Report: How do people use Citi Bike ?

Arnaud Stiegler (aes2329), Redouane Dziri (rd2853) and Corentin Llorca (cl3783)

Fall 2018

I. Introduction

The general topic of transportation and commuting is prevalent in the world of data analysis, because it is a perfect way to get large-scale data and extract trends of which we can easily understand the meaning, and because it touches everyone, since everyone travels everyday in a way or another. We have seen studies done with subway data, or taxi data. However, we wanted to study an alternative means of transport, that is often overlooked: cycling. Analyzing the Citi Bike bicycle sharing system, which accounts for a large part of the bicycle traffic in New York, and which is the largest bicycle sharing system in the United States, was the logical continuation to that idea. Moreover, the fact that there is a large quantity of data of good quality directly available on Citi Bike's website, coupled with the status of Corentin, one of our group members, as an annual Citi Bike subscriber, steered us towards that subject among the other ideas we had.

Our team is made of three members. Redouane Dziri (rd2853) took charge of the main part of the data gathering, pre-processing and data quality analysis and cleaning, as well as analyzing seasonal patterns and the effect of weather in the main analysis part. Arnaud Stiegler (aes2329) analyzed the geographical patterns in the data, as well as coding the interactive component of the project. Finally, Corentin Llorca (cl3783) undertook the user types analysis, with age, gender and customer/subscriber status, as well as writing the executive summary and organizing the report.

II. Description of the data

Our data was collected from Citibike system data. We downloaded trip data (.csv files) from the months of February, July and October 2018 from trip data and daily ridership and membership data from the last quarter of 2014 to the third quarter of 2018 included on the home page (also .csv files). We also gathered climate and weather data from the National Centers for Environmental Information (part of the National Oceanic and Atmospheric Administration) from the last quarter of 2014 to the third quarter of 2018 included by ordering the data on NOAA portal and selecting precipitation, temperature, snow and wind information for the Central Park station.

To prepare the data for use, we concatenated the daily ridership and membership data that was available in one file per quarter per year. We kept decided to remove information on the number of 3-days and 7-days passes bought per day as these were not available across most quarters and we weren't convinced of the usefulness of keeping that information. We also had to deal with inconsistencies in terms of date format and had to level them out across all files in order to merge them with no distorsion and false readings.

We chose to refactor the trip data from the three months we selected in 2018 by building a separate file saving the information regarding Citibike stations and keeping only stations id in the main table to avoid redundancy - which saved a considerable amount of memory and made computations in R smoother.

All of those steps were stored in a script called preprocessing.R, which should be run within a directory containing a "data" folder which includes :

- the February, July and October 2018 trip data
- the weather data
- a folder named "raw_summary_stats" containing the daily ridership and membership data files, from the 4th quarter of 2014 to the third quarter of 2018

Please be careful as the R working directory must be changed to the directory containing this file, the script and the data folder.

The script is available in the project's GitHub.

The files that are produced by the script are the following :

- "concise_trips.csv" for the February, July and October 2018 trip data
- "concise_weather.csv" for the weather data
- "summary_stats_per_day.csv" for the daily ridership and membership data
- "stations_info.csv" for the stations info data

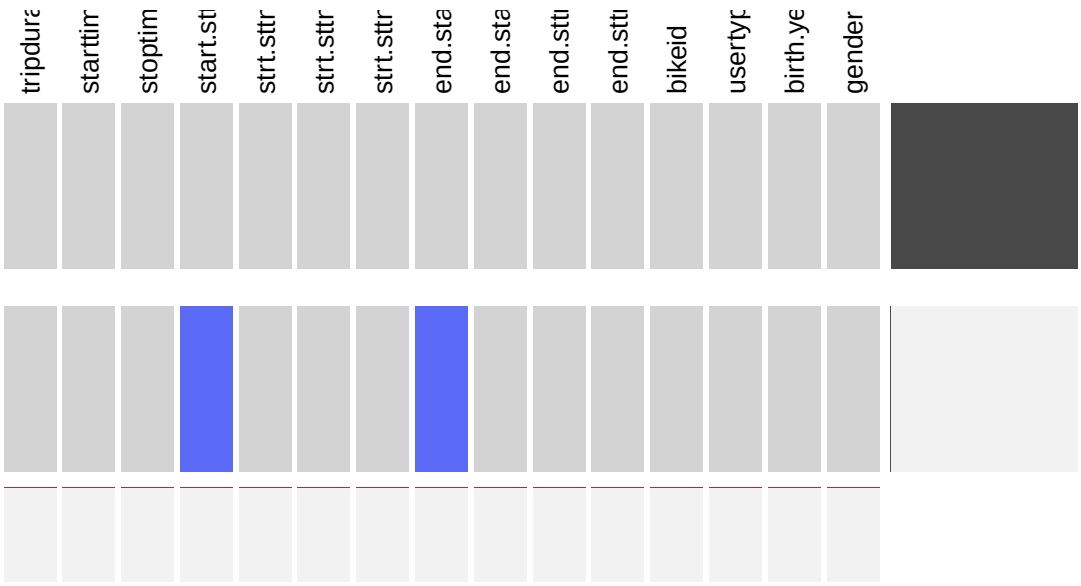
III. Analysis of data quality

Before conduction any exploratory analysis we need to assess data quality. A quick survey of the weather data reveals that the data is tidy, with no outliers and no missing values.

The trip data fetched from February, July and October 2018 on the other hand has a few notable mishaps. The coding for categorical data is consistent - no multiple codings for the same item - and the data dictionary available online explains each individual variable clearly as well as the levels of categorical variables.

1. Missing data

We started data quality analysis by looking at the **missing data patterns**.



The only missing data pattern apart from no missing data appears to be missing ids for the start and end stations. The first question we asked ourselves when seeing this is: **why is only the id of the station missing and not the station name or longitude/latitude**?

1.1. Missing station ids

To understand it better we peeked at the rows with missing start and end station ids.

We noticed two things: - it would seem that missing ids for the start and end stations is also associated with missing names for those stations. R just did not recognize those as missing values as they were encoded as strings 'NULL' and not NA. - those stations all seem associated with high latitudes

So we tried to look up those stations by cross referencing their longitude/latitude with the online Citibike stations map. We were surprised to find that there were no stations to be found. Instead we discovered that those trips corresponded to a new **dockless bike area** launched by Citibike in the Bronx in August 2018. Bikes are available in certain areas of the Bronx and can be left anywhere, without docking - which explains the lack of start and end stations. The position of the bikes is still registered (longitude and latitude of where the bike was picked up and dropped off registered). All the data for such trips is dated after August 2018, which seals the coffin.

We will separate this data from the rest and explore it separately as it relates to a whole other way of envisioning bike sharing.

1.2. Missing gender

After closer inspection we realized some gender information was also missing. It was not detected early on because the missing data was not encoded as NA. Rather missing genders were input as 0. Males are 1s and Females are 2s. We first changed the encoding to a more transparent one and looked at how many values for gender are missing.

There is a significant number of missing values in the gender column. This should be kept in mind when analysis requires faceting on gender or studying variables in relation to gender.

There is no reason to discard the data with missing gender; the trips' validity is not brought into question. We can not think of any satisfactory way to impute missing genders without risking to distort the data significantly so we will just filter it out whenever gender comes into play.

2. Outliers

Next we dived into detecting outliers.

2.1. Age

Citibike clearly states in its Help Center: "You have to be 16 years or older to ride Citi Bike." While we are not convinced this is really enforced and is more of a liability protection for Citibike than anything else, we do know that one requires to log in their smartphone and pay with credit/debit card to ride a bike. So we will assume that the bulk of riders is 16 or older. The App should not let you log in if you indicate a younger age anyways. There shouldn't be younger riders in our data. This can only be checked grossly as data contains only the birth year of the user and not their month or day of birth.

It turns out that there are no such young riders.

An important thing to note is that **Citibike does not enforce identity verification**. Therefore one can input whichever age they want when registering. This casts some doubt on the validity of the data at hand, and we are unsure that assuming that most people don't lie about their age when registering is reasonable.

We have to be watchful of people that register with a very old age.

There are around 5,000 trips for people over 90 years old. Whilst it is very possible that a senior 90 or older would rent out a bike, the validity of the age data in this range is seriously doubtful. Nevertheless there is no reason the data on the trips themselves should be wrong or corrupted so we will keep them (we could remove them as they represent less than 0.2 % of the data). We should just remember to filter them out whenever age is a factor in our analysis.

2.2. Trip Duration

Now more about the trips themselves: we wish to examine outliers in terms of trip duration. **Ideally we would remove trips under 90 seconds and over 2 hours**.

Why 90 seconds? We believe most of those trips will correspond to a change of heart by the user or more frequently, bikes that don't work properly and are re-docked. A person using a bike to get from point A to point B would probably walk instead if the trip was under 90 seconds so the risk of filtering out valid data is small. It should not take more than 90 seconds for most users to become dissatisfied with a bike to the point of choosing to dock it and taking another one or choosing an alternative means of transportation. Therefore it seems reasonable that there should not be a significant number of invalid trips, on the basis of short trip duration, left in the data after filtering out those under 90 seconds.

Why 2 hours? Customers using passes are allowed 30-minute rides and then pay each extra 15 minute until they dock their bike. This extends to 45 minutes for Subscribers. Given the density of the Citibike station network, as their expiration approaches, a user would probably dock their bike and take another one rather than pay more for longer trips. 2 hours is a safe bet: we will probably retain some unproperly registered trip durations but the risk of removing valid data is small. Reasons for longer registered trip durations can be failure to dock the bike, stolen bikes, or broken docks.

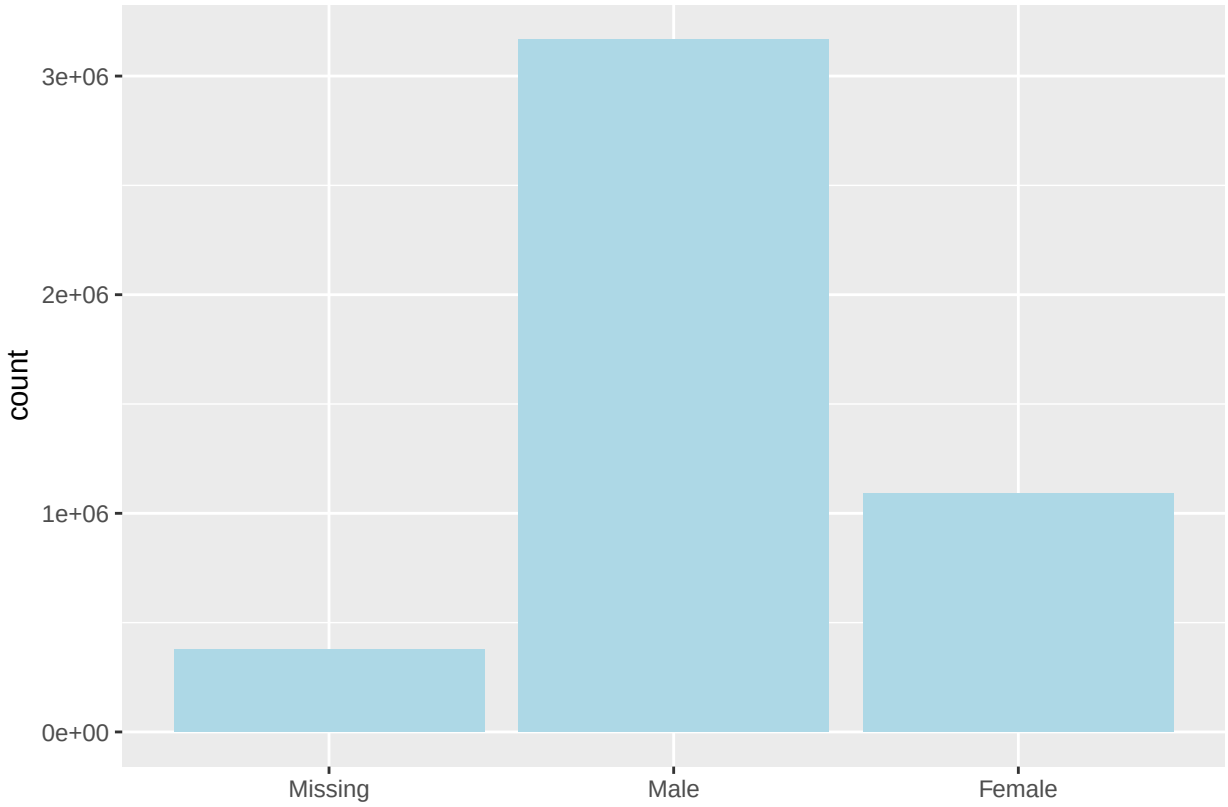
There are a little less than 30,000 trips registered that last less than 90 seconds. This may seem like a lot but from personal experience, getting an unsatisfactory bike is not a rare occurrence so we are not too surprised. They represent a little less than 1 % of the data and we choose to remove them as we believe they would only add noise to our analysis.

There are a little less than 13,000 trips which last more than 2 hours. We will also remove those.

3. User characteristics distribution

To get a better idea of the data at hand we decided to also include the **distribution of user characteristics** in our preliminary data analysis.

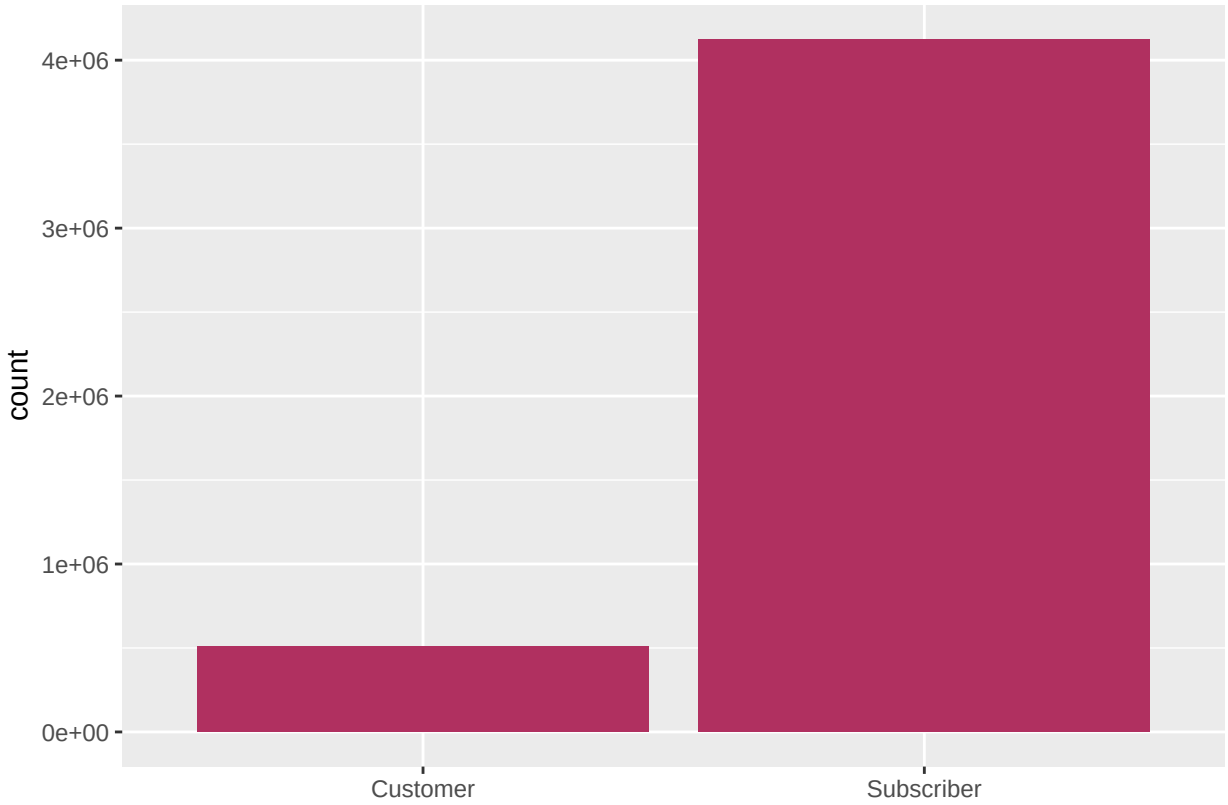
3.1. Gender



Again we see there is a significant number of missing gender values.

We could make the observation that users can input whichever gender they wish, there is no identity verification. Nevertheless we feel that most people would not lie about their identified gender.

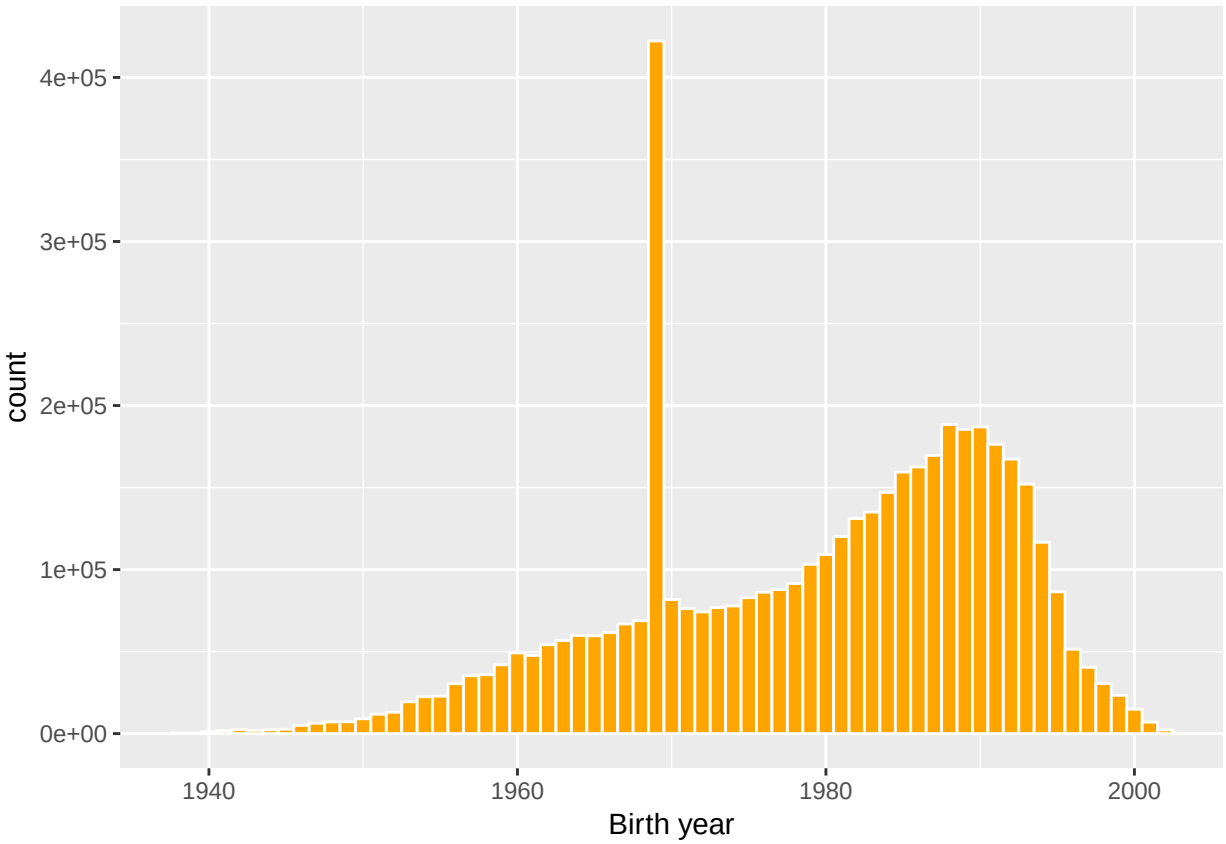
3.2. User status



We don't notice any anomalies here. The Subscriber / Customer status is a feature that's always going to have a value, and that won't have many errors, since the Citi Bike system knows which type of user is using the bike when one is rented - it isn't a value that depends on the user entering it.

3.3. Age

After removing outliers as discussed above:



The first thing we notice is the outlier value for 1969. We think this might have been the default age value for which users did not have to interact with the system to change their age. Many would have agreed to that value without a second glance. Any analysis based on the age of users or their birth years will take this into account, probably by altogether ignoring this value since it probably includes data for all ages.

The data cleaning steps that we have selected (removing short and long trips, assigning more explicit values to genders, removing dockless bikes and turning birth date to age) can be executed by running the data_cleaning.R script. Again, the script must be placed in the same directory as the "data" folder previously mentioned, and R's current directory must be set to that same directory. The script is available in the project's GitHub.

The output file for the trip data is then "data_concise_cleaned.csv".

IV. Exploratory Data Analysis

Let's first load the data we will need :

- The trip data file for Feb-Jul-Oct 2018
- The daily ridership and membership summary data
- The stations info
- The weather data

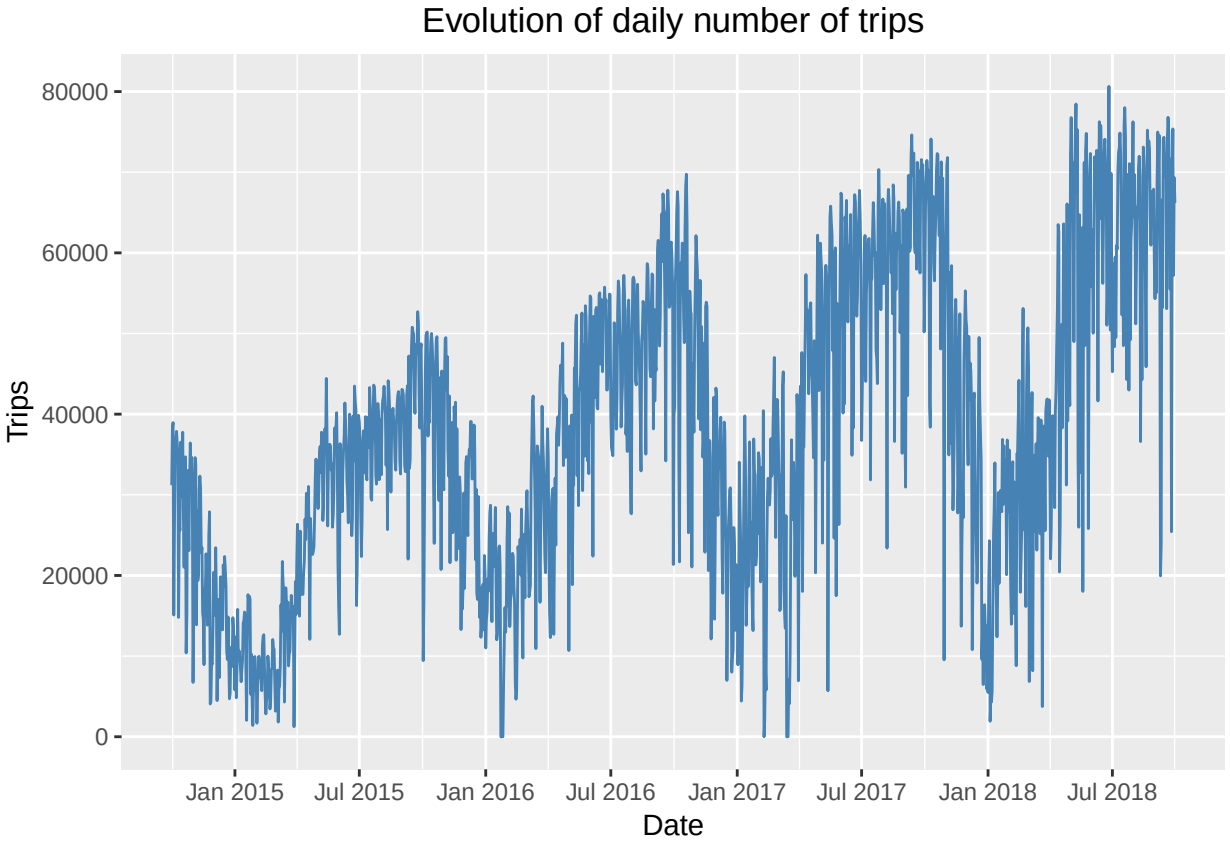
The aim of the exploratory data analysis step is to gain insight on the general question of "How do people use Citi Bike ?". This includes looking at the general behavior of the users against outside parameters like seasons, days of the week, weather, etc.; byt also looking at the different user types for the bike sharing system, and checking whether they have different behaviors in similar outside circumstances.

1. General usage trends through time

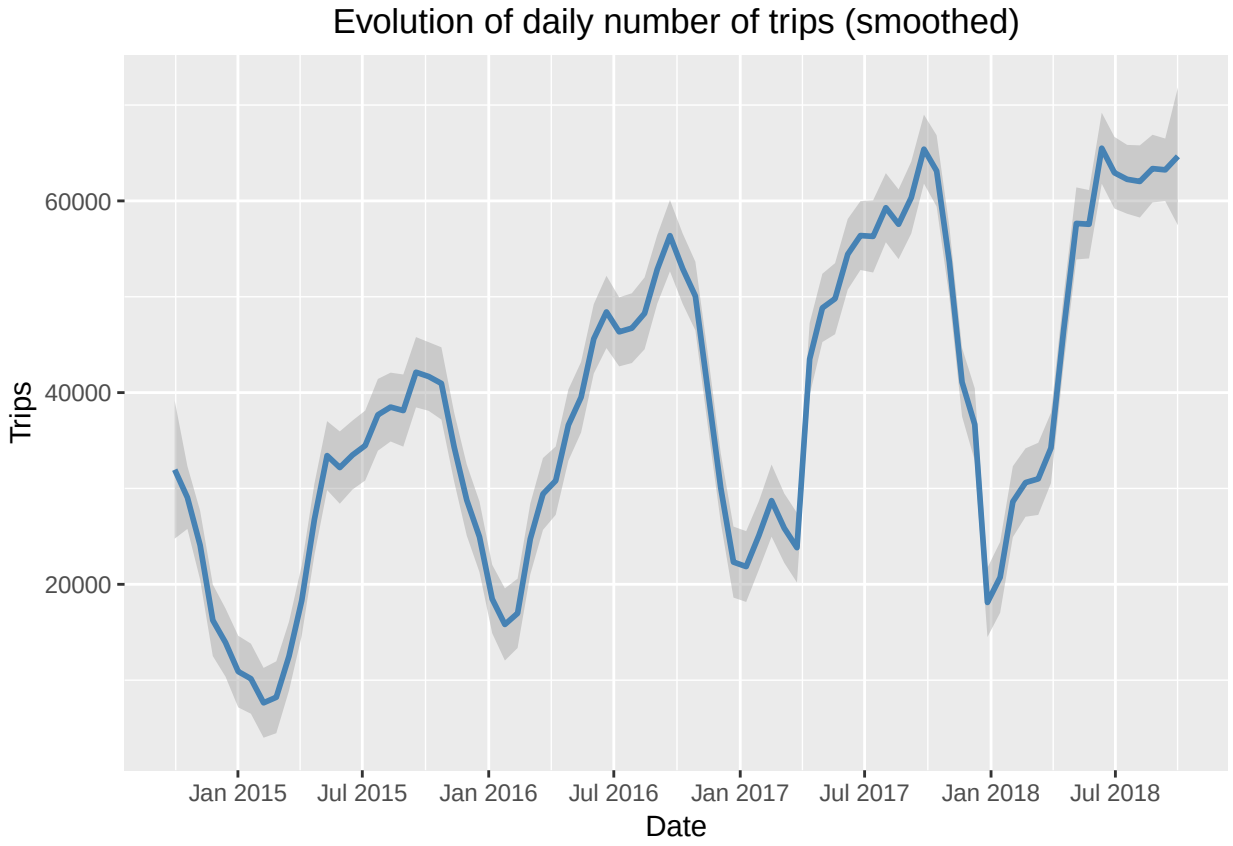
In this part, we aim to briefly showcase some general trends depending on time. We'll plot the distribution of trips taken on three different scales : yearly, weekly and daily. This will serve as baseline information for the rest of the project, and will be expanded in later parts.

1.1. Yearly trends

First of all, let's simply plot the number of trips taken per day across all the daily data that we have.



This curve is mostly unreadable, so we've decided to replace it with a smoother that could show us some yearly trends. We used a loess smoother, and toyed around with the span until the curved was smooth enough while still displaying detailed trends.

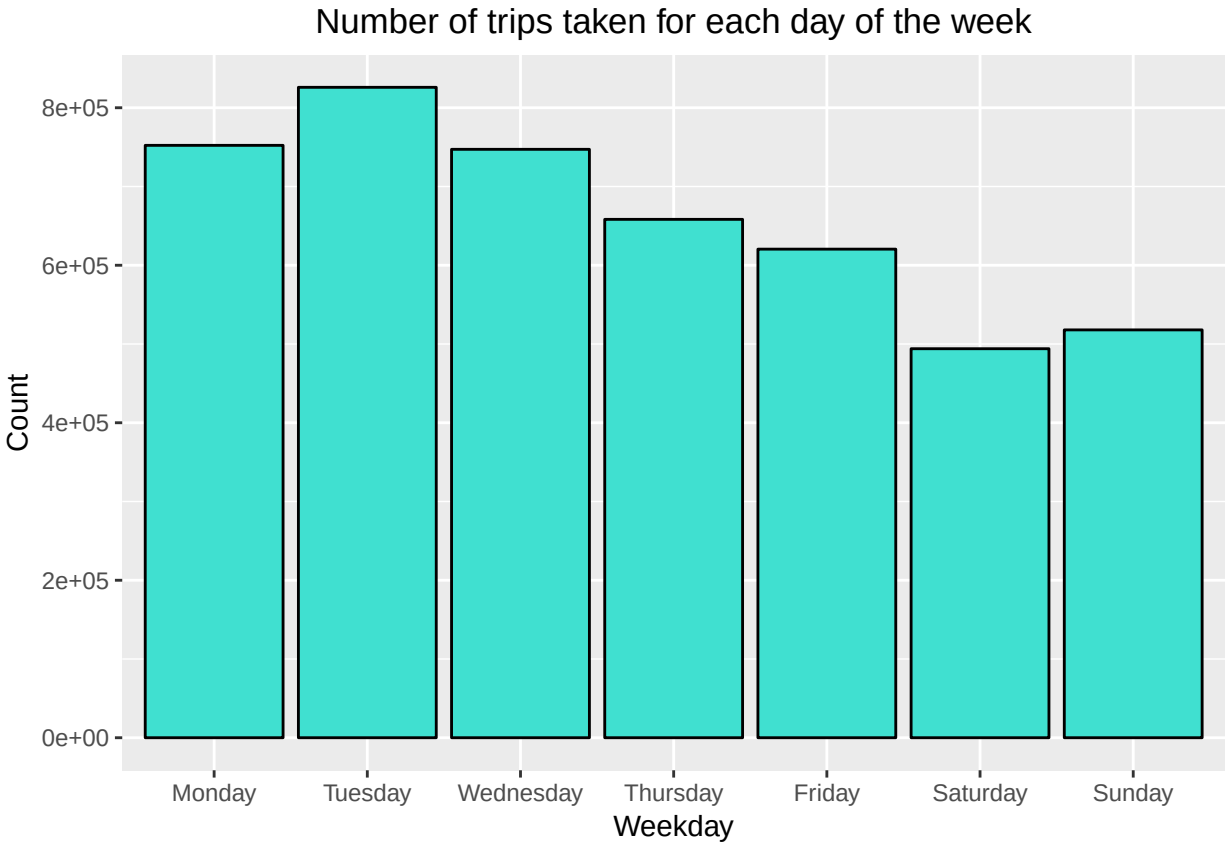


As expected Winter months have much fewer trips than the Summer months, on which the number of trips peak. The number of trips was lowest in February 2015 - around 8,000 daily trips - and highest in June 2018 - around 67,000 daily trips. There is no denying there is an overall global increasing trend, as well as a seasonal component.

One might wonder what the sharp increases between August 2015 and September 2015, July 2016 and September 2016 represent. We looked into it and found that those were critical months for Citibike that saw the deployment of a large number of bikes and new stations as part of their expansion plan. This probably led more people to at least try the bikes from new stations close to their home.

1.2. Weekly trends

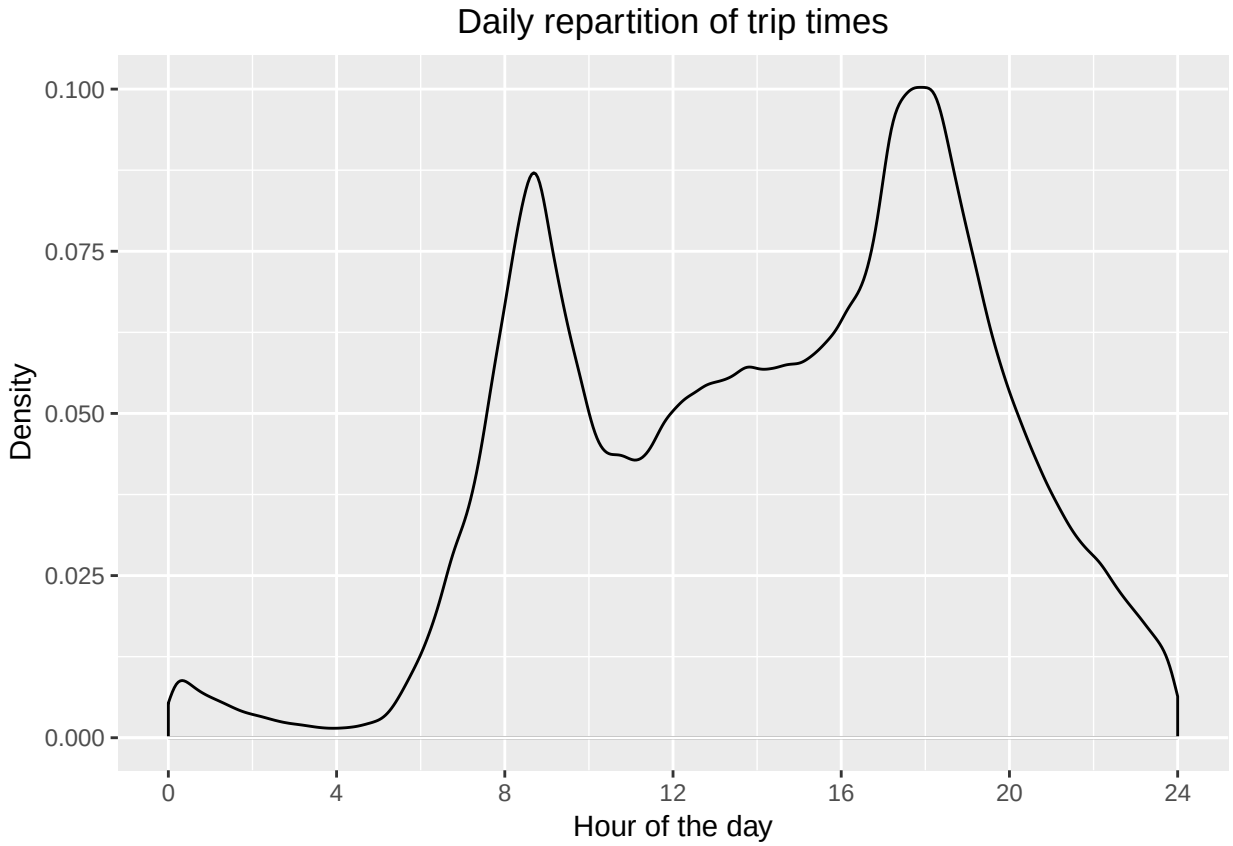
Let's now look at the distribution of the number of rides according to the day of the week : we're looking for a distinction between the weekend and the weekdays.



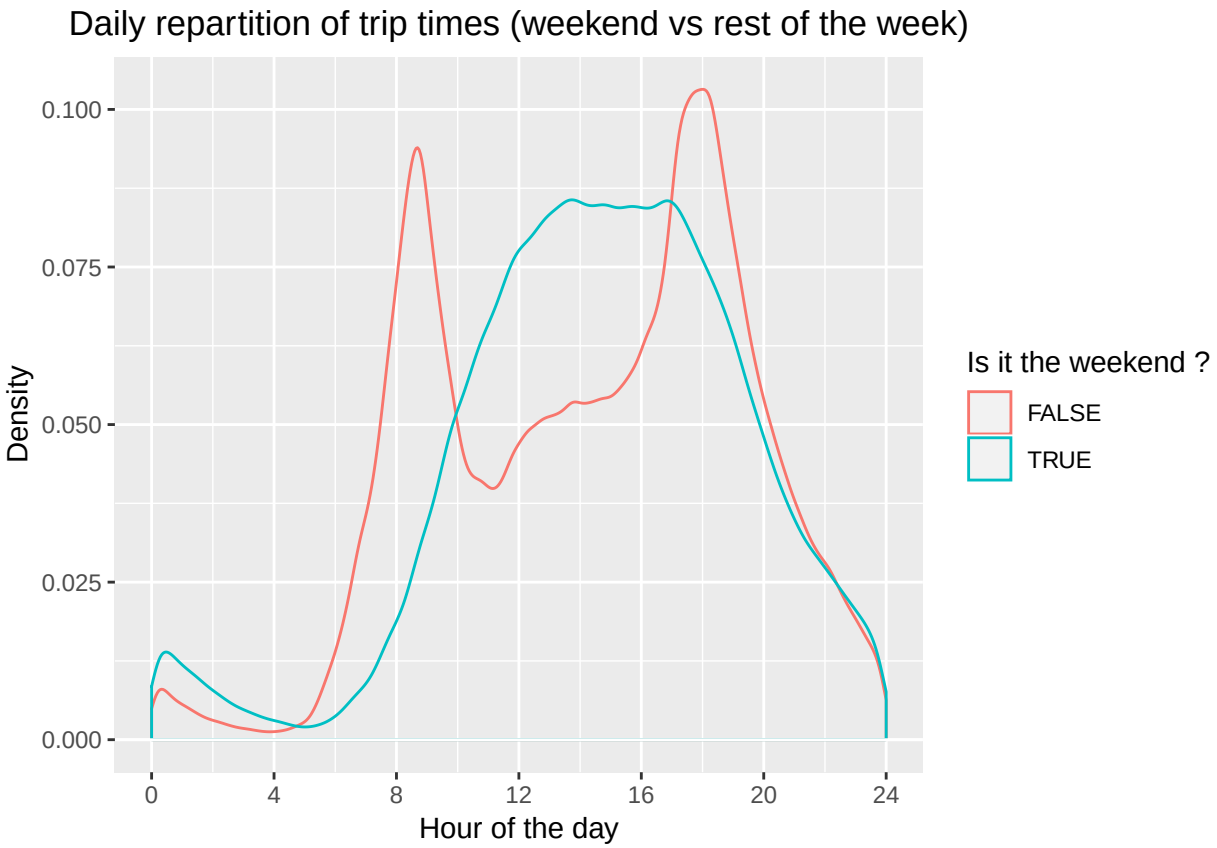
Surprisingly enough, there are less people cycling on the weekends than during the week. We would have expected people to be getting out more during the weekend, which would have meant using Citi Bike more to get around the city. However, the higher number of trips taken during the week might point to a large number of people using Citi Bike to commute.

1.3. Daily trends

Finally, for the last plot of this part, let's plot the density distribution of the time of day at which trips are taken. Here, we use the starting time of the trip as our data, and plot its density. We're hoping for our commuting hypothesis to be confirmed.



And, indeed, our hypothesis gets confirmed : there is a large part of the user base that uses Citi Bike for commuting, as evidenced by the two clear modes of the repartition of daily trip times, around the traditional commuting hours of 8 am and 6 pm. In order to strenghten that hypothesis, we also want to display the same plot, but faceted between weekdays and the weekend. We expect the modes to be less apparent in the density for the weekend.



Indeed, we can see that the modes are much less present during the weekend - even invisible - thus confirming that they correspond to commuting. Instead, we get an unimodal curve with a flat mode from 1 pm to 5 pm approximately. Moreover, we can observe that people bike more at night during the weekend.

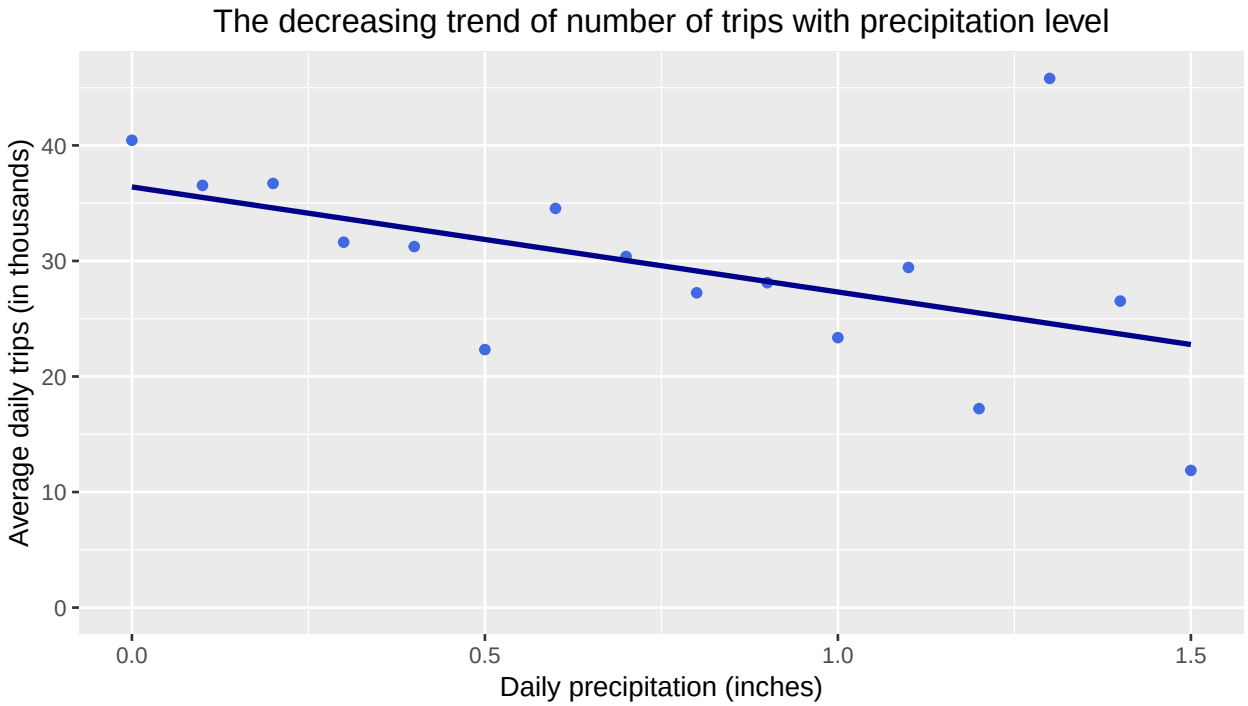
In this part, we have gained the insight that a majority of Citi Bike users use the system in order to go to work every day, as well as observing that there are a lot less people using the system in the winter. We'll keep that in mind as we go through the rest of our research.

2. The effect of weather on Citi Bike users

Here, we want to look at some more general trends that may affect all Citi Bike users. Surely people's use of Citi Bikes is greatly dependent on weather conditions: there will probably be a lot less trips on rainy days than on dry days, the more snow on the ground the less likely people are to rent a bike. Those are hypotheses that we want to test through exploration of the data collected from the Central Park weather station and Citibike's data.

2.1 The effect of rain

We will first explore how the number of trips varies with the precipitation level. The data used here is the number of trips and precipitation level in inches from October 2014 (included) until October 2018 (not included).



On top of the aggregated data points, we plotted a linear smoother that visually highlights the decreasing trend in the average number of daily trips against precipitation. There is a linear-like behavioral trend but this should be interpreted as a ground-truth uncovered in the data. The higher the value of daily precipitation, the less data was available for gathering (there are few number of days with high levels of precipitation in New York past a certain threshold, at least during the past 5 years). The attempt at modelling represented by the straight line fails to capture this weakness in the data - which also explains the outlier around 1.3 inches of daily precipitation.

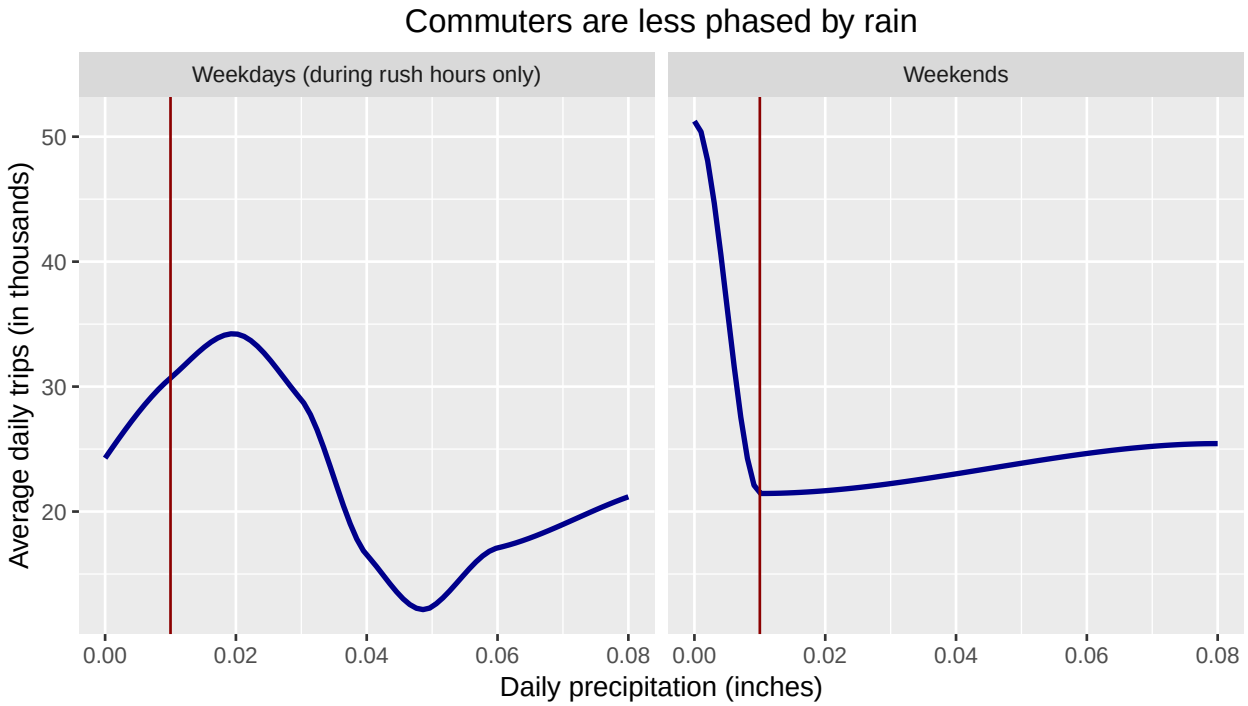
Nevertheless, we can interpret this plot as telling us that the average number of daily trips decreases with precipitation, in approximation linearly. One could expect the decrease to be quite sudden as daily precipitation slowly increases from 0 and then the trend to be less obvious, or less abrupt. After all, the difference between no rain and rain, and rain at 0.7 inches and rain at 0.8 inches is not the same for the human experience of biking. The former is much more informative and can act as a powerful detterent for bike use while the latter is harder to evaluate and is less likely to have a big impact on a person's decision to rent a bike.

The plot suggests that some people either aren't phased by rain very much or that renting Citi Bikes is a necessity for some users - either for timing reasons or unavailability of transport alternatives for their routes. Therefore, we expect that the effect of rain might be stronger during the weekends, during which people generally rent bikes for fun, versus morning and evening rush hours during weekdays when people have to leave home to work and vice-versa within a restricted timeframe.

We will proceed to test this hypothesis in the following, for the months of February, July and October 2018 for which we have data on every trip - including the exact timing within the day.

2.2 Does rain have the same effect on usage during the weekend and during the week's rush hours ?

We plotted the average number of trips against precipitation, distinguishing trips done during morning rush hour and trips done during the weekend. The aim here is to check whether Citi Bike users are more discouraged by precipitation when taking trips during the weekend, which would correspond to leisure trips.



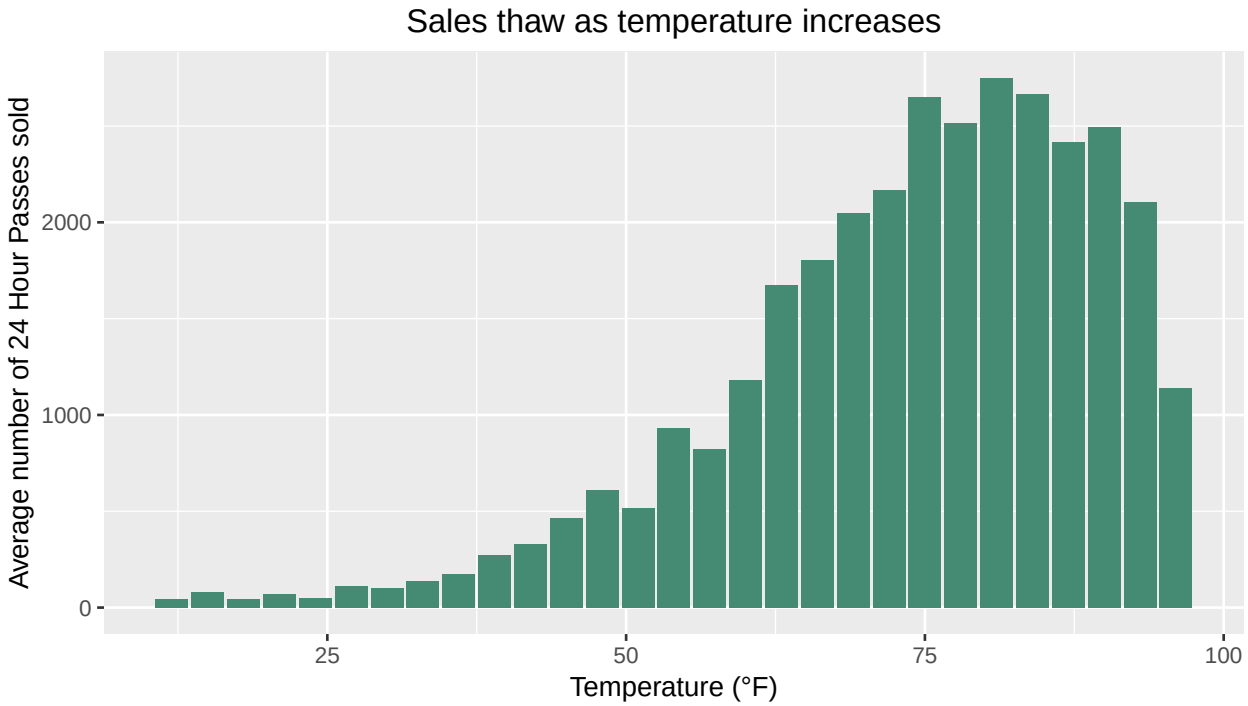
We see that the first signs of rain have much more of an impact on bikers during the weekend than it does on commuters. Commuters that have to get from point A to point B are probably expecting to spend less time on their bikes than bikers for fun on the weekends. The first drizzle does not discourage most of them, which is reflected by a non-decreasing trend as daily precipitation slowly increases from 0 - even increasing !

On the other hand, tourists and bikers for fun on the weekends probably expect to spend a lot of time biking and enjoying their ride. The first signs of rain are seen as bad omens for the weather to come and the number of daily trips on those days quickly decreases as daily precipitation increases from 0 to 0.01 inches. It increases after 0.01 inches but there are very few data points with high precipitation on weekends in our data so we should not blindly accept this smoother model as an accurate representation for reality for those values of daily precipitation.

Commuters are rapidly discouraged by the rain starting at 0.02 inches until 0.05 inches. There aren't too many days with more precipitation in our data so the following increasing slope could very well just be an artefact of overfitting the little data available.

2.3. Are visitors sensitive to temperature? How does temperature impact 24 hour passes sales?

After examining the effect of precipitation on the number of trips by Citibike users we will turn to the effect of another climatic feature which we believe might have a significant impact on riding patterns: temperature. We will focus on whether temperature affects the number of 24-hour passes bought each day. Those are mostly purchased by tourists and therefore, we expect our exploration to help us gain new understanding of how New York City visitors use the ride share network.



We see that as temperature increase from 12 degrees to around 80 degrees, average sales of 24 hour passes increase (first slowly, and then more rapidly). We expected low sales for low temperatures as fewer people venture out in the cold and riding in the cold can worsen its effects by amplifying wind speed. Furthermore, snow is often associated with low very low temperatures, which makes biking around the city even more unpractical.

Use of Citi bikes starts to decrease around 80 degrees, with extreme temperatures seemingly yielding less 24 Hour passes: visitors are less likely to bike under the harsh Summer New York sun. Although this could also be a discrepancy due to the lack of sufficient data for those hot days that are in small numbers across the five last years compared to the other temperature bins.

3. Age and gender : the demographics of Citi Bike users

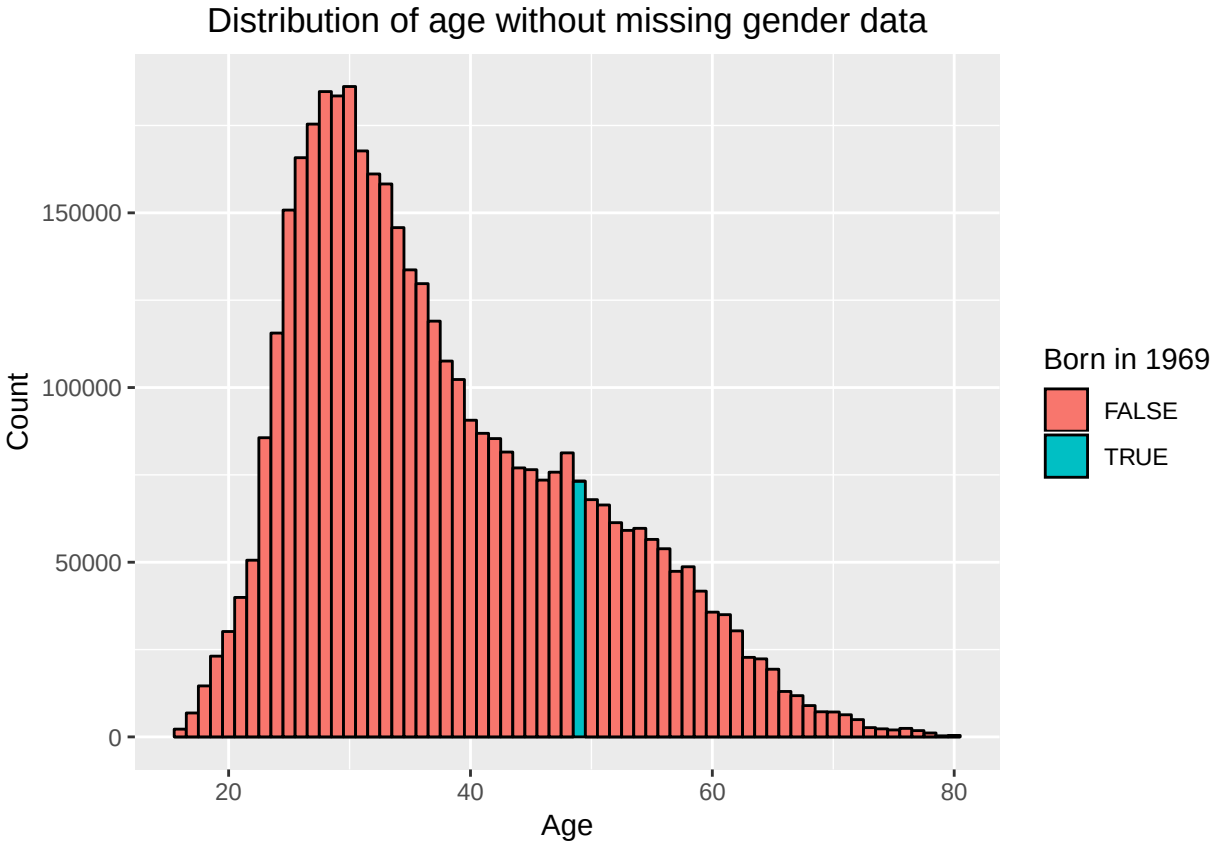
Our next analysis will be to look at the age and gender of Citi Bike users. Please note that user information is only required within the yearly subscription plan, meaning that the data that will be shown in this part will only be the trips done by subscribers. However, those represent about 90% of all trips, so there should be no issue in generalizing the insights gained here.

3.1. Anomalies in the data

A big issue with this data is that the users making the trips aren't individually identified, meaning that what we're plotting isn't the age/gender distribution of Citi Bike subscibers, but rather the number of trips taken by people from each individual age/gender. Getting for instance a high count for men compared to women doesn't necessarily mean that there are more men than women using Citi Bike, but rather that there are more trips being taken by men than women.

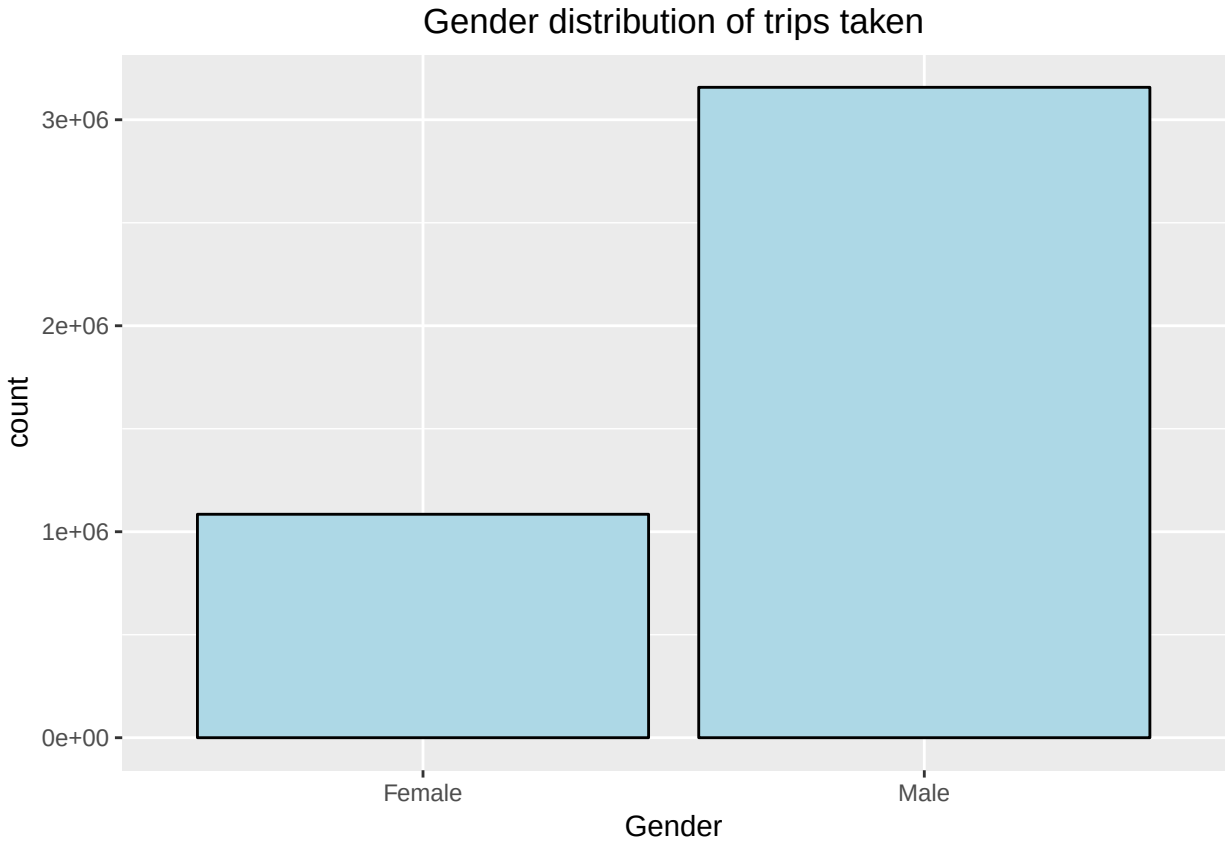
As found in our data cleaning, there is a lot of missing data on gender, and there are a lot of outliers and wrong-looking data on age. For this section, we will then be using a reduced version of the trip database. The data we'll be excluding will be the trips for which the gender is not filled in, and those for which the user is older than 80 years old. We also had a problem with the birth year 1969 (age 49) being over-represented,

but let's see if removing the data in which gender is missing solves the problem. We first apply our filter, then plot a histogram of user age, with age 49 being highlighted.



We see, indeed, that the 1969 anomaly is removed when we exclude data in which gender is missing. This could be explained by the fact that people who don't fill out their gender might be the same people who would leave the default birth year when subscribing. We'll continue on with that data for this section.

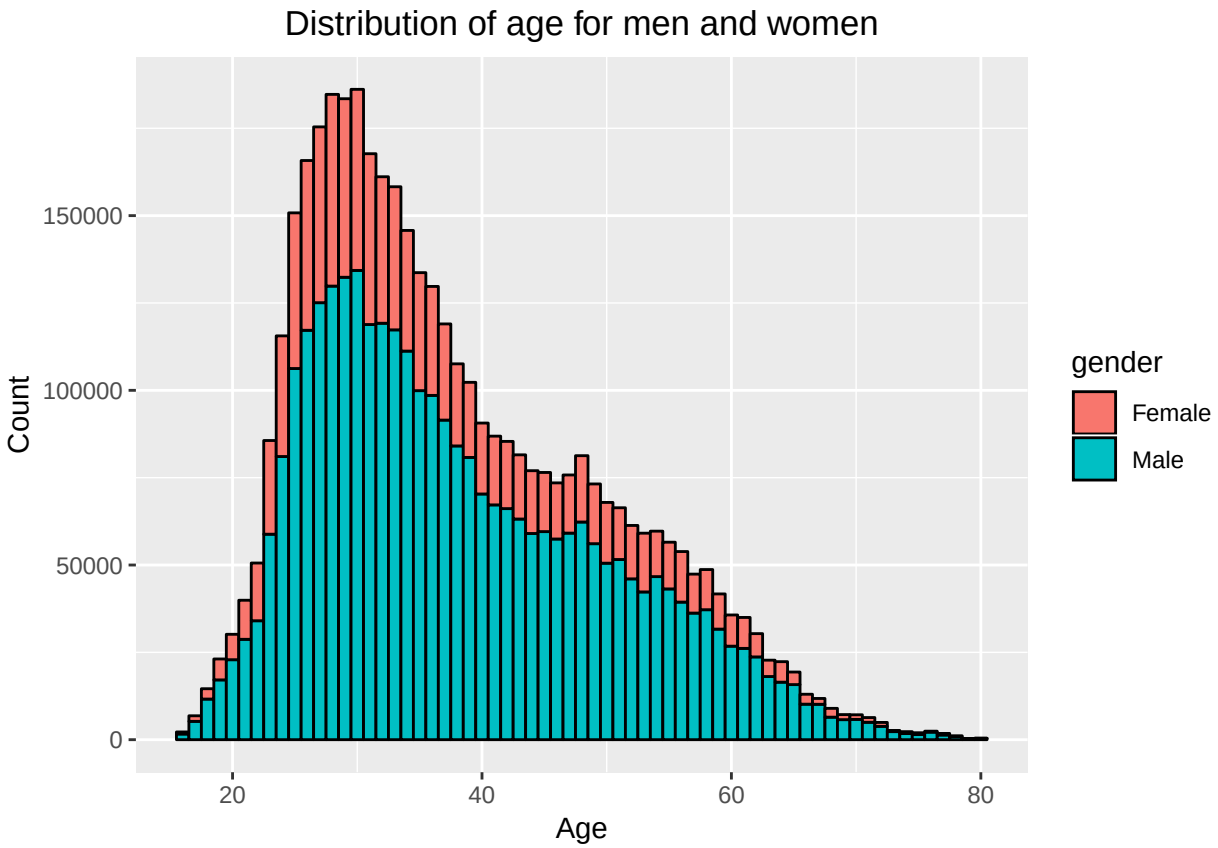
We can also observe the age distribution thanks to that plot. It shows that there are few users under 20, a mode around 30 and then a long tail. We can also note that there are more adults over 60 than what we expected.



There are significantly more males using Citibike than females. We will keep this in mind as it might provide insight in some findings.

3.2. Age repartition

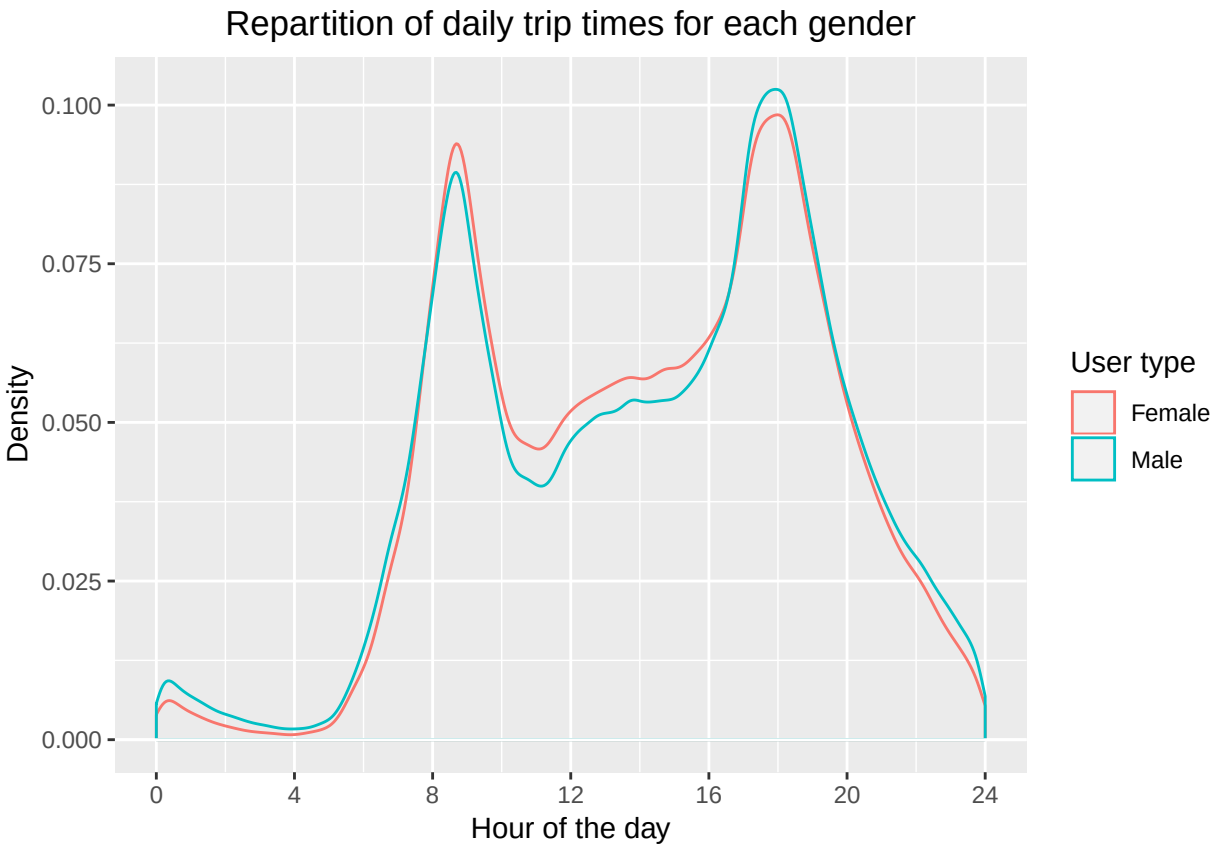
Let's then display a stacked bar chart displaying the age repartition, but separated between men and women - in order to find out from which age the discrepancy between men and women comes from.



We see that a majority of users in terms of gender are men, and in terms of age, most users are between 20 and 50. Moreover, almost all users above the age of 70 are men - which is even more surprising when you consider the fact that women have a higher life expectancy on average, meaning that there are much more women than men above the age of 70.

3.3. Do women commute as much as men ?

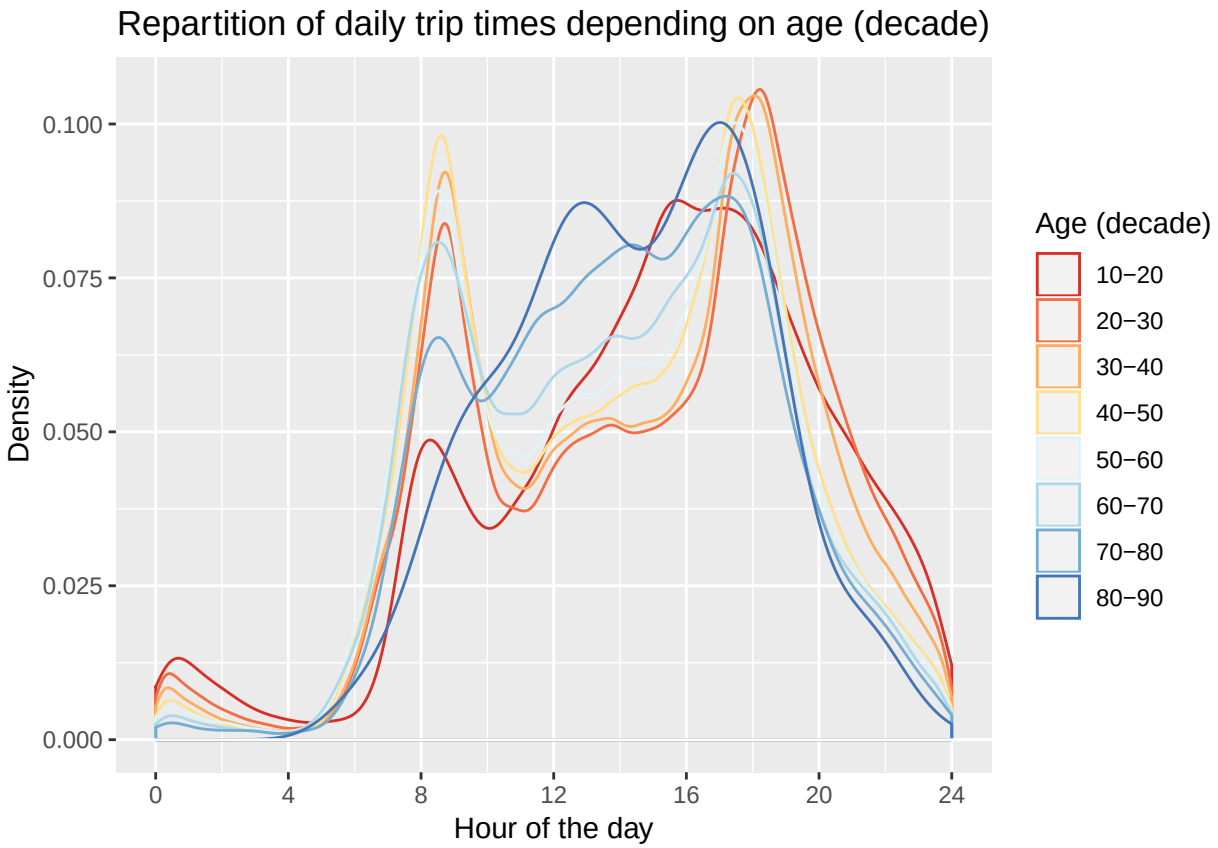
Let's now check if we notice any difference in the time of day at which the trips are taken depending on the gender of the user.



We don't see any difference between the daily trip times of men and women, which points towards the fact that women commute as much as men in New York City. The only difference we might notice is that women are slightly more active during the day, while men are more active in the evening and at night.

3.4. How do people of different ages behave daily ?

Let's now look at the repartition of trip times depending on the age of the user : in order to do that, we plot the density of daily trip time for each possible age of the user rounded to the nearest decade.



Here, we see what we expected to see : users from 20 to 60 generally use Citi Bikes to commute, as evidenced by the modes around 8 a.m. and 6 a.m., but people in their 70s and 80s, who are generally retired, use Citi Bike for leisure a lot more. Moreover, teenagers using Citi Bike tend to travel more at night (9 p.m. to 4 a.m.) than others, while being much less active in the morning - which indeed sounds a lot like teenagers.

In this second part, we have briefly seen what kind of demographic Citi Bike subscribers were part of. The average subscriber is a male between the age of 20 and 50, and he uses the system to commute, which corroborates our finding from the first part.

4. Customers and subscribers : two very different ways to use Citi Bike

Here, we will be looking at the "Customer Type" variable in the trip database. It is an unordered factor corresponding to the type of customer having taken the trip. The different levels are "Customer" and "Subscriber". The Customers are defined as the people using a 1 or 3 day pass, thus more likely to be occasional users or tourists, and the Subscribers are those with the annual pass, thus more likely to be regular users.

There are significantly more Subscriber trips than Customer trips, as we can wee below :

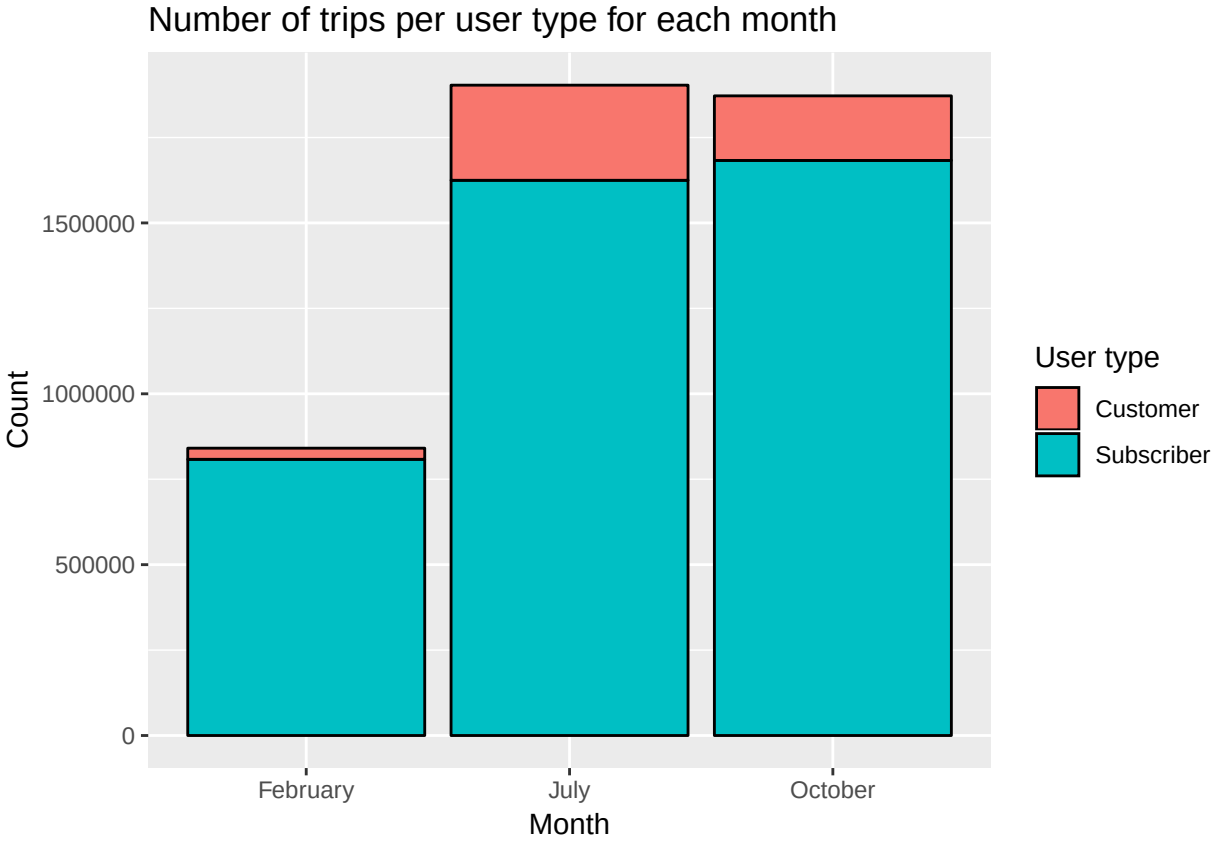
```
## # A tibble: 2 x 2
##   usertype         n
##   <chr>        <int>
## 1 Customer    499805
## 2 Subscriber 4116044
```

We would expect a service like Citi Bike to be used by subscribers much more than by occasional users, since the annual subscription (170 dollars) is much more cost-efficient than the 1-day pass (12 dollars). We observe

a roughly 10%/90% repartition. However, the size of our data (more than 4.5 million trips) makes it so that we can still safely assume that the trends observed among customers are actually meaningful.

4.1. Seasonal patterns

First, we want to observe seasonal patterns in the usage of Citi Bikes by customers and subscribers. We can't use the daily trip summary for that because the trip count isn't separated between customers and subscribers, so the only thing we can plot is the number of trips taken by customers and subscribers in the three months that form our trip database. We do that in the form of a stacked bar chart.



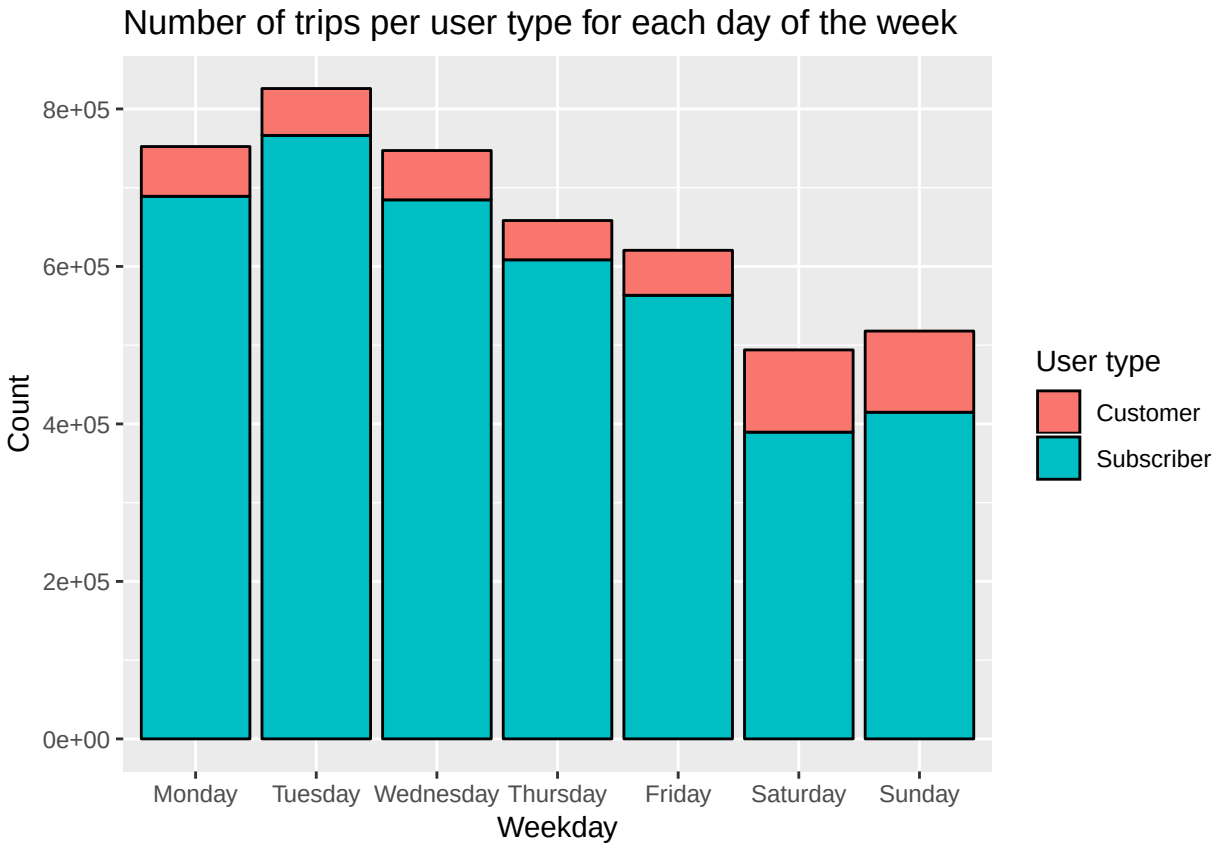
First of all, this confirms that an overwhelming majority of trips are taken by subscribers with the annual pass. As for the variation in the repartition between customers and subscribers, we can clearly see an evolution. We can see that there is a much lower proportion of customers compared to subscribers in the winter than in the summer. Moreover, between July and October, we see the number of customers decrease while the number of subscribers increases: this is a clear indicator that there is a lower proportion of customers in October than in July, although that proportion is still much higher than in February.

This seems logical, as we would expect occasional users and to use Citi Bike when the weather is more forgiving. On the other hand, Subscribers probably keep using the bikes in the winter by necessity more than for leisure (commute, getting from point A to point B on time for various commitments) and are probably more prepared to rent bikes for the whole season with proper gear to face the cold and often snowy streets.

Also, we might follow with the hypothesis that Citi Bike is used by tourists - and tourists would in all likelihood be customers rather than subscribers. Then, the increasing number of tourists in the summer compared to the winter might be another cause for that observed difference.

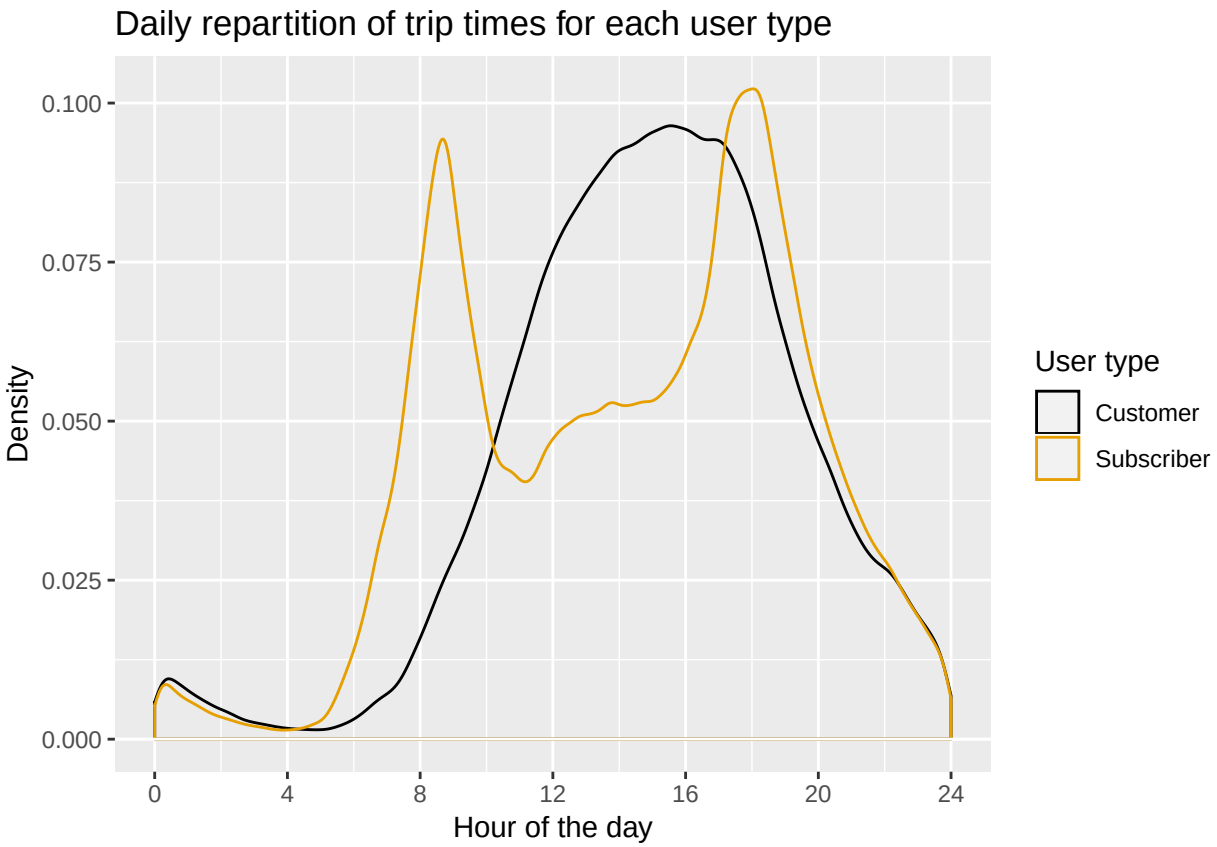
4.2. Weekly patterns

We've already seen that Citi Bikes were being used less on weekends. Let's now see whether the weekly usage differs between the two categories of customers. In order to do that, we plot another stacked bar chart.



We see here that customers actually use Citi Bike on weekends more than they do during the week, while it's the opposite for subscribers. This, once again, supports the hypothesis that most subscribers use Citi Bike to commute, and adds the information that customers would use Citi Bike for leasure, hence the increased numbers on weekends. Please note that this doesn't necessarily corroborate the hypothesis that customers are mostly tourists, as tourists generally wouldn't care whether or not they're visiting during a weekday or during a weekend day. Instead, it probably means that there are some NYC residents who use the system to go out during the weekend, but not often enough for it to be worth it to get a subscription.

4.3 Daily patterns

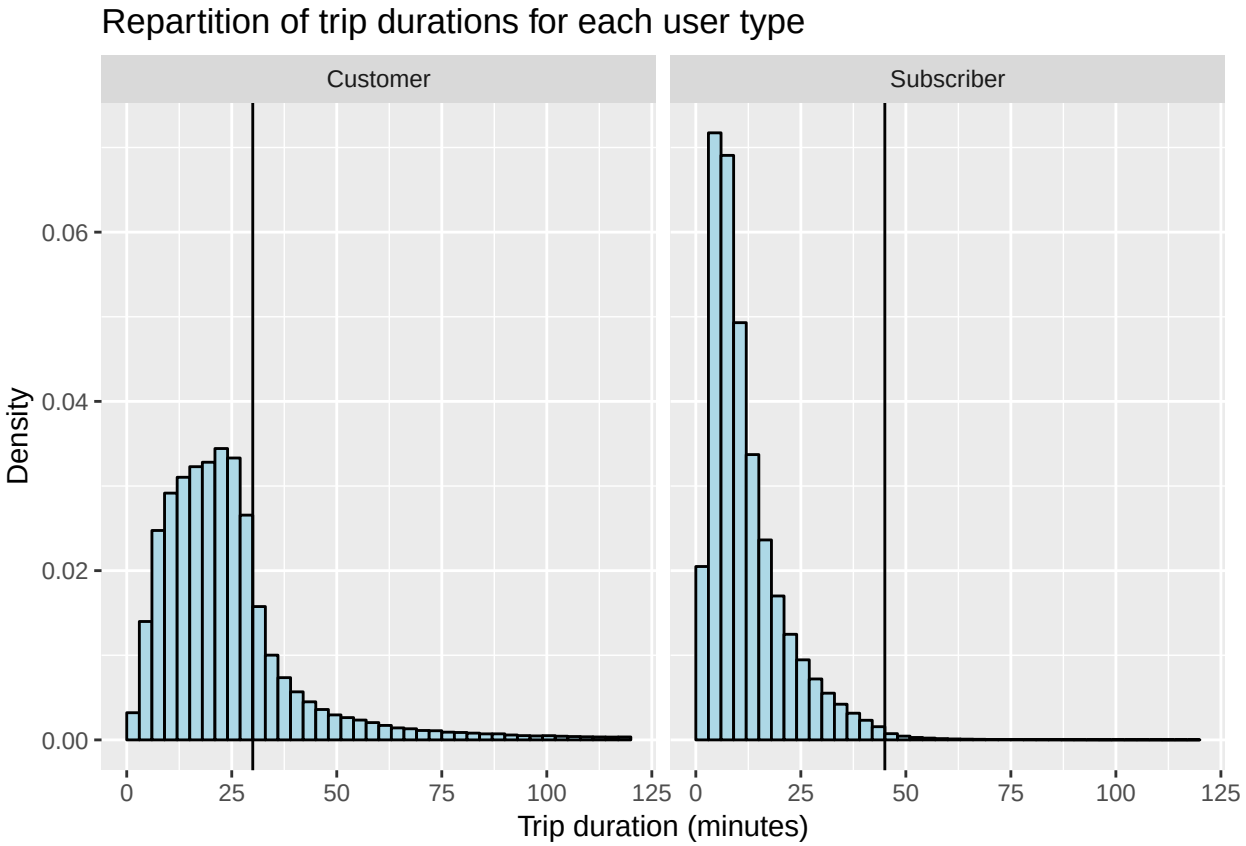


We notice a clear difference in the density curves between customers and subscribers, and this difference is another strong piece of evidence in favor of the hypothesis that customers might be in majority tourists. Indeed, we see a curve that's roughly bi-modal for subscribers, with modes aroung 8 a.m. and 6 p.m., which are times usually associated with commuting. On the other hand, the curve for customers is unimodal, with the mode roughly around 3 p.m., and a much slower start to the day; which indicates that the bikes are being used for leisure.

4.4. Types of trips taken

We want to highlight the difference in the types of trips taken by customers and subscribers. First of all, we want to see if there's any difference in the duration of those trips.

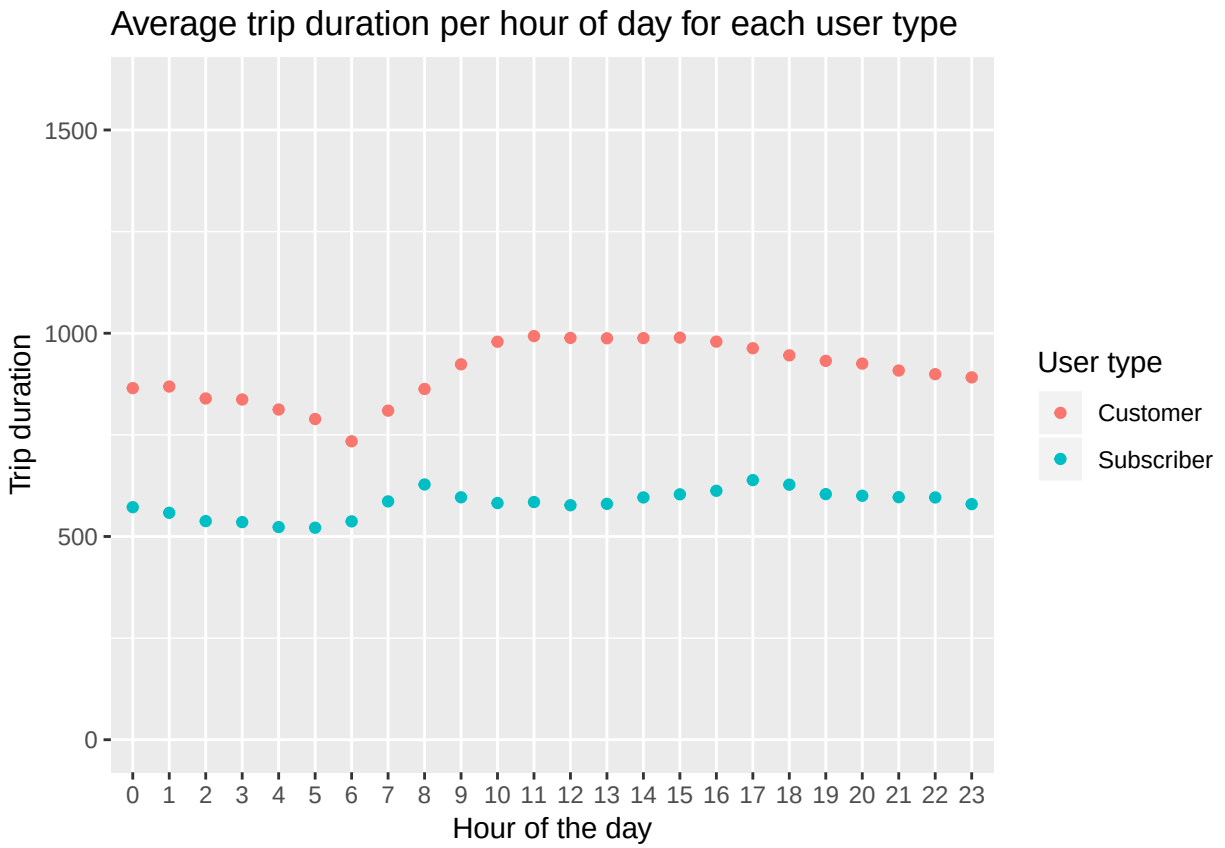
Here, we plot vertical lines for both plots that signal the end of the time included in the plan, which is 30 minutes for customers and 45 for subscribers.



Once again, we can notice a clear difference between both repartitions. Indeed, subscribers tend to favor shorter trips, with the overwhelming majority of trips taken being around 10 to 15 minutes long - that might correspond to the time for commuting. Customers, on the other hand, take longer trips, with the mode being around 25 minutes. This can be explained by the fact that we consider some customers to be tourists, and tourists would explore the city much more than residents, thus taking longer-lasting trips.

Moreover, we can notice that customers are generally much keener on taking long trips, going as far as to make them longer than the time that is included in the plan. Indeed, even though we do notice a significant dropoff between the 25-30 and 30-35 minutes bins, a significant part of customer trips are still longer than 30 minutes, with some of them going well past 1 hour. As for subscribers, they usually take very few trips longer than 45 minutes, which is their alloted time limit.

Something that might also reveal something would be the average trip length for each hour of the day. Let's show that plot :



First of all, we get a confirmation that the average trip is longer for customers than for subscribers. Other than that, we can't see that much on this plot, since the trip durations are mostly similar throughout the day; however, we do notice two small modes around commuting hours for subscribers, indicating that their trips to commute might be longer than their trips during the rest of the day. As for customers, we see that their trip are longest between 10 a.m. and 6 p.m., which corresponds to the times at which tourists might use Citi Bike to wander around the city without really caring for getting from point A to point B.

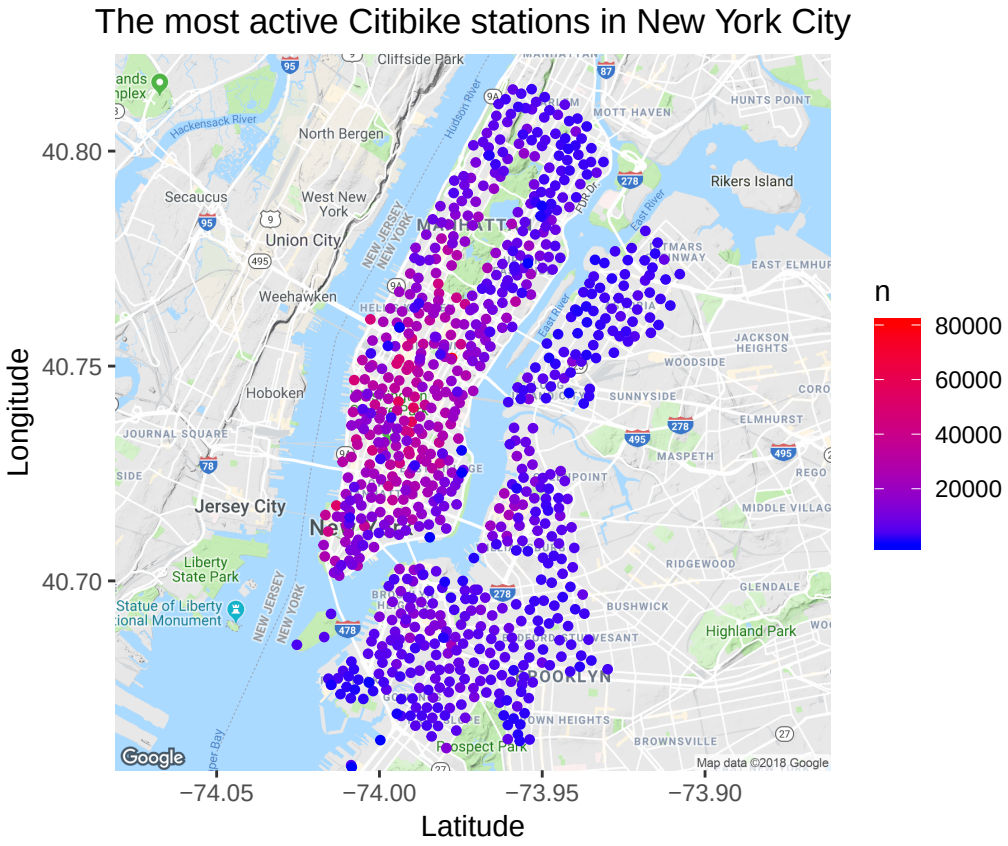
5. Geographical trends

In this part, we will try to identify geographical trends to try and understand what use people have of the cibibikes. Based on New York geography and its subway tracks, we expect to find some proofs that Citi Bikes are a means of transportation for routes that are harder to do with subways, for instance East-West commutation. Furthermore, we will try to identify trends that shows that bikes are both a recreational and a practical way of commuting in New York. For the following study, we will focus on geograpical features, such as the station locations, as well as the route and the distance done by users (we do discard any seasonality effect by aggregating over the study over the 3 months we chose previously).

First of all, we need to join the station info data with the trip data, in order to get the correct latitudes and longitudes for each trip.

5.1. Most active stations

First of all, we can visualize the stations that are the most 'active': those are the stations that have the highest traffic of Citi Bikes (in terms of start and end for each station).

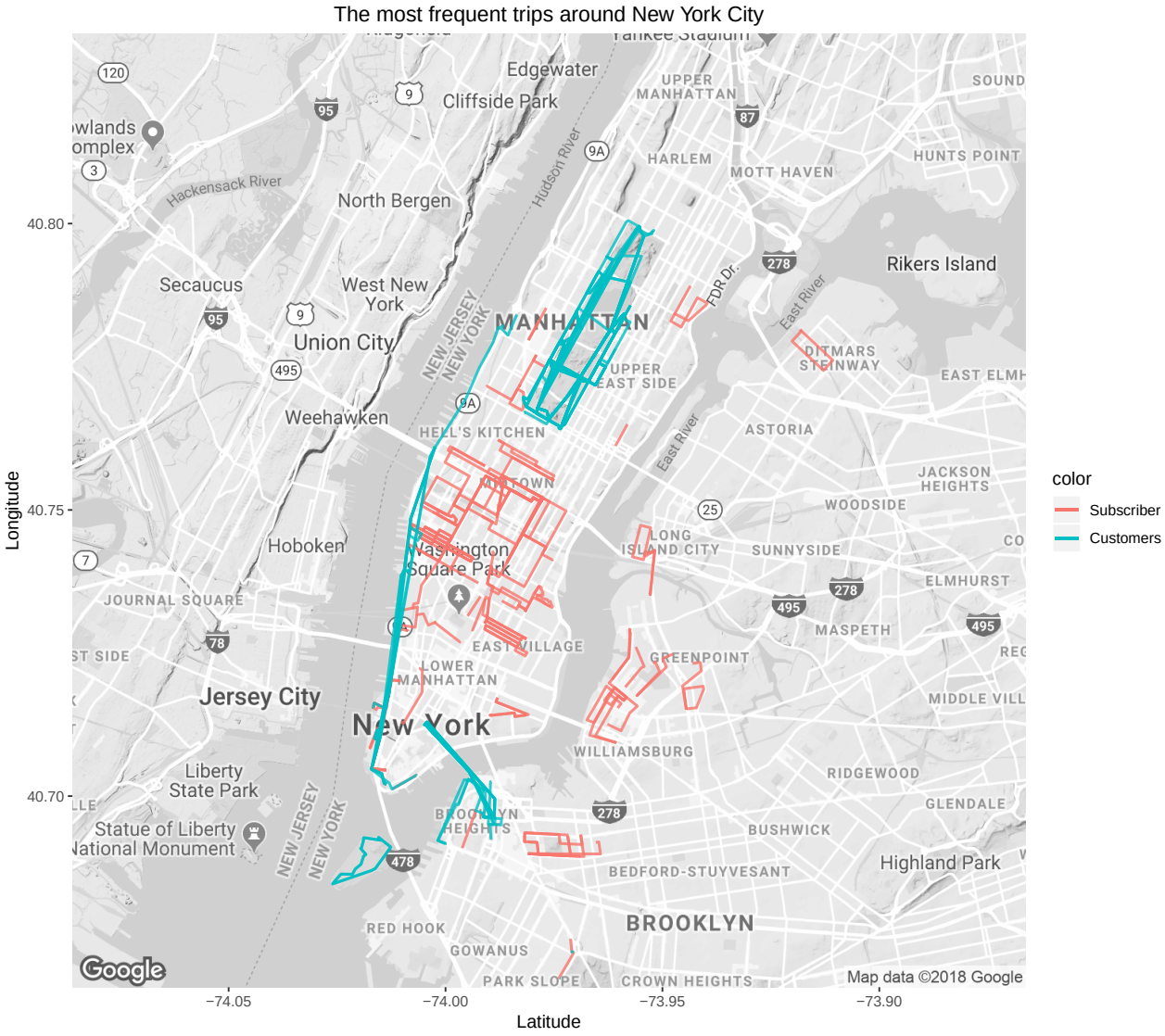


From the map, we can see that the there are multiple 'active' zones but the general trend is that Midtown is the most active part of Manhattan. There is also a lot of activity in Downtown, along Central Park and in the neighborhoods of Brooklyn that are close to Manhattan. Areas close to Central Park, or in Brooklyn are zones that we can consider to be mostly visited by tourists. However, the high activity of citibikes in Midtown could be further evidence that citibikes are used for commuting since Midtown is a working area of Manhattan. Furthermore, the activity in Downtown also seems more likely to be the result of commuting then visiting.

A natural follow-up question would be to try and find whether the most active station change between customers and subscribers, and also whether they change depending on the season. We can do just that by displaying the same map, with markers of different colors for customers and subscribers, and faceting it by season.

5.2. Most commons routes

To go further in our analysis, we can visualize the most common trips made by citibikes users. To do so, we will gather for each trip the start station and the station of arrival, and compute the most likely route the users took by using the route function from ggmap package (and using the option so that the routes are tailored for bike use). We first have to create adequate dataframes for customers and subscribers.

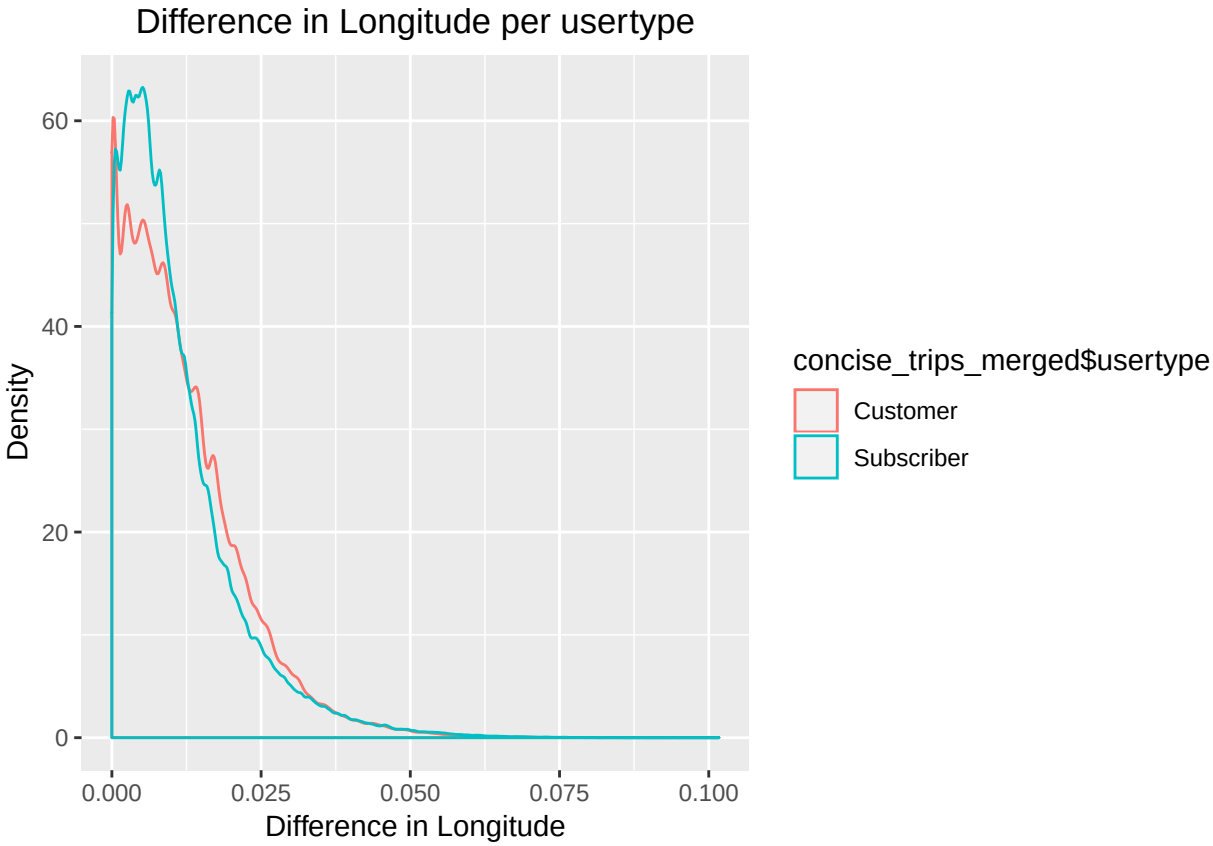


The above map provides some interesting information:

- There are some main zones where the trips are done: around Central Park, in Midtown as well as along the Hudson River.
- Touristic zones can be identified: Central Park, Governor's island for instance
- The Hudson River route is particularly interesting because it is the very example of a hybrid route: it is widely used by new yorkers since it is the fastest way for going north/south. But it is also a very scenic route, thus a lot of tourists use it to see the shore
- Finally, it seems that an important part of the trips that are done in Midtown are mostly longitudinal: this might be a clue that people are using citibikes to compensate for the lack of subways in the East-West direction.

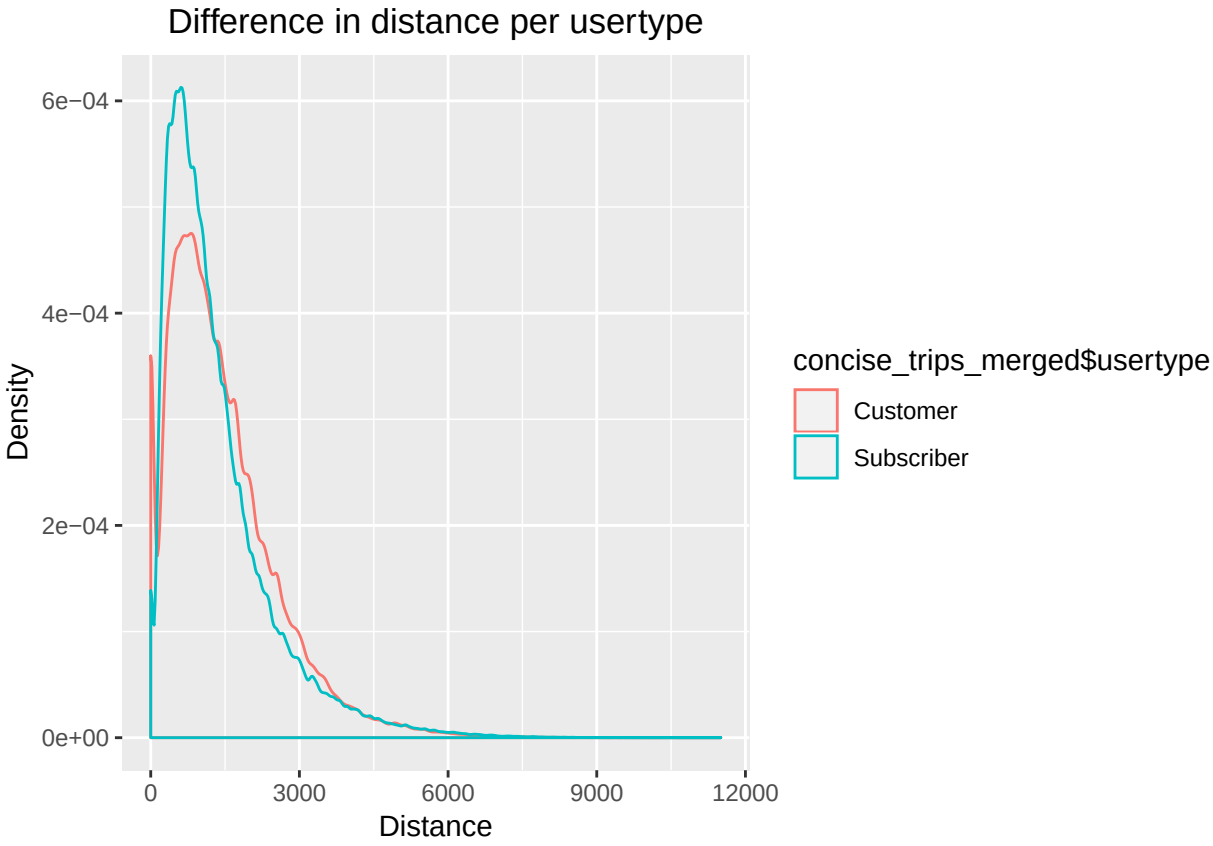
5.3. Geographical patterns for trips

Finally, we would like to study whether there are some geographical patterns related to latitude/longitude features. More specifically, we would like to see if there are some discrepancies between subscribers and customers on this particular matter.



Here we have plotted the absolute difference in longitude between start and finish. Because of the layout fo New York, we can approximate longitude difference as a North/South difference and latitude difference as a East/West difference.We can see that for both latitude and longitude, subscribers have a much higher peak close to low values.

It reveals yet another difference in the behavior of subscribers and customers: subscribers usually do shorter rides (in terms of distance) specially if it is for a North/South difference. This seems to follow the logic of commuting: commuters will not do lengthy North/South rides since it is slower than taking the subway. However, for short rides, it can be quicker to take a bike.



If we plot the distribution of distance for both customers and subscribers, it confirms what we previously saw with latitude and longitude: subscribers tend to do shorter rides (in terms of distance).

Therefore, there is a genuine discrepancy between subscribers and customers which attest that there are two different ways of using citibikes: short rides or more lengthy rides, and it confirms that citibikes are not only used by tourists, but also by New Yorkers as a way of commuting.

V. Executive summary

VI. Interactive component

Here is the link to the interactive component for this project :

(link)

VII. Conclusion