Olaf Wied, Nanodegree “Data Analyst”, 05/14/2015

Thank you for your very helpful and for the interesting “optional” comments.

Analyzing the NYC Subway Dataset
Questions
Overview

Section 0. References
Please include a list of references you have used for this project. Please be specific - for example, instead of including a general website such as stackoverflow.com, try to include a specific topic from Stackoverflow that you have found useful.

- Python for Data Analysis by Wes McKinney, Publisher: O’Reilly Media, 2012

- I used http://stackoverflow.com/questions/19991445/run-an-ols-regression-with-pandas-data-frame on how to use the statsmodel package for linear regression

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?

Mann-Whitney U-test. Two-sided P-value. The null hypothesis: The two samples come from the same distribution (are the same). P-critical value (two-sided): (common default value) 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.

A histogram quickly shows that the data is not normally distributed. However, the Mann-Whitney-U is non-parametric and works also for non-normal data.

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.

p: 2*0.0249999 = 0.0499998 < 5%, mean for rainy days: 1105.4463767458733 mean for non-rainy days: 1090.278780151855

1.4 What is the significance and interpretation of these results?

The “U-test” gives a p-value below the critical p-value (if assumed 5%) and we can reject the null hypothesis, implying that the distribution of the ridership on rainy days is statistically different than the one on non-rainy days. Together with the higher mean on rainy days this could be a hint that there might be more subway rides on rainy 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:
Gradient descent (as implemented in exercise 3.5)

OLS using Statsmodels
Or something different?

OLS using statsmodels.

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

UNIT (dummy) and Hour.

```
df = pd.read_csv('turnstile_data_master_with_weather-2.csv',header=False,index_col=0)

linreg = sm.ols(formula = "ENTRIESn_hourly ~ Hour + UNIT",data=df).fit()
```

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.”

I used Hour because Hour and ENTIRESn_hourly showed a correlation factor of around 0.18 (>> 0.0) which was higher than most of the other combinations. UNIT as a proxy for stations and so for the location of the entry was used as it significantly improved the R-squared. I did not use weather categories as they seemed to have little impact on R-squared/the goodness of fit.

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

67.3948

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

0.458

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?

R-squared is the ratio of explained variation over total variation. Therefore, goodness of fit is best for R-squared = 100% = 1 and lowest for R-squared = 0. A value of 0.458 is not very convincing as more than 50% of the variability around the regression line remains unexplained. However, low R-squared are often found with real-world data. With a value of 0.458 the linear regression should not necessarily give accurate predictions but we can still draw conclusions

about the linear relationship between input and output variables (for example is it positive or negative). I would call it appropriate.

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.

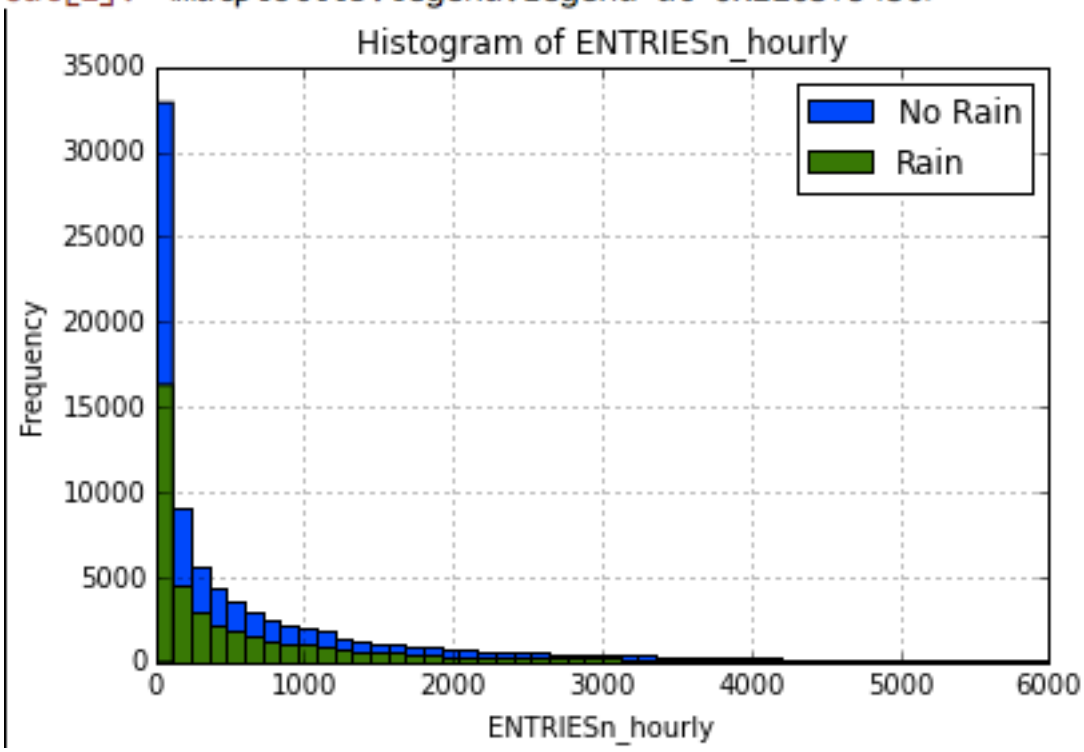
Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.

```
import matplotlib.pyplot as plt

df = pd.read_csv('turnstile_data_master_with_weather-2.csv',header=False,index_col=0)
rain = df[df['rain']==1]
norain = df[df['rain']==0]

plt.figure()
norain['ENTRIESn_hourly'].hist(bins=50,label='No Rain',range=[0,6000])
rain['ENTRIESn_hourly'].hist(bins=50,label='Rain',range=[0,6000])
plt.title('Histogram of ENTRIESn_hourly')
plt.xlabel('ENTRIESn_hourly')
plt.ylabel('Frequency')
plt.legend()
```

```
Out[2]: <matplotlib.legend.Legend at 0x1105f9450>
```



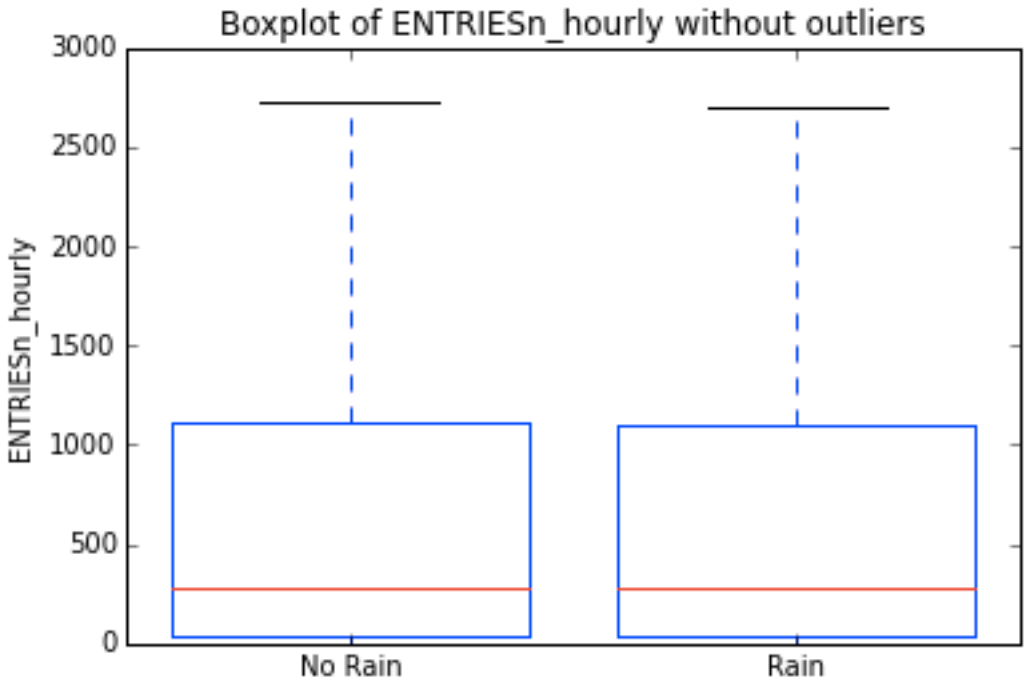
3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

Ridership by time-of-day

Ridership by day-of-week

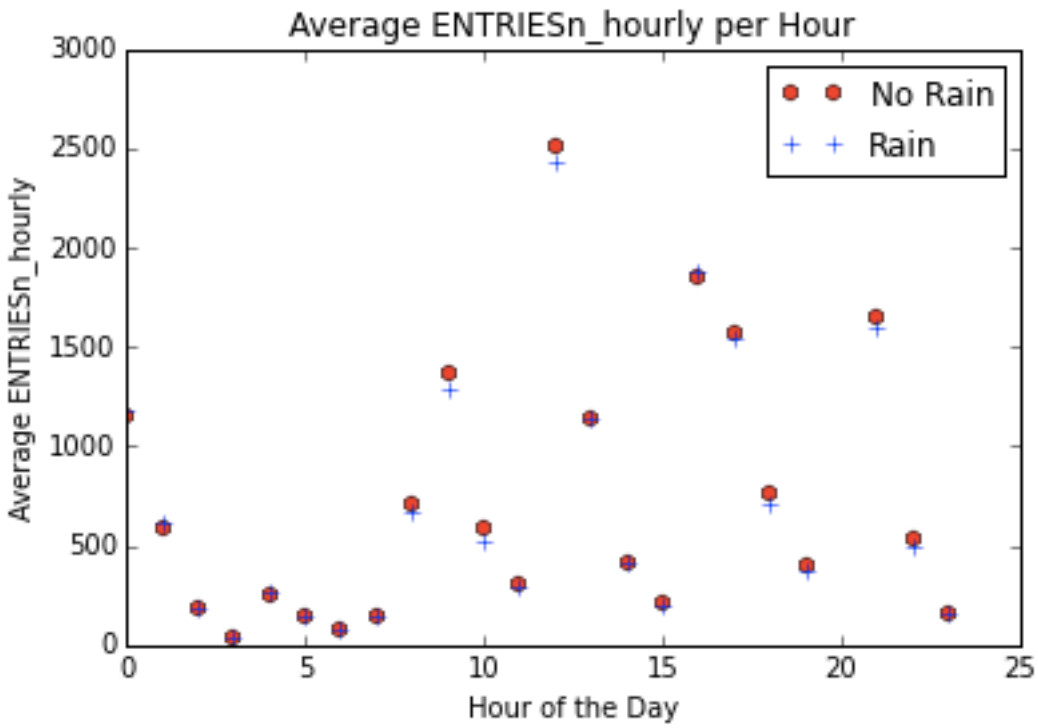
Boxplot of ENTRIESn_hourly per Hour (Rain vs No Rain) without outliers for more clarity:

```
fig = plt.figure()
ax = fig.add_subplot(111)
bp = ax.boxplot([norain['ENTRIESn_hourly'],rain['ENTRIESn_hourly']],sym='',widths=0.8)
ax.set_xticklabels(["No Rain","Rain"])
ax.set_ylabel("ENTRIESn_hourly")
ax.set_title("Boxplot of ENTRIESn_hourly without outliers")
```




```
rain_avg = []
norain_avg = []
for j in range(0,25):
    rain_avg.append(np.mean(rain[rain['Hour']==j]['ENTRIESn_hourly']))
    norain_avg.append(np.mean(norain[norain['Hour']==j]['ENTRIESn_hourly']))

plt.figure()
plt.plot(range(0,25),rain_avg,'rh',label="No Rain")
plt.plot(range(0,25),norain_avg,'b+',label="Rain")
plt.title('Average ENTRIESn_hourly per Hour')
plt.ylabel('Average ENTRIESn_hourly')
plt.xlabel('Hour of the Day')
```




```
plt.legend()
plt.show()
```

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?

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 “U-test” tells us that the distribution of ridership differs on rainy and non-rainy days. Also, the mean on rainy days is (only) slightly higher. However, samples sizes differ and ridership might depend on other variables like the time of the day. If we would include “rain” into our model, we would see that it would not improve the goodness of fit. The coefficient would be positive which means our model would predict more entries on a rainy day. However, since rain does not improve the goodness of fit (here R-squared), this is not enough to conclude that there are more hourly entries on rainy days. If there are is a difference, it is small (as we can also see from the boxplot which shows a quite similar picture for rainy and non-rainy days).

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

a. The U-test's null hypothesis is very general, so even we obtain statistically significant results, we might not be able to answer our question.
b. The linear regression encounters problems as the ENTRIESn_hourly variable seems to depend a lot on UNIT, i.e. the station. Therefore, the impact of the weather variables is covered as we do not differentiate between stations.
c. Linear regression (if we do not transform variables) might be too simplistic as it is not very intuitive that for example hour should have a linear relationship with the number of subway riders. Same holds for rain as a little bit of rain might people that tend to walk take the subway but a lot of rain, on the other hand, might let people stay at home.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?