Citibike User Behavior

Corentin Llorca (cl3783)

24 November 2018

Who uses Citi Bike ?

Customers and Subscribers, and their usage of Citi Bike

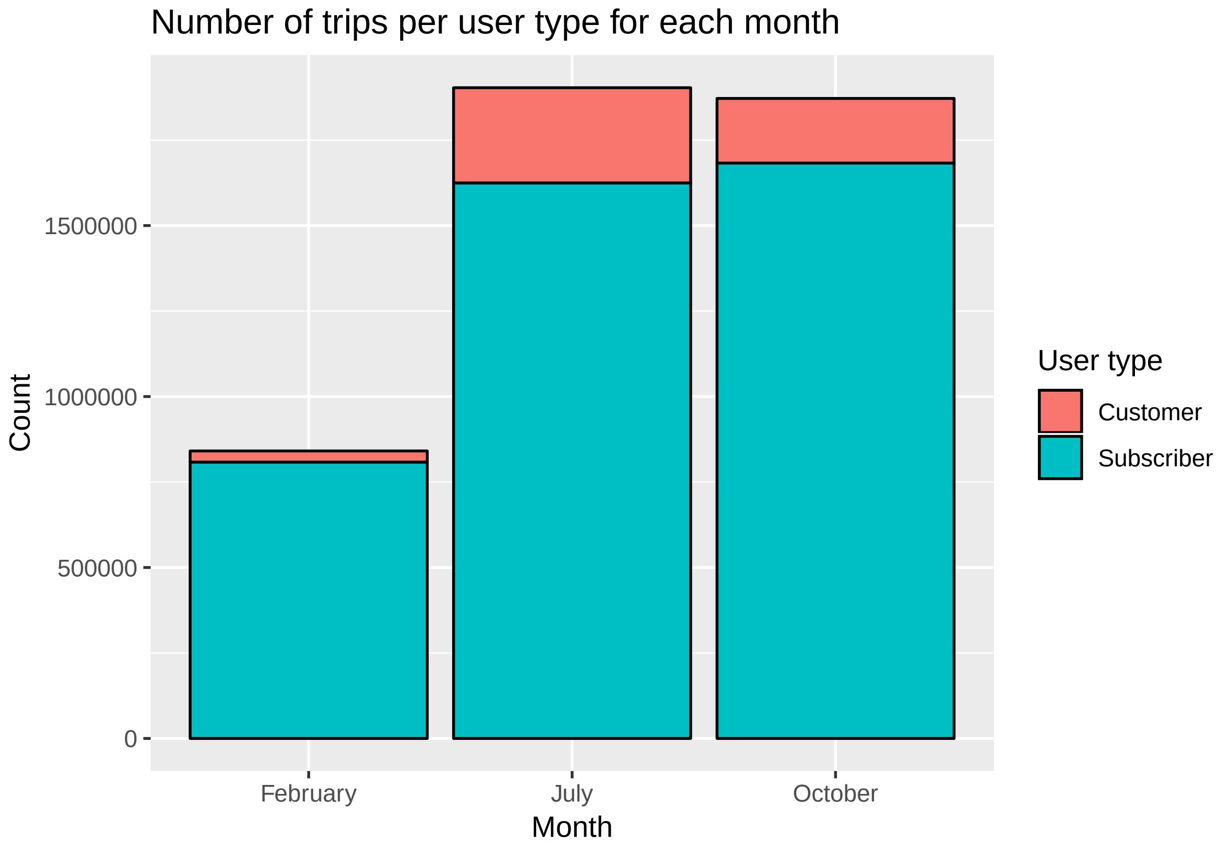
Here, we will be looking at the “Customer Type” variable in the trip database. It is an unordered factor corresponding to the type of customer having taken the trip. The different levels are “Customer” and “Subscriber”. The Customers are defined as the people using a 1 or 3 day pass, thus more likely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

The two levels are highly unbalanced, as we can see 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.

Seasonal patterns

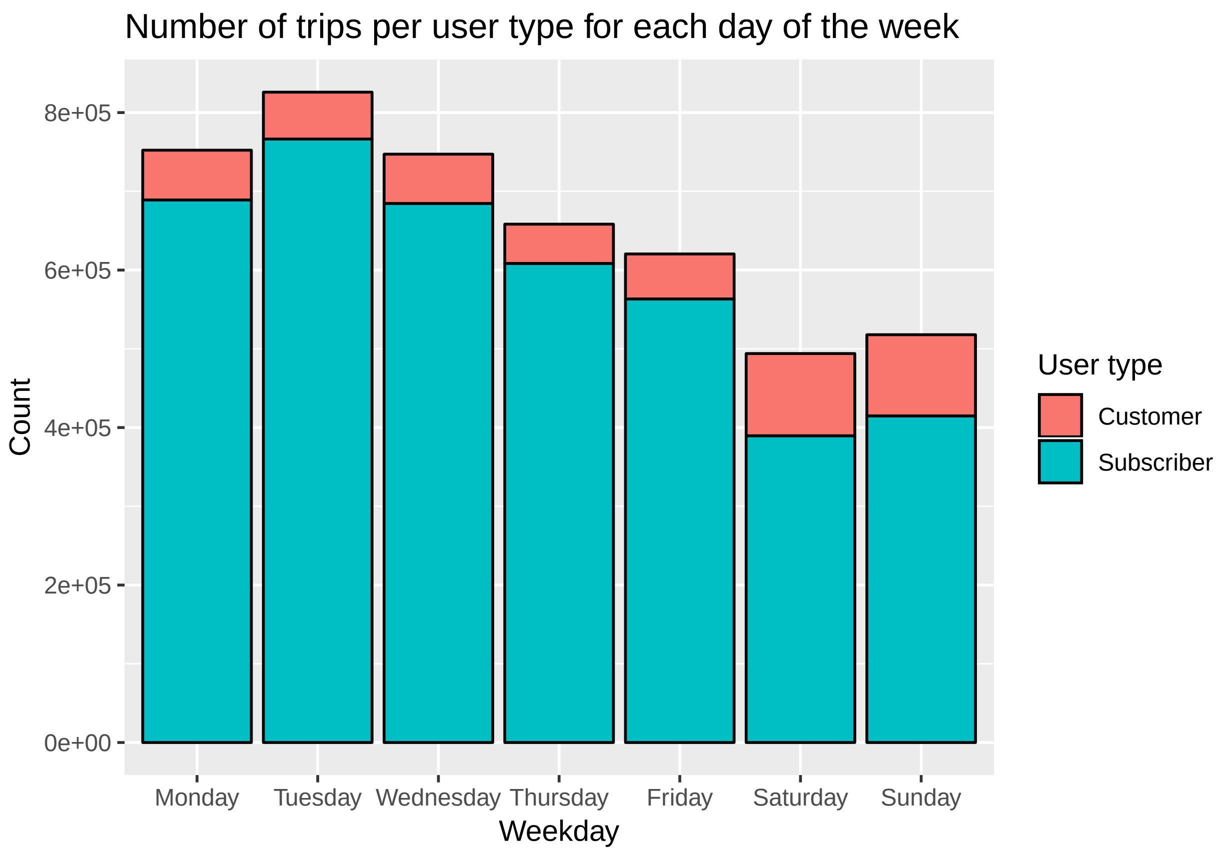


First of all, this confirms that an overwhelming majority of trips are taken by subscribers with the annual pass. As for the variation in the repartition between customers and subscribers, we can clearly see an evolution. We can see that there is a much lower proportion of customers compared to subscribers 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 to use Citi Bike when the weather is more forgiving. Also, we might emit 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.

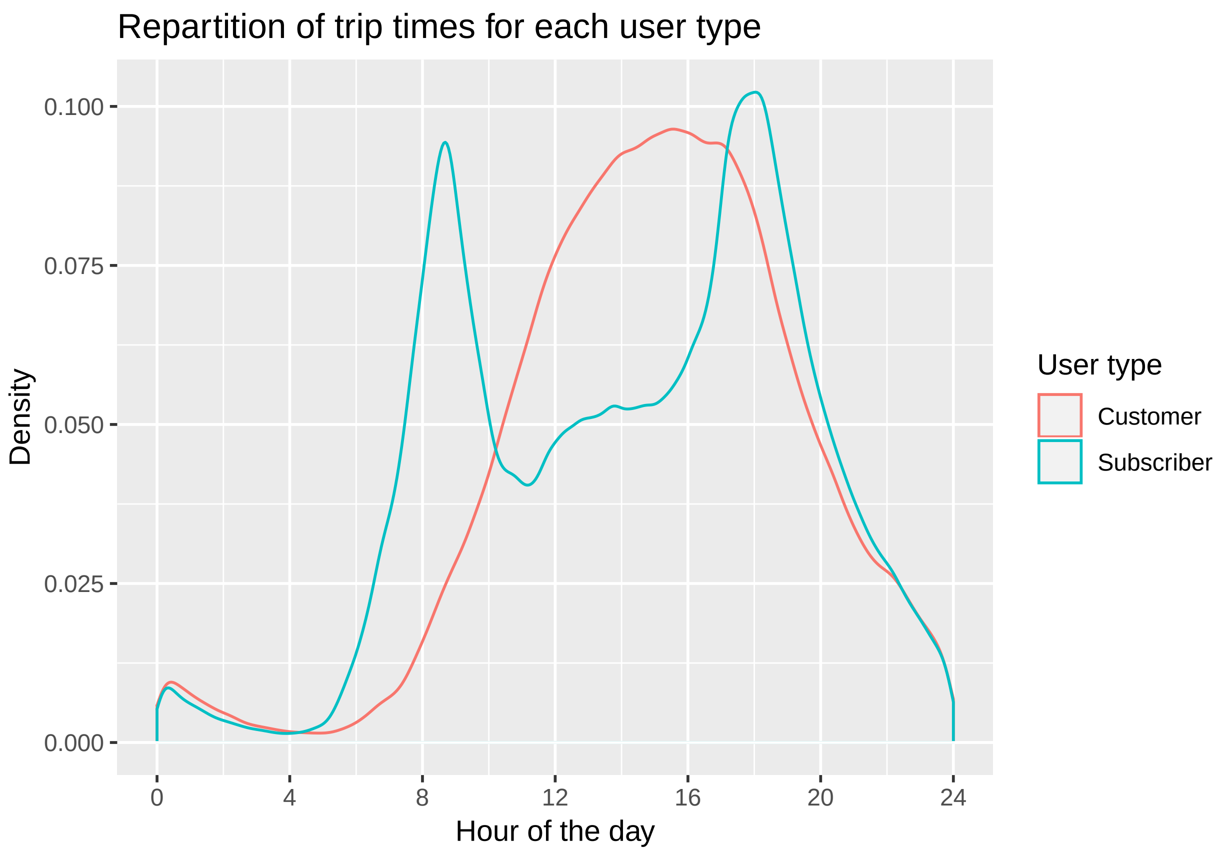
Weekly patterns

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.



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 suggests that subscribers might use Citi Bike to commute to work, while customers use it more for leisure, hence the important use on weekends.

Daily patterns



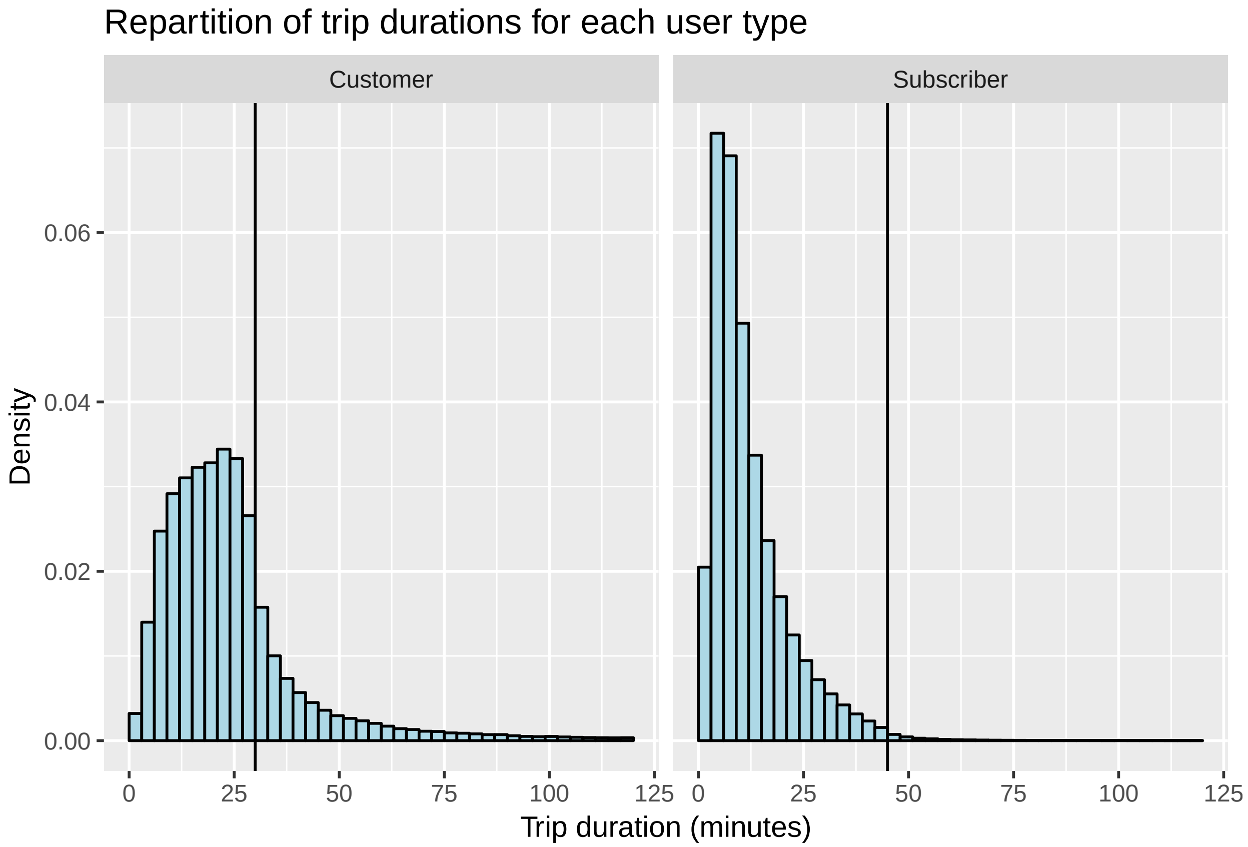
We notice a clear difference in the density curves between customers and subscribers, and this difference is another strong piece of evidence in favor of the hypothesis that customers might be in majority tourists. Indeed, we see a curve that's roughly bi-modal for subscribers, with modes aroung 8 a.m. 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.

Types of trips taken

Trip duration

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.

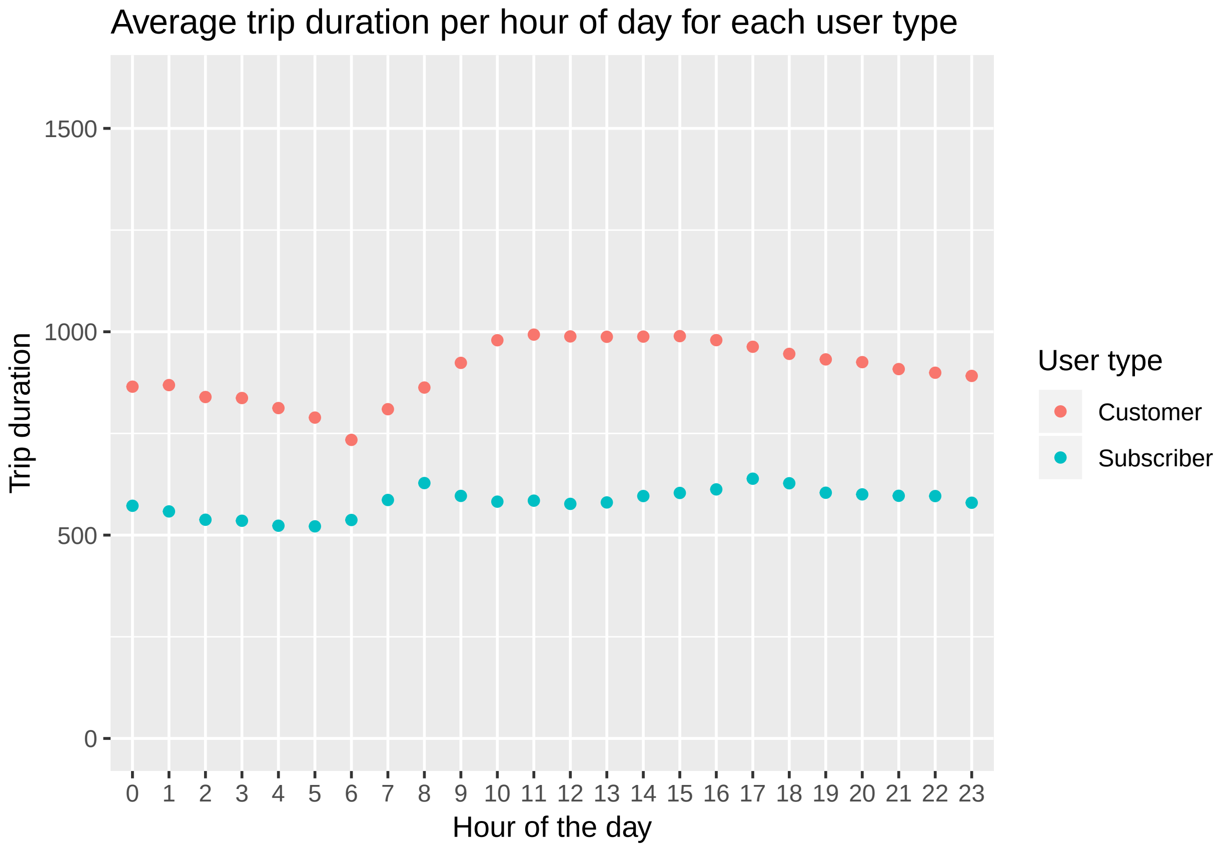
Here, we plot vertical lines for both plots that signal the end of the time included in the plan, which is 30 minutes for customers and 45 for subscribers.



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 :

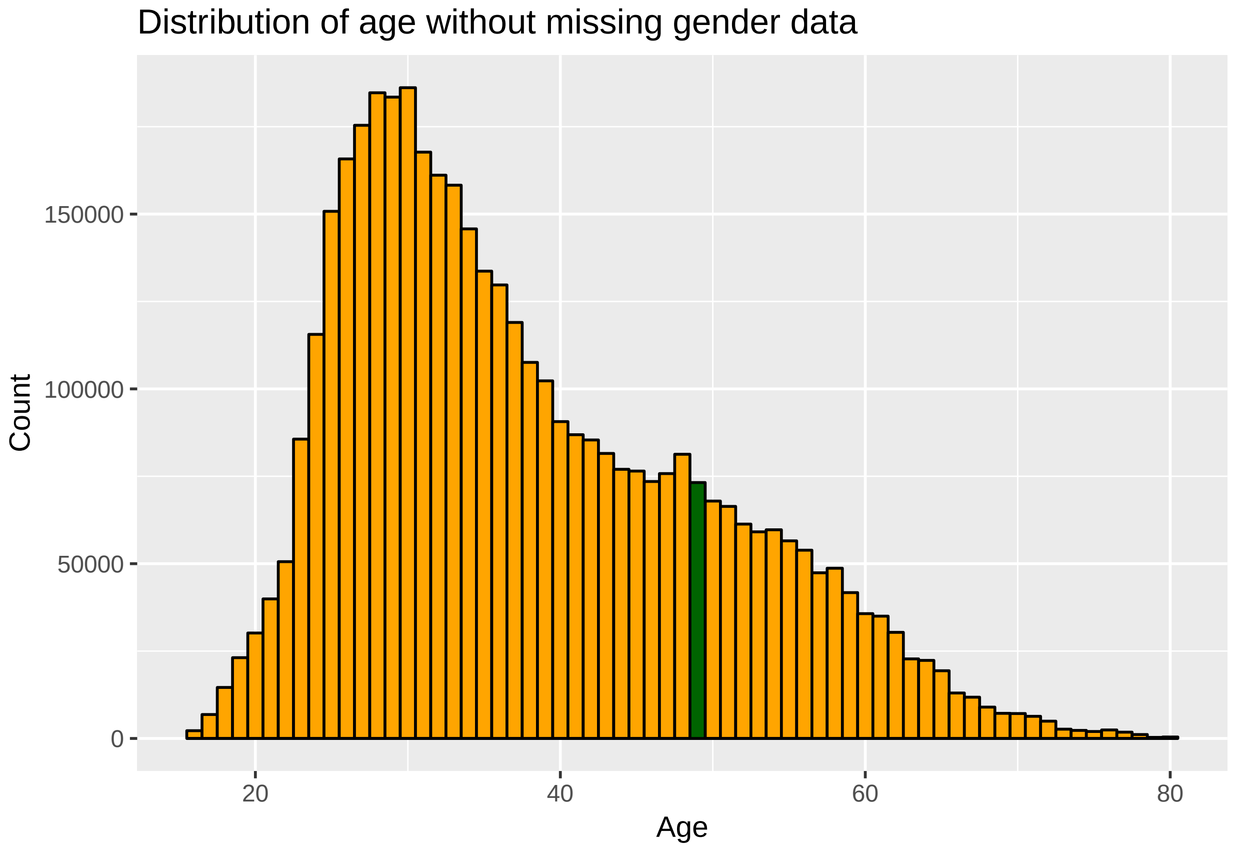


First of all, we get a confirmation that the average trip is longer for customers than for subscribers. Other than that, we can't see that much on this plot, since the trip durations are mostly similar throughout the day; however, we do notice two small modes around commuting hours for subscribers, indicating 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.

Age and gender of users

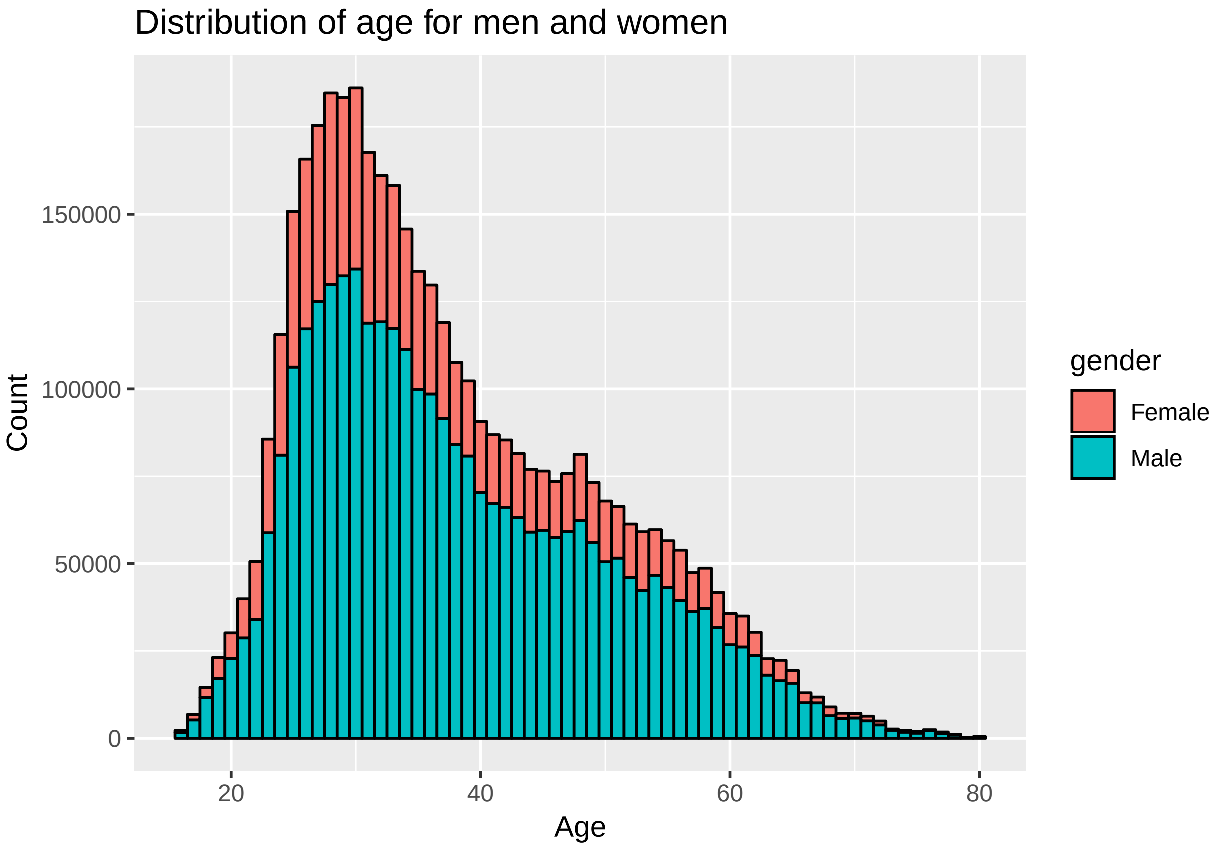
The Citi Bike public database contains informations about the age and gender of the users. However, that is only the case for users with the subscriptions, as simple customers with short-term passes aren't required to fill in any information. Still, we have seen that subscriber trips represented about 90% of all trips, so the following studies on age and gender can be safely extended to the whole user base of Citi Bike.

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.



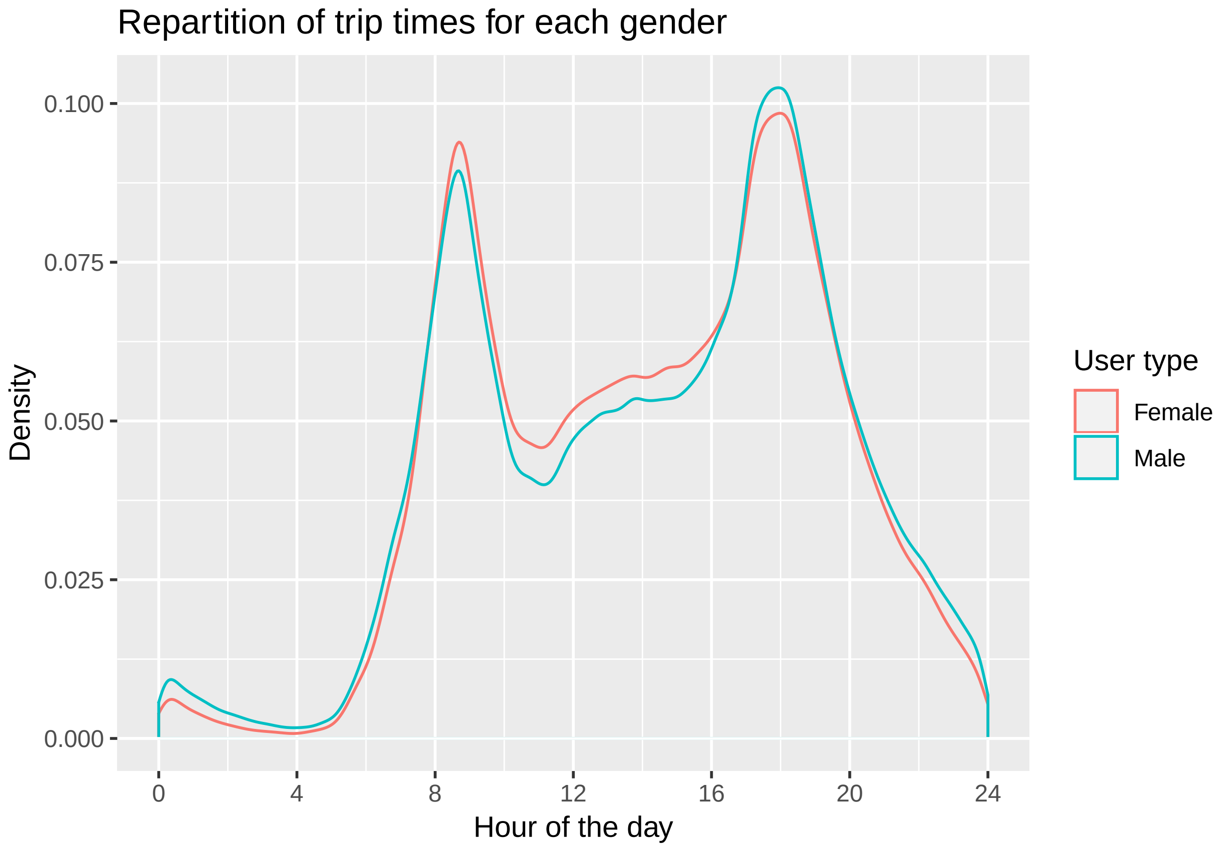
We see, indeed, that the 1969 anomaly is removed when we exclude data in which gender is missing. This could be explained by the fact that people who don't fill out their gender might be the same people who would leave the default birth year when subscribing. We'll continue on with that data for this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

Let's first display a stacked bar chart displaying the age repartition, but separated between men and women.



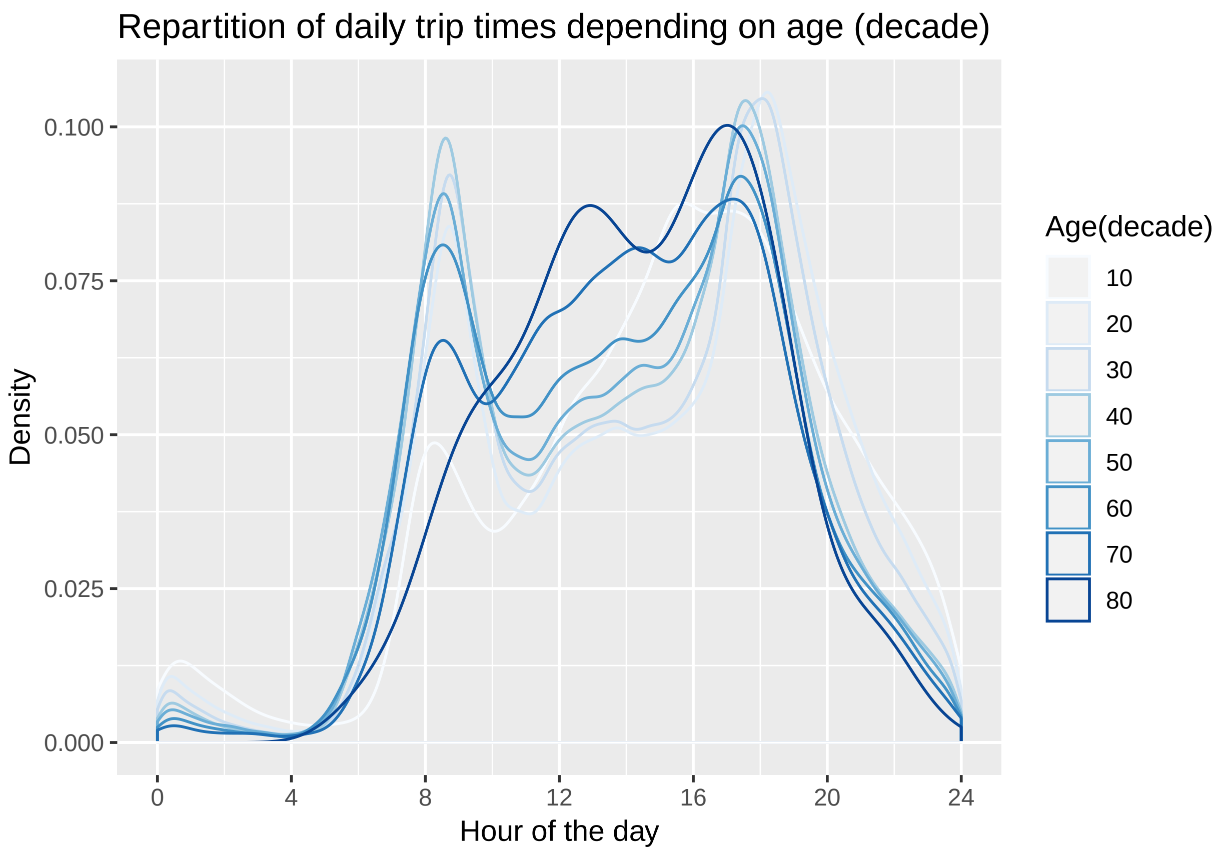
We see that a majority of users in terms of gender are men, and in terms of age, most users are between 20 and 50. Moreover, almost all users above the age of 70 are men - which is especially significant when you consider the fact that women have a higher life expectancy on average, meaning that there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

Let's now check if we notice any difference in the times 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 more active during the day, while men are more active in the evening and at night.

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.

We now want to look at the difference in speed between people of different ages and genders. In order to do that, we add a “speed” column, which corresponds to the distance divided by the trip duration. The speed is then in m/s.

Todo : avg speed per age with separation of genders