The purpose of this project is to explore whether weather conditions make a sizable impact on the ridership of NYC Subway System.

SECTION 1. STATISTICAL TEST

1.1 Which statistical test did you use to analyze the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

The Mann-Whitney U Test was used to determine whether group's medians are statistically significantly different. The null hypothesis is that both populations are equal, ea. there is no significant difference in populations' medians. The two-tail P value was selected in accordance with the null hypothesis. In this test we use the significance level of 0.05.

1.2 Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

The choice of statistical test type is based on the characteristics of the data. The datasets for both rainy and non-rainy days are not normally distributed. The data in both groups are independent of each other, and the responses are ordinal.

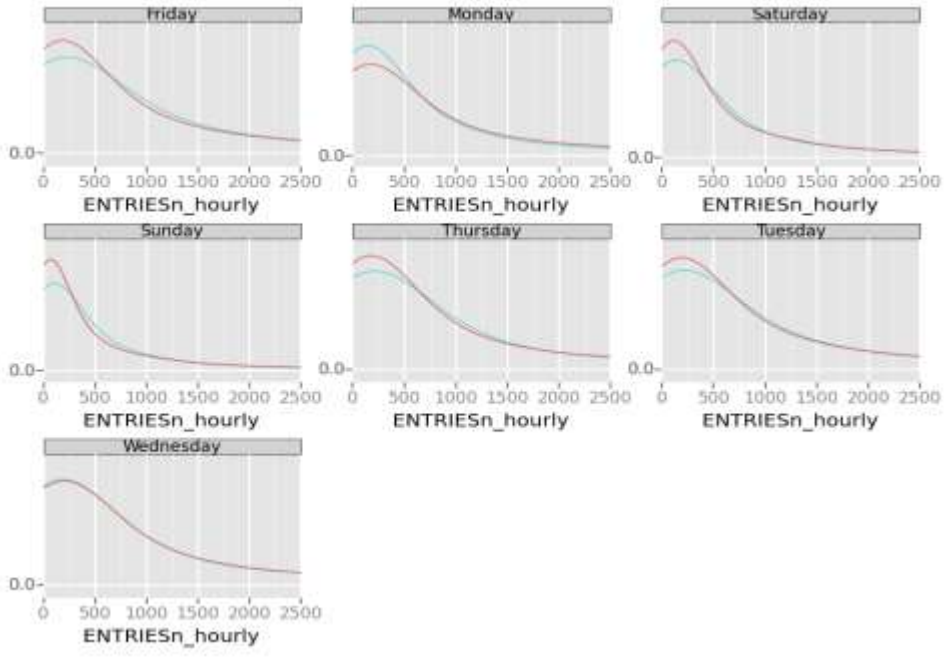

Figure 1. Entry per hour during rainy and non-rainy days

1.3 What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.

Mean of rainy days	Mean of non-rainy days	U Value	P for two-tail test
1105.45	1090.28	1924409167.0	0.039

Table 1. Results of Mann-Whitney U Test

1.4 What is the significance and interpretation of these results?

There was a statistically significant difference between the two data sets. Because of that we have to reject the null hypothesis that there is no impact on the ridership.

SECTION 2. LINEAR REGRESSION

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model?

The Gradient descent and OLS methods were chosen to model the data.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features? Why did you select these features in your model?

The variables *Hour, weekday_or_holiday, rain, Unit*, *meantempi, and mintempi* are used in the model. This selection is based on everyday life observations. As Figure 1 shows, there are different patterns of ridership on weekdays and weekends. The time of day makes a big difference too. It seems that weather related variables might impact on riders' decision to use a subway as well.

2.5 What is your model's R2 (coefficients of determination) value? What does this R2 value mean for the goodness of fit for your regression model?

After many iterations, we were able to raise R2 value for Gradient descent to 0.469. This value is relatively low, that means that the model will not be very successful at predicting.

Obtained R2 value for OLS using Statsmodels is a little bit higher, 0.570. It means that the model will predict NYC subway ridership with 57% accuracy.

Dep. Variable:	ENTRIESn_hourly	R-squared:	0.570
Model:	OLS	Adj. R-squared:	0.567
Method:	Least Squares	F-statistic:	186.0
Date:	Tue, 10 Mar 2015	Prob (F-statistic):	0.00
Time:	17:41:12	Log-Likelihood:	-1.1551e+06
No. Observations:	131951	AIC:	2.312e+06
Df Residuals:	131018	BIC:	2.321e+06
Df Model:	932		

Table 2. OLS Regression Results

SECTION 3. VISUALIZATION

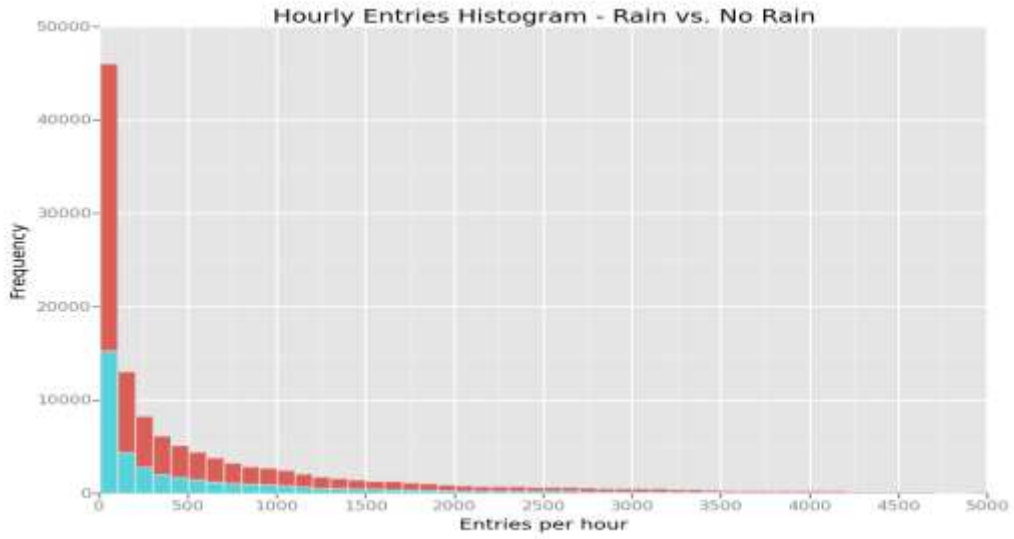

Figure 2. Entries per hour on rainy and not rainy days

The histogram of hourly entries shows us distribution of ridership on rainy and non-rainy days. As we can see, the shape of the distribution is not a "*bell* curve" of a normal distribution.

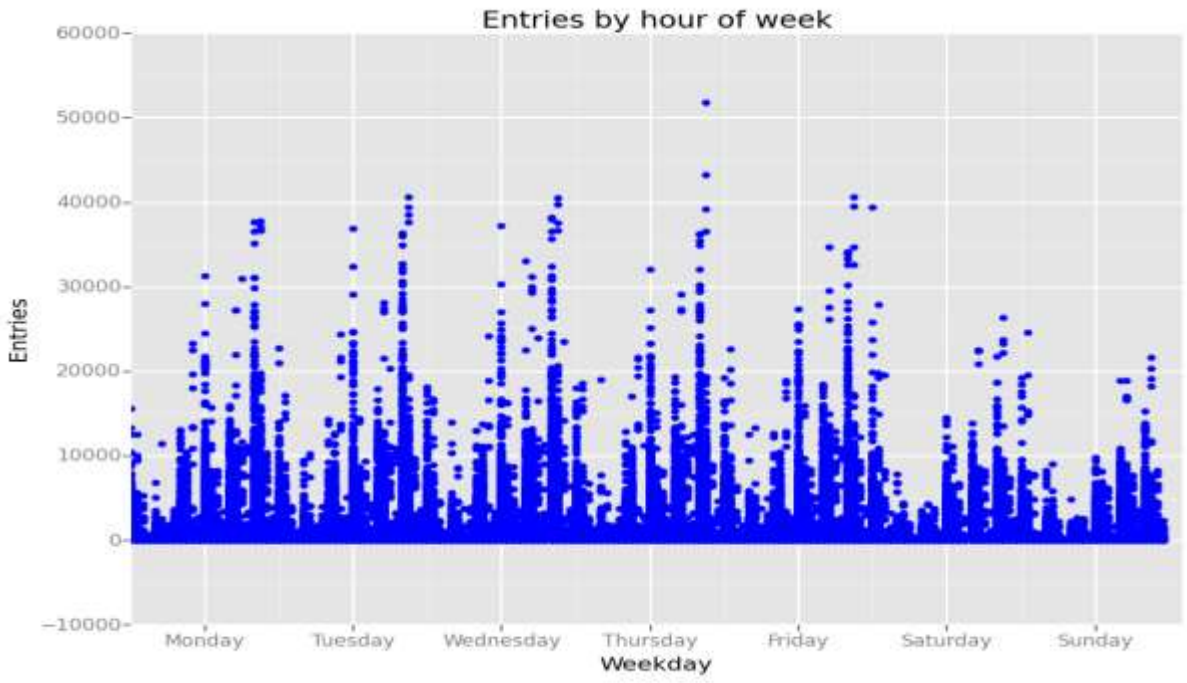

Figure 3. Entries by the hour of week

The Figure 4 demonstrates the weekly pattern of NYC subway ridership with noticeable spikes of rush hour traffic and lower numbers of the weekend riders. The afternoon spikes in ridership might be explained by 4 hour time periods between data recording. That makes a morning traffic to be recorded as an afternoon traffic.

SECTION 4. CONCLUSION

The analysis of the provided dataset shows that the NYC subway ridership slightly increases on a rainy day. The results of the Mann-Whitney U Test provide us with a reason to believe that rain makes an impact on NYC subway ridership.

5. Please discuss potential shortcomings of the methods of your analysis.

With the dataset that contains 31 days of data with only 10 rainy days we choose to use the same data points for building the model and for testing it. Having the separate testing set would let us see if the model performs similarly well on data that was not used to build it. Also, it would be interesting to run the same test on a dataset combining data recorded during other seasons.

The more detailed data, like hourly entries combined with weather conditions for the same period of time, I think, could significantly improve the predictive power of our models. In the current dataset day might be marked as rainy even if it was raining for a short period of time, so rainy time data might be diluted.

https://statistics.laerd.com/premium-sample/mwut/mann-whitney-test-in-spss.php

https://github.com/glamp/ggplot-tutorial

http://statsmodels.sourceforge.net/devel/index.html

http://people.duke.edu/~rnau/rsquared.htm