Divvy Bike System

ISQS 4370-Data Mining

Fall 2020

Created by:

Aileen Morales

Braxton Parrott

Hemant Sharma

Analyzing Decision Making in the Divvy Bike System

1.) Introduction

About Bike Sharing System

Divvy is Chicago's bike sharing system; it is a fun and affordable way to get around town. The customer is able to pick up a bike at one of hundreds of stations around the Chicago area. Bikes are used to commute to work/school, run errands, explore the city and so much more. The ride ends by the customer returning the bike to any station around town. Divvy has over 600 stations with approximately 6,000+ bikes.

The single ride is just $3 and includes 30 minutes of ride time to get you anywhere you need to go. If a customer needs to ride longer, they are able to keep the bike out for more than 30 min and they will be charged an extra $0.15/ minute. This rate can be avoided through the purchase of a premium subscription service, which replaces the per-ride payment type to an annual fee type of payment. Users who find themselves using the service frequently might be inclined to purchase the premium service as a long-term based decision in order to save money. Free trials and such are also offered to entice people to try out the service who might not have otherwise would.

One may be wondering how they can get access to the service. The best way to get access to the service is by downloading the 'Divvy' app and directly buying a pass there. Afterwards, the customer will need to find a 'Divvy' bike, scan its QR code to unlock and then be on its way in minutes. The customer can also keep in mind that a smartphone is not necessary to make a purchase. Passes can also be purchased at any 'Divvy' kiosk station, but should also be aware that if they incur any usage fees, their card will be charged. The fee is $1200 if the customer were to lose or steal a bike from the company, however; the case would be investigated thoroughly to make sure the client is truly at fault for damages.

a.) Business Problem

The business problem we wish to investigate is the correlation between the use of Divvy products by members vs. the use of Divvy products by single ride paying customers or customers who purchased a day pass.

Though part of the focus for Divvy is to expand their outreach with the introduction of electric bikes, another issue they must focus on is what type of customers are using their bikes & how often are their bikes being used by a single customer.

b.) Analytics Problem

In order to address the rate of success for the Divvy bikes being used, we will be analyzing the daily patterns of how many users and what type of users on a daily basis use Divvy, what distance they travel, how much time it takes them to travel that distance & how much possible foot traffic passes by their stations. To be put simply, what factors and/or travel routines might lead to an individual becoming a premium member of our service?

Based on the analyses, we will be working to provide the best possible solution to Divvy about what locations they should expand their operations, to bring the best return on investment. Along with each analysis we will be providing certain steps on how to best target the issues for the best solution possible.

2.) Data Description

Techniques Used

We will be utilizing a dataset obtained from the Divvy site itself in order to arrive at the results needed to formulate possible solutions.

To analyze the data properly, we decided to use several visualization and data analyzing tools such as Tableau and Python to understand the correlations between variables and our analytics problems. While our team used Tableau to visualize the data, we chose to use Python to perform multiple linear regressions as well as sort each characteristic through the use of classification trees.

Impactful and New Variables

We created new variables in the original database titled *"miles"* and *"minutes"* that calculate the distance traveled in each trip by utilizing the *"start_lat"*, *"start_lng" "end_lat"* and *“end_lng”* columns, and the time taken in minutes by utilizing the *“started_at”* and *“ended_at”* columns (respectively). These variables aided in the construction of meaningful visualizations by organizing data into a format that is easier to use by calculating the information conveyed by the data and compiling it into single columns. Note: *“start_lat,” “start_lng,” “end_lat,” “end_lng”* were omitted due to their data being better organized by the *“miles”* variable. All impactful variables are shown below:

“Ride_id”: The ID assigned to a ride/trip. Can be utilized to gain a count of the total number of rides.

“Rideable_type”: Identifies the type of vehicle (electric/basic) that was used in the trip.

“Started_at”: The start time of the trip, organized in M:D:YYYY H:MM formatting.

“Ended_at”: The end time of the trip, organized in M:D:YYYY H:MM formatting.

“Start_station”: The geographically accurate name (location) of the station at which a user began a trip.

“Start_station_id”: The unique ID assigned to the station at which a user began a trip.

“End_station”: The geographically accurate name (location) of the station at which a user began a trip

“End_station_id”: The unique ID assigned to the station at which a user began a trip.

“Member_casual”: Identifies whether the user of the trip was a member or non-member.

“Miles”: The total distance traveled in the trip, found from the coordinate variables.

“Minutes”: The total time taken in minutes during the trip, found from the start and end time variables.

Reason behind predictor selecting

We considered our intended predicted variable: “member_casual.” The variable included values as either “member,” which held that a trip was taken by a member of Divvy’s Premium Membership Program, or a “casual,” which held that a trip was taken by an individual not under Premium Membership. We feel as though the contrast between the tendencies of each type of user will allow us to gauge what factors and routines might entice an individual to purchase the Premium Membership.

3.) Analyses Methodology

Regarding the dataset that was provided to us, the topic that we are mainly focusing on would be on the decision making of the 'Divvy Bike System' customers. We will be taking into consideration the location of the customers and the time frame the bicycle was used for each customer whether they are considered 'Member' or 'Casual.'

In order to analyze this problem our team decided to conduct variable visualizations, sentimental analysis, and classifying the analysis. Using variable visualizations will help uncover a whole new dimension of underlying information within our dataset. The sentimental analysis will assist in decoding the data and extract subjective information in source material and help understand the social sentiment of the 'Divvy' service. Lastly, classifying the analysis will aid our team in assigning categories to the collection of the data to allow for the data analyzation to be more accurate.

a.) Variable Visualization

Moving forward to the description of our visualizations, a map ***[Figure 1]*** is able to show geographical "hotspots" in the city of Chicago to pinpoint where specifically the highest concentration of ride is found. In addition to pinpointing the locations of individual rides, the distance traveled during the ride is also portrayed via circle size and color, which shows how frequency of rides might relate to distance of individual rides. The highest amount of concentration can be found in the mid-east of the city, with higher miles-driven closer to the outskirts of the city.

The bar chart ***[Figure 2]*** showcases the comparison between members of the premium service versus non-members (casual user) of the service. Specifically, the average miles driven and minutes used by each party are compared in an effort to convey what traits each user type has when it comes to usage. Surprisingly, casual members have higher average miles driven and average minutes taken; it is theorized that this is the case due to premium members having a more streamlined approach to their usage with the service, as well as being more versed with the intricacies of the service.

The box chart ***[Figure 3]*** shows the top 50 locations for users to begin a ride at (the sample size is reduced for clarity and simplicity). This chart works in addition with the map chart

[Figure 1] to help more accurately pinpoint what areas get high amounts of usage in comparison to other locations. These high usage locations may be areas of interest for attempts at converting nonmember users to members or to avoid marketing too highly in these areas in an attempt to branch out to areas of lower usage. According to this chart, out of all stations, the Millennium Park station is the station where the greatest number of riders begin their journey(s).

[Figure 4] is able to portray that there is a vast difference on deciding whether to use a docked or electric bike for premium members. According to our chart premium members usually picked docked bikes. This information does not necessarily mean that the premium members have a preference for docked bikes. Ridership for users has a possibility of being more focused on the central business district of Chicago and there are more stations to dock the bikes compared to electric bikes. By the same token, we arrived to the conclusion that premium members may live in the popular area of Chicago.

A box-and-whisker plot ***[Figure 5]*** helps display the distribution of most popular locations and average miles for each customer. According to the box-and-whisker plot a popular start station for customers is Ashland Ave & Blackhawk St with an Avg. Mile of 1.517 and 133 Discount of Start Station. Neighborhood scout states that this area is a densely urban neighborhood(based on population) located in Chicago, Illinois. Rent is also currently lower in price compared to 42.2% of Illinois neighborhoods in ***[Figure 6].***

The packed bubbles show the details about a ‘Casual’ and ‘Premium’ member ***[Figure 7]***. The sizes of the bubbles portray the count of miles and the color green and blue show the difference between the two. As expected, ‘Premium’ members have more miles compared to its counterpart. Premium members may also gain health benefits since they are on the bike longer. Benefits include low impact which causes less strain and injuries than most other forms of exercise. It is also good for strength and stamina because cycling increases stamina, strength, and aerobic fitness. Cycling can also be time efficient because since Chicago is heavily populated, members will not have to worry so much about traffic. As a matter of fact, this can also help reduce one’s carbon footprint. We are in an era where one’s carbon footprint is higher than other eras and cycling can assist with this dilemma.

b.) Sentiment Analysis

Our team created new variables for the ‘Divvy Data set’. The variables that were created to better understand our raw data were “Miles” and “Member”. The miles category were used to see the average mileage of each customer. In addition, the member category was used to see if the customer was a ‘casual’ or ‘premium’ customer. According to our visualizations we were able to create meaningful relationships. These categories helped depict whether being a casual or premium may make a greater difference in the mileage created in every ride.

Multiple Linear Regression Analysis

We conducted an analysis of all available variables found in our dataset prior to conducting our regression analysis and formulation of a regression model in order to confirm that such variables are accurately represented by their assigned data type, as well as to gain a better understanding about what data we are working with.

We utilized Python within our regression models to extract quantitative data, which means that data pertaining to *started_at, ended_at, ride_id, and type variables* was omitted. Rows with missing values were also omitted as they skewed the final results. Prior to the omission of the datasets presented in ***[Figure 8],*** we can concur that the conditional number is significantly large for the dataset being compared, as opposed to the results in ***[Figure 9]*** post omission of the insignificant data where the conditional number is fairly reasonable and helps provide a better insight into what data should be considered by Divvy to bring the maximum change in their business strategy to expand.

d.) Classification Analysis

Given the data collected by Divvy whenever any of their bikes are used, we decided to add our own data analysis to calculate the miles & minutes each bike was driven for, from one destination to another. The rationale behind this is, that if Divvy is able to identify a possible pattern in the miles each bike is driven, how often it’s driven & for the amount of time its driven, they can look into expanding their stations to vicinities near those areas, giving more people more options for riding a Divvy bike.

In order to ascertain the factors influencing the use of Divvy bikes, our team chose to utilize classification trees for the most accurate data possible. Since a lot of the data in our dataset is continuous, we chose to use the “*member_casual*” column to create a classification tree which will provide an insight into what type of users (*casual or members*) are using the bike and exactly how many miles have they driven as presented in [***Tree Iteration 1-2]***.

This will help us to understand which user types are more active on the Divvy rides, and how to possibly attract more of the same user types or how to attract the user types whom they believe could help to increase their revenue intake in the long run.

4.)Impact & Implementation

a.) Limitations

One of the few challenges we faced while preparing the data visualizations and analysis included the confusion of how to correlate the data with each other for the best possible result, which could give Divvy an idea of what changes they should bring within the functionality model to provide the best output.

Another major constraint we had in our dataset was the exclusion of revenue generated details per each ride taken by users. Had this data been included into the dataset, we could have had a better view of what practices must Divvy change or even continue to maintain or possibly increase the revenue generated. To tackle this issue, we simply chose to focus on the *"member_casual"* column in relation to *"miles"* driven. The sole purpose of this was to get a rough estimate of the usage habits per person, and how that could possibly affect the revenue generated for Divvy.

As we outlined in sections above, since we didn't have many options to correlate the data to, we chose to incorporate the use of *"started_at", "ended_at"* and create an additional data column for *"minutes"* to get an idea of the distance from one station to another in time. Similarly, to incorporate the use of *"start_lat", "end_lat"* and *"start_lng", "end_lng"* we chose to add an additional column to calculate the proper distance in *"miles"*.

b.) Recommendations to Divvy

The findings report that if an individual frequently travels a distance of roughly a mile or more, their personal cost-benefit analysis would lead them to the conclusion that they would be better off using the membership service rather than to continue to use the casual service in order to save money. Casual users who fit the role of this type of individual are users that should be focused on through advertising (whether it be discounting, typical advertising, or presenting information that might lead them to believe they are better off purchasing a membership).

In addition to this insight and possible solution, we may find that offering special discounts in “hotspot” locations might boost premium membership sales by pulling in users who previously might not have been motivated enough to purchase the membership. The company will be able to utilize this information to create geographically targeted advertising campaigns suited to user types based on the use-demographics shown in the results.

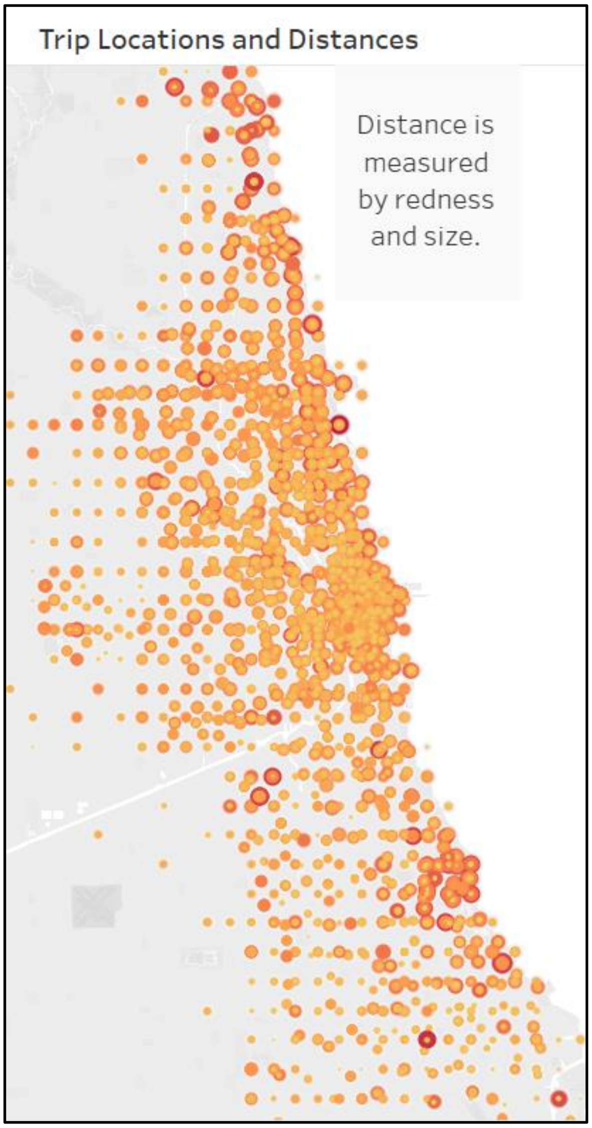
Though Divvy’s partnership with Lyft has significantly uplifted their brand and image in society, I believe they could further their reach to people by diversifying their options for rides being offered, similar to LIME; as seen in several cities, LIME first started out with just the LIME scooters and soon added the LIME electric bikes to their portfolio; or maybe follow in the footsteps of providing similar services as ZipCar. But a more unconventional idea which Divvy could add to their business strategy to build their portfolio could the the fusion of business strategies followed by both, LYFT and ZipCar; by this I mean they could offer ride services such as access to electric scooters and bikes similar to LIME, all while providing access to rental cars for further travelling distances similar to ZipCar.

If Divvy were to pick up all or either of those options in the passage above, they could improve their business model and portfolio significantly, all while expanding and diversifying their clientele and presence in the market.

5.) Conclusion

All techniques utilized in the research project allowed for a better comprehension of our business problem and formulation of possible solutions to solve it. Information gathered pertained to characteristics and tendencies of members versus casual users, the level of use of each party, and the locational “hotspots” of high usage within the city of Chicago. This information will be put forth into attempts at trying to increase Premium Member numbers and, in turn, revenue. Geographical approaches to increasing consumer base are only effective when you know the correct location to pour efforts into, and our findings present such location(s).

Specifically, insights gathered from the information presented allows us to know what travel tendencies of individuals might make them more likely to purchase a membership, as well as the types of individuals who benefit heavily from the membership. This information will ultimately allow the company to reduce the amount of marketing "slack" and will be better suited to target efforts at a more specific user demographic.


Trip Locations and Distances
Distance is measured by redness and size.

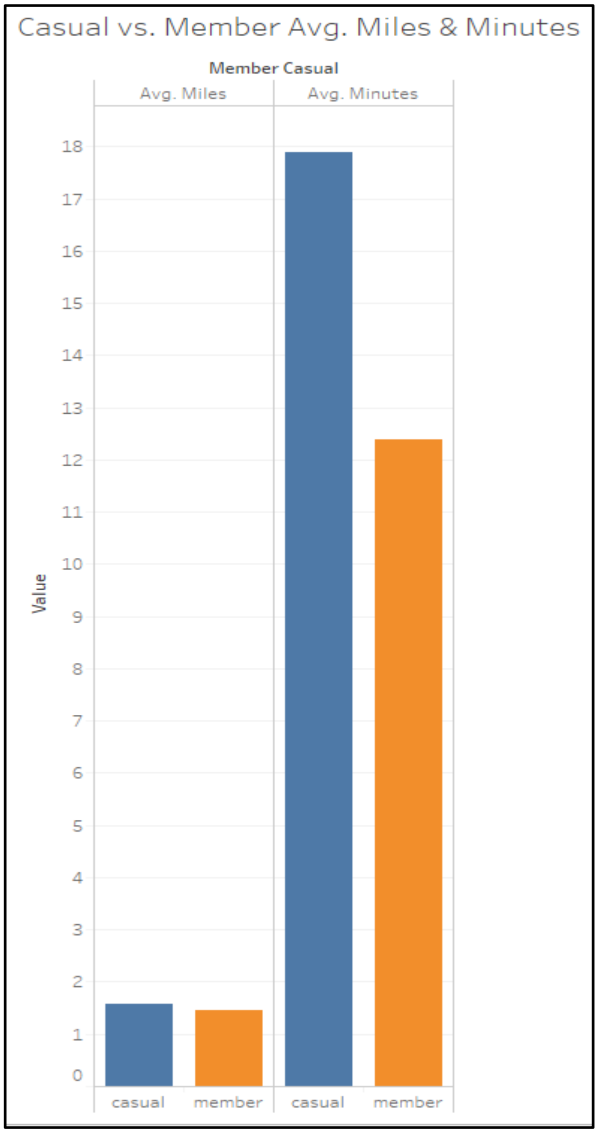


Figure 1 and 2 show the geographical locations of ride as well as distance traveled, and the average miles traveled and minutes used for members and non-members (respectively).

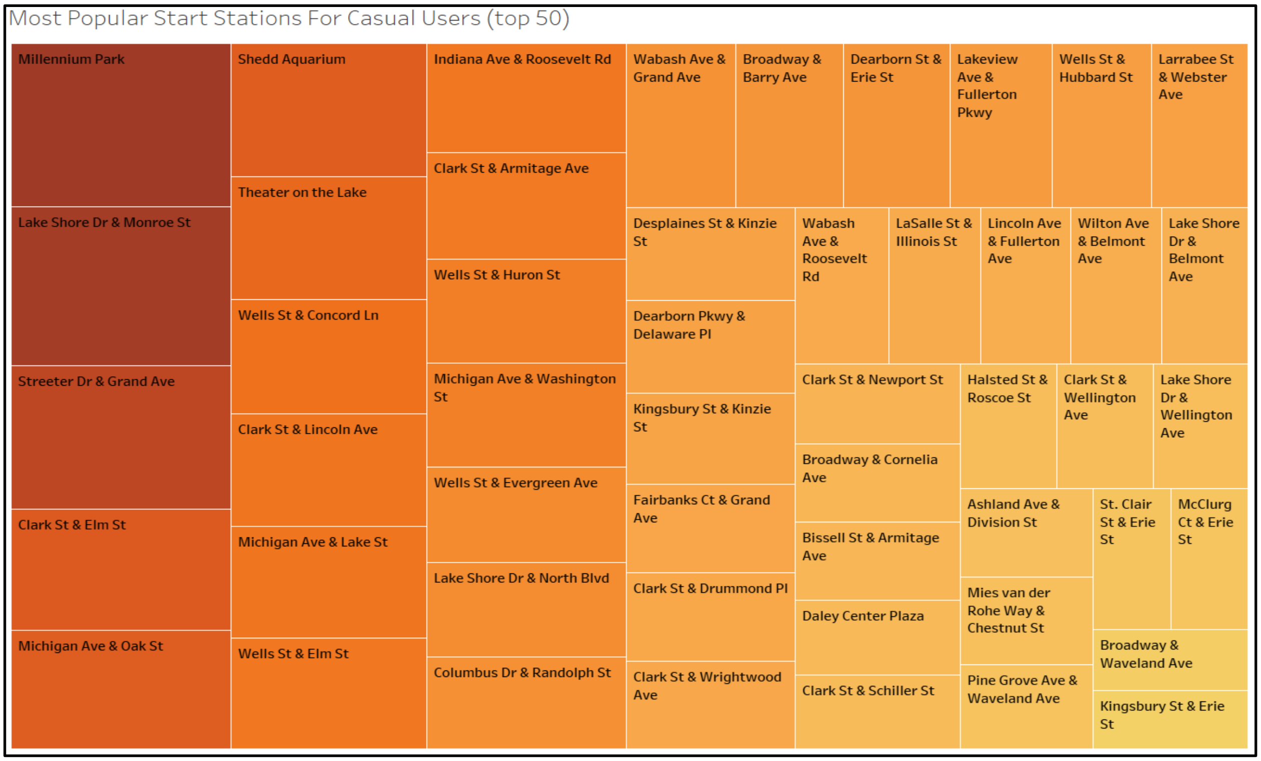

Most Popular Start Stations For Casual Users (top 50)

Figure 3: Organizes Station by a count of riderID’s that used them (high to low)

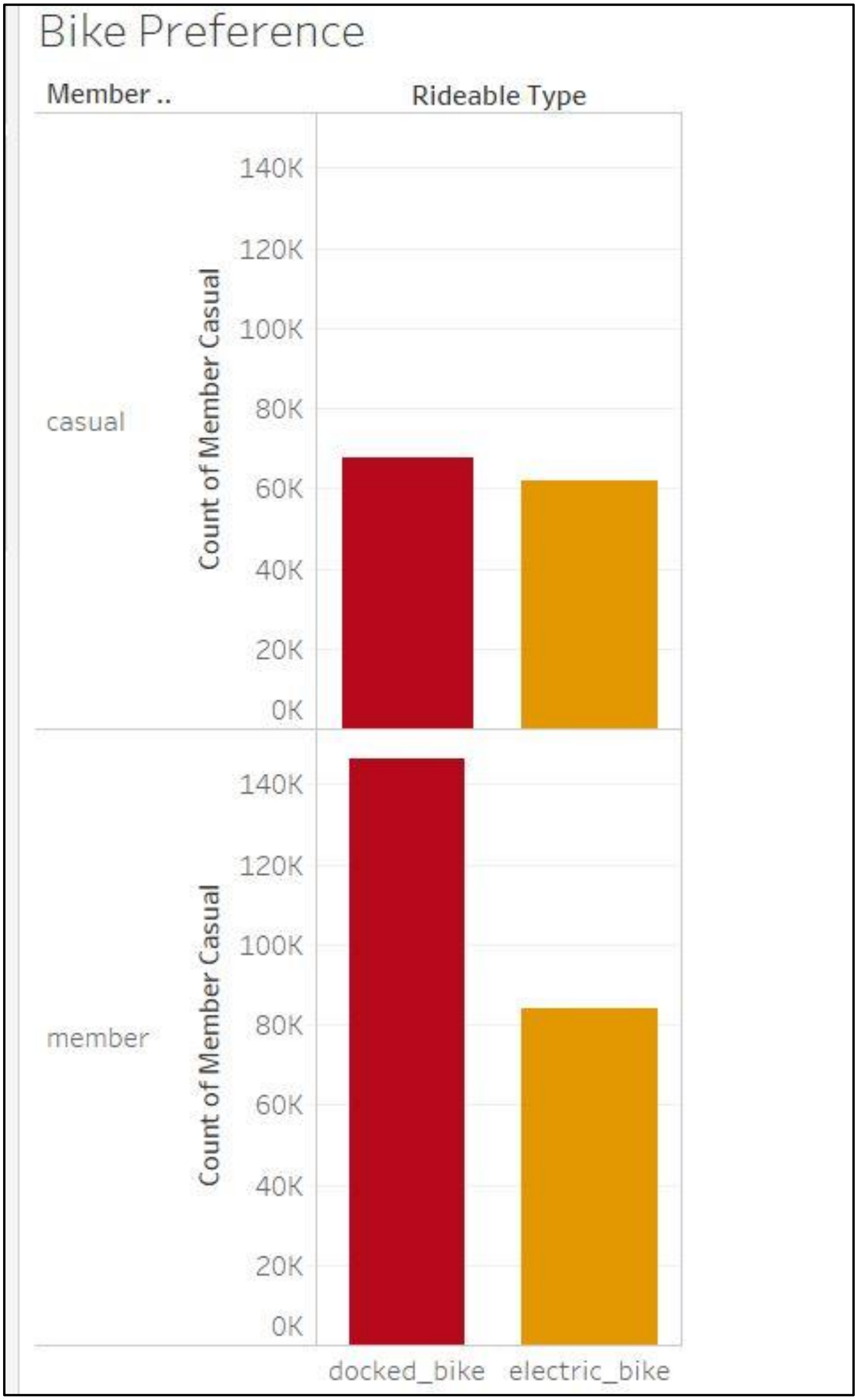

Bike Preference
Member ..
Rideable Type
casual
member
Count of Member Casual
Count of Member Casual
140K
120K
100K
80K
60K
40K
20K
0K
140K
120K
100K
80K
60K
40K
20K
0K
docked_bike
electric_bike

Figure 4

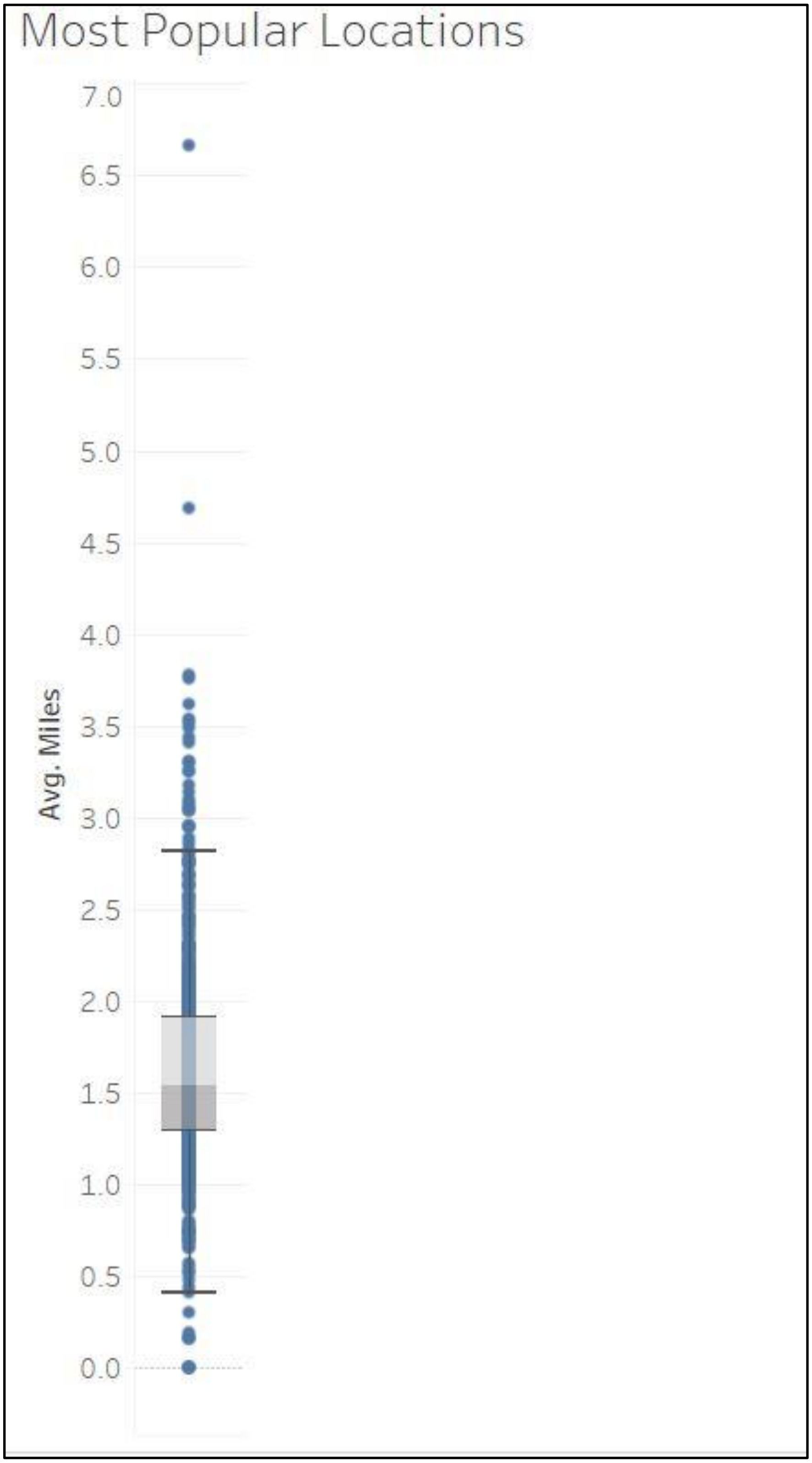

Most Popular Locations
Avg. Miles
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

Figure 5

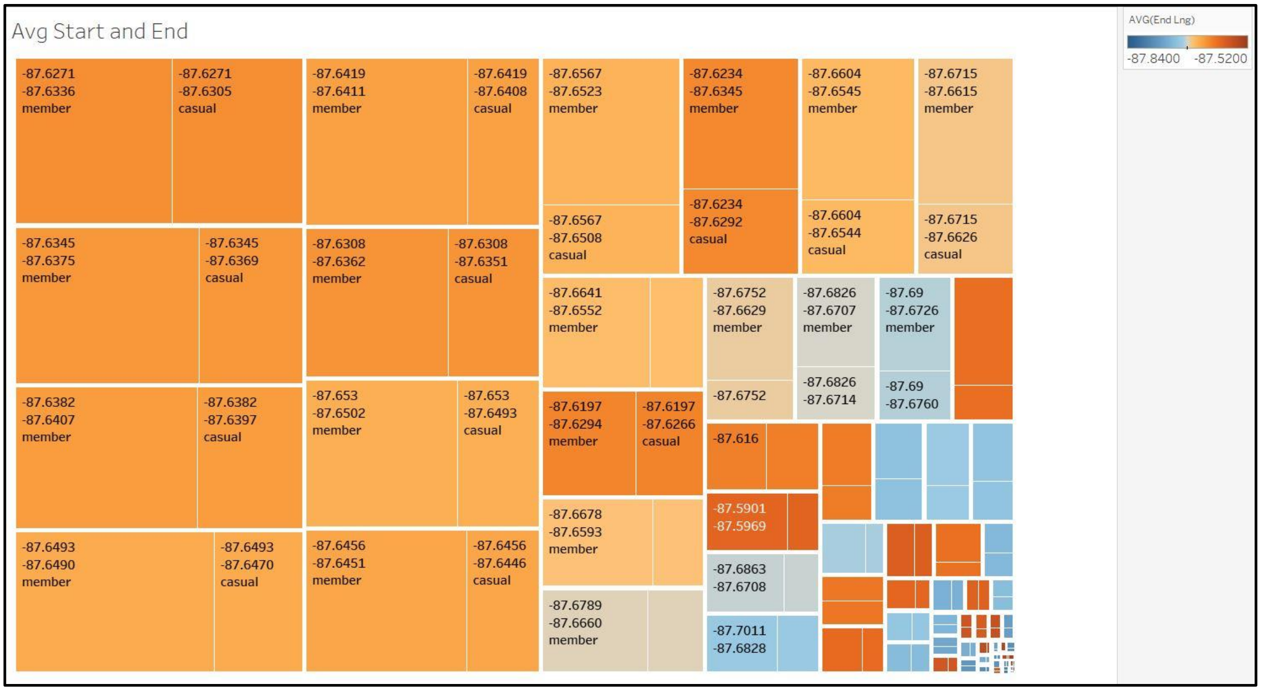

Avg Start and End
AVG(End Lng)
-87.8400
-87.5200
-87.6271
-87.6336
member
-87.6271
-87.6305
casual
-87.6419
-87.6411
member
-87.6419
-87.6408
casual
-87.6567
-87.6523
member
-87.6234
-87.6345
member
-87.6604
-87.6545
member
-87.6715
-87.6615
member
-87.6567
-87.6508
casual
-87.6234
-87.6292
casual
-87.6604
-87.6544
casual
-87.6715
-87.6626
casual
-87.6345
-87.6375
member
-87.6345
-87.6369
casual
-87.6308
-87.6362
member
-87.6308
-87.6351
casual
-87.6641
-87.6552
member
-87.6752
-87.6629
member
-87.6826
-87.6707
member
-87.69
-87.6726
member
-87.6752
-87.6826
-87.6714
-87.69
-87.6760
-87.6382
-87.6407
member
-87.6382
-87.6397
casual
-87.653
-87.6502
member
-87.653
-87.6493
casual
-87.6197
-87.6294
member
-87.6197
-87.6266
casual
-87.616
-87.6678
-87.6593
member
-87.5901
-87.5969
-87.6863
-87.6708
-87.6493
-87.6490
member
-87.6493
-87.6470
casual
-87.6456
-87.6451
member
-87.6456
-87.6446
casual
-87.6789
-87.6660
member
-87.7011
-87.6828

Figure 6

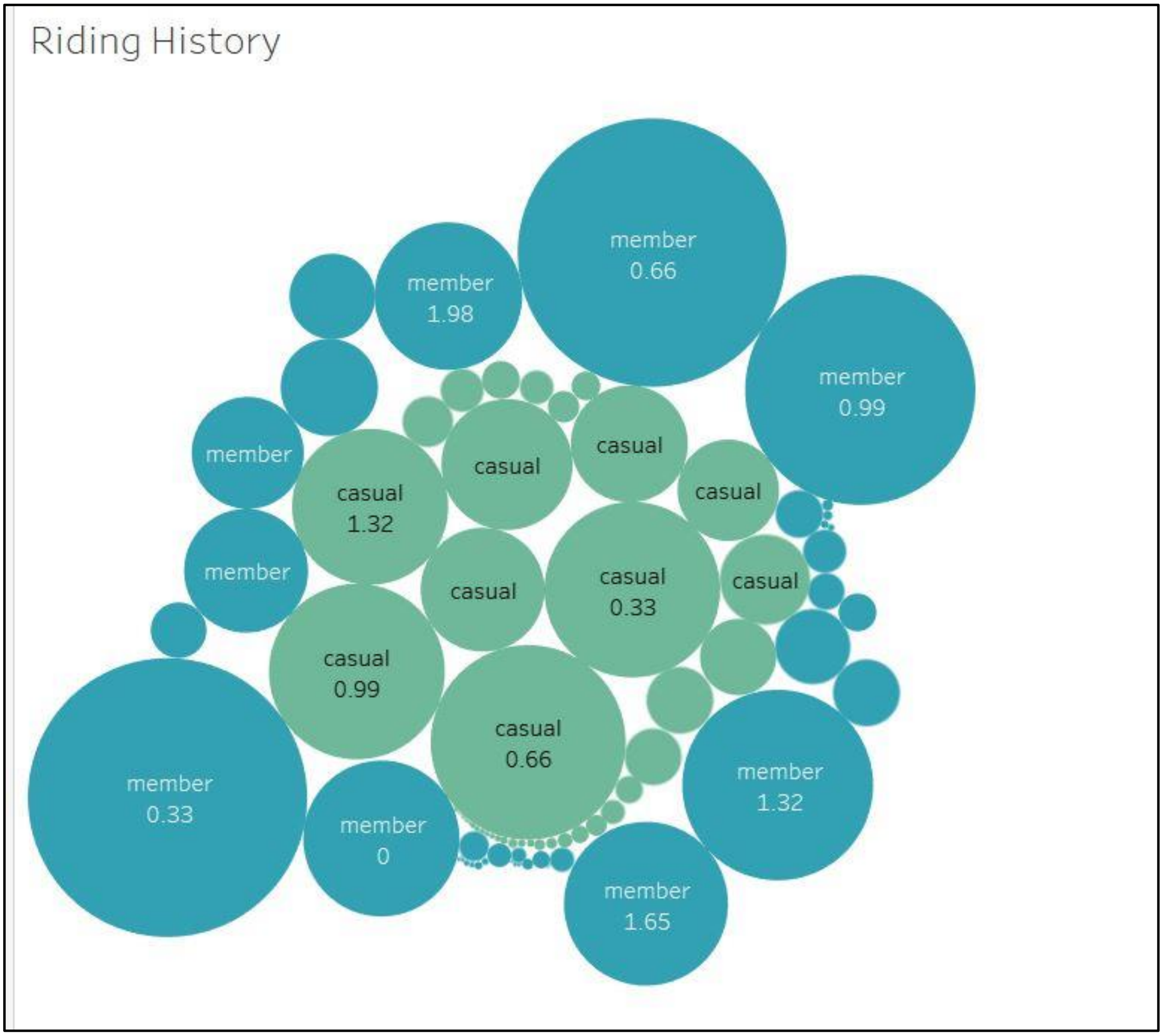
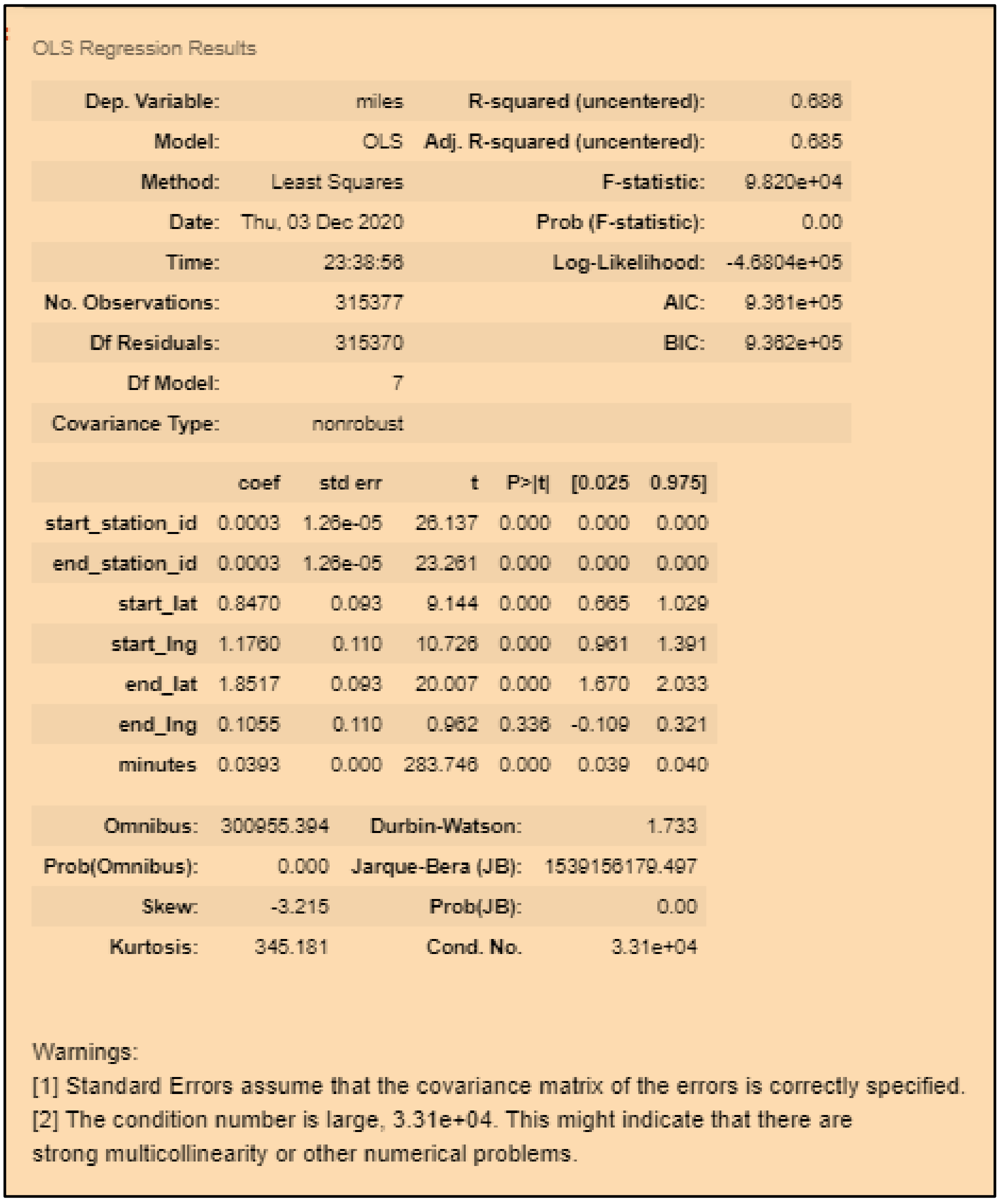

Riding History
member
0.66
member
1.98
member
0.99
member
casual
casual
casual
casual
1.32
casual
member
casual
casual
0.33
casual
casual
0.99
casual
0.66
member
0.33
member
0
member
1.32
member
1.65

Figure 7

First Regression



OLS Regression Results

Dep. Variable:	miles	R-squared (uncentered):	0.686
Model:	OLS	Adj. R-squared (uncentered):	0.685
Method:	Least Squares	F-statistic:	9.820e+04
Date:	Thu, 03 Dec 2020	Prob (F-statistic):	0.00
Time:	23:38:56	Log-Likelihood:	-4.6804e+05
No. Observations:	315377	AIC:	9.361e+05
Df Residuals:	315370	BIC:	9.362e+05
Df Model:	7		
Covariance Type:	nonrobust		

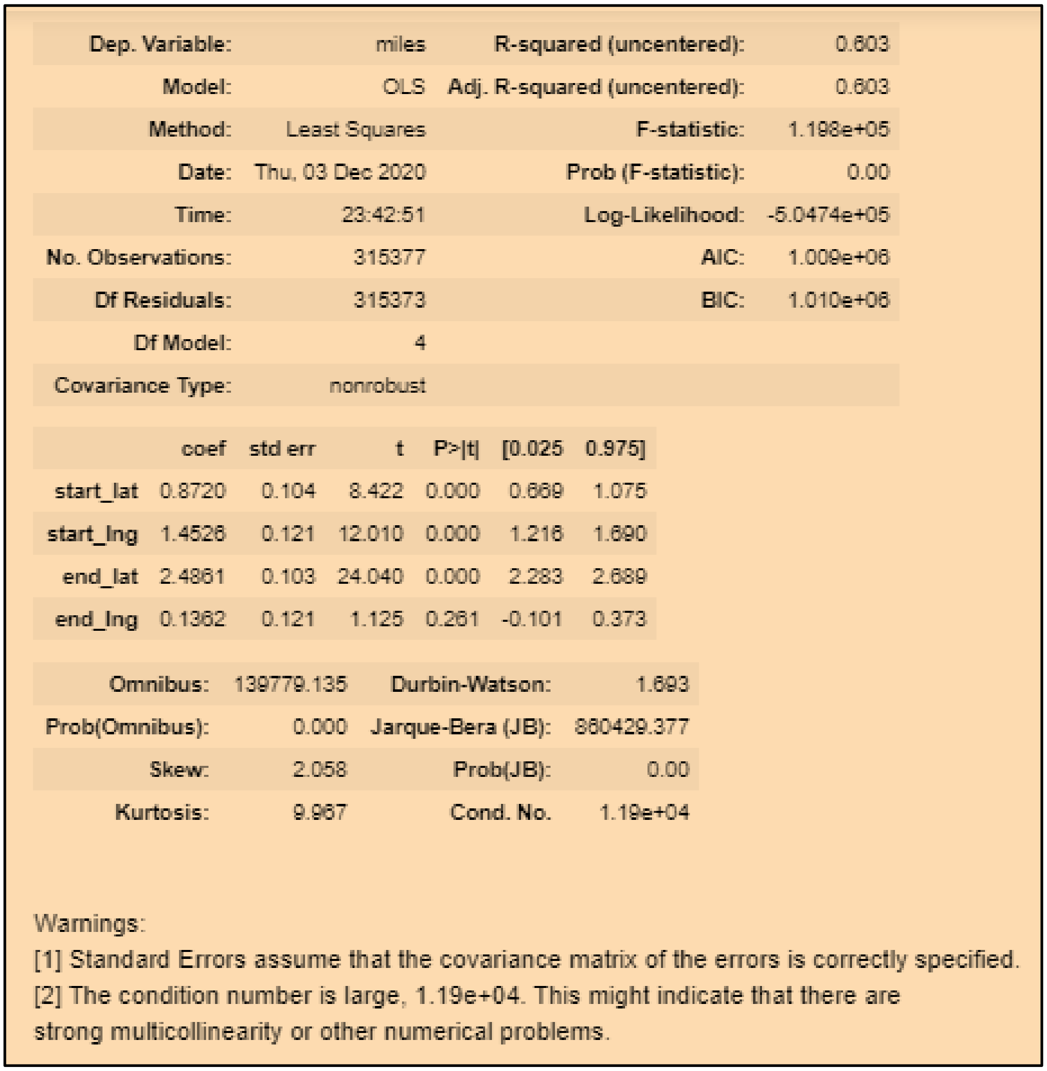
	coef	std err	t	P>|t|	[0.025	0.975]
start_station_id	0.0003	1.26e-05	26.137	0.000	0.000	0.000
end_station_id	0.0003	1.26e-05	23.261	0.000	0.000	0.000
start_lat	0.8470	0.093	9.144	0.000	0.665	1.029
start_lng	1.1760	0.110	10.726	0.000	0.961	1.391
end_lat	1.8517	0.093	20.007	0.000	1.670	2.033
end_lng	0.1055	0.110	0.962	0.336	-0.109	0.321
minutes	0.0393	0.000	283.746	0.000	0.039	0.040

Omnibus:	300955.394	Durbin-Watson:	1.733
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1539156179.497
Skew:	-3.215	Prob(JB):	0.00
Kurtosis:	345.181	Cond. No.	3.31e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.31e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

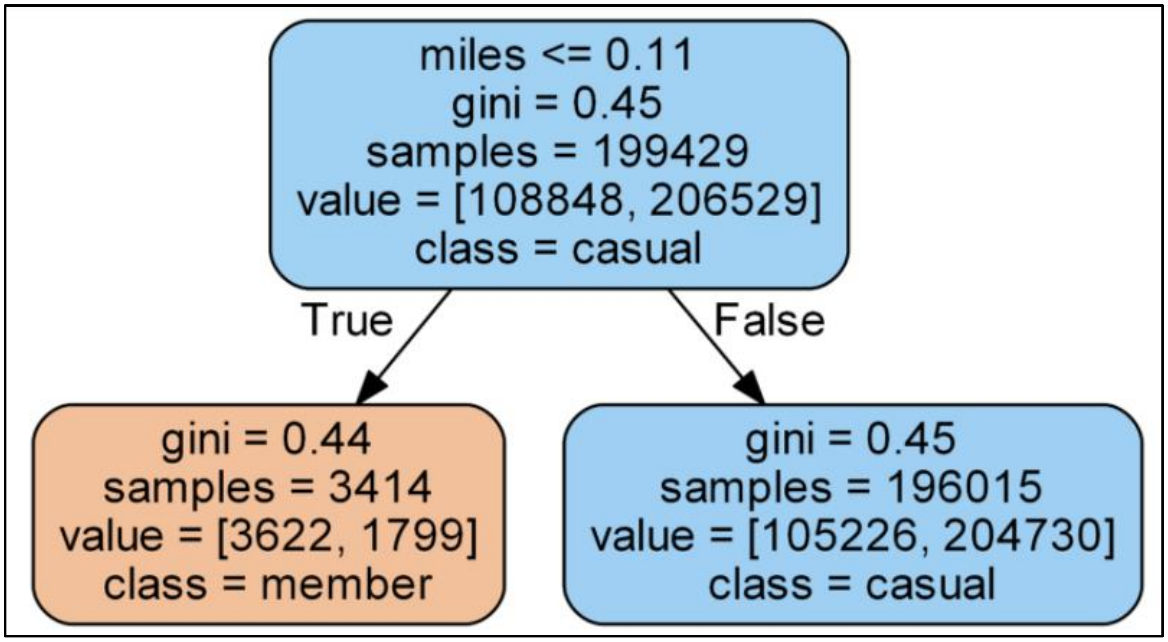
Figure 8

Second Regression

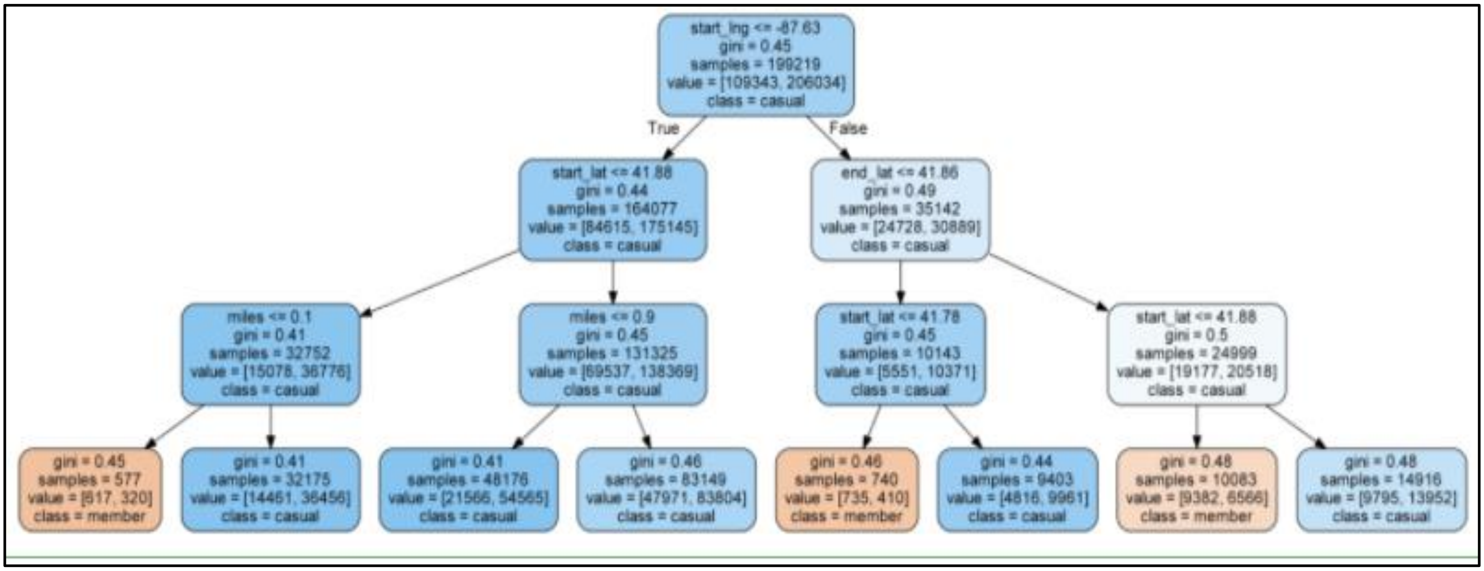


```
       Dep. Variable:           miles      R-squared (uncentered):       0.603
              Model:             OLS  Adj. R-squared (uncentered):       0.603
             Method:   Least Squares                 F-statistic:   1.198e+05
               Date: Thu, 03 Dec 2020           Prob (F-statistic):        0.00
               Time:        23:42:51               Log-Likelihood: -5.0474e+05
   No. Observations:          315377                          AIC:   1.009e+06
       Df Residuals:          315373                          BIC:   1.010e+06
           Df Model:               4
    Covariance Type:       nonrobust

               coef  std err       t   P>|t|  [0.025  0.975]
  start_lat  0.8720    0.104   8.422   0.000   0.669   1.075
  start_lng  1.4526    0.121  12.010   0.000   1.216   1.690
    end_lat  2.4861    0.103  24.040   0.000   2.283   2.689
    end_lng  0.1362    0.121   1.125   0.261  -0.101   0.373

        Omnibus:  139779.135      Durbin-Watson:      1.693
  Prob(Omnibus):       0.000   Jarque-Bera (JB): 860429.377
           Skew:       2.058           Prob(JB):       0.00
       Kurtosis:       9.967          Cond. No.   1.19e+04


Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.19e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 9



Tree Iteration 1



Tree Iteration 2

6.) References

"Divvy System Data." *Divvy Bikes*, Motivate International, 26 May 2020, www.divvybikes.com/system-data.

"Chicago, IL (S Ashland Ave / W 76th St)." *NeighborhoodScout*, 2020, 5:15, www.neighborhoodscout.com/il/chicago/ashland-76th.

Department of Health & Human Services. "Cycling - Health Benefits." *Better Health Channel*, Department of Health & Human Services, 30 Nov. 2013, www.betterhealth.vic.gov.au/health/healthyliving/cycling-health-benefits.