Citibike New York City

Oct 2015-Oct 2020 statistical data analysis

Group 16

Wang, Su, s399wang
Mabbayad, Mabi, mmabbaya
Xue, Kira, w27xue
Jajcanin, Igor, ijajcani
Cebula, Daniel Adam, dacebula
Shi, Wenwen, w65shi

Project Objective

The primary goal of this project is to analyze data gathered for New York City Citi Bike Program from Oct 2015 to Oct 2020.

Data collection is based solemnly on rental start and end time information i.e., duration of the trip for each bike. It does not collect other factors that could have been related to weather conditions, traffic congestions or trip path information.

The process includes data gathering, data cleaning, grouping and data analysis based on the bike ID (type), origin and destination stations, as well as the time analysis grouped by the day of the week and month. Analysis should also include distance travelled based on the geo information provided for start and end stations. Distance traveled calculation is simplified as a straight line between start and end stations using Haversine formula, rather than calculation of the actual path travelled.

The main aspects of python library and techniques that we used are as follow:

- Data cleaning/manipulation: pandas, NumPy
- Visualization: matplotlib, seaborn
- Statistical Modelling: pymc3

Introduction

Data Analysis

Data analysis is a process of capturing, cleaning, transforming and modeling data in order to gain insight on the available dataset. The focus here is to understand the data, find patterns, discover relationships within a dataset to answer some questions one might have. Data analysis has a longer history, traditionally data analysis was primarily reactive and focused on reporting, over time, it has evolved from dealing with smaller, structured dataset to steam and process more complex and unstructured data sources. More and more organizations are relying on data analysis to make informed business decisions. In this report, our team will be building models using SARIMA, Bayesian using Markov Chain Monte Carlo and Ordinary Least Squares regression.

Background

The New York Citi Bike Rental program is one of the most popular bike sharing programs in the United States. With 14,500 bikes and 950 stations across Manhattan, Brooklyn, Queens and Jersey City, it provides bike riders an affordable way to get around the city. The team will be exploring its publicly available ridership data to answer some of the following questions. Where do Citi Bikers ride? When do they ride? How far do they go? Which stations are most popular and can we make conclusions as to why? What time of the day, days of the week, month, season are most popular to ride the bike? What gender and age group bikes are most popular and used by? There are many more questions that arise for the growing popularity of city bike riders that is becoming worlds wide phenomena.

Bike ridership analysis can be very complex. It depends on multiple factors, obviously weather being one of the key factors, but also safety, road and trails infrastructures, bike qualities, etc.

Dataset Overview

There are multiple data sets that we used for our data analysis. They are downloaded from https://www.citibikenyc.com/system-data. This data is originally collected by Citibike New York City. These datasets contain all the bike rentals from Oct 2015 to Oct 2020.

Here are the attributes information:

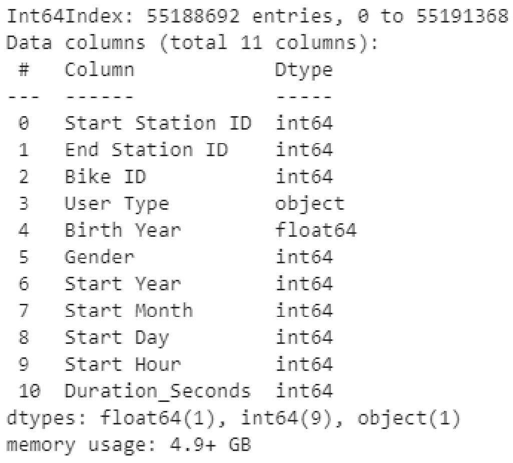
TRIP_DURATION': Duration of the trip (seconds)
'START_TIME': Trip Start date and time
'STOP_TIME': Trip End date and time
'START_STATION_ID': Trip Start Station ID
'START_STATION_NAME': Trip Start Station Name
'START_STATION_LATITUDE': Trip Start Station Latitude
'START_STATION_LONGITUDE': Trip Start Station Longitude
'STOP_STATION_ID': Trip End Station ID
'STOP_STATION_NAME': Trip End Station Name
'STOP_STATION_LATITUDE': Trip End Station Latitude
'STOP_STATION_LONGITUDE': Trip End Station Longitude
'BIKE_ID': Bike ID
'USER_TYPE': User_type (Customer – 24 hrs or 3-day pass user, Subscriber – Annual Member)
'BIRTH_YEAR': Year of Birth
'GENDER': Gender (0 – unknown, 1 – male, 2 – female)

According to the data source documentation staff test trips, system inspection trips and any trip below 60 seconds in length were removed from the dataset.

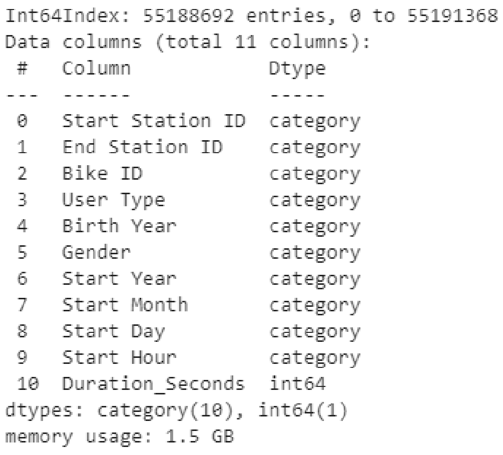
Data Preparation

An exploratory analysis was done on a subset of the data to better understand the main characteristics of the dataset. Features available from data included geographical coordinates of the stations which the team used to calculate the distance of a bike ride from its starting location to the recorded ending location using the Haversine formula. The records that did not contain both Start Station ID and End Station ID were excluded from the analysis.

Several features with fixed number of possible values were converted using pandas. Categorical function to reduce the size of the on-memory data for analysis. This conversion trimmed the dataframe size from 4.9 GB to 1.5 GB. The resulting dataset was then used for all the analysis and prediction models.



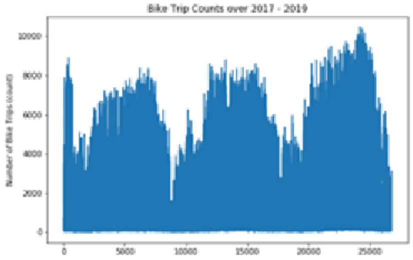
```
Int64Index: 55188692 entries, 0 to 55191368
Data columns (total 11 columns):
 #   Column            Dtype
---  ------            -----
 0   Start Station ID  int64
 1   End Station ID    int64
 2   Bike ID           int64
 3   User Type         object
 4   Birth Year        float64
 5   Gender            int64
 6   Start Year        int64
 7   Start Month       int64
 8   Start Day         int64
 9   Start Hour        int64
 10  Duration_Seconds  int64
dtypes: float64(1), int64(9), object(1)
memory usage: 4.9+ GB
```

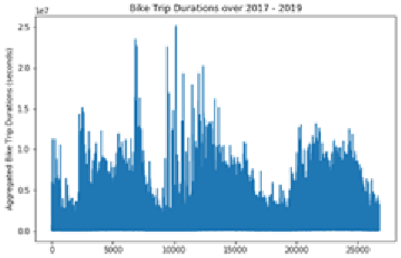


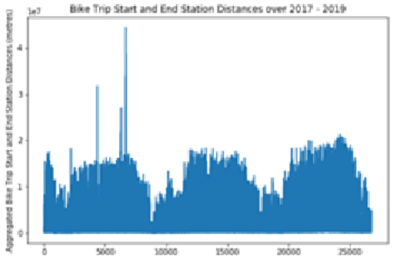
```
Int64Index: 55188692 entries, 0 to 55191368
Data columns (total 11 columns):
 #   Column            Dtype
---  ------            -----
 0   Start Station ID  category
 1   End Station ID    category
 2   Bike ID           category
 3   User Type         category
 4   Birth Year        category
 5   Gender            category
 6   Start Year        category
 7   Start Month       category
 8   Start Day         category
 9   Start Hour        category
 10  Duration_Seconds  int64
dtypes: category(10), int64(1)
memory usage: 1.5 GB
```

Dataset Visualizations

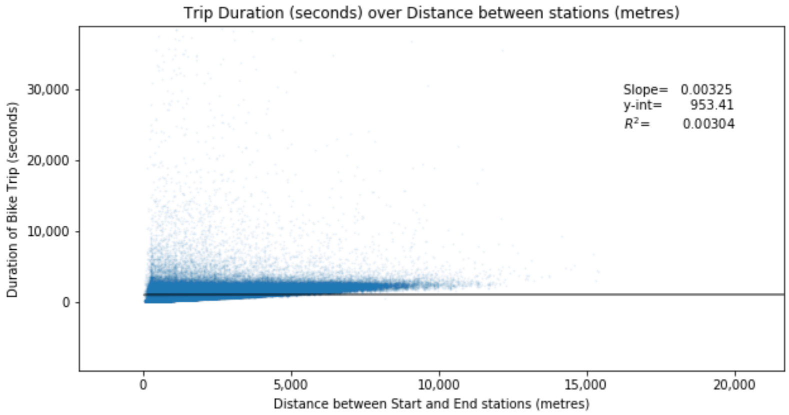
One can plot the entire count, duration and distance of bike trips to see the overall features of the data.





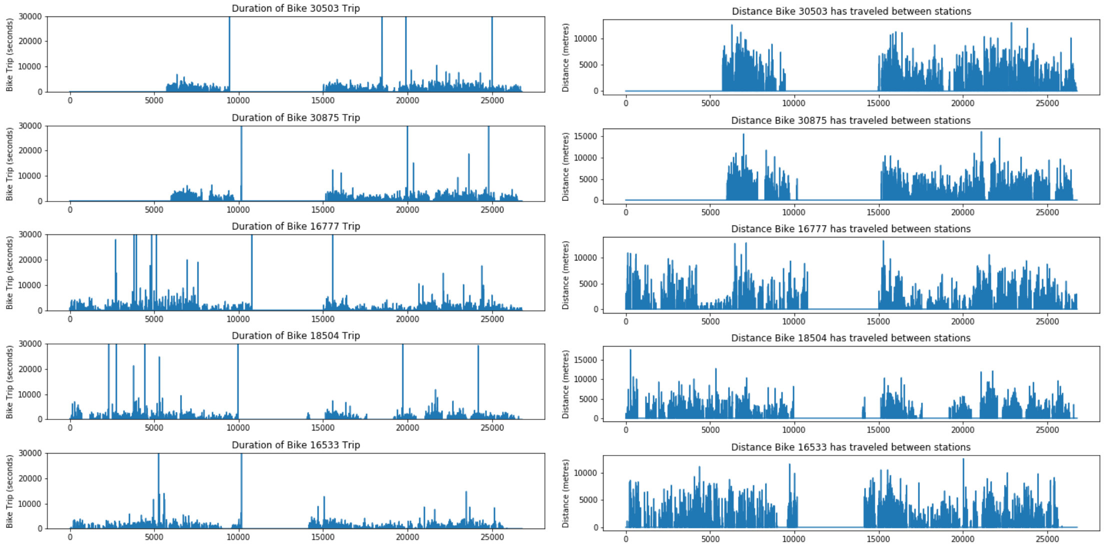


Initially a relationship between bike trip duration and distance between start and end stations was investigated.



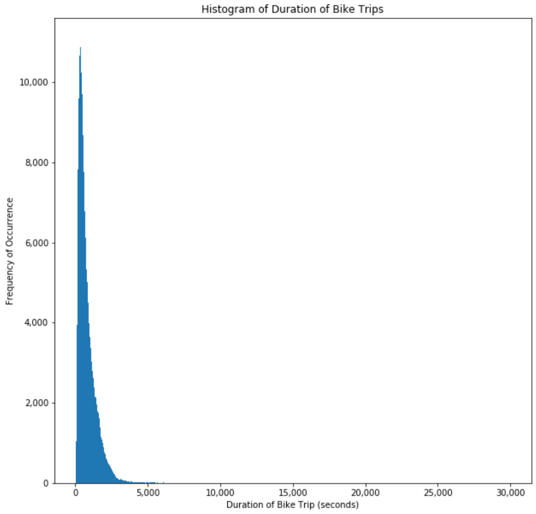
It appears that no linear relationship can be found in the above scatter plot; whereas common sense would indicate as bike trips take longer durations one can travel to further end stations from start stations.

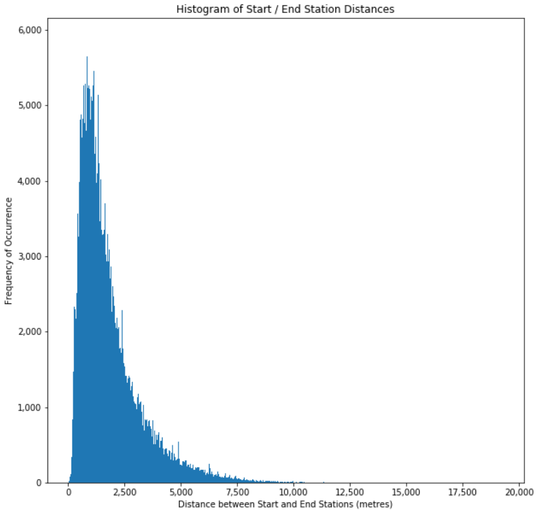
Top 5 performing Bike IDs have very unique plots for both duration and distance measurements.



It appears to be a typical Poisson distribution.

Histograms of Bike Trip Duration and Start and End Station distance are quite revealing.



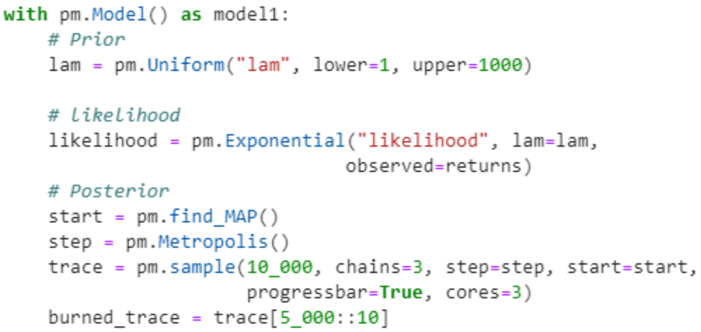


Both histograms encompass the whole dataset and is indicative of a combined Gaussian (normal) and exponential distribution. This is necessary information for creating Bayesian models.

Bayesian Models

3 models were created for the count, duration and distance of bike trips.

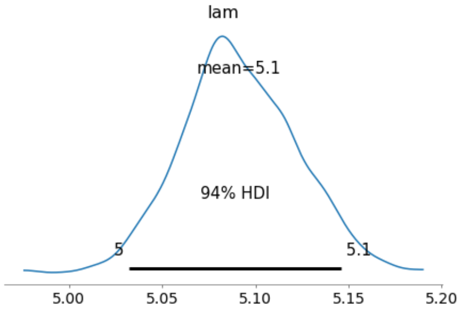
1. Total Count of Bike Trips

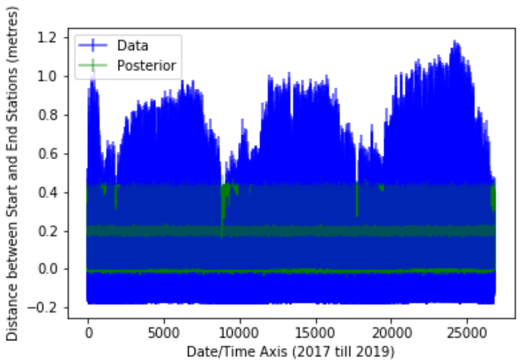


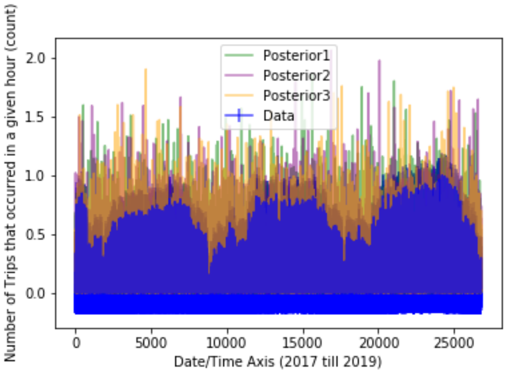
```
with pm.Model() as model1:
    # Prior
    lam = pm.Uniform("lam", lower=1, upper=1000)

    # Likelihood
    likelihood = pm.Exponential("likelihood", lam=lam,
                                observed=returns)

    # Posterior
    start = pm.find_MAP()
    step = pm.Metropolis()
    trace = pm.sample(10_000, chains=3, step=step, start=start,
                      progressbar=True, cores=3)
    burned_trace = trace[5_000::10]
```

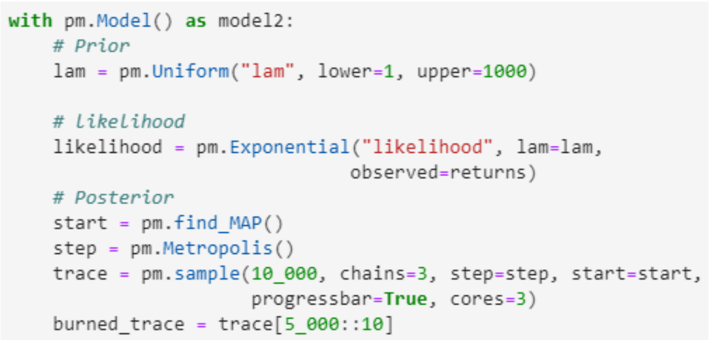






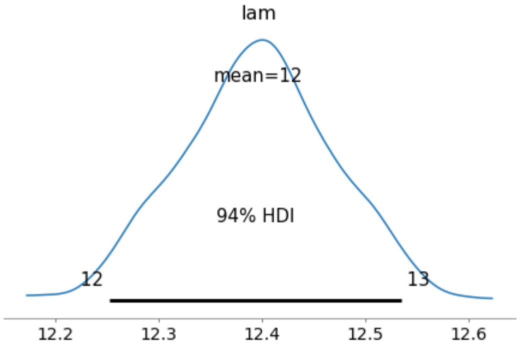
It appears the exponential distribution is a good for fitting posterior randomly sampled values.

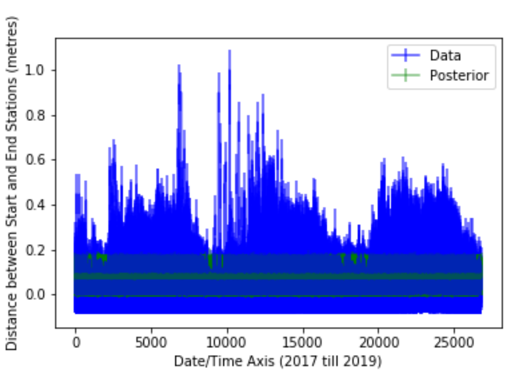
2. Total Duration of Bike Trips

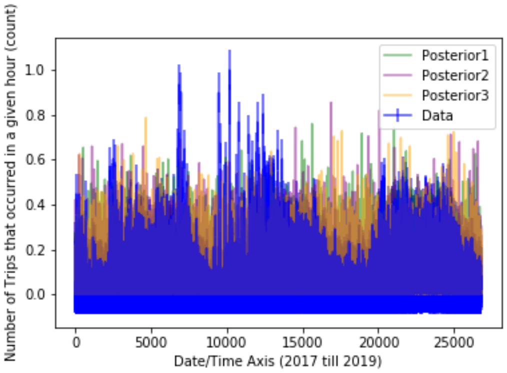


```
with pm.Model() as model2:
    # Prior
    lam = pm.Uniform("lam", lower=1, upper=1000)

    # Likelihood
    likelihood = pm.Exponential("likelihood", lam=lam,
                               observed=returns)
    # Posterior
    start = pm.find_MAP()
    step = pm.Metropolis()
    trace = pm.sample(10_000, chains=3, step=step, start=start,
                      progressbar=True, cores=3)
    burned_trace = trace[5_000::10]
```

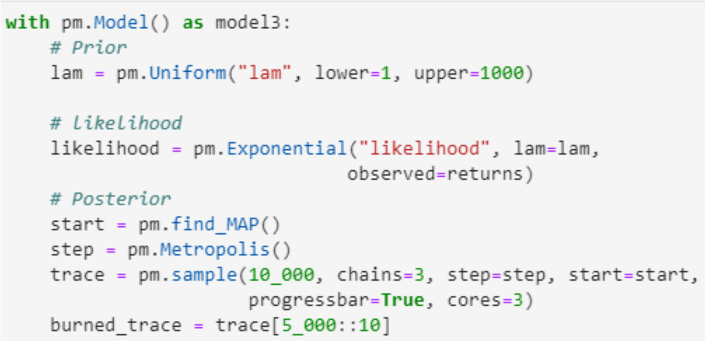






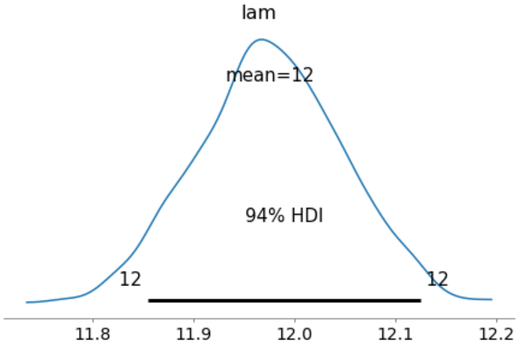
It appears the exponential distribution is a good for fitting posterior randomly sampled values.

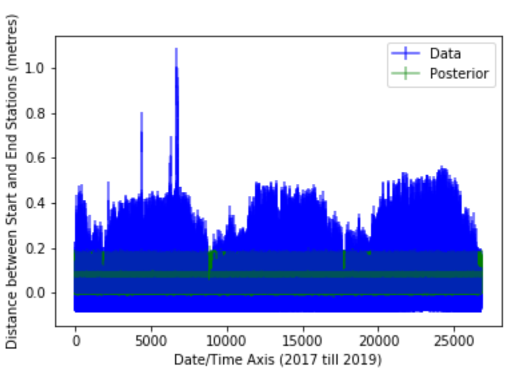
3. Total Distance of Bike Trips

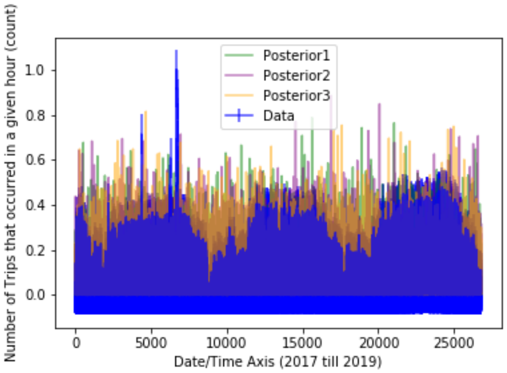


```
with pm.Model() as model3:
    # Prior
    lam = pm.Uniform("lam", lower=1, upper=1000)

    # Likelihood
    likelihood = pm.Exponential("likelihood", lam=lam,
                                observed=returns)
    # Posterior
    start = pm.find_MAP()
    step = pm.Metropolis()
    trace = pm.sample(10_000, chains=3, step=step, start=start,
                      progressbar=True, cores=3)
    burned_trace = trace[5_000::10]
```







Statistical Data Analysis

Predictive Model

Time series (SARIMA) analysis

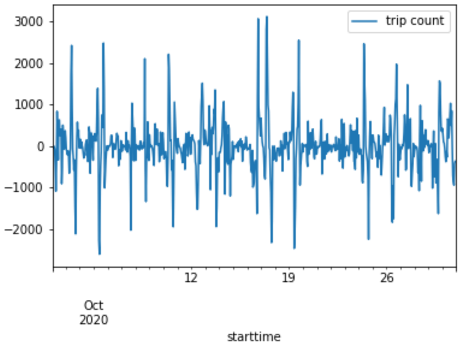
One predictive model used was SARIMA time series analysis. Since the data provides timestamps for each trip, we wondered if we could use the timestamps in order to predict how many trips will be expected in the future.

Time series data transformation

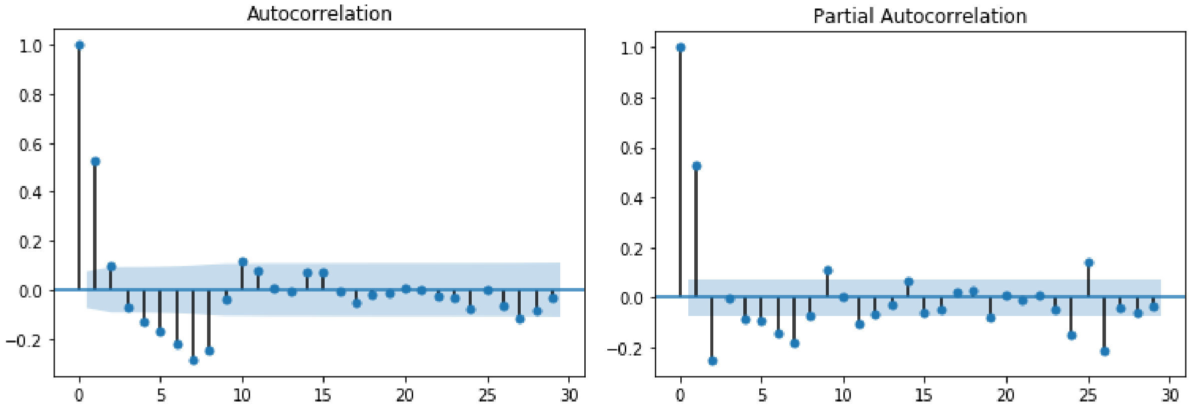
We extracted the data for the period of October 1st, 2020, to October 31st, 2020. As we were only concerned with the timestamps, we dropped all columns except the “starttime”, and added a column to indicate that 1 trip was initiated for each row of data. From there, we aggregated the table based on hour, and sum of trips initiated within that hour. This created a time series of number of Citibike trips initiated every hour. Based on this data, we can create a model to forecast, on an hourly basis, what’s the expected number of trips in the future.

Time series visualization & differencing

First, we plotted the time series to visually inspect if there are any trends or seasonality. While it is hard to determine if a trend exists, the high fluctuation in trip counts imply that there is a seasonality at play here.



Next, we inspect the ACF and PACF plots.

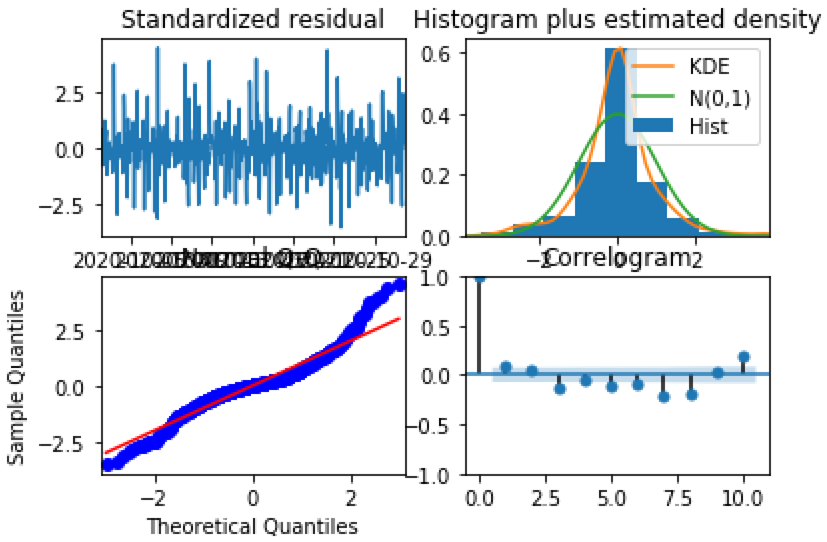


ACF plot shows that there is a high autocorrelation at lags 1, 12 (negative autocorrelation), and 24. Viewing the PACF plot shows only partial autocorrelation that is not already accounted for in mutual correlations with other variables.

Model creation

Based on the ACF and PACT plots, we will use a SARIMA model with order: ARIMA (0,1,1) x (0,1,1)24.

Once the model has been determined, we need to ensure the model is valid by analyzing the residuals.



The model validation plots show that:

- The residuals plot shows no trend in the residuals
- The histogram shows that residuals are normally distributed and centered on zero
- The normal Q-Q plot is somewhat linear
- The correlogram plot shows that some values are outside of the 95% confidence bounds

To forecast the data, we will use this model and forecast it against November data (October being the training data).

We can see that while the model is able to capture general trends (movement of trip counts), it is not very accurate at predicting the exact values. We may need to change the orders of differencing, or switch to a model which can account for exogenous factors, for even more accurate forecasting.

One factor to consider may be that during COVID situations, ridership trends may be drastically different month to month, based on lockdowns. Factors like that would not be considered in a SARIMA model.

Two Sample T-test

Data Preparation

The data required in this statistical analysis are the duration time periods and distances between stations grouped by gender. There are two datasets (duration time periods by gender & distances between stations by gender) used in the T-test, and each has two independent samples for women and men, sample size is 1000

Hypothesis

Null Hypothesis: The data for duration time periods and distances between stations have the same distribution regardless of the gender.

Alternative Hypothesis: The data for duration time periods and distances between stations have different distributions for women and men.

Test Results

Below are the statistics found in the T-test process for each sample:

1. **For Duration Time Periods (by seconds):**

	Men	Women
Mean	606,079	264,581
Standard Deviation	969,463	490,139
Standard Error	16,274	7,882

Standard error on the difference between the samples is 18,083

T Value is 18.88

2. **For Distances Between Stations:**

	Men	Women
Mean	1,175,408	405,713
Standard Deviation	2,019,068	673,540
Standard Error	63,848	21,299

Standard error on the difference between the samples is 67,307

T Value is 11.44

The further away the T value is from 0, the lower the chance that the null hypothesis is true. Because both results show large T values, the null hypothesis can be rejected, and the conclusion is that the data for duration time periods and distances between stations have quite different distributions for women and men.

Correlation Analysis

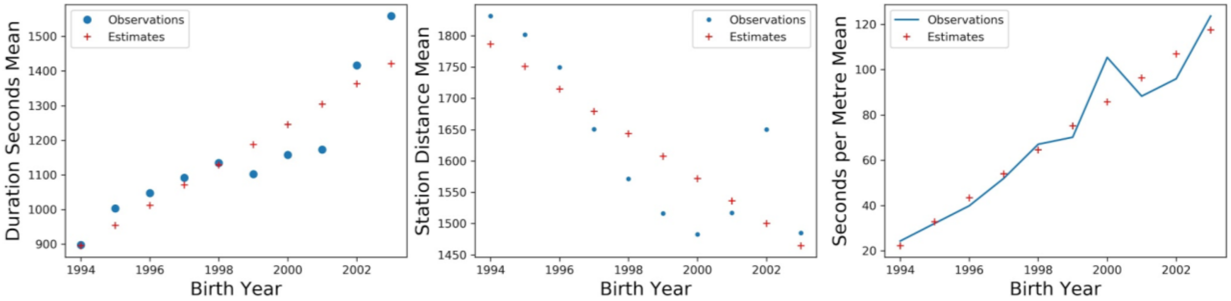
Ordinary least squares (OLS) method

Data Preparation

The data required in OLS analysis are the birth year of users, and station distance, duration time, and the speed of biking. The birth years of each user are provided in the dataset. The speeds of biking are calculated from the distance of stations and the duration time period. After preparing the data, the mean of “seconds per meter” was calculated for each birth year. Noticing there’s a big part of the birth year is the 19th and 20th century, so the analysis would be only focused on the users who are born after 1994.

Model and Visualization

The hypothesis for the OLS analysis is to prove the relationship between birth year and duration seconds, birth year and station distance, birth year and seconds per meter. Compared with duration seconds and station distance, there’s a strong linear correlation between birth year and seconds per meter. The younger people have a higher seconds per meter, which means they spend more time on biking compared with older people.



Conclusion

Categorical function proved to be very efficient in reducing the size of the dataframe. Overall dataset was fairly scattered and it appear to be no linear relationship between indicating that longer duration would mean traveling longer destinations. Analysis indicate of combined Gaussian and exponential distribution. Based on Bayesian models it appears the exponential distribution is a good for fitting posterior randomly sampled values. On time series analysis there was showing of a seasonality trend. Using SARIMA model was good to capture general trends. However, the it was lacking in predicting exact values. That was most likely due to the differences in test data due to COVID lockdowns and downturns. T-test to prove relationship in distance traveled and gender showed larger T values. Therefore, the conclusion is that the data for duration time periods and distances between stations have quite different distributions for women and men. Finally, OLS method used to show dependences of age group (i.e., birth year) and distance and speed show that younger population were traveling on higher speed.