BikeStorms!

An Analysis of Bike and Taxi Rides in Relation to Weather Patterns

Josh Benzon, Shivani Mendiratta, Sam Minars, Chance Emerson

Background

Recently, CitiBikes have become increasingly popular among New Yorkers as an affordable and eco-friendly mode of transportation. However, we want to investigate if weather conditions affect the number of people who use CitiBikes on any given day and how that may relate to the number of people who opt for Yellow Cabs instead.

Hypothesis

- If the weather description shows "heavy rain", the average bike ride distance will decrease.
- If the weather description shows "clear", the average taxi ride duration will increase.
- If the weather description shows "heavy rain", the total number of bike rides and taxi rides will both decrease.

Data

We collected taxi data from the NYC Taxi and Limousine Commission website and bike data from the Amazon News - CitiBike website. As shown in Figure #1, we cleaned our data to get the average durations/distances, and total trips for each date for both taxi and bike rides.

We web scraped and created an API call to the Meteo website to collect our weather data. We organized our data to get the weather descriptions, average temperatures, and total precipitations for each date.

Methodology

For our hypotheses, we used "two-sided t-tests" in order to compare one data from the taxi and/or bike to one weather description. By simply separating the data into two groups, we were able to calculate the mean and variance of the bike and taxi data for each group prior to conducting the t-test.

For our first analysis, we theorized that higher temperatures in the summertime would lead to people taking longer taxi rides by using a linear regression model.

For our second analysis, we wanted to classify the weather on a given day from the average duration, average distance, and total number of CitiBike trips by using the K-Nearest Neighbors model.

Results and Diagrams

Sample Data

date	weatherDescription	avgTemp	totalPrecip	avgBikeDuration	avgBikeDistance	totalBikeTrips	avgTaxiDuration	avgTaxiDistance	totalTaxiTrips
2021-03-26	Light Rain	61.0	0.8	1339.0	1.972150077619876	71615.0	962	6.992840875536818	73535
2021-03-27	Clear	54.2	0.0	1561.0	2.213448604706594	84712.0	904	4.756122678187322	67392
2021-03-28	Heavy Rain	53.3	18.0	1447.0	1.823250308863806	25256.0	779	5.449800243272944	40404

Figure #1: A sample of the data for three consecutive days between 2021-03-26 to 2021-03-28.

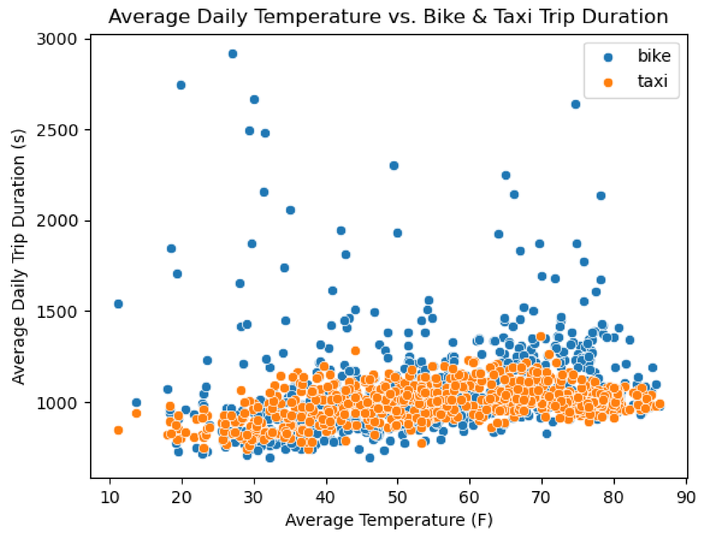

Figure #2: A scatter plot of the average trip duration and average temperature between bikes and taxis.

Based on Figure #2, the scatter plot provides a relationship between the average daily trip duration (in seconds) and the average temperature (in Fahrenheit) among the data points between bike and taxi rides.

Furthermore, the graph reveals information from the result of our linear regression machine learning analysis model of the data. There is a slight positive correlation between bike trip duration and temperature, which yielded a testing R-Squared value of 0.18 and a training R-Squared value of 0.20.

Based on Figure #3, the line plot provides a monthly snapshot of the total amount of bike and taxi trips. As such, the graph reveals a seasonal pattern in regards to whether the total number of trips either increases (during the warmer months) and decreases (during the colder months).

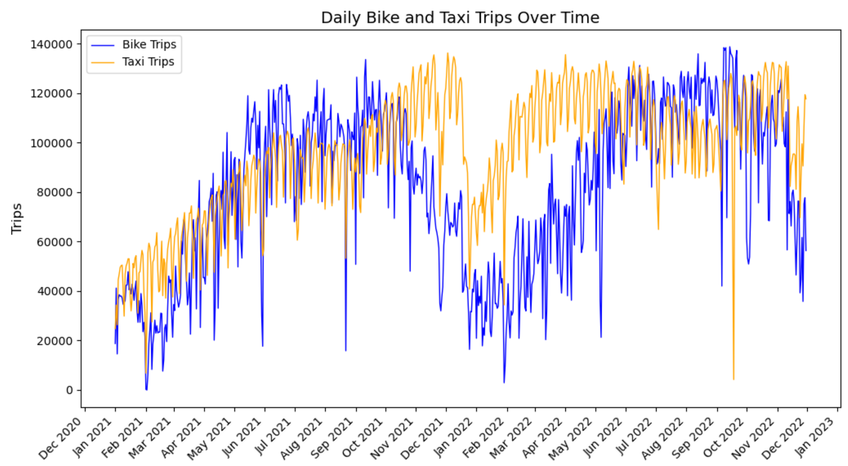

Figure #3: A line plot of the total trips with bikes and taxis between Dec 2020 to Jan 2023.

Significance

For our first hypothesis, there does not seem to be a significant difference in the average bike distance when the weather description is "Heavy Rain" compared to when it's not.

For our second hypothesis, there does not seem to be a significant difference in the average taxi duration when the weather description is "Clear" compared to when it's not.

For our third hypothesis, there was a significant decrease in the mean number of bike rides during heavy rain compared to non-heavy rain. However, there was no significant difference in the mean number of taxi rides during heavy rain and non-heavy rain.

Conclusions

For our linear regression model, there's a possibility of a positive linear relationship between the taxi duration and the weather description. While the scatterplot data appears to be fairly linear, the analysis led to a low correlation (with only an R-value of roughly 0.20). Likewise, our intuition of higher temperatures leading to longer taxi rides may be due to the lack of data points from only one to two years.

For our classification model, the training accuracy of 0.59 and testing accuracy of 0.48 suggests that the weather on a given day in New York City cannot not be accurately predicted from the average duration, average distance, and total number of bike trips on that day. Likewise, the weather labels corresponded may be too broad in order to predict these labels from three metrics alone (average duration, average distance, total number of trips).

Challenges

- Both CitiBike and New York Taxi source data only covered roughly a one year span (with the December 2022 being missing as well).
- If there was only one small part of the day that was raining, snowing, etc., the entire weather description would be labeled as that label, which could lead to incorrect weather representations.