

# Analyzing the NYC Subway Dataset

## Overview

*In this analysis, I am taking data from May 2011 NYC MTA Subway as provided in the Udacity 'Introduction to Data Science' class and testing the hypothesis that whether rainy days impacts ridership based on problem sets 2 to 5.*

## 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?

The NYC subway data was analyzed with the Mann-Whitney U-Test. A two-tail p-value was used as no prior assumptions are made about the contrast in the distributions of ridership on rainy and non-rainy days.

The p-value returned by `scipy.stats.mannwhitneyu` is one-tailed as noted here:

<http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.mannwhitneyu.html>

In order to use a two-tailed test, the one-tailed p-value returned by `scipy.stats.mannwhitneyu` must be multiplied by 2.

The p-critical value used was 0.05 (i.e. 5% chance of observing a result as least as extreme).

### 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.

The question being asked is whether subway ridership varies with the weather. The provided dataset allows hourly entries to MTA turnstiles to be spliced into two different samples, entries with and without rain. The Mann-Whitney U-Test tests the null hypothesis that the two samples being compared are derived from the same population. This null hypothesis allows us to test whether there is a statistically significant difference in ridership on rainy and non-rainy days (i.e., are the hourly entries derived from the same population).

Furthermore, exploratory data analysis (see Section 3.1) has shown that the data is not normally distributed. The Mann-Whitney U-Test does not assume normality of the data, making this test appropriate.

### 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.

The numerical results of the Mann-Whitney U-Test are the following:

Mean entries, with rain	Mean entries, without rain	Mann-Whitney Test Statistic	p-value
1105.44637675	1090.27878015	1924409167.0	0.0249999127935

#### 1.4 What is the significance and interpretation of these results?

```
# Significance level defined in Section 1.1
alpha = 0.05

# two-tailed test
if (p * 2) < alpha:
    print 'Reject the null hypothesis'
else:
    print 'Fail to reject null hypothesis'
```

##### Reject the null hypothesis

With this small p-value, we reject the null hypothesis of the Mann-Whitney U-Test. In other words, the distribution of the number of entries is statistically different between rainy and 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?

- Gradient descent (as implemented in exercise 3.5)
- OLS using Statsmodels
- Or something different?

Ordinary Least Squares (OLS) is chosen to solve the linear regression model. The normal equation is implemented directly without the use of the Statsmodel package.

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

The following features are taken directly from the turnstile\_master dataset:

- rain
- fog
- Hour
- Meantempi

```
# select features
features = turnstile_master[['rain', 'fog', 'Hour', 'meantempi']]
```

The following features are created by transforming data:

- **precipi** – The square root of the 'precipi' field in the turnstile\_master dataset

```
# needed to prevent warning about chained assignment
features.is_copy = False

# create polynomial features
features['precipi'] = turnstile_master['precipi']**0.5
```

Dummy variables are used to include categorical data as features in a linear model. **The dummy variables include:**

- **UNIT** - Taken directly from the turnstile\_master dataset
- **weekday** - The day of the week as an integer [0:6] (Monday = 0, Tuesday = 1, ...)

ref - <https://docs.python.org/2/library/datetime.html#datetime.datetime.weekday>

```
# UNIT
dummy_units = pd.get_dummies(turnstile_master['UNIT'], prefix='unit')
features = features.join(dummy_units)

# weekday
from datetime import datetime

# use lambda function to get weekday from 'DATEn' field
f = lambda x: datetime.strptime(x, "%Y-%m-%d").weekday()

dummy_units = pd.get_dummies(turnstile_master['DATEn'].apply(f), prefix='weekday')
features = features.join(dummy_units)
```

**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<sup>2</sup> value.”

meantempi is chosen as a feature as temperature is a component of the weather that affects people's decision making. A given temperature may affect how long and how much effort it takes to clothe in the morning. A person may simply choose to stay indoors due to discomfort with a given temperature.

Hour and weekday features were chosen as it is easily observed how ridership varies with time of day and day of week. A phenomenon supporting the Hour feature is the daily rush hours that mass transit systems and roadways exhibit and accommodate. An observation supporting the use of weekday is how the MTA train schedule varies between weekdays and weekends, with trains arriving with more infrequency on Saturday and Sunday.

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

- Rain: 0.14
- Fog: 0.17
- Hour: 0.017
- Meantempi: 0.0051

## **2.5 What is your model's R<sup>2</sup> (coefficients of determination) value?**

The R<sup>2</sup> of the model is 0.470491013765

## **2.6 What does this R<sup>2</sup> 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<sup>2</sup> value?**

This R<sup>2</sup> value is relatively low, meaning that the fit of this regression model is not very good. A linear model is probably inappropriate given this low R<sup>2</sup> value. Furthermore, the categorical nature of data in the given data set, such as 'rain', makes a linear model inappropriate.

R-squared (R<sup>2</sup>) statistic, some of its limitations, and uncover some surprises along the way. For instance, low R-squared values are not always bad and high R-squared values are not always good!

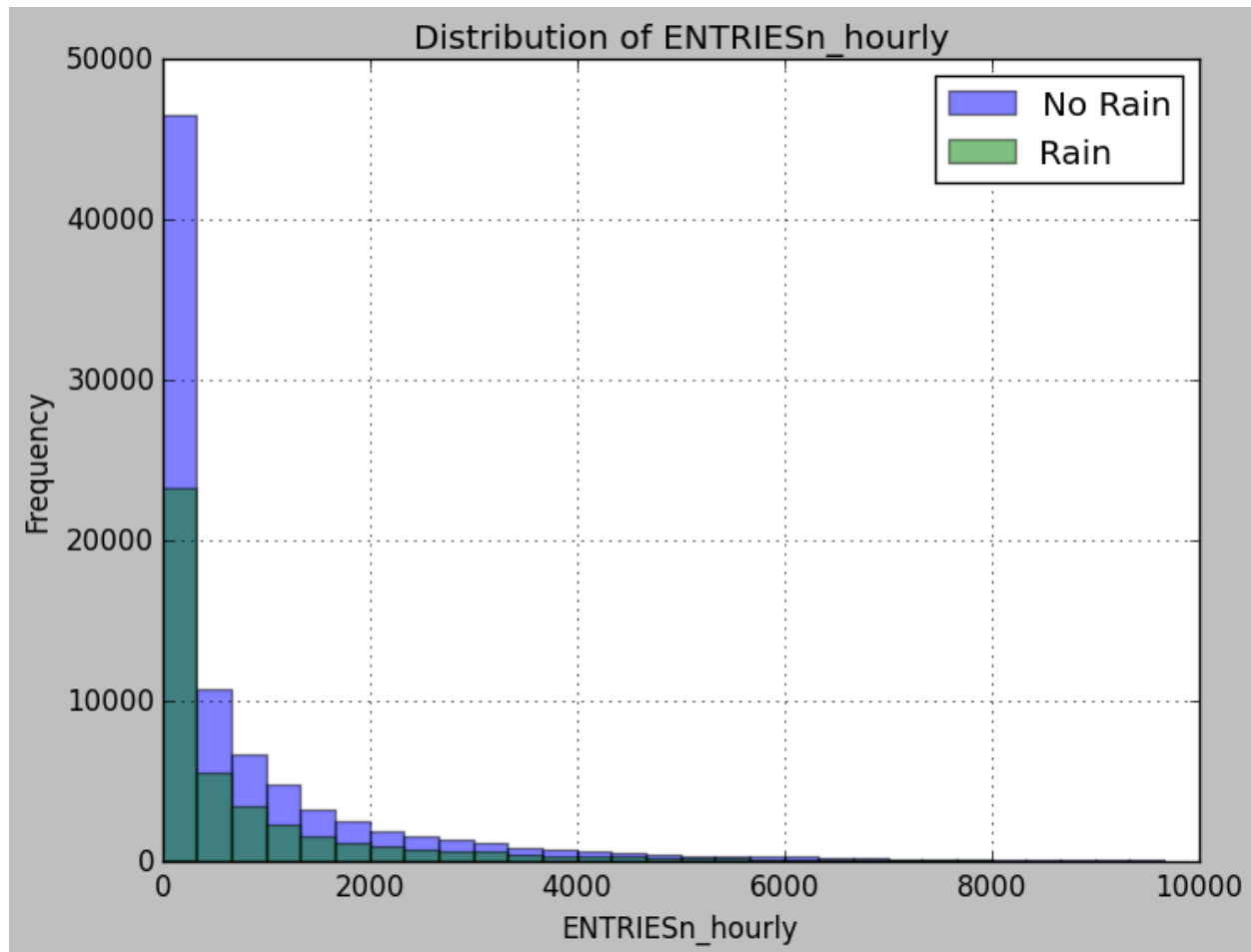
Linear regression calculates an equation that minimizes the distance between the fitted line and all of the data points. Technically, ordinary least squares (OLS) regression minimizes the sum of the squared residuals.

In general, a model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased.

## 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.

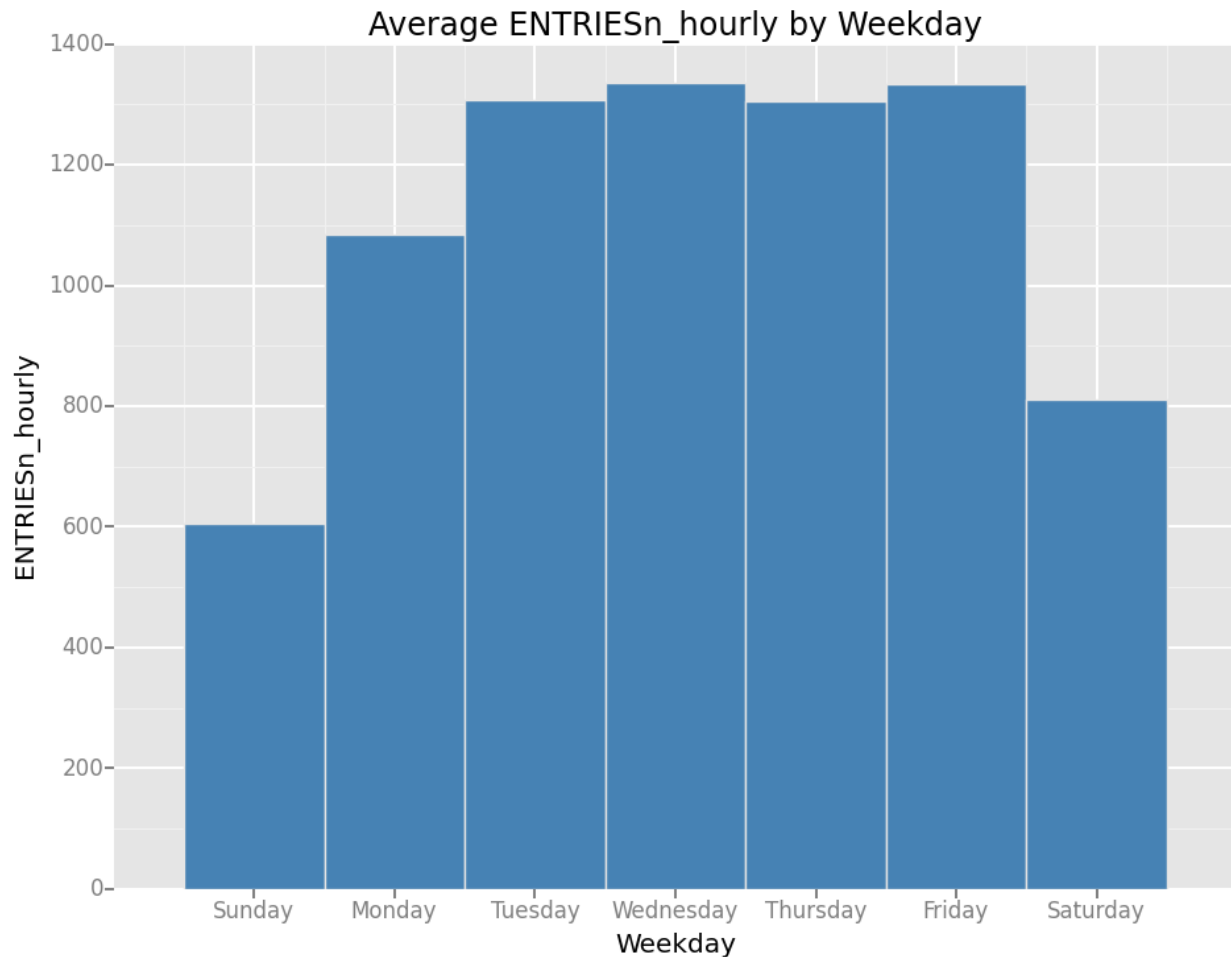
**3.1 One visualization should contain two histograms: one of `ENTRIESn_hourly` for rainy days and one of `ENTRIESn_hourly` for non-rainy days.**



The distribution of `ENTRIESn_hourly` appears to not be normally distributed and skewed to the right on both rainy and non-rainy days. The mode appears to be within the smallest bin for both distributions. There are far fewer observations on rainy days than non-rainy days.

### 3.2 One visualization can be more freeform.

#### Average Hourly Ridership by Day of Week



The above bar chart shows the average hourly ridership by day of week. The sum of ENTRIESn\_hourly by day of week was divided by the count of rows for a given day of week (as each row represents an hour's worth of data). The bar chart shows that the average hourly ridership is higher on weekdays than weekends, with Saturday seeing significantly higher ridership than Sunday. It appears that the average hourly ridership on Monday is significantly different than the rest of the weekdays. This may be due to a seasonal effect of Monday holidays. The given data set is a sample from May 2011. There is at least one major holiday that falls on Monday in the month of May: Memorial Day.

## 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?

Based on this analysis, I believe that more people ride the NYC subway when it is 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.

If we consider the regression results it remains unclear if rain and ridership have a linear relationship. The model that we have created, with a coefficient of determination of ~47%, seems rather insufficient to prove that weather has any correlation with ridership. But if we take a look at Figure in section 3.1, we can observe that there is a small correlation.

As seen in Section 1.3, the mean of `ENTRIESn_hourly` is greater for hours with rain than without (1,105 vs. 1,090). Additionally, a Mann-Whitney U-Test shows that the `ENTRIESn_hourly` sample with rain appears to be drawn from a different distribution (or population) than the `ENTRIESn_hourly` sample from hours without rain.

Thus, although this does not prove that rain causes more entries. It proves that on rainy hours, there is a higher chance of entries. **It is conscientious to claim that rain increases subway ridership.**

## 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.

The data set provided contains only one month of MTA data. This smaller data set is subject to effects of seasonality, as the time of year may also affect ridership. Additionally, the Month of May contains a Monday holiday, which appears to have an effect on the data in the visualizations in Section 3.2 and Section 4.

The largest shortcoming I see with this data set is that it appears that the MTA component of the data is produced on an hourly basis, but it is joined to daily weather data. For example, if it rained at any point in a given day, every hour of that day will reflect that it rained. This prevents a truly granular analysis of how the weather can affect ridership within a day.

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

I think that an interesting investigation would be to use gradient descent with logistics regression to see if one might be able to predict if it rained or not given various parameters, to include turnstile location/ID, time of day and subway entries. Intuitively, this might produce false positives or negatives on special days (e.g. holidays, city events, marathon runs, etc).

**REFERENCES**

- Welch's t-test  
[http://en.wikipedia.org/wiki/Welch%27s\\_t\\_test](http://en.wikipedia.org/wiki/Welch%27s_t_test)
- Mann-Whitney test  
<http://www.real-statistics.com/non-parametric-tests/mann-whitney-test/>  
<http://ocw.umb.edu/psychology/psych-270/other-materials/RelativeResourceManager.pdf>  
<http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html>
- R2 - linear regression  
[http://www.graphpad.com/guides/prism/6/curve-fitting/index.htm?r2\\_ameasureofgoodness\\_of\\_fitoflinearregression.htm](http://www.graphpad.com/guides/prism/6/curve-fitting/index.htm?r2_ameasureofgoodness_of_fitoflinearregression.htm)