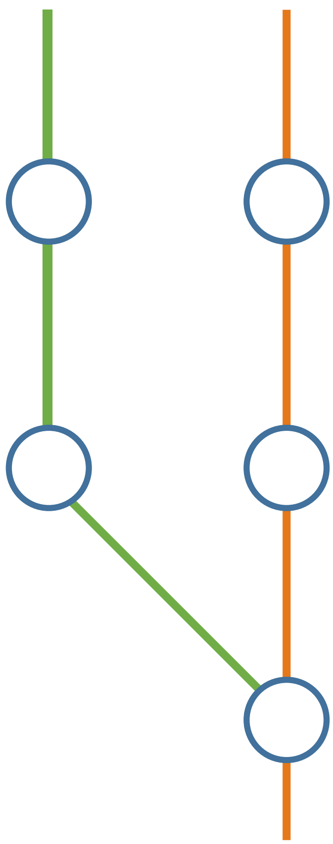
Udacity Data Analyst Nanodegree



Project 2:

Analyzing the NYC Subway Dataset

Daniel Mercier

1 TABLE OF CONTENTS

2 REFERENCES

1. **Udacity.** Understanding the MannWhitney U Test. [Online] 2014. https://goo.gl/Bhe2Up.
2. **Wikipedia.** Dummy variable (statistics). *Wikipedia.* [Online] [Cited: Sep 15, 2015.] https://en.wikipedia.org/wiki/Dummy_variable_(statistics).
3. **Wikipedia.** Mann—Whitney U test. *Wikipedia.* [Online] [Cited: Sep 15, 2015.] https://en.wikipedia.org/wiki/Mann–Whitney_U_test.
4. **McKinney, Wes.** Python for Data Analysis. Sebastopol: O'Reilly Media, Inc., 2013. ISBN: 978-1-44931979-3.
5. **Essa, Alfred.** Python Pandas Cookbook. *Youtube.* [Online] Aug 11, 2013. [Cited: Aug 30, 2015.] https://goo.gl/46x9Kc.
6. **Grus, Joel.** Data Science from Scratch, 1st Edition. Sebastopol: O'Reilly Media, Inc., 2015. ISBN: 978-1-491-90142-7.
7. **Diez, David M, Barr, Christopher D and Rundel-Cetinkaya, Mine.** OpenIntro Statistics. 3rd. 2015.
8. **Kormanik, Katie, Laraway, Sean and Rogers, Ronald.** Udacity Inferential and Descriptive Stats. *Udacity*. [Online] 2015. https://www.udacity.com/course/intro-to-inferential-statistics--ud201
9. **Evans, Dave.** Udacity Intro to Computer Science. *Udacity.* [Online] 2015. https://www.udacity.com/course/intro-to-computer-science--cs101.

3 STATISTICAL TEST

3.1 Test choice

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

I used the Mann-Whitney U test to determine the effect of rain on NYC subway ridership.

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

I used a two-tail P value, as we only want to know if rain has an effect on ridership and are not concerned with the effect's direction.

What is the null hypothesis?

For the Mann-Whitney U test, the default null hypothesis is *p(x,y) = 0.5*. My null and alternative hypothesis are thus expressed as:

$$H_0: p(ENTRIESn\_Hourly \mid rain > ENTRIESn\_Hourly \mid no\ rain) = 0.5$$

$$H_A: p(ENTRIESn\_Hourly \mid rain > ENTRIESn\_Hourly \mid no\ rain) \neq 0.5$$

In other words, the null hypothesis states that if we separate NYC hourly ridership data into two groups according to the occurrence of rain, the probability of drawing a sample from the rain group that is larger than the no rain group is 50%.

What is your p-critical value?

I chose a p-critical value of 0.05 for the test.

3.2 Test applicability

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 distribution of NYC ridership subway data for both the rain and no rain group has a significant positive skew. This skew likely reflects concentration of ridership at certain periods of the day. Skewed distributions like these usually require a non-parametric test. As the Mann-Whitney U test makes no assumptions about population distributions and only tests for bias between random draws from two groups of observations, it is ideal for testing the ridership data.

3.3 Test results

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 results of the test (using *turnstile_data_master_with_weather.csv*) are as follows:

Figure 1: Mann Whitney U Results

	RESULT
with_rain_mean	1105.4463767458733
without_rain_mean	1090.278780151855
U	1924409167.0
p-value	0.0386192688276

What is the significance and interpretation of these results?

As the p-value returned from the Mann Whitney U test is below the 0.05 critical value, we can reject the null hypothesis. This result suggests that rain does have some effect on subway ridership, but does not tell us the effect's direction.

4 LINEAR REGRESSION

4.1 Approach

What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model?

When deciding which approach to use, I weighed the benefits of an approach using OLS vs. one using gradient descent. OLS offers a closed form solution, but also comes with additional computation and time cost. Gradient descent benefits from lower computation cost and quicker execution, but risks settling on a local minimum or maximum instead of a global one.

As the project's dataset is of a fixed and relatively small size and is not computationally intensive, I opted for an OLS approach using Statsmodels to compute the θ coefficients and *ENTRIESn_hourly* predictions.

4.2 Features

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

In my final regression model, I used the following dummy variables:

- *UNIT* – Unit that collects turnstile information at a given location
- *day* - day of the week (Sunday – Monday)
- *Hour* - hour of the day (24 hour clock)

Why did you select these features in your model?

When testing my model, I considered variables as being either temporal, geographical or environmental in nature.

Temporal variables.

My heat map visualization suggested a strong effect on ridership from temporal variables. I decided on using *Hour* and *day* - as part of a two variable analysis they return an R^2 0.134.

I also considered including *DATEn* in the regression. However, the dataset only spans one month, giving us only a single cycle of data for *DATEn*, but four cycles for *day*. When also considering the potential for multi-collinearity between *day* and *DATEn*, I felt it was best to use only *day*.

Geographical variables

Geographical variables include all of the turnstile units (*UNIT*) from the dataset. Due to the large number of turnstile units (522), I treated them as a single grouped variable.

The turnstiles were the largest predictor of subway ridership - running a single variable analysis with UNIT returns an R^2 of 0.418.

Environmental variables

Environmental variables include any weather-related phenomenon, such as rain, wind speed, fog and temperature. I tried multiple combinations of environmental variables in my model, but none of them had any significant impact on R^2.

4.3 Parameters

What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?

Although I tried many model iterations that included variables other than the dummy variables I settled on, none of the non-dummy variables had significant coefficients or p-values. I would consider included them in the model if I were using a dataset containing a full year's worth of data.

4.4 Coefficients of determination

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

Scikit's OLS analysis returns an R^2 of 0.514:

Figure 2: Statsmodels OLS Regression Summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         ENTRIESn_hourly   R-squared:                       0.514
Model:                             OLS   Adj. R-squared:                  0.512
Method:                  Least Squares   F-statistic:                     281.7
Date:                 Sat, 21 Nov 2015   Prob (F-statistic):               0.00
Time:                         19:17:36   Log-Likelihood:            -1.1632e+06
No. Observations:               131951   AIC:                         2.327e+06
Df Residuals:                   131457   BIC:                         2.332e+06
Df Model:                          493
Covariance Type:             nonrobust
==============================================================================
```

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

An R^2 of 0.514 means that turnstile location (*UNIT*), day of the week (*day*) and hour of the day (*Hour*) explain 51.4% of the variability in the NYC ridership.

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

An R^2 of 0.514 indicates that the variables in the regression model predict a meaningful amount of the variability in subway ridership.

However, it is worth noting that the dataset only contains data points from the month of May. As such, it fails to take into account seasonal patterns that might affect subway ridership.

Given this, I believe that the predictive capability of this model is appropriate for the limitations of the dataset, but not for subway ridership models that attempt to generalize outside of those limitations.

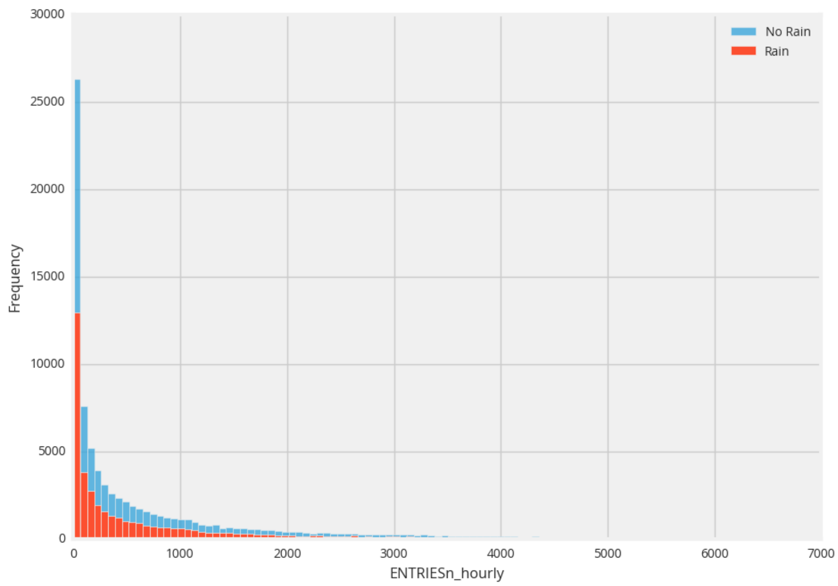
5 Visualizations

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

5.1 Visualization 1 - Histograms

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

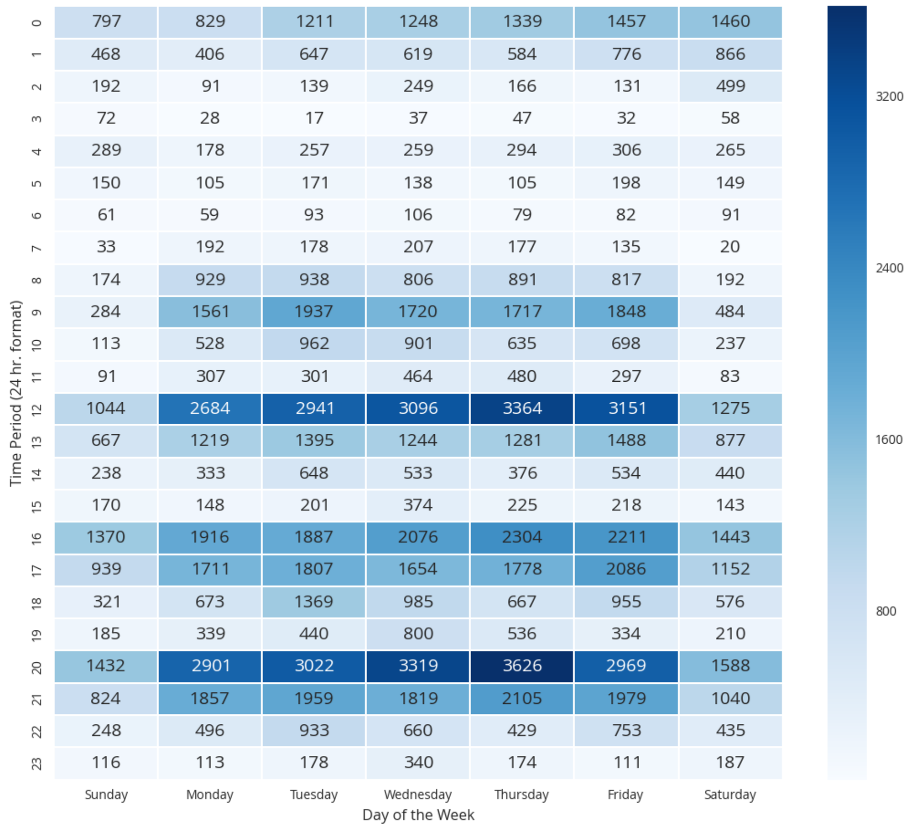
Figure 3: ENTRIESn-hourly Histogram



In the above combined visualization, the histograms for *ENTRIESn_hourly* for rainy days and non-rainy days both present a strong positive skew. This suggests that turnstile traffic is concentrated within certain time periods. The skew also suggests that testing our hypothesis will require a non-parametric test.

5.2 Visualization 2 – Heat Map

Figure 4: Heat Map of NYC Subway Turnstile Entries



The heat map for *ENTRIESn_hourly* by *day* (day of the week) shows a ridership pattern similar to that suggested by the histograms. Ridership peaks during four distinct time periods: 8-10 am (commute to work), 12-2 pm, (lunch break) 4-6 pm (commute back home) and 8-10 pm (social outings or late commute back home). These peaks are most pronounced from Monday to Friday. On the weekend, we still see a similar pattern, but with reduced intensity and the disappearance of the peak in the 8-10 am time period.

6 CONCLUSION

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 my analysis, the number of people riding does not increase when it is raining.

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 Mann-Whitney U tests result does fall within the p-critical value of 0.05, suggesting that rain does have an effect on ridership.

However, when we perform a single variable regression with rain, we see the following:

Figure 5: Statsmodels OLS Single Variable (rain) Regression Summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        ENTRIESn_hourly   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     1.237
Date:                Sat, 21 Nov 2015   Prob (F-statistic):              0.266
Time:                        21:32:15   Log-Likelihood:            -1.2107e+06
No. Observations:              131951   AIC:                         2.421e+06
Df Residuals:                  131949   BIC:                         2.421e+06
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const       1090.2788      7.885    138.274      0.000      1074.824  1105.733
rain          15.1676     13.638      1.112      0.266       -11.564    41.899
==============================================================================
Omnibus:                   146011.622   Durbin-Watson:                   1.032
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         14970624.827
Skew:                           5.690   Prob(JB):                         0.00
Kurtosis:                      53.926   Cond. No.                         2.41
```

The t-score and p-value for rain show that its effect is not statistically significant. It also has no discernable effect on R^2.

To conclude, we can say that rain has no effect on ridership, and that the primary drivers of people’s subway usage are by far the geographical (turnstile locations) and temporal ones (time of day and day of the week).

7 REFLECTION

Please discuss potential shortcomings of the methods of your analysis. Including Dataset, Analysis, such as the linear regression model or statistical test.

My regression model is unable to predict the effects of some structural and seasonal lurking variables not present in the dataset.

Structural

We could expect to see changes to ridership due to a number of structural factors such as:

- periods of construction or maintenance on subway lines or stations.
- increases or decreases to fare prices.
- operational differences between peak and off-peak operation.

Seasonal Variations

As the dataset used with the regression model only covers the month of May, this limits its ability to predict seasonal patterns that would occur during a 12-month calendar year. Given a dataset that covers a full year, we might expect to see changes in ridership during:

- heavy snow or long cold strings in winter.
- the Christmas to New Year's holiday period.
- peak tourist season at stations near Central Park or Times Square.

8 Appendix – Source Code

The source code used for this project is available on Github at https://github.com/dwmercier/Data-Analyst-Nanodegree---Project-2.