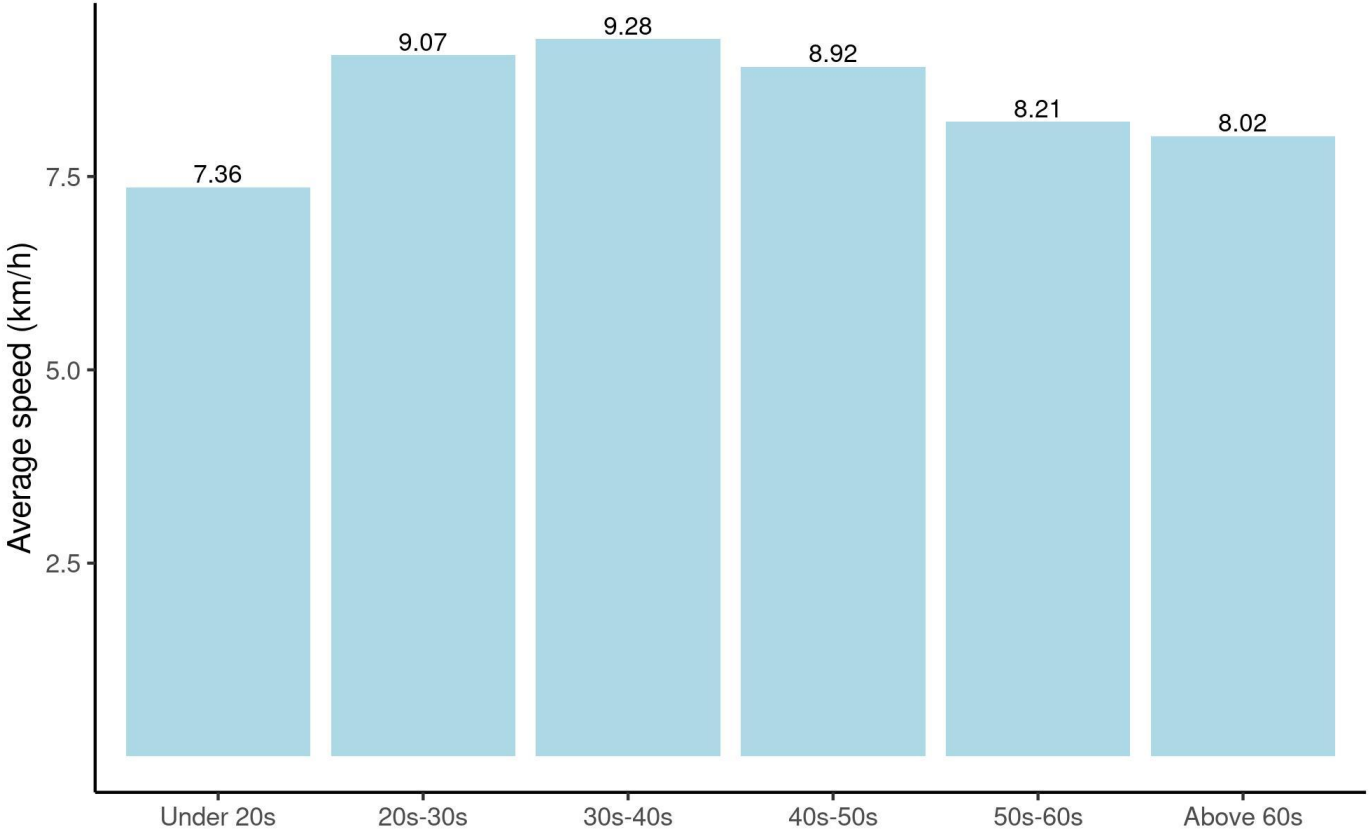
Citibike NYC Data Analysis

R Project - Jackie Pham

Background

- This project is to explore the data associated with the New York City bike share program, Citi Bike.
- There are over 850 Citi Bike stations in New York City; users check a bike out from a starting station and then dock that bike at a different station when they reach their destination.
- Hereby dataset contains information about 2,248,869 individual trips for October 2020.

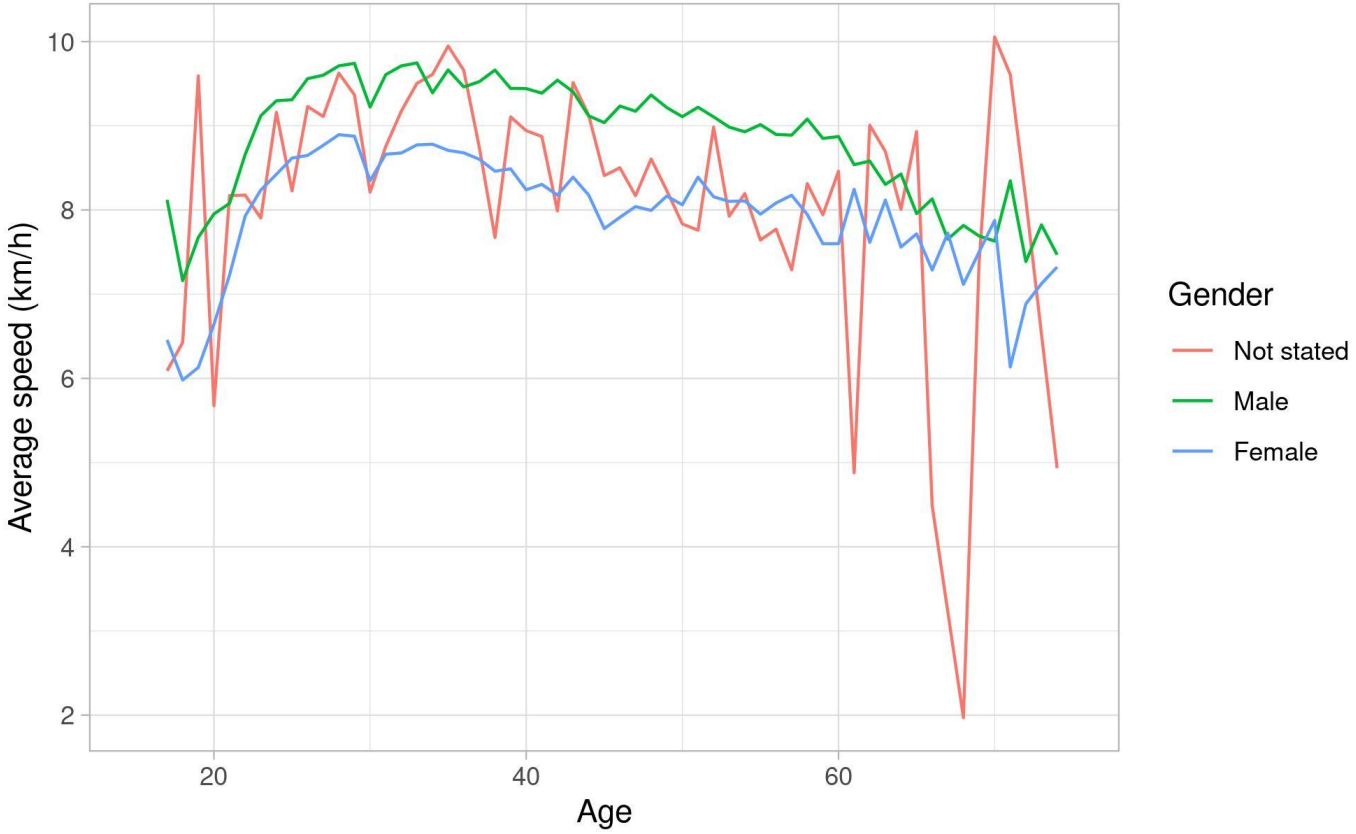
Average speed of different age groups



The 20-30 yo group has the highest average speed (9.28 km/h)

It's understandable that group aged under 20s ranks bottom, even lower than above 60s group. It's because the dataset only includes 16 yo plus users.

Average speed by age and gender (1)

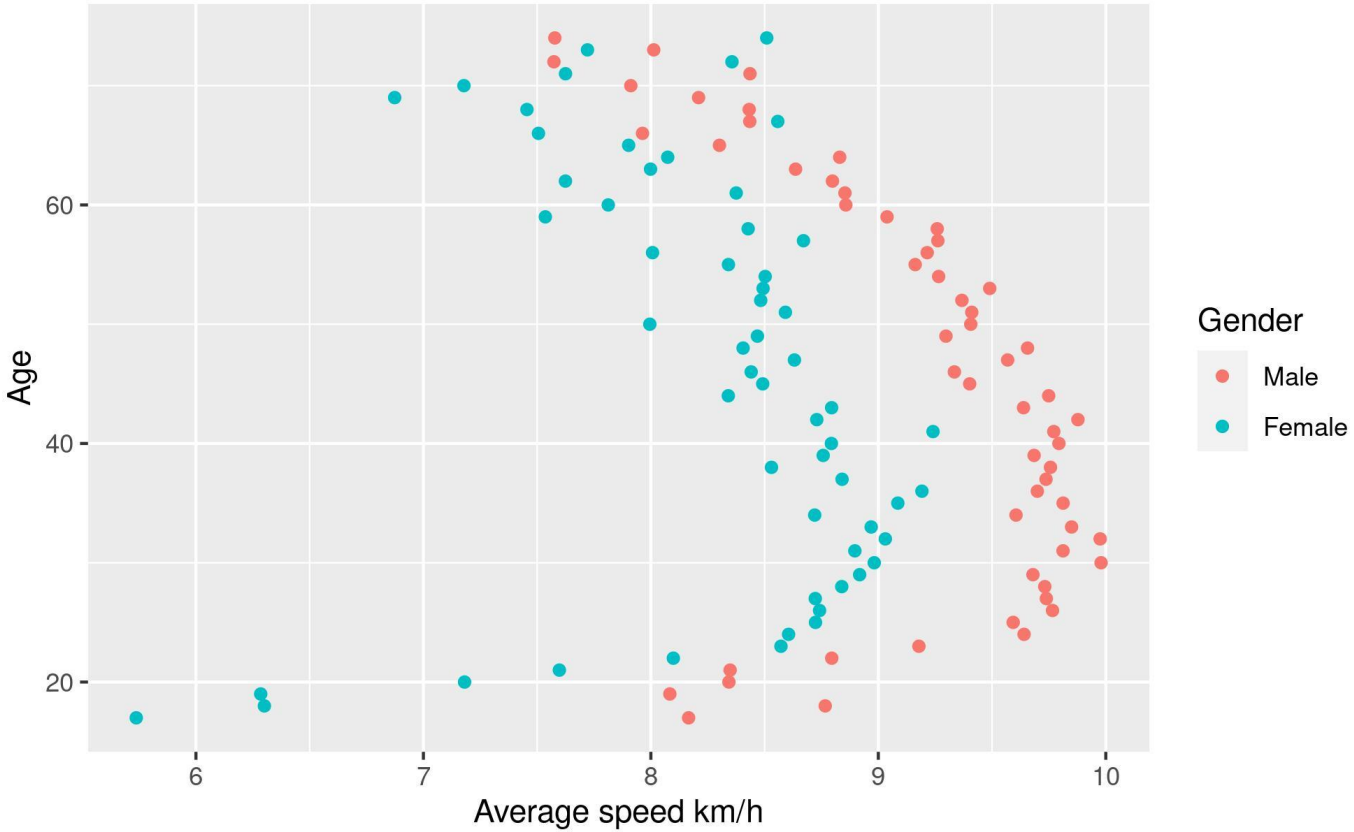


Male users, on average, have higher average speed than female users in all groups.

When doing more calculations we find that average speed of male group is approximately 8.846 km/h while female group's is 7.941km/h (1.114 times higher)

Users with an unknown gender do not follow any specific pattern. It is likely that there isn't enough data to properly visualise those users.

Average speed by age and gender (2)

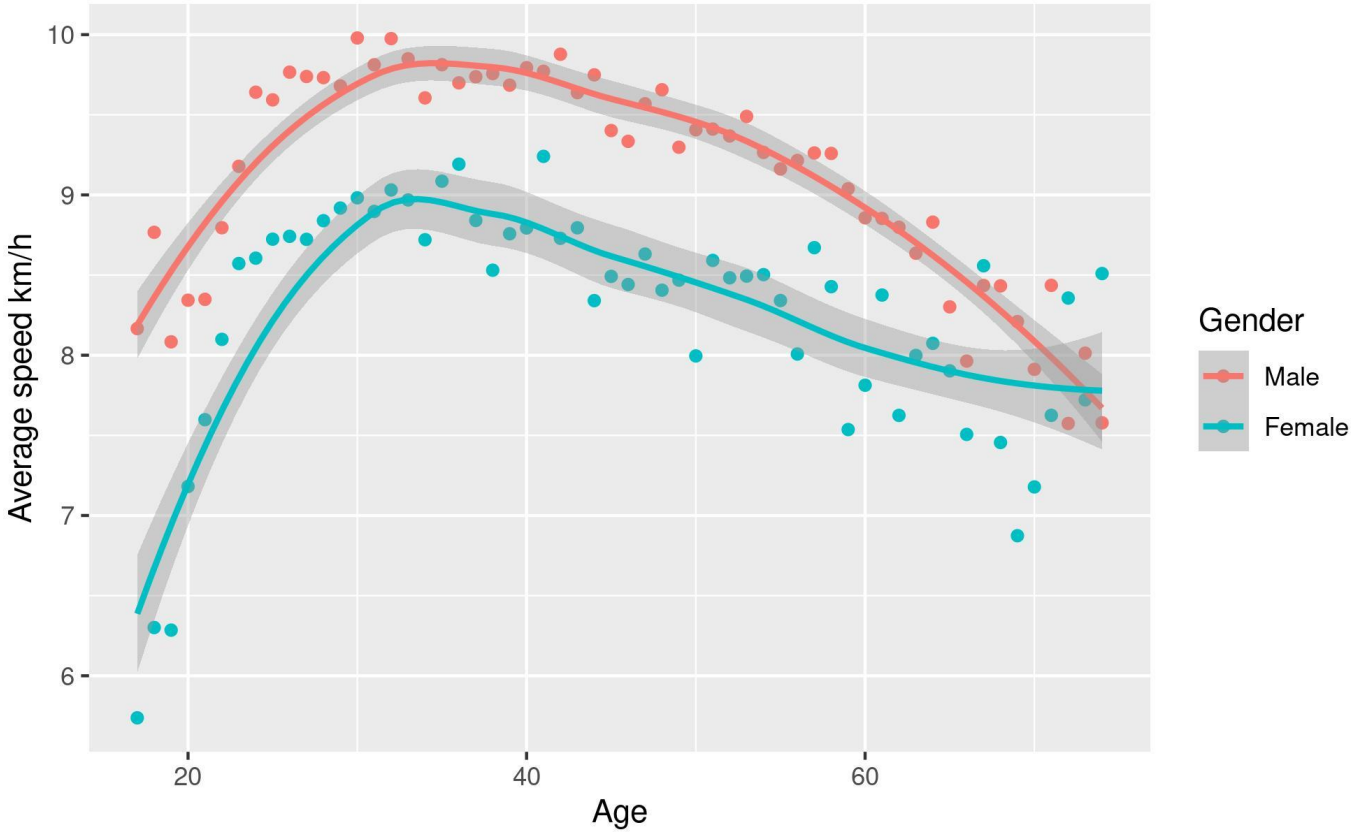


The group that didn't state its gender has been removed from the plot.

This scatter plot helps observe the average speed that most users circle around.

We can see that the average speed of both group are around 8 km/h - 10 km/h. Although male riders cycle slightly faster than their female counterparts.

Average speed by age and gender (3)



This plot shows that both groups reach the peak of their speed at the aged around 35-40 yo.

Is the average speed affected by age and gender?

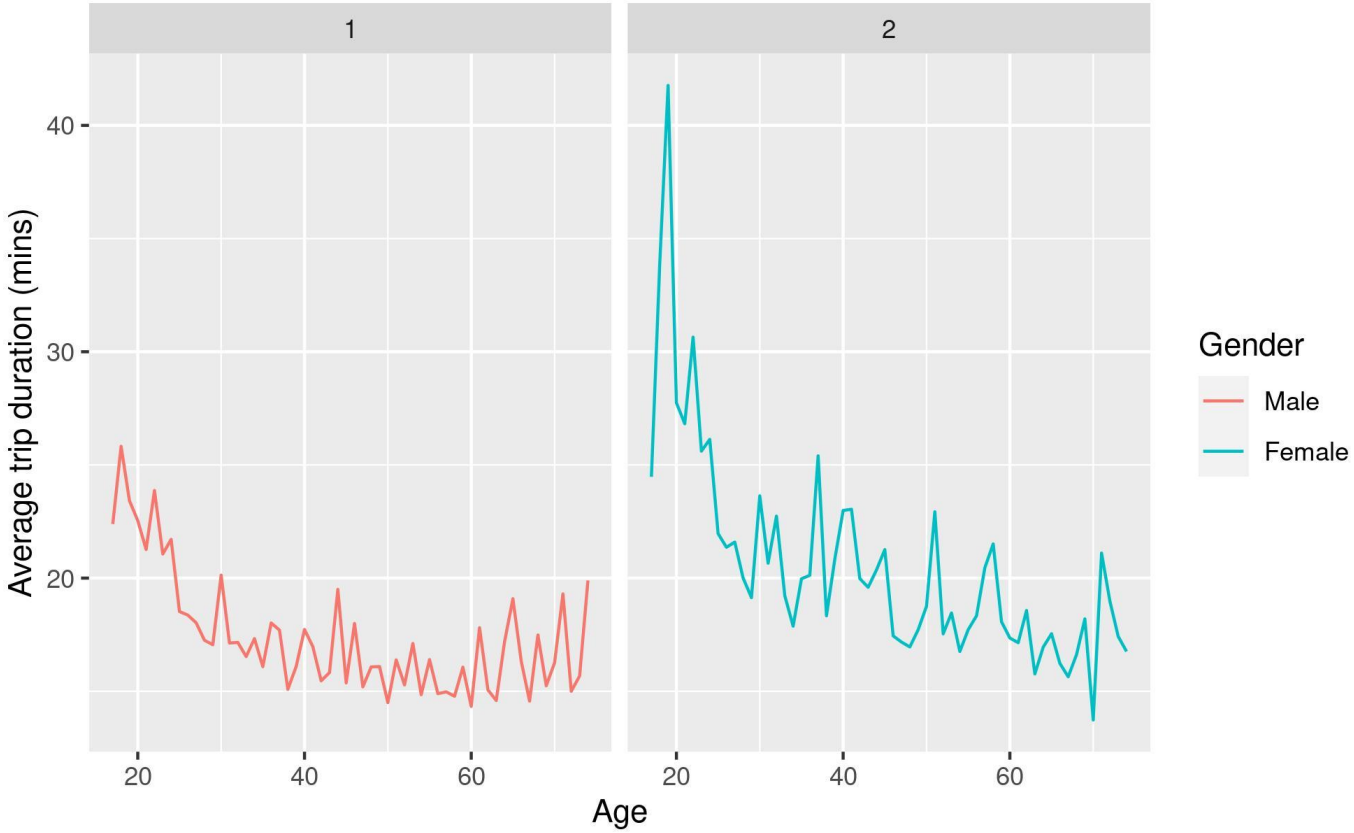
Relationship between average speed, age and gender

```
Call:
lm(formula = mean_speed ~ gender + age, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5325 -0.5855  0.2192  0.5412  3.2967

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.988591   0.300102  29.952  < 2e-16 ***
gender1      0.692340   0.250349   2.766  0.00664 **
gender2     -0.158002   0.249326  -0.634  0.52755
age         -0.013421   0.005961  -2.252  0.02628 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.077 on 113 degrees of freedom
Multiple R-squared:  0.1382,    Adjusted R-squared:  0.1153
F-statistic: 6.041 on 3 and 113 DF,  p-value: 0.0007472

[1] 0.1382187
```

Running a model to see the relationship between outcome variable ***average speed*** and 2 predictor variables ***age*** and ***gender*** for the training data (70%) produces ***p-value of 0.0007472*** and ***R- squared of 0.1382187.***

The result suggests that ***age*** and ***gender*** can explain ***13.82%*** of the variability in the ***mean average*** value. It’s hence necessary to consider a lot more factors that can affect the speed, for example weather or traffic.

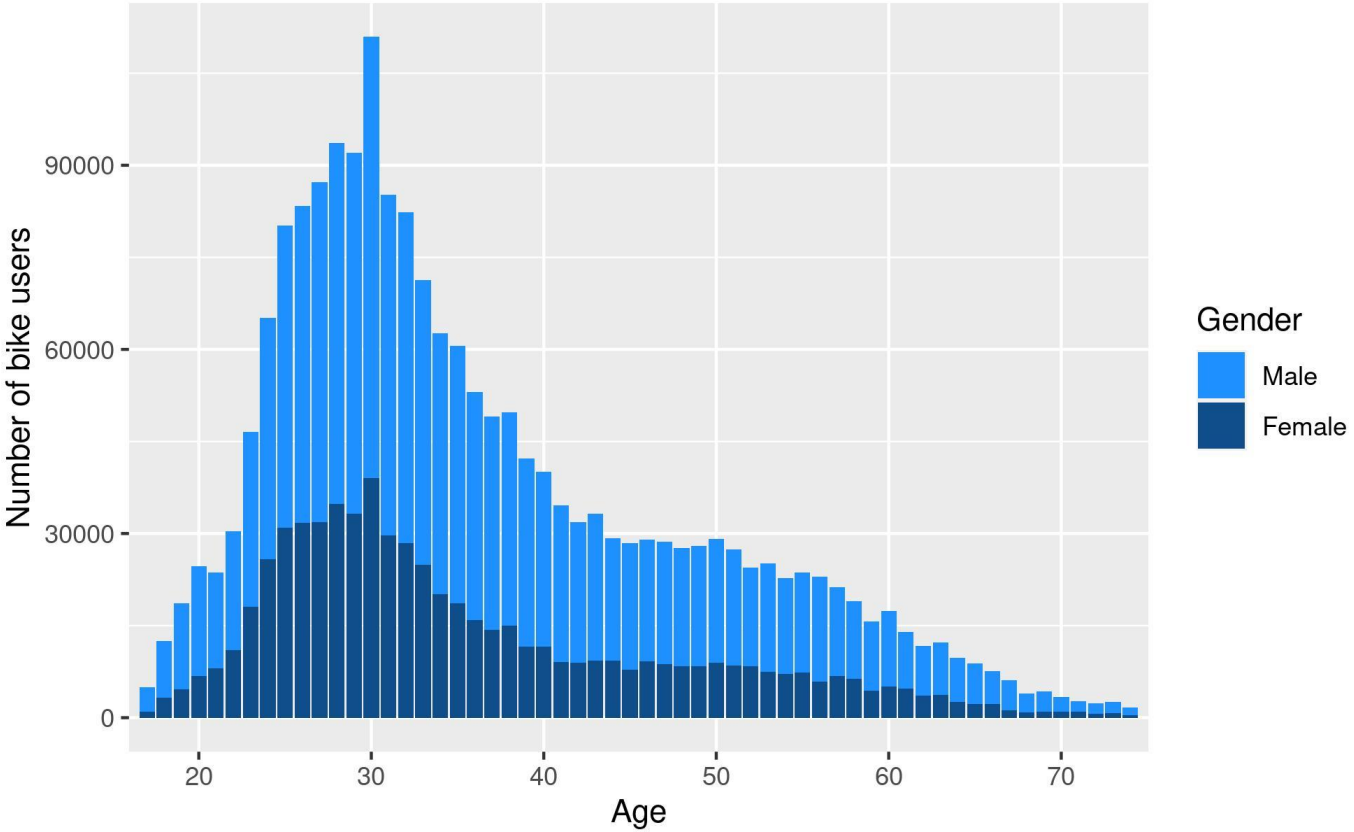
Average trip duration by age and gender



Female riders on average spend more time on their trips compared to their male counterparts. This could be explained by their speed which tends to be slower than men's.

The trip duration of female aged 20 yo especially high. This is quite interesting as when looking at the previous chart we find that the speed of this group is not that low! Perhaps they just love to cycle, hence, cycle for longer time?

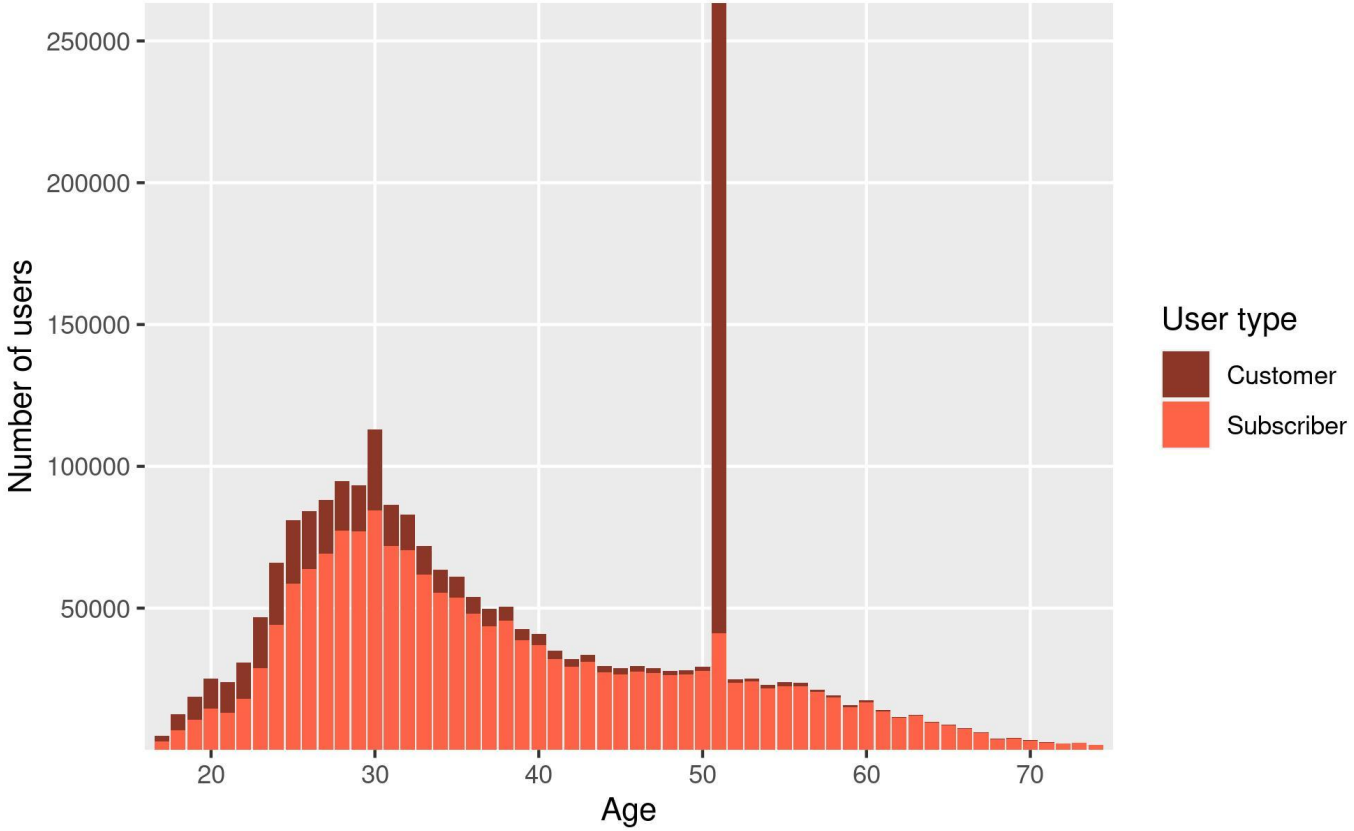
Distribution of Citibike users' gender



At all age groups the number of male users are always higher than female users, especially at the age of around 30 yo.

Unsurprisingly the groups aged 25 yo - 35 yo show the highest number for both male and female users.

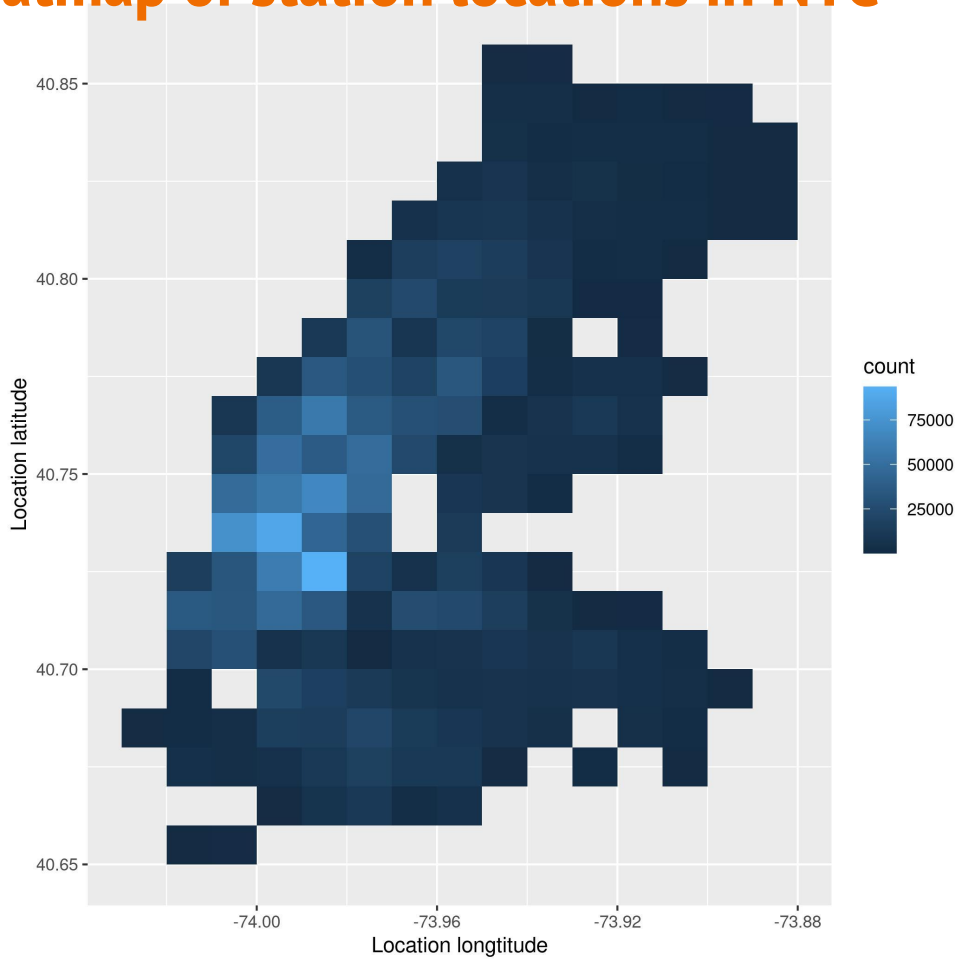
Distribution of Citibike user types by age and gender



The number of subscribers of all ages is much higher than those of customers.

However at the group aged 51 yo, the number of customer is exponentially high. It's likely that there is an error in data that needs to be re-checked, or perhaps there was an event causing the jump.

Heatmap of station locations in NYC



The map illustrates NYC which shapes Manhattan, Brooklyn, and Queens.
The rectangle in Manhattan with no stations is the Central Park.

Manhattan area observed a high number of trips (more than 75000 trips recorded).

Distance between 2 stations

Max distance between 2 stations is 21.21627km which links 14St &7Ave St and W181 St & Riverside Dr

Min distance is 0.03677km links Pershing Square North and Pershing Square South