Don Garofalo

5/30/15

Introduction to Data Science

This document addresses part 2 of the project. Please use this document as a template and answer the following questions to explain your reasoning and conclusion behind your work in the problem sets. You will attach a document with your answers to these questions as part of your final project submission.

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.

The following primary references were used:

- Maksoudian, Y. L, Probablity and Statistics with Applications, International Textbook Company, Scranton, PA, 1969.
- McKinney, W., Python for Data Analysis, O'Reilly, Sebastapol, CA, 2013.

Section 1. Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data?

The Mann-Whitney U test was used to analyze subway turnstile entries data for rainy days and non-rainy days.

1.2 Did you use a one-tail or a two-tail P value?

A one-tail P value was used.

What is the null hypothesis?

Is there no statistical difference in the distribution of the number of turnstile entries between rainy and non-rainy days? This hypothesis is used to test whether the turnstile entries population is the same on rainy and non-rainy days.

What is your p-critical value?

When the Mann-Whitney U test was applied to the improved subway turnstile data, a p-critical value of 2.74106957124e-06 was computed.

1.3 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 following assumptions make the Mann-Whitney test applicable:

- Turnstile entries data is an independent variable
- Categorizing rainy versus non-rainy days is a dependent variable
- Rainy days and non-rainy days are mutually exclusive, and therefore independent observations

1.4 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.

For the original subway turnstile data, a p-value of 0.025 was calculated. The mean turnstile entries were 1105.4 and 1090.3 for rainy days and non-rainy days, respectively.

For the improved subway turnstile data, a p-value of 2.74106957124e-06 was calculated. The mean turnstile entries were 1845.5 and 2028.2 for rainy days and non-rainy days, respectively.

1.5 What is the significance and interpretation of these results?

Both the original and improved subway data support a conclusion that ridership is increased on non-rainy days. The test is significant for both the original and improved subway turnstile data (the p-value must be less than an a priori alpha value of 0.05). Only the improved subway turnstile data has high significance, which typically requires the p-value to be less than an alpha of 0.01.

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?

Gradient descent was used to compute the coefficients theta and predict hourly entries.

2.2 What features (input variables) did you use in your model?

The following table shows the input features used in the gradient descent model.

Input variable	Meaning
Precipi	Precipitation in inches at the time and location.
Hour	Hour of the timestamp from TIMEn. Truncated rather than rounded
meantempi	Daily average of tempi for the location.
Wspdi	Wind speed in mph at the time and location.
day_week	Integer (0 6Mon Sun) corresponding to the day of the week.
Fog	Indicator (0 or 1) if there was fog at the time and location.
pressurei	Barometric pressure in inches Hg at the time and location.

Did you use any dummy variables as part of your features?

Yes, the turnstile unit (substation) was introduced as a dummy variable.

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."

The features were selected upon both intuition (i.e. people will tend to ride the subway less when the weather is bad) and data exploration (i.e. observing how the R2 value is affected by the inclusion of a feature).

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

Variable	Coefficient (theta)
Precipi	-89.68750998
Hour	832.16578665
Meantempi	-117.50827566
Wspdi	45.33565206
day_week	-302.81999275
Fog	-34.94602728
Pressure	-73.77776976

2.5 What is your model's R2 (coefficients of determination) value?

The R-squared value for the improved data set is 0.472039455496.

2.6 What does this R2 value mean for the goodness of fit for your regression model?

The R-squared value is a good fit. One would be a perfect fit, and zero would not explain any variability about the mean.

Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

The R-squared value is reasonable for a couple of factors:

1. Human behavior cannot be easily modeled mathematical equations
2. Ridership is not a linear function of certain input variables. For example, ridership tends to be higher during the work week. Hourly ridership would have multiple peaks during rush hours.

Section 3. Visualization

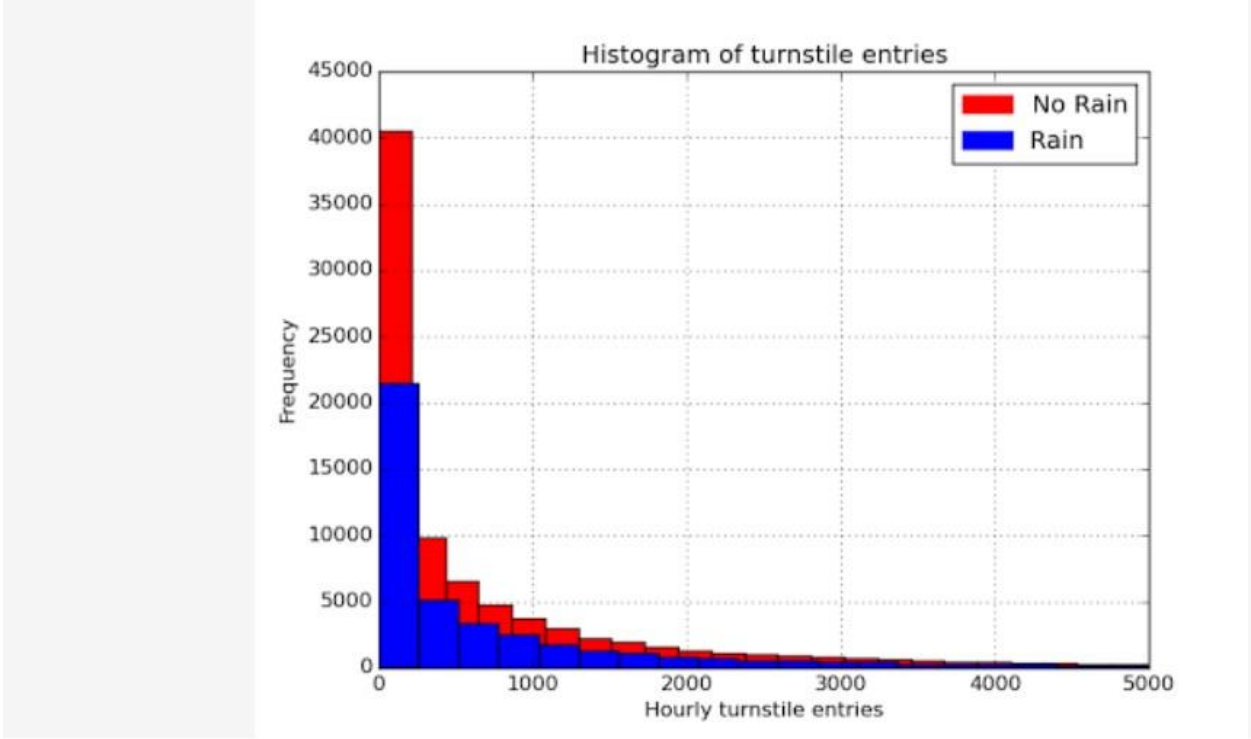
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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.

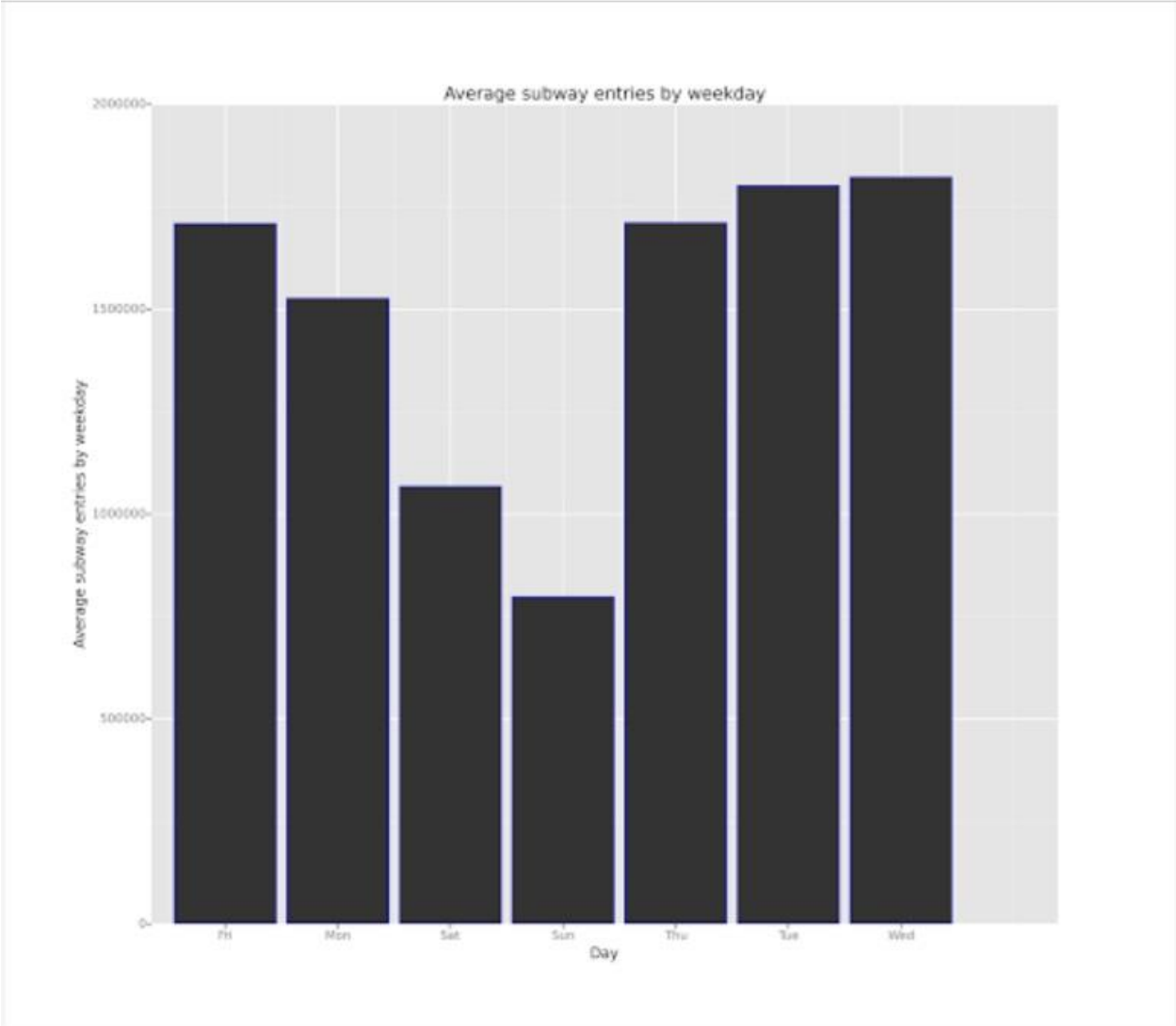
- Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.



The purpose of the above diagram is to characterize the distribution of subway ridership. It should be noted that the non-rainy day sample size is 87,847 and the rainy day sample size is only 44,104. Thus, it is not apparent from the visualization that there is more ridership on rainy versus non-rainy days. It is apparent the ridership distribution does not follow a normal distribution.

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



The above bar chart shows that average ridership is greater on during the normal work week (i.e. Monday through Friday) versus on the weekend (i.e. Saturday and Sunday).

It day labeling should be improved to show the days sequentially rather than a seemingly random hash order.

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?

More people will tend to ride the NYC subway when it is not raining. Comparing the average ridership on rainy and non-rainy days supports this conclusion. Also, this conclusion is consistent with my intuition as people will tend to avoid exposure to bad weather conditions.

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.

Both the statistical tests as well as the linear regression model support the conclusion that ridership will be greater on non-rainy days. The mean turnstile entries are greater for non-rainy days when compared to rainy days. From the Mann Whitney U test, this mean difference is highly significant, and it is very unlikely that this is due to a sampling error. As for the linear regression, the negative coefficient for the "precipi" input variable suggests that the ridership will decrease as the amount of rain increases.

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.

   Both the data set and analysis have potential shortcomings.

   The turnstile data may have missing values or turnstile counters that reset. A more thorough analysis would cleanse such anomalies from the data prior to analysis.

   The linear regression model has input variables that are really not linear in nature. For example, reviewing the output of the map reducer shows a "busiest hour" occurs when the corresponding hour is neither a maximum nor minimum. Furthermore, rush hours typically occur in daily pairs (i.e., morning and evening events), and therefore are multimodal. The day of week visualization shows that the day of the week variable would be a better input variable is categorized (e.g. normal weekday versus weekend).

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

I think studying the propensity of physics majors going into data science would be an interesting study. This is my humorous feedback on the many videos shown in the course featuring data science professionals with an educational background in physics.