Analyzing the NYC Subway Dataset

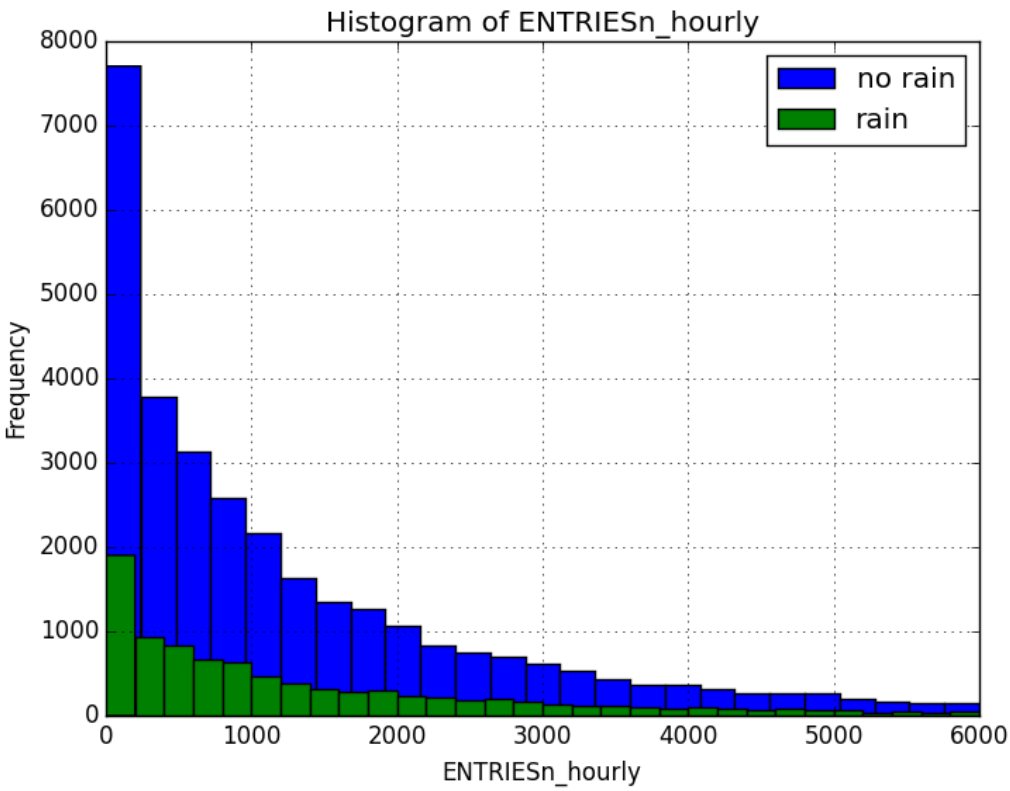
This project aims to analyze how ridership varies in the NYC Subway data set using statistical tests, linear regression and visualization.

Statistical Test (see mann-whitney.py)

Initial data exploration was carried out by visualizing the distribution of 'ENTRIESn_hourly' for days with rain(Group R), and days without rain(Group NR). The mean 'ENTRIESn_hourly' was calculated for each group:

with_rain_mean, **xR = 2028.20**

without_rain_mean, **xNR = 1845.54**



For both samples, distribution of 'ENTRIESn_hourly' is right-skewed and not normal. R has lower frequencies recorded at all 'ENTRIESn_hourly' values, compared to NR. Given that we know that **xR > xNR** , this observation in the histogram is a result of there being fewer rainy days than non-rainy days in the dataset.

As R and NR are are independent and the distribution of ridership is not normally distributed, the **Mann-Whitney U test** was used to further study if there was a significant difference in ridership between the 2 groups. Two-tail P value was used with a p-critical value of 0.05.

Null Hypothesis, H0 : the probability of the with-rain group(R) having a higher ridership than the with- no-rain group(NR) is 0.5 i.e. P(r > nr) = 0.5,

where P: probability

r: random draws r from R

nr: random draws nr from NR

Alternative Hypothesis, HA : the probability of the with-rain group(R) having a higher ridership than the with-no- rain group(NR) is not equal to 0.5 i.e. P(r > nr) ?0.5

Results:

The two-tailed p-value was < 0.01. As p-value < p-critical, the null hypothesis is rejected. Statistically, the probability of the rain group(R) having a higher ridership than the no-rain group(NR) is not equal to 0.5.

Linear Regression (see linear_regression.py)

Linear regression was done using sm.OLS.

The following variables were used:

rain, tempi, wspdi, conds – weather variables are likely to affect ridership. In less comfortable weather conditions such as rain, high or low temperatures, gusty winds or an overcast sky, people are more likely to use the subway than walk. If it's raining, traffic jams could be more likely, so subway ridership may also increase as people avoid taking road transport. tempi and wspdi were used instead of the daily average values meantempi and and meanwspdi as the instantaneous readings at each data point collected would give greater granularity.

UNIT – ridership would be affected by location of station i.e. some stations located near densely populated areas or offices would see higher ridership

hour – ridership could be expected to peak at different hours of the day e.g. In the morning and evening as people commute to work or during mealtimes when people travel to eat

DATEn and day_week – there could be higher subway ridership on special events or holidays or on weekdays when more people need to commute to work.

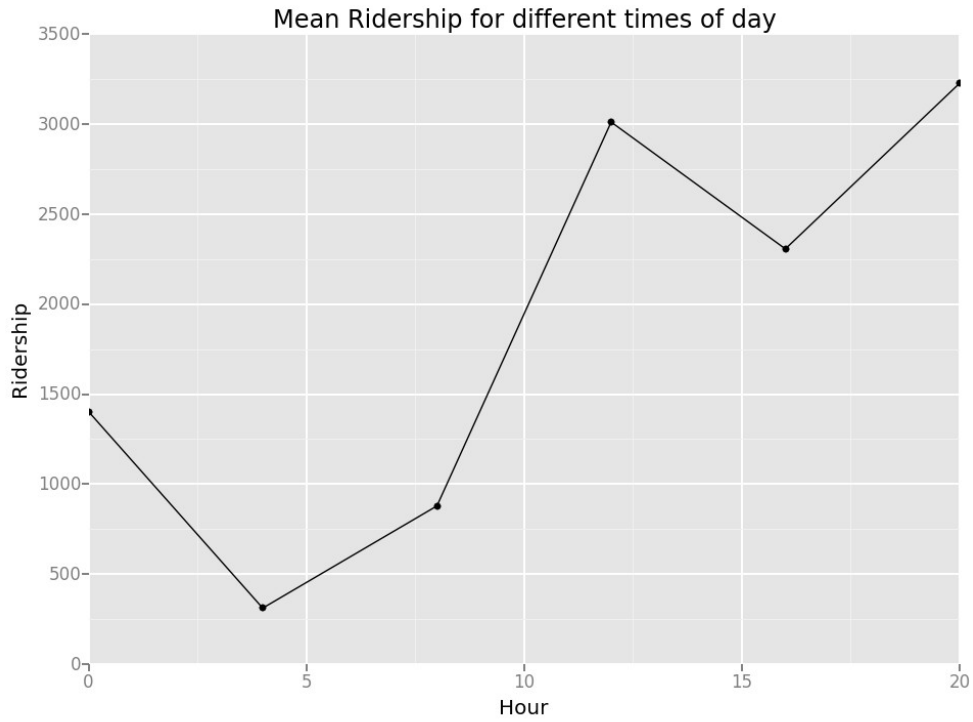
Model Coefficients:

hour	107.253112
tempi	44.420680
wspdi	9.469858
rain	64.824156
day_0	-266.574014
day_1	-127.093989
day_2	123.916797
day_3	128.398753
day_4	69.701561
day_5	-696.371930
day_6	-888.323439

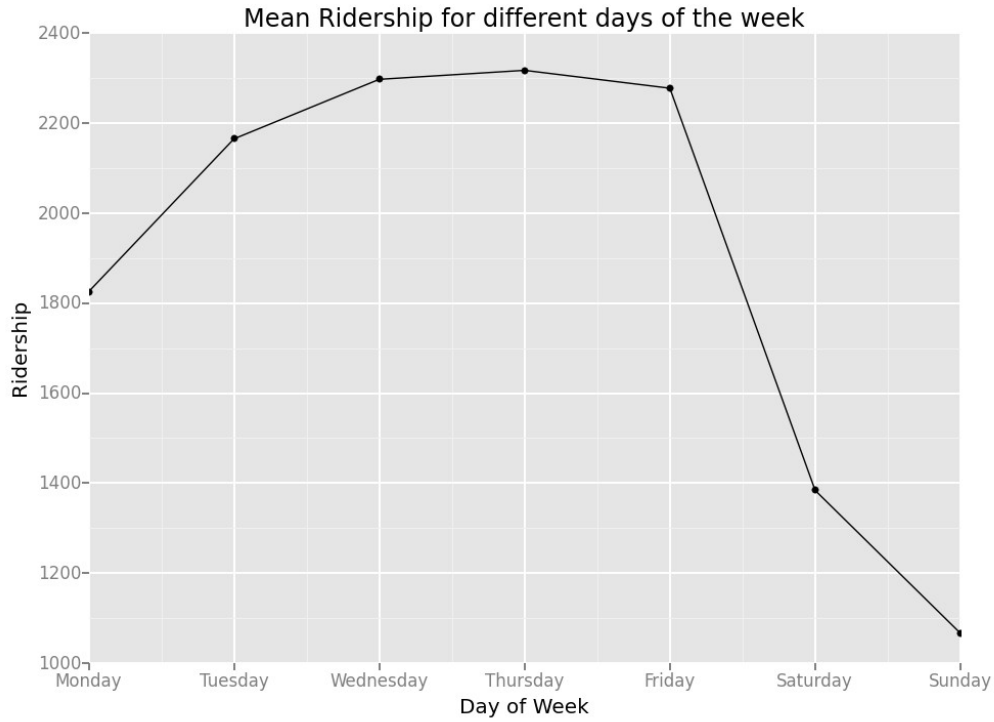
R2 value - 0.495

At R2 = 0.495 the model explains 49.5% of the variance seen in the data. Given this relatively low R2 value, this linear model may not be suitable for predicting ridership.

Visualization (see visualization.py)



The mean ridership varies for different hours of the day. There are peaks at 12 and 20, which could correspond to lunch time and the end of the work day which would in turn correspond to more people commuting and higher subway ridership. Interestingly, ridership is not as high at Hour 8, which would correspond to the start of the workday. This could mean that people start work at varied times between 8am and 12 noon, so ridership could be distributed over an interval for the morning rush, rather than concentrated at a specific time. As the data only captures ridership at 4-hourly intervals, we are unable to comment on how ridership would vary between 8 am and 12 noon.



Ridership drops dramatically on Saturday and Sunday. The higher ridership on weekdays could be due to people having to commute to work or school, whereas weekends, which would be 'off' days for most, would see lower ridership.

Conclusions

Ridership is affected by rain, hour, day of the week.

The statistical test comparing ridership on rainy days vs non-rainy days is significant (with p-value < p-critical) and shows that the likelihood of the group with rain (R) having a higher ridership is not equal to the likelihood of the group without rain (NR) having a higher ridership i.e. P(r > nr) ? 0.5. The with_rain_mean was also calculated to be higher than the without_ rain_mean. Taken together, it appears that more people ride the subway when it's raining.

Linear regression also showed positive correlations between rain and ridership.

The trends showed in the visualization show that ridership varies with hour and day of the week and this corresponds well with the positive correlation obtained by linear regression.

Challenges

The dataset only captures data at 4-hourly intervals, so important trends in ridership could have been missed.

The data set has many different stations, which makes it difficult to fit a model. An alternative could have been to assign the stations to groups based on their locations and use the assigned groups for linear regression.

Many variables were related such as rain/precipi and day_week/weekday. By using only linear regression and calculating R2 value, it was difficult to choose variables to be used. A different model could be tried, or the data set could be split into test and cross-validation sets in order to calculate the error of the predictive model in order to select the best variables.