# LA Metro Bike Analysis: Integration of the Past, Present, and Future

2019 TAMIDS Data Science Competition

Texas A&M University

400 Bizzell St

College Station, TX 77843

By Isaac Ke, Michael Lee, Pranika Madhavan, and Stephanie Sims

Advisor: Dr. Jeffrey Hart

April 3, 2019

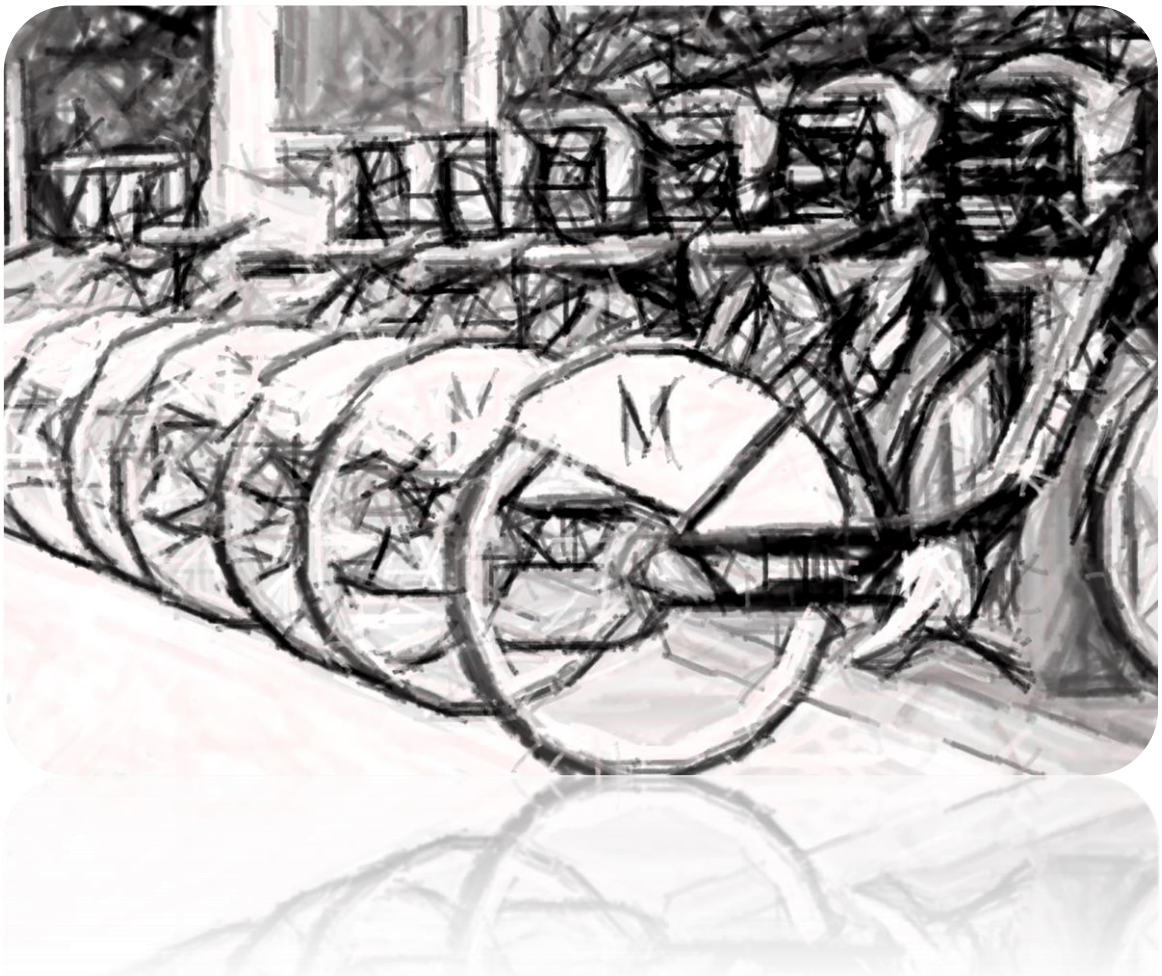


**Table of Contents**

## I. Introduction (Executive Summary)

a. Problem

With the main data that was provided to us, we decided that the main problem that had arisen was the uncertainty of which regions/stations would flourish with more stations/bikes at specific times used by certain populations. In short, the problems were akin to the complexity of how to differentiate between the many different factors and how to maximize the data in order to formulate suggestions that would benefit both the society and the company. With many factors in consideration such as the starting/ending points, number of rides per day, and analyzing which months were more popular than others, we were able to create graphs and charts while utilizing particular statistical methods that would help find visible and data-driven patterns. This accumulation of data could later be used for generating suggestions.

We sought to make inferences about new variables and generate new parameters based off of information that was not provided, but these inferences would assist in drawing conclusions about the data. Using the limited amount of data provided as well as extracting the right balance of variables to analyze certain problems was the underlying theme of our analysis. Identifying outliers as well as missing and problematic values were also a dilemma that was encountered.

b. General Approach Used to Address the Problem

Using R (in the RStudio IDE), we started off by plotting the data in various ways to see initial trends. From there, new parameters were added to aid in more detailed analysis. We focused on analyzing ride duration and location data as a function of pass type, region data, and time period. Furthermore, certain statistical methods such as ANOVA were implemented to both validate and draw conclusions from the data.This allowed us to delve into possible explanations of patterns that we saw within this time frame and possibly allow for suggestions of when and where bikes should be implemented as this was plotted with region taken into account as well. We also divided this data into the type of rides between walk-in rides or membership rides. Regions and times with more membership rides would mean that a more regular number of customers rely on a consistent service provided by Bike Share. Then, we analyzed the previous years of the same quarters and tried to find general patterns to predict the next cycle for the year after. In addition, we took into account the overall pattern of yearly growth and how this would translate into forecasting for next year.

This section was largely similar to the previous as we had to rely on models used before but instead of number of rides, we took into account the revenue generated from each quarter and region to find where and when most opportunities would arise in relation to ticket sales. Again, these were plotted against a time versus income plot in order to find a general trend in how ticket sales would decrease or increase based on past precedents. We were also able to find some crests and troughs within the graph that could help either utilize or avoid these times that would prove to be fruitful or fruitless.

c. Recommendations and Estimated Benefits for Adopting these Recommendations

Some of the main recommendations based off of the data analysis for improving the business included lowering the cutoff for extra ride charge to 25 minutes, implementing a 6-month pass, and concentrating bicycle placement in certain downtown LA stations. We discussed possible solutions and suggestions based on what a successful or unsuccessful region would look like in regards to Bike Share profit. The general idea for a successful region is one that promotes

a bike-friendly zone in which there are dedicated bike areas between popular destinations and an acknowledgement to environmental factors that would otherwise discourage or even harm users. If placed strategically near new companies with many employees as well as finding hotspots of new tourist sites, there would be a dramatic increase in usage of Bike Share. On the other hand, there were some problems that would make it an unsuccessful region with problems such as air pollution, more efficient means of transportation like buses, and areas where buildings are more spread out that would require an automobile over a bike that would be too inefficient.

A strategic placement for new bikes would be to place them near tourist locations where a demand for bikes would be high. In addition, college campuses should have bikes located near them as it is another major consumer population. Appealing to citizens and tourists alike in reducing gas emissions can also stand to improve the number of bikers. Finally, the last suggestion is to remove docking and import a GPS system as other bike companies have found large success in being able to leave a bike that could be found by others easily instead of having to relocate to simply put the bike back into a rack.

**II. The Problem and Data Collection**

The data that we received contained the information regarding the time, date, starting point, ending point, and type of pass specified to the duration of which the pass would be valid for per ride. From this, we were able to fill in the "missing spaces" such as the duration of each ride by writing a program that would extract the start time and end time to find the difference between the two in minutes to later on calculate the revenue earned from the specific ride. Another "missing space" included the region name corresponding to each start station and this was also accounted for by a program that would pair the starting station to locate the name of the region so that later, more analysis that may have some relation to the region could be discovered.

After filling in these "missing spaces," analysis of the data could formally begin, starting with the division of region and quarters so that a more specific pattern could be found that would otherwise not be uncovered if all of them were placed into one giant compilation. An important step included having to separate each ride into the type of pass used and finding a pattern in regards to the region or time to find information such as what region tended to attract the most local customers who most likely purchased passes rather than using walk-up biking. Another consideration would be what time of the year it would be the most popular for tourists.

As the data was now divided into these two categories, a more effective method for calculating total profit could occur in that a program would be able to determine the total money spent by customers using this service. For those that have purchased a pass, the first 30 minutes of their rides would be free while every 30 minutes after included a $1.75 fee so the program would search out every ride over 30 minutes for those with a pass and calculate the revenue from those trips. For those that did not purchase a pass, R code was written to multiply the minutes biked with the base rate of $1.75 for every 30 minutes. These two results would then be added and distributed according to pass, region, and the designated quarter to find additional patterns to find the most popular combination of time and place for Metro Bike Share.

The next step included sorting the rides by the day they occured to be put in the form of an easily-readable graph that could give insight into if there were specific days where bike usage did particularly well or bad. It would also give us a general pattern for each region and quarter so we could pinpoint where the most number of rides occurred. The regions could also be compared to each other to determine which regions did better in which time frame. Moreover, a continuation of this idea was extended to graphing the data between the day versus the average minutes traveled per trip on that day to find which days in each region were with higher or lower durations than usual. The idea here is that looking for patterns could allow for a model that could point to certain time quarters and regions that were successful in generating profit and similarly, unsuccessful days that could be attributed to some potential barriers which could be fixed to increase the number of users.

**III. Data Preprocessing**

To better model the data, the addition of the parameters "ride duration," "day of ride," and "start station region" were created to assist in drawing our final conclusions. In specific, the duration parameter was an integer value of the total length of time in minutes for each trip from the start time to the end time. This was helpful when it came to calculating the revenue for both walk-ups and rides that were over 30 minutes which were not walk up. Both of these cases' costs were purely based on time. The day parameter was an integer value ranging from 1-908 that corresponded with the number of days since the data was released on July 7 from Q3 of 2016 to the end of Q4 of 2018 (see Appendix F. *Figure F2*). This new variable was used in separating the data into quarters and different regions to see trends across quarters. Lastly, the start station region parameter was added as a categorical factor to analyze trends across regions (see Appendix F. *Figure F3*).

Regarding data preprocessing, missing, null, or problematic data values were found when attempting to add the three parameters listed above. Running multiple iterations through the data to extend certain variables to new parameters led to some null values in the output. For example, it was found that stations with ID's 3009 and 3039 had no corresponding region in the station table data sheet. To solve this problem, we replaced these start station region values with "DNE" for does not exist. Furthermore, station ID 4276 had a row of information in the station table excel sheet but was not used in the big bike data spreadsheet.

For outlier identification, this was most utilized in the ride duration parameter that was newly constructed. Using the R package ggplot2 (see Appendix F. *Figure F8*), the graph of the different ride duration frequencies were plotted (see Appendix D. *Figure D1*). Knowing that the ggplot command auto scales the x-axis, we inferred that there must have been an outlier for the

ride duration because of the clustering of the data near the shorter times. We found that the maximum ride length of 1,440 minutes (60 hours or ~2.5 days) occurred 2,126 times, thus hinting at the fact that rides most likely max out at this duration, having occurred due to a passenger forgetting to lock his/her bike after they were done using it. Hence, these points were removed (see Appendix F. *Figure F8*) and the new data was plotted (see Appendix D. *Figure D2*).

**IV. Data Analysis and Interpretations and Applications**

The initial analysis consisted primarily of plotting the data in various ways. For one, the number of rides per quarter were plotted (see Appendix A. *Figure A1*). This showed that the number of rides reached a minimum during 2017 Q1 with a total number of 33,786 rides. On the other hand, during 2018 Q3, a maximum of 95,283 rides occurred. It was noticed that there was a general cycle of rising and falling in regard to bike usage with an overall positive sloping trend.

Next, the percentage of rides that used certain pass types was found (see Appendix B. *Figure B1*). From the visual, it was evident that the majority of riders used a monthly pass with 367,146 total rides. Next, walk-up passes consisted of 222,104 rides. Surprisingly, the number of annual pass users was so small that it barely showed up on the pie chart. However, since rider IDs were not provided, using monthly, daily, flex or annual passes as a way of calculating revenue was not safe as there might have been overlap across trip ID's for people who have paid one time but take multiple trips. Hence, we figured some revenue information can be calculated using the walk-up data since each walk-up trip ID corresponded with one separate payment of $1.75. Moreover, all rides, no matter the pass type, that went over 30 minutes were charged $1.75. So, the total revenue from walk-ups was calculated (see Appendix F. *Figure F4*) as well

as the total revenue for rides that went over 30 minutes (not including walk-ups) (see Appendix F. *Figure F5*). These values were found to be $1,026,767 and $284,704 respectively, for a total of $1,311,471. Furthermore, the average revenue from each walk-up was found to be $4.62 with each rider riding an average of 79 minutes (see Appendix F. *Figure F6*).

Next, as seen in Appendix C, the number of rides was plotted vs. other variables including start station ID, day, and start region plotted (using R code found in Appendix F). As seen in *Figure C3*, a disproportionately large amount of rides occur in the DTLA region. Furthermore, the majority of the riders in DTLA and Pasadena used a monthly pass while the riders in the Port of LA and Venice were walk-ups. *Figure C2* illustrates how the stations outside of DTLA did not open up until after around day 370. Past certain days, certain stations in Pasadena, in particular, began to decline in usage. Overall, bike stations in Pasadena and the Port of LA were rarely used.

The graph D2 shows the popularity of the ride durations in different regions (see Appendix D. *Figure D2*). In specific, Venice is consistent in terms of popularity for all different ride durations. However, Port of LA has rarely any riders in general, prompting a possible removal of bike docks from this region. Pasadena and Downtown LA both seem to have more favorable riders in shorter durations, specifically, the 0-25 minute range. With a more closer look, Pasadena's overall population of riders is significantly lower than Downtown LA's. Thus, the perk of pass-holders having the first 30 minutes free should be reduced to 25 minutes to generate more profit.

Regarding bike allocation, analysis of the data showed that more bikes should be added or concentrated in the stations listed in (Appendix D. *Figure D3*). The majority of these stations are located in the downtown LA region. Adding more bikes to these more popular stations would

allow for more riders to have access to bikes. Furthermore, the surrounding suburbs of downtown LA are another target area for more bike stations. The rationale is that commuters will be able to have easier access to the congested downtown, where many people work.

For more in-depth analysis, we decided to conduct a multi-factor ANOVA test (see Appendix F. *Figure F13*) to determine if there is an interaction between factors of the region, pass type and whether these factors impact the ride duration. An ANOVA table was produced (see Appendix E. *Figure E3*) and we thus concluded that there is indeed an interaction between the factors and the duration response. Furthermore, we conducted Tukey's multiple comparisons test (see Appendix F. *Figure F14*) to see which specific factors differ based on the mean ride duration. Only three confidence intervals did not contain 0; so, we rejected the null hypothesis and concluded that there is a difference between the mean ride duration for Venice, the virtual station (of which had no interest to us), flex pass, annual pass holders, monthly pass and annual pass holders. We can then conclude there is a difference between these factors on the mean ride duration. Thus it is reasonable to implement a pass that meets in the middle, as we will discuss later in network management.

**V. Forecasting Bicycle Demand**

We split this process into two parts in order to find an estimate of the bicycle demand during the time period Q3 of 2018 to Q1 of 2019: comparing by the quarters respective to the year and looking at the preceding months.

By comparing the time period for the past two years that Metro Bike Share data was available to the public, it allowed us to more accurately provide a forecast as the time of year is a vital factor in determining the number of users for bikes. For example, comparing Q3 of 2017 to

Q3 of 2018 would be a more accurate comparison than Q4 of 2017 to Q3 of 2018 (see Appendix A *Figure A1*). Presumably more people who purchase monthly, annual, or flex passes return to their homes in LA as compared to during the summer when there were noticeably less monthly passes bought. In the summer, people typically use the time to go on vacations, and college students in LA would return to their hometowns. However, comparing the quarters at around the same time of the year allows us to eliminate external factors such as these and allow us to see the patterns more accurately. What we could notice is that for Q3 of 2016 to Q1 of 2017 and Q3 of 2017 to Q1 of 2018 followed the same pattern where they would experience a higher number of rides for the former part of the time period, steadily decrease, then resurface the following year. Using this model, we could determine that Q1 of 2019 would likely be slightly lower than Q4 of 2018.

On the contrary, we still acknowledged the fact that comparing the preceding years would give us a more in-depth pattern of the overall growth of the number of people who would be using Metro Bike Share. Since we took this into account when formulating the model for the time frame, it also maintains consistency in terms of overall growth at an annual level. For example, we can conclude that later years will have a higher number of rides with respect to previous years of the same time period. The Q1 of 2019 would likely be higher than Q1 of 2018.

**VI. Forecasting Ticket Sales**

Similar to the process of forecasting bicycle demands, we noticed that we would have to compare similar times of the year for a more accurate depiction of the predicted model to eliminate external factors that may affect the model. However, this time we were more focused

on extracting the total revenue for each month as opposed to the entirety of the data to follow a monthly pattern rather than a general pattern to more accurately describe the model.

We ran two different programs that would run through each quarter based on whether the user had a walk-up encounter or biked with a pass. While Q3 2016 through Q1 2017 followed the logical pattern of decreasing revenue from walk-ins, there was an increase from Q4 2016 to Q1 2017 in revenue for those with passes. However, in the end, the data still managed to have an overall decline during this time in the combined revenue (See Appendix A *Figures A2, A3, and A4*). Interestingly, all three models reflected an increase from Q3 2017 to Q4 2017, but a decline from Q4 2017 to Q1 2018, suggesting a rather inconsistent pattern in these time frames. However, similar to the graph analyzing the number of rides per quarter, a decline is to be expected from Q3 2018 to Q1 2019 as shown by all three graphs. Thus, there will most likely going to be less revenue generated in Q1 2019 than the prior two quarters.

**VII: Success Characteristics and Ways to Achieve Them**

The placement of bike stations is imperative and directly correlated to the usage of the biking system. The ideal bed of bike-usage hotspots in terms of a regional perspective would be metropolitan areas with large companies scattered within a surface area. In these areas, it is common for the workers to commute daily, and many will use bikes for their 10-20 minute rides on a daily basis. In addition, since that region has plastered companies with many employees, there is a chance that a significantly larger number of companies constantly moving into the area would garner more demand for LA Bike Share for working commuters.

In addition, the infrastructural add-in of bike-friendly paths to certain areas would promote the usage of a bike-share program within the city. It allows for a more convenient trip

for the public and encourages them to use bikes. This is usually seen in areas where environmental awareness is a priority. Places that focus more on clean air, recycling, and reducing pollution tend to entice more bike riders as more people are encouraged to enjoy the health benefits of biking outside. For a more closer look, in heavily populated areas such as Downtown LA, there is a lot of traffic which includes cars, buses, and other forms of transportation that can allow them to get to places faster and cheaper. However, these pro-fossil fuel vehicles are a detrimental con to the environment. Therefore, bike companies are generally looked highly upon while the public sees them as a cheaper and more environmentally friendly option.

Another great example of densely populated areas are commuters on college campuses. In these places, many students use bike sharing systems to avoid walking while simultaneously opting out of financing a car, making bikes the preferred method of transportation. Additionally, campuses in the urban regions of LA, where the gas prices are high continue to force students to use other modes of transportation.

Tourism is another profitable sector for bike sharing to boom. Whether it is for people touring from another region or simply locals enjoying some time off, plugging Metro Bike Share into these regions would be a great incentive for Bike Share. For example, Long Beach is an area where an abundance of tourists and locals come to visit and travel around in order to get to a hotel, restaurant, etc. Thus, tourist-dense areas are a successful region for Metro Bike Share to place an abundance of docking stations.

A major cautionary look Bike Share should have is on places where competing forms of transportation and companies are present. For example, in LA, there any many forms of alternative transit, be it busing or the rail system. Therefore, the competition and tensions to beat

these transporting systems and become more popular is also high. However, with the right approach, it will be easy to tackle these areas that are generally seen as unsuccessful regions or bike dead-zones to place bikes. For an insightful source, Matthew Tinoco, a journalist for the LAist, wrote a recent article by City Metric; he focused on issues with urban planning regarding both busing and the rail system and how this demonstrated that these modes of transportation are becoming less and less popular. In fact, when looking at the "population of LA, barely even 20 percent of the population uses them." If bus service was more consistent, or rail service more ubiquitous, I think Angelenos would flock to transit," Tinoco says.

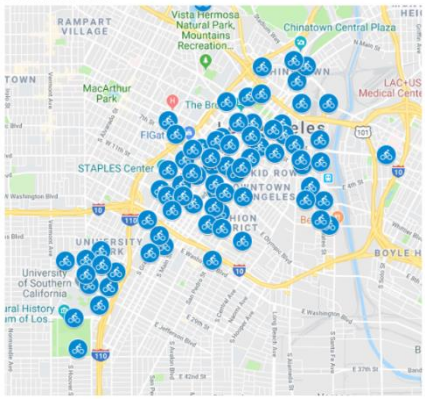
With this in mind, biking systems can confidently place more docks near subway stations as many people would use it to get to a rail stop and from there make long-distance commutes. Not only would this curb the effort of customers by saving them more steps after a long metro ride, but this would also provide a means of a unique city experience where bikes would be available all around the city while touring or simply enjoying the urban scene. This would aid the bike shares and even allow them to partner with the rail system. Both would benefit from each other thanks to the use of bikes. Busing however, is a direct competitor, but biking weeds buses out as bikes are always present at set stations which brings out the consistency that buses lack. Even then, bus stations can serve as good biking stations similar to how metro stations can be.

Another competitor to look out for in terms of regions is where competing bike-share companies' docks are located, an indication of the fad of electric bikes. There are still many benefits to simply using a standard bike with the biggest being that electric bikes need to be recharged and therefore, inevitably, need charging docs or stations. In fact, if they are parked anywhere especially the further away it is from a charging station, it can be disastrous for the company since many bikes would be left uncharged, and searching and connecting them to

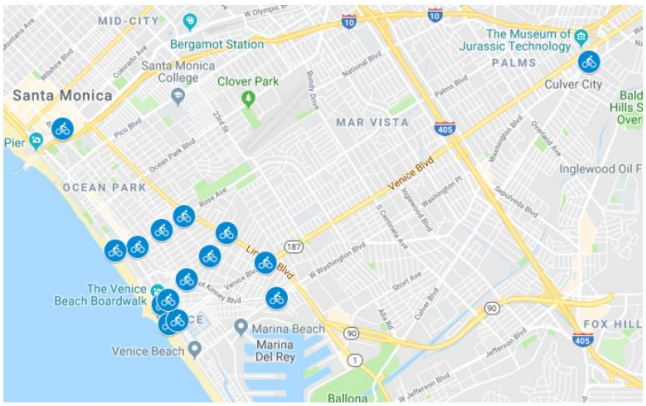
stations would serve as a financial burden. Meanwhile, the conventional method does not require any such thing. Thus, being near electric bikes is actually not an unsuccessful region to have these bike stations. With this one point alone, the normal bikes would likely become the victor as a better option for their flexibility. Keeping this in mind, a suggestion would therefore be to make a move to become dockless which will be elaborated more in detail in the following section.

## VIII: Network Management and Business Recommendations

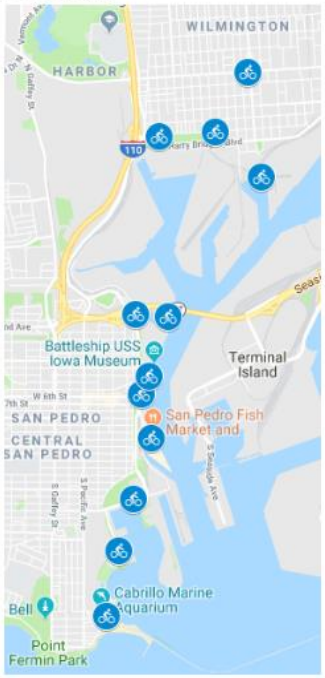
Downtown LA:



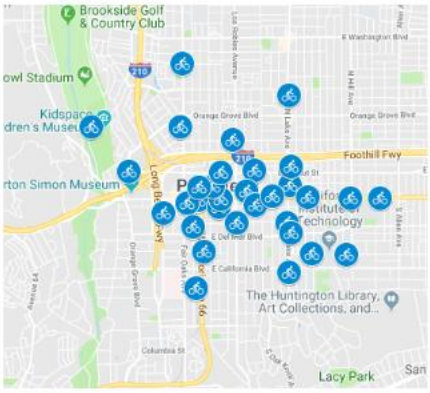

Venice:




Port of LA:




Pasadena:



Based on our analyses, there are many ways that the Metro Bike Share could expand and find lucrative results. There are many areas of high population density around that aren't covered by the Metro Bike Share. Within these areas and existing areas, the places with the heaviest tourism are the perfect candidates for new bike docks. Some examples of tourist spots to add bike ports are The Westin Bonaventure Hotel, Grand Park, and the Sports Museum of LA in Central LA; Loews Santa Monica Beach Hotel, Ocean View Park, Marina Beach in Venice; The Westin Pasadena Hotel, Grant Park, and Norton Simon Museum in Pasadena; and Best Western, Fort MacArthur Museum, and Point Fermin Park in the Port of LA. These examples of course are just some among many other attractions.

There are also places outside of the four areas that Metro Bike Share covers that are highly populated that would be good candidates for expansion. These include Beverly Hills, Hollywood, Long Beach, and East LA, among others. These areas attract a lot of tourism in addition to the current residents. Therefore, it is foreshadowed to be successful future candidates for the implementation of docks.

As mentioned earlier, college campuses are also a hot spot for quick, inexpensive transportation, and putting multiple docking stations on a few campuses would create more business for the company. Examples of universities to expand to are California State University LA, Long Beach and the University of Southern California. Within each of these universities, strategic placement for bike docks would be around libraries, student centers, rec centers, stadiums, and dorms. These are all places where students gather often and in large volumes for activities or studying, thus there is an increased likelihood of needing a bike.

To add on to the most noteworthy regions to consider, analysis of the data shows that more bikes should be added or concentrated in the stations listed in Appendix D, *Figure D3*. The majority of these stations are located in the downtown LA region. Adding more bikes to these more popular stations would allow for more riders to have access to bikes. Furthermore, the surrounding suburbs of downtown LA are another target area for more bike stations. The rationale is that commuters will be able to have easier access to the congested downtown, where many people work. When in comes to downtown though there is another important factor to consider, that being advertisement.

The most obvious form of advertisement is simply the sight of people riding the bikes. But, when it comes to paper or electronic ads, certain areas can be emphasized in order to appeal to audiences. For example, the eco-friendliness of the bike-share program, as mentioned earlier is a huge aspect to be promoted. As the desire for an emission-less way of moving arises, the bikes' part to play in reducing humanity's carbon footprint should be preached in order to appeal to those concerned with the environment. This will also highlight this selfless approach to transportation.

Another thing that could potentially be beneficial to the Metro Bike Share company is a partnership with the Los Angeles Metro Rail and bus services. With docks located at every subway station and bus stop, customers can easily shift from the bus to bike to subway, or any combination of the three, using one card. Rewards or discounts could be given to frequent users. This partnership would benefit everyone involved and bring more ease of access in general to the LA transportation system. There is currently a card system known as TAP which does bridge the three for payment, as it allows for flexibility in travelers to switch over. However, to promote more bike usage in these areas, rewards and discounts can be emphasized more. Later on, if there

is a significant increase in the usage of the rail system with correlation to the bike docks nearby, Metro Bike Share can even potentially secure a percentage of sponsorship by the rail system in constructing more docks by rail systems. This, in turn, will foster a chain of gaining sponsorship money, creating more docking and in turn more customers.

One feature that could be taken away in order to improve the efficiency of the bike share system is the need for a dock. Companies like Ofo, Spin, and LimeBike are all dockless bike share companies that operate in the LA area. A definite benefit to having a dockless system is the disappearance of the need to forecast the locations of the docks. With the dockless system, some locations will be natural spots for bikes to gather or spots that nobody could predict. What this would mean for the company is that GPSs' would need to be added to the bikes so that they can be found by the customers and more people would need to be employed in order to organize and distribute the bikes every so often. This added expense won't be too much of a burden on the company, though, because it will gather more business due to the convenience of the bikes around the city. The dockless system also gives an advantage over electric bike companies, whose bikes have to be docked in order to charge so that they can actually be ridden.

Looking at the pass duration, we noticed that there is a huge gap between the 30-day pass and the 365-day pass. One way to bridge the gap is to include a 6-month pass for those who don't necessarily want a pass for a whole year. This additional pass would encourage customers to make the purchase since the duration is closer to what they want. Additionally, many people take vacations and college students leave town for the summer. These are long periods of time that a pass will not be used. These sort of customers would be pleased to see the option to hold or terminate their pass so they wouldn't be spending money on something they're not actually using.

This would, in turn, decrease the amount of uncertainty people might feel before committing to a year-long pass.

**IX: Summary**

To summarize our findings, we were able to organize the information given to us into charts, graphs, and statistical tests in order to more clearly find trends and spot flaws. In order to increase revenue and expand the impact of the LA Bike Share company, we believe that a six month pass should be added, the cutoff duration for extra charge should be lowered to 25 minutes, more bikes should be placed in the downtown LA stations, and new stations should expand outward into the suburbs. Our analysis shows revenue is on track to increase, but new innovations would help prevent elongated troughs in the revenue cycle as well as spur new possibilities for customer interaction.

Locations to add bikes and stations include highly populated areas, tourist locations, college campuses, and near subway stations and bus stops. An advertising technique would be to focus on the eco-friendliness of the system. Adding bike-paths would also greatly increase the demands for Bike Share. A dockless system can improve the convenience of the bike share.

## X. Citations

Panagiotopoulos, Vas. “The Los Angeles Metro Is Great – so Why Aren't People Using It?” *CityMetric*, www.citymetric.com/transport/los-angeles-metro-great-so-why-aren-t-people-using-it-2742.

“Data.” *Metro Bike Share*, 25 Mar. 2019, bikeshare.metro.net/about/data/.

“Los Angeles Population Density.” *Arcgis.com*, www.arcgis.com/home/webmap/viewer.html?webmap=5913b5311e6449909e4139117c96a878.

“Trying Out L.A.'s Three Dockless Bike-Share Systems.” *Streetsblog Los Angeles*, 22 Dec. 2017, la.streetsblog.org/2017/12/21/trying-out-l-a-s-three-dockless-bike-share-systems/.

# Appendix A

## Revenue Analysis by Quarter

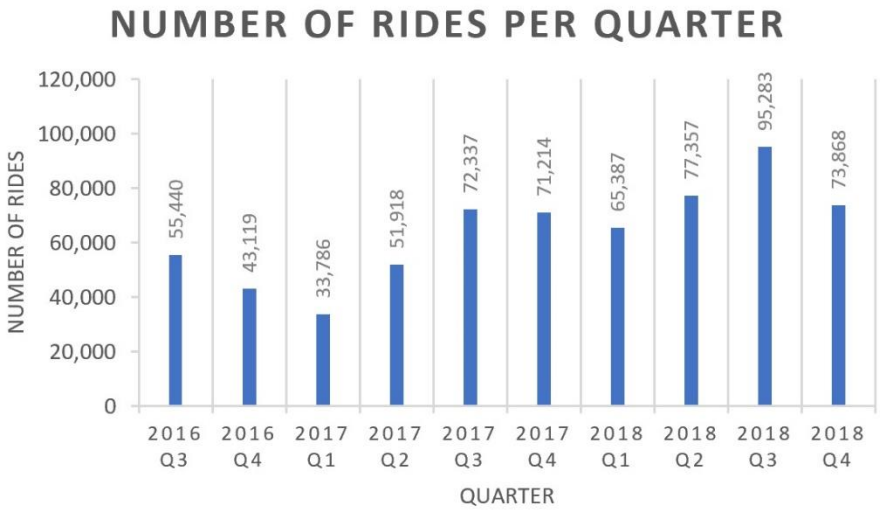



*Figure A1.* Number of rides per quarter.

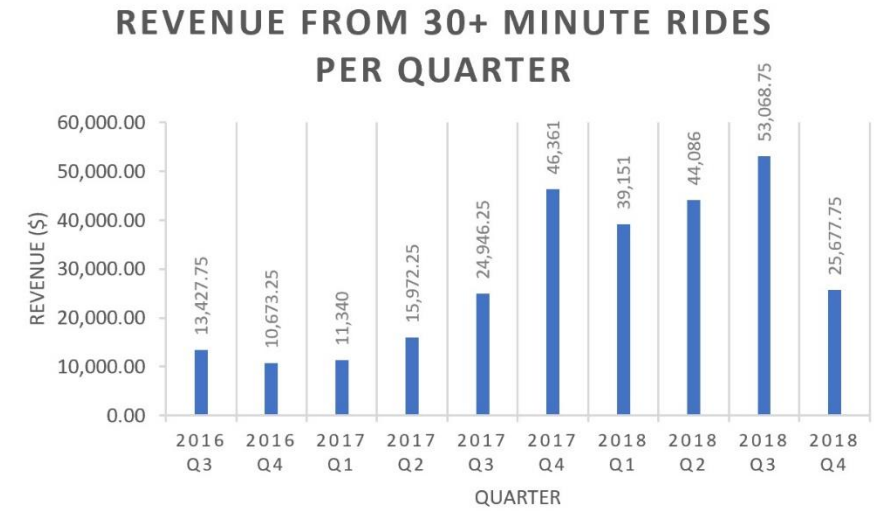



*Figure A2.* Revenue from 30+ minute rides per quarter.

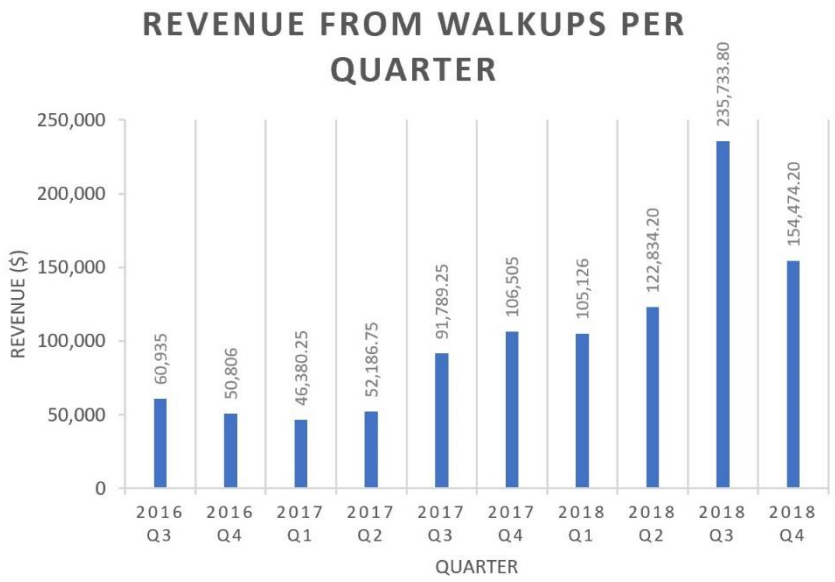



*Figure A3.* Revenue from walk-ups per quarter.

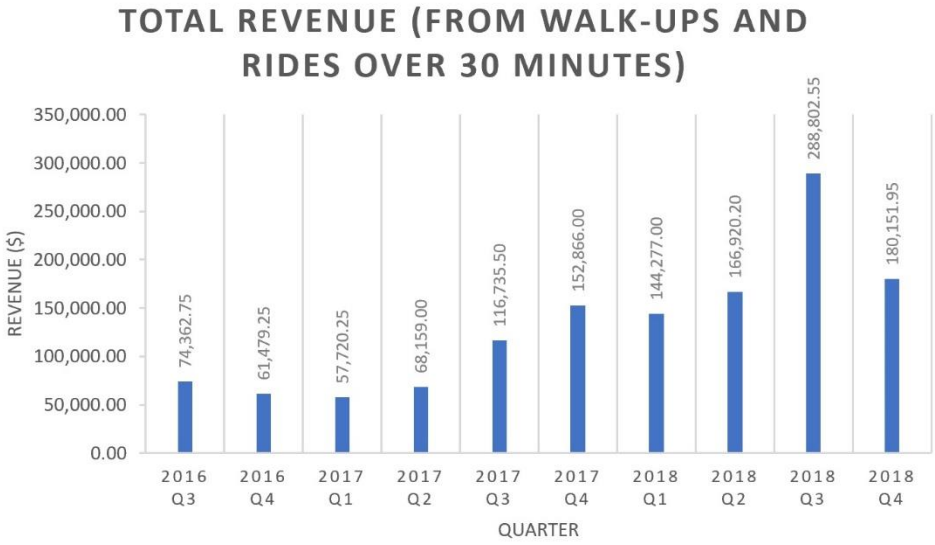



*Figure A4.* Total known revenue per quarter.

# Appendix B

## Type of Passholder Type

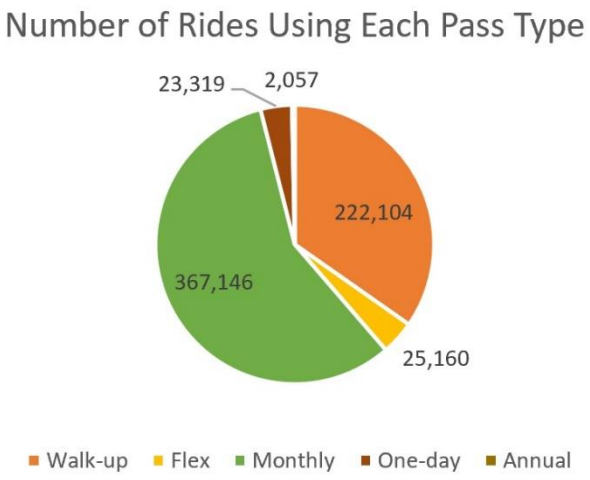



*Figure B1*. Number of Rides Per Pass Type.

# Appendix C

## Passholder Type and Start Region Analysis

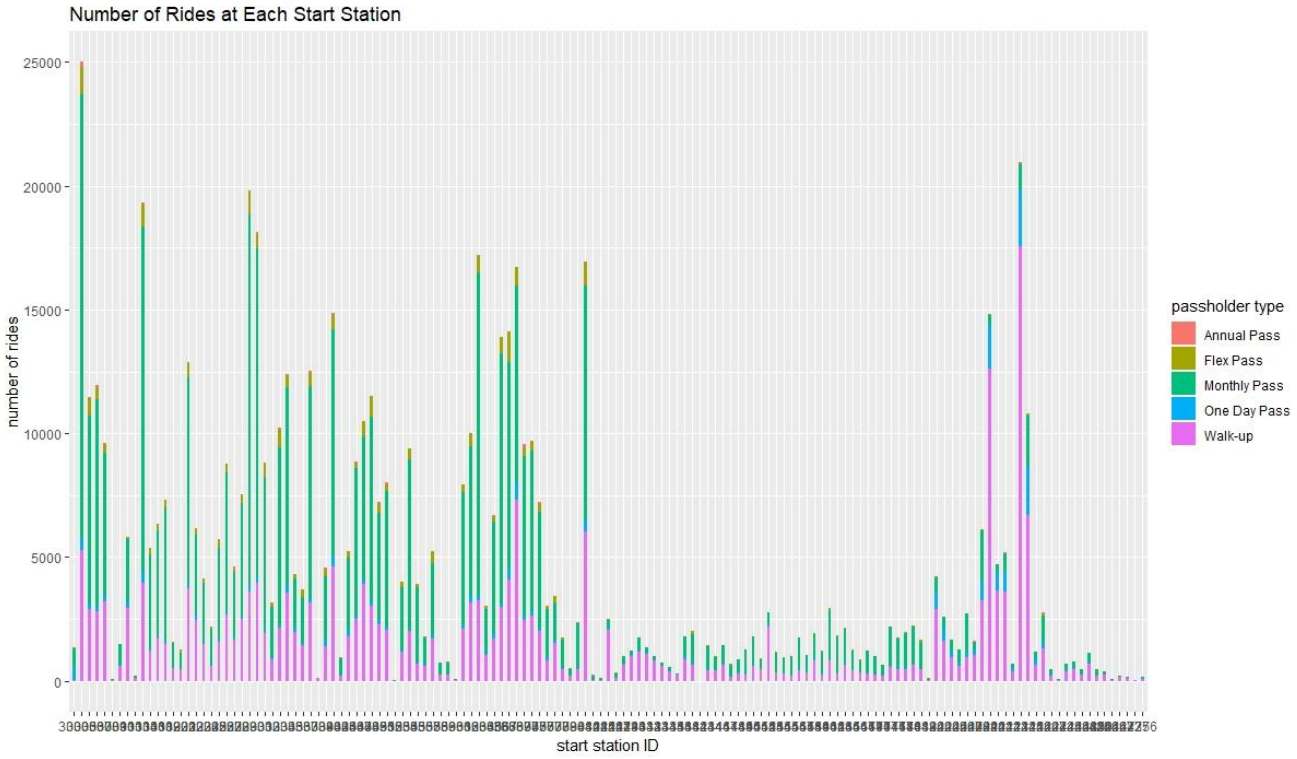



*Figure C1.* Number of Rides at Each Start Station Filled with Passholder Type.

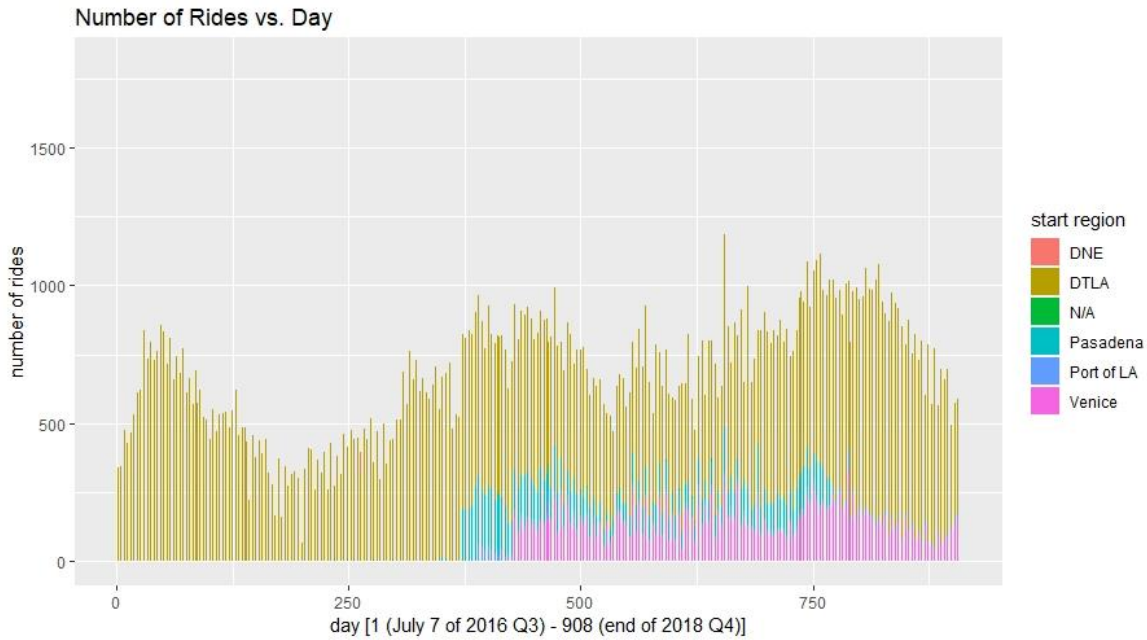



*Figure C2.* Number of Rides vs. Days Filled with Start Region.

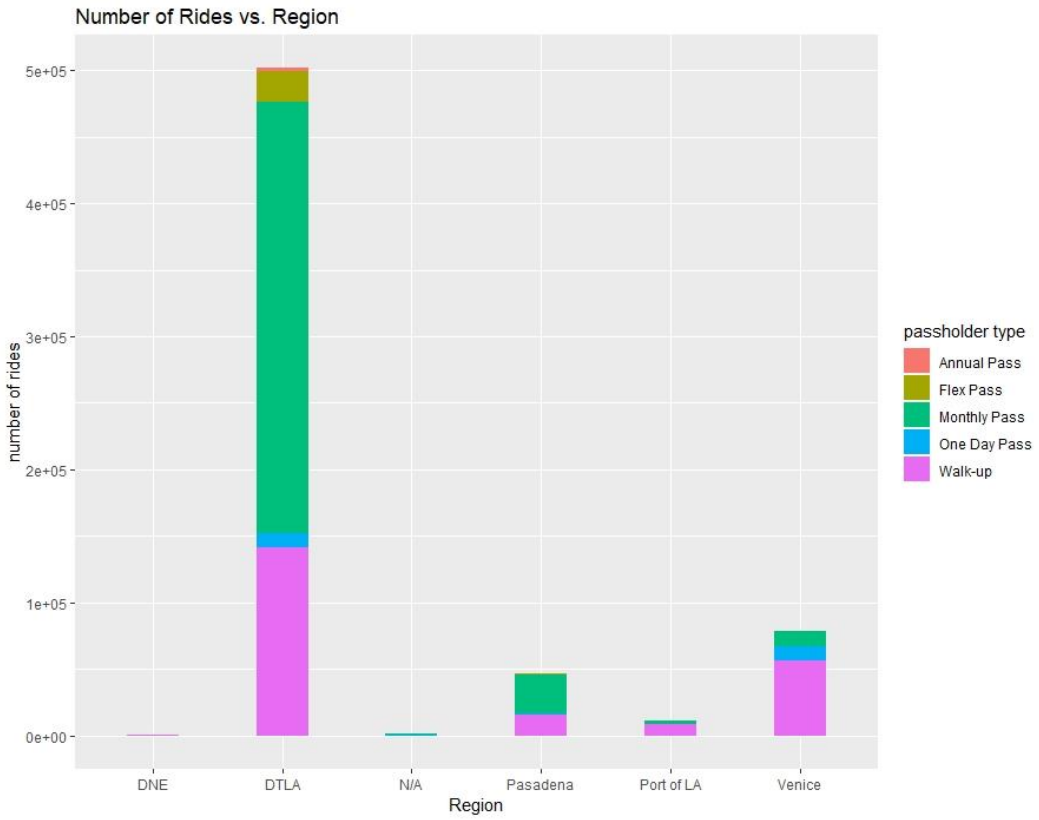



*Figure C3.* Number of Rides vs. Region Filled with Passholder Type.

# Appendix D

## Ride Duration

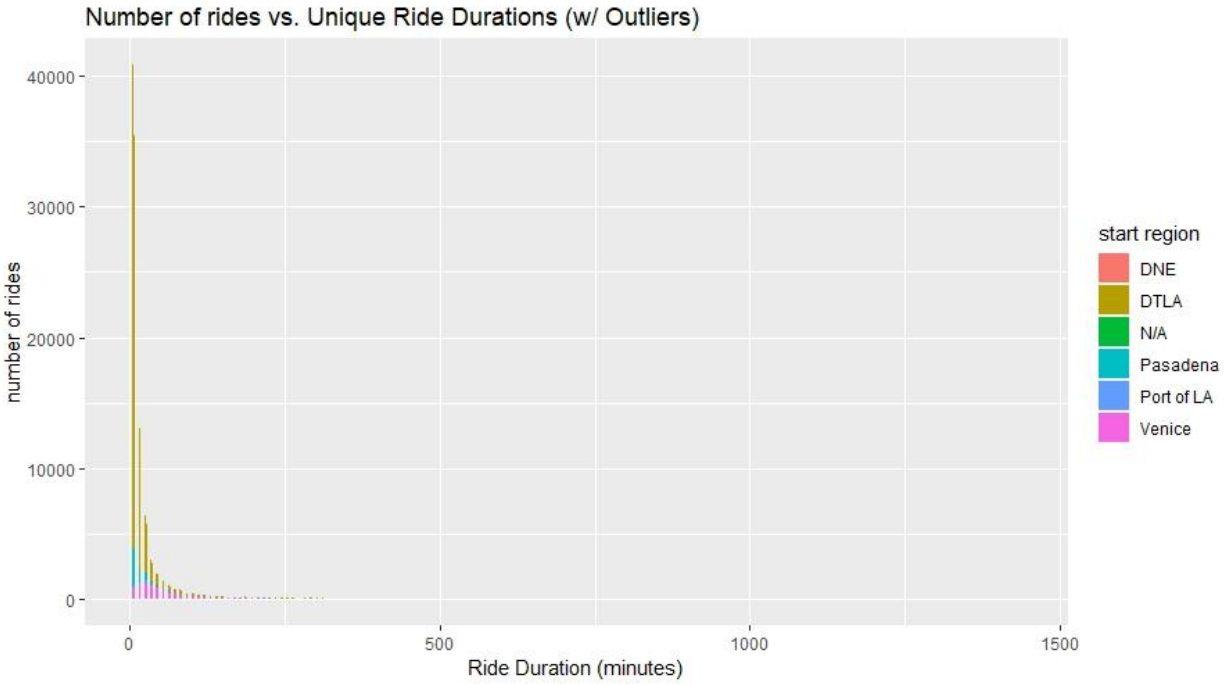



*Figure D1.* Number of Rides vs. Duration Filled with Start Region (with outliers).

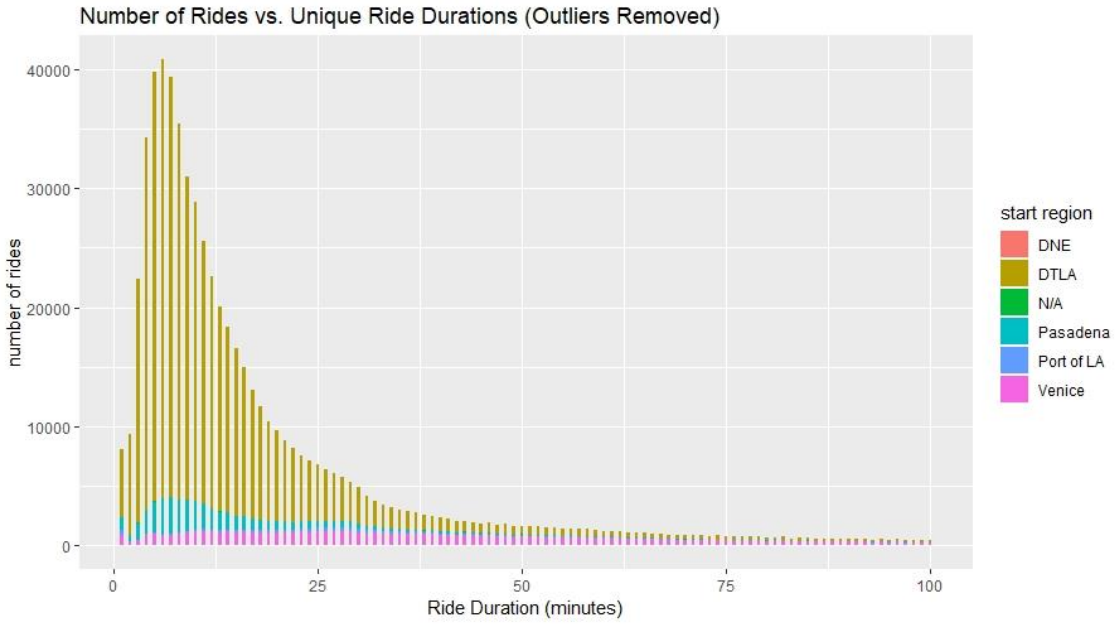



*Figure D2.* Number of Rides vs. Duration Filled with Start Region (with outliers removed).

| start_station_id | total number of rides |
|---|---|
| 3063 | 10020 |
| 3034 | 10223 |
| 3048 | 10490 |
| 4215 | 10812 |
| 3006 | 11450 |
| 3049 | 11513 |
| 3007 | 11928 |
| 3035 | 12412 |
| 3038 | 12510 |
| 3022 | 12857 |
| 3067 | 13912 |
| 3068 | 14131 |
| 4210 | 14816 |
| 3042 | 14847 |
| 3069 | 16735 |
| 3082 | 16953 |
| 3064 | 17197 |
| 3031 | 18144 |
| 3014 | 19328 |
| 3030 | 19824 |
| 4214 | 20935 |
| 3005 | 25000 |

*Figure D3.* Most Popular Stations to Add Bikes To.

# Appendix E

## Multifactor ANOVA

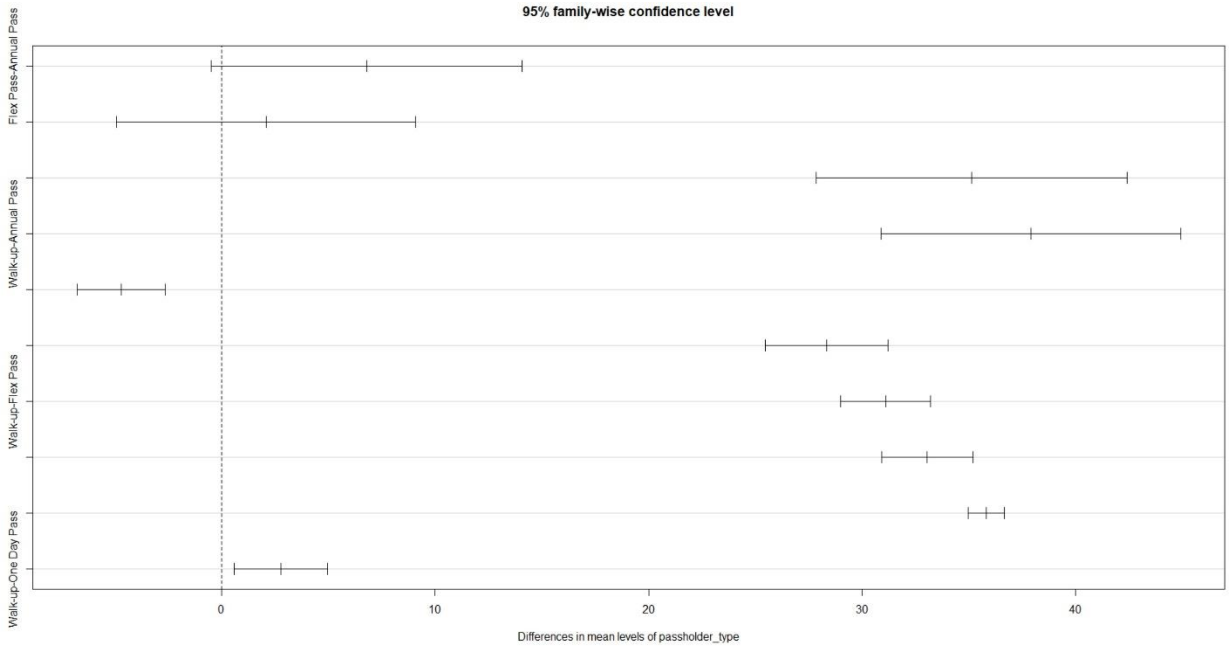



*Figure E1.* Plot of 95% confidence intervals for all of Tukey's multiple comparisons for difference in mean ride duration for different pass type.

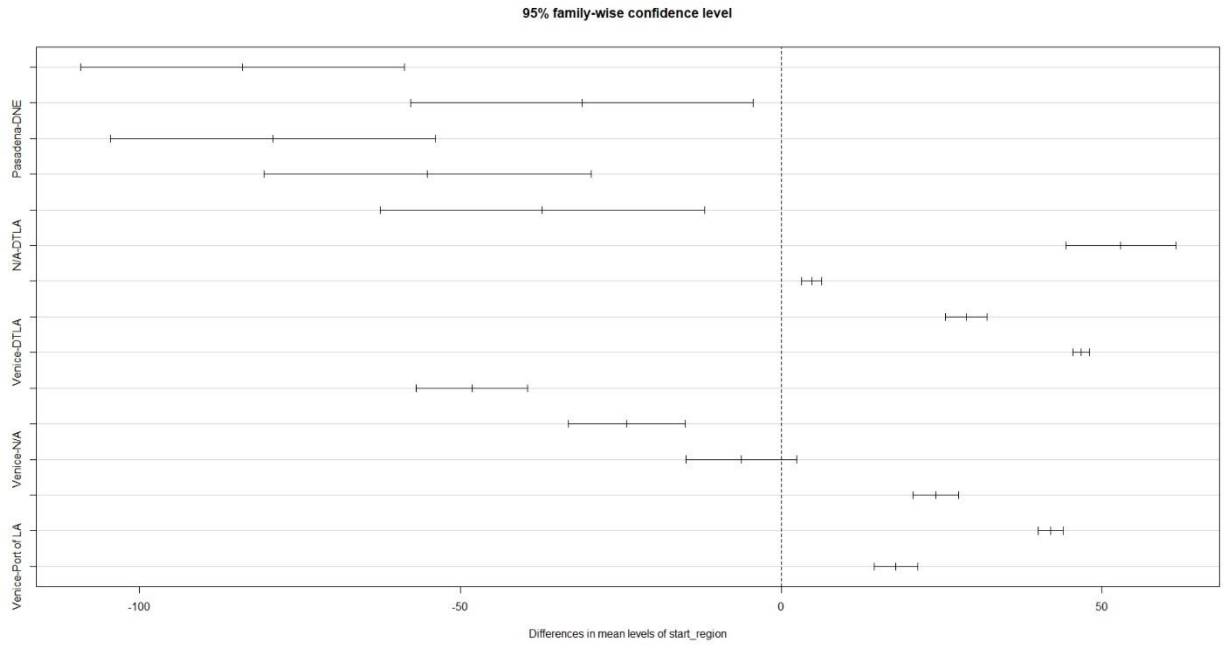



*Figure E2.* Plot of 95% confidence intervals for all of Tukey's multiple comparisons for difference in mean ride duration for different ride start regions.

```
Analysis of Variance Table

Response: duration
                              Df     Sum Sq  Mean Sq  F value    Pr(>F)
start_region                   5  158216259 31643252 2342.763 < 2.2e-16 ***
passholder_type                4  221249103 55312276 4095.140 < 2.2e-16 ***
start_region:passholder_type  16    3089241   193078   14.295 < 2.2e-16 ***
Residuals                 639760 8641115540    13507
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure E3.* ANOVA table for the ANOVA F-test for ride duration vs. factors of start region and pass type.

# Appendix F

## R Code

```
#read in data as LABikeData
LABikeData <- read.csv("~/Data Science
Competition/Data/LABikeData.csv")
View(LABikeData)
```

*Figure F1.* R code to read in the bike data into a manipulatable data frame and view it for initial analysis.

```
#extract just the day from the start_time into a seperate data frame
Days=as.Date(LABikeData$start_time, "%m/%d/%Y")
Days=as.numeric(Days)
View(Days)
```

*Figure F2.* R code to separate the specific day from all of the start time information.

```
#create a data frame called Region that holds the start region for each trip
N=nrow(LABikeData)
Region <- data.frame("Region" = 1:N)
for (j in 1:N) {
  for (i in 1:143) {
    if (Station_Table$Station_ID[i]==LABikeData$start_station[j]) { #Station_Table holds station data
      Region[j,1]=Station_Table$Region[i]
    }
  }
}
```

*Figure F3.* R code to match each start station for each trip to its particular region.

```
#calculate revenue from walk-ups (changing indices to see revenue for certain quarters)
revenue_walkup <- 0
N <- nrow(LABikeData)
for (i in 1:N) { #change indice bounds to corresponding indices for each quarter
  if (LABikeData$passholder_type[i] == "Walk-up") {
    if (durations$V1[i] <= 30) {
      revenue_walkup = revenue_walkup + 1.75
    }
    if (durations$V1[i] > 30) {
      revenue_walkup = revenue_walkup + ((((durations$V1[i])%/%30)+1)*1.75)
    }
  }
}
```

*Figure F4.* R code to calculate the revenue from walk-ups for different periods of time.

```
#calculate revenue from non-walk ups for rides over 30 minutes (per quarter as well)
revenue_over30min <-0
N <- nrow(LABikeData)
for (i in 1:N) { #change indice bounds to corresponding indices for each quarter
  if (LABikeData$passholder_type[i] != "Walk-up") {
    if (durations$V1[i] > 30) {
      revenue_over30min = revenue_over30min + ((((durations$V1[i])%/%30)+1)*1.75)
    }
  }
}
```

*Figure F5.* R code to calculate the revenue from non-walk-up rides that exceeded 30 minutes.

```
> #Average revenue per walk-up
> avgRevWalkUp <- revenue_walkup/nrow(walkUpPass)
> avgRevWalkUp
[1] 4.622911
> #Number of minutes for each walk-up ride
> avgMin = (avgRevWalkUp/1.75) *30
> avgMin
[1] 79.2499
```

*Figure F6.* R code to calculate the average revenue per walk-up and the average number of minutes per walk-up ride.

```
#plot types of passholder per region
library(ggplot2)
ggplot(LABikeData, aes(x=start_region, fill = passholder_type)) +
  geom_bar(width=0.4) +
  xlab("Region") +
  ylab("number of rides") +
  labs(fill = "passholder type") +
  ggtitle("Number of Rides vs. Region")
```

*Figure F7.* R code to plot the number of rides vs. start region.

```
#plot times of rides per region
library(ggplot2)
ggplot(LABikeData, aes(x=duration, fill = start_region)) +
  geom_bar(width=0.4) +
  xlab("Ride Duration (minutes)") +
  ylab("number of rides") +
  labs(fill = "start region") +
  ggtitle("Number of rides vs. Unique Ride Durations (w/ Outliers)")
```

*Figure F8.* R code to plot the number of rides vs. duration (including outliers).

```
#find the maximum length of a ride, how often it occurs and at which rides
max(LABikeData$duration)
which(LABikeData$duration==1440)
length(which(LABikeData$duration==1440))
```

*Figure F9.* R code to find the maximum length of all of the rides and to find how many times this occurs.

```
#removing these 1440 minute ride outliers and all other rides that went over
300 minutes
#chose 300 after looking at graph
outlier_duration <- which(LABikeData$duration > 300)
length(which(LABikeData$duration > 300)) #how many rides were over 300
minutes
duration_no_outliers <- LABikeData[-outlier_duration, ]
```

*Figure F10.* R code to remove duration outliers.

```
##plot number of rides vs. ride duration (filled with start region)
library(ggplot2)
ggplot(duration_no_outliers, aes(x=duration, fill = start_region)) +
  geom_bar(width=0.4) +
  xlab("Ride Duration (minutes)") +
  ylab("number of rides") +
  labs(fill = "start region") +
  ggtitle("Number of Rides vs. Unique Ride Durations (Outliers Removed)")
```

*Figure F11.* R code to plot the number of rides vs. duration (without outliers).

```
#plot the number of rides per day in each region
ggplot(duration_no_outliers, aes(x=day, fill = start_region)) +
  geom_bar(width=0.4) +
  xlab("day [1 (July 7 of 2016 Q3) - 908 (end of 2018 Q4)]") +
  ylab("number of rides") +
  labs(fill = "start region") +
  ggtitle("Number of Rides vs. Day")
```

*Figure F12.* R code to plot the number of rides per day in each region.

```
> #ANOVA for duration vs. start region and pass type: Is there a difference
between
> #the ride duration vs. region and ride duration vs. pass type and ride
duration vs. both factors?
> fit=aov(duration~start_region*passholder_type, data = LABikeData)
> anova(fit)
```

*Figure F13.* R code to conduct an ANOVA F-test for the difference in mean duration time due to factors of start region and passholder type.

```
#since we rejected the null and concluded there is a difference in means between durations
#across regions and/pass type, we conduct Tukey's procedure to see specific interactions
tukey = TukeyHSD(aov(duration~start_region+passholder_type, data = LABikeData),
conf.level=0.95)
plot(tukey)
```

*Figure F14.* R code to conduct Tukey's multiple comparisons and plot the corresponding confidence intervals for analysis.