Project P1: Analyzing the New York City Subway Dataset

Project submission by Tony Malerich as partial completion of the Udacity Data Science Nanodegree program

Overview

The Udacity Introduction to Data Science course involves an analysis of a dataset of New York City subway ridership over the course of a month. The dataset includes the number of entries and exits each hour, a Boolean value for rain, the date, time, and hour, the subway turnstile involved, and a number of other factors. Our analysis attempts to determine whether or not more people ride the subways in inclement weather.

This project consists of two parts. In Part 1 of the project, we completed the questions in Problem Sets 2, 3, 4, and 5 in the Introduction to Data Science course.

This document addresses part 2 of the project. In this part of the project we answer guided questions to explain our reasoning and conclusions behind our work in the problem sets.

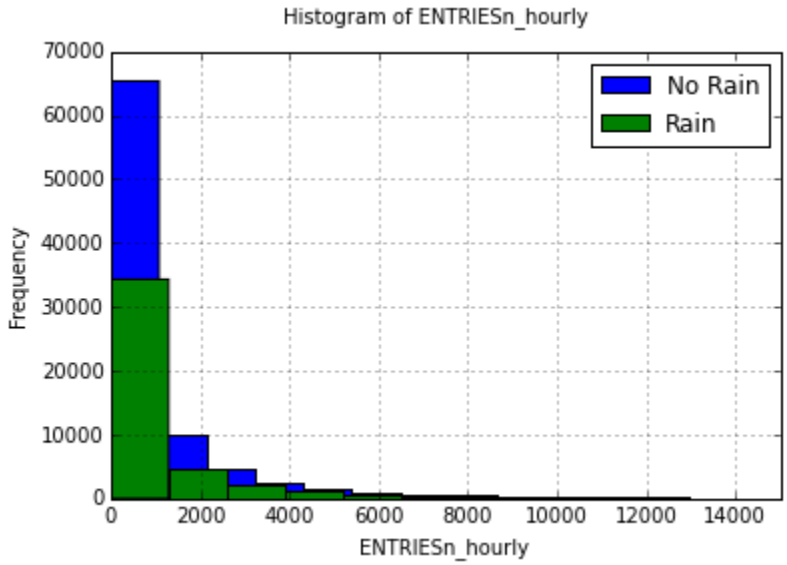
Section 1. Statistical Test

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?

We used the Mann-Whitney U-Test to determine whether or not there is a statistically significant difference between NYC subway ridership on days when it rained or did not rain. This is a non-parametric statistical test that we use to compare the underlying distribution of subway ridership on days with rain, and the distribution on days without rain. The null hypothesis of this test is that the probability of a number drawn from distribution 1 being greater than a number drawn from distribution 2 is 0.5; in other words, that the distributions are indistinguishable. In this case we used a two-tailed P value, since we are interested in determining that any difference exists between ridership on rainy vs. non-rainy days (although we believe it is unlikely that less people ride the subway on days when it rains). The p-critical value to reject this null hypothesis at the 95% confidence level is 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.

There are two statistical tests commonly used to compare distributions, Welch's T-test and the Mann-Whitney U-Test. Welch's T-test is applicable when the data is normally distributed. To get a feel for the distributions of ridership on rainy vs. non-rainy days, we plotted the histogram of each distribution using the python script `entries_histrogram.py` and produce this plot:



The plot indicates that the ridership is not normally distributed; hence the Mann-Whitney U-Test is more appropriate for this data set.

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.

We conducted the Mann-Whitney U-Test on the ridership data with numpy, pandas, and scipy with the script `mann_whitney_plus_means.py`. Running the analysis, we found the following mean riderships:

Rainy days: 1105.44

Non-rainy days: 1090.27

The mean ridership is approximately 15 more riders on rainy days, or nearly 1.4% higher. That implies that more people may be riding the subways on rainy days, but is it

statistically significant? To answer that question, we calculated the Mann-Whitney U value and found 0.024999912793489721. Since this is a two-tailed test, we need to multiply the result by two and we have p = 0.049999825586979442.

1.4 What is the significance and interpretation of these results?

Since the result of the Mann-Whitney U-Test is less than the p-critical value of 0.05, we reject the null hypothesis that the underlying distributions are identical, and conclude that there is a statistically significant difference in ridership on rainy days vs. non-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:

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

We used the gradient descent technique to produce a predictive model for the number of hourly riders on the NYC subways.

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

We used rain, hour, a dummy variable for the UNIT, and a term of ones (a constant term) as the input variables to the model.

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."

We tested many variations of input variables in an effort to get a well correlated model. In our final model, we use the features rain, hour, plus the dummy UNIT variable, and a constant term of ones as the independent variables. We started the model with the variables suggested by the Udacity: rain, precipitation, hour, and mean temperature, plus the dummy variable for the UNIT. We played with combinations of these and other

variables, such as fog and min temperature, then tested the R^2 value of the model. Comparing the R^2 value with various combinations of features to the initial set, we found that rain alone was not as good of a predictor. Including precipitation in addition to rain did not improve the model, which is not surprising as it is nearly a duplicate of rain itself. Adding the hour as an input improved the result, while adding mean temperature and fog resulted in nearly the same R^2 as just rain and hour. Hence, we went with just rain and hour since it is a simpler model with nearly the same predictive power as adding any of the other features.

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

The non-dummy coefficients are { 8.2, 459.2, 1088.7 } for the variables { RAIN, HOUR, 1's }. We observe that since the weight for HOUR is 459.2 and the weight for RAIN is only 8.2. This shows that the time of day accounts for many more subway riders than does the occurrence of rain. Intuitively this makes sense, as the bulk of the population maintains a regular schedule where they would work the traditional hours of 9-5, so the time of day is a big predictor of how many people are on the move, and some or many of them will take the subway,

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

The coefficient of determination that we get using the input variables of rain, hour, the UNIT dummy, and a constant, with learning rate alpha = 0.4 and setting the number of iterations to 10, is $R^2 = 0.457$.

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?

The R^2 value is an indicator of how well the model fits the data, with a value of 1.00 being ideal. The suggestion in the exercise is to get an $R^2 > 0.20$, so our value of 0.46 is at least higher than that, but it is not that close to 1.00. So that says that our linear model, while estimating trends in a general sense, will not yield a highly accurate prediction of the day-to-day subway ridership.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

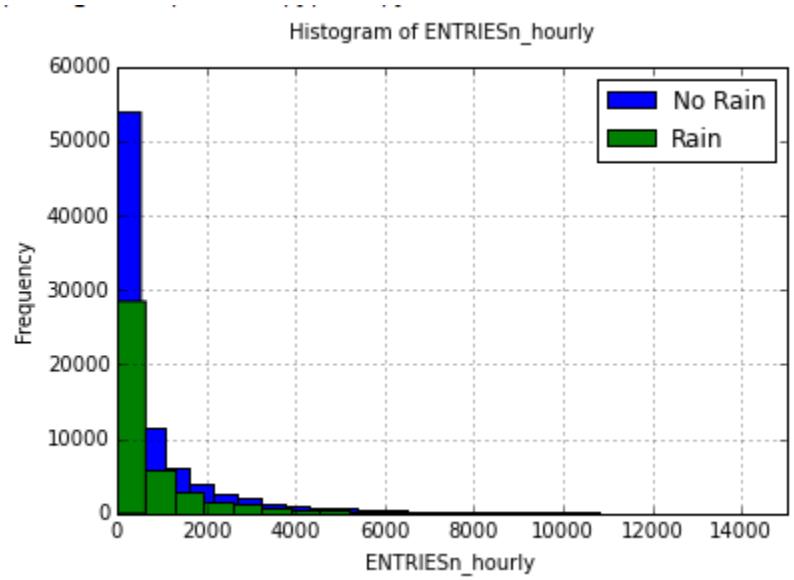
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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.

Here is the plot of the histograms of ridership when it rains vs. when it does not (this is the same plot we presented in section 1 with an increased number of bins).

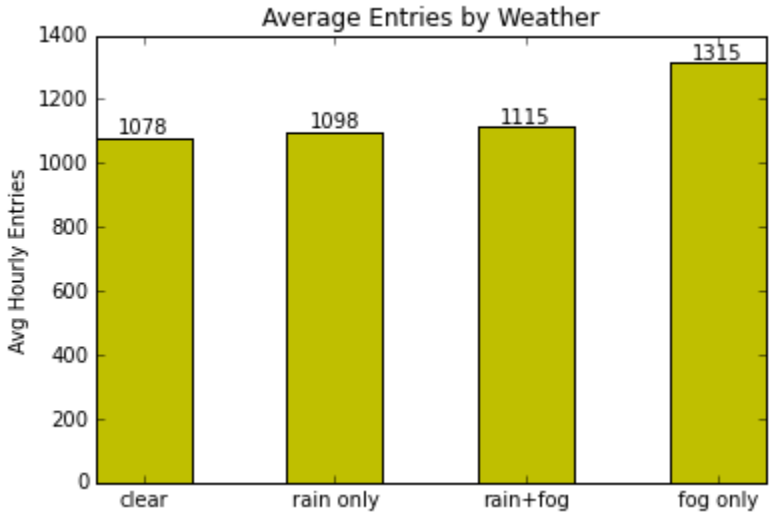


Visual inspection of the plot shows more overall ridership on non-rainy days; however, this is simply total ridership, and does not consider the number days that have rain.

3.2 One visualization can be more freeform. Some suggestions are:

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

Here we present a plot that shows the average number of hourly entries, broken out across different types of weather.



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?

In the final analysis, we conclude that more people ride the subways when it rains than when it does not. The descriptive statistics argue for this conclusion. The regression model is inconclusive. The Mann-Whitney U-Test gives 95% confidence that ridership is different on rainy days, and it is a rigorous statistical test; in the end it is best to rely on that rigor.

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.

Based on the statistical and visual analysis of the NYC subway data, it is rather difficult to make definitive conclusions regarding the ridership on rainy vs. non-rainy days. The

various comparisons we made show mixed results. So we endeavor to draw the best conclusion possible based on the data.

We have a histogram that shows more total riders on non-rainy days; however, as we mentioned above, the histogram does not correct for the number of days with and without rain, so it is not conclusive.

We have the Mann-Whitney U-Test to compare the underlying distribution of ridership on rainy and non-rainy days. The critical test value for 95% confidence is 0.025, and our result is 0.024999. This does imply a statistically significant difference between ridership, although it is very close.

We fit a linear regression model to the data using gradient descent, and found a coefficient of determination $R^2 = 0.46$. This does not imply a strong correlation between ridership and rain. To be confident in the predictive power of our model, we would prefer an R^2 closer to 1.

Additionally, we generated the descriptive statistics (using numpy) measuring ridership across different weather patterns:

Weather	Mean	Median	Stdev
No Rain	1090.3	278.0	2320.0
Rain	1105.4	282.0	2370.5
Fog	1154.7	297.0	2474.0
Min Temp < 50	1312.0	334.0	2698.7

Looking at the averages, both mean and median, one would suspect a trend saying that ridership increases the worse the weather conditions. However, the standard deviations are also large, larger than the difference between mean ridership across weather conditions, so it is dangerous to draw conclusions without applying statistical methods to determine significance of the differences.

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.

A potential shortcoming of the dataset is that it is drawn from only one month of subway ridership. The month in particular is May, which is potentially another shortcoming. We have min/max temperature data, and it would be interesting to look for temperature correlations to ridership numbers when the temperature range is more extreme than the 46 to 86 deg F presented in this data.

There are also several shortcomings in the analysis. The coefficient of determination of the regression model calculated by gradient descent did not show a strong correlation between rain and ridership, but it did imply some correlation. This is rather vague, which in turn is rather disquieting in an analysis.

The Mann-Whitney U statistical test was more reassuring, although it too had some shortcomings. The critical score to reject the Null hypothesis was 0.025 and our score was 0.024 with some 9's, so it's right on the border. We could not have rejected the Null hypothesis at the 99% confidence level.

References:

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
http://pandas.pydata.org/pandas-docs/stable/visualization.html#histograms
http://en.wikipedia.org/wiki/Coefficient_of_determination
https://github.com/yhat/ggplot/
http://mathematicalcoffee.blogspot.com/2014/06/ggpie-pie-graphs-in-ggplot2.html
http://matplotlib.org/examples/api/barchart_demo.html