BikeStorms

Samantha Minars (sminars), Josh Benzon (joshbenzon), Shivani Mendiratta (mendiratta27), Chance Emerson (ceboop)

Goal

We wanted to analyze the relationship between the daily weather conditions in New York City with CitiBike and taxi usage.

Data

We collected the CitiBike data by downloading CSV files from Amazon News and the taxi data by downloading parquet files from the NYC Taxi and Limousine Commission. We made API calls to Meteo to retrieve our weather data. Our data's time frame covers from January 2021 to November 2022. We cleaned our data by dropping unnecessary columns and combined others to calculate daily averages. Our final data table has 699 rows with the following columns: date, weatherDescription, avgTemp, totalPrecip, avgBikeDuration, avgBikeDistance, totalBikeTrips, avgTaxiDuration, avgTaxiDistance, and totalTaxiTrips. The weather data is fairly uniform in regards to the time of the month, but generally, 'skewed' depending on the time of season. The bike distances are fairly uniform, centering from 1 to 3 miles. The taxi distances are slightly skewed towards distances from the 10 miles range with the minimum being around 2 miles and maximum around 20 miles.

Hypothesis Test Results & Analysis

Hypothesis Claim #1: There's statistically significant evidence that the average number of CitiBike rides decreases during heavy rain, but heavy rain conditions do not signal a significant decrease in the average number of taxi rides.
Support for Hypothesis Claim #1: We used a two-sided t-test to evaluate whether there was a significant difference in the mean number of bike rides and taxi rides during heavy rain and non-heavy rain. Since the p-value for the bike data is less than 0.05, we conclude that there's a significant decrease in the mean number of bike rides during heavy rain as compared to non-heavy rain conditions. Since the p-value for the taxi data is greater than 0.05, we conclude that there's not a significant difference in the mean number of taxi rides during heavy rain and non-heavy rain weather.

Two-Sided t-Test		
Statistic	**Bike Data**	**Taxi Data**
t-statistic	-3.404	-0.135
p-value	0.001	0.893

Hypothesis Claim #2: There's not statistically significant evidence to support that the average duration of taxi rides on a given day in New York City increases when weather conditions are clear in comparison to non-clear conditions.
Support for Hypothesis Claim #2: We used a two-sided t-test to evaluate whether there's a significant difference in the average taxi duration when the weather description is 'Clear' compared to when it's not. Since the p-value is higher than 0.05 (5%), we fail to reject the null hypothesis, and cannot accept the alternate 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.

Two-Sided t-Test	
Statistic	**Value**
t-statistic	0.995
p-value	0.320

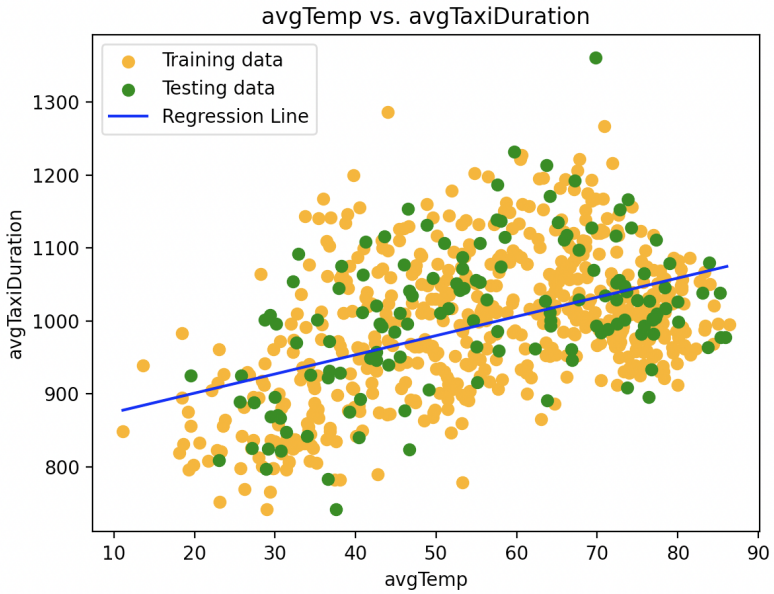
Hypothesis Claim #3: There's not statistically significant evidence to support that the average distance of CitiBike trips on a given day in New York City decreases in heavy rain when compared to non-heavy rain conditions.
Support for Hypothesis Claim #3: We used a two-sided t-test to evaluate whether there's a significant difference in average bike distance when the weather description is 'Heavy Rain' compared to when it's not. Since the p-value is higher than 0.05 (5%), we fail to reject the null hypothesis, and cannot accept the alternate 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.

Two-Sided t-Test	
Statistic	**Value**
t-statistic	-1.801
p-value	0.072

Machine Learning Results & Analysis

Machine Learning Claim #1: There is a positive correlation between the average temperature and average duration of taxi trips on a given day in New York City.

Support for Claim #1: We ran a linear regression model to determine the relationship between average temperature and average duration of taxi trips. Looking at the figure below, we can see that as the average temperature increases, the average duration of taxi trips also increases. We got a regression line of $\hat{y} = 2.63x + 848.43$ for our model, and a testing R^2 value of 0.18 and a training R^2 value of 0.20. Although these R^2 values are relatively low, there's still support that a positive relationship exists between average temperature and average taxi trip duration.



Machine Learning Claim #2: 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 CitiBike trips.

Support for Claim #2: We ran a K-Nearest Neighbor analysis to predict the weather label (Heavy Rain, Light Rain, Clear, & Snowing) from the average duration, average distance, and total number of CitiBike trips for a given day in New York City. After plotting the training and validation accuracies against the number of neighbors, we found that the optimal k value was 7. After running the KNN model with k=7, we got a training accuracy of 0.59 and testing accuracy of 0.48 suggesting that there's not enough evidence to support weather classification from our collected CitiBike data.

KNN Model (k=7)	
Type of Data	**Accuracy**
Training	0.586
Testing	0.479