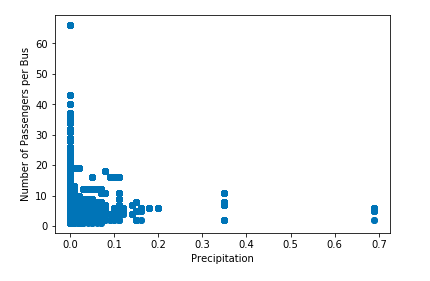
Why using weather data?

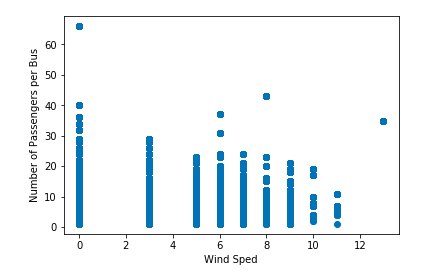
The main goal of using weather data was to see whether there is any correlation between different weather conditions and bus data, and if yes, than what kind of. The source of the data was NOAA, in particular NOAA weather station in Central Park, New York. Our original hypnosis stated that bad weather condition would have a positive correlation with number of passengers on the bus. In other words, busses would not come on time and our calculated time delta would be bigger, what results in higher crowdedness of the bus.

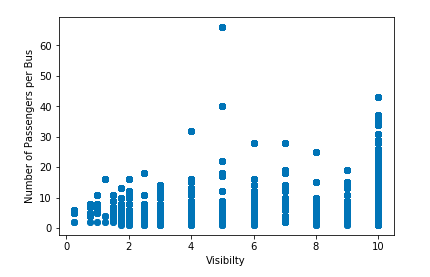
Cleaning the data:

The original data contained empty cells as well as string characters mixed with numbers. Since dropping all na values would result in inconsistent time value periods, it was decided to substitute those ‘bad’ rows with such values that would not influence further analysis (ex: 0 for precipitation or 15 for visibility).

Merging weather and bus data resulted in combined data set that allowed us to create scatter plots that demonstrated correlation between our variables.





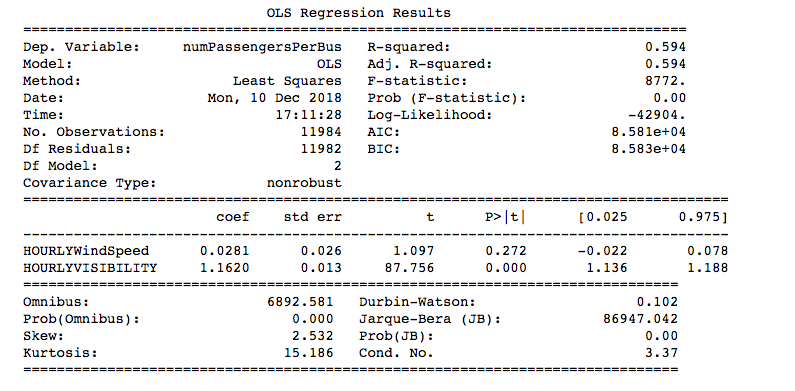


Analysis of scatter plots:

Our scatter plots roughly demonstrated that the days with the most crowding have 0 precipitation. In addition,  lower wind speed tends to mean more passengers. Finally, higher visibility generally means a higher number of passengers on the bus.

How strong is the correlation between independent (precipitation, wind speed, and visibility) and dependent variables (crowdedness)?

In order to answer this question, we performed OLS regression analysis.



Explanation of OLS:

R squared is the proportion of variance (%) in the dependent variable that can be explained by the independent variable. Our R-squared has a value of 0.594 what generally considered to have a moderate effect size. In addition, our p values indicate that there is 0% or 2% probability that that the result would have come up in a random distribution. Finally coefficient values demonstrate that when the independent variables increase by 1, dependent variable also increases. As a result, we can state that our correlation analysis is quite reliable.

Findings:

We found very interesting and unexpected results that contradict our original hypothesis of positive correlation between bad weather conditions and crowdedness of the bus. It seems like bus comes on time more often during bad weather conditions. The reason for that requires further research. Our assumption is that moderately bad weather stimulates regular drivers to stay at home what has a positive affect on bus arriving time.