Capstone 1 Statistical Methods

Background

The purpose of this project is to predict the quantitative target variable, price, of an Uber or Lyft ride based on Boston Uber/Lyft ride data and weather data taken from a real-time API, ranging from November 26 to December 18, 2018. To achieve this, supervised regression methods will be employed on a clean dataset containing the weather data appended to the ride data, along with a dummy variable column representing the occurrence of a sports game/event that occurred within one hour of the game's start time and end time. This section details the methodology involved in the statistical analysis process, performing tests to determine unique relationships between features and feature groups.

Dataset Description

Real time ride data, collected using Uber and Lyft API queries, containing 10 features including: ride application used (Uber, Lyft), distance between pickup and dropoff, timestamp of ride, pickup location of ride, destination of ride, price of ride, surge multiplier of ride (over much price was increased), specific type of ride (Uber Black, Lyft Lux XL), and corresponding id is included. There are over 690000 ride data instances recorded for this dataset.

Weather data, containing shared features with the ride dataset such as the location of the ride pickup, at timestamps ranging between November - December, contains unique features such as rain, clouds, humidity, wind, and pressure measurements.

Sports data, which was manually created, contains features for NBA Celtics and NHL Bruins games occurring between November 26 to December 18, 2018. Features include location of the game's stadium (e.g. North Station), start time, stop time. This data was collected from NBA and NHL 2018-2019 schedule.

Ride and Weather data provided from:
https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices and is relatively clean.

Statistical Methods

To identify which features are particularly significant in determining the target variable, price, of the Uber or Lyft ride, a measurement of correlation between the price and each quantitative features in the dataset was performed. Specifically, the Pearson correlation coefficient was calculated to determine which feature positively (value greater than 0) or negatively (less than 0) correlates with price. The results are shown below:

Feature	Pearson Correlation Coefficient (feature against price)
distance	0.345
surge_multiplier	0.241
temp	0.002
clouds	-0.003
pressure	0.043
rain	-0.006
humidity	-0.021
wind	0.023

The coefficients indicate a positive correlation relationship between price and distance, as well as, price and surge_multiplier. The rest of the features are fairly close to 0, indicating no significant correlation between these features and the price.

An assumption for regression states that multicollinearty cannot occur, a violation of the assumption that no independent variable is a perfect linear function of one or more variables. The presence of multicollinearity will affect the regression and prediction of the target variable negatively. To test for this, we calculate the Pearson correlation coefficient for each of the independent variables. These are the results:

feature	distance	surge_mult	temp	clouds	pressure	rain	humidity	wind
distance	1	0.026	0.003	-0.006	0.002	-0.012	-0.011	-0.003
surge_ mult	0.026	1	-0.013	0.028	0.112	0.022	0.031	-0.059
temp	0.003	-0.013	1	**0.414**	**-0.367**	0.135	**0.364**	0.113
clouds	-0.006	0.028	**0.414**	1	-0.202	0.207	**0.533**	0.101
pressure	0.002	0.112	**-0.367**	-0.202	1	-0.079	-0.146	**-0.589**

rain	-0.012	0.022	0.135	0.207	-0.079	1	0.235	0.205
humidity	-0.011	0.031	**0.364**	**0.533**	-0.146	0.235	1	-0.213
wind	-0.003	-0.059	0.113	0.101	**-0.589**	0.205	-0.213	1

Coefficients values in bold somewhat or do display either a positive or negative linear relationship between their respective independent variables. This may affect the interpretation of the regression model, and will be monitored for multicollinearity when modelling. The pairs of independent variables with a high degree of correlation occur in weather variables: temp and clouds, temp and pressure, temp and humidity, clouds and humidity, pressure and wind.

Next, we want to determine if the prices charged between similar Uber and Lyft ride types are the same or different. Because we do not know the population standard deviation, we performed five independent samples t tests, to compare the means of the two ride-type's prices, using a 0.05 significance level and not assuming equal variance. The null and hypothesis test for the two-tailed test of the five pairings are as such:

$H0$: *uber ride-type price = lyft ride-type price*
Ha: *uber ride-type price ≠ lyft ride-type price*

Two Groups	statistic	P-value
UberPool and Shared	209.790	0.0
UberX and Lyft	10.001	1.55e-23
UberXL and Lyft XL	13.139	2.11e-39
Black and Lux Black	-71.293	0.0
Black SUV and Lyft Black XL	-53.661	0.0

Because all of the calculated p-values were either 0, or less than the alpha level 0.05, we reject the null hypothesis for all five cases, and conclude that UberPool, UberX, and UberXL rides are generally charged more compared to their Lyft counterparts, based on the significantly positive t-statistics. Lux Black and Lyft Black XL rides are generally charged more than their Uber counterparts, based on significantly negative t-statistics generated.

Finally, we want to determine if weather plays a role in determining the price, and compare the prices of rides under rainy, cloudy, and cool weather. Once again, we perform three independent samples t-test for each of the three pairings: rainy and cloudy, rainy and cool, cool and cloudy:

*H*0: *weather 1 prices = weather 2 prices*
Ha: *weather 1 prices ≠ weather 2 prices*

Two Groups	statistic	P-value
Rainy[1] and Cloudy[2]	-3.272	.001
Rainy and Cool[3]	-8.959	3.309e-19
Cloudy and Cool	-2.962	.003

1. 'Rainy' weather determined when a measurement of rain > 0 occurs within the dataset.
2. 'Cloudy ' weather determined when the measurement of rain is 0, and measurement of clouds is 1.
3. 'Cool' weather determined when measured rain value is 0, and clouds measurement is < 1. Because the weather in Boston's area ranges from 40 to 50 degrees F, the weather is considered 'cool'.

The results of the t-test indicate that prices are different depending on the weather, as we would reject the null hypothesis in all three cases, because the p-value is less than 0.05. The statistic indicates that prices under rainy weather are generally lower than under cloudy and cool weather, and prices under cloudy weather are generally lower than under cool weather, according to the negative statistics calculated for all three cases.