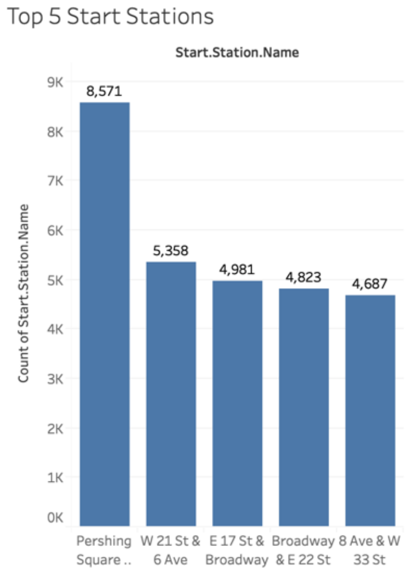
Citi Bike Operation Report – January 2017

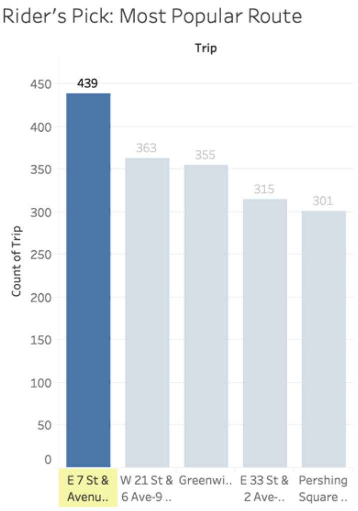
Summary by bike stations

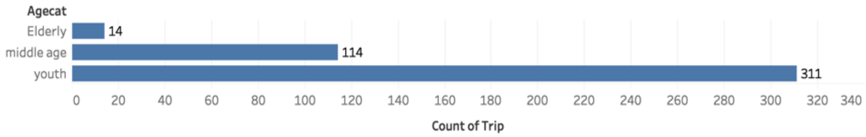
Despite the harsh weather of January, Citi Bike still has a steady ridership from its customer base and other users. Based on the data collected from the kiosks, the most frequently visited bike station is the one at Pershing Square North with 8,571 times of visits, as it's standing right in front of the Grand Central and surrounded by a couple of hotels. As is shown in the bar chart, the station at Pershing Square digests 60% more of the traffic than the second most popular station. The full list of top 5 stations are as follow:

- Pershing Square North
- W 21 St & 6 Ave
- E 17 St & Broadway
- Broadway & E 22 St
- 8 Ave & W 33 St



In terms of future bike allocation, it is critical to put more focus on these stations and others that are in similar districts or with similar surroundings (eg. high density of office buildings, close to tourist sites), to ensure that sufficient bikes will be transported back.

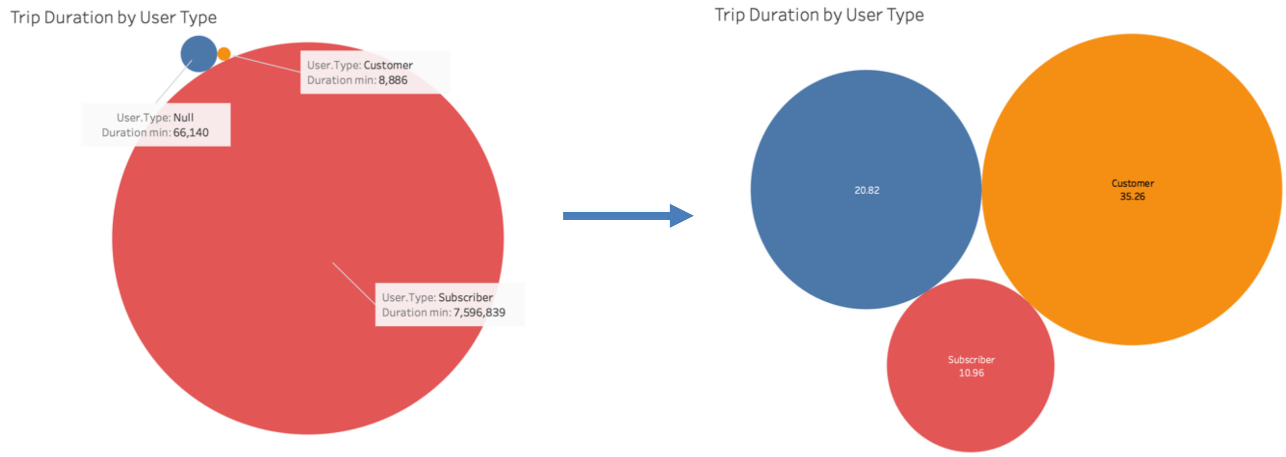




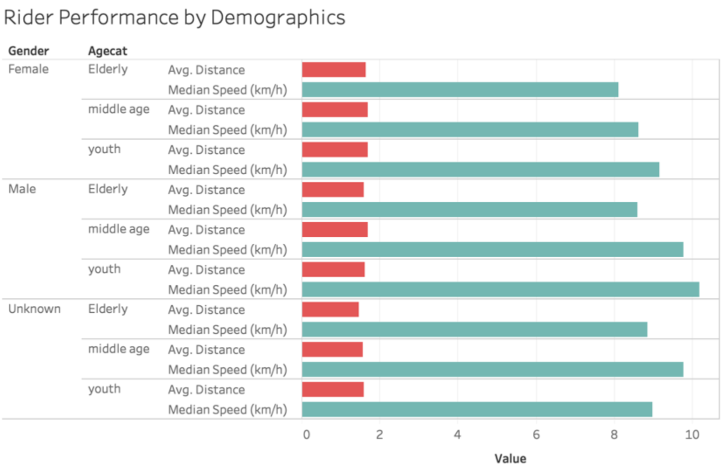
On the other hand, the busiest stations did not necessarily overlap with the stations in the most popular routes. The route starting from E 7 St & Avenue A to Cooper Square & E 7 St was repeated 439 times in January. Within the 439 times, 70% of the trip are taken by the "Youth" group (below 35 years old); besides, 390 of the trips happen during weekdays. This route is probably most utilized by students and commuters, thus enough supply should be ensured along the route especially during weekdays.

User analysis

The following chart characterizes total rider duration in minutes on the bike based on the extent of ridership. Duration data was cleaned up by taking out the extreme values based on 99.95% quantile, which will be elaborated in the appendix. Looking at the total duration, the subscriber group is the biggest driver of total duration, spending 7.6 million minutes on the bike thanks to its large user base; while on average, customers, who are less “attached” to Citi Bike services actually spent 25.26 minutes per person, followed by the single riders, who on average spent 21 minutes riding the shared bikes. In general, the trip duration is pretty short, which partly explained why many people do not subscribe to the program but rather spend $3 on the single ride. Therefore, it is crucial to optimize the offered programs in order to better retain and develop people from the customer and null groups.

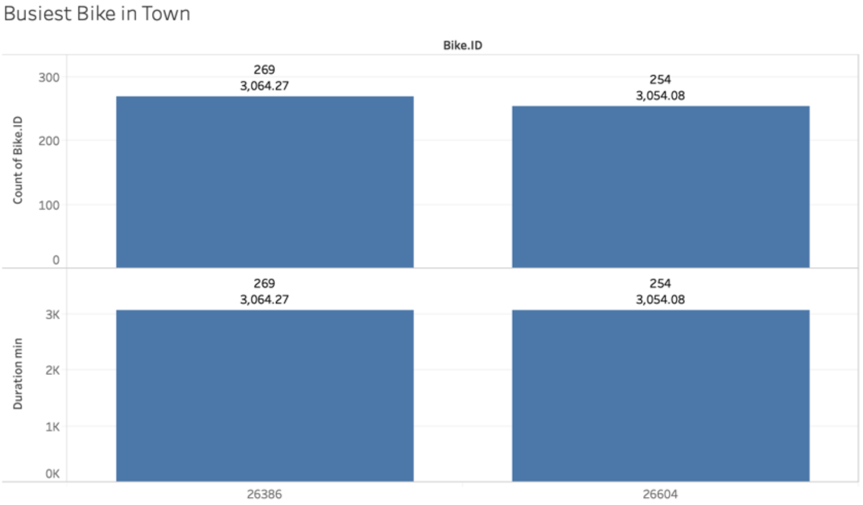


Given the coordinates of start and end stations, we are able to calculate the distance between the two as a proxy of the distance travelled. We used gender and age group to categorize the users. After cleaning up the outlier age records (values beyond 100), we set three age groups where people who are 35 years old or younger are in the youth group, 35~60-year-old would be labeled as middle age group and the rest make up the elderly group. Overall, riders only use the bike for short-distance travels, as the average distance did not exceed 2 kilometers across all groups. Setting aside the unknown group, the pattern of riding speed in Male and Female groups is quite similar – the younger the rider, the faster he/she can ride.



Bike Usage

Based on the dual criteria of usage frequency and usage duration, Bike No. 26386 and No. 26604 popped up on the list; the No.1 busy bike was borrowed 269 times during the month and was ridden for 3064 minutes. In order to improve customer experience, this information provides guidelines for future maintenance with higher efficiency, by focusing on a subset of bikes that are used most.



Predictive Model

In order to produce more reliable predictions, we use random forest as the learning algorithm, which functions efficiently on large dataset and does well in avoiding overfitting based on some features, compared with a single decision tree.

Given the start and destination input, we are able to produce the distance between the two points, which will serve as one of the predictors for the model. Except for distance, 6 other predictors are derived based on rider demographics, trip features and weather conditions to predict trip duration.

<u>Rider demographics</u>

- Gender: contains three levels, male, female and unknown
- Age group: Three age categories – "Youth", "Middle Age" and "Elderly" are derived from the continuous variable "Age". Age is determined by the difference between current year (2017) and birth year information; outliers where Age greater than 100 were eliminated from the dataset.

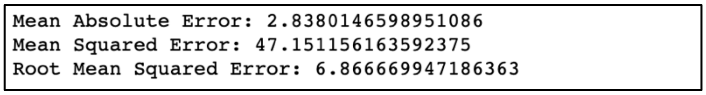
<u>Trip features</u>

- Distance: take the longitude and latitude data to calculate the direct distance between start and end stations, using Haversine formula
- Wday: day of the week, indicator of whether the trip happens during weekday or on weekends
- Peak: dummy indicator that takes the value "Peak" only when Wday has a value from weekdays and the start time captured is during either 8-10 AM or 4-7 PM

<u>Weather conditions</u> (data collected by weather station # 744860, Kennedy International Airport)

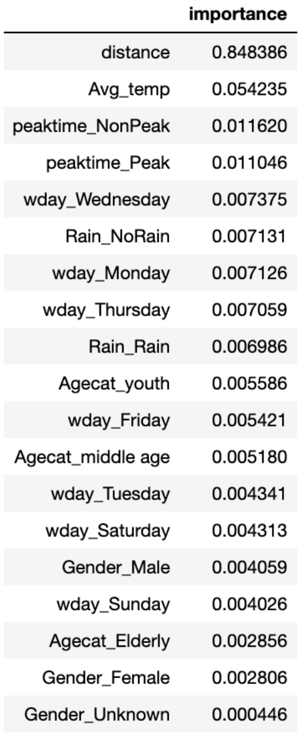
- Daily average temperature (in Celsius degree)
- Rain: dummy indicator of whether it rained on a specific day

The initial model sets the total tree number to 80 trees and the model performance is shown below. In general, the model performs well as the root mean squared error at 6.87, which means the predictions are pretty concentrated around the actuals.



```
Mean Absolute Error: 2.8380146598951086
Mean Squared Error: 47.151156163592375
Root Mean Squared Error: 6.866669947186363
```

As is also presented, the importance of each feature is listed in descending order, in which the distance between stations and average temperature are the two most important features. Therefore, in terms of future implementation, it is critical to connect to government weather data for more accurate duration predictions.



	importance
distance	0.848386
Avg_temp	0.054235
peaktime_NonPeak	0.011620
peaktime_Peak	0.011046
wday_Wednesday	0.007375
Rain_NoRain	0.007131
wday_Monday	0.007126
wday_Thursday	0.007059
Rain_Rain	0.006986
Agecat_youth	0.005586
wday_Friday	0.005421
Agecat_middle age	0.005180
wday_Tuesday	0.004341
wday_Saturday	0.004313
Gender_Male	0.004059
wday_Sunday	0.004026
Agecat_Elderly	0.002856
Gender_Female	0.002806
Gender_Unknown	0.000446

Appendix

1 New / recoded variables

Distance (km)

- distance between start and end station measured in kilometers
- Haversine formula:

$$a = \sin^2(\Delta\text{latitude}/2) + \cos \text{latitude1} \cdot \cos \text{latitude2} \cdot \sin^2(\Delta\text{longitude}/2)$$

$$c = 2 \cdot \text{atan2}(\sqrt{a}, \sqrt{(1-a)})$$

$$d = 6371 \cdot c \text{ (radius of Earth in meters = 6371 km)}$$

Age_category

- Age = 2017 – [Birth Year]
- Youth: people aged below 35
- Middle Age: people aged between 35 and 60
- Elderly: people aged above 60

Peak (assign value “peak” if both of the two requirements are met)

- The trip happens during weekday
- The start time either falls in 8-10 AM or 4-7PM

2 Outlier deletion

Age:

Deleted values above 100

Duration_minute:

Based on the 99.95% and 0.05% quantile, deleted all values above 289.98 and below 1.084 minutes

3 External data source

The data is downloaded from the following website https://en.tutiempo.net/climate/ws-744860.html and is collected from weather station # 744860, Kennedy International Airport