Fan Zhang

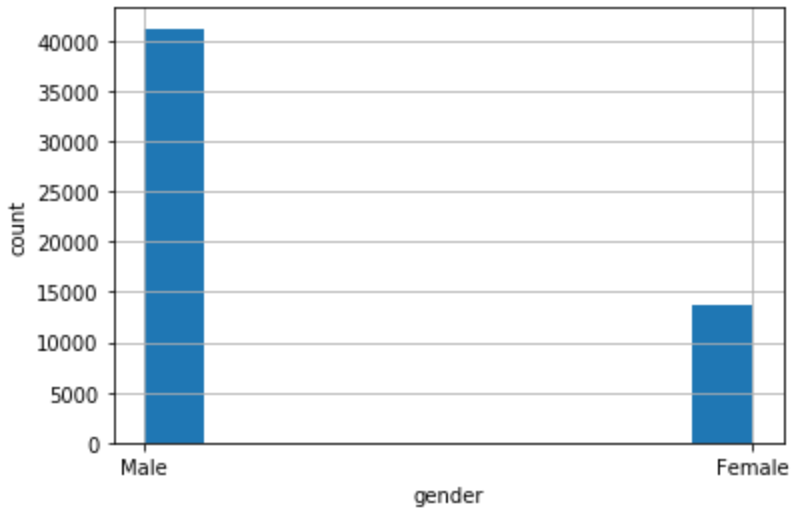
3031922700

Data102 Project Part I

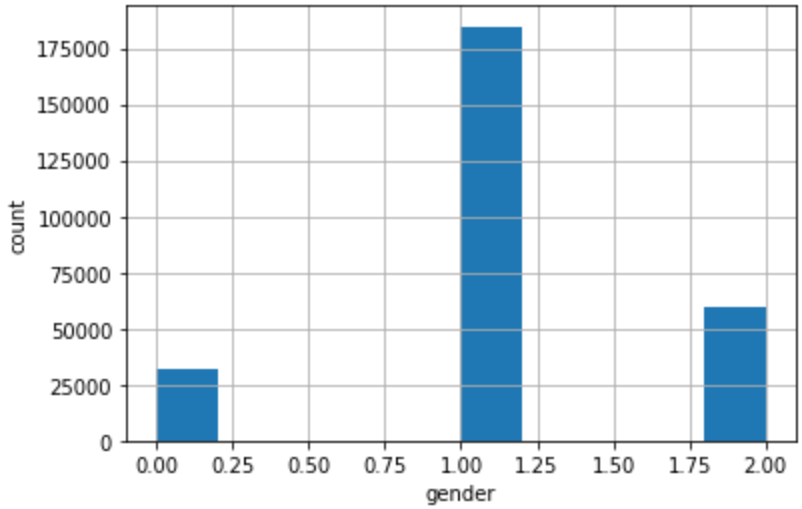
1 Preliminary Data Analysis

1.1 Demographic Information

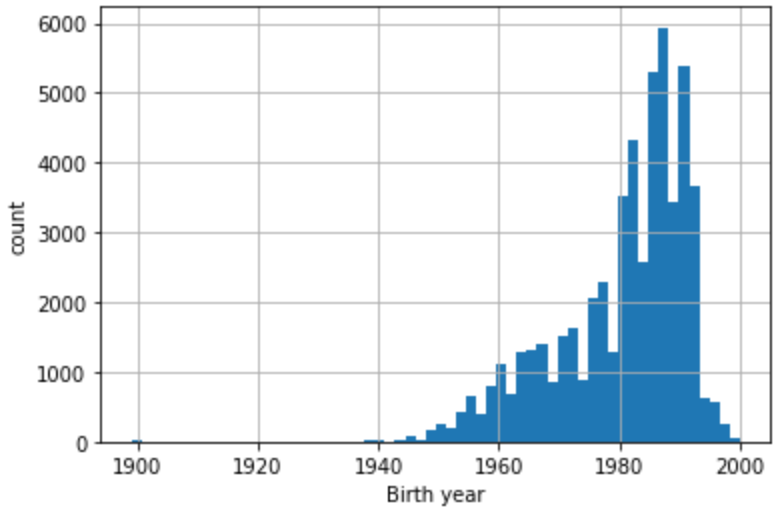
1. Chicago data set gender distribution:



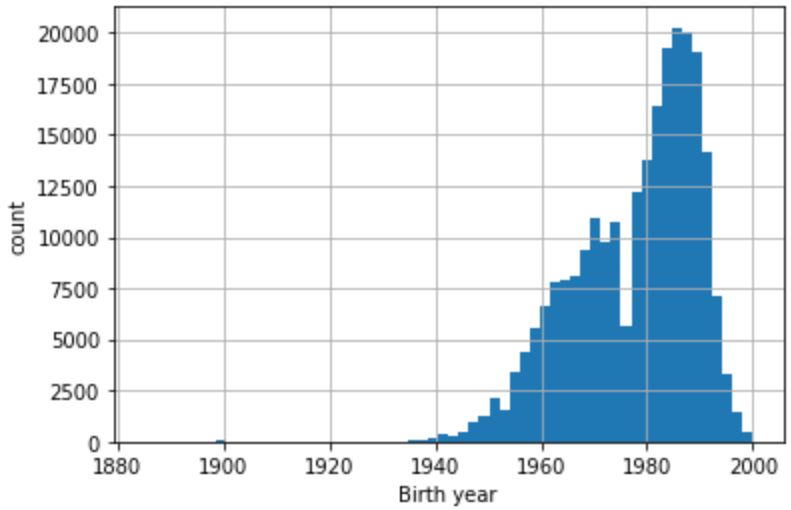
2. New York data set gender distribution:



3. From chicago data we can see the ratio between female and male is roughly 1:3. Therefore, mapping to ny dataset, I guess that male is marked as 1, female is 2 and uspecified is 0.
4. Distribution of the birth year of bike renters in Chicago:



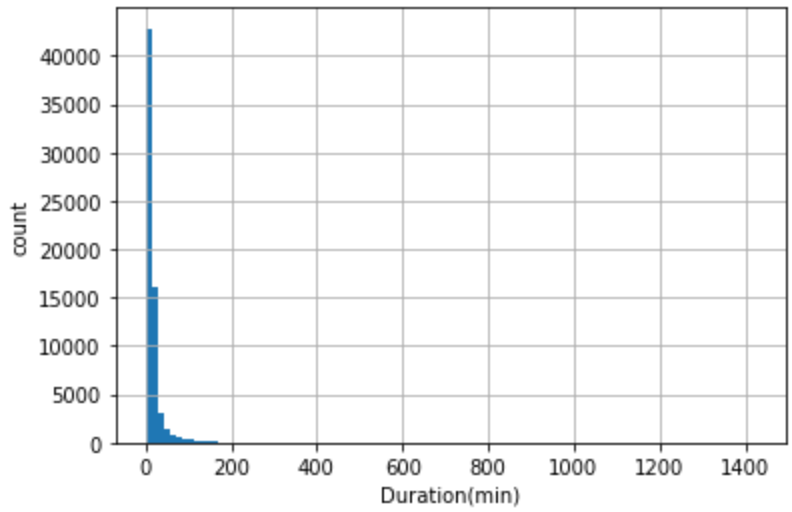
Distribution of the birth year of bike renters in New York:



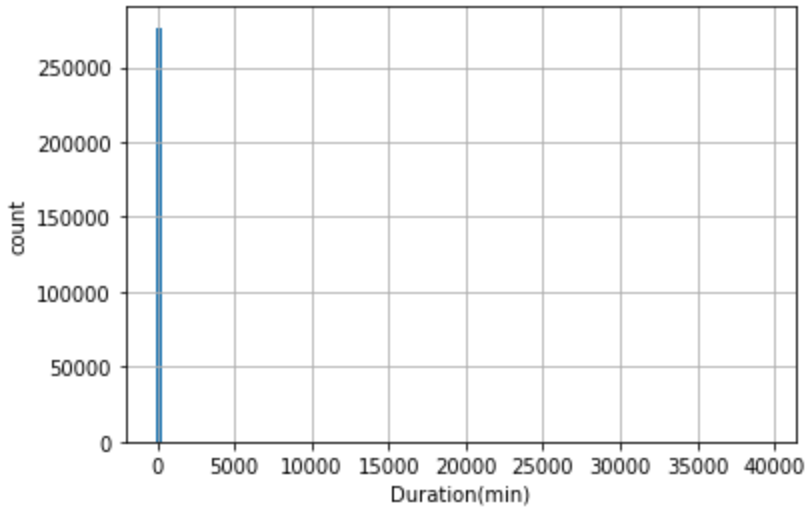
5. The result plot has outliers. The oldest renter was born in 1880 and the youngest renters are only two-year-old. There must be some mistakes in recording. Also, from the plot we can see that most renters are young people who were born after 1980. This is reasonable because intuitively, they are the ones that ride the most in big cities.

1.2 Rental Times

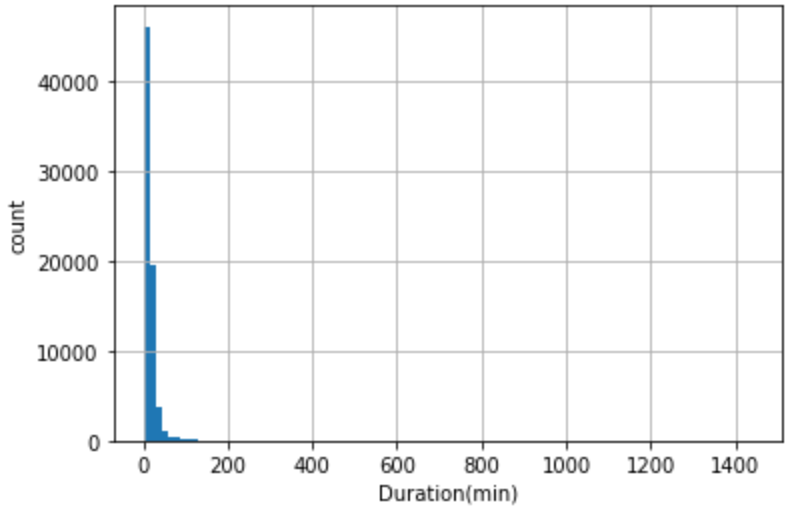
1. DC trip duration distribution:



NY trip duration distribution:

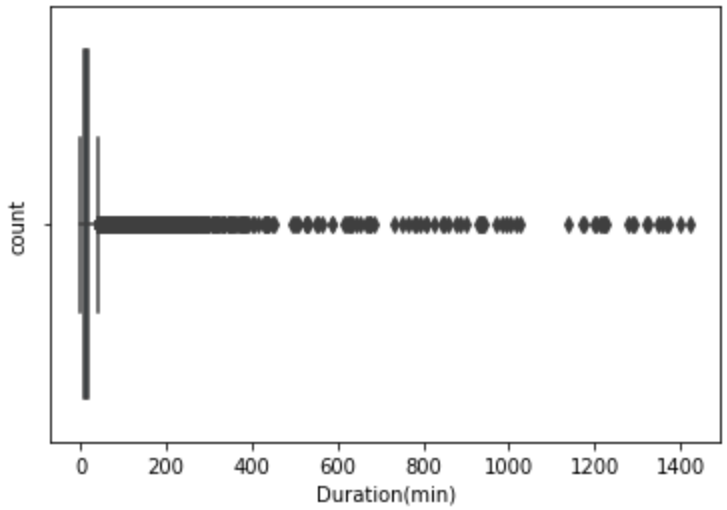


Chicago trip duration distribution:

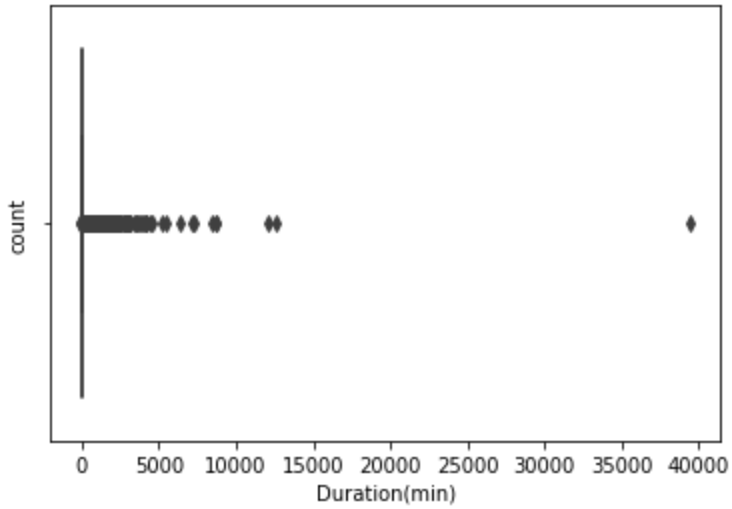


2. They are not useful because there are too many outliers. I use boxplot to plot them again.

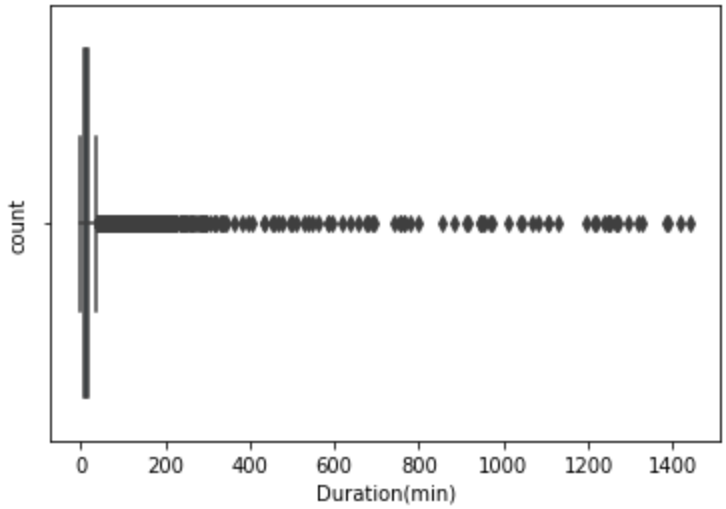
   DC trip duration distribution:



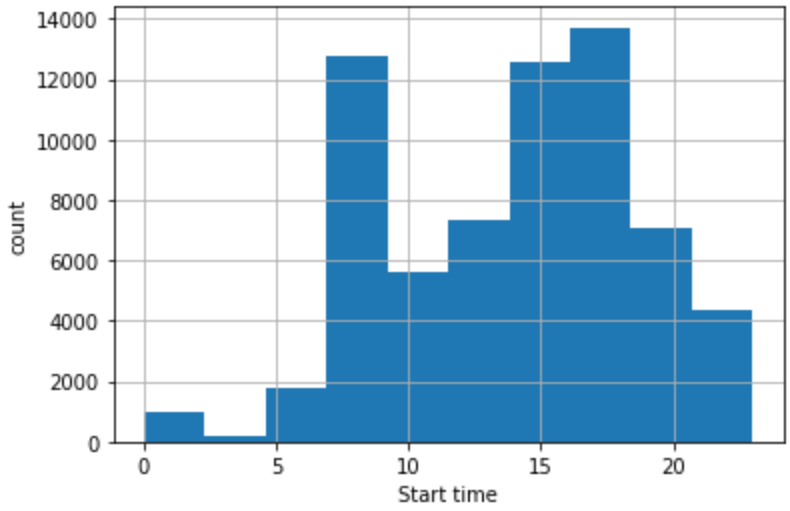
NY trip duration distribution:



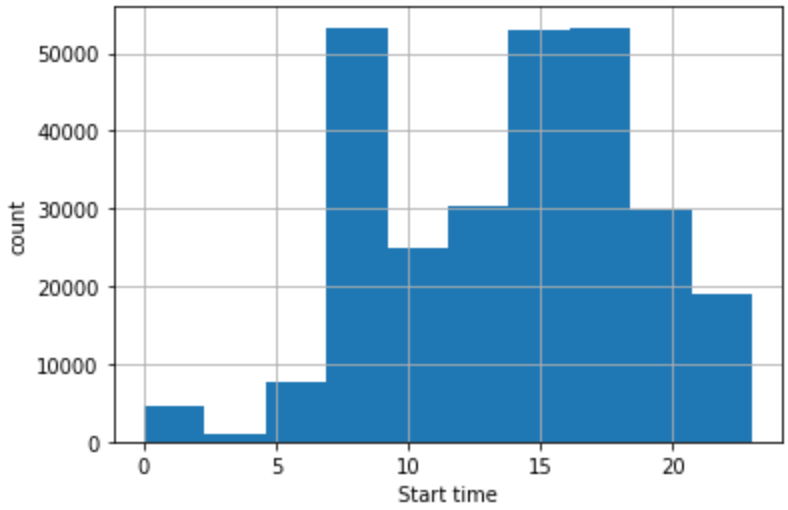
Chicago trip duration distribution:



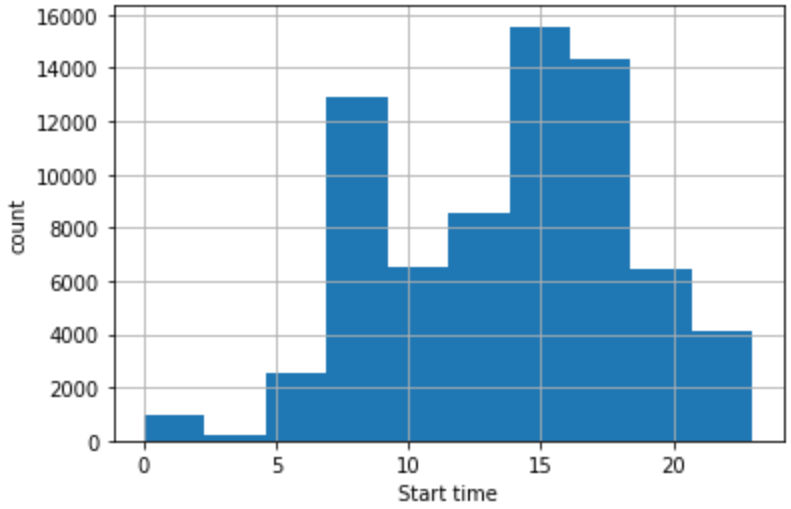
3. Start time distribution in DC:



Start time distribution in NY:



Start time distribution in Chicago:

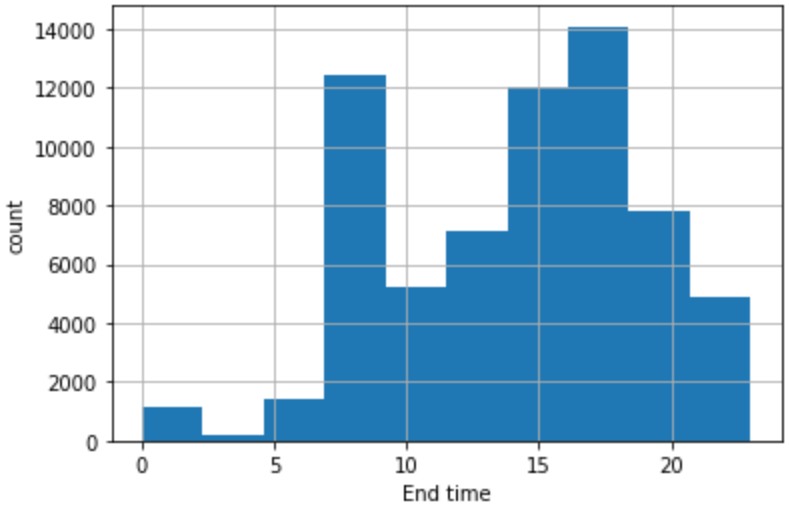


4. Most rentals occurs during peak hours when people are commuting between home and work. There are few rentals in late night. The pattern is consistent in all three cities. The results fit my intuition.

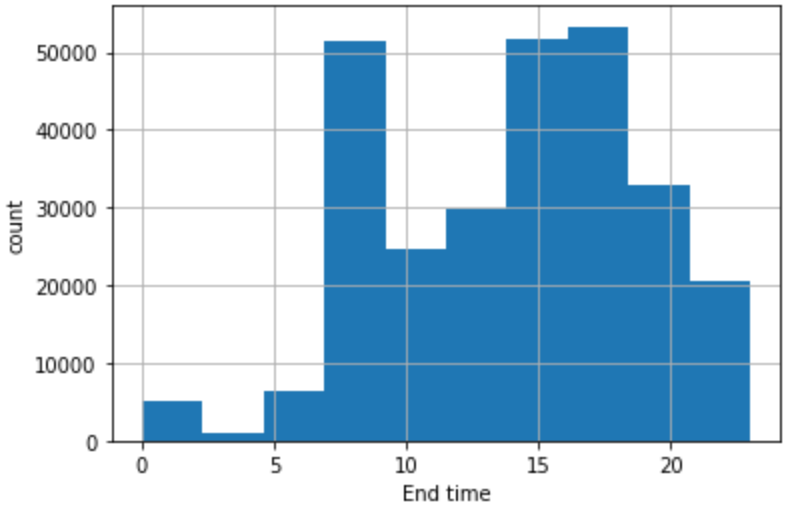
1.3 Further Exploration

1. I choose to visualize the end time in hour, user type distribution, and average trip duration by user type in all three cities.

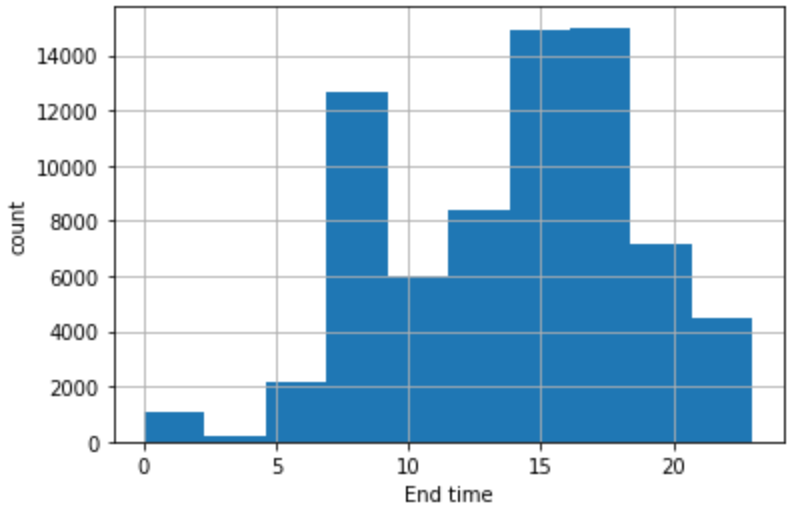
   End time distribution in DC:



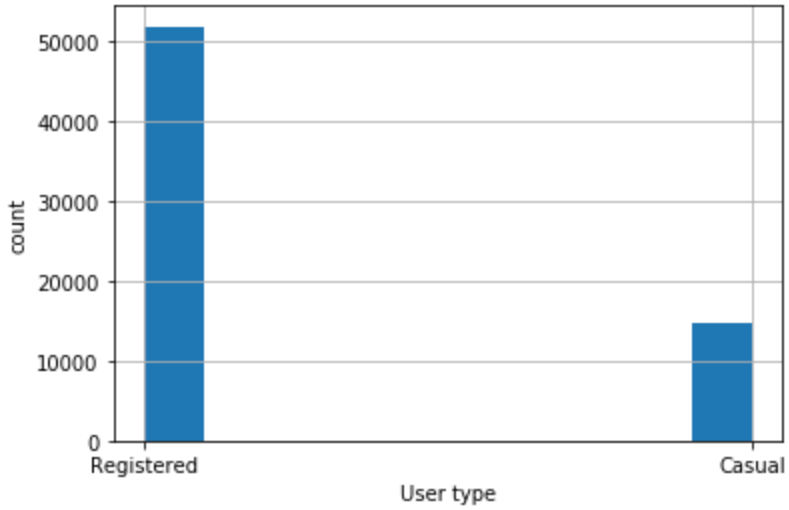
   End time distribution in NY:



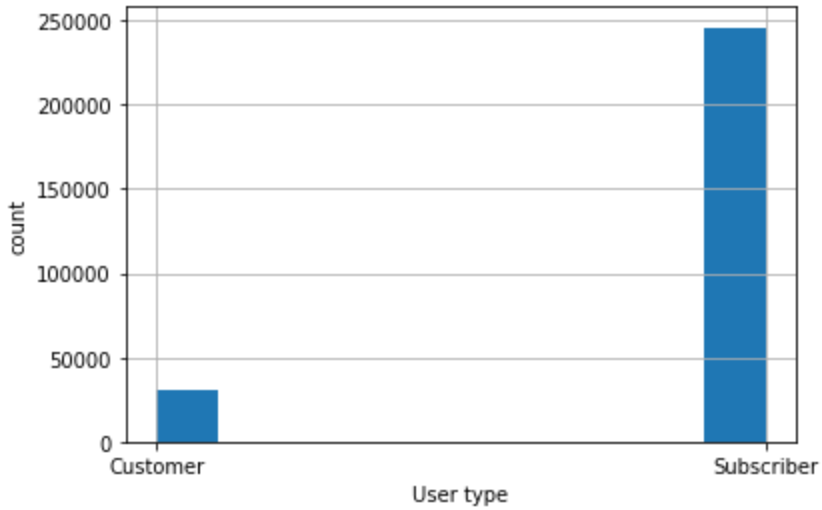
   End time distribution in Chicago:



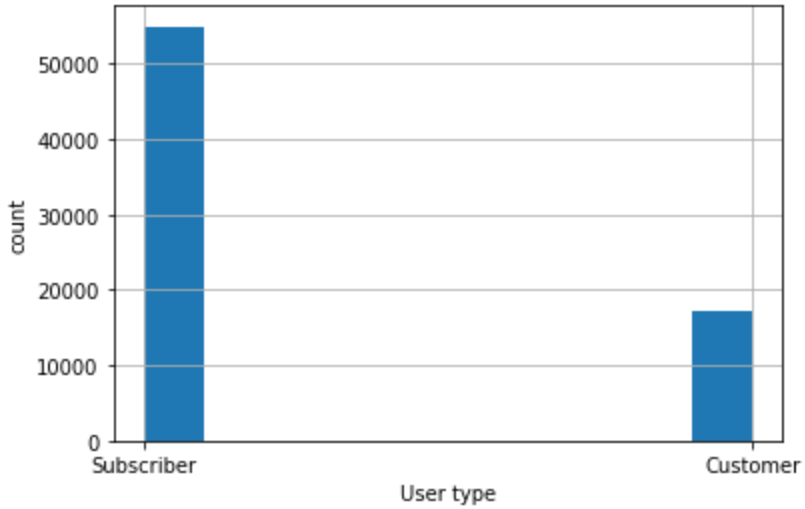
User type distribution in DC:



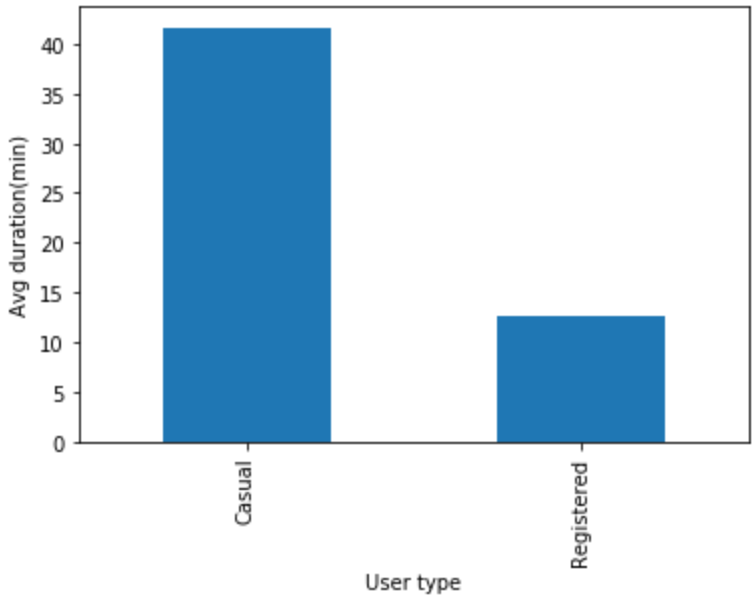
User type distribution in NY:



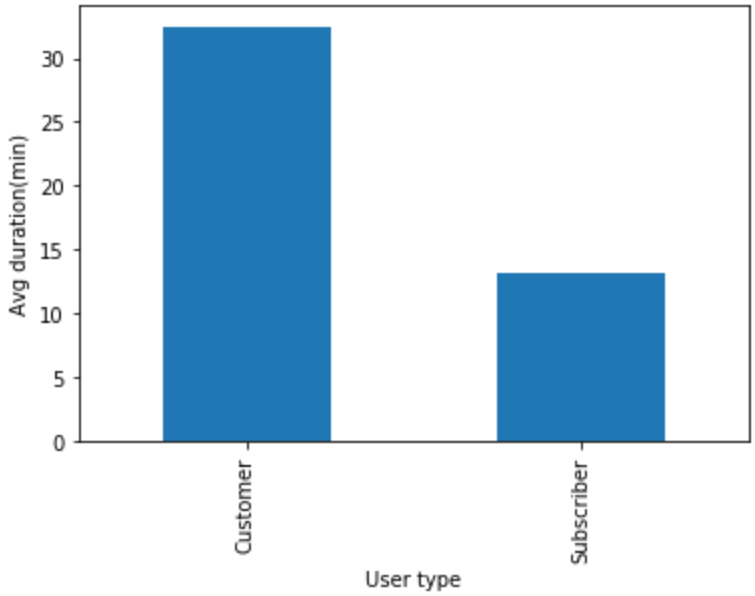
User type distribution in Chicago:



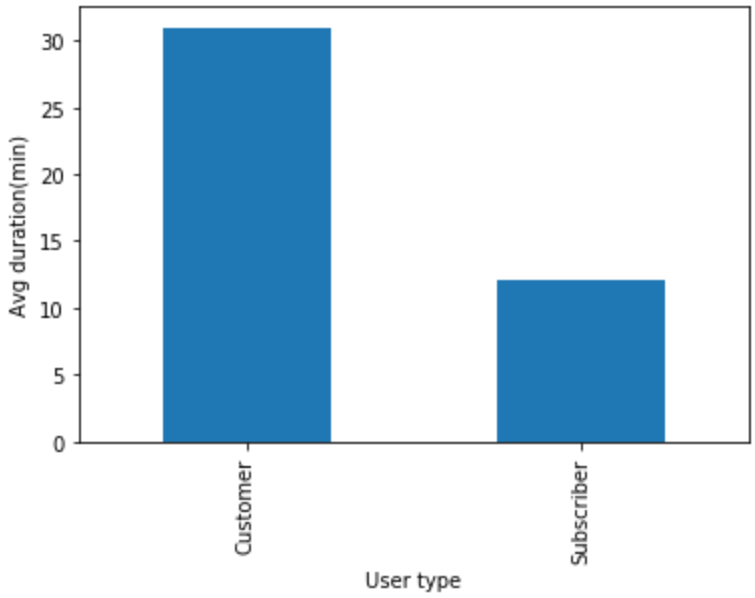
Trip duration in minutes average in DC by user type:



Trip duration in minutes average in NY by user type:



Trip duration average in minutes in Chicago by user type:



2. From the above analysis, I found that similar to start time, stop time is also correspond to the commuting hours. Most renters are subscribers(members) across all three cities. Subscribers has shorter rental durations compared to causal renters.

3. For the plots above I found that average trip duration is correlated with the user type. The average rental durations for subscribers are ⅓ of the causal renters.

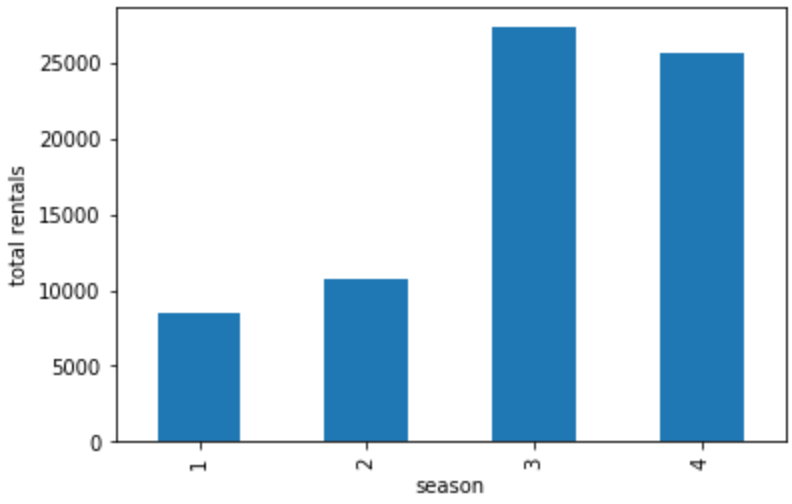
4. Hypothesis: the subscribers are mostly people who use rental bikes for commuting purposes.

   How to test: set null hypothesis as member type doesn't correlated with the start and end times of the working hours. Also, if the null hypothesis hold, every trip of each subscriber would vary in length. Calculate the test score to see if the p_value is significant for null hypothesis.

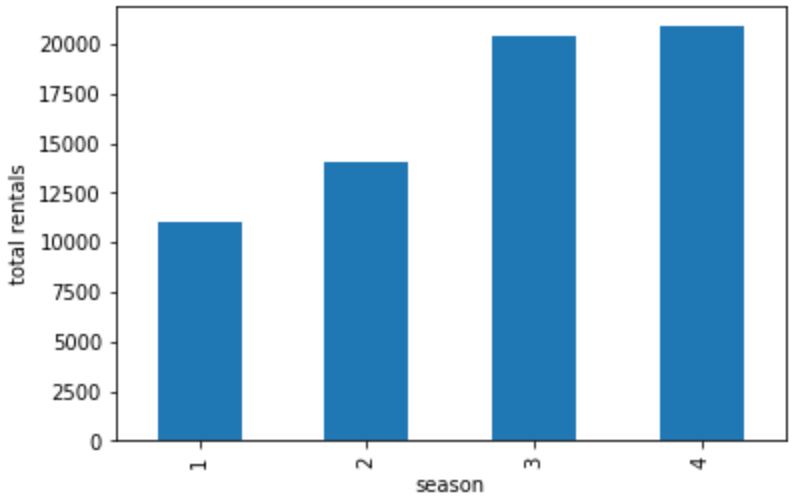
1.4 Optional*: Creating a new dataset

ny_daily: https://drive.google.com/open?id=1ULHff6OHPquvdLaxa7g7BYlDCju4DpGU
chicago_daily:
https://drive.google.com/file/d/1-0zcFwSsoZLy-SoU7WnbqNR2KgRVhxbU/view?usp=sharing
dc_daily:
https://colab.research.google.com/drive/1NUxVLdsdOWowtutySNGcrsflRyHeqJpl

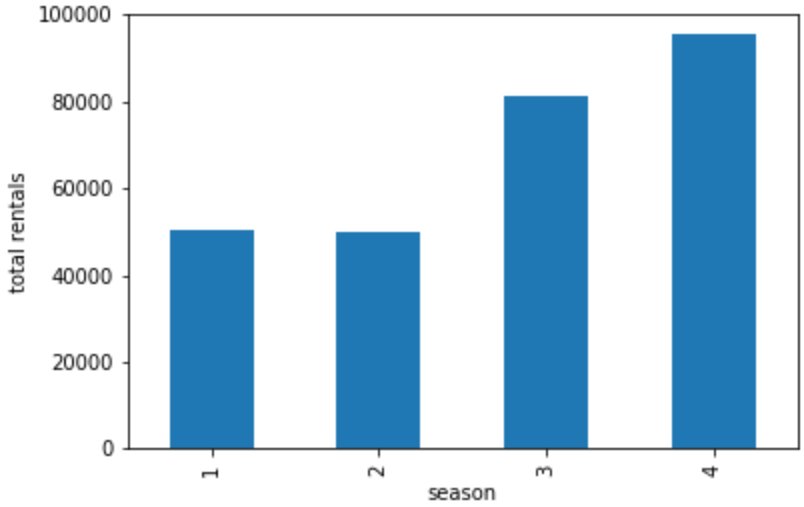
Creation procedure: I first dropped the irrelevant columns of the data set and changed start time into standard format using pd.to_datetime. After that, I split dates from the start time. Then I used one hot encoding to separate the two user types. I used groupby date and found the sum of both user types on any given date. Then, I call pd.DatetimeIndex to find month and weekday. I also import USFederalHolidayCalendar from pandas.tseries.holiday to find holidays. After that, I can filter out working days by setting conditions that working days are not weekends or holidays. Also, I used condition on month to find seasons. Finally, I added instance and count the total number of bike rentals on a given day.
After comparison, I found that the distribution of number of rentals in each season varies in different cities.
In Chicago, most rentals are in summer. In winter, the total amount of rentals are really small.



In DC, however, the most rentals occurred in Fall. And the difference between the least and most are not as big as in Chicago.



In NY, same as in DC, most rentals are in Fall. People rent the bike least in spring. And the difference between the most popular and least popular season are even smaller.

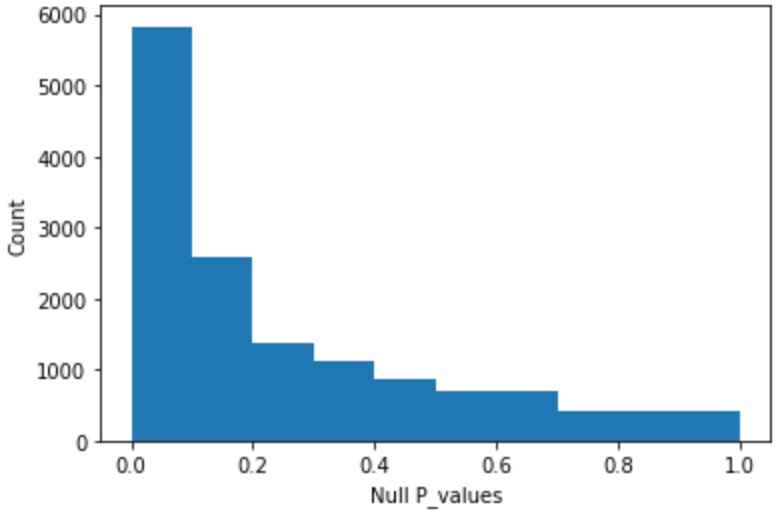


These are possibly due to weather. In Chicago, maybe winter is too cold to ride bikes. But for NY and DC, they have similar weather for each season so their patterns are the same.

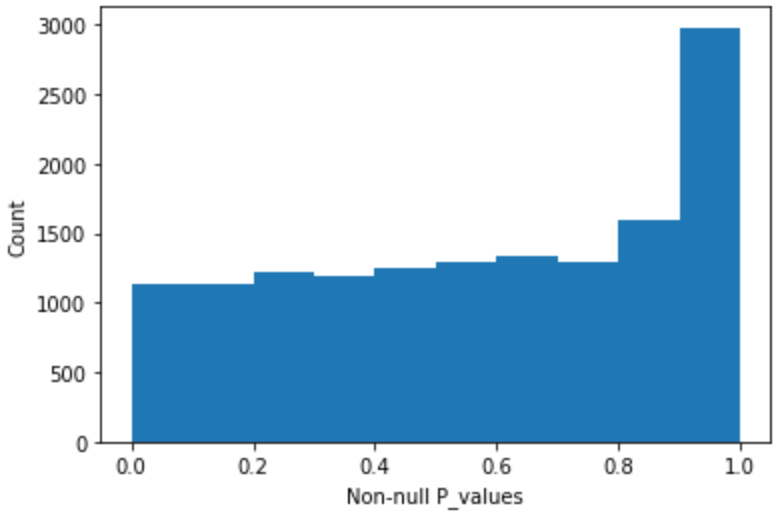
2 Hypothesis Testing

1. I chose Chicago data set to do hypothesis testing. I split the data using np.split into s1, s2 and s3. and called sklearn.logisticregression()to fit s1's user type by s1's trip duration, start time and stop time. After getting the learned parameters, I plugged them with s2 and s3's data points into the sigmoid function given to get si(2) and sj(3), and use the condition in question 2.2 to calculate the p-values.

2. Null p-values:



Non-null p-values:

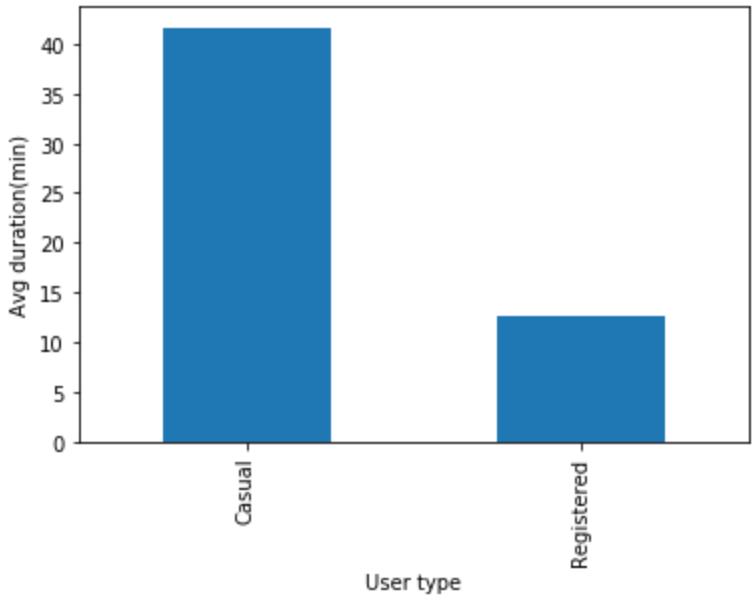


The two histograms have reversed shape. For null p values, there is a peak around 0, which shows alternative hypotheses. The flat distribution is all null p-values, which are uniformly distributed between 0 and 1. For non null p values, the peak around 1 is the null hypothesis and the rest are uniformly distributed.

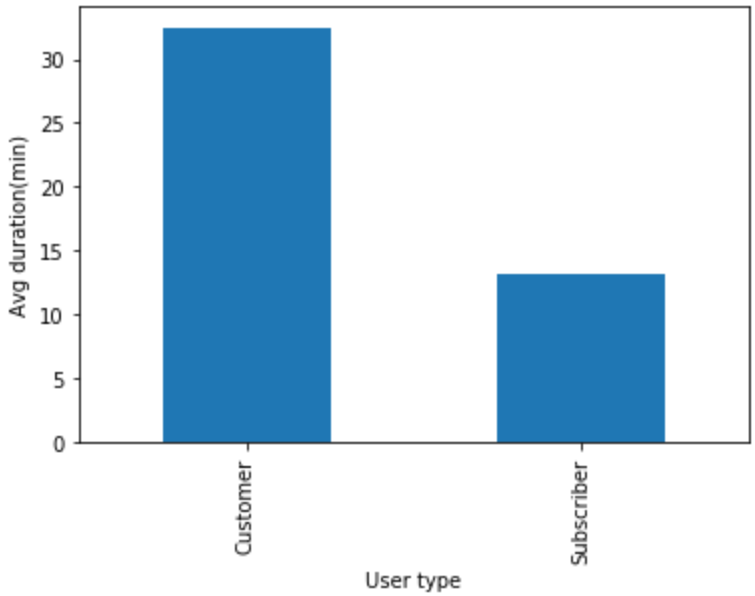
3. The average FDP is 0.311 and the average sensitivity is 0.656. The average FDP is above 0.2. This is because our si(2)s and sj(2)s are estimated from s1, not the true values. Therefore, FDR control is not guaranteed.

3 Gaussian Mixture Models of Trip Durations

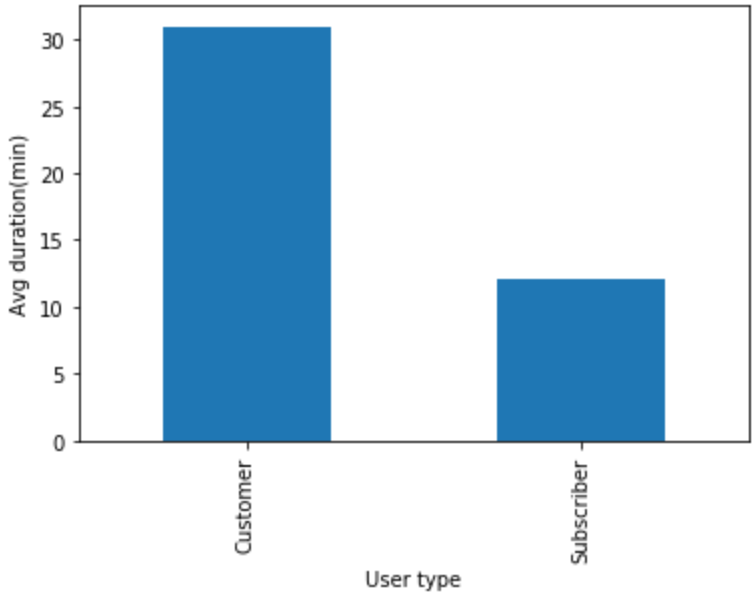
1. Trip duration in minutes average in DC by user type:



Trip duration in minutes average in NY by user type:

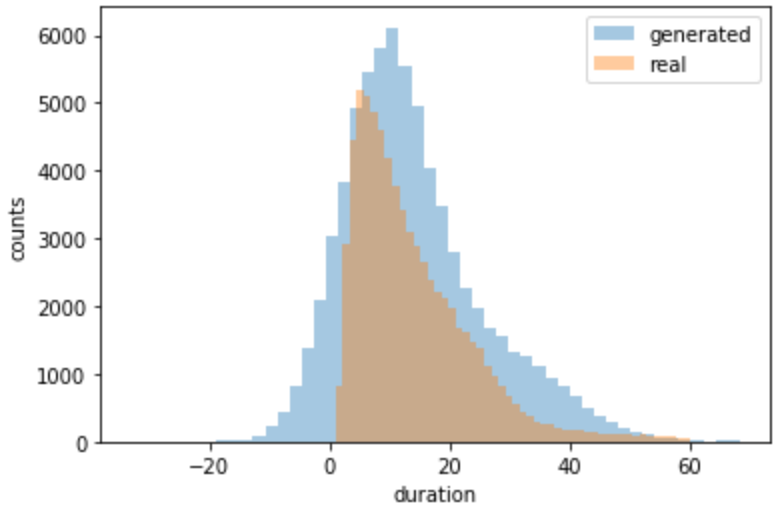


Trip duration average in minutes in Chicago by user type:



From above, we can see customer has an average trip duration of 35 minutes, whereas for subscribers, it is around 12 minutes across all three cities. The average rental durations for subscribers are ⅓ of the causal renters.

2. My initialization for EM algorithm is pi_0=0.2, mu_0=30,pi_1=0.8,mu_1=10, num_steps=100. The code I used for EM is from the lab.



My result doesn't change drastically depending on initialization. This is weird because I thought EM algorithm depends heavily on initialization. Once steps exceed 10, the predicted values converge to the same value. The means of fitted gaussian are 25.83 and 8.82, variances are 11.79 and 7.25.

3. Given the output of the E-M Algorithm, Normal(8.82,7.25^2) captures the behavior of the subscribed customers. Posterior probability that the customer is from this distribution:

	usertype	tripdurationm	post
0	Subscriber	15.433333	9.468857e-01
1	Subscriber	3.300000	1.000000e+00
2	Subscriber	2.066667	1.000000e+00
3	Subscriber	19.683333	5.364368e-03
4	Subscriber	10.933333	9.999700e-01
...	...	...	...
72126	Subscriber	3.883333	1.000000e+00
72127	Subscriber	7.866667	9.999997e-01
72128	Customer	17.816667	8.242397e-01
72129	Subscriber	24.866667	7.513330e-08
72130	Subscriber	10.400000	9.999868e-01

70619 rows × 3 columns

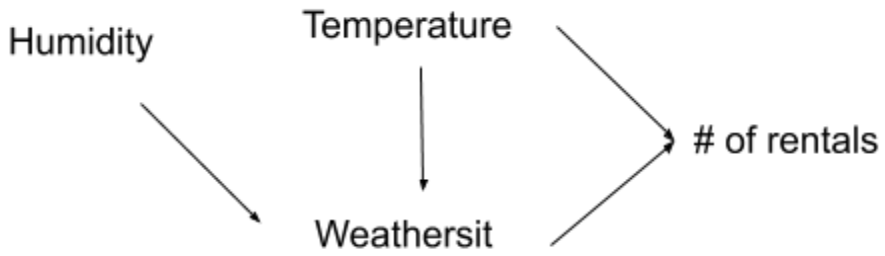
4. The error is around 37% of this classifier given the true user types in the chicago.csv dataset.

5. The error is around 30.9% for New York data set and 29.1% for DC data set. The performance is better compared to the previous part. All trip durations are measured in minutes.

4 Causality and Experiment Design.

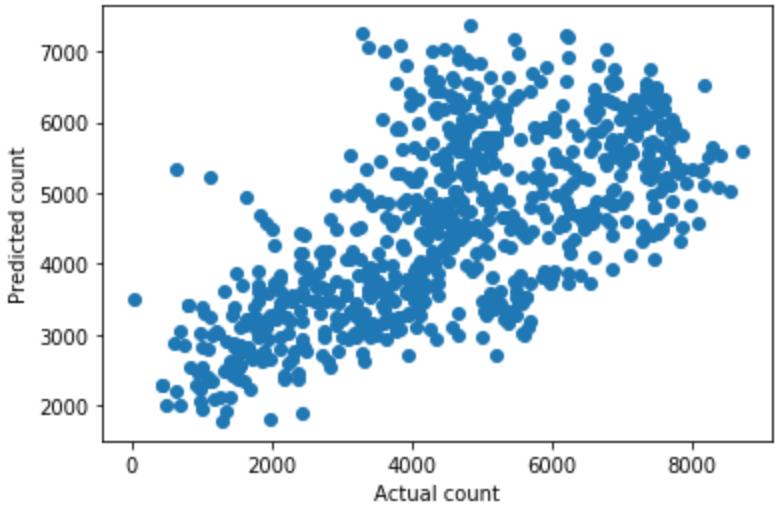
4.1.1 The causal model.

1.

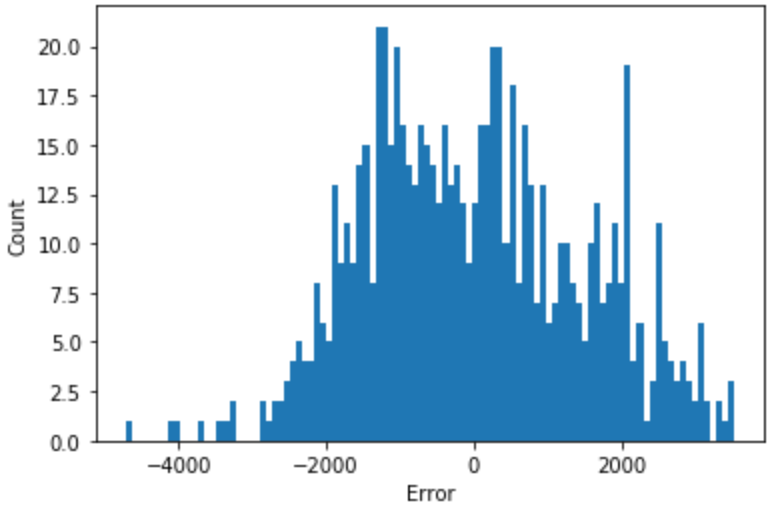


2. Assumptions:
    1. Models (equations) should be correctly identified.
    2. The error variance of all the variables should be equal.
    3. Error terms should be normally distributed.
    4. It is assumed that the outlier(s) is removed from the data.
    5. Observations should be independents of each other.

    For 1 and 5, we assume they are satisfied. For 4, I plot the data and there is no outliers. For 2, I plot the error terms and it seems to be heteroskedasticity. 2 is not satisfied.



For 3, I plot the error terms and it seems to be normally distributed.



4.1.2 2 Stage Least Squares

1. In ordinary least squares (OLS), the treatment is often correlated with the noise; this is often due to the effect of some variables which are important for the generation of Y, but were not measured. One way to get around this issue is by using instrumental variables (IVs). A valid instrument Z is a variable which is independent error term, and affects Y only through X*. Then, one way to estimate alpha is to first "guess" X* from {Z,X}, noted as X*_hat and then regress Y onto { X*_hat , X}. If both the initial "guess" X*_hat and Y regressed onto the {X*_hat, X} are obtained using ordinary least squares, this procedure is known as two-stage least squares (2SLS). In our case, we want to estimate the causal effect of weather situation on number of bike rentals. Correlation between the weather situation and bike rentals does not necessarily imply that weather affects bike rentals, because other variables, such as temperature, might relate to both of them. However, we can estimate the causal effect of weather situation on bike rentals by using instrumental variables. Suppose that weather situation is sufficiently correlated with humidity. Then, we can use humidity as an instrumental variable, as it is fairly reasonable to assume that it doesn't directly affect bike rentals but only through weather situations.In the first stage, we "predict" weather situation from temperature and humidity. Then, in the second stage, we regress number of bike rentals onto temperature and the predicted weather situation.

2. For number of total bike rentals, the resulting alpha of weather situation is -1507. Weather situation negatively affects bike rentals. Bad weather situation would reduce the number of rentals.

3. For number of casual bike rentals, the resulting alpha is -303. For registered, the resulting alpha is -753.

4.1.3 Discussion

1. Using two stage least square method and humidity as an instrumental variable, we "predict" weather situation from temperature and humidity in the first stage. Then, in the second stage, we regress number of bike rentals onto temperature and the predicted weather situation. we found that weather situation has a negative relationship with number of bike rentals.

2. For number of casual bike rentals, the resulting alpha is -303. For registered, the resulting alpha is -753. With 1 unit increases in weather situation, bike rentals amount decreases 303 for casual bike rentals, 753 for registered bike rentals. Weather situation negatively affects more on registered user behavior. This difference could be that registered bikers need bikes for routine usage. If the weather is bad, they have other forms of transportation as backup plans. But for casual users, they rent bikes for entertainment purposes, so their behaviors are not related to weather situation that much.

3. In our model, we omit factors like if the date is a holiday or weekend, or the price of the rental. Those factors also have effects on number of rentals. For the arrows, temperature and humidity could be related. “As air temperature increases, air can hold more water molecules, and its relative humidity decreases. When temperatures drop, relative humidity increases. High relative humidity of the air occurs when the air temperature approaches the dew point value.” -- https://sciencing.com/temperature-ampamp-humidity-related-7245642.html

4. If I had the opportunity to design my own study, I would have more degree measures on weather situation Instead of just 1 and 2. I would keep records on different weather situation like rain, snow, thunder, etc. This would make the resulting alpha smaller and becomes more informative as we can use it to find incremental changes and what kind of weather affects the bike rental amount the most.