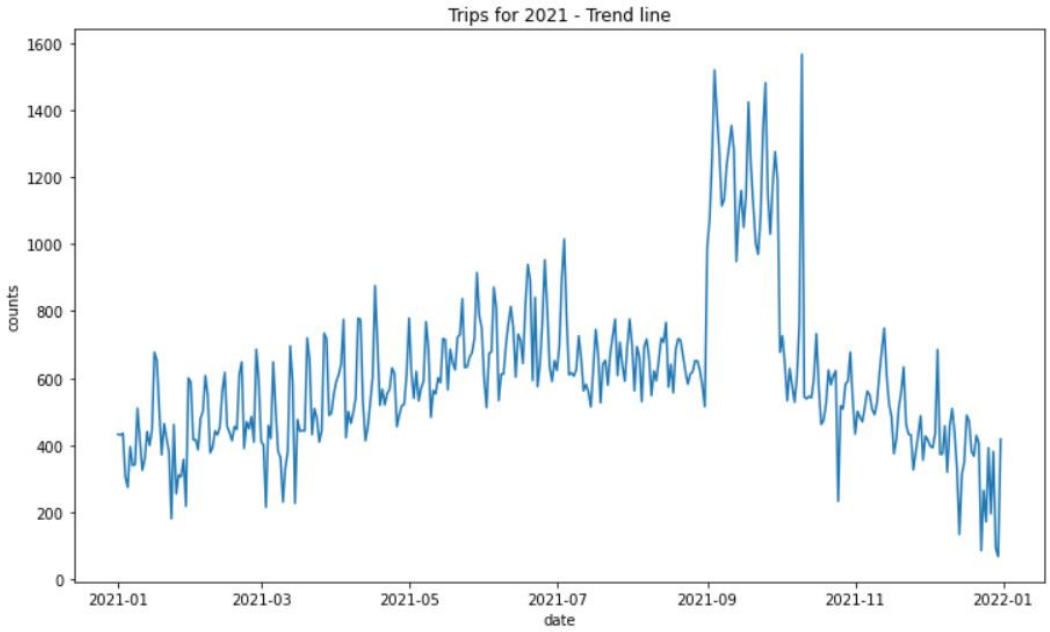


**Figure 2: Line chart of trips between January and December 2021**

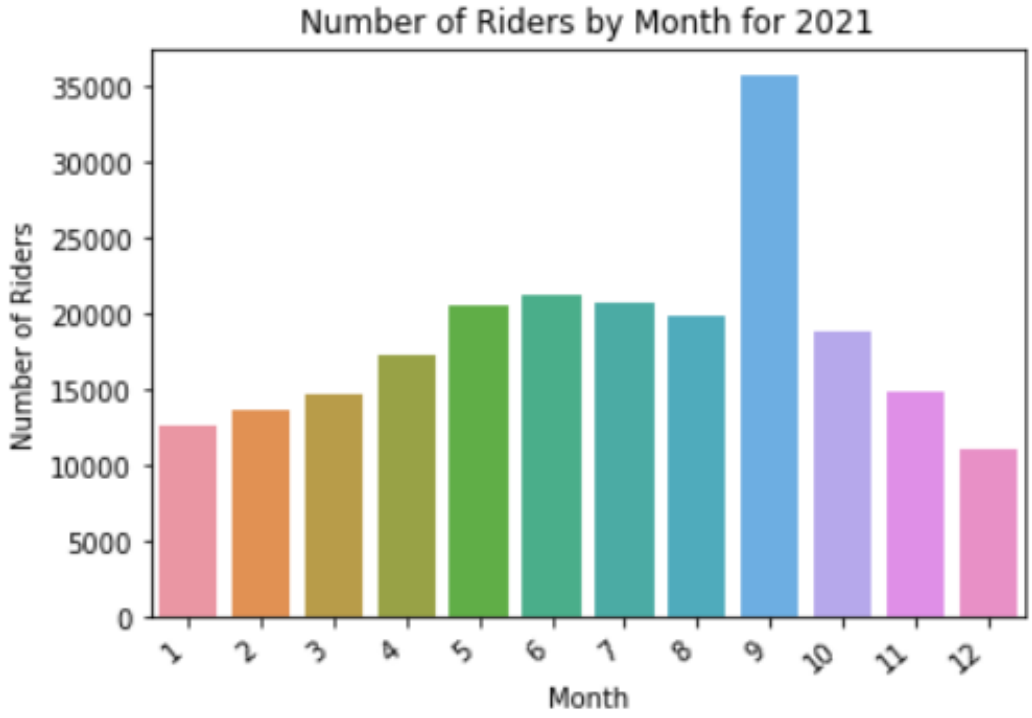


**Figure 3: Bar chart of number of Riders by month for 2021**

The two figures above (Figure 2 and Figure 3) are different visuals of the number of trips riders have taken from the January till December of 2021. Based on the bar chart as weather warms up the number of trips increase. Usage remains stead until September where it spikes. That spike is followed by valley coinciding with the holiday season and colder weather.

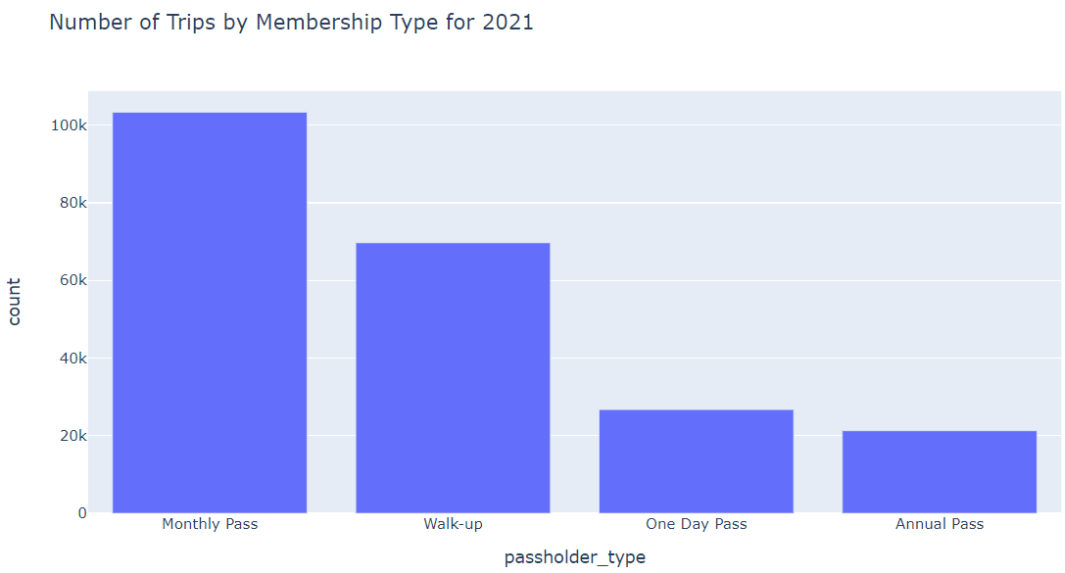


**Figure 4: Bar chart of Passholder Types compared to Usage**

The above bar chart displays passholder types to the number of rides taken. The number of annual pass holder are smaller compared to the monthly pass holders. An annual pass is $150.00 compared to $17.00 for a monthly pass.

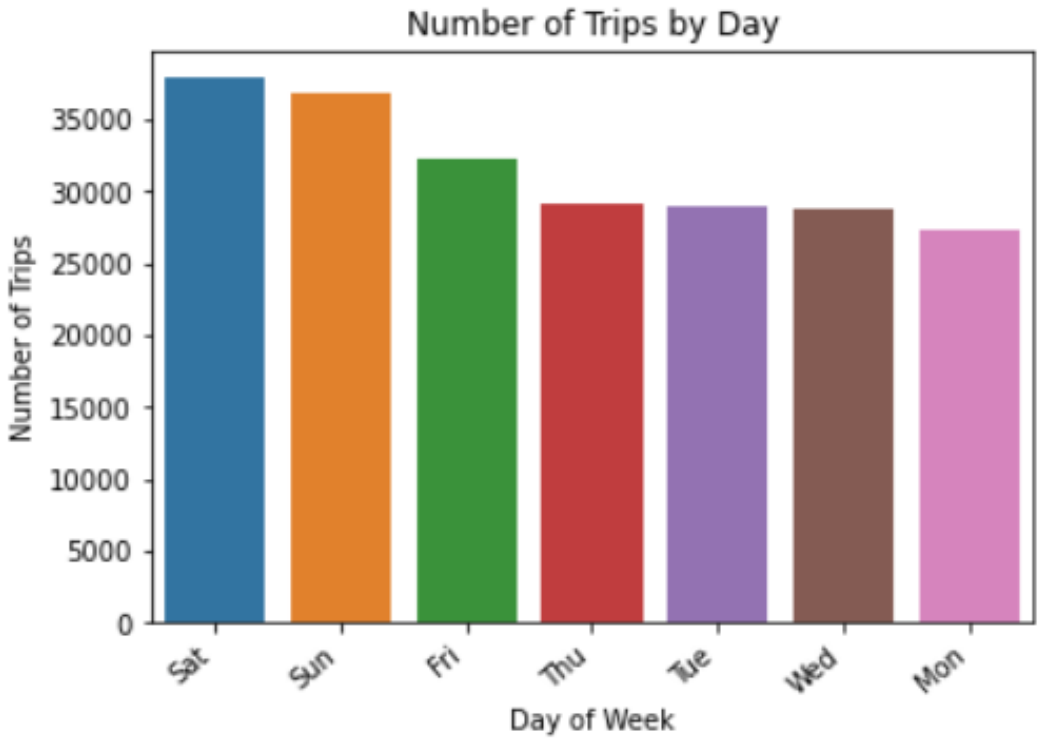


**Figure 5: Number of Trips by Day of Week**

In Figure 5 I answer the question of, does the day of the week effect trips? Based on the data Saturdays and Sundays are the highest with Mondays being the lowest usage days. Overall bike usage is stead for the whole week.

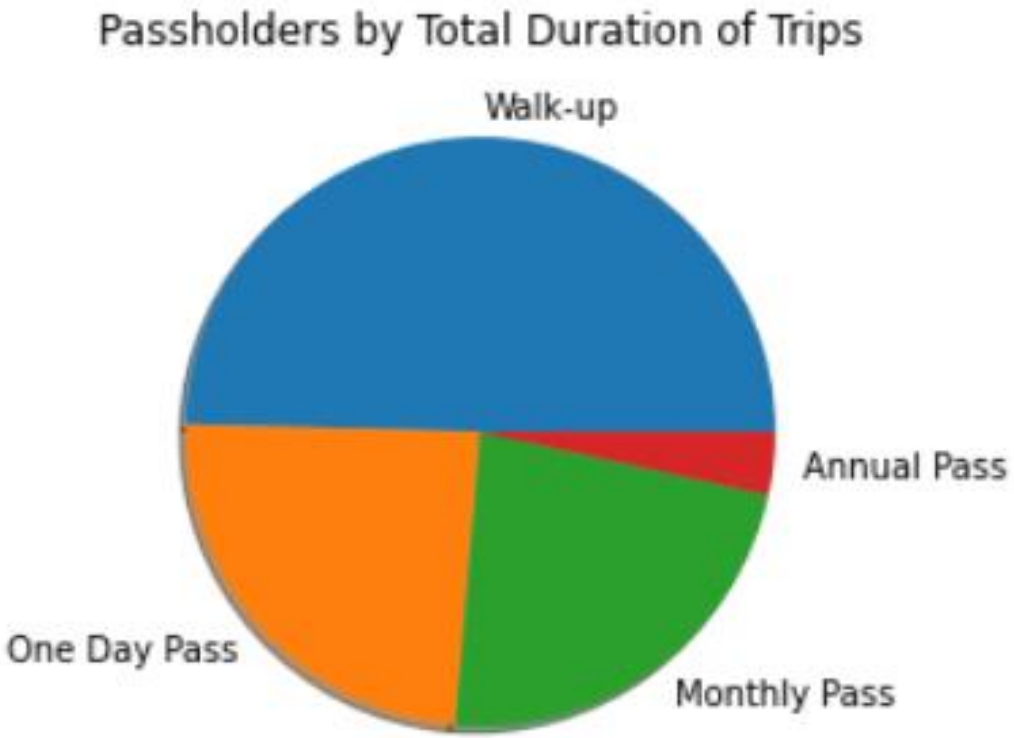


**Figure 6: Pie chart of Total Duration of Trips by Passholder Type**

From Figure 6 we can see that Walk-up passholder make up nearly 50 percent of the user with the largest duration. The true number is 49.58%. One Day passholders make up 24.05 percent of the users. Monthly passholder make up 22.92% and just 3.44% of Annual passholder.

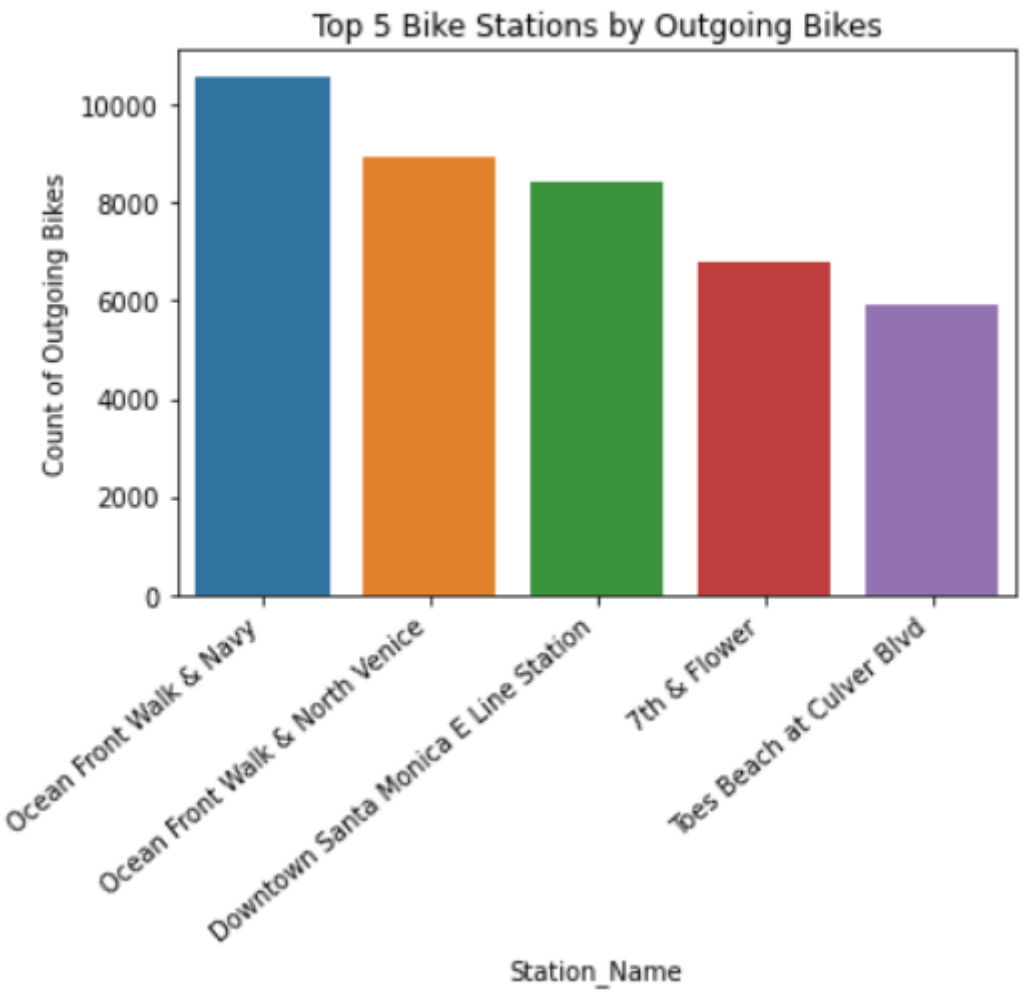


**Figure 7: Top 5 Busiest Bike Stations**

Figure 7 shows the top five busiest stations. Station Ocean Front Walk & Navy and Ocean (4214) Front Walk & North Venice (4210) are both located at Venice Beach. This is the most icon places in Los Angeles with Muscle Beach and Venice Public Art Wall. Seen in countless movies, the location it is a prime spot. The next station is 4215 which is in downtown Santa Monica near the E-Train Line.

### Assumption/Limitation/Challenges:

The data seems straight forward. It was obtained from the source meaning no combining of data with other information that might cause unknown arrogation or interpretation problems. I combined the data with day or week information. My biggest concern is how would I deal with the large volume of data? It could cause memory issue or other unknown issue I have yet to confront. Those concerns were alleviated when I built out the outgoing data feature. This feature totaled up the trips by date reducing the amount of rows.

## Ethical Considerations:

This data is publicly available due to the partnership with city of Los Angeles. The data does not contain any personally identifiable information. The information can be used to determine bike usage and consideration for additional stations. The data might be used to determine better bus routes or other public transportation options. Metro Bike does do some scrubbing of the data prior to releasing it. If financial information was provided like type of credit card used or age or birth dates this information could be used for rider segmentation.

If age data was available it could be used to determine age range of riders. This would allow for information to promote on the Metro Bike app. This could be used to generate other add revenue similar to gas station adds as your filling up your tank. This information could be used to tie in other sponsors like local restaurants or bars or shopping venues. This would increase Metro Bikes revenue from alternative sources.

Outside of the promotion aspect of the data if more attributes were provided one could determine better bike stations locations based on address of users. The data could be used to increase pedestrian and bike lanes. Usage of the data could be used to possibly close street to cars to increasing foot traffic. This would reduce parking needs and may lead to increase business for nearby business.

## Test/Training the Model:

Once completing my analysis, I started the splitting the data into test and train for my predict model. I split the model in a few different ways first using the 80/20 train and test split. Later using the 70/30 split for the training to testing dataset. I settled on using 80/20 as I got a slightly better accuracy number.

```
# import sklearn train_test split
from sklearn.model_selection import train_test_split
# split into train - test sets
X_train, X_test, y_train, y_test = train_test_split(df_features,
                                   df_target,
                                   test_size=0.30,
                                   random_state=42)
```

**Figure 8: Python code for splitting data between test and train**

## Model Evolution:

Based on the evolution of the target variable outgoing bikes trips (no_outgoingbikes) I understood this as a non-linear model so I did not try using a linear model to start. Once the data had been prepared I ran various version of Decision Tree model and Random Forest model. Below are the results.

The Decision Tree model and the Random Forest model came out on top of each other. This is a little concerning and the accuracy numbers are very low. I expected higher numbers.

The appendix contains a table of the top 20 records from the Decision Tree model. The values appear very far off.

```
Train data - mean Absolute Error  2.8733050508948157
Train data - mean Squared Error  29.033386627709646
Train data - R-square  0.0654078410706781

Test data - mean Absolute Error is :  2.9222000992752646
Test data - mean Squared Error :  30.852367757339447
Test data - R-square  :  0.04605687190970775
```

**Figure 9: Show the Results from Train and Test for Decision Tree model**

The above Figure 9 show the results returned for the Decision Tree model. The results are lower than expected. The R-squared for training data was 6.54% while test data was 4.61%.

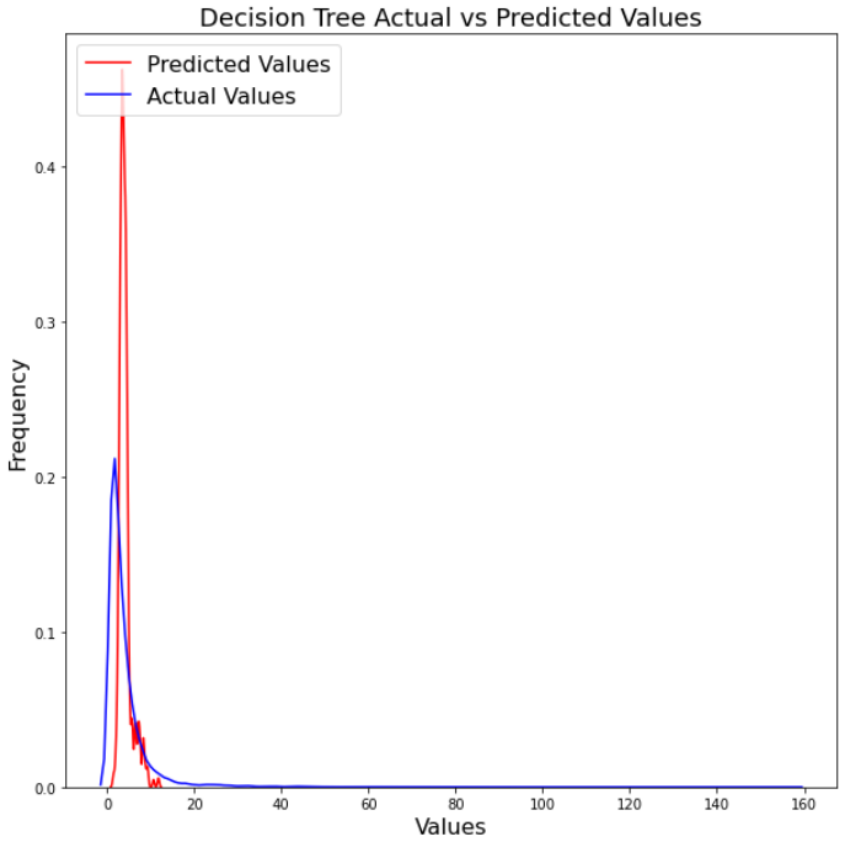


**Figure 10: Show the results of the Decision Tree model**

If you look at Figure 10 the graph shows the difference between predict values shown by the red line and the actual values shown by the blue line.

The Random Forest model results where very similar as well. Below in Figure 11 you can see R-squared value for train set was 6.54% and for test is was 4.62%. Figure 12 show the visual using a scatter plot.

```
Train data - Mean Absolute Error is :  2.8726156403858805
Train data - Mean Squared Error :  29.033736316166788
Train data - R-square :  0.06539658450959185

Test data - Mean Absolute Error is :  2.9211486490770917
Test data - Mean Squared Error :  30.84645506506924
Test data - R-square:  0.04623969003906625
```

**Figure 11: Show the Results from Train and Test for Random Forest model**

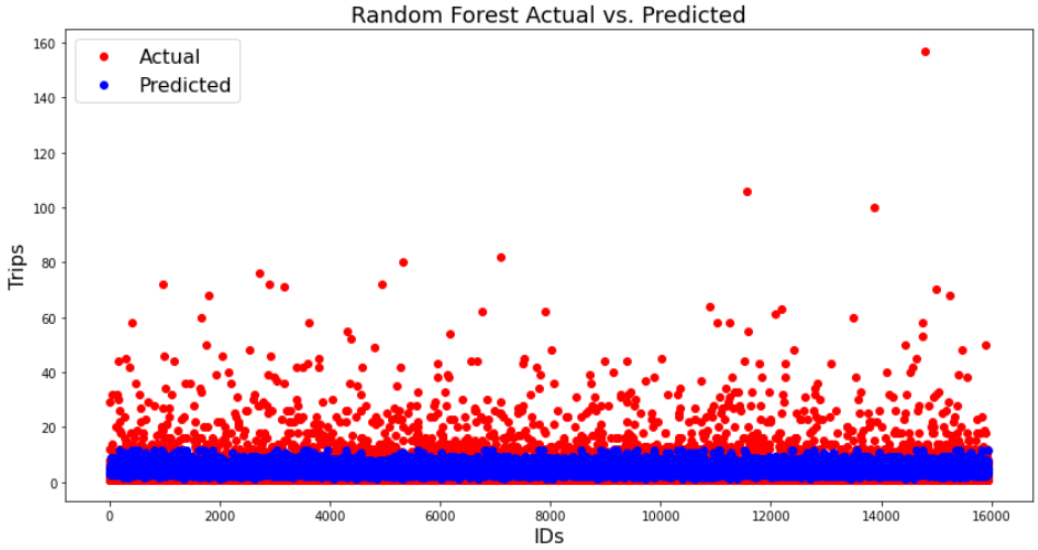


**Figure 12: Scatter plot showing Actual values in red and Predicted values in blue**

## Conclusion:

The outcome of this project was to predict the count of outgoing bike trips from each station based on features that were available in the dataset. The project started by extracting the data using Pandas library and performing exploratory data analysis. The data was in quarters so needed to be merged in to a signal dataframe. I brought in additional elements from a calendar file. I engineered a feature called outgoing bike, since it did not exist as my dependent variable. There appeared a need to make the values a uniform range. I used Standard Scaler from Sklearn library to perform this task. The data was than split into training and test data. I ran two models a Decision Tree model and a Random Forest model. The results of the Decision Tree model and Random Forest model were identical. The Decision Tree training data had an $R^2$ of 0.06540 and the testing data set returning $R^2$ 0.04605. The Random Forest training data had an $R^2$ of 0.06539 and testing had an $R^2$ of 0.04623. Very similar if not the same.

**Reference:**


1. LA metro bike share network https://bikeshare.metro.net/about/
2. McEvoy, J. (January 2022). “Bike Sharing Market Accelerating to $4.4 Billion by 2027”. From Startup Savant. https://startupsavant.com/news/bike-sharing-market
3. “Bicycle sharing System”. (December 2021). From Wikipedia.com. https://en.wikipedia.org/wiki/Bicycle-sharing_system
4. Walker, A. (December 2019). “The quiet triumph of bike share”. From curbed.com. https://archive.curbed.com/2019/12/16/20864145/bike-share-citi-bike-jump-uber
5. Malouff, D (January 2017). “All 119 US bikeshare systems, ranked by size”. From Greater, Greater Washington.org https://ggwash.org/view/62137/all-119-us-bikeshare-systems-ranked-by-size
6. Hosford, K & Winters, M & Sersli, S. (January 2020). “More people are using bike share programs, but the gender gap persists”. From Green Biz.com https://www.greenbiz.com/article/more-people-are-using-bike-share-programs-gender-gap-persists
7. Perry, S. (May 2020).” More people commute by bike in cities with bike-sharing programs, study finds”. From MINN POST.com https://www.minnpost.com/second-opinion/2020/05/more-people-commute-by-bike-in-cities-with-bike-sharing-programs-study-finds/

## Appendix:

Bike Usage Table:

| Column Name | Column Description |
|---|---|
| trip_id | Locally unique integer that identifies the trip |
| Duration | Length of trip in minutes |
| start_time | The date/time when the trip began, presented in ISO 8601 format in local time |
| end_time | The date/time when the trip ended, presented in ISO 8601 format in local time |
| start_station | The station ID where the trip originated (for station name and more information on each station see the Station Table) |
| start_lat | The latitude of the station where the trip originated |
| start_lon | The longitude of the station where the trip originated |
| end_station | The station ID where the trip terminated (for station name and more information on each station see the Station Table) |
| end_lat | The latitude of the station where the trip terminated |
| end_lon | The longitude of the station where the trip terminated |
| bike_id | Locally unique integer that identifies the bike |
| plan_duration | The number of days that the plan the passholder is using entitles them to ride; 0 is used for a single ride plan (Walk-up) |
| trip_route_category | "Round Trip" for trips starting and ending at the same station or "One Way" for all other trips |
| passholder_type | The name of the passholder's plan |
| bike_type | The kind of bike used on the trip, including standard pedal-powered bikes, electric assist bikes, or smart bikes. |

Station Information Table:

| Column Name | Column Description |
|---|---|
| Station ID | Unique integer that identifies the station (this is the same ID used in the Trips and Station Status data) |
| Station Name | The public name of the station. "Virtual Station" is used by staff to check in or check out a bike remotely for a special event or in a situation in which a bike could not otherwise be checked in or out to a station. |
| Go live date | The date that the station was first available |
| Region | The municipality or area where a station is located, includes DTLA (Downtown LA), Pasadena, Port of LA, Venice |
| Status | "Active" for stations available or "Inactive" for stations that are not available as of the latest update |

Calendar Information Table:

| Column Name | Column Description |
|---|---|
| sasdate | SAS date field from the application the data was pulled from |
| date_key | Primary Key for the table |
| word_date | Date with the month written out in alphabetical characters. The format is dd/mmm/yy. |
| date | Actual date corisponding to word date. The format is dd/mm/yyyy. |
| year | The year coresponding to word date. The format is yyyy. |
| quarter | The quarter corespinding to the word date. The format is qq. |
| month | The month corresponding to the word date. The format is mm. |
| day_of_month | The day of the month counting from 1 to 365. |
| week | The week in which the word date corresponds to in a 52 week year. |
| day_of_week | The day of the week in 3 alphabetical characters. The format is ddd. |
| weekday | The day of the week based on number 1 - 7 where 1 is Sunday and 7 is Saturday. |
| month_and_year | The month and year of the corresponding word date. The format is mmm-yy |
| holiday | Name of the holiday that occurs on the word date. |

Image of the Top 20 rows of the Decision Tree model:

| | Real Values | Predicted Values |
|---|---|---|
| **24735** | 1.0 | 4.281818 |
| **7903** | 2.0 | 2.928571 |
| **32179** | 4.0 | 4.059406 |
| **35526** | 29.0 | 4.377049 |
| **26946** | 2.0 | 5.264957 |
| **44887** | 1.0 | 3.564356 |
| **31387** | 3.0 | 3.851852 |
| **22253** | 6.0 | 4.542373 |
| **28961** | 5.0 | 3.892562 |
| **28342** | 1.0 | 4.382609 |
| **2593** | 8.0 | 4.522936 |
| **2992** | 1.0 | 4.063063 |
| **31266** | 6.0 | 4.127451 |
| **6242** | 12.0 | 3.153846 |
| **16461** | 2.0 | 4.085714 |
| **43128** | 2.0 | 3.680000 |
| **41202** | 3.0 | 4.144231 |
| **29098** | 7.0 | 4.260000 |
| **11648** | 2.0 | 3.403670 |
| **14039** | 1.0 | 3.233645 |

## Questions:

1. Can you factor in weather conditions and holidays also while drawing the trend line on the number of trips? They would also have impact on the bike trips.
2. Can this model be used for other bike sharing networks in other cities?
3. Who is using the bike sharing most? Is it casual users or members?
4. When are bike sharing trips occurring the most? What season or what month of the year?
5. How are the trips split across the stations in the city? Are they evenly spread out or concentrated around one station the most?
6. Do we have the information on fare and payment data? Can you calculate average price per trip?
7. What can you do to improve the predictive model for outgoing bikes further?
8. Any other predictive analytics use cases can be done on bike sharing network dataset?
9. What is the customer retention rate? Do bike sharing have repeat users?
10. What information would assist in helping biking sharing networks better position for growth?