ANALYSIS OF METRO BIKE SHARE: LOS ANGELES

GROUP 4

ANKIT JAIN

CASEY KAN

DIDO CHANG

SIJIA LI

SONG HAN

Table of Contents

Executive Summary

Shared mobility as a new trend of transportation is taking root and has become the pick for people to gain short-term access to transportation modes on an “as-needed” basis. According to CNN, Los Angeles consistently tops lists of world’s most congested cities, which has recently adopted a new urban mobility implementation like **Metro Bike Share**. Metro Bike Share is the first bike share system in the Los Angeles, California metropolitan area. The service was launched on July 7, 2016, covers 3 areas in Downtown LA, Westside, and North Hollywood. Since it was first put into operation in 2016, over 1 million rental trips have been accomplished till date. People use this service for transportation, as well as for leisure and exercise purposes. This bike share rental system was being used a lot but as the pandemic started, they started bringing on some changes in order to offer relief to essential workers and people impacted during this time. We thought it would be interesting to see the underlying changes on this bike share system which took place this year and make some recommendations for the company based on our findings.

Business Ideas

By exploring and examining the data, we will have a comprehensive understanding of this service by getting what factors might influence users to rent a bike during a time period based on different factors. In addition, we are able to analyze and provide valuable business insights according to our findings of the data.

Data Acquisition

Source: https://bikeshare.metro.net/about/data/

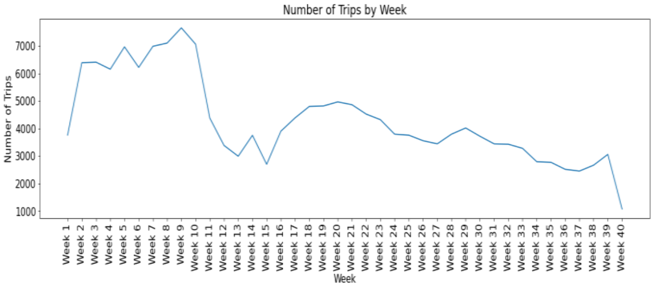
The official Metro Bike Share website has datasets uploaded every quarter for rentals made in Los Angeles, with each dataset being over 50,000 rows. It also has the dataset which has information for the different stations present in the dataset. We used the datasets for the station, and we used the data for every quarter until now in 2020 [January to September]. Each instance in the quarterly datasets contains quality variable information including start-end time, station number, bike ID, pass type and etc. In order to get more factors and information which may influence the number of rentals in a day, it was essential to get more information about the weather conditions. To get the data for weather conditions, we used an API called Weather Visual Crossing [https://www.visualcrossing.com/weather-api]. We used the Urllib package as well as the CSV packages on Python to create the url with the API Key and then extract the data, and convert it usable form for our analysis. We got information about the average temperature, conditions, maximum temperature in the city of LA for the time period we were making observations for.

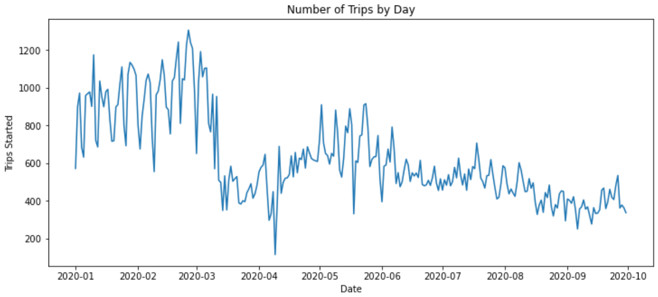
Data Preparation

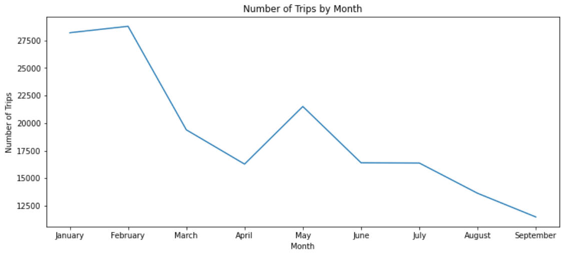
- For all three quarter data, we first converted the start date and end date attributes to Datetime form, and also added another column for month. We also added the station names based on the station ids for each row from the stations dataset in order to merge the station and quarterly datasets. We also added the weather data for each day to the dataset as well in order to merge the weather dataset with the quarterly dataset as well.
- We then merged all the 3 quarterly datasets to make one dataset to use. We then dropped the longitude and latitude and columns as they were all the same for almost all of them and had a lot of missing values and wouldn’t add much to our analysis.
- The final dataset which we used for our analysis which included the weather data, station data and data for the three quarters in 2020 had 172097 rows and 16 columns.

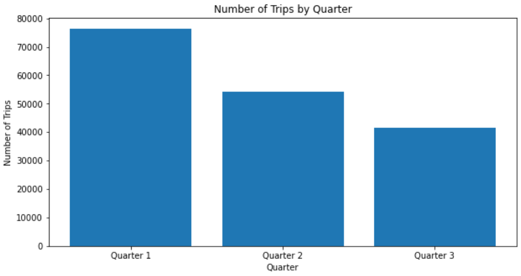
Data Processing

Our data had a lot of categorical based data, so it was essential for us to quantify our data into usable form in order to make useful visualizations and models. In order to understand which degree of time periods would be most useful for us, we made cohesive datasets which encompassed the number of trips, the different frequencies for the different bike types, passholder types and trip categories, weather information, popular stations etc for different degrees of time periods. We made datasets based on daily, weekly, monthly and quarterly. In order to understand which time degrees would be most useful for us to observe further, we decided to graph the number of trips in terms of day, week, month and quarter. These were the results we got:







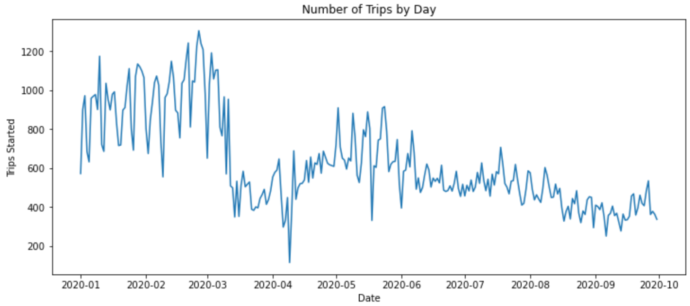


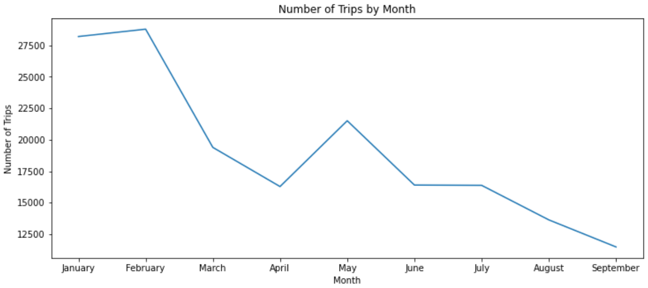
What we noticed from these graphs is that we seem to get the most cohesive and detailed trends from the Monthly and Daily graphs so we decided to explore them further in our analysis.

Data Visualization and Analysis

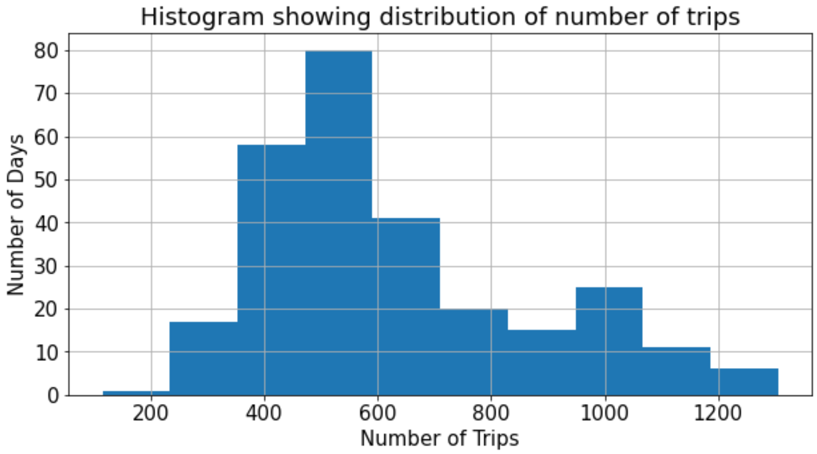
- Observed Time Period: **January 2020 to September 2020 - 274 Days**
- Total Number of Rental Trips Observed: **172097 Trips**
- Average Number of Trips per day in time period: **628 Trips**
- Average Duration of Trips in time period: **37.15 mins**

Analysis: Number of Trips over the time period



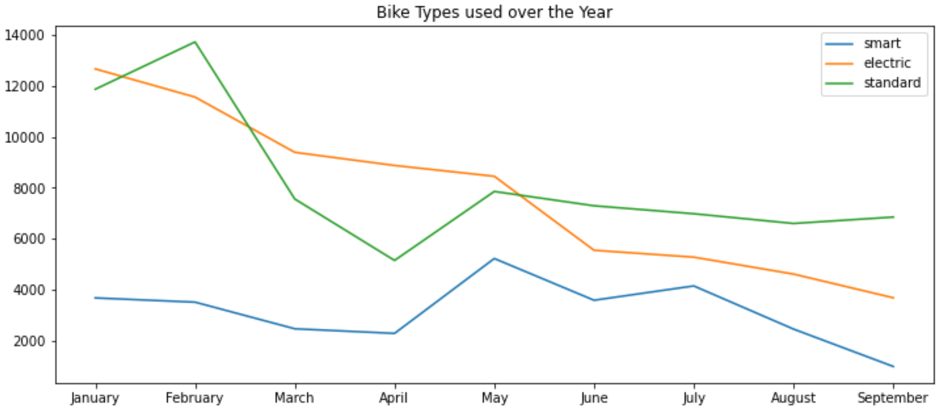


The number of rental trips essentially indicates how much consumers were engaged with the bike share system and basically a major chunk of their revenue stream. What we noticed is that after the month of March which is when the COVID-19 Pandemic lockdown picked up in the city, the number of trips noticeably dropped due to the fact that people were staying home and were going out less. What is interesting is that in May there seems to be a small peak, which is specifically due to the surge in trips in the week of May 24th, 2020. This surge could be a result of the events and protests happening in the city during that week, due to which the service was being used more. In general, the trend seems to be downward for the number of trips which can mainly be attributed to the fact that especially in Los Angeles, the pandemic has just become more and more serious without much improvement, so people are going out less and using the bike share system less. The system was often used by people for getting to work, but because companies shifted to a Work from home format, a number of regular users were lost. They do have some users who use the system for leisure purposes or simply to get some exercise. The company has started implementing safety precautions for the pandemic, and maybe for their marketing purposes they could emphasize on the precautions they're taking to bring in consumers and increase the number of trips.



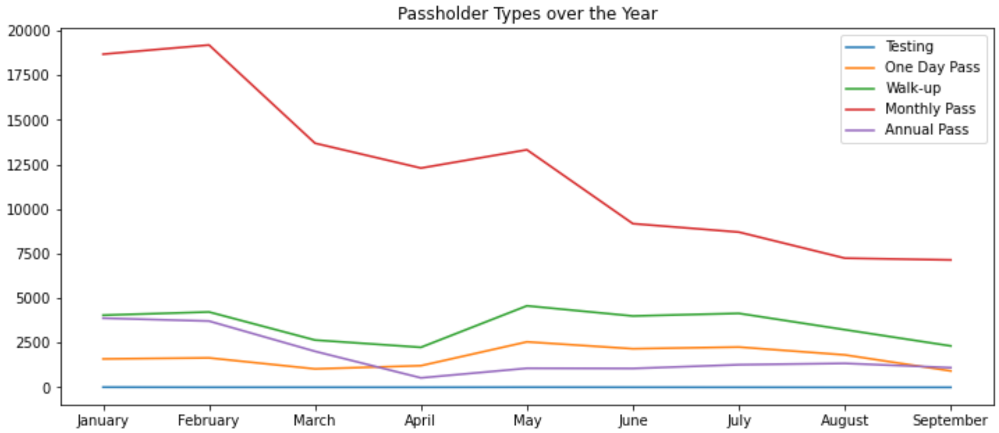
In general, from this histogram it seems like most days seem to have trips mainly in the range of 400 to 600, with only a few days in the higher range. This is possibly due to the fact that the pandemic took place this year and the number of trips per day were heavily influenced by that.

Analysis: Bike Types over the time period



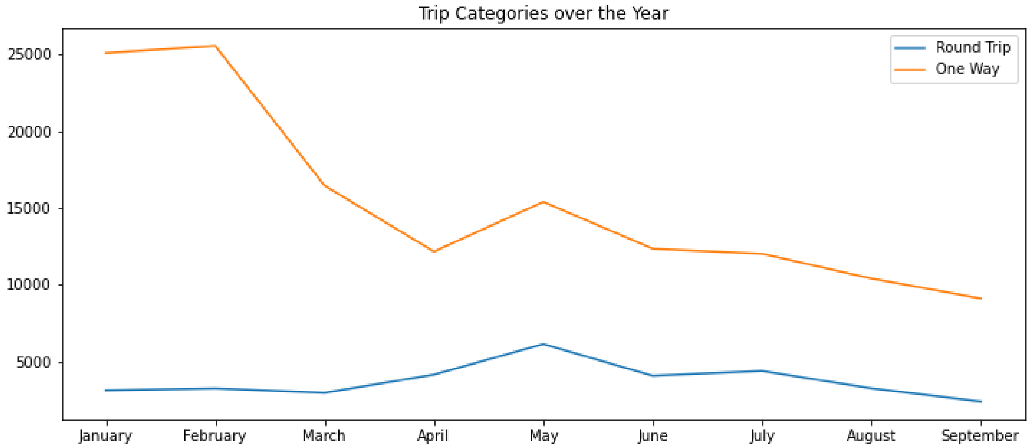
In terms of bike type, the electric and standard bikes are being used more compared to the smart bikes mainly because the smart bike rentals are more expensive compared to the other two. What’s interesting is that between standard and electric bikes, there are some months where the electric bikes seem to be used more and somewhere the standard bikes seem to be used more. For the electric bikes, the trend for usage seems to be downward, but in case of the standard bikes, there seems to rise after the month of May, and it ultimately overcomes the electric bike usage. For improvement in their usage for smart bikes, the company could possibly incentivize its usage to get more people to use it.

Analysis: Passholder Types over the time period



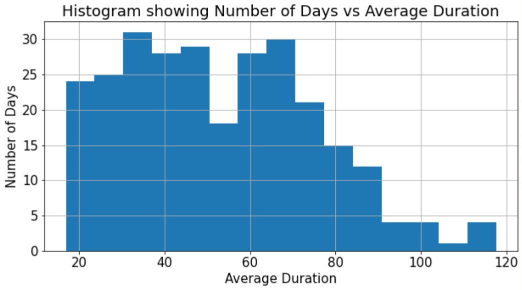
It seems that the Monthly Passes are the most popular of the passholder types and are used significantly more than the rest of the types. What is interesting is that in general, there is a downward trend noticed for all pass types, but there is a rise in the walk-up pass types. The drop is also significantly the most for the annual passes which may be due to the uncertainty in LA due to the pandemic, and people are less likely to make a long term commitment with their passes if they don’t plan on using it regularly. For future improvements, the company could focus on promoting the annual pass more.

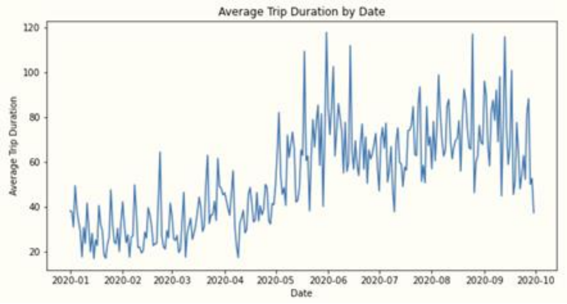
Analysis: Trip Categories over the time period



From this graph, it can be noticed for both trip categories, there is a downward trend. One Way Trips are noticeably used more than round trips, but what is interesting is that the graph for the one-way trips is very similar to the monthly trends for the number of trips. For future improvements, the company should focus on encouraging people to use their systems for round trips by offering possible incentives and promotions or focusing on the most popular routes.

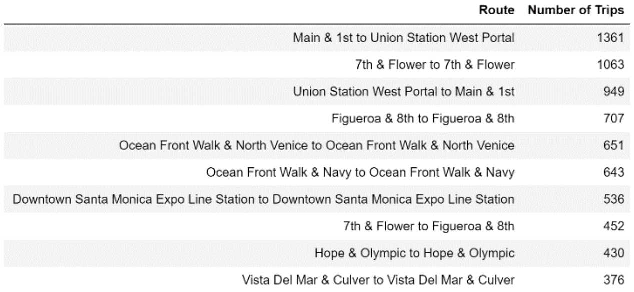
Analysis: Average Duration over time period





From these graphs, it can be seen that most days have an average duration of about 20 minutes to 60 minutes, and that could be noticed that the overall average duration noticed was 37.15 minutes. Also the interesting fact is that the average duration seems to have an upward trend and is increasing over time, despite the number of trips having a downward trend. This could be because people are indoors most of the time with the pandemic, so when they do use the system it's often for leisure purposes or exercises, so they try to prolong their overall time outdoors using the bikes, hence the duration increases. The company could focus on the physical exercise aspect provided by the bike rentals to bring in more consumers.

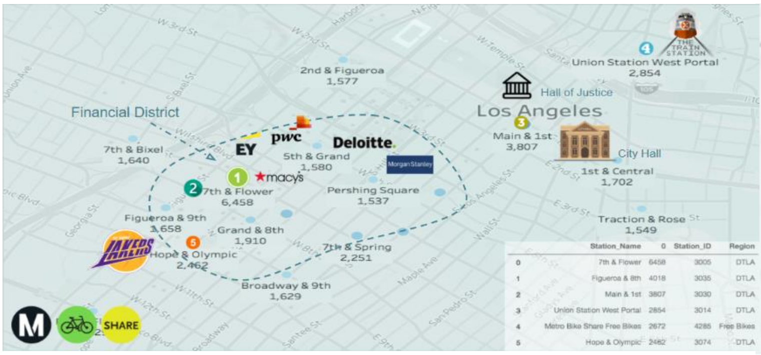
Analysis: Popular Routes

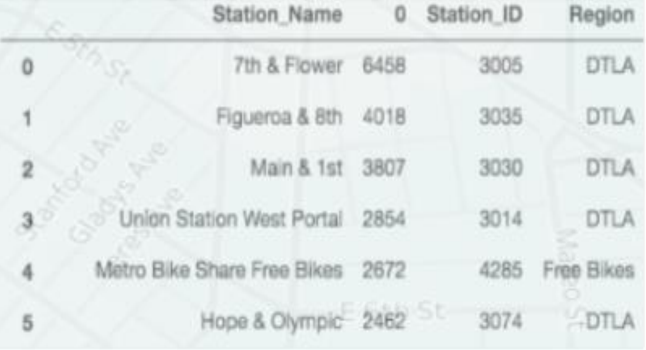


Route	Number of Trips
Main & 1st to Union Station West Portal	1361
7th & Flower to 7th & Flower	1063
Union Station West Portal to Main & 1st	949
Figueroa & 8th to Figueroa & 8th	707
Ocean Front Walk & North Venice to Ocean Front Walk & North Venice	651
Ocean Front Walk & Navy to Ocean Front Walk & Navy	643
Downtown Santa Monica Expo Line Station to Downtown Santa Monica Expo Line Station	536
7th & Flower to Figueroa & 8th	452
Hope & Olympic to Hope & Olympic	430
Vista Del Mar & Culver to Vista Del Mar & Culver	376

These are the Top 10 most popular routes noticed over the time period. What’s interesting is that many of these routes are round-trip routes, but the overall number of round trips as we noticed earlier was significantly lower than the one-ways. The company could potentially focus on using these routes more to bring more consumers through a possible competition and race element, this will engage users more and also bring in more users for the company.

Analysis: Popular Stations



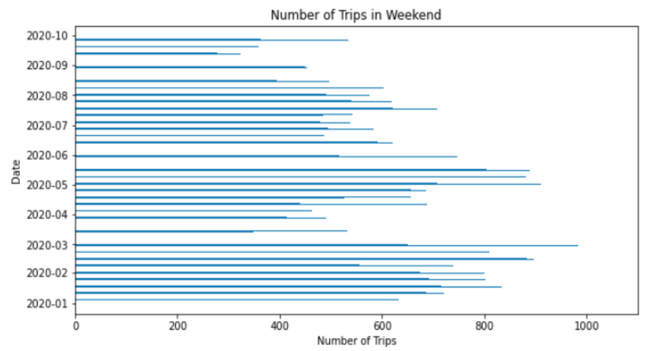


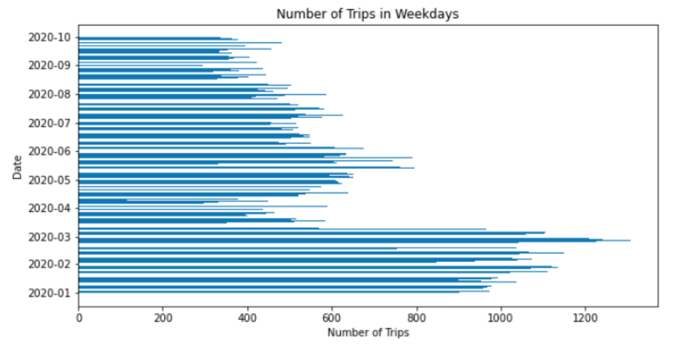
	Station_Name	0	Station_ID	Region
0	7th & Flower	6458	3005	DTLA
1	Figueroa & 8th	4018	3035	DTLA
2	Main & 1st	3807	3030	DTLA
3	Union Station West Portal	2854	3014	DTLA
4	Metro Bike Share Free Bikes	2672	4285	Free Bikes
5	Hope & Olympic	2462	3074	DTLA

The visual on the right side shows the top 6 most popular stations over the time period, to analyze what makes these stations so popular, we build a visual map (left side) based on the location of the stations. In the visual map, the common fact of these stations is that these stations are either located at business centers or near government agencies. In the future, the company could focus on setting up more stations near large commercial districts and companies to attract more customers. The company can also provide coupons, place advertisements, and provide free trial bikes near these popular stations to engage more potential users.

Analysis: Weekends VS. Weekdays

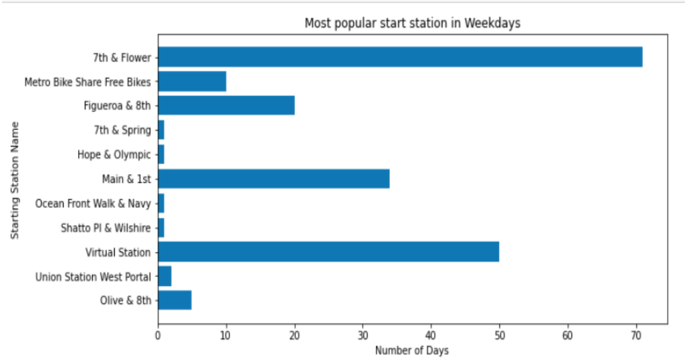
1. **Weekend VS. Weekdays: Number of Trips**

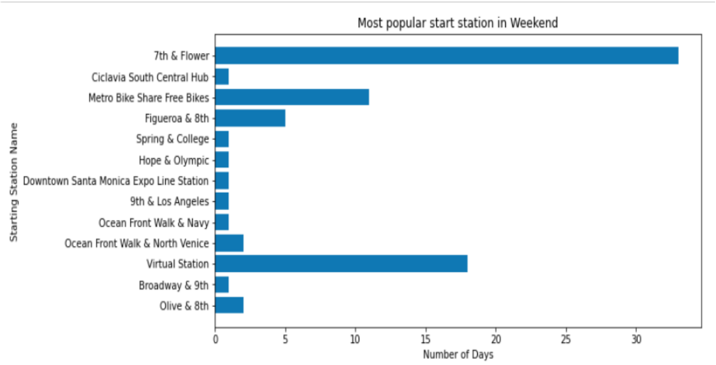




In order to analyze the characteristics of the number of trips in more detail, we compared the difference between the number of trips on weekdays and on weekends. From the bar charts above, the density of columns in the weekdays graph is obviously larger than that on weekends. taking a close look at the x-axis for both graphs, we can see that the total amount of trips on the weekdays is also more than the amount on the weekend. These can indicate that shared bikes are used more often on the weekdays.

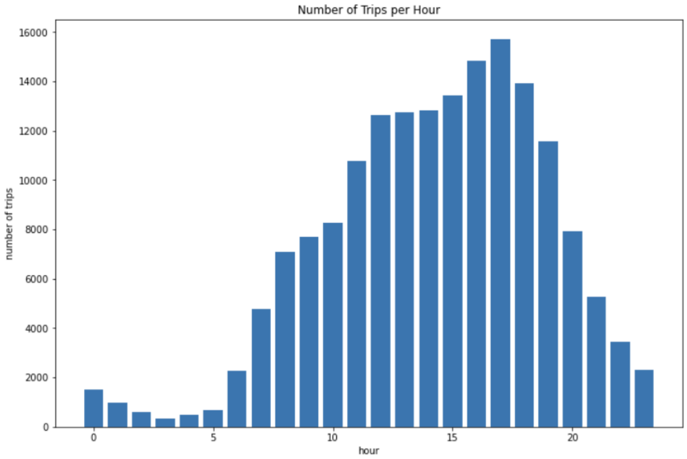
2. Weekend VS. Weekdays: Most popular station





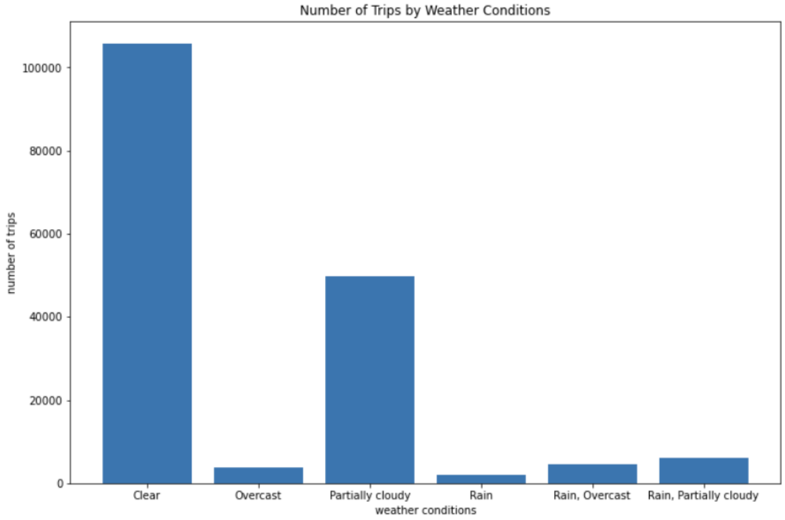
Exploring whether the popular stations are different on weekdays and weekends is another interesting topic. From the above bar charts we can see that 7th flower station is the most popular one for all time, because it is located in the central business district and it also has large shopping centers around it. Main and 1st station is popular on weekdays, because it is located around government agencies and on weekends those agencies are usually closed, so our potential users will decrease in a large scale during the weekend. Virtual station is on the second place for both weekdays and weekends, because the virtual station is the most flexible station, when the city has a big event or festival, different virtual stations will be created depending on demand.

Analysis: Hourly Activity

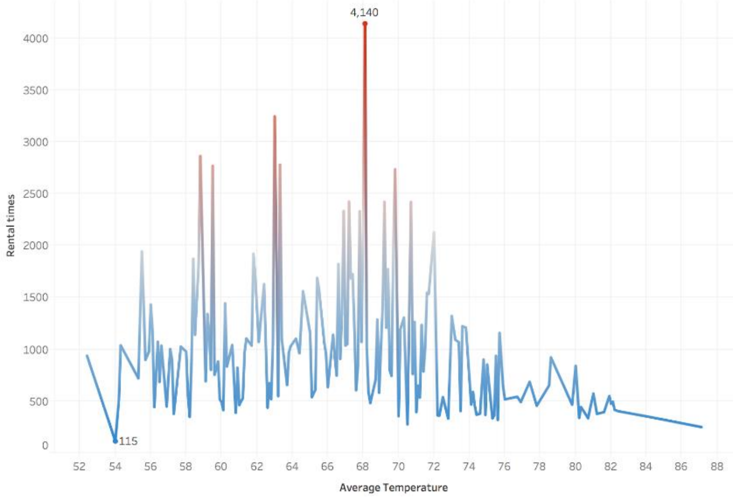


In terms of the number of trips per hour, our result shows that starting from 11 am, there is an increase in the use of the service. The most active time period is between 3pm to 6pm, with 5pm as the peak, which may be due to rush hours when people get out from work.

Analysis: Number of Trips by Weather Conditions and Temperature



For the number of trips by weather conditions, from the graph, we can see that most of the trips are on clear and partially cloudy days. The most active days are clear and the most least active days, which match with our predictions of people prefer to ride a bike in good weather than in bad weather days.

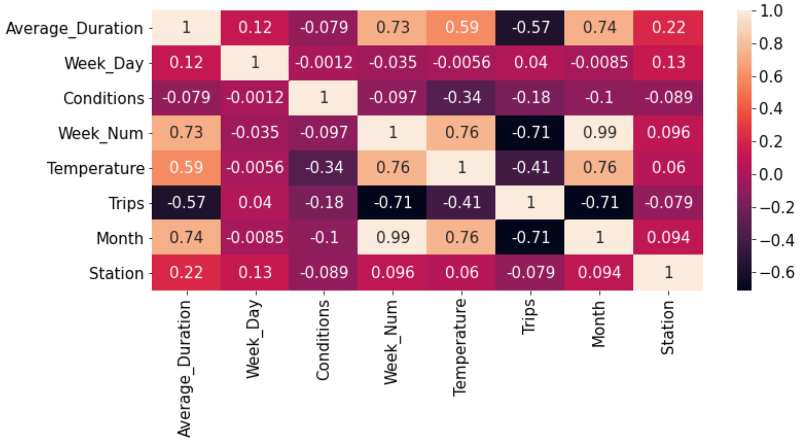


From the graph, we can see that most of the trips are on cooler days, in the range of high 50 degrees to mid 70 degrees. An interesting thing we can see is that when the temperature is higher than 68 degrees, there is a lower use of the service, which may be due to people not prefer to ride a bike in the heat.

Modelling

To quantify our observations and better understand what more information we need to understand what factors influence the number of trips started in a day, we decided to perform some machine learning models on the data we have.

The first thing we did is create a heatmap of how some important factors influence one and other, and this is the results we got:



It seems that the Trips variable is negatively correlated with all the variables except Week Day, the highest correlation being with the Week_Num and Month, followed by Average Duration and lowest being the Station variable. Based on this we decided to using the following variables in our modelling to predict the number of trips in a day:

- Average Duration: the average time bikes have been used in the day for which the number of trips is being predicted
- Month: the month number from the year of the day for which the number of trips is being predicted
- Temperature: average temperature of the day for which the number of trips is being predicted

We selected these attributes because:

- With average duration, we will be able to predict if there is a certain range of the bikes being used which brings in the most number of consumers in a day
- With month, better understand if certain months have some increases and focus marketing to those months and focus on events occurring in certain months to bring in more consumers
- With temperature, understand what temperature ranges brings in more consumers and rental trips

Linear Regression

We first ran a linear regression model with our data, and got the following results:

Linear Regression: Coefficient of Determination: **0.5502091001164389**

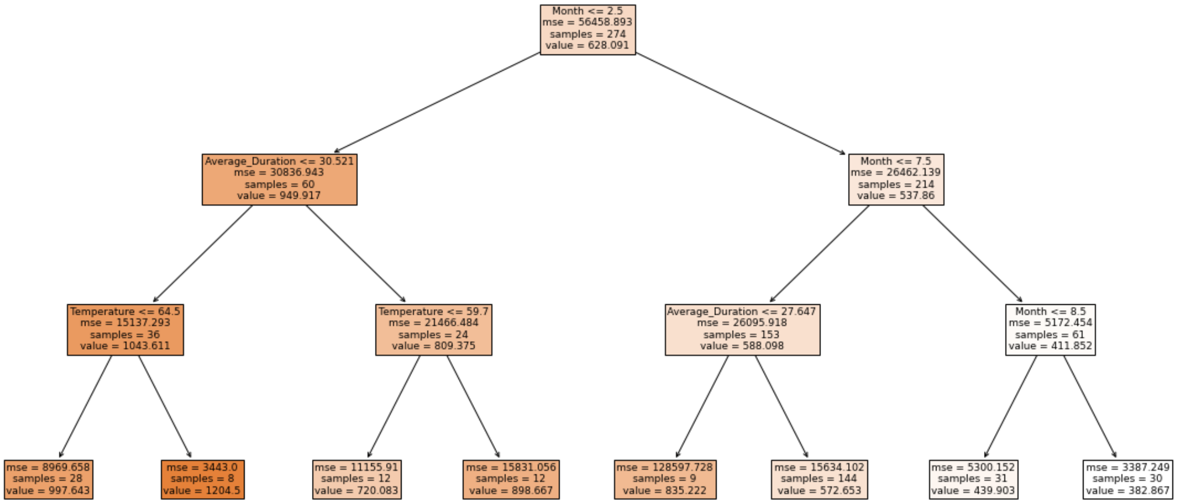
Linear Regression Equation:

Number of Trips = 380.1453386535637 - 1.31465495x - 78.26284064y + 10.54253354z

Where x = Average Duration, y = Month, z = Temperature

The R^2 value for this model isn't that high, so it doesn't seem to show strong variability in the data. We found the number of rows where the predicted value was within 10% of the actual number of trips that day, and it was 86 days of the 274 we noticed. In general, this model seems to be doing a below to average job in predicting the number of trips, so it would be useful to see another model.

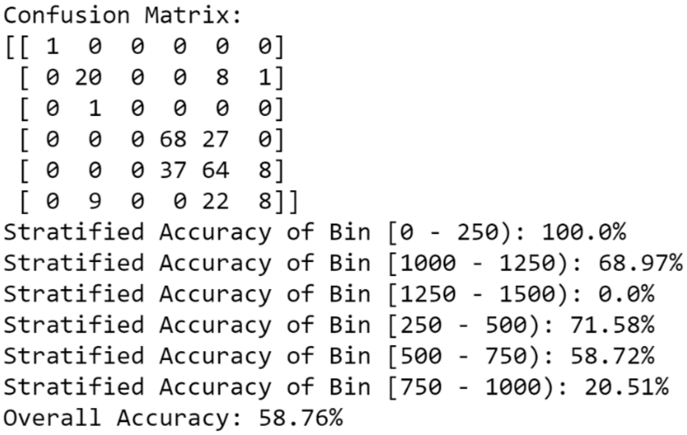
Decision Tree



This was the decision tree which we obtained, the first split seems to be on the Month variable followed by Average Duration, and then finally temperature. This shows that the maximum information gain comes from the Month variable pertaining to the number of trips, but the data may be somewhat biased due to the pandemic and the earlier months having more trips. It might be useful to compare this to data from previous years to notice the difference in trends. We found the number of rows where the predicted value was within 10% of the actual number of trips that day, and it was 38 days of the 274 we noticed. In general, this model seems to be doing a below to average job in predicting the number of trips, so it would be useful to see another model.

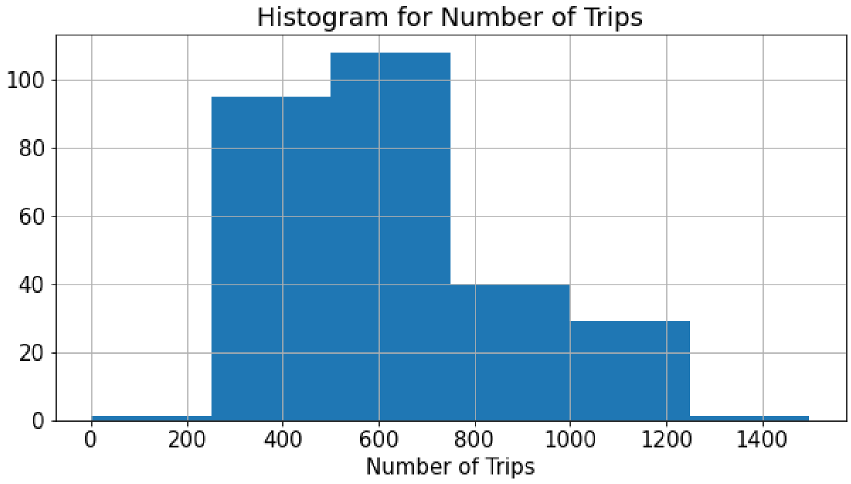
Logistic Regression

Based on our understanding for what kind of model might be the best fit for our data, we decided to run a logistic regression model. Before we did so, we also noticed that it might not be viable predicting the exact number of trips, so we decided to discretize the trips variable into bins, with each bin being of equal size of 250 trips. We ran two models one for the exact number of trips, and the other for the binned number of trips. For the exact number of trips our coefficient of determination was only 12.4%, whereas with the binning it was 58.76%, hence we decided to explore more with the binned attribute for number of trips. With this new model, 161 out 274 rows were accurately predicted with our logistic model. Our model was again based on Average Duration, Month and Temperature. This model seemed to have produced more accuracy compared to the other models we did, so we decided to dive in deeper and explore the accuracy for each bin.



```
Confusion Matrix:
[[ 1  0  0  0  0  0]
 [ 0 20  0  0  8  1]
 [ 0  1  0  0  0  0]
 [ 0  0  0 68 27  0]
 [ 0  0  0 37 64  8]
 [ 0  9  0  0 22  8]]
Stratified Accuracy of Bin [0 - 250): 100.0%
Stratified Accuracy of Bin [1000 - 1250): 68.97%
Stratified Accuracy of Bin [1250 - 1500): 0.0%
Stratified Accuracy of Bin [250 - 500): 71.58%
Stratified Accuracy of Bin [500 - 750): 58.72%
Stratified Accuracy of Bin [750 - 1000): 20.51%
Overall Accuracy: 58.76%
```

What can be noticed is that the lowest stratified accuracy is for the maximum range for number of trips, this could be due to the fact that there aren't many values for number of trips in that range, which can be in this histogram below for the number of trips. Contrary to that, for the minimum range, there seems to be a 100% accuracy again due to the fact that there is only 1 case for that range, it might be useful to maybe explore a different set of binning which is more evenly distributed in terms of frequency to get more results for all ranges.



Overall, the accuracy of this model is the highest from all the models we created and we would recommend using this model to predict the ranges of number of trips based on average duration, month, and temperature. For future analysis, we could maybe try different binning. We tried a variation of the attributes used for predicting, but these attributes gave the best results so we would recommend using them in future analysis as well.

Conclusions and Recommendations

For conclusions, firstly, we can see that the pandemic has a huge influence on the bike rental business. The Number of trips keeps decreasing since February with only one small peak in May. They should promote the different safety precautions they have taken to assure consumers and bring in more users. They could also focus on the exercise aspect of using these bikes, which most users are looking for during this pandemic when they have to spend most of their time indoors. They can measure whether their strategies are effective based on the number of trips and use our Logistic Regression model to see what attributes from duration, temperature and time of the year need to be focused on in order to bring a change in number of trips per day.

Secondly, most trips happen in days with the temperature between 50 degrees to 70 degrees. Besides, most popular stations on both Weekdays and Weekends tend to be located in the CBD of LA. For improvements, since most people tend to use the electronic bikes with one-day passes on weekdays, we should add more season pass and weekend promotions to give possible incentives for usage of smart bikes, round trips, and the purchase of annual passes to bring in more revenue from those outlets.

They should also consider adding a possible competitive element for users to race with cycles in most popular routes with incentives and prizes given to the winner is also a good way to attract more users. This will increase user engagement and incentivize usage of the rental program, bringing in more consumers and hence more revenue.

Limitations and Improvements

Our dataset has some limitations in terms of the data we lack. First of all, the dataset lacks user demographic information, like gender, age, job, with which we can do more analysis on bike users and learn about the reasonings for them to use the bike share system and get a better grasp of the demographic using the system. We also don’t have the price information and it is also essential for us to do the analysis in the financial domain, like revenue review. This data will allow us to better understand which revenue streams need to be improved and give us more in-depth data about what areas need to be focused on more. For future analysis, we would like to acquire this information to gain more valuable insights of the bike share business model and make even more conclusive statements and recommendations.

Bibliography

Chen, Anna. “Metro Bike Share LA Relief program offers reduced cost passes for the month of August.” *The Source - Metro*, 31 July 2020, https://thesource.metro.net/2020/07/31/metro-bike-share-la-relief-program-offers-reduced-cost-passes-for-the-month-of-august/.

Metro Bike Share - LA. “Data - Metro Bike Share LA.” *Metro Bike Share LA*, https://bikeshare.metro.net/about/data/.

Metro Bike Share - LA. “Official Website.” *Metro Bike Share - LA*, https://bikeshare.metro.net/.

Visual Crossing. “Weather API.” *Visual Crossing API*, https://www.visualcrossing.com/weather-api.

Visual Crossing. “Weather API Usage Sample.” *Github*, Visual Crossing, https://github.com/visualcrossing/WeatherApi/blob/master/python_samples/loading_historical_weather_data_into_python.py.