Citibike Bikeshare usage demand forecasting

https://www.citibikenyc.com/

A social experiment for predicting future demand of Citibike shared bicycle system in New York city

Sachid Deshmukh [a] | Ann Liu-Ferrara [a] | Ahmed Sajjad [a]

a M.S. in Data Science student CUNY SPS

Highlights

- Paper focuses on analyzing Citibike data and identifying important patterns to support business viability and overall profitability of Citybike bikeshare system [c]
- Paper proposes Machine Learning and Timeseries based forecasting methodologies for predicting future usage demand of Citibike bikeshare system in New York city [a,b]
- Paper proposes various feature engineering techniques for predicting future usage demand of Citibike bikeshare system in New York city
- Paper evaluates feature importance and proposes key features which contributes towards predicting future usage demand of Citibike bikeshare system in New York city

a https://www.citibikenyc.com/ b Hyndman & Athanasopoulos. https://www.otexts.org/fpp2/ c https://www.citibikenyc.com/system-data

Article Info

Keywords
Bikeshare, Weather, Cycling, Citibike, New York City

Abstract

Bicycling is an activity which yields many benefits: Riders improve their health through exercise, while traffic congestion is reduced if riders move out of cars, with a corresponding reduction in pollution from carbon emissions [5]. In recent years, Bike Sharing has become popular

Article History
Created:
02/16/2020

Revision-1
04/03/2020

in a growing list of cities around the world. The NYC "Citibike" bicycle sharing scheme went live (in midtown and downtown Manhattan) in 2013, and has been expanding ever since, both as measured by daily ridership as well as the expanding geographic footprint incorporating a growing number of "docking stations" [9-10] as the system welcomes riders in Brooklyn, Queens, and northern parts of Manhattan which were not previously served. One problem that many bikeshare systems face is money. An increase in the number of riders who want to use the system necessitates that more bikes be purchased and put into service to accommodate them. Heavy ridership induces wear on the bikes, requiring for more frequent repairs. However, an increase in the number of trips does not necessarily translate to an increase in revenue because riders who are clever can avoid paying surcharges by keeping the length of each trip below a specified limit (either 30 or 45 minutes, depending on user category.) We seek to examine Citibike ridership data [2-3], joined with daily NYC weather data [4], to study the impact of weather on shared bike usage and generate a predictive model which can estimate the number of trips that would be taken on each day [6-8]. The goal is to estimate future demand which would enable the system operator to make expansion plans. Our finding is that ridership exhibits strong seasonality, with correlation to weather-related variables such as daily temperature and precipitation [6-8]. Additionally, ridership is segmented by user type (annual subscribers use the system much more heavily than casual users), gender (there are many more male users than female) and age (a large number of users are clustered in their late 30s). [11-12]

1. Introduction

Since 2013 a shared bicycle system known as Citibike has been available in New York City [1]. The benefits to having such a system include reducing New Yorkers' dependence on automobiles and encouraging public health through the exercise attained by cycling [5]. Additionally, users who would otherwise spend money on public transit may find bicycling more economical – so long as they are aware of Citibike' s pricing constraints. There are currently about 12,000 shared bikes which users can rent from about 750 docking stations located in Manhattan and in western portions of Brooklyn and Queens. A rider can pick up a bike at one station and return it at a different station. The system has been expanding each year, with increases in the number of bicycles available and expansion of the geographic footprint of docking stations. For planning purposes, the system operator needs to project future ridership to make good investments. The available usage data provides a wealth of information which can be mined to seek trends in usage 2-3]. With such intelligence, the company would be better positioned to determine what actions might optimize its revenue stream.

- Because of weather, ridership is expected to be lower during the winter months, and on foul-weather days during the rest of the year, compared to a warm and sunny

summer day. Using the weather data, we can seek to model the relationship between bicycle ridership and fair/foul or hot/cold weather [6-8].

- What are the differences in rental patterns between annual members (presumably, residents) vs. casual users (presumably, tourists?) [12]
- Is there any significant relationship between the age and/or gender of the bicycle renter vs. the rental patterns? [12]
- What is the seasonal component of bikeshare pattern exhibited by consumers (weekday vs weekend) [11]

2. Data Collection

We are leveraging two major sources of dataset for this scientific experiment

1. **Citibike trip dataset [2-3]**
   Citibike makes a vast amount of data available regarding system usage as well as sales of memberships and short-term passes. For each month since the system's inception, there is a file containing details of (almost) every trip. (Certain "trips" are omitted from the dataset. For example, if a user checks out a bike from a dock but then returns it within one minute, the system drops such a "trip" from the listing, as such "trips" are not interesting.) There are currently 77 monthly data files for the New York City bikeshare system, spanning July 2013 through November 2019. Each file contains a line for every trip. The number of trips per month varies from as few as 200,000 during winter months in the system's early days to more than 2 million trips this past summer. The total number of entries was more than 90 million, resulting in 17GB of data. Because of the computational limitations which this presented, we created samples of 1/1000 and 1/100 of the data. The samples were created deterministically, by subsetting the files on each 1000th (or, 100th) row.

2. **Central Park daily weather data [4]**
   Also we obtained historical weather information for 2013-2019 from the NCDC (National Climatic Data Center) by submitting an online request to https://www.ncdc.noaa.gov/cdo-web/search. Although the weather may vary slightly within New York City, we opted to use just the data associated with the Central Park observations as proxy for the entire city's weather. We believe that the above data provides a reasonable representation of the target population (all Citibike rides) and the citywide weather.

3. Methodology

For Citibike bikeshare usage demand forecasting, we are leveraging Citibike trip dataset [2-3] along with Central Park daily weather data [4]. Goal of the project is to analyze Citibike bikeshare usage data at daily level combined with Central Park weather data and predict future bikeshare usage demand in terms of no of trips per day for next 30 days. Future usage

demand is very critical to Citybike system operator for planning increase in the number of bicycles availability and expansion of the geographic footprint of docking stations. This enables data driven decision system for making informed business expansion plans resulting into increased ROI and long-term business profitability and viability of Citibike operations

Dataset Snapshot:

CitiBike data

An example record from the CitiBike dataset includes the following features:

feature name	value
tripduration (seconds)	527
starttime	10/1/2019 00:00:05.6
stoptime	10/1/2019 00:08:52.9
start station id	3746
start station name	6 Ave & Broome St
start station latitude	40.72430832
start station longitude	-74.00473036
end station id	223
end station name	W 13 St & 7 Ave
end station latitude	40.73781509
end station longitude	-73.99994661
bikeid	41750
usertype	Subscriber
birth year	1993
gender	1

Weather data - NCDC (National Climatic Data Center)

An example of the key weather data elements includes:

Feature	description	|Random date 1	|Random date 2
STATION	Station ID number	USW00094728	USW00094728
NAME	Name of station	NY CITY CENTRAL PARK, NY US	NY CITY CENTRAL PARK, NY US
LATITUDE		40.77898	40.77898
LONGITUDE		-73.96925	-73.96925
ELEVATION		42.7	42.7
DATE		1/30/2019	7/22/2019
AWND	Average Wind Speed		2.68
PRCP	Amount of precipitation	0.01	1.66
SNOW	Amount of snowfall	0.4	0
SNWD	Snow Depth	0	0
TAVG	Average temperature		
TMAX	Maximum temperature	35	90
TMIN	Minimum temperature	6	72
WDF2	Direction of fastest 2-minute wind		10
WDF5	Direction of fastest 5-second wind		340
WSF2	Fastest 2-minute Wind Speed		|14.1
WSF5	Fastest 5-second Wind Speed		|25.1
WT01	Fog, ice fog, or freezing fog?	1	1
WT02	Heavy fog or heavy freezing fog?		1
WT03	Thunder?		|1
WT06	Glaze or rime?		
WT08	Smoke or haze?		

Data Cleaning and Imputation

Citibike dataset is huge. Processing big volume data like this demands sophisticated big data platform with distributed data processing capability. To resolve system scaling challenge, we employed stochastic random sampling technique for this analysis [13]. Underlying dataset is selected for scientific experimentation and analysis using uniform random sampling technique reducing the volume of data for faster computation [13]

In addition to that following data cleaning and data imputation techniques are used for data preparation.

1] Data column names changes slightly from month to month data file. Need to add column name standardization transformation for data processing
2] In some months, Citibike specifies dates in YYYY-MM-DD format, while in other months, dates are Specified in MM/DD/YYYY format. Need to add date standardization transformation for data processing
3] Dataset has some records with unusually high recording of bikeshare session duration. (StartTime and EndTime) Need to employ outlier detection techniques to identify such records and data imputation techniques to normalize the data
4] Data Aggregation and join: Aggregate individual Citibike trip data by day, and join to daily weather data

Data Analysis and inference

Citibike users tend to book shorter bikeshare trips, most of them are less than 10 mins

Table 1: Summary of Trip durations before truncation

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
supplied_secs	61.0000000	375.0000000	621.0000000	911.5364944	1064.0000000	1688083.0000
calc_secs	60.0000000	375.0000000	621.0000000	912.0153679	1065.0000000	1688083.0000
calc_mins	1.0000000	6.2500000	10.3500000	15.2002561	17.7500000	28134.7167
calc_hours	0.0166667	0.1041667	0.1725000	0.2533376	0.2958333	468.9119
calc_days	0.0006944	0.0043403	0.0071875	0.0105557	0.0123264	19.5380

The above indicates that the duration of the trips (in seconds) includes values in the millions which likely reflects a trip which failed to be properly closed out.

We removed cases with unreasonable trip_duration values Let's assume that nobody would rent a bicycle for more than a specified time limit (say, 3 hours), and drop any records which exceed this:

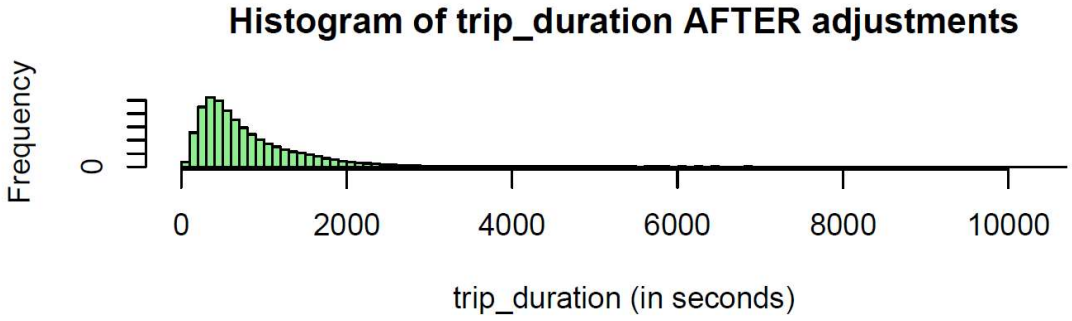
[1] "Removed 157 trips (0.174%) of longer than 3 hours."
[1] "Remaining number of trips: 90213"

Summary of trip durations AFTER censoring/truncation: After we eliminate cases which result in extreme values, the duration of the remaining trips is more reasonable.

Table 2: Summary of trip durations AFTER truncations:

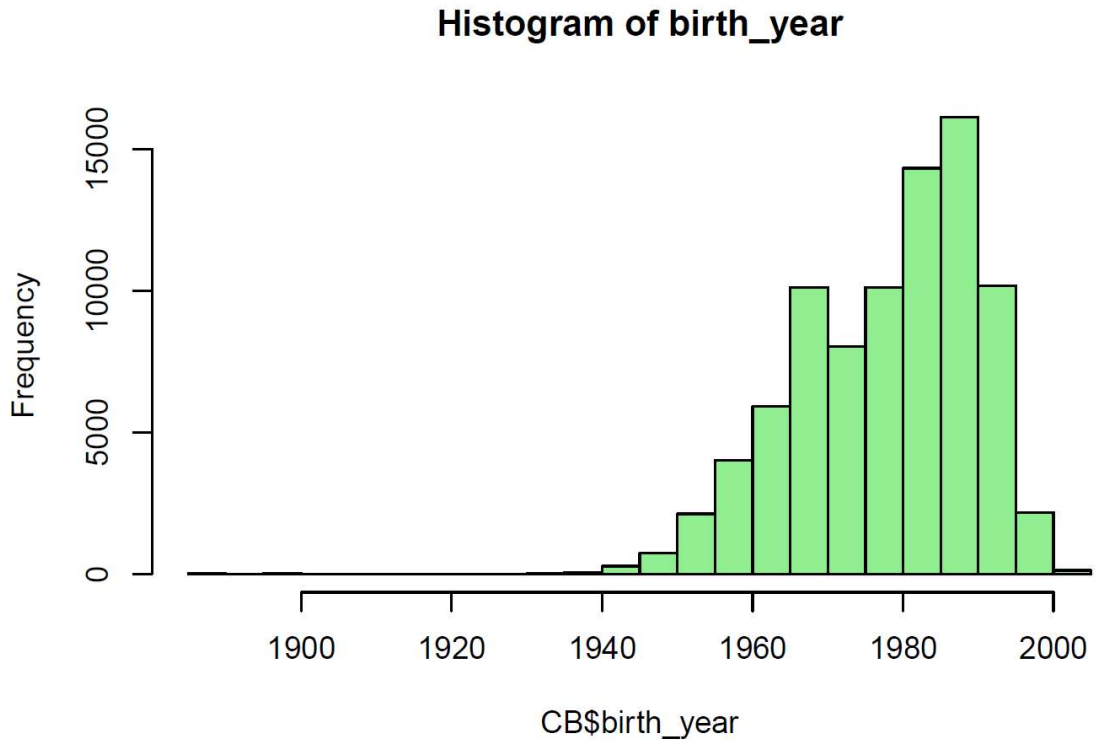
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
supplied_secs	61.0000000	364.0000000	594.0000000	777.6994724	997.0000000	10617.0000000
calc_secs	60.0000000	364.0000000	594.4970000	778.1873234	997.7920001	10617.5160000
calc_mins	1.0000000	6.0666667	9.9082833	12.9697887	16.6298667	176.9586000
calc_hours	0.0166667	0.1011111	0.1651381	0.2161631	0.2771644	2.9493100
calc_days	0.0006944	0.0042130	0.0068808	0.0090068	0.0115485	0.1228879

Trip duration statistics after outlier removal looks more realistic



Most of Citibike users are in their 30s/40s

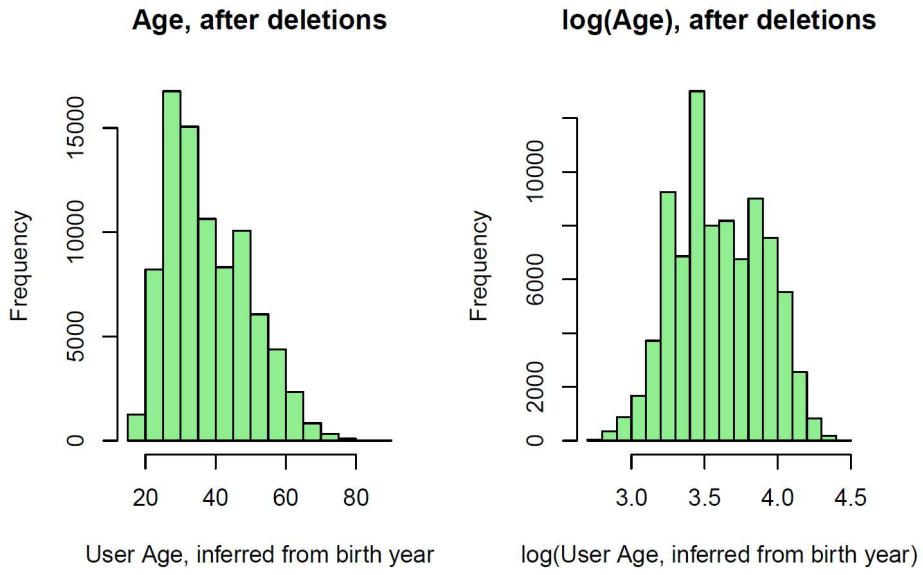
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1885    1969    1981    1978    1988    2003    5828
```



Birth year recorded for some of the users is 1885 which can't be true

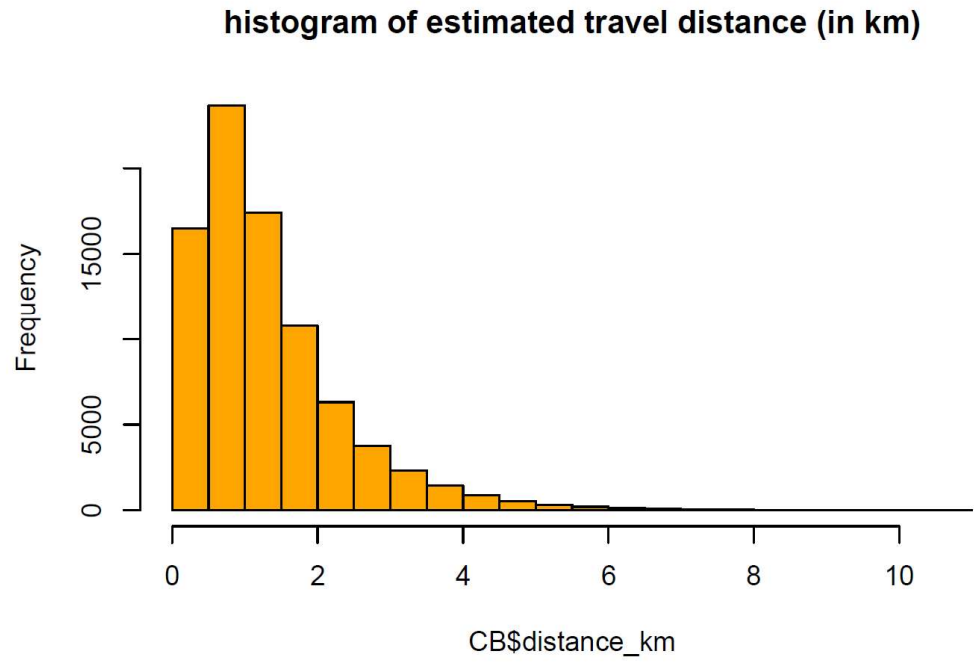
Remove trips associated with very old users (age>90) and remove trips associated with missing birth_year)

[1] "Removed 40 trips (0.044%) of users older than 90 years."
[1] "Removed 5828 trips (6.449%) of users where age is unknown (birth_year unspecified)."
[1] "Remaining number of trips: 84345"



Citibike consumers mostly use bikeshare service for shorter trips < 2 KM [a]

a This is straight-line distance between (longitude, latitude) points – it doesn't incorporate an actual bicycle route.

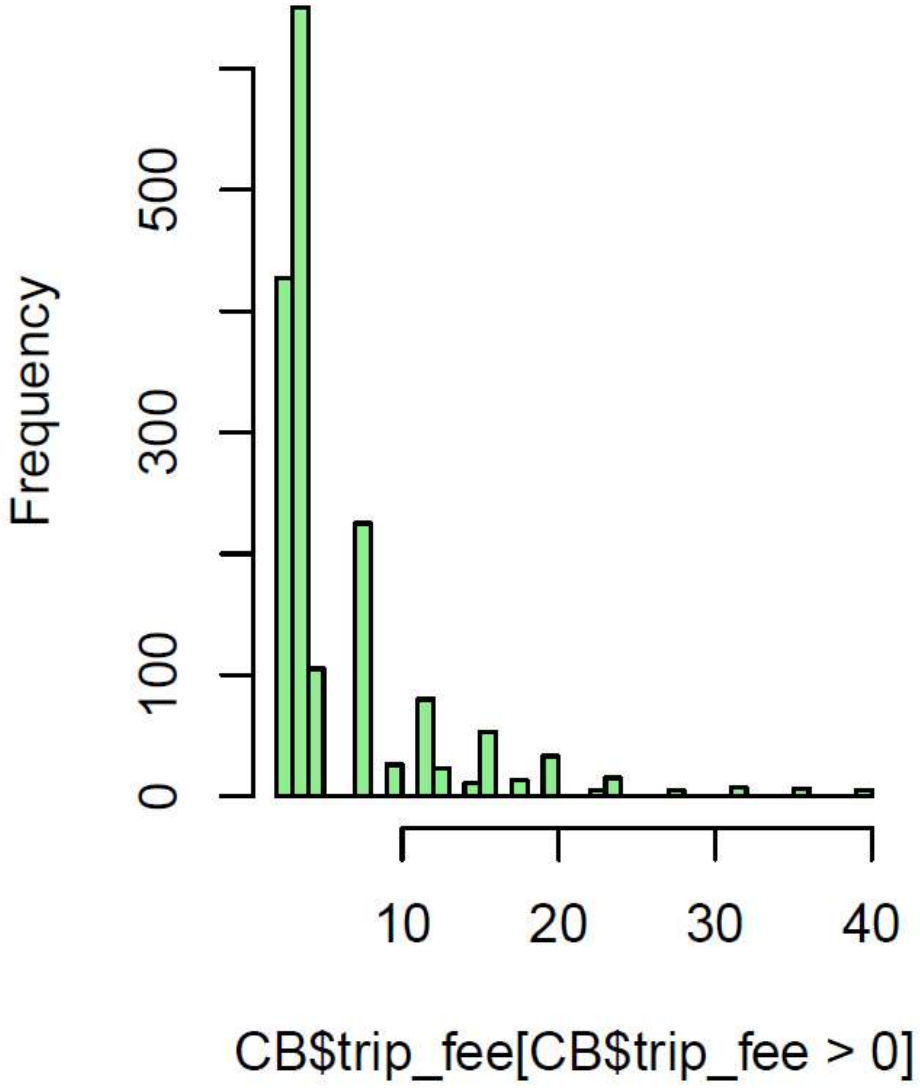


Analyze Trip Fee

Compute usage fee There is a time-based usage fee for rides longer than an initial period:

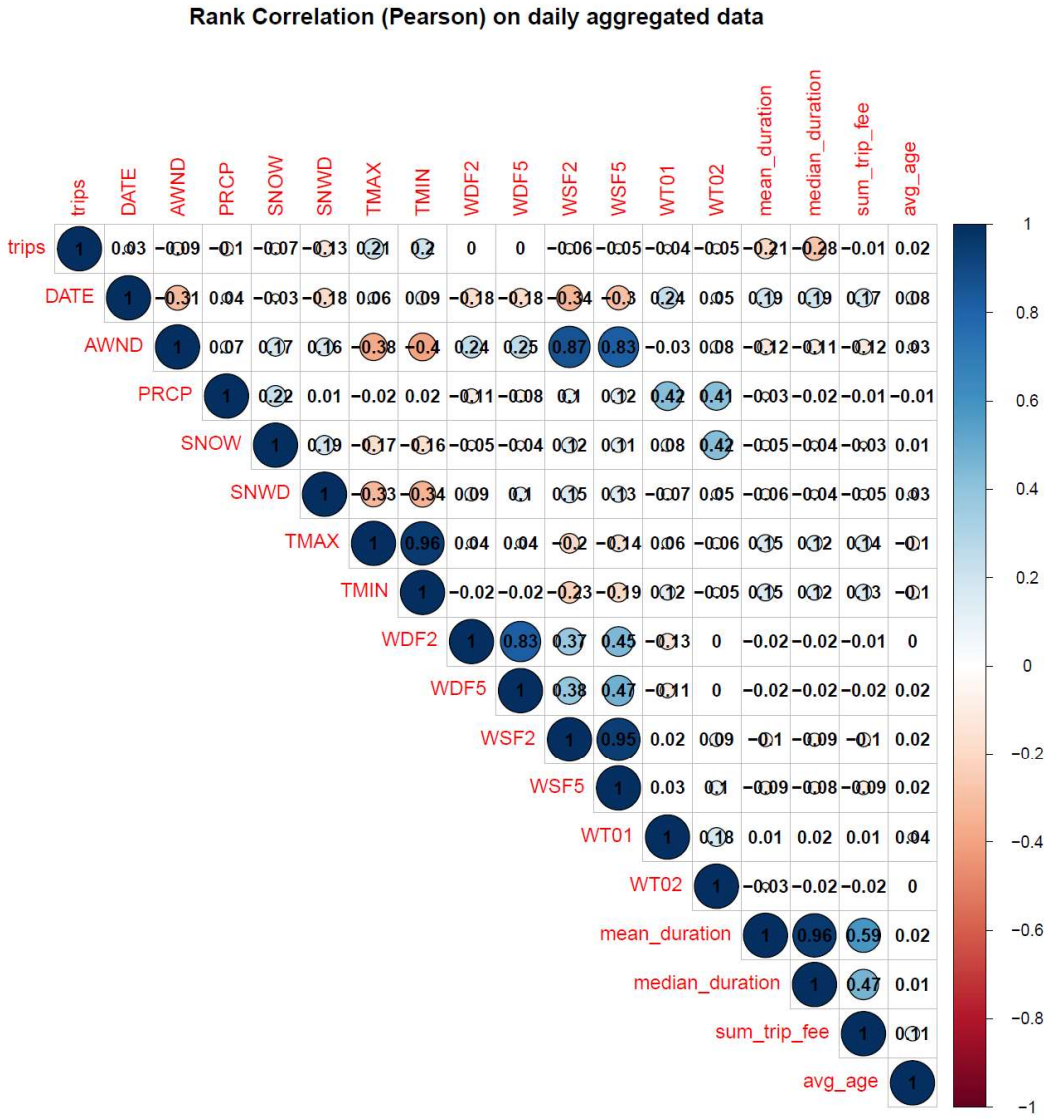
- For user type=Subscriber, the fee is $2.50 per 15 minutes following an initial free 45 minutes per ride.

• For user type=Customer, the fee is $4.00 per 15 minutes following an initial free 30 minutes per ride.
• There are some cases where the user type is not specified (we have relabeled as "UNKNOWN", and we do note estimate usage fees for such trips.)



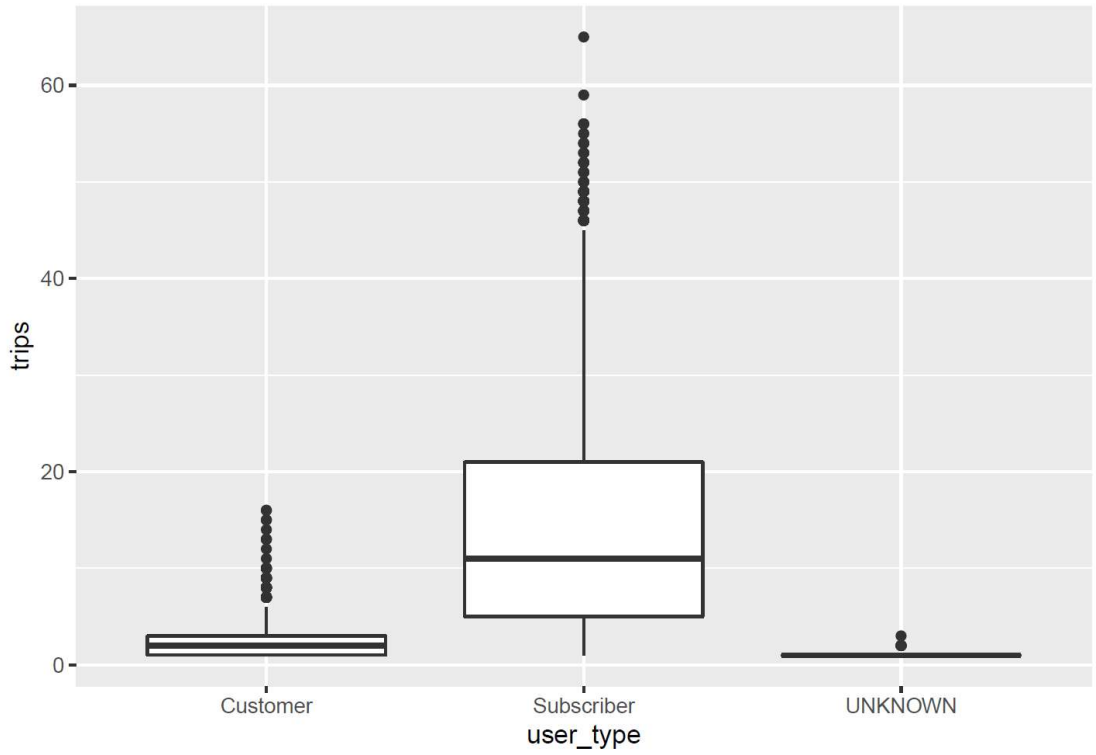
Most of Citibike users spent less than 10$ per bikeshare trip

Feature correlation: Correlation between input feature set reveals very interesting relationship between variables, and surfaces out potential problems of multicollinearity [14].



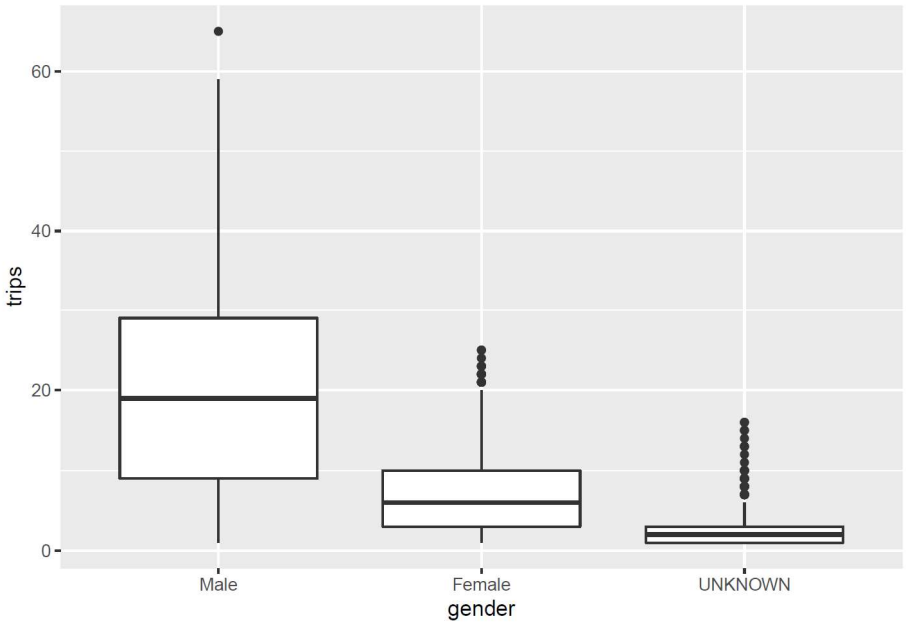
1] Negative correlation between trips and AWND suggests that people tend to take fewer bike trips on a windy day
2] Negative correlation between trips and SNOW suggests that as amount of snowfall increases number of bikeshare trips decreases
3] Negative correlation between trips and SNWD suggests that as amount of snow depth increases number of bikeshare trips decreases
4] TMAX and TMIN has positive correlation with trips. This indicates that on a warm sunny day people tends to take more bikeshare trips

Number of trips by user types
Citibike users who have monthly subscription tends to take a greater number of trips compared to other user types

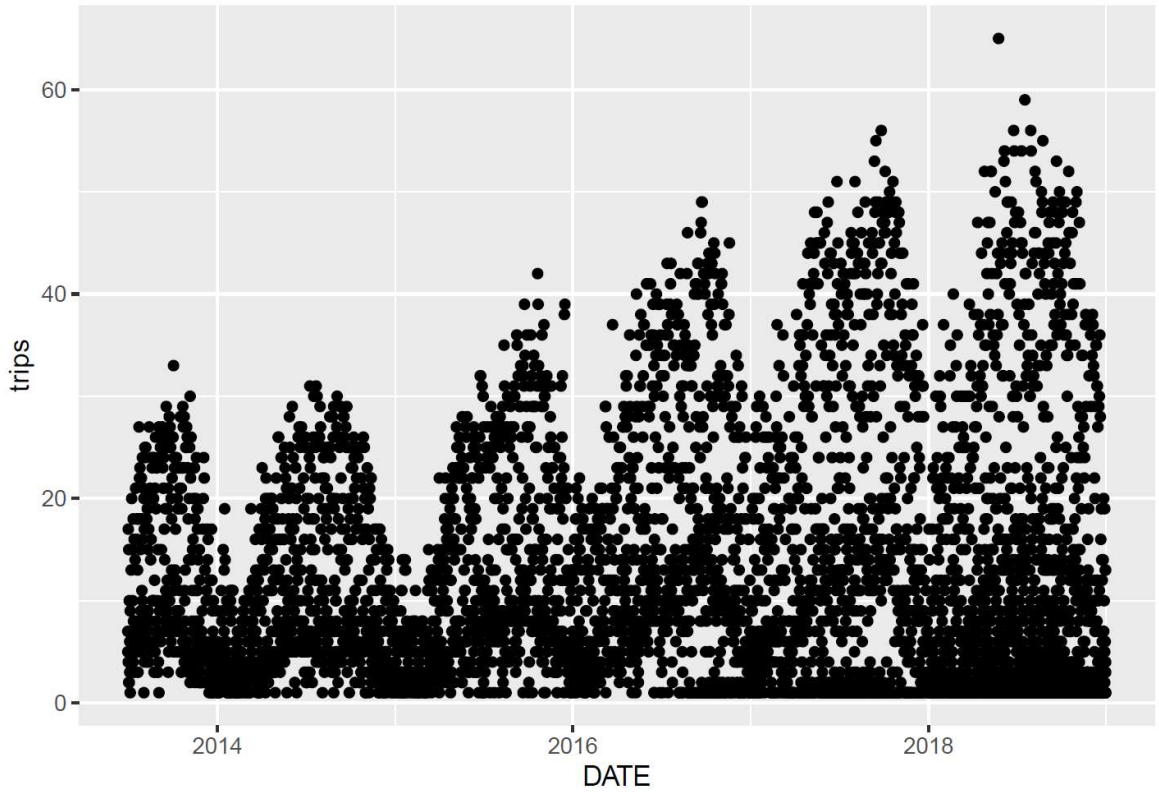


Number of trips by gender

Male customers tend to take more bikeshare trips compared to female counterpart



Year over year trips



Trips taken year over year shows a clear evidence of seasonality. Number of trips decreases in winter and peaks in summer. We can see that magnitude of seasonal component in a time series is increasing year over year. This indicates that Citibike bikeshare service became popular and more and more consumers are using the service. Time series also show increasing trend and we can see that it is non-stationary

Feature importance and variable selection

We leveraged random-forest based algorithm known as "Boruta" to assess variables "importance" in estimation of the target variable as "Confirmed", "Tentative", or "Rejected." The graph from the Boruta algorithm indicates the variables confirmed as important in green (on the right) while rejected variables are in red (on the left.) The variables in the middle (in yellow) are "tentative."

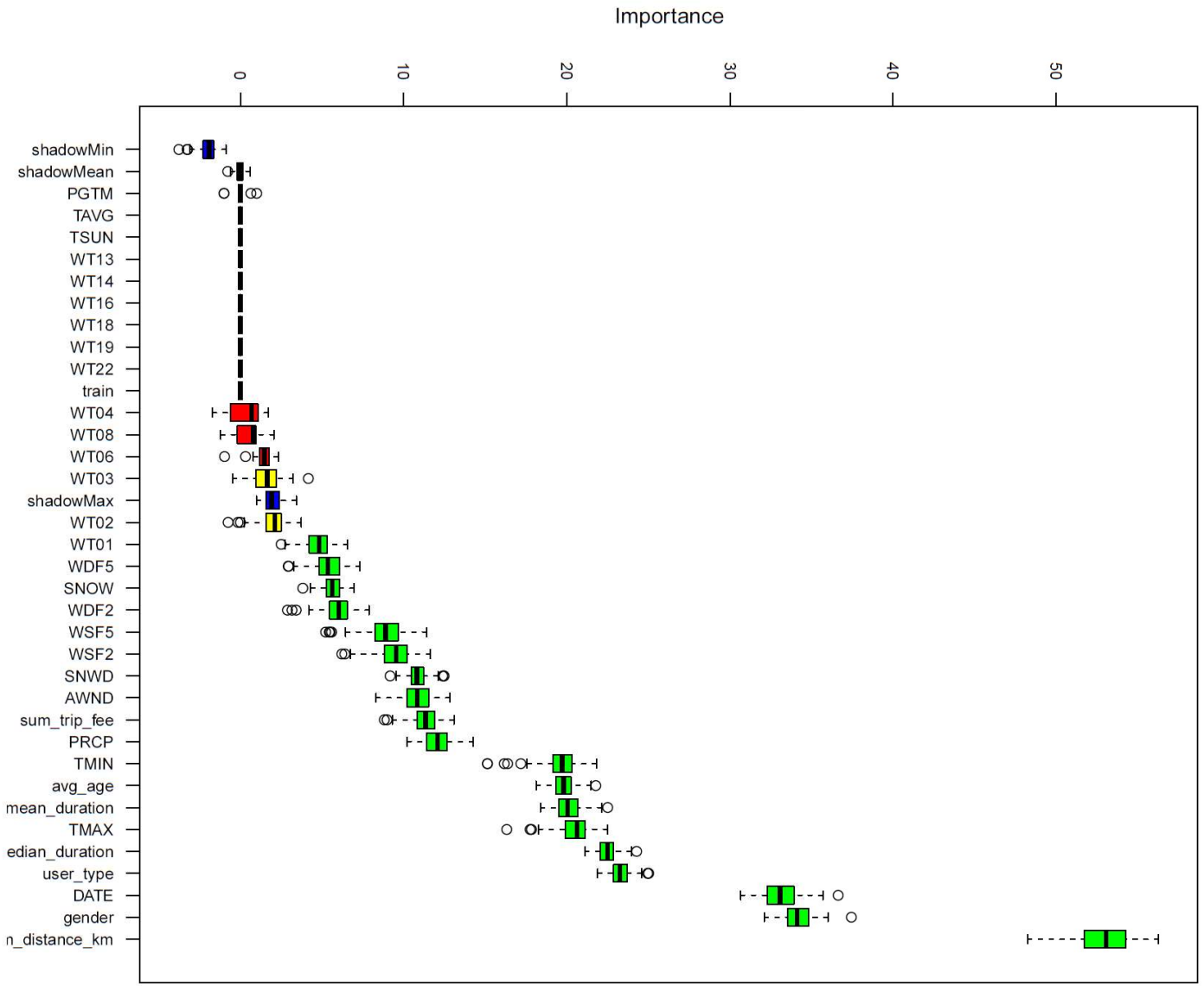

Importance
0
10
20
30
40
50
shadowMin
shadowMean
PGTM
TAVG
TSUN
WT13
WT14
WT16
WT18
WT19
WT22
train
WT04
WT08
WT06
WT03
shadowMax
WT02
WT01
WDF5
SNOW
WDF2
WSF5
WSF2
SNWD
AWND
sum_trip_fee
PRCP
TMIN
avg_age
mean_duration
TMAX
edian_duration
user_type
DATE
gender
n_distance_km

Table 3: Boruta results sorted by median value

BorutaFinalAlphaNames	BorutaFinalAlphaResults	BorutaMedianAlphaNum
PGTM	Rejected	-Inf
TAVG	Rejected	-Inf
train	Rejected	-Inf
TSUN	Rejected	-Inf
WT04	Rejected	-Inf
WT06	Rejected	-Inf
WT08	Rejected	-Inf
WT13	Rejected	-Inf
WT14	Rejected	-Inf
WT16	Rejected	-Inf
WT18	Rejected	-Inf
WT19	Rejected	-Inf
WT22	Rejected	-Inf
WT03	Tentative	1.64945194243239
WT02	Tentative	2.11633350230996
WT01	Confirmed	4.83249465978541
WDF5	Confirmed	5.36695674002365
SNOW	Confirmed	5.6229979537176
WDF2	Confirmed	6.01907170623564
WSF5	Confirmed	8.90302463691063
WSF2	Confirmed	9.54648436389202
SNWD	Confirmed	10.7949118384744
AWND	Confirmed	10.8365323005427
sum_trip_fee	Confirmed	11.3369707957975
PRCP	Confirmed	12.0672514525478
TMIN	Confirmed	19.6963854571892
avg_age	Confirmed	19.8013140670482
mean_duration	Confirmed	20.0313281605762
TMAX	Confirmed	20.6156043555391
median_duration	Confirmed	22.4861858330595
user_type	Confirmed	23.2447374328727
DATE	Confirmed	33.0493493321453
gender	Confirmed	34.0878629495875
sum_distance_km	Confirmed	53.0821655307592

Based on above statistical analysis following variables are identified as important for predicting number of trips

- WT01: Fog, ice fog, or freezing fog?
- WDF5: Direction of fastest 5-second wind
- SNOW: Amount of snowfall
- WDF2: Direction of fastest 2-minute wind
- WSF5: Fastest 5-second Wind Speed
- WSF2: Fastest 2-minute Wind Speed
- SNWD: Snow Depth

- AWND: Average Wind Speed
- Sum_trip_fee : Total trip fee
- PRCP: Amount of precipitation
- TMIN: Minimum temperature
- Avg_age: Avg age
- Mean_duration: Mean trip duration
- TMAX: Maximum temperature
- Median_duration: Median trip duration
- User_type: User type
- Date: Date
- Gender: Gender of user
- Sum_distance_km: Sum of distance in KM

4. Model building

TBD: Will be added in final submission

5. Conclusion

TBD: Will be added in final submission

6. References

1. Citibike website: https://www.citibikenyc.com/
2. Citibike data description: https://www.citibikenyc.com/system-data
3. Citibike trip data repository: https://s3.amazonaws.com/tripdata/index.html
4. National Climatic Data Center (NCDC) weather data: https://www.ncdc.noaa.gov/cdo-web/search
5. **https://en.wikipedia.org/wiki/Cycling**
6. Westland et al. examined consumer behavior in bike sharing in Beijing using a deep-learning model incorporating weather and air quality, time-series of demand, and geographical location; later adding customer segmentation.
7. [@Faghih-Imani_Eluru_2018] An et al. examine weather and cycling in New York City and find that weather impacts cycling rates more than topography, infrastructure, land use mix, calendar events, and peaks. They do so by exploring a series of interaction effects, which each capture the extent to which two characteristics occurring simultaneously exert a combinatorial effect on cycling ridership – e.g., how is cycling impacted when it is both wet and a weekend day or humid day in the hilliest parts of the cycling network?
8. In the 1990s, Nankervis examined the effect of weather and climate on university student bicycle commuting patterns in Melbourne, Australia by examining counts of parked bicycles at local universities and correlating with the weather for each day, finding that the deterrent effect of bad weather on commuting was less than commonly believed (though still significant.) [@Nankervis_1999]

9. [@Westland_Mou_Yin_2019] Jia et al. performed a retrospective study of dock less bike sharing in Shanghai to determine whether introduction of such program increased cycling. Their methodology was to survey people in various neighborhoods where the areas were selected by sampling, and the individuals were selected by interviewing individuals on the street.
10. [@Jia_Ding_Gebel_Chen_Zhang_Ma_Fu_2019] Jia and Fu further examined whether dock less bicycle-sharing programs promote changes in travel mode in commuting and non-commuting trips, as well as the association between change in travel mode and potential correlates, as part of the same Shanghai study.
11. [@DellAmico_Iori_Novellani_Subramanian_2018] Zhou analyzed massive bike-sharing data in Chicago, constructing a bike flow similarity graph and using a fast-greedy algorithm to detect spatial communities of biking flows. He examined the questions 1. How do bike flow patterns vary because of time, weekday or weekend, and user groups? 2. Given the flow patterns, what was the spatiotemporal distribution of the over-demand for bikes and docks in 2013 and 2014? [@Zhou_2015]
12. Hosfordetal surveyed participants in Vancouver, Canada and determined that public bicycle share programs are not used equally by all segments of the population. In many cities, program members tend to be male, Caucasian, employed, and have higher educations and incomes compared to the general population. Further, their study determined that the majority of bicycle share trips replace trips previously made by walking or public transit, indicating that bicycle share appeals to people who already use active and sustainable modes of transportation
13. https://research-methodology.net/sampling-in-primary-data-collection/random-sampling/
14. https://en.wikipedia.org/wiki/Multicollinearity