Section 0.
References

I used the following links:

http://discussions.udacity.com/t/short-question-3-1-totally-lost/13472/3

http://discussions.udacity.com/t/2-2-sql-querry-quick-formatting-problem/13070/2

http://discussions.udacity.com/t/5-3-mapper-fine-reducer-not-printing-anything/12913

http://discussions.udacity.com/t/5-2-ridership-by-weather-problems-understanding-the-key-function/12393/8

http://discussions.udacity.com/t/titanic-3-custom-filter-any-analytical-approach/12437/3

http://discussions.udacity.com/t/5-1-mapper-and-reducer-problem-any-pro-can-help/11425/8

http://stackoverflow.com/questions/28971058/mapreduce-easy-probably-doing-a-udacity-course-any-help-from-a-pro

http://stackoverflow.com/questions/28848516/python-writing-content-of-2-files-into-1-file

http://pandas.pydata.org/pandas-docs/stable/visualization.html#histograms

Besides that I watched a lot of YouTube videos and read udacity forum posts on different topics needed for the course.

Section 1.
Statistical Test

1.1. 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?

A:

I used a two sided Mann Whitney U-test since it is unclear in what direction how rain will affect the ridership.

The null hypothesis asserts that the medians of the two samples are identical. This does not have to mean that both distributions have to be identical.

My critical p value is:

p= 2 *$0.024999912793489721 \approx 0.05$

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.

A: The Test is applicable because it does not require that the data points have any underlying probability distribution. In Contrast, such an underlying probability distribution is necessary for a t-test.

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.

A:

Means on days where it does not $\approx$ 1105.44

Means on days where there is rain $\approx$1090.27

P$\approx$0.02499

1.4 What is the significance and interpretation of these results?

A:

The calculated p-value is in the middle of the threshold values given in the documentation (0.05 and 0.01). The U-value is very large which would point in the direction of the null-hypothesis not being true. A further hint for that is the difference between the means on rainy and nonrainy days.

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:

1. Gradient descent (as implemented in exercise 3.5)
2. OLS using Statsmodels
3. Or something different?

A: 1. I used Gradient Descent for my regression model.

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

A:

The following features where included in my model:

'rain', 'Hour', 'meantempi','fog'

Unit was added using a dummy variable as a feature.

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."

A:

I chose fog as a feature for the reason that a fog might persuade people not to use their car and choose the subway instead.

Along the same lines I chose Meantempi and rain, since those features could convince people not to wait at a station where it might be cold. Furthermore a rainy day might lead them to not want to walk to the

Choosing "Hour" as a feature seemed obvious due to the strong influence on ridership. It increased the value of r^2 from 0.42 to 0.46.

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

'rain', 'Hour', 'meantempi','fog'

```
-1.81385128e+01   4.68385118e+02  -7.31884345e+01   6.27692258e+01
```

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

R^2 ≈ 0.4644

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?

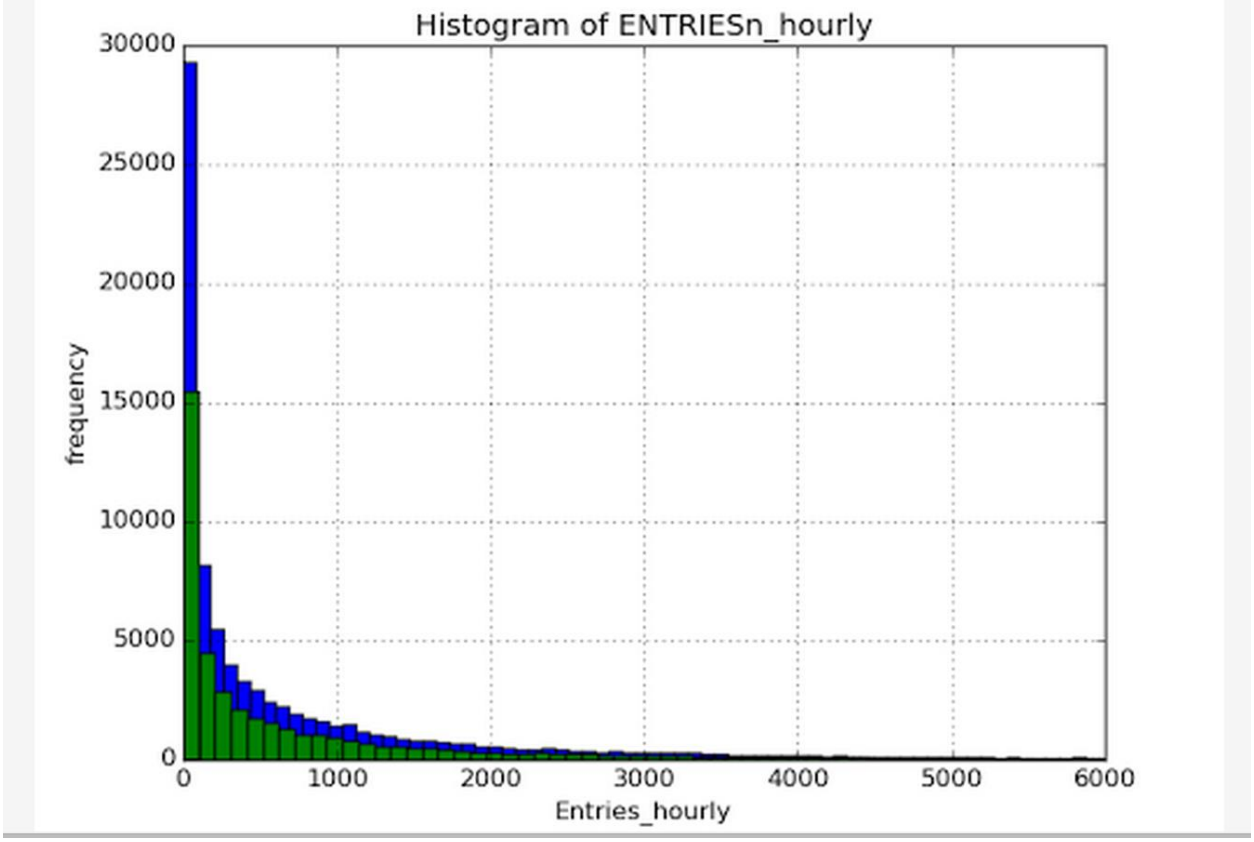
Since it is well above 0.2 (the value mentioned as a threshold) I think the value is decent enough to draw conclusions from it and view it as significant.

I think a linear model is possible since I don't see any effects that would immediately qualify as having a higher or different exponent and thus making a linear model not appropriate.

Still, a value of 0.4644 indicates that 46.44% of the original variability is explained, leaving around 53% of residual variability.

Section 3.
Visualization

3.1.

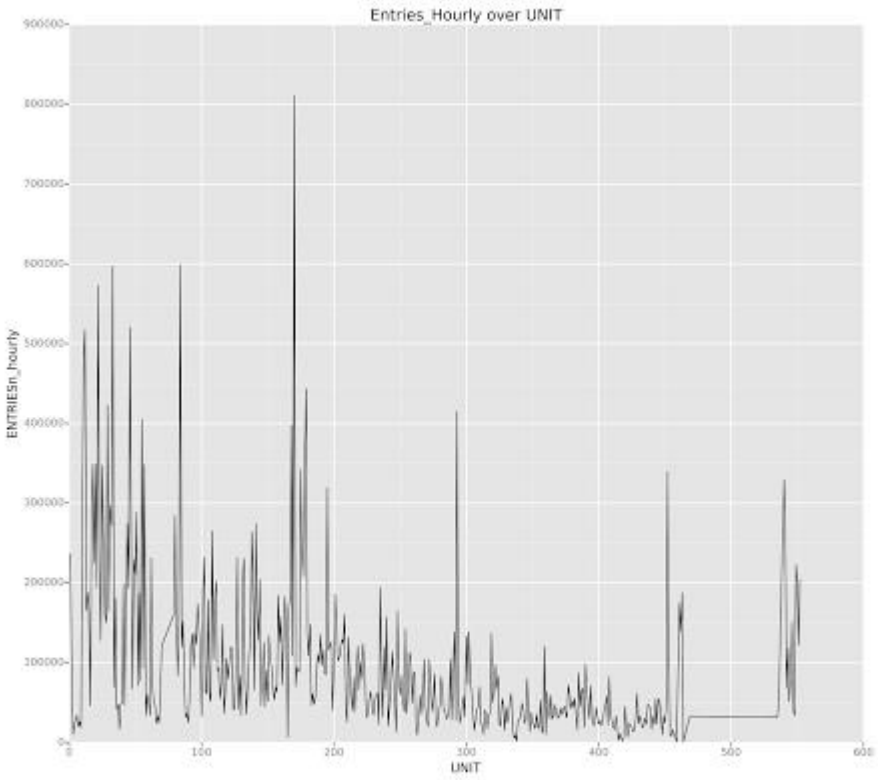


The blue bars are the values for the days where it did not rain.

The green bars are the values for the days it did rain.

3.2.

Original is 4.2. In my work:



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?

A:
The results indicate that more people ride the subway on days without rain.

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.

A:

As described in the answer to question 1.4. the Mann Whitney test indicates that more people ride the subway on days without rain. Factoring in the means of both cases (rain and no rain). It is very probable that there is a significant difference on ridership in the context of being a rainy day or not.
Taking the regression model into account, the value for the coefficient for the feature rain is negative (-1.81). That further suggest that there is a higher number of people riding the subway on days without rain.

Section 5. Reflection

5.1.

In my opinion one of the weaknesses of the used dataset is its relatively small size due to a very limited timespan. The Dataset only includes entries from the month of May. It would be easy to imagine that the ridership numbers change during the winter for example, where there are a lot less construction and seasonal workers going to work.

A shortcoming of the linear regression model that was used, was that it was not guaranteed to find the absolute maximum. Even when the step size (alpha) is sufficiently small, we can only say we certainly found a local minimum. To find the global minimum we had to randomize our starting point sufficiently so with several regressions we could be reasonably sure, that we found the global minimum.

Another even more general problem with the linear regression model is nature of the model itself. The model assumes a priori a linear relationship between its entries and the value that is analyzed. It is easy to imagine situations where an entry has a nonlinear relationship, because for example it asymptotically nears itself to a limit and has a logarithmic shape.

I think that is why it is always essential that before analyzing data with the various tools described in the lectures to take a good look at the data itself and think about what someone should expect from the analysis and of the logical connections in the data. Blindly just executing algorithms on the data seems not very efficient and is leading to wrong conclusions at least some of the time.