Data Analyst Nanodegree
Project 1
Short Questions

Submission #2
Author: Konstantin Kunz
Date: 2015-04-22

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?
I used the Mann Whitney U-test. In this test I used the two-tail p-value.
The p-critical value is 0.05.
The null-hypothesis can be formulated as:
H_0: P(sample_no.rain > sample_rain) = 0.5

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.
This statistical test is applied to this dataset because the distribution of ridership with and without rain is not normally distributed. This calls for a non-parametric statistical test, such as the Mann Whitney U-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.
U-statistic = 1924409167
p = 0.02499991*2 = 0.04999982 (two-tail p value)
mean ridership with rain = 1105
mean ridership without rain = 1090

1.4 What is the significance and interpretation of these results?
The p value of 0.04999982 is just slightly below the p-critical of 0.05. This shows us that the difference of the two samples is of statistical significance allowing us to reject the null-hypothesis.
It can be assumed that it is most likely not a random effect that the two samples have different means. This means that most likely more people take the metro when it rains as when it doesn’t rain.
Note that the difference of ridership was not big. This also reflects in the p-value beeing very close to the p-critical value.

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?

I used the gradient decent method to compute the coefficient theta.

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

Features:
Rain
Precipitation
Hour
Mean wind speed

Of which were dummy variables:
UNIT

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.”

One basic intuition was that the ridership of the subway will vary with the climatic conditions. If it is raining or snowing (precipitations) more people will take the subway rather than walk or take a car (in case of snow). Also the hour (time of day) plays a big role in predicting ridership as there are surely effects of rush-hour. The UNIT feature takes care of the different ridership of the different stations. There are stations with more and less ridership in general.

Small improvements could be made by adding the Mean Wind Speed. Here the improvement of the prediction was found through exploration of the dataset. An improvement of the R2 value could be observed.

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

Rain: 10.8093935

Precipitation: 7.70462870

Hour: 458.817474

Meanwindspeedi: 68.6889198

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

Your r^2 value is 0.464033975766

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 is quite low when taking in to account that a perfect prediction of all variations in observations gives r^2 = 1 and an absolute wrong prediction of the variations in observations gives r^2 = 0. This means that our predictions are not very accurate but not absolutely wrong. In turn we must say that our model is not very accurate. This could be the case because the ridership is not linear in some variables, especially with respect to time (two peaks, in the morning and in the evening).

Section 3.
Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

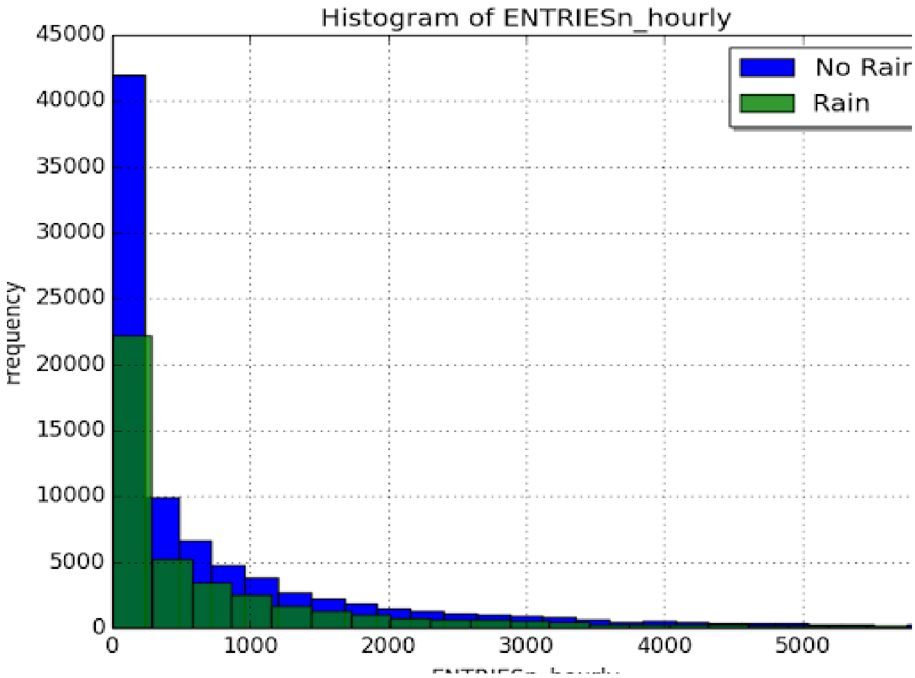
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

- You can combine the two histograms in a single plot or you can use two separate plots.

- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.

- For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.

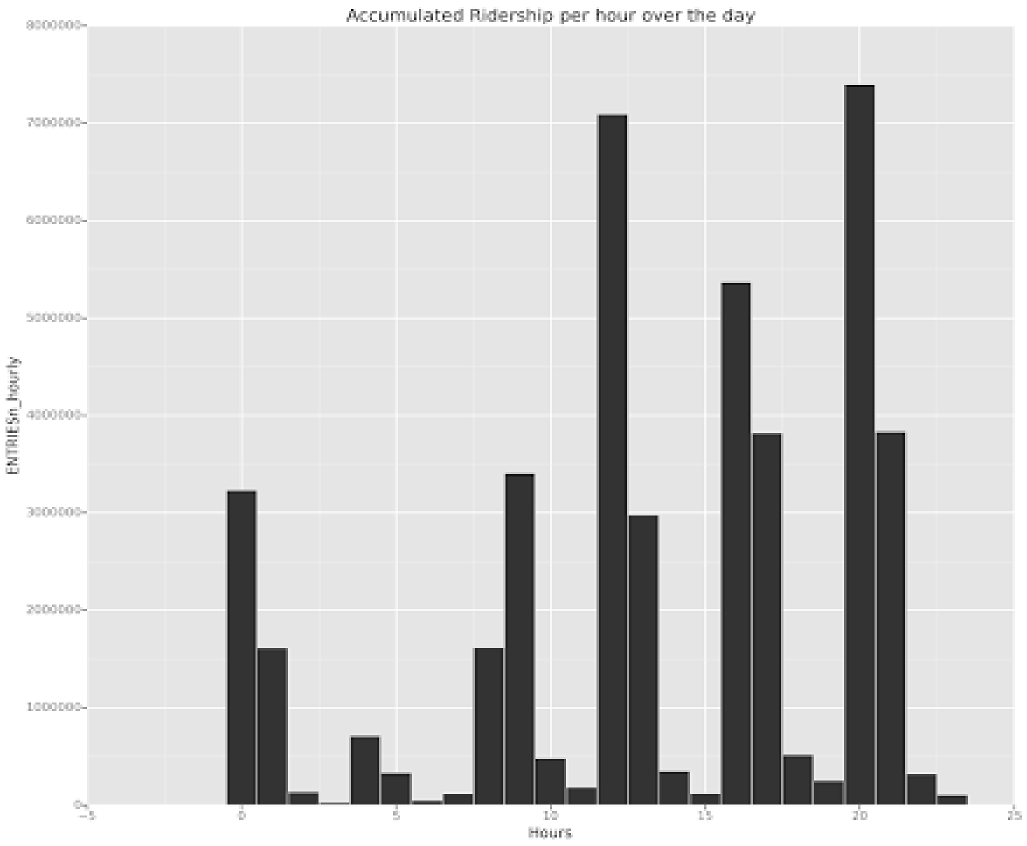
- Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.



This figure shows the occurrence frequency of a number of hourly entries in the NYC subway. Both rainy and not rainy days are compared. One cannot conclude that there are more entries when it is raining or not raining from this figure since the sample amount for rainy days is much smaller than for not rainy days, there are simply more not rainy days in NY than rainy days. One can see that the distribution is not normal.

3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

- Ridership by time-of-day
- Ridership by day-of-week



This figure shows the accumulated number of entries to the NYC subway over the day. The accumulated value is the value of all entries over the entire data set at that time of day. One can see that there are peaks around midnight, the morning rush-hour, noon, afternoon and after 8 pm. This would also coincide with the conclusion that the ridership is not linear with respect to time.

Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.
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?
The NYC subway is used more often on rainy days than on non-rainy days. The mean difference in hourly entries for rainy days vs non-rainy days is (1105 – 1090) 15. Also any kind of precipitation, such as snow, hail, etc) results in a higher usage of the subway system.
https://lh3.googleusercontent.com/gfiy2IaNaxHbedz007KWWbbDbY1hUbIej-8F4HM2ScC64PI6PbK3jYxso3L_xgIwy6BiOvj_5vX11MIVhBo4.2 What analysis lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.
First we performed a Mann Whitney U-test on the ridership samples for rainy days vs non-rainy days. The test concluded that the two samples are from a different population. This means that it is not due to reasons of random selection that the mean ridership on rainy days is higher than on non-rainy days (1105 vs 1090). This is the first insight that the ridership is higher on rainy days than on non-rainy days.

The second analysis performed on the data, was to generate a linear model of the ridership in order to predict the behavior in future scenarios. The following features were used as parameters of the model: rain, precipitation, hour & mean wind speed. When looking at the coefficients of our model with respect to these parameters we can see the effect they have on the ridership in qualitative way. The coefficient for rain is 10.8, and for precipitation it is 7.7. Both values are above 0 and multiples larger than 1 meaning that their effect on ridership is positive. The more it rains, snows, hails in New York the more people take the NYC subway.

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.
5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. Dataset,

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

Potential shortcomings of the analysis:
- The dataset is limited. More data would allow an analysis with greater precision.
- The analysis and data do not show why people actually use the subway more when it is raining. Is it due to worsened traffic, do they not want to walk or are more cabs taken? This is not clear but could give more insight into the topic.
- The linear regression model is a bad fit for the time parameter. The ridership cyclically varies over the day (morning & evening rush-hour, calmer in-betweens) and this cannot be captured with a linear model.

My Sources: (various subentries of these websites)
(1) http://pandas.pydata.org/pandas-docs/stable/groupby.html
(2) http://stackoverflow.com/questions/12555323/adding-new-column-to-existing-dataframe-in-python-pandas
(3) http://docs.ggplot2.org/current/index.html
(4) http://en.wikipedia.org/wiki/Student%27s_t-test
(5) http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
(6) http://www.graphpad.com/quickcalcs/statratio1/
(7) http://www.pythoncentral.io/pythons-range-function-explained/
(8) www.python.org

And various more which I have sadly not noted down during the preparation of this project. I will keep close track of the resources for further projects.