Project Number 2

Analyzing the NYC Subway Dataset
Answers to the Short Questions

Section 0: References

- Information to Pandasql: https://github.com/yhat/pandasql/blob/master/README.md
- Mann-Whitney U test: Udacity Paper
- Coefficient of Determination: R^2: https://en.wikipedia.org/wiki/Coefficient_of_determination; http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
- Gradient Descent: Udacity Paper
- MWU test: https://en.wikipedia.org/wiki/Mann–Whitney_U_test#Assumptions_and_formal_statement_of_hypotheses

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 (MWU) test (non parametric) was used because after plotting the subway data to compare the use of the subways in NYC on rainy and non rainy days, the distributions of the data did not look bell shaped (Normal Distributions). In addition the sample size was not equal. So the MWU test was needed. It shows whether one distribution of two different population is more likely to generate higher values. For two samples x,y from different populations X,Y the Null Hypothesis is:

Null Hypothesis: The probability of an observation from the population X which exceeds an observation from the second population Y **is equal** to the probability of an observation from the population Y which exceeds an observation from the population X.

$$P(X>Y) = P(Y>X)$$

Alternative Hypothesis: The probability of an observation from the population X which exceeds an observation from the second population Y **is different** (two tailed test) to the probability of an observation from the population Y which exceeds an observation from the population X.

$$P(X>Y)\neq P(Y>X) \text{ (Two Tailed)}$$
$$P(X>Y)>P(Y>X) \text{ (One Tailed)}$$

The P value is two tailed and the value after calculating it with python is: 2.5% *2 = 5%
My critical Alpha level is 5%, so the result is statistical significant.

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 MWU-test is applicable because of the reasons mentioned it 1.1. The data does not seem normal distributed. As mentioned in the Udacity paper about the MWU test, the Null Hypothesis is True if the most common assumption, that the distributions are equal, is true.

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.

means of ENTRIESn_hourly on rainy days:	1105
means of ENTRIESn_hourly on non rainy days:	1090
U-value:	1.9* 10^9
p-value (one tailed):	2.5%
p-value (two tailed):	5%

1.4 What is the significance and interpretation of these results?

Concerning the p-value of 5% and our Alpha criteria level of 5% one can reject the Null-Hypothesis and conclude that it is more likely to get a different result of “Entriesn_hourly” when comparing two samples of this different populations.

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:
- OLS using Statsmodels or Scikit Learn
- Gradient descent using Scikit Learn
- Or something different?

I used the OLS method with the Statsmodels in Problem-Set 3 Exercise 5 Linear Regression.
And in addition the Gradient descent method on Exercise 8 with SciKit Learn

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

Features:
For method OLS: (I used mostly the improved data set)

fog,precipi,hour,rain,tempi,weekday,wspdi,latitude,longitude,day_week,Minute,
*pressure*i

For the Gradient Descent method:

rain, precipi, hour, fog, weekday, wspdi

Both times dummy variables [‘UNIT’] were used

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that

the selected features will contribute to the predictive power of your model.

Your reasons might be based on intuition. For example, response for fog might be: “I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often.”

Your reasons might also be based on data exploration and experimentation, for example: “I used feature X because as soon as I included it in my model, it drastically improved my R^2 value.”

Table 1

Feature	Reason
Fog,Rain,pressurei,tempi,	All this features were used because I thought these could influence the behaviour of people using the subway more often if it is for example raining or too warm. NYC has a lots of tourists thats why I also think the tempi correlates with the numbers of tourists going into the city and using the subway
Location(Lat,Lon)	This features might indicate whether people with different incomes, comparing for example Bronx and Upper Manhattan, tend to different useage of the subways. Moreover the location could have an influence whether people can just walk to work (living the the city centre) or depend on the subway to reach their working place
day_week, Minute	These features reflect whether people use the subway more often during the week to get to work. Moreover the Minute or especially the Time underline working hours of companys and might have an influence
rain, day_week	These values improved the R^2 value significant

2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?

Coefficient of the OLS model:

```
fog         -3.597404e+02
precipi     -3.318035e+03
hour        -3.002576e+13
rain         4.598954e+01
tempi        2.834999e+00
weekday      1.292525e+03
```

```
wspdi        1.805364e+01
latitude     7.761600e+10
longitude    4.358225e+10
day_week     8.787949e+01
Minute       3.002576e+13
```

Coefficient of the Gradient Descent model:

```
rain:       73.52032676
percipi:  -3593.40043903
hour:       121.52490011
fog:       -488.76946145
weekday:    1005.19681343
wspdi:      23.23729544
```

2.5 What is your model's R^2 (coefficients of determination) value?

OLS Model: R^2: 0.484
Lineare Descent: R^2: 0.454

2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

The R^2 value also known as Coefficient of Determination is a indicator for the accuracy for the regression. It is a statistical measure how well the data fits the regression line. It is also referred to the R^2 value as

R^2= explained variation/total variations

The R^2 value show how much variability of the response data around it's mean is explained by the Regression model.

A value of around 0.5 as in our calculations is not bad. In addition the R^2 value does not mean, that our regression is inadequate. You can have a low R^2 but a good regression model that fits your data because the regression optimisation process (as I need when trying different features and looking at the R^2 value and how it was influenced) is influenced by chance correlation and is kind of biased response.

Moreover in the field of human behaviour R^2 values tend to be low, because humans are harder to predict. Regrading our project, we can only guess why somebody decides against or pro subway use.

In addition by plotting the Cumulative Probability Curve one can see that the residuals don't seem to follow a normal distribution, especially the long tails. So our regression model is not a good estimate for the predict ridership.

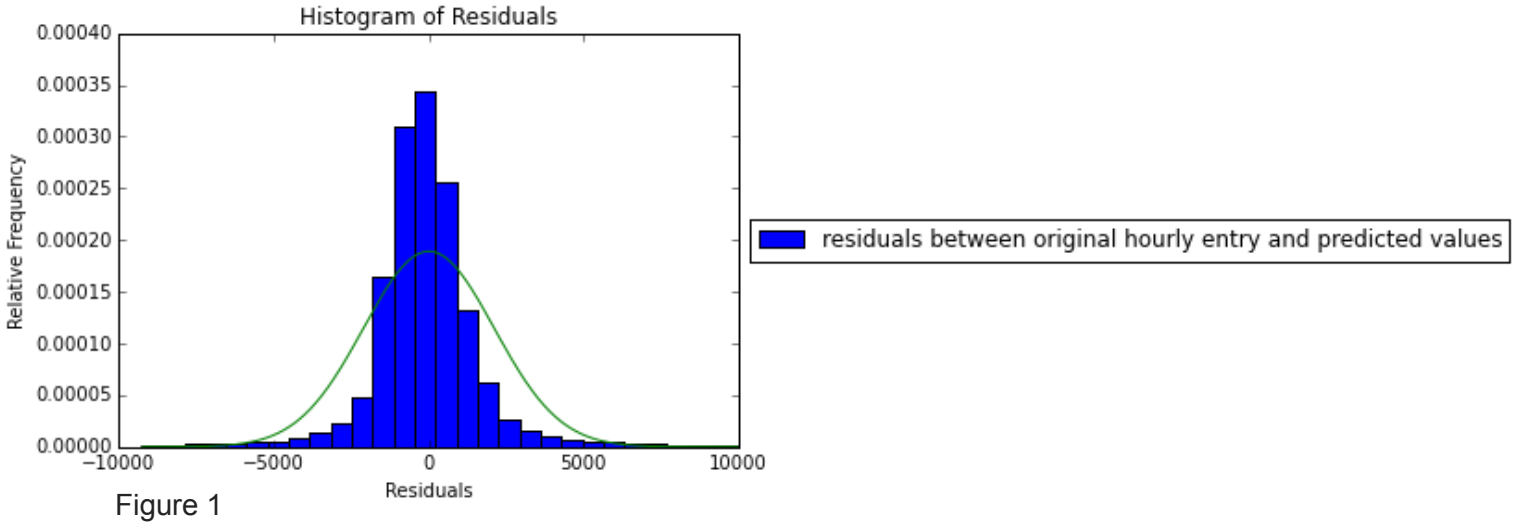

Figure 1

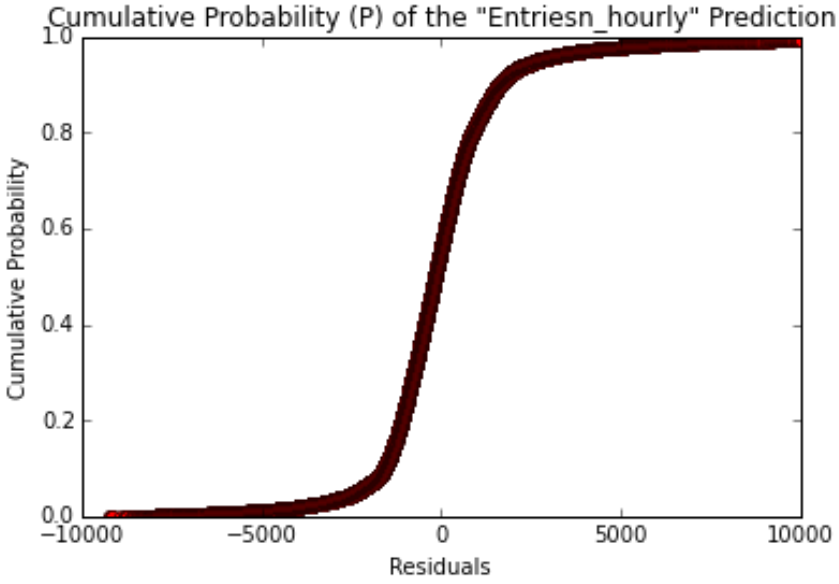

Figure 2

Section 3: Visualization

3.1 Histogram Rain/Non rainy days

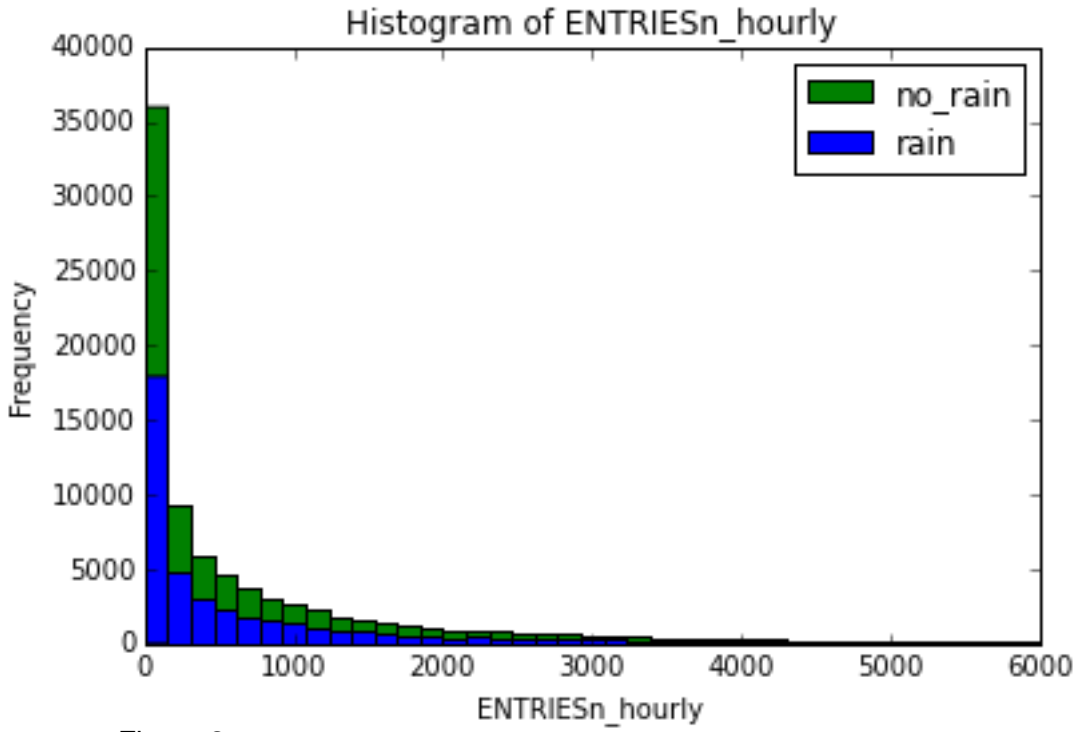

Figure 3

The two distributions of the samples don't correspond with a normal distribution, thats why we used the Mann-Whitney-U test. The histogram also reveals that the sample size of 'rainy days' was smaller. The Frequency bars for non rainy days are much higher.

3.2. Other visualisations:

Figure 4 depicts the Entriesn_hourly regrading the subway station and the time of the day. I only used 10% of the dataset here because otherwise there would have been to many station and data to get some information out of the chart. One can see that there is a difference between the subway stations and the Entriesn_hourly. I used the colour distinction to give a rough overview, to highlight the significant of the location in correlation to the subway entries. The stations depicted with green colour tend to have more Entriesn_hourly that the pink ones (i.e. Stuffing Blvd, Wall St). The Time of the day has also and impact on the amount of people using the subway. We have the highest values of Entriesn_hourly at 12:00 pm and 8 pm (20:00). Maybe at 12 pm people have their lunch break at work and around 8 pm people

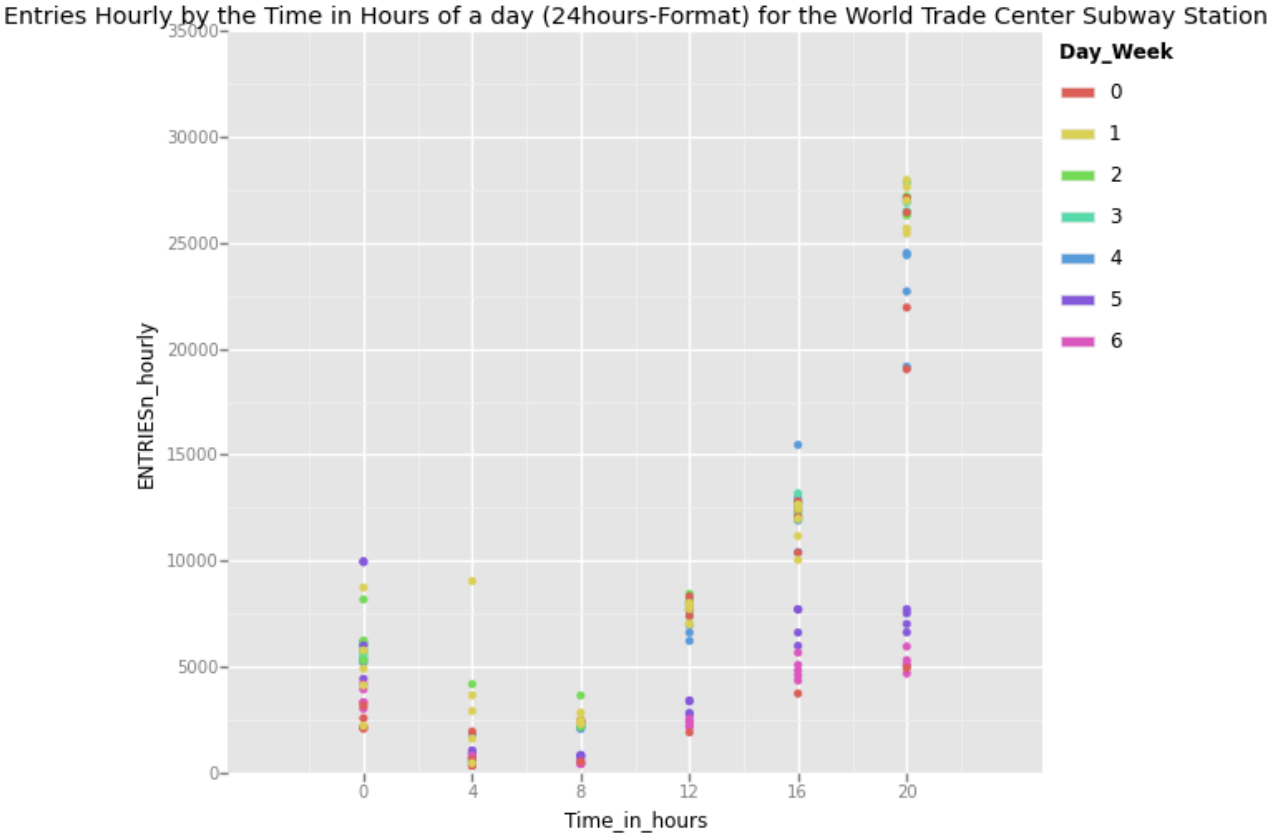

Figure 4

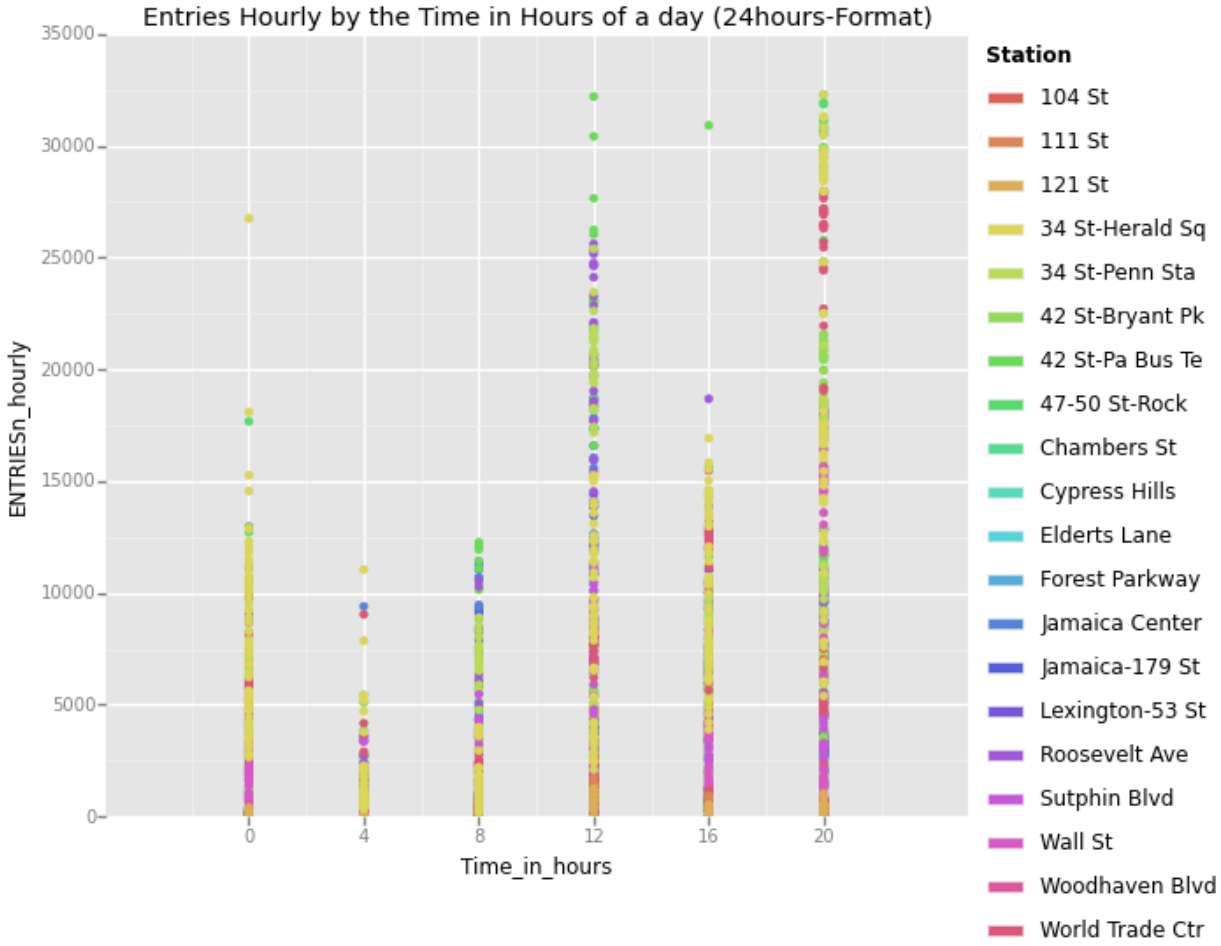

Figure 5

In Figure 5 only the data for the World Trade Ctr station is plotted and one can see that it is very hard to make some conclusions about the ridership during the week. On weekends people take less the subway from 12 to 8 pm than during the week. Moreover most people take the subway around 4 pm (16:00 o'clock) to 8 pm than in the morning

Section 4: Conclusion

4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

The analysis started with plotting the distribution for the rainy and non rainy days. The Mann-Whitney U test was used to find a correlation between the two samples and their population. We found out that there is a difference between the number of people using the subway on rainy and non rainy days. A two tailed test was used, so we know that the probability that x (number of entriesn_hourly on non rainy days) tend to be bigger than y (number of entriesn_hourly on rainy days): We rejected the Null-Hypothesis and stayed with the Alternative one:

Alternative Hypothesis: $P(x>y) \neq 0.5$

But one can only say the number of people who ride the subway differs when it is raining or not, regarding our statistical analysis.

The Linear Regression showed that a good model, which fits the data depends on more than one feature. For the OLS model 11 features were used to generate the regression. Comparing the R^2 value of the OLS models with the feature rain and without showed no significant difference in the regression model (It stayed the same 0.484 (rounding to 3 decimals). This fact underlines that the rain feature has not that much influence on the variation of the data. Regarding the coefficient (weights) in Question 2.4 shows, that the "Time" feature is more significant than the "rain" one

The Gradient Descent method used 6 features. Including the rain feature increased the R^2 value of 5%.

So the Linear regression model underlines that the rain has a small influence on the numbers of people using the subway in NYC, but there are more important criteria especially the Time of the days.

Section 5: Reflexion

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

- Dataset,
- Analysis, such as the linear regression model or statistical test.

One shortcoming when analysing the data concerning the influence of rain on the NYC subway usage was the huge difference of the sample sizes. It was hard to find some interesting facts using a visualisation.
Generally speaking this experiment corresponds a lot with the human psychology, because there might be so many different reason why a person won't or will use the some. So many individual reasons, thats why it is better to look at a subset of people, for example compare employed to unemployed people or the different station corresponding to the companies which are near by. So there are so many more facts than just weather factors.

The Linear Regression especially the R^2 value is a controversial factor in the data analysis (as mentioned in Question 2.6) especially when the behaviour of a human being is involved. The linear regression might still be good, although the R^2 value is very low.

Referring to the statistical test the Alpha criteria can be set smaller. When one limit it to 1 % the analysis would have kept the Null-Hypothesis. So maybe one should connect the Alpha Criteria to the size of the dataset. In this case where the two sample sizes differ that much one should limit the Alpha criteria to the smallest kind available to avoid any variation by chance.