Analyzing the NYC Subway Dataset

Section 0. References

- http://docs.ggplot2.org/current/ R ggplot2 documentation
- https://github.com/yhat/ggplot Python ggplot description page
- http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm Interpreting results: Mann-Whitney test
- http://pandas.pydata.org/pandas-docs/stable/generated/pandas.melt.html
- https://en.wikipedia.org/wiki/Mann–Whitney_U_test
- http://www.graphpad.com/guides/prism/6/curve-fitting/index.htm?r2_ameasureofgoodness_of_fitoflinearregression.htm Interpreting R2
- http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit Regression Analysis: How Do I Interpret R-squared

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?
For analysis of NYC I used the Mann-Whitney U-test. My null hypothesis was that the distributions of both groups of data (number of entries in rainy and non-rainy days) are identical.
If p-value will be less than p-critical = 0.05, I can reject null hypothesis.

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.
For this analysis I used improved dataset provided in file **turnstile_weather_v2.csv**. This dataset have equal intervals between data points - 4 hours without gaps.
Mann-Whitney test is robust and can be used for 2 population with unknown distributions.
The only property of dataset required for Mann-Whitney is the number of data points in sample.
Length of the each sample should be >= 5. We have 33064 samples for non-rainy weather and 9585 samples for rainy.

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.
My results:
mean value for rainy days: 2028.2
mean value for days without rain: 1845.5
U: 153635120.5
p: 2.74e-06

1.4 What is the significance and interpretation of these results?
My one-tail p-value is *2.74e-06*. This value is several orders of magnitude smaller than critical. Such a small value means that I can reject null hypothesis.

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?
I've tried several approaches to compute coefficients theta of linear regression and produce predictions for **ENTRIESn_hourly.** Three models from scikit-learn package (Ordinary Least Squares, Stochastic Gradient Descent and Bayesian Ridge Regression) and self-developed algorithm for gradient descent. Although for large datasets with (n>10000) stochastic could be significantly faster, for our dataset (42649 rows, 245 features) all methods are acceptable.
I've measured wall time using %time magic function built in IPython and running on Core-i5 based laptop.
Self implemented **Gradient descent** took **2.98 s** to complete with R2: **0.4766**
Stochastic gradient descent with 10 iterations took **2.01 s**, but R2 is smaller: **0.4484**
Stochastic gradient descent woth 50 iterations took in total **3.85 s** but R2 still smaller than with gradient descent R2: **0.4592**
Ordinary Least Squares showed time and R2 very similar to gradient descent
time: **2.57 s** R2: **0.4776**
Bayesian Ridge Regression took more time for same result
time: **3.36 s** R2: **0.4775**
Although results were achieved with **Ordinary Least Squares** were slightly better, I've used **gradient descent** for my model.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?
For the model I used these features:
rain: 1 for days with rain, otherwise 0
fog: 1 for days with fog, otherwise 0
hour: hour of the record
hour2: square hour of the record
meantempi: Mean temperature during the day
day_week: Day of week
Also I've used dummy variables which represent UNIT column (240 dummy columns in total)

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 R2 value.”
As I realised in previous chapter, I couldn’t ignore relation of **ENTRIESn_hourly** with rain parameter, so I added **rain**.
It’s pretty obvious that subway traffic depend on day of week and time of the day, so I used **day_week** and **hour**. Because dependence on hour could be non linear I added **hour2** - square of hour to feature list. Addition of mean temperature to feature list (**meantempi**) slightly improved my results.

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
My theta values for non-dummy variables are:

rain	fog	hour	hour2	meantempi	day_week
24.59	-53.81	340.33	551.83	-101.43	-301.58

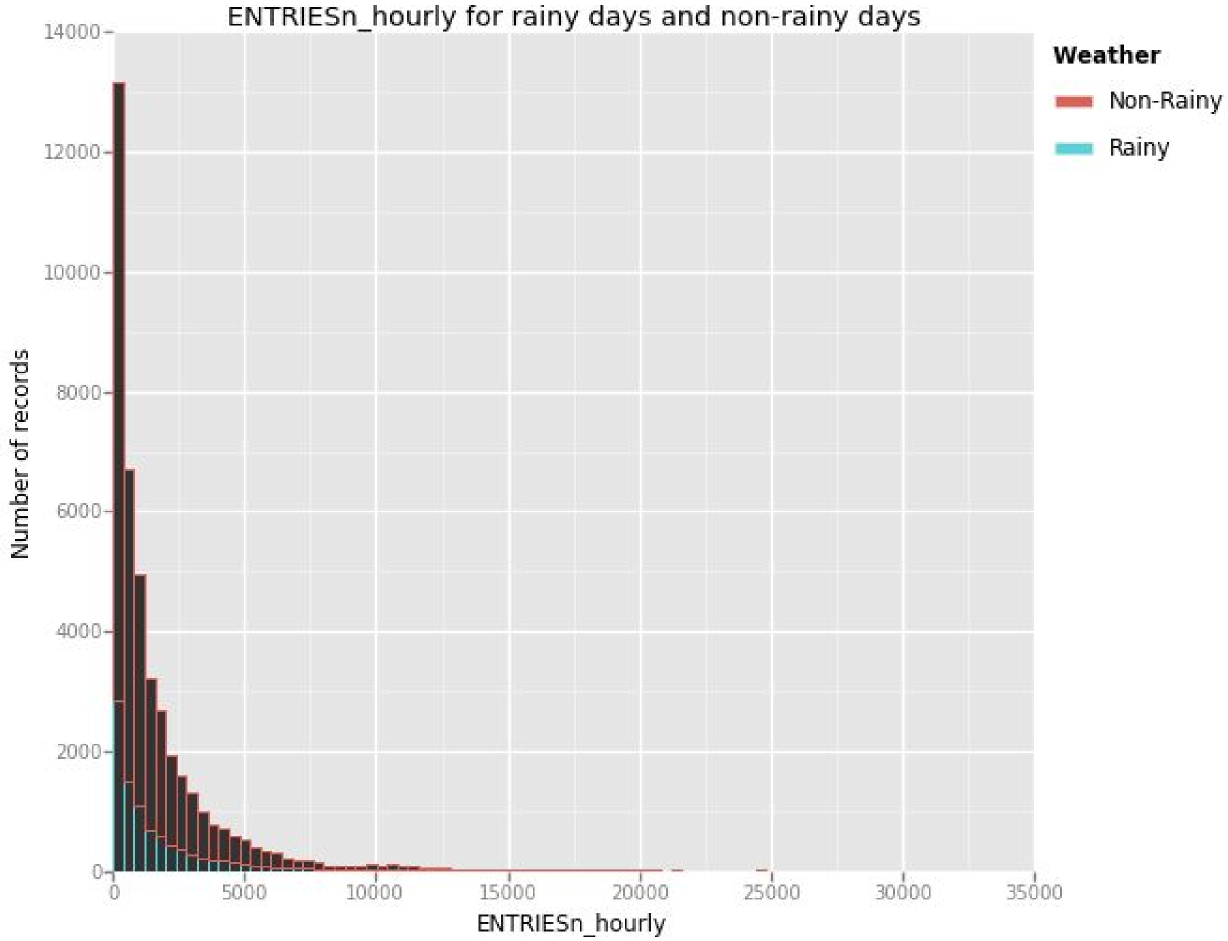
2.5 What is your model’s R2 (coefficients of determination) value?
My R2 is 0.477

2.6 What does this R2 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 R2 value?
Although larger R2 values (close to 1.0) are usually mean better fitting of data to model, large R2 not always mean that model will produce better predictions on new data. Nevertheless R2=0.477 means that data points are scattered, and our predictions will not be reliable.

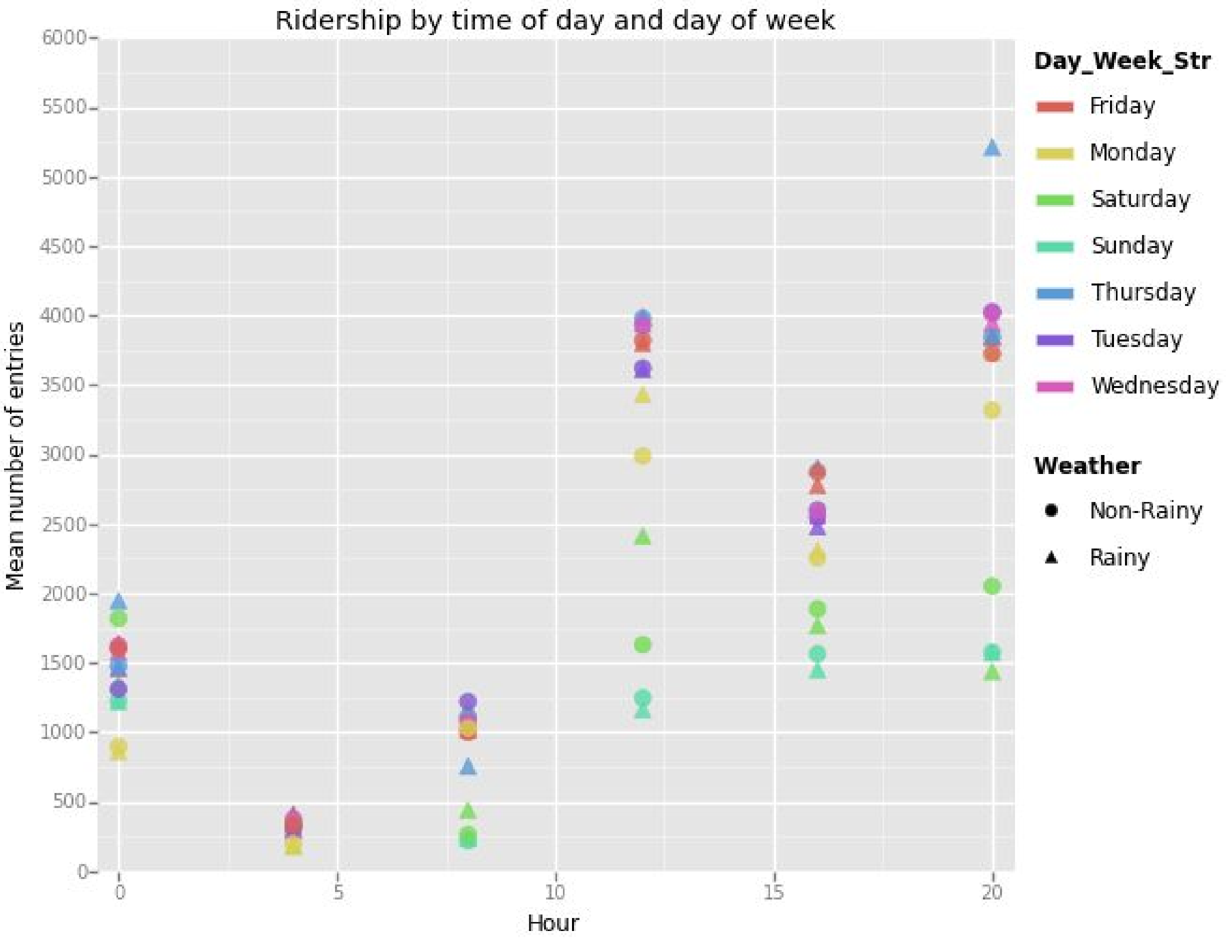
Section 3. Visualization

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.



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



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?

Yes, ridership in NYC subway is depend on weather, and it depend on rain. More people prefer subway to other transport when it rains.

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.

Statistical analysis showed that number of entries to NYC subway is depend on rain. After that we build a regression model. Theta value for rain parameter is 24.59
and it's greater than zero. This means, that for days when rain is set to 1, predicted value will be greater, than in case of rain = 0. Although, this additional value 24.59 is not very high.

Full source of my ipython notebook with calculation for this project is available through nbviewer and secret github gist page: http://nbviewer.ipython.org/gist/pinya/e276317b366b94ed5d28

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including: Dataset, Analysis, such as the linear regression model or statistical test.
Mann-Whitney test is a good starting point for testing hypotheses, but we couldn't use it for short datasets. This statistical case couldn't say anything in case when null hypothesis is true. Linear regression is a fast way to build prediction model, but it's hard to provide good results for complex data. Researcher should spend time for selection and/or introduction features for building model.

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

Timestamps in improved dataset rounded to 4 hour interval, and this data too smooth for analysis ridership in rush hours.