Geraldo B. Padilla F.

# Introduction

In this assignment, we will explore some relationship between variables in a data set referred to Uber and Lyft rideshares in Boston, MA. The data set is a random sample of n= 2.500 observations from 693.071 observations in the original database, available at <https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma>

Specifically, we have the following guiding question: does the distance of the ride explain variation in the price of the ride?

1. In this vein, the first thing to do is exploring the distribution of the price (dependent variable) of the rides (shape, center, variation, and/or extreme values).

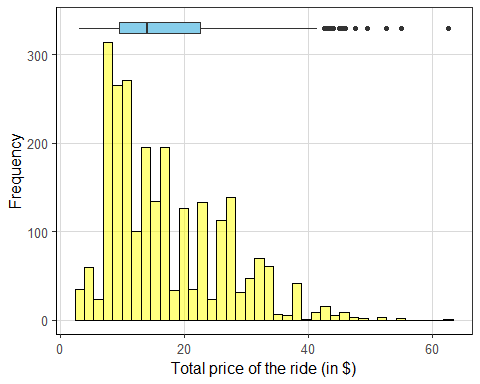
The mean is $16.84, with a standard deviation of $9.37.The range of values is $59.5, and the IQR is $13.00.

Table 1 shows descriptive statistics in detail:

## response mean min max median sd skewness kurtosis 25% 75%  
## 1 price 16.8438 3 62.5 14 9.370861 0.9473665 0.4683887 9.5 22.5

When exploring the Figure 1, price shows a positive skew. Moreover, the 50% of the values go from $9.50 up to $22.5.

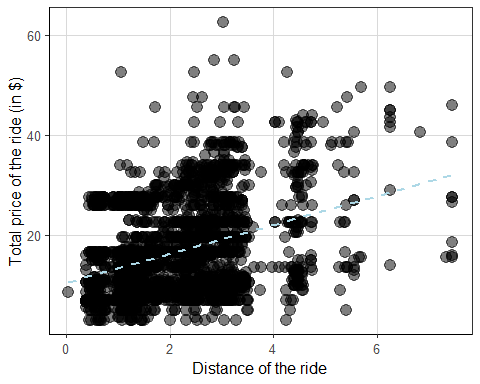
Figure 1



1. Second, we have to explore the bivariate distribution of the price of ride and distance of rides.

Based on Figure 2, it appears that there is no relationship between the two variables, or not a linear one at least. If we draw a line across the Figure, high variability is observed. On the other hand, there is a high concentration of both variables in the left-down part of the graph; maybe this is produced for the short length of the variable on the x-axis (distance), which is expressed in miles from 0.3 up to 7,46

Figure 2



We can also explore this relationship observing the Pearson’s correlation coefficient, which is 0.353. This reinforce the previous analysis: the relationship is positive but the magnitude is regular or even small if we consider that among the set of variables in the data set, distance seemed the most relevant characteristic.

To dig deeper into this assumption, we could explore other correlations between different pairs of variables in the data set. According to the Table 2, the high correlation coefficient is between our two variables already analyzed (price and distance), followed by the surge multiplier for the cost of rides (surge\_multiplier).

Table 2

## hour day month price   
## 0.04 -0.02 0.02 1.00   
## temperature distance surge\_multiplier humidity   
## 0.00 0.35 0.21 -0.02

1. Third, based on the previous analysis, we can claim that the relationship between the two variables is non linear.

At this point, it is worthy to recall that one thing is statistics results and another one is practical significance of the results. This means, according to the correlation coefficient, in fact, a positive linear relationship it is possible, but if we consider the visual representation of the relationship, this looks far away from a linear pattern.

On top of that, the magnitude of the relationship is small, especially if we take into account the huge size of the sample (n=2.500).

1. To answer the guiding question, we can perform a linear regression under the assumption that price (response) is a function of distance (predictor). The intercept is 10.44 and the slope is 2.89. But, what does this mean?

Based on the sample, we can assume that price increases as does the distance; more precisely, we could claim that for every single unit change on distance (portions of miles), price increases, on average, $2.89 (slope). In absence of distance, the price of rides would be, on average, $10.44; this is the intercept, the expected mean value of Y when X = 0

However, in this case the intercept does not make much sense, because rides do not have distance equals zero; in other words, if there is no distance, thus there is no payment. This a common problem in the interpretation of the intercept of linear regressions.

## (Intercept) distance   
## 10.439142 2.889949

We could visualize some predicted values to better understand the regression results:

|  |  |  |  |
| --- | --- | --- | --- |
| Intercept | Slope | Values (#description) |  |
| 10.44 | + 2.89 | \* 0.030 #min | 10.5267 |
| 10.44 | + 2.89 | \* 1.300 #1st quart | 14.197 |
| 10.44 | + 2.89 | \* 2.275 #median | 17.01475 |
| 10.44 | + 2.89 | \* 2.216 #mean | 16.84424 |
| 10.44 | + 2.89 | \* 2.860 #3rd quart | 18.7054 |
| 10.44 | + 2.89 | \* 7.460 #max | 31.9994 |

Again, the linear regression shows that as the distance increases so does the price of the ride.

1. Other way to analyze the relationship between price and distance is using the r-square and sigma, measures of variability.

First, the r-squared equals 12,4%, that is, the distance just explains, approximately, the 12% of the price variability. On the other hand, the standard deviation of the observed prices (sigma) respect with the regression line is 8.77.

These results make more sense in the context of the problem. First, the r-squared looks small considering, first, the size of the sample and, second, the kind of relationship expected between price and distance. When we analyze something like the price of rides, it is normal to assume that distance is one of the main factors involved. Therefore, a higher r-square was expected.

In addition, the sigma estimates expresses the high dispersion observed in the first plot between price and distance. The difference between the observed prices and predicted ones is large (8.77), which means that the percentage error is 52.3%. Once more, the variability is higher than expected and opens the question about what variable can aid to improve the linear regression estimation.

1. Finally, in a couple sentences, provide a summary of the overall model. Does the model appear to be useful to predict the price of a ride with only the distance of the ride? Use statistics from the analysis steps above to support your answer.