Final Project

625.492 - Probabilistic Graphical Models

Molina Nichols

Due: 15 May 2018

Contents

List of Figures

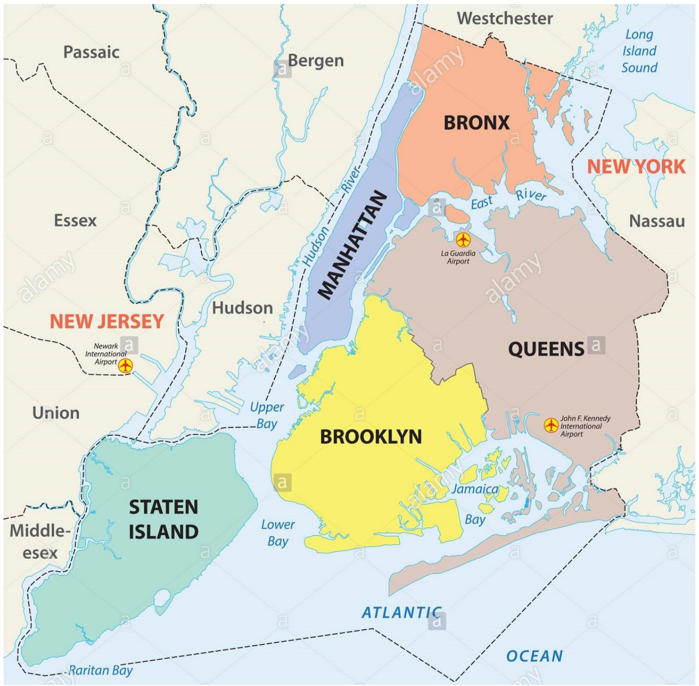

Figure 1: The five boroughs of New York City. Retrieved 23 April 2018, from [13].

1 Introduction

1.1 New York City

1.1.1 Overview and Demographics

New York City is a major city on the East Coast of the United States of America in the most south eastern region of the state of New York. As of July 2017, the population of the city was estimated to be 8,622,698 by the United States Census Bureau, an increase from the 8,175,133 measured by the April 2010 United States Census [3]. New York City outstrips the next closest U.S. city in population, Los Angeles, by about 4.3 million people, and makes up approximately 43% of the total population of the state of New York. With a population density of approximately 27,000 people per square mile over its 302 square miles, New York City also holds the title of the most densely populated city in the United States [4]. The city is spread over five "boroughs" - Brooklyn (approx. pop. 2.5 million), Queens (approx. pop. 2.2 million), Manhattan (approx. pop. 1.6 million), The Bronx (approx. pop. 1.4 million), and Staten Island (approx. pop. 0.5 million) [3]. Along with the boroughs of New York City, the surrounding areas of Long Island, the south eastern Hudson River Valley, north eastern New Jersey, and south western Connecticut all combine to form the largest metropolitan population in the United States, comprising approximately 20 million people [5]. The boroughs and some of the surrounding metropolitan area can be seen in Figure 1.

New York City, particularly the island borough of Manhattan, is known as a global center of finance, media, and culture, prompting this large population concentration. Financial institutions like Wall Street and the New York Stock Exchange, media such as Broadway and many other performing arts venues, visual arts and fashion, many educational institutions, and a plethora of restaurants and fine dining all define the character of New York City. Many people in the metropolitan area commute into the city to work in these industries, as well as cross-borough commuters. Additionally, New York City sees many tourists each year, drawn to the culture and lifestyle that can be experienced there.

1.1.2 Commuting and Tourism

We will hereafter focus primarily on the borough of Manhattan, given its dominance as the center of both commuting and tourist destinations. Estimates put the number of daily weekday commuters into Manhattan at approximately 1.6 million, while only about 130,000 of the 1.6 million population of Manhattan leaves the borough for work elsewhere [6]. The population of Manhattan can up to double on a weekday, encompassing crowds for special events, day-trippers, and vacations in addition to regular commuters. Around 80% of those 1.6 million commuters use public transit to enter Manhattan. Public transit options to serve that population are abundant, primarily rail, including the New York City Subway, New Jersey Transit, PATH (Port Authority Trans-Hudson), Long Island Railroad, Metro-North Commuter Railroad, and Amtrak trains and MTA (Metropolitan Transit Authority) buses. Other commuting options may include walking, bike, ferry, personal car, taxi, and ride sharing (Uber, Lyft, and similar services). While not the most prevalent method, some of the commuters do choose to bike, numbering around 14,500 per day inbound to Manhattan in 2007 [7] and gradually increasing over the years from 2000 to 2007 as seen in Figure 2.

In addition to its large commuter population, New York City also welcomes a tremendous number of domestic and international tourists each year. This number has increased each year since at least 2010 with 48.8 million visitors, to 2015 with 58.5 million visitors, to 2016 with 60.5 million visitors, and to 2017 with 62.8 million visitors [8] [9]. Many of these tourists are from the metropolitan area and the surrounding areas as well as from across the country (especially more affluent states), including Florida, Massachusetts, Maryland, California, Virginia, and Texas. A majority of international visitors come from countries such as the United Kingdom, China, Canada, Brazil, and France [8] [9]. These tourists come each year to observe the culture and entertainment offerings in New York City, and play a vital role in contributing revenue to these types of industries.

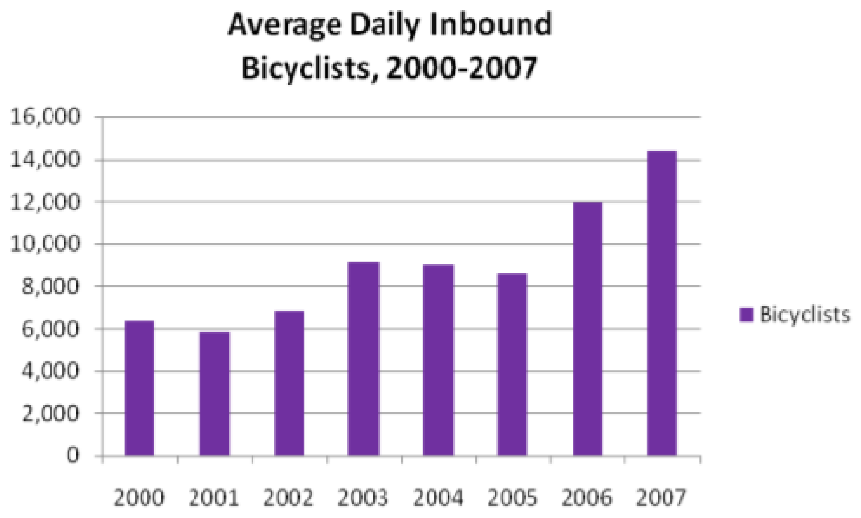

Figure 2: Average number of bicyclists, inbound daily to Manhattan, 2000-2007. Retrieved from [7].

1.2 Citi Bike

1.2.1 Inception and Growth

Citi Bike is a public bike sharing system in New York City. Similar sharing systems exist around the world and in U.S. cities such as Washington D.C., Boston, Chicago and many others, with Citi Bike being the largest of all of these in the U.S. [2]. Citi Bike is spread over parts of the New York City metropolitan area, with stations in the boroughs of Manhattan, Brooklyn, and Queens and across the Hudson river in Jersey City, New Jersey. The introduction of Citi Bike was announced by the New York Department of Transportation in September 2011. By May 2012, more plans were underway with major sponsors Citibank and Mastercard. The system officially launched in May 2013 with 6,000 bikes and stations in only Manhattan and Brooklyn. By August 2015, the system had expanded to include new stations and 8,000 bikes. A year later in August 2016, the system encompassed 10,000 bikes with added stations in Jersey City. By September 2017, the system boasted 12,000 bikes and new stations in Queens, and in October 2017 marked its fifty millionth trip taken on a Citi Bike. Currently, the system still has 12,000 bikes with 750 stations [1].

1.2.2 The System

The bike sharing system consists of bike rack stations throughout the city where users can pay to take out a bike for a given amount of time (in the case of Citi Bike, rides are permitted to be up to 30 minutes, or more for an additional fee). Users will pay or swipe their membership card, “undock” the bike from its locked station, ride as they please, and then “dock” the bike again at any station of their choosing near their destination. See Figure 3 for an example of a docking station

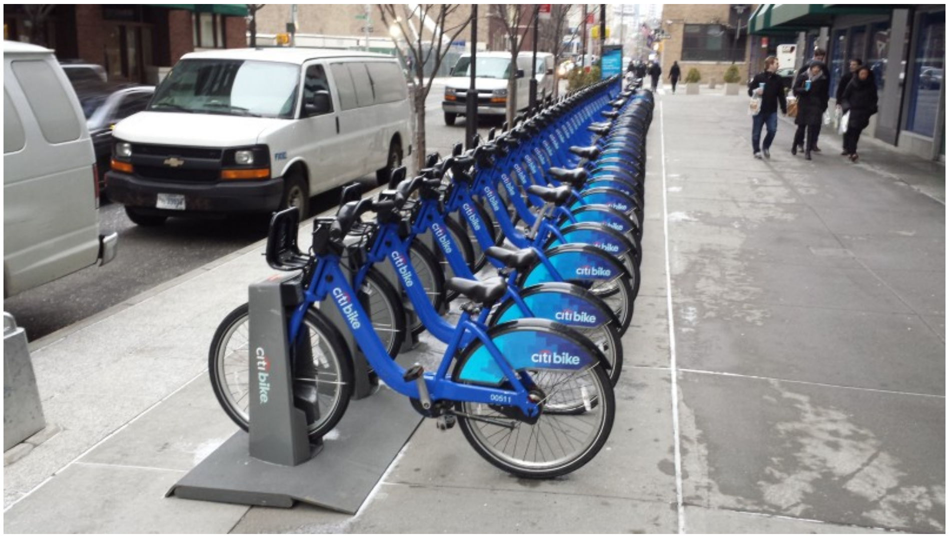

Figure 3: A Citi Bike docking station.. Retrieved 24 April 2018, from [12].

with bikes.

Citi Bike offers four pricing levels - a single ride for $3, 1 day (24 hours) of unlimited rides for $12, 3 days (72 hours) of unlimited rides for $24, or 1 year of unlimited rides for $14.95 per month, approximately $180 annually (users must commit to the full year plan even if they are paying monthly). This type of bike sharing system is appealing to city residents and tourists alike for a variety of reasons. For city residents, Citi Bike may be cheaper than other transportation options; taxis in New York City are quite pricey and a monthly card for unlimited rides on the subway costs $121 each month [10]. Given dense vehicular traffic in the city and occasional disruptions in subway service, biking could also be more reliable and quicker than other options. In addition to these reasons, a bikeshare could be more effective than a personally owned bike because the bike does not need to be stored in a small city apartment and does not need to be protected against vandalism like a personal bike would be. The flexibility of Citi Bike docking stations also means that one may choose biking for one direction of a trip but switch to another mode of transportation for the other direction of the trip if it is more convenient. For tourists, a bikeshare like Citi Bike offers a fun way to see New York City or outdoor areas such as Central Park, may be a way to get around faster, and a unique experience. For both city residents and tourists, a bikeshare system provides the general benefits of reducing vehicle congestion, availability of a form of exercise, and a fun activity.

1.2.3 The Data

Citi Bike publicly releases their biking data on their website. The data has been collected and posted since the launch of the system in May 2013 until the present as of March 2018 [11]. The data is provided per month, and appears to all be posted on the Citi Bike website with no significant gaps. This data is collected into comma-separated value (CSV) files with the following fields:

- Trip Duration (integer number of seconds)
- Start/End Time (24 hour clock, HH:MM:SS)
- Start/End Station ID (integer identifying the station)
- Start/End Station Name (plaintext of the intersection where the station is located)
- Start/End Station Latitude (float identifying the latitude of the station)
- Bike ID (integer identifying the bike in use)
- User Type (plaintext for type of user, Subscriber=annual subscription user, Customer=single-ride, 1-day, or 3-day pass user)
- Birth Year (integer birth year of the user, may be unknown)
- Gender (integer for gender of the user, 0=unknown or not provided, 1=male, 2=female)

It is important to note that the exact route traveled on the bike is not provided - the data simply provides the beginning and ending points of the trip at each docking station.

2 Data Analysis

2.1 Overview

For this analysis, the dataset from June 2017 was used as the sample from the Citi Bike data. This subset was chosen to keep the size of the data manageable. Additionally, it was assumed that a month like June would exhibit a wide variety of ridership and the most representative sampling of Citi Bike use. This is because the weather in June in New York City tends to be warm and favorable, encouraging those who use a bike to commute to do it more regularly, and inviting city residents to simply take a "joyride" on the bikes for day. Additionally, while June is not the height of tourism in New York City, it is still a month with a significant number of tourists as compared to other less busy months, with good weather contributing to a higher likelihood of them utilizing the bikes. In processing the data, it was found that the data was mostly complete with a few adjustments needed. A derived data point for age was calculated from the birth year provided for each customer; however, it was found that in some places the birth year was either missing or unreasonably early. To fix these issues, any customers with birth years making them 85 years of age or older were reassigned to have an 'unknown' age. Even this threshold may have been overly

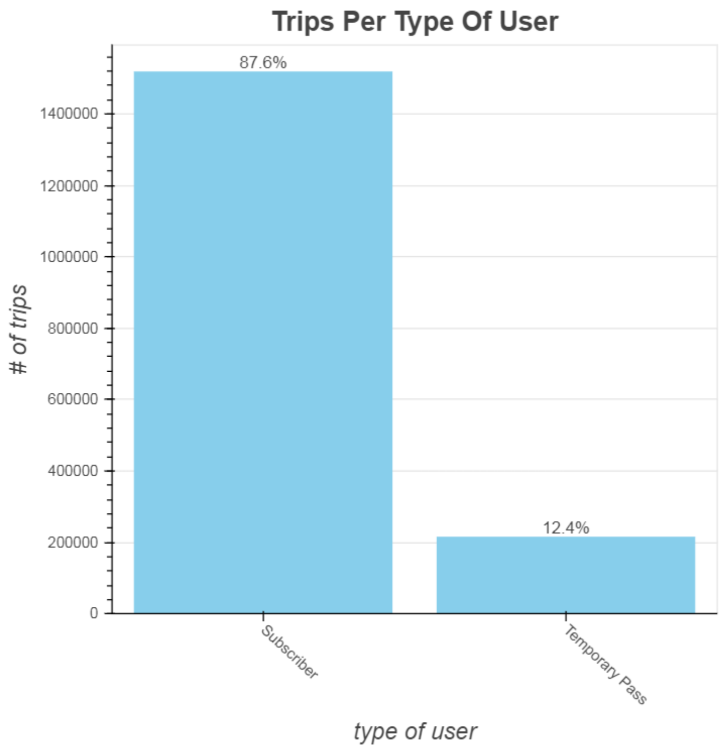

Figure 4: Trips taken per type of user.

generous to older riders, but there was no true way to know where the cutoff should have been. The vast majority of riders were under 65 however, so no significant issues were associated with this estimation. Other derived data points were created for ease of analysis, such as the day of the week on which a ride occurred, and a few others. Otherwise, the data was used “as-is” for analysis.

2.2 Data Characteristics

Within the June 2017 data set, there are 1,731,594 rows (rides taken) total. For all of the following figures, the percentage over each bar indicates that data field’s percentage of total rides taken in the entire data set unless otherwise indicated. It seems most appropriate to first examine the data-set by type of user so as to get a sense of who participates in the bikeshare the most. In Figure 4, we see that there are far more subscribers (87.6% of total rides taken) than 1-day, 3-day, or single ride short-term customers (12.4% of total rides taken). Later in our analysis, this may point to searching for more patterns among commuters using the bikes than among daytrippers and tourists.

In Figure 5, we see that the gender distribution of the riders is more imbalanced than expected, with ∼65% of rides taken by male users, ∼23% of rides taken by female users, and ∼11% of rides taken by users with the gender unspecified. Perhaps male users are more likely to take bike trips at any hour of the day, whereas female users would choose more limited hours for safety. Perhaps also female users are less likely to report their gender to the system or perhaps male users are just

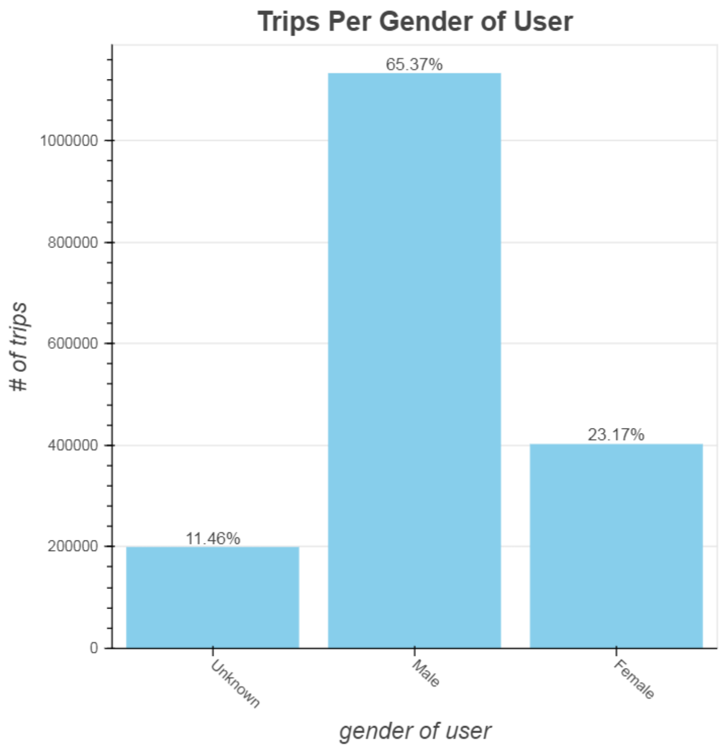

Figure 5: Trips taken per gender of user.

more likely to use the system overall - we have no real insight into why this disparity might be.

We also examine the trips taken by their duration to determine what the bikes could be used for. As seen in Figure 6, the vast majority of trips taken (∼77%) fall into reasonable commute times of 0-20 minutes. Plenty of trips, ∼8.7%, incur the additional fee imposed by Citi Bike for trips lasting longer than 30 minutes.

Age of the users may also provide insight into their riding patterns, as shown in Figure 7. Unsurprisingly, most rides were taken by 25-35 year olds. This demographic tends to be the most physically fit, would be the most likely to hold a regular weekday 9-5 job, and be the most likely candidates for using a bike to commute to work. Due to their physical fitness and general hobbies, this group may also be some of the most likely to use the bikes recreationally. Those in the 35-45 year old range also represent a significant chunk of ridership. As discussed in later sections and evidenced by the large subscriber to short-term customer ratio as seen in Figure 4, the dominance by these age groups is likely due to their use of Citi Bike as a commuting option.

2.3 Temporal Trends

It is also important to understand when the bikes are ridden to understand trends in the data. First, we examine how often the bikes were ridden during each of the 30 days of the month of June 2017, shown in Figure 8. Ridership was generally higher on weekdays, pointing to heavy use by

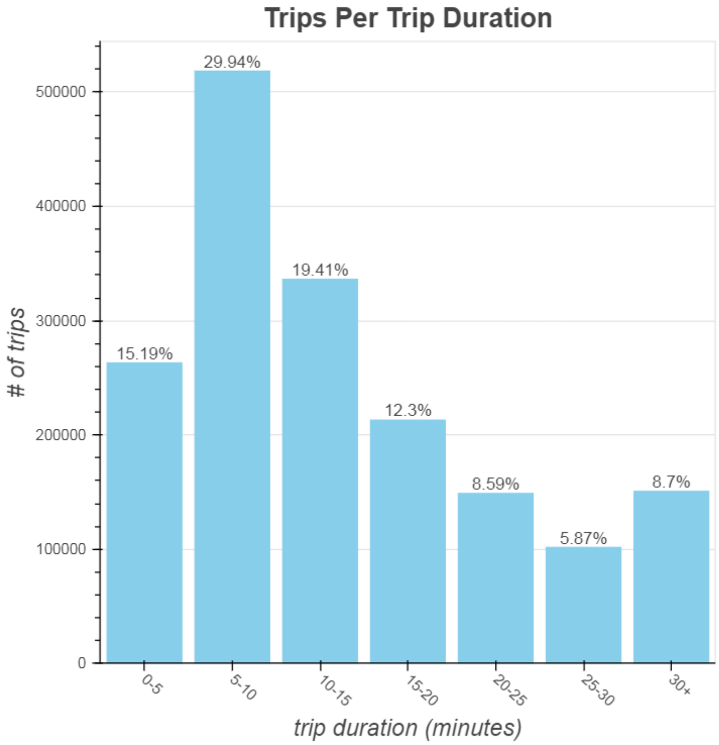

Figure 6: Trips taken binned by trip duration.

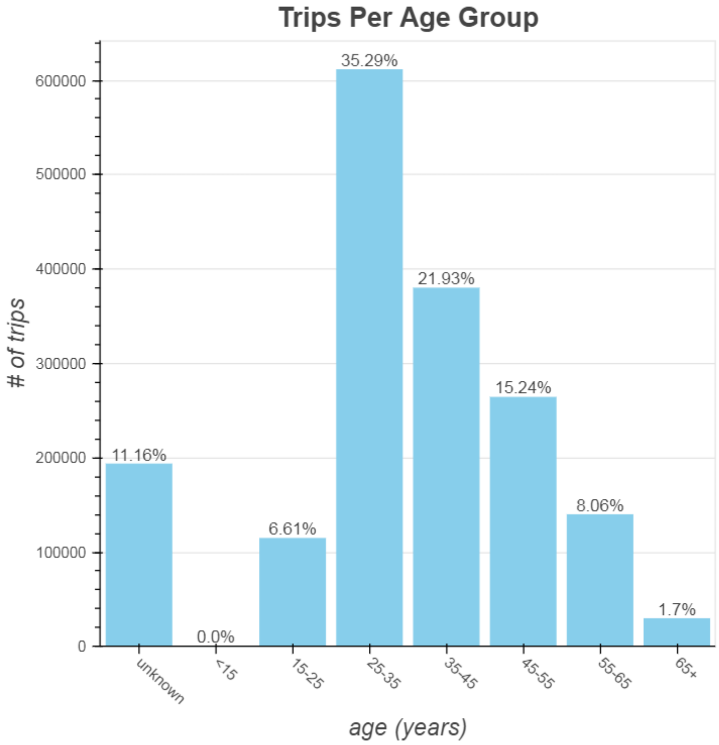

Figure 7: Trips taken binned by age of user.

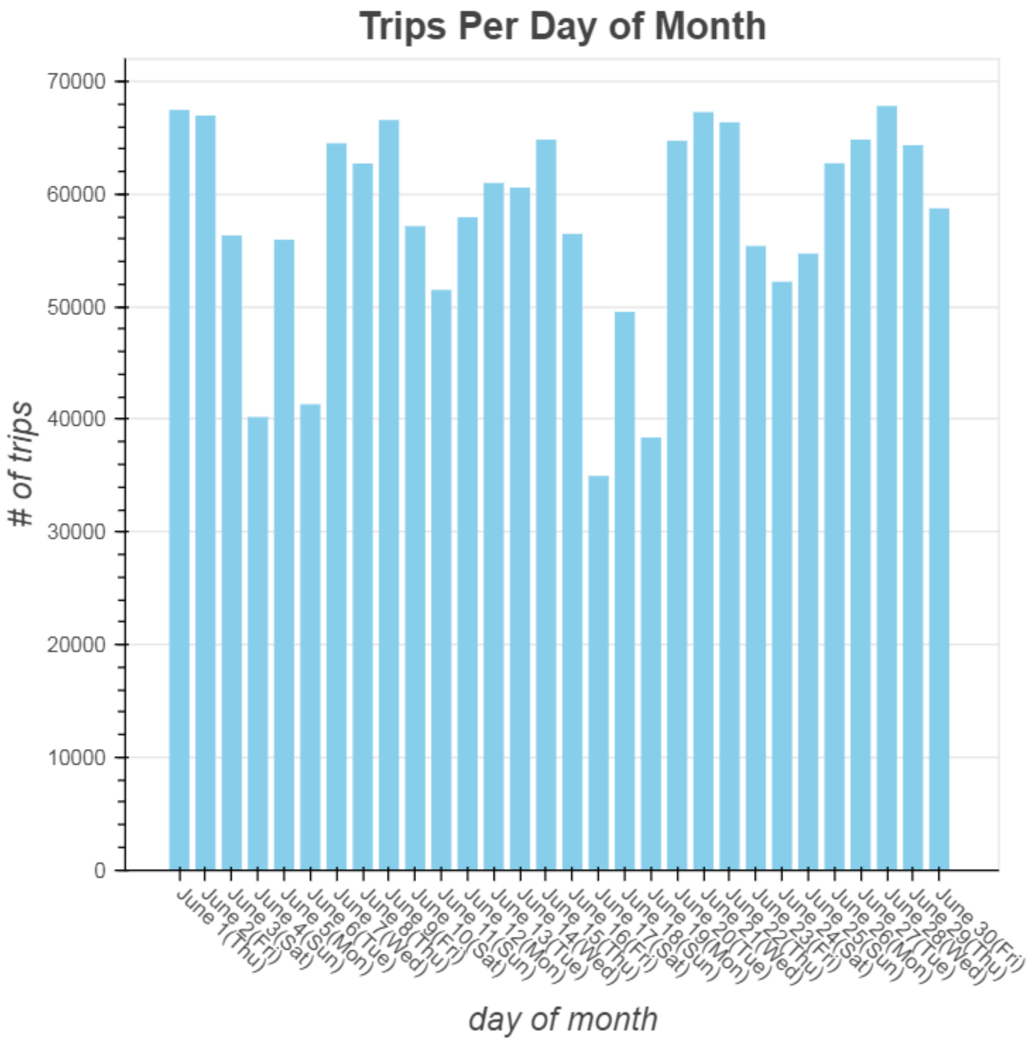

Figure 8: Trips taken per day of the month.

commuters since subscribers make up 87.6% of all Citi Bike rides taken. Otherwise, there is nothing abnormal or especially interesting to note about ridership throughout the month (as we would have noted if there was an out of the ordinary occurrence, such as a special event or a holiday).

Going along with the weekday trends we noticed across the whole month, we also examine trips per day of the week Monday through Sunday, as seen in Figure 9. We note that this data had to be "normalized" to represent 4 instances of each day during the month, as the data initially included 4 Saturday-Wednesday and 5 Thursday-Friday, artificially inflating the number of trips that appeared to take place on Thursday and Friday. To perform this normalization, we totaled the number of rides occurring on Thursday and multipled by $\frac{4}{5}$ to obtain a representative quantity for if the month had included only 4 Thursday's. We performed this reassignment for Friday as well. We see now that this data supports our day of month data, showing the highest ridership on

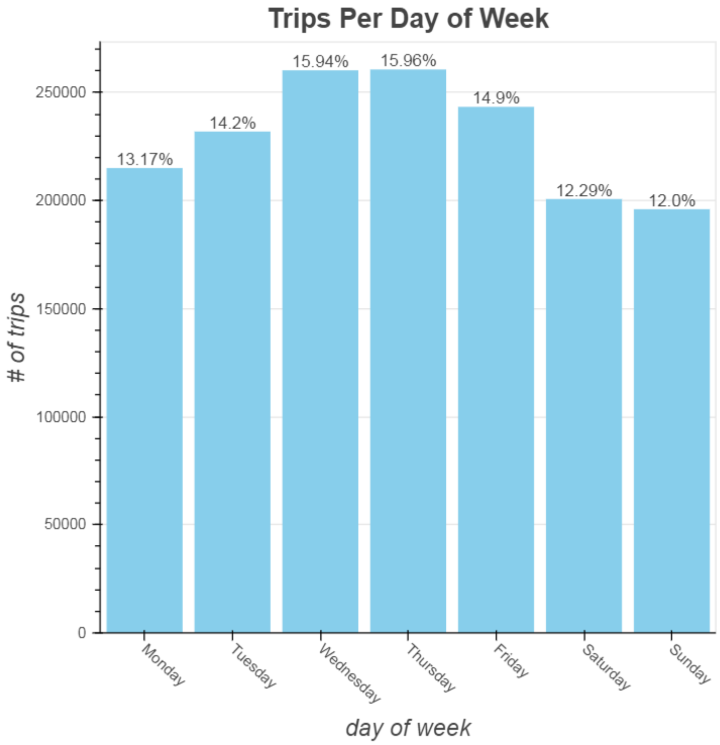

Figure 9: Trips taken per day of the week.

Thursday, Wednesday, Friday, Tuesday, and Monday, respectively, with a slight dip on Saturday and Sunday. The Saturday and Sunday percentages are higher than one might expect having just seen Figure 8, but a few higher than average days on the weekends (perhaps extra nice weather?) drove those percentages up a bit. However, we see that ridership is still dominated by weekdays, due to subscribers using the Citi Bikes to commute.

To further examine the reasons behind our results in Figure 9, let us investigate which hours of the day the bikes are most likely to be used. We refer to a 24-hour clock, running from 00 (midnight) to 12 (noon) to 23 (11:00 P.M.) in Figure 10. Again supporting our popular weekday trips evidence, ridership peaks during the hours of 7-9 and 16-19, with the most ridership occurring between 8-9 and 17-18. This is clearly driven by the weekday commuters we saw in Figure 9 with the usual 9-5 work schedule of New York City professionals, with some leeway on each end for different working hours or for commuting time. Figure 10 is also separated into the hour in which the given trip started or stopped. For example, if a user departed their workplace and took a bike from a nearby station at 16:45, rode home, and docked the bike near their residence at 17:10, that trip will have one point counted for the "start" bar during 16 and one point counted for the "stop" bar during 17. In this way, we may see bike usage with more granularity than simply approximating when trips were taken and we do not need to bin trips that overlap hours into just one of those hours. We also note the nontrivial number of trips between the hours of 9-15. This could particularly be the

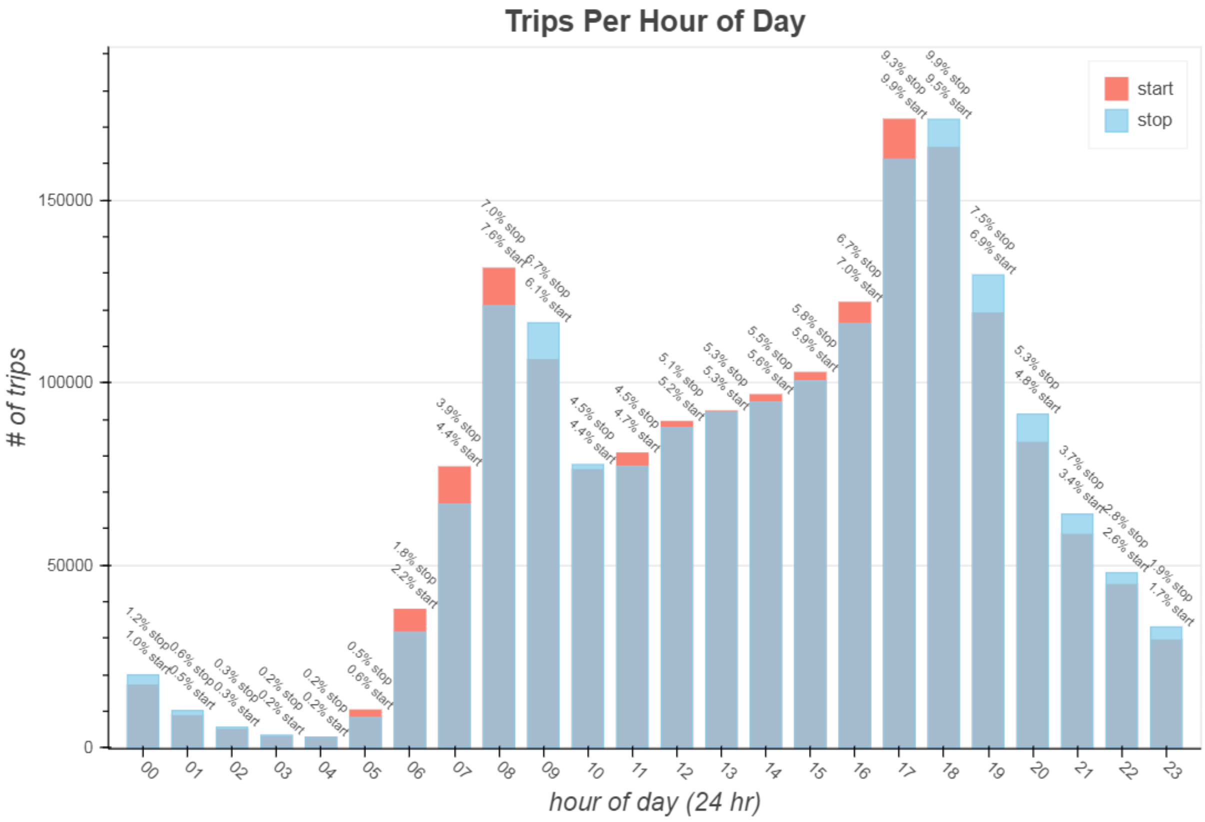

Figure 10: Trips taken per hour of the day.

influence of short-term (or possibly subscriber) trips taken on weekends, those commuters taking a bike out for a break or trip during their lunch hour, or a number of other cases.

2.4 Geographic Trends

We next examine the geographic trends in Citi Bike usage, particularly as they pertain to commuting habits and tourist or joyriding destinations within New York City.

We first consider the most popular starting and ending locations for trips taken on the Citi Bikes. Please refer to [17] throughout this section for a thorough understanding of these locations and to put them in the context of neighborhoods, businesses, and attractions within New York City. Each labeled layer of the Google Map in [17] will allow interactive exploration of these sets of stations in context of New York City, as well as compared to each other.

The street intersections of the most popular origin stations can be seen in Figure 11 along with the percentages of all rides starting at these stations. We of course ask - why these stations? We use the top 5 stations to understand what makes them so popular as locations to start a Citi Bike trip. The first, Pershing Square, with ∼1% of all rides originating there, is directly outside of Grand

Central Terminal, a major transportation and commuter hub, as well as a tourist draw. The second, West St & Chambers St, is located in the south-eastern Manhattan neighborhood of Tribeca, near ball fields and a waterfront park, as well as only 4 blocks from both the World Trade Center and New York City Hall, making it a likely choice for commuters and tourists alike. The third, E 17 St & Broadway is located at the corner of the popular Union Square Park in Manhattan, and could be the beginning of a ride north on Broadway toward Midtown, a neighborhood with many tourist and business opportunities. The fourth, Broadway & E 22 St, is directly next to the Flatiron Building, an office space and very popular tourist attraction. The fifth, 12 Ave & W 40 St, is located next to the Hudson River Greenway, a bike and walking path that runs the entire west side of the island of Manhattan. As seen in [17], several of the other stations on this top 20 list are near Central Park (popular for recreation), outside Pennsylvania Station (another tourist and commuter transit hub), in the vicinity of Times Square (the most popular tourist attraction in all of New York City), and near the World Trade Center (where many offices are housed). All of these locations make sense as popular stations at which to begin a Citi Bike ride. The most popular destinations from these top 20 most popular origin stations fall largely into the list of most popular destinations overall (these are denoted by a ' * ' preceding the station name), and additionally include several parks such as Pier 40 on the Hudson River Greenway, the High Line walking trail, and Central Park. In fact, $\sim$17.7% of trips originating from these stations end at the stations shown in Figure 12 with the percentage of trips from the top destinations ending at those stations displayed on the figure.

We next move to the most popular destinations stations as seen in Figure 13 with the percentages of all rides ending at these stations shown. As with the most popular origin stations, we examine the geographical reasons for the results we see. The top 5 destination stations are actually the very same as the top 5 origin stations, though in a slightly different order. The most popular origins of these top 20 most popular destination stations, analagously to before, fall mostly into the set of most popular origins overall (these are denoted by a '$\sim$' preceding the station name) with a few others in various neighborhoods. $\sim$18.6% of trips ending at these top 20 destinations originated at the stations shown in Figure 14 with the percentage of trips ending at the top destinations originating at those stations displayed on the figure.

With this analysis, we also see that 18 of the top 20 stations for starting and ending are in fact overlapping. The two unique top 20 starting stations are Grand Army Plaza & Central Park S (likely due to its location at the south-eastern tip of Central Park, closest to Midtown and Times Square) and Cooper Square & E 7 St (possibly due to its location near the Facebook New York City headquarters and the very residential East Village neighborhood). The two unique top 20 ending stations are E 7 St & Avenue A (in the exact center of the East Village, a likely place to return home from work or a fun excursion) and Christopher St & Greenwich St (on the west side of Greenwich Village, a very residential area which is home to New York University).

We also examine trends we have alluded to previously, distinguishing subscribers from short-

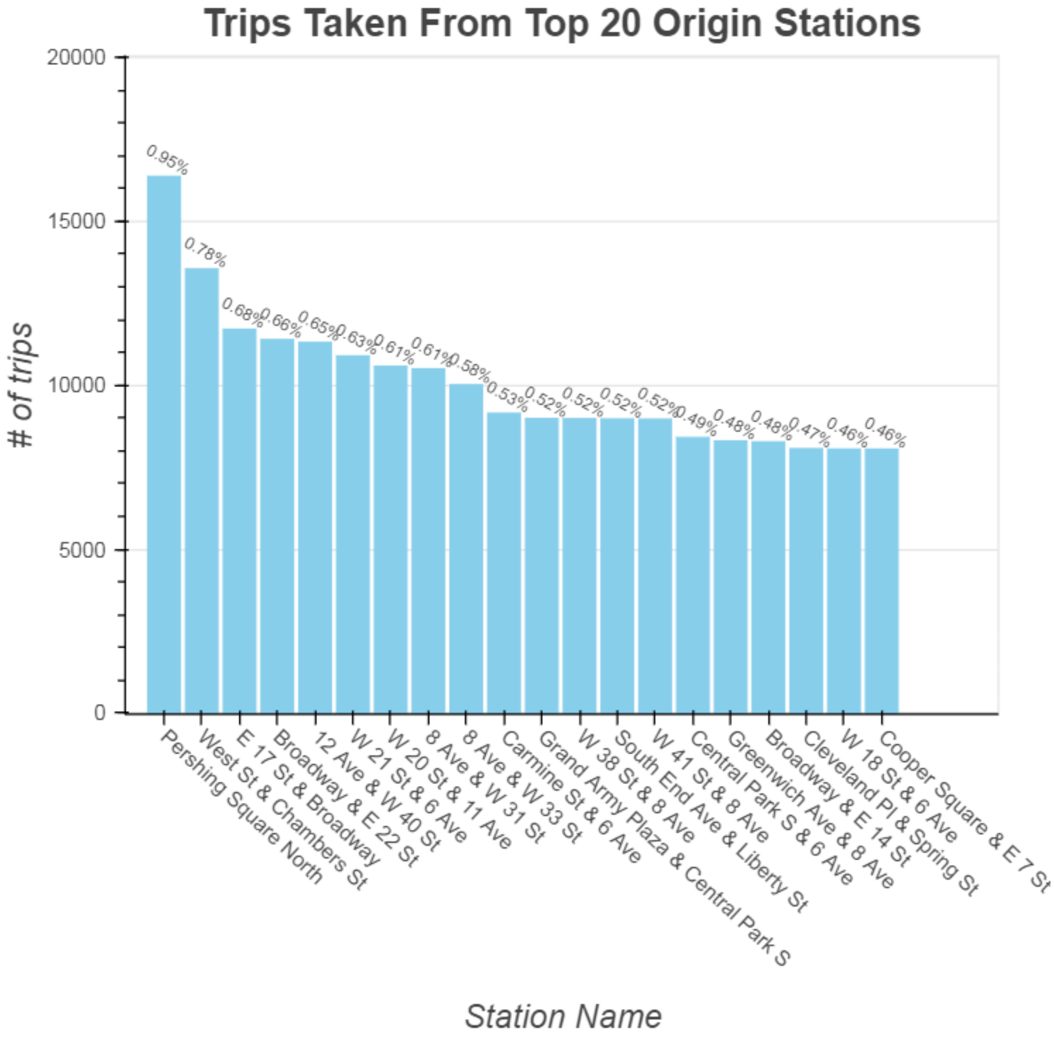

Figure 11: Trips taken from the top 20 most-used origin stations.

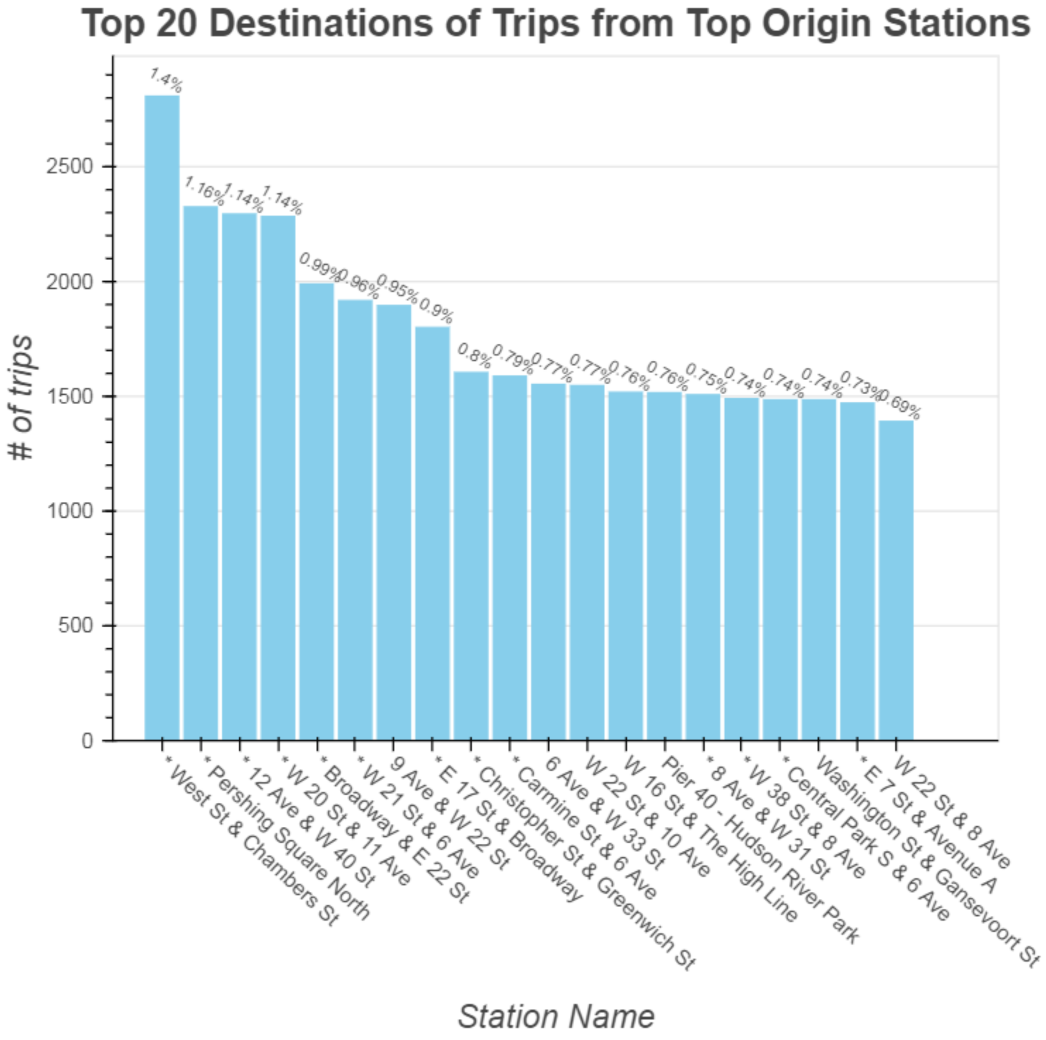

Figure 12: Top destinations of trips taken from the top 20 most-used origin stations.

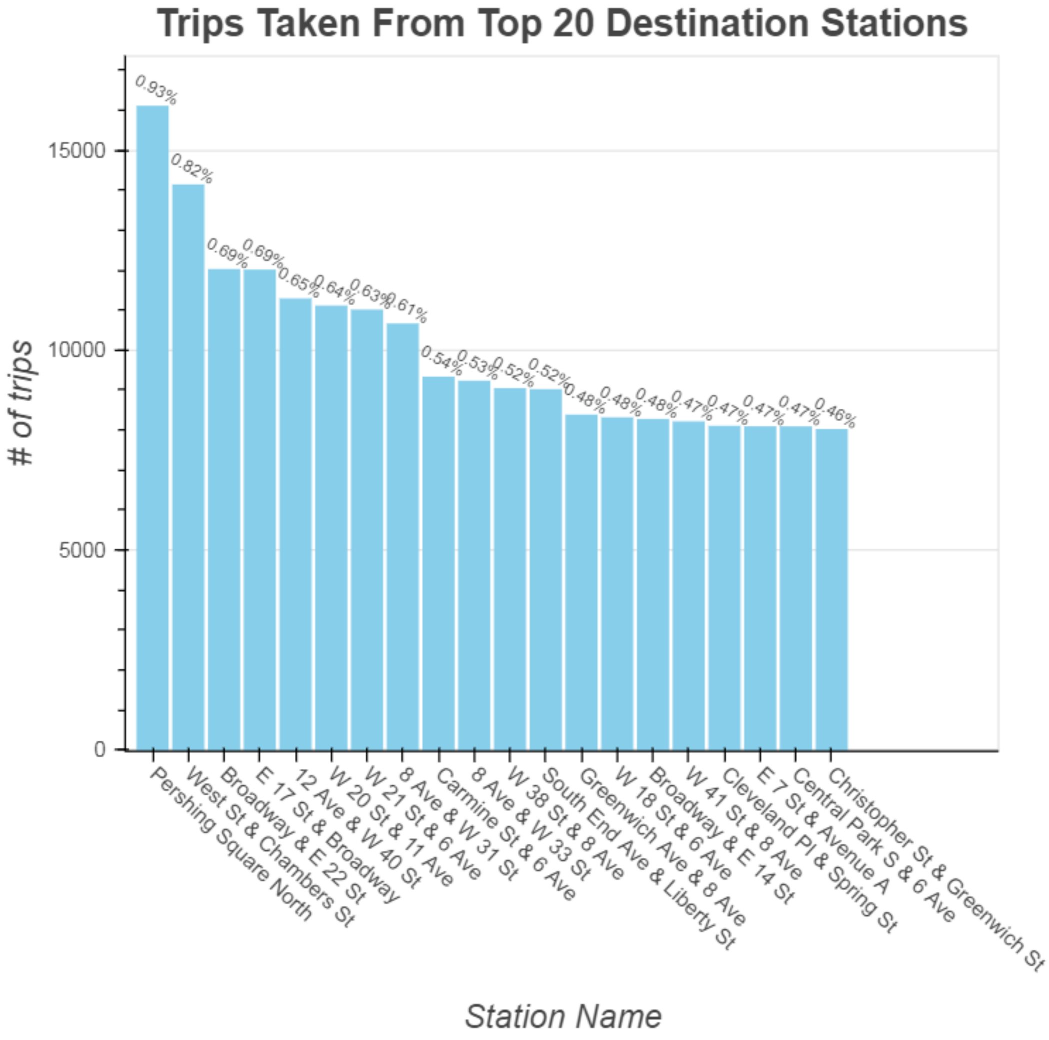

Figure 13: Trips taken to the top 20 most-used destination stations.

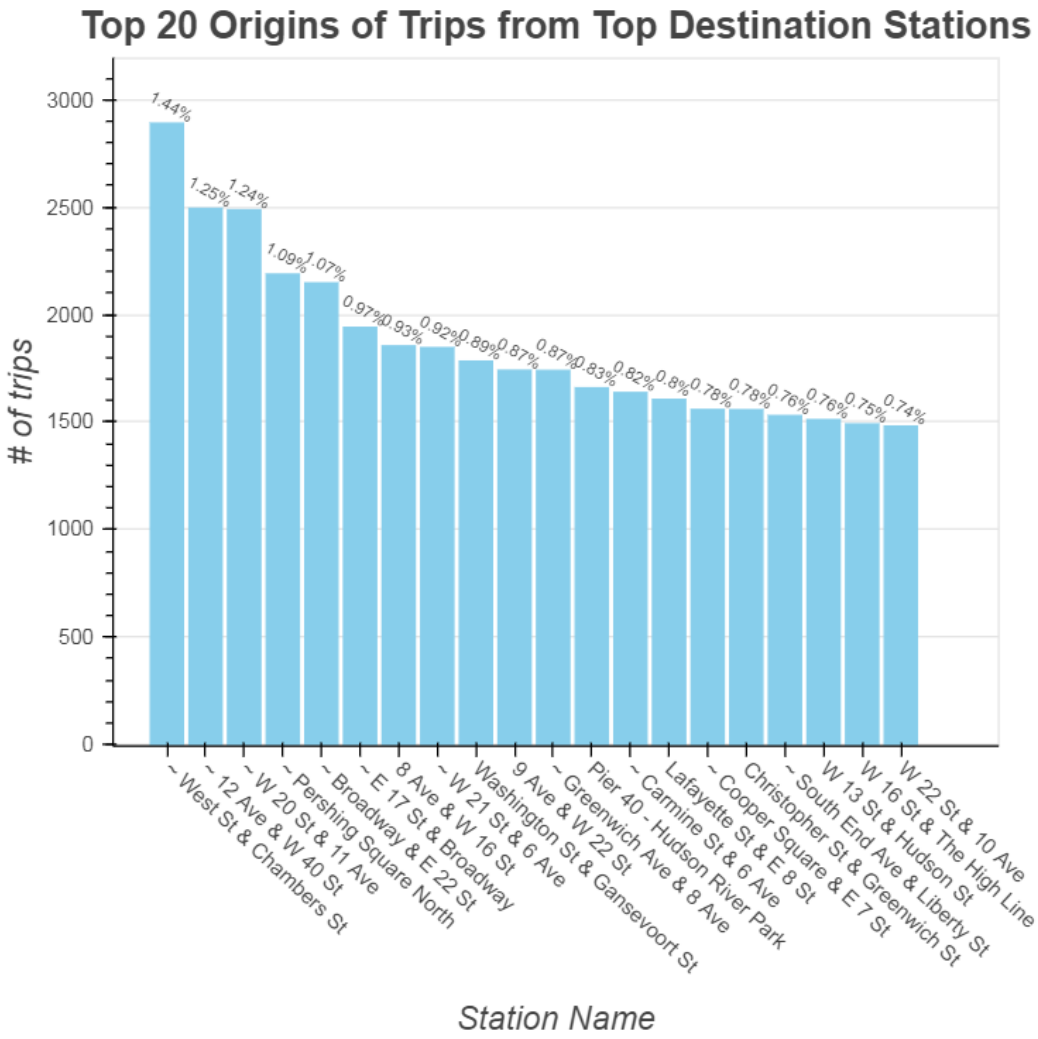

Figure 14: Top origins of trips taken to the top 20 most-used destination stations.

term customers. This provides an even more interesting look at the way the Citi Bike stations are used since the trends between the groups of users are more starkly divided. Please refer to [19] throughout this section to use the interactive Google Map for context. The stations most used by short-term customers are shown in Figure 15. These stations are clearly very concentrated around Central Park, the Hudson River Greenway and piers and attractions on the Hudson River on the west side of Manhattan, the southern tip of Manhattan at Battery Park, Wall Street, and the Statue of Liberty ferry, and at each end of the Brooklyn Bridge. All of these locations are top tourist attractions, as well as attractive and fun destinations for local New York City residents to take a ride for fun. The stations most used by subscribers are shown in Figure 16. These stations are clustered near commuter hubs at the World Trade Center, Pennsylvania Station, and Grand Central Station. They are also primarily located in majority residential areas in Greenwich Village and the East Village, likely living arrangements between possible locations of work in the Midtown or Wall Street areas.

2.5 The Rebalancing Problem

The so-called "rebalancing" problem is is a challenge that Citi Bike has previously encountered and continues to work through in their daily operation [14] [15]. This problem refers to a shortage or surplus of bikes that may occur at certain stations throughout the city as a result of imbalanced riding patterns. This problem manifests itself in one of two ways - either someone has come to return their bike, has found a full dock, and must find a different station at which to return their bike or someone has come to take out a bike only to find an empty station with no bikes available. For example, stations near commuting hubs tend to empty out quickly at the morning rush hour as users take the bikes and ride to their place of work, while those very same stations may become overcrowded with no docking locations come evening rush hour as commuters return to the commuting hubs to travel home. Of course both of these situations are sure to induce customer frustration and it is in the best interest of Citi Bike to understand the riding patterns to aid in avoiding these situations. Methods to remedy the problem include offering incentives to "bike angels" to ride the bikes to very empty stations or away from very full stations and moving the bikes on trucks. To more fully understand this problem, we investigate which stations have the worst "balancing ratios", simply defined as $\frac{numberofridesoriginatingatstation}{numberofridesendingatstation}$. Thus, a high balancing ratio indicates a station where more rides start than end (leading to empty docks), a low balancing ratio indicates a station where more rides end than start (leading to full docks), and a balancing ratio near 1 indicates a near equal number of rides starting and ending at the station. Please refer to [18] throughout this section to see the stations discussed placed on a Google Map for full geographical understanding.

We first consider the stations with the highest balancing ratio. Note that the 3 stations with

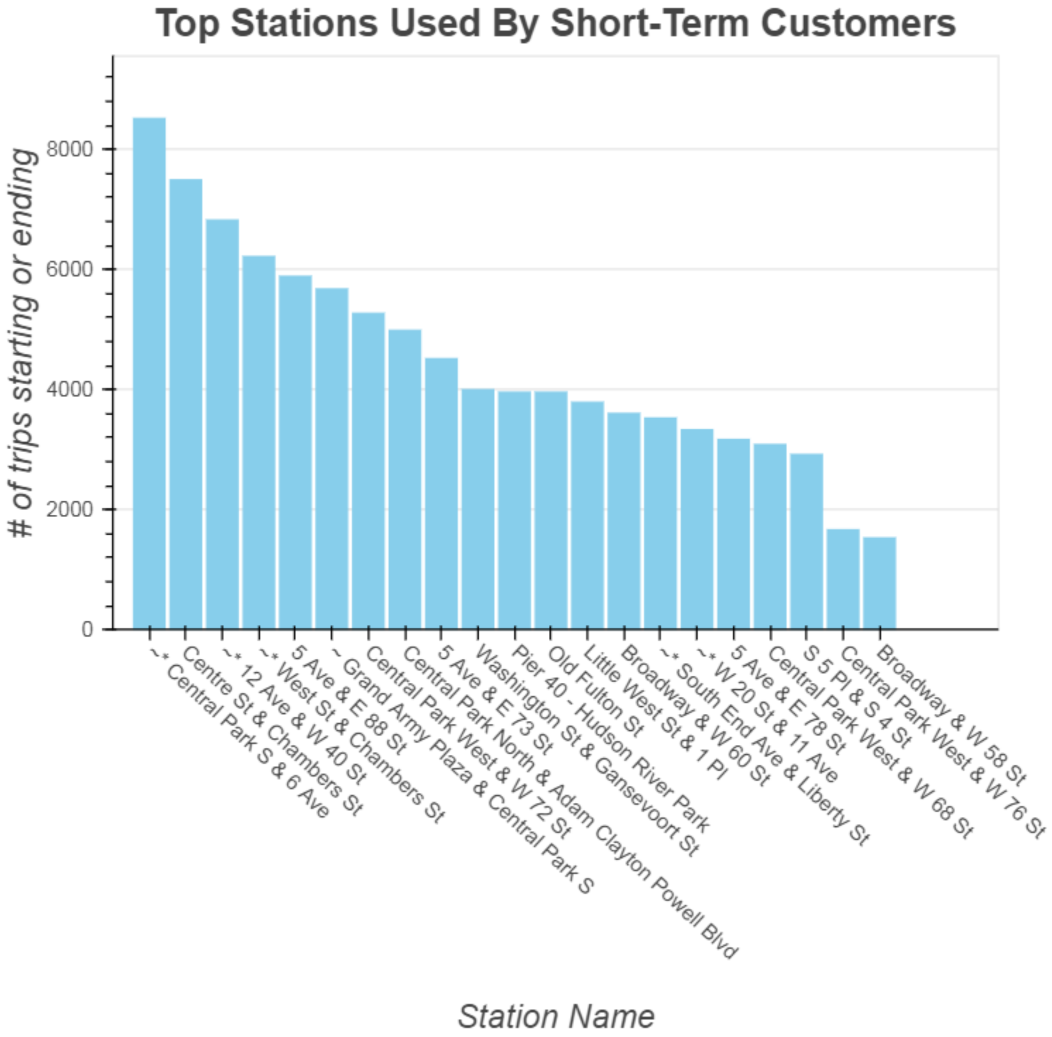

Figure 15: Most-used stations by short-term customers.

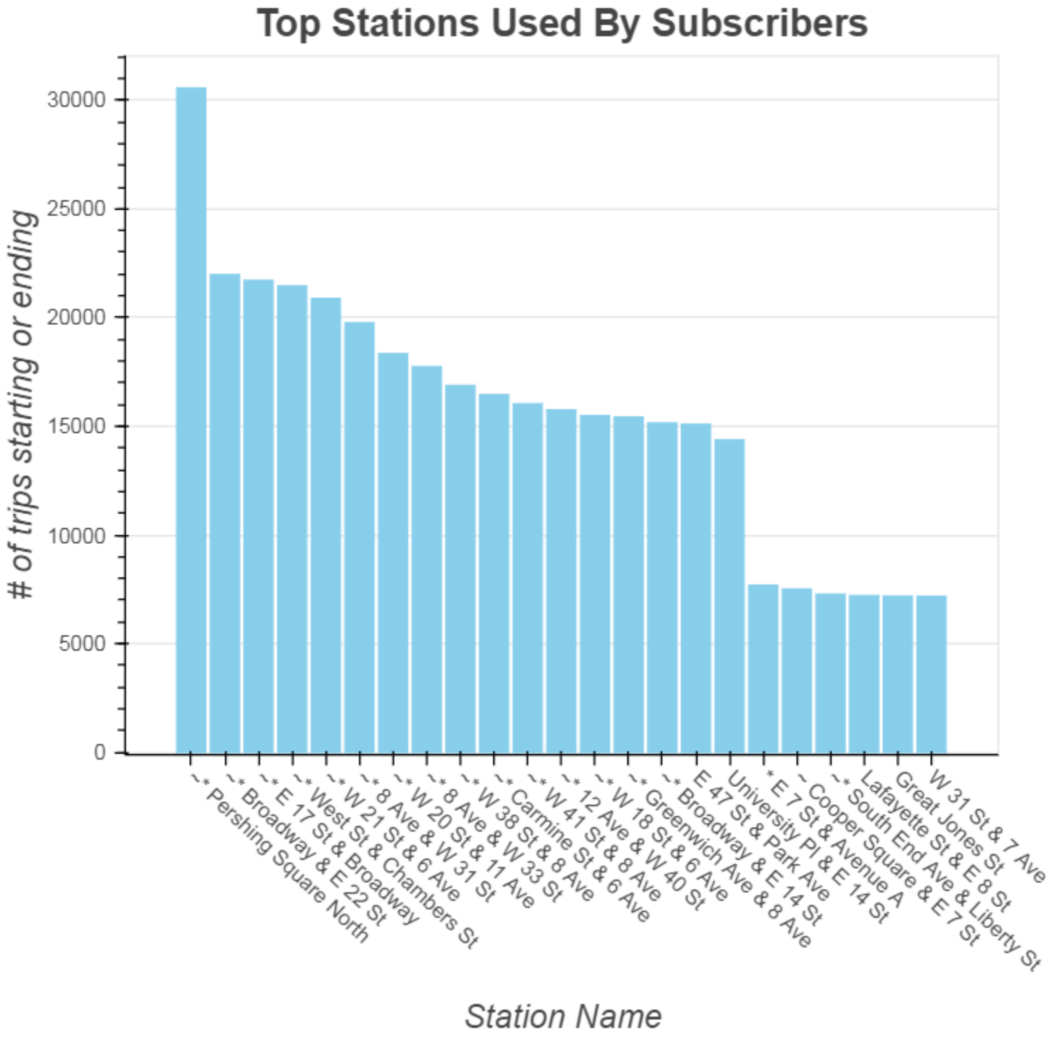

Figure 16: Most-used stations by subscriber customers.

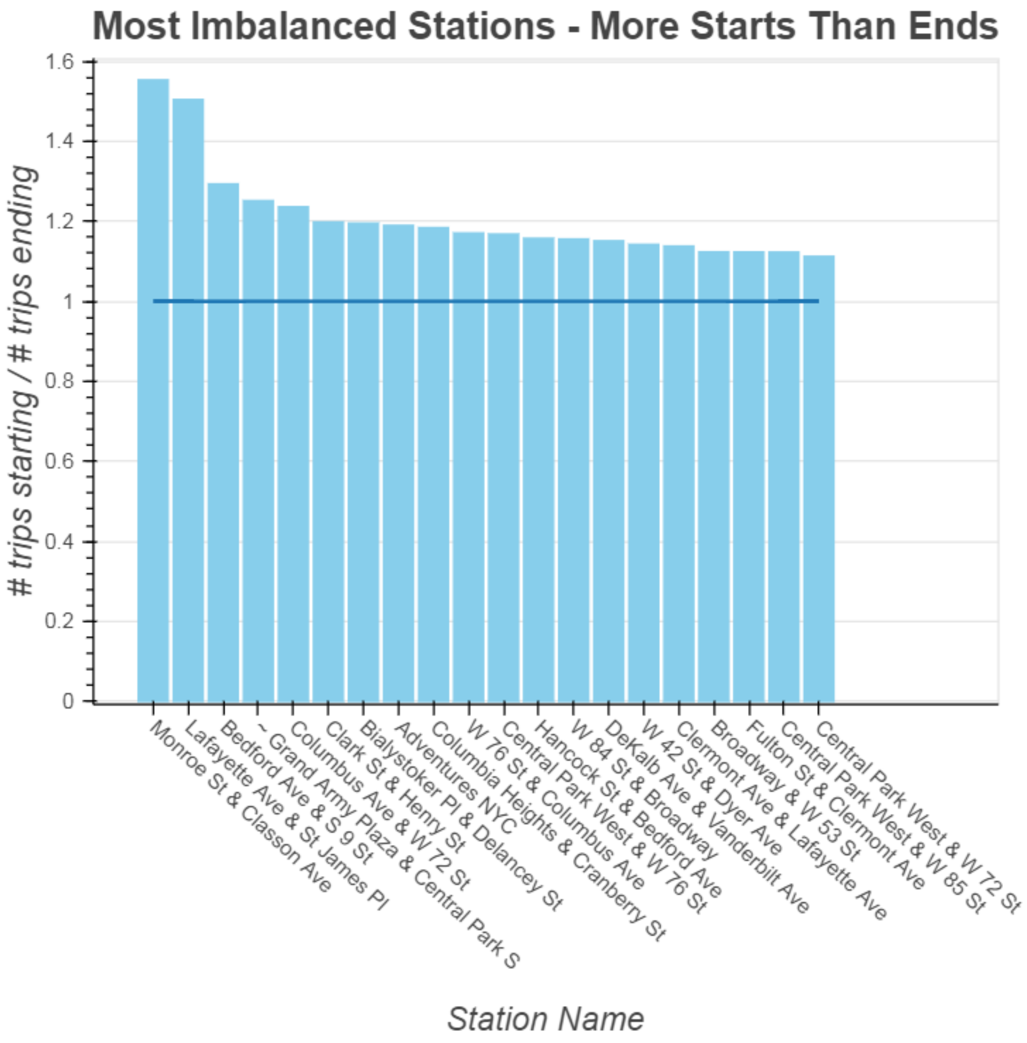

Figure 17: Stations with the highest balancing ratio.

the highest ratios were excluded from Figure 17 because they were found to be service stations for Citi Bike personnel only. We have also included a horizontal line at the ideal balancing ratio of 1 for reference. It is clear when looking at the stations in Figure 17 on the map in [18] that many of these rides originate near Central Park (common for joyrides among tourists and city residents). There are also many rides starting in Brooklyn, presumably to commute into Manhattan, as well as the fact that most do not start in Manhattan with intent to ride to Brooklyn. One station in this set, “Adventures NYC”, was noted as anomalous for not being a street intersection in Citi Bike’s usual stations. Upon searching, it was discovered that this was actually a special outdoor event held in Central Park on June 17, 2017 (right in the middle of our data set) at which Citi Bike presumably offered bikes for rent, hence many rides originating there [16].

We next consider the stations with the lowest balancing ratio. Again we see many of these

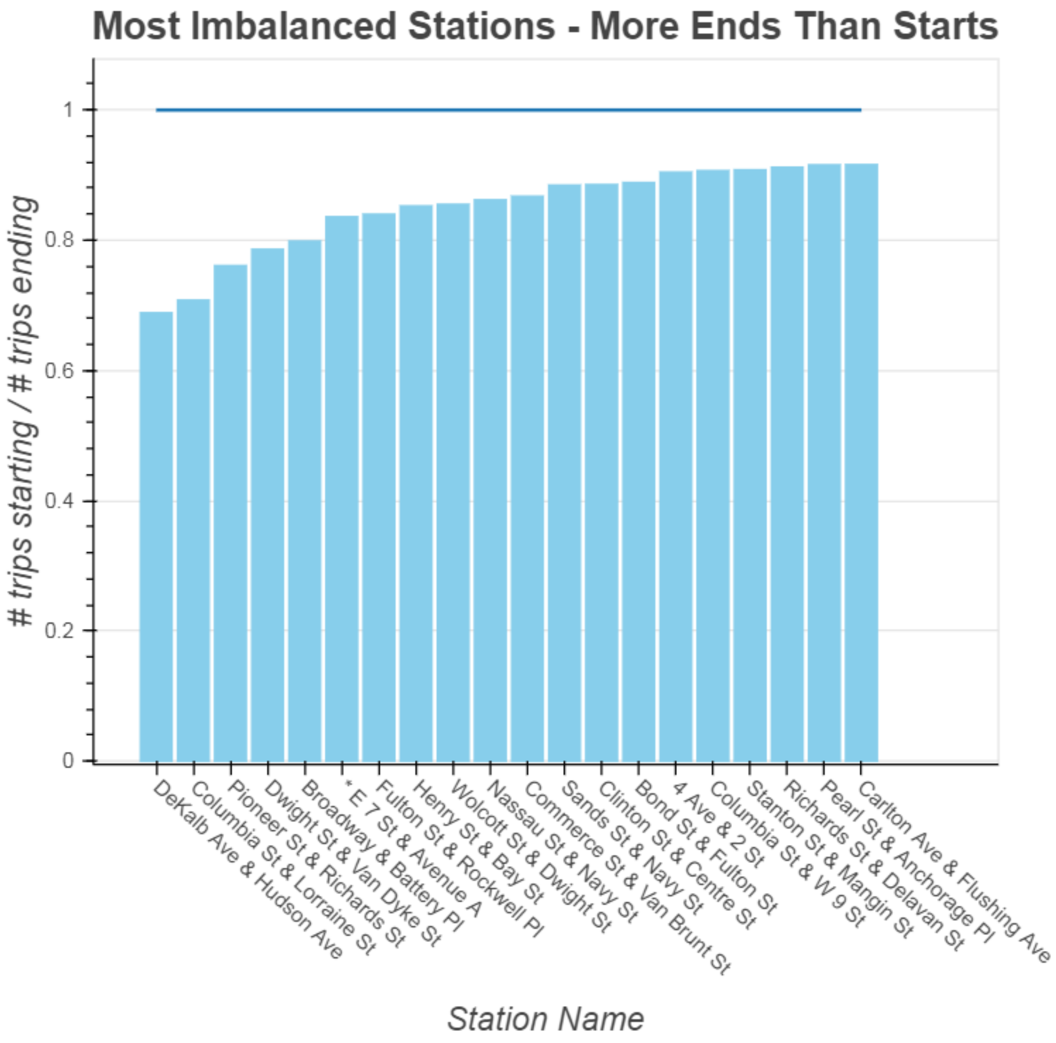

Figure 18: Stations with the lowest balancing ratio.

stations can be found in Brooklyn - perhaps those riding home after commuting into Manhattan. We also see that Battery Park, near the Statue of Liberty ferry departure point, Wall Street, and a generally enjoyable outdoor park space, is a popular destination for many to ride to.

We last consider the stations with the most ideal balancing ratio of 1 - note the very small y-scale in Figure 19. Again, several Citi Bike service stations were found in this set - they are denoted next to the station name on the x-axis. Examining these stations on [18], the locations are less predictable than expected. They are mostly scattered throughout the city, including some in multiple residential areas as well as the bustling Midtown area. We note that this analysis did not take into account temporal factors, which is a key piece to understanding the rebalancing problem. Considering this parameter would likely provide a better indicator of the true balancing ratio at each station throughout the day, rather than the overall indicator that has been given in

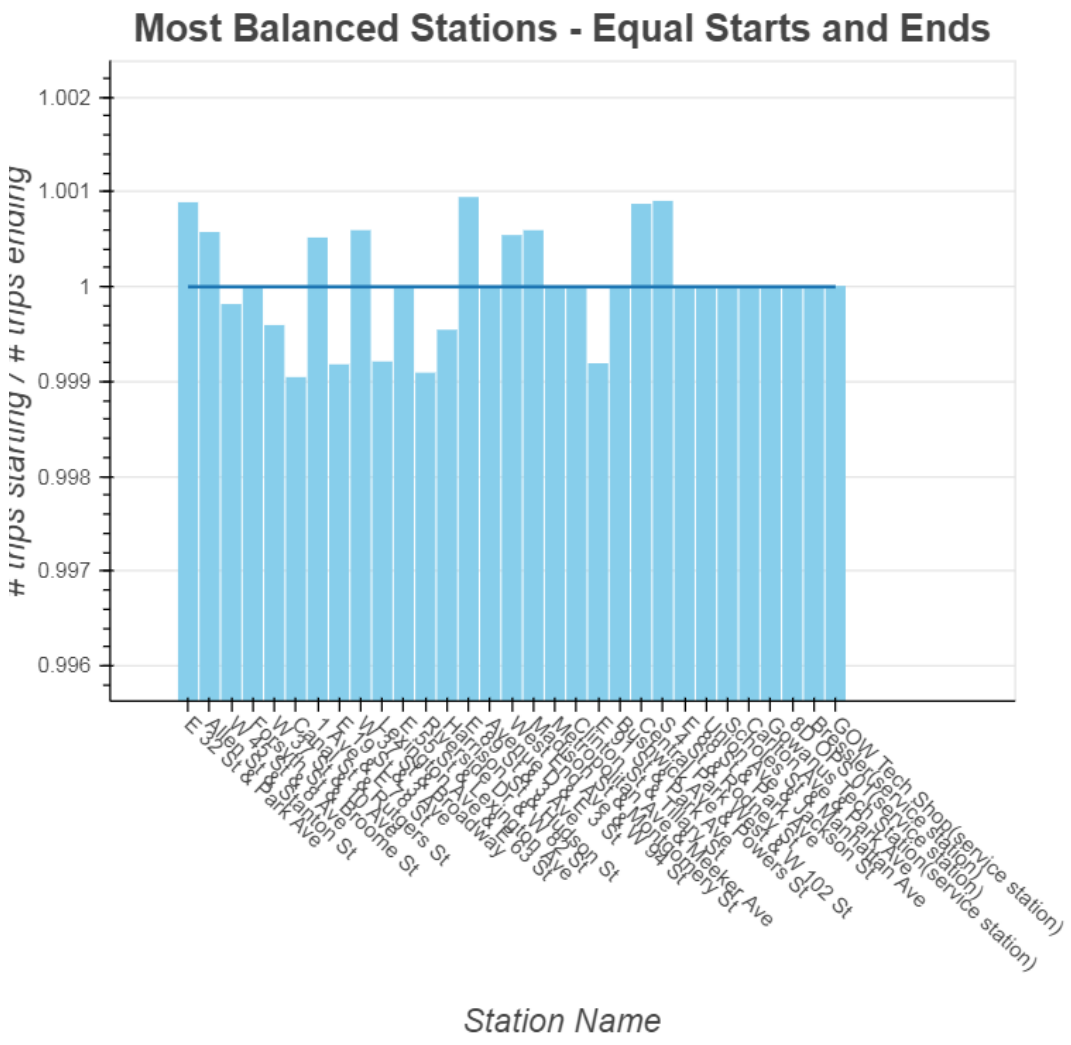

Figure 19: Stations with the most ideal balancing ratio.

this section. For example, some of the stations that experience the worst swings in balance may actually be "balanced" when viewed from a global perspective, as they experience approximately the same amount of traffic during the morning and evening rush hours, but simply in a peaked behavior pattern. This temporal parameter would need to be represented with some sort of information flow taking these peaks into account for a more thorough analysis.

2.6 The Rebalancing Problem Model

After examining the most frequently used stations and considering the rebalancing problem as described in Section 2.5, we now turn to constructing a model to represent these relationships. Key to the model will be stations involved for consideration and some thought to the flow of bikes and customers between them. This problem will be constructed using a heat map as seen in Figure

20. This figure was constructed using the intersect of the top 25 stations that were used as start or end destinations. This intersect consists of the 23 stations seen in Figure 20. We put all of the stations on both the x and y axes of our heatmap. We now let the (x, y) entry color represent the total number of bike rides taken between stations x and y. This is an undirected graphical model. The "heat" color can be imagined as the weight of the edge that would connect node x to node y in a graph. Additionally, we have symmetry in this matrix heatmap representation due to the undirectedness of the graph (we do not care in which direction the ride was taken, simply the *total* rides taken between them) and the diagonal entries are all zero since no trip can be taken from a station to itself. What we have in Figure 20 is really just an alternate visualization of a graphical model of these station nodes connected with weighted edges. From our heatmap, we see that the most rides are between West St & Chambers St and 12 Ave and W 40 St, followed by West St & Chambers St and W 20 St and 11 Ave. These two routes are seen in Figure 21, each clearly forming a route along the Hudson River Greenway on the bank of the Hudson River, with the West St & Chambers St near popular destinations on Wall Street and the World Trade Center. This distance is also well within reason for one trip, approximately a 21 minute, 3.7 mile ride along the entire route.

Now, let us take a directed model approach. We now let the (x, y) entry color denote the trips taken from station x to station y, in that direction. We no longer have the symmetry property in our matrix, as the entry at (y, x) now represents the trips from station y to station x and is not necessarily equal to the entry at (x, y). For example, we see in Figure 22 that more trips travel from W 20 St and 11 Ave to West St & Chambers St (south) than in the opposite direction (north). Please note that these plots do not share a color scale when comparing the magnitude of the the number of trips taken between each station.

3 Conclusion

We have now seen how dynamic of a metropolitan area New York City is and understood some of its commuting, tourism, and recreational trends. We have insight into an innovative system of bikesharing that is used throughout the city and how that bike usage can reflect patterns in the city at large. Bikeshare is not without its challenges however; temporal trends in usage must be characterized, user demographics must be understood, supply and demand must be managed and balanced, and different types of users must be catered to appropriately for their needs. To this end, we have examined the many different ways in which Citi Bike is used. Citi Bike could use this type of analysis to improve their operations in several ways. Perhaps by analyzing usage data at each station, they could understand which stations bear the most burden and will need the most maintenance attention. The complicated rebalancing problem requires constant evaluation but if executed successfully, taking into account the many factors at play, will have enormous benefits

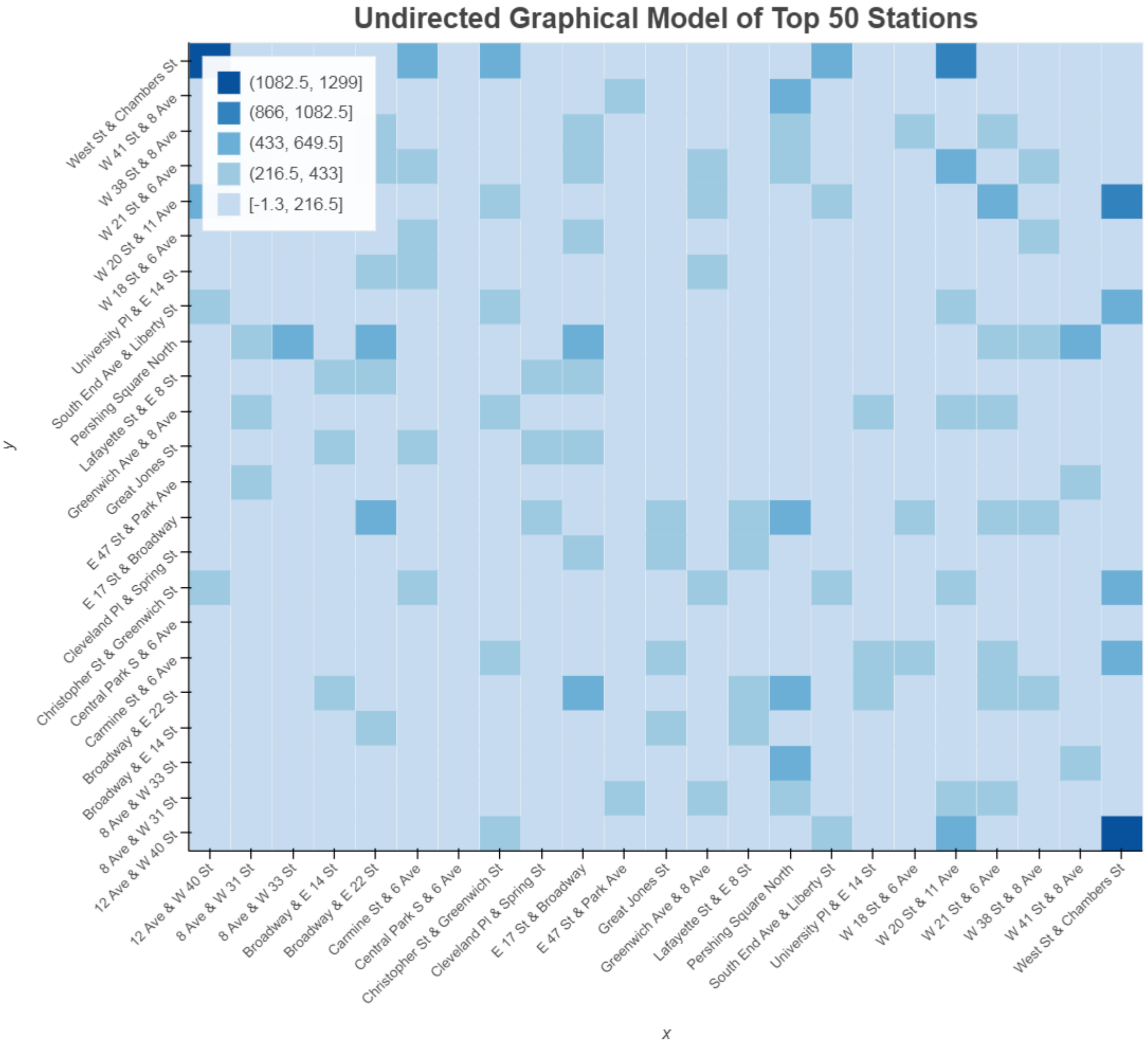

Figure 20: Undirected model of most-used stations.

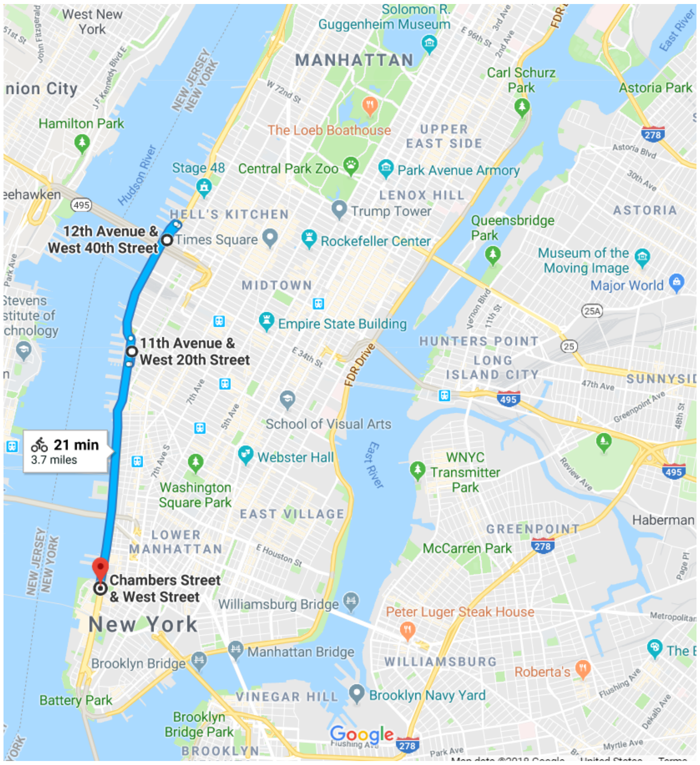

Figure 21: Google map of the most popular rides in undirected model.

for usability and customer satisfaction. By analyzing user demographics, Citi Bike could perhaps target marketing in certain locations or at certain times based on their most likely user demographic, whether subscribers vs short-term customer, male vs female, commuter vs recreational, or any other combinations. The city of New York could use this information for their urban and city planning, as well. Through analyzing popular bike routes used by customers of Citi Bike, the city could have more insight into future planning and understand where dedicated bike lanes might be useful or wanted by the bike-riding population of New York City. Additionally, they could use this data for current analysis to understand the success and usage rates of existing infrastructure that supports biking throughout the city. Future work to expand this data analysis to improve utility to Citi Bike and other customers would include a more thorough temporal analysis with respect to seasons and months of the year, as well as day of the week and time of day. If users were willing to provide more demographic information than is already collected, this could further expand Citi Bike's ability to market to new customers and New York City's ability to understand the demands of the public based on their characteristics. Biking is just one small part of transportation in New York City, but much can be learned from studying the data provided by it.

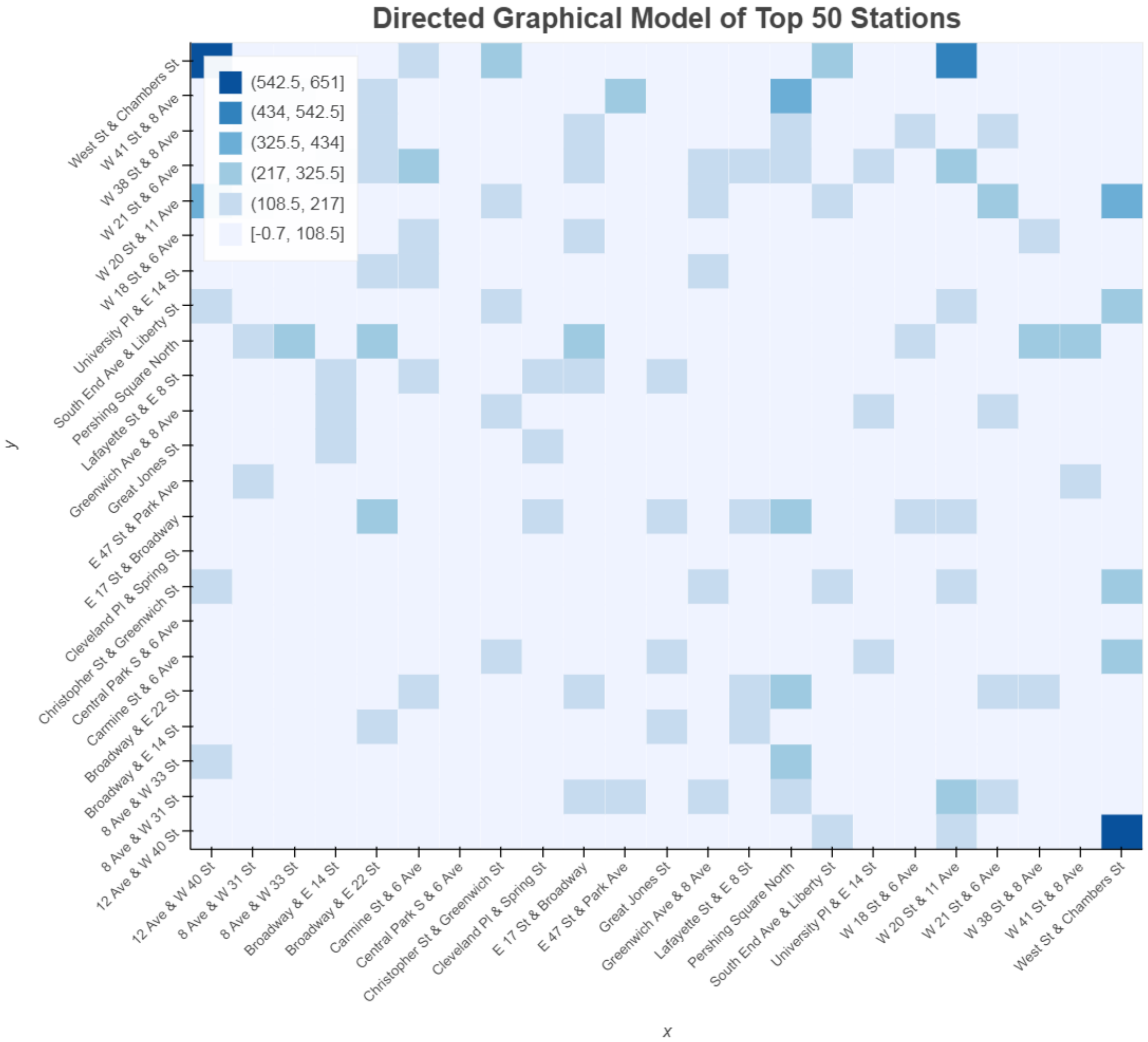

Figure 22: Directed model of most-used stations.

4 Personal Reflections

I thought I would take a small section as well to step away from the paper and provide my personal thoughts on the project! I'm very glad that we were allowed to pick our own topic for this paper - it made the project so much more fun! Getting to explore something I was genuinely interested in combined with the fact that I love visiting New York City (I'm a big Broadway fan) and find geographic analyses and trends so fascinating made this enjoyable to work on. With respect to the data analysis, one piece of data I really found myself wishing for was a label for which borough a station was in. The data provides latitude, longitude, station id (meaningless to me, I'm sure Citi Bike has some use for it) and station name, but no borough. The only way for me to know which borough the station name (intersection) was in was to look it up on Google maps. This wasn't terribly prohibitive since I was dealing with limited scope sizes for stations, but it would have been much more useful or invited more data classification abilities if I had that borough label. Additionally, near the end of my analysis I found myself wishing that I had combined the effects of more of the parameters together. I really enjoyed seeing the connections between all the data analysis I did complete, but realized as I was thinking through all of it and concluding the project that there is so much more possible insight to be gained with this additional analysis. I suppose this is why Citi Bike has to hire software engineers and mathematicians to help with keeping the system up and running smoothly!

References

[1] Motivate International, LLC. (2013, May - 2018, March). *Citi Bike* (and subpages). Retrieved from https://www.citibikenyc.com/

[2] National Association of City Transportation Officials (NACTO) Bike Share Initiative. (2018, May 1). *Bike Share in the US: 2010-2016.* Retrieved from https://nacto.org/bike-share-statistics-2016/

[3] New York City Department of City Planning. (2017, July 1). *NYC Population - Current and Projected Populations.* Retrieved from http://www1.nyc.gov/site/planning/data-maps/nyc-population/current-future-populations.page

[4] United States Census Bureau. (2017, July 1). *QuickFacts: New York City, New York.* Retrieved from https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork/PST045217#viewtop

[5] Census Reporter. (2016). *New York-Newark-Jersey City, NY-NJ-PA Metro Area.* Retrieved from https://censusreporter.org/profiles/31000US35620-new-york-newark-jersey-city-ny-nj-pa-metro-area/

[6] Moss, M.L. and Qing, C. (2012, March). *The Dynamic Population of Manhattan.* Retrieved from https://wagner.nyu.edu/files/rudincenter/dynamic_pop_manhattan.pdf

[7] New York City Department of City Planning. (2008, December). *Changes in Employment and Commuting Patterns among Workers in New York City and the New York Metropolitan Area, 2000-2007.* Retrieved from https://www1.nyc.gov/assets/planning/download/pdf/data-maps/nyc-population/census_commute_patterns0007.pdf

[8] NYC & Company. (2018, March). *2017-2018 Annual Summary.* Retrieved from https://res.cloudinary.com/simpleview/image/upload/v1/clients/newyorkcity/2017_2018annualsummary_02_MR_017d3826-f55d-45a6-b10e-b52e84c06334.pdf

[9] NYC & Company. (2017 December). *NYC Travel and Tourism Overview.* Retrieved from https://res.cloudinary.com/simpleview/image/upload/v1/clients/newyorkcity/NYC_Company_NYC_Travel_Tourism_OverviewEW_dcf2eeb0-2f7b-4dfa-be7f-c4721564b60b.pdf

[10] New York City Metropolitan Transportation Authority (NYC MTA). (2018, March 19). *Fares at a Glance.* Retrieved from http://web.mta.info/nyct/fare/FaresatAGlance.htm

[11] Motivate International, LLC under NYCBS Data Use Policy. (2018, March). *System Data.* Retrieved from https://www.citibikenyc.com/system-data

[12] Anderson, L. (2014, January 2). *Bike-share to keep rolling in blizzard - but who's riding? Some cycles are shifted off streets*. Retrieved from http://thevillager.com/2014/01/02/bike-share-to-keep-rolling-in-blizzard-for-now-but-some-cycles-to-be-shifted-off-streets/

[13] Alamy. (2018, December 5). *New York City, 5 boroughs map*. Retrieved from https://www.alamy.com/stock-photo-new-york-city-5-boroughs-map-96927034.html

[14] Motivate International, LLC. (2018). *Bike Angels*. Retrieved from https://bikeangels.citibikenyc.com/

[15] Motivate International, LLC. (2018). *Redistribution*. Retrieved from https://help.citibikenyc.com/hc/en-us/articles/115007197887-Redistribution

[16] New York City Department of Parks & Recreation (2017, June). *Adventures NYC*. Retrieved from https://www.nycgovparks.org/events/2017/06/17/adventures-nyc

[17] Nichols, M. (2018, May 15). *Popular Bike Stations Google Map*. Created at https://drive.google.com/open?id=1FCjRkk30qxZ7S7p_1iCFA480_JduYe5i&usp=sharing

[18] Nichols, M. (2018, May 15). *Bike Balance at Stations Google Map*. Created at https://drive.google.com/open?id=1i6_42GIG_vRNIyrRfQuUOOj8mYMhd_YN&usp=sharing

[19] Nichols, M. (2018, May 15). *Popular Bike Stations by Type of Customer Google Map*. Created at https://drive.google.com/open?id=1pC6DptQ119HqGKfH0hMk88oiBGeJzFvG&usp=sharing