Analyzing the NYC Subway Data

Questions

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 the NYC subway data, I used a two-tailed Mann-Whitney U test. The null hypothesis is that the probability of an hourly entries observation from a non-raining period being greater than an observation from a raining period is equal to the probability of a raining period hourly entries period being greater than a non-raining period for any combination of raining and non-raining period observations. The p-critical value for this test 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.

The Mann-Whitney U Test is applicable to this dataset because the data does not have a normal distribution. Because the distributions are not normal, t-tests cannot be used to judge the two samples and the Mann-Whitney U Test is used because while, it is not as efficient as t-tests, it can handle non-normal distributions.

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.

Mann-Whitney U Test for raining entries compared to not raining entries

Mean ENTRIESs_hourly Raining	1105.4
Mean ENTRIESs_hourly Not Raining	1090.3
P Value	0.0499
U Value	1924408167

1.4 What is the significance and interpretation of these results?

Since the p-value for raining vs. not raining hourly subway entries was 0.0499 (4.99%) which is smaller than 0.05 (5%), it refutes the null hypothesis and shows there is a meaningful difference in ridership where more people ride when in rains.

Since the ridership difference is meaningful we can interpret the results and claim that, when it rains, NYC subway ridership increases by 1.375%.

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:

I used OLS using Statsmodels to determine theta and produce predictions for ENTRIESn_hourly in my regression model

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

I used 6 input variables for the model:

1. **Rain – A binary indicator of whether it was raining or not**

2. **Precipitation-A measure of how hard it was raining at the time. The T-test indicated that it was the intensity of the rain, and not just the presence of rain that impacted ridership so this input variable is extremely important.**

3. **Hour –The time of day of the measurement to account for lower ridership in non-peak hours.**

4. **Mean Temperature—This variable is to account for lower or higher ridership attributed to temperature and not whether its raining or not or whether it is a combination of rain and temperature that impacts ridership.**

5. **Mean Wind Speed-This variable is to account for lower or higher ridership on account of how windy it is. Since the amount of precipitation has an impact on ridership, as shown in the T-Test, and higher wind speeds usually accompany higher precipitation, this variable could further show the significance that more precipitation impacts ridership.**

6. **Mean Pressure –Another indicator of the whether it was raining and to account for riders predicting of rain in the near future who would then take or not take the subway. The pressure drops when it is about to rain and when it is raining**

I did use a dummy variables filling in for column labeled 'UNIT' to account for any instances when not all the turnstiles stations were reported.

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

Input Variable	Weight
Rain	-27.19
Precipitation	18.11
Hour	67.4
Mean Temperature	-5.57
Mean Wind Speed	21.03
Mean Pressure	-264.79
Intercept	9017.64

2.5 What is your model's R^2 (coefficients of determination) value?

My model's coefficients of determination value is 0.458

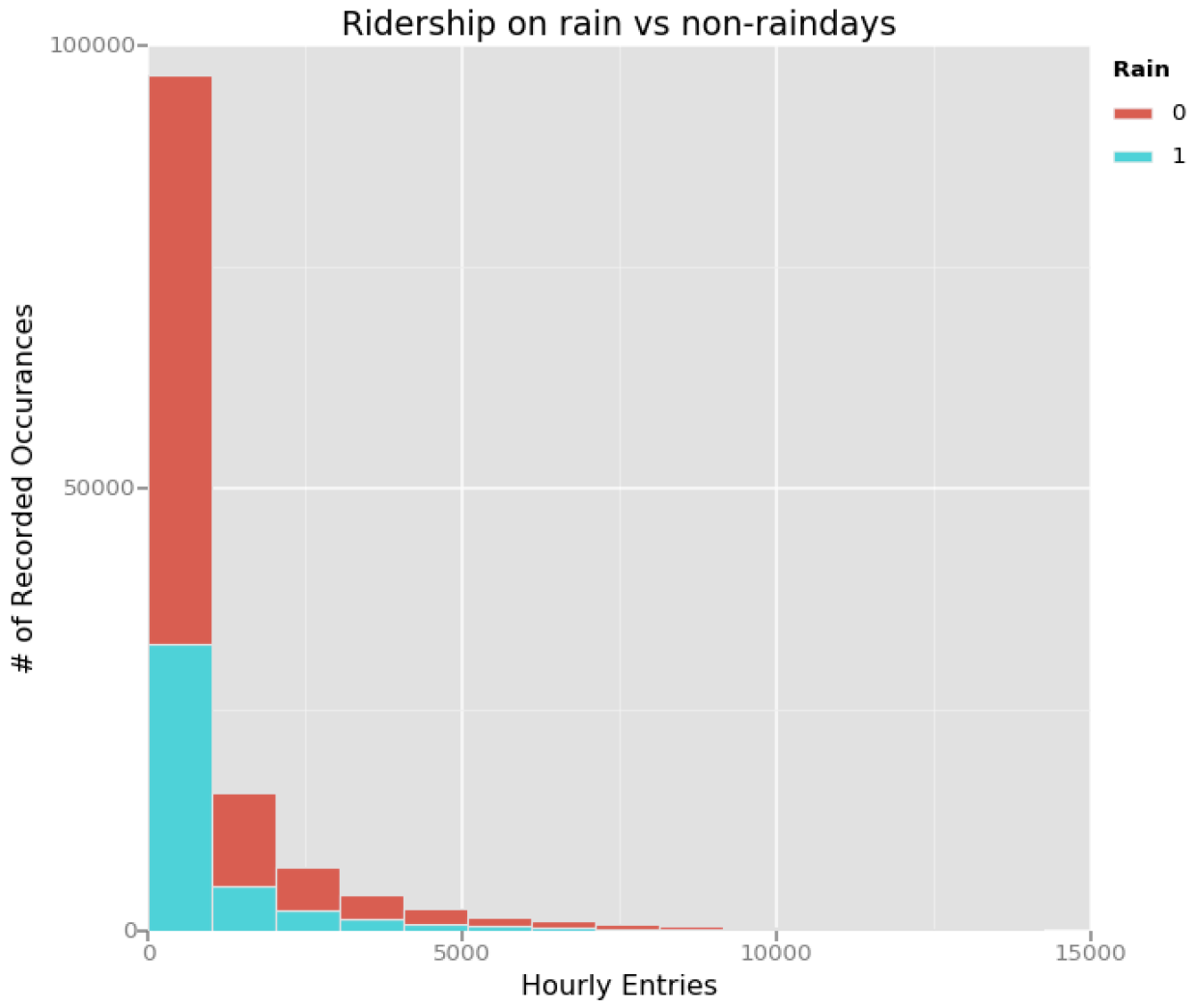
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

This R^2 value means that the regression model is a moderate fit. While the R^2 value should be as close to one, 0.45857 is not necessarily a bad value as a linear regression is probably not the best model for predicting ridership due to the high number of variables. Also, since the linear regression was created to determine only whether rain impacts ridership, there is not a need for an extremely precise prediction model (high R^2 value).

Section 3. Visualization

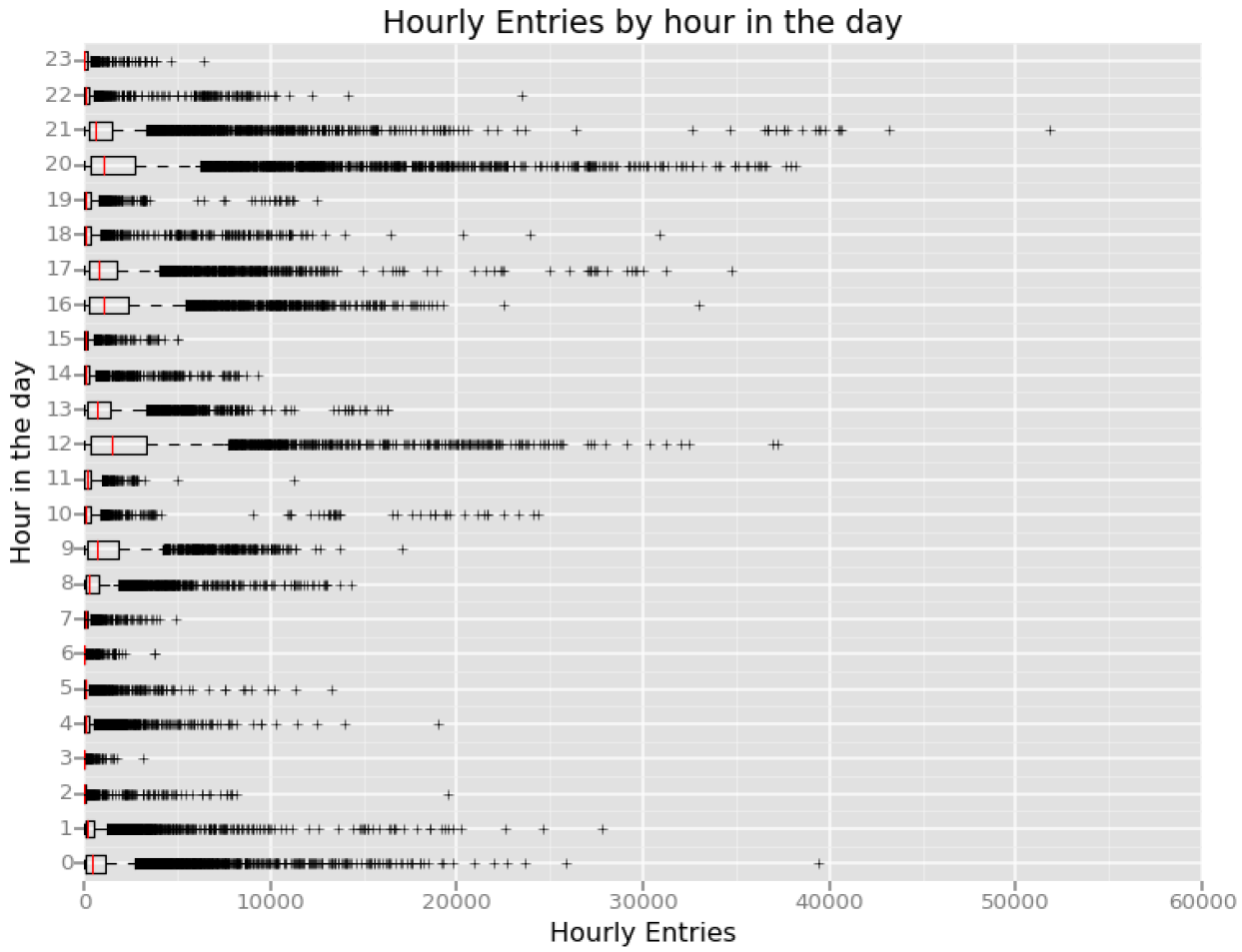
Please include two visualizations that show the relationships between two or more variables in the NYC Subway data

Visualization #1: Histogram of ENTRIESn_hourly for days with rain (1) and days without rain (0)



This Histogram shows how much more frequently people ride when it is not raining. The teal highlighted boxes represent the number of occurrences of the hourly entries for when it rains and the red highlighted boxes are for the hourly entries for when it is not raining. While both groups diminish in size as 'Hourly Entries' increases, the 'rain' group is not registered after 8000 while the 'not rain' group still have recordings when hourly entries are around 14,000. From this we infer that ridership generally increases in when it is not raining.

Visualization #2: Boxplot of how ridership is impacted by the hour in the day



This boxplot graph is a deeper analysis on the how the hour of the day impacts ridership on the subway. Each hour has the same number of data points over the same span of the month of May in 2011. From this chart we can see that ridership is at its greatest around 20:00 (8pm) and second greatest at 12:00 (12pm).

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?

From my analysis, it appears that people ride the subway more 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.

Using a Mann-Whitney U test to compare the rain vs. non-rain hourly entries returned a P-value of 4.99% which does not exceed the P-critical value of 5%. Because of this, the null hypothesis was disproven and that ridership does go up by 1.375% when it rains.

The Mann-Whitney U test results are corroborated by the linear regression which shows that also shows that ridership increases when it rains. This is shown by the weight the linear model applies to the 'Mean Pressure' variable which measures air pressure. Rain occurs when air pressure is low so a high pressure means that it is not raining. Since the 'Mean Pressure' weight is larger than the next most predictive variable 'Hour (-264.79 to 67.4). Since the 'Mean Pressure' weight is a negative value, increased air pressure, which correlates with not raining, would lower the expected ridership. Because of this the linear model also shows that ridership increases when it rains.

The linear regression test is further validated by comparing its R^2 value of 0.45857 to a linear model that does only takes into account time and not weather conditions has an R^2 value of 0.4575. Since 0.45857 is greater than 0.4575, having taking into account variables that deal with rain and rain indicating variables does make the R^2 value higher and increases the accuracy of the model.

In conclusion, both the Mann-Whitney U statistical test and the linear regression model both support the claim that ridership increases when it rains.

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.

Regarding the dataset, there are two values I believe would have increased the accuracy of the results and would lead to a stronger conclusion:

1. **Thunder—While 'thunder' was part of the data set, every single value was '0' meaning that at no point during all of data collection and 44,000 recording of rain, was there ever also a recording of thunder. By including data where thunder occurred, I could further determine whether heavier rain alone decreases ridership or if there were other inclement weather conditions that affect ridership.**

2. **Location—The NYC subway is an extremely large system with diverse ridership that extends includes commuter stations from the outer boroughs of NYC, and large number of stations catering to tourists. As such, the ridership behavior for the Brooklyn subset of the NYC subway system could be fundamentally different than that of Queens or Manhattan. The data set offers no way to determine the location of the turnstiles and so being unable to isolate turnstile by location unnecessarily muddies the data.**

Regarding the analysis, the main potential shortcoming was relying on a linear model for analysis. A linear model is very dependent on linear relationships among the variables which could not be applied to several important variables in the NYC subway data. These non-linear variables include the hour of day and binary variables like thunder and fog. Because of this, quadratic regression or other more complex models could have provided more accurate predictions than linear regression as they can better account for non-linear variables. Also, while Mann-Whitney U tests are useful, the amount of statistically significant data an observer can draw from them is limited since they can only disprove the null hypothesis and not fully affirm any other hypotheses. With that in mind, there could be improvements to the analysis of NYC subway ridership with more data, more rigorous statistical tests and more complicated models but the current analysis as is does still show that rain does in fact negatively impact ridership.

Section 6. References:

1. **All examples for the code used in this project was provided by the Udacity "Intro to Data Science" course.**
2. **The example code for the box plot was found on the Udacity discussion forum, *https://discussions.udacity.com/t/creating-boxplots-in-python-ggplot/28363***
3. **Information used to determine the worth of the R^2 value was found here: http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis**
4. **Information regarding the Mann Whitney U Test was found here: https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test#Assumptions_and_formal_statement_of_hypotheses**
5. **Information regarding the decision to have a one-tailed P value was found here: http://www.graphpad.com/guides/prism/6/statistics/index.htm?one-tail_vs__two-tail_p_values.htm**

Python code

```
import numpy
import scipy
import scipy.stats
import pandas
import statsmodels.api as sm
from ggplot import *

#with open('turnstile_data_master_with_weather.csv') as csvfile:
        #spamreader = csv.DictReader(csvfile)
        #for row in spamreader:


#f = open('turnstile.csv')
#turnstile = csv.reader(f)
turnstileweather = pandas.read_csv('turnstile.csv')
turnstile = pandas.DataFrame(turnstileweather)

print ('mean for everyone is %s' % numpy.mean(turnstile['ENTRIESn_hourly']))

rain = turnstile[turnstile.rain == 1]
no_rain = turnstile[turnstile.rain ==0]
print('length of rain is %s' %len(rain))
print('length of no rain is %s' %len(no_rain))
print ('mean for rain is %s' % numpy.mean(rain['ENTRIESn_hourly']))
print ('mean for no rain is %s' % numpy.mean(no_rain['ENTRIESn_hourly']))

u,p = scipy.stats.mannwhitneyu(rain["ENTRIESn_hourly"],no_rain["ENTRIESn_hourly"])
p = p*2
print ("u is %s" % u)
print ("p is %s" % p)

print ("p <= 0.05 so I reject the null hypothesis")

print('//////')
print('//////')


print('//////')
print('//////')

def linear_regression(features, values):
        features = sm.add_constant(features)
```

```
        model = sm.OLS(values, features)
        results = model.fit()
        intercept = results.params[0]
        params = results.params[1:]
        return intercept, params


features = turnstile[[ 'Hour']]
dummy_units = pandas.get_dummies(turnstile['UNIT'], prefix='unit')
features = features.join(dummy_units)

values = turnstile['ENTRIESn_hourly']

intercept, params = linear_regression(features, values)

predictions = intercept + numpy.dot(features, params)

print("the intercept is %s" % intercept)
print("the weights are %s" % params)
#print (intercept, params)
print ("my predictions")
print (predictions)


top = ((predictions - turnstile['ENTRIESn_hourly'])**2).sum()
bottom = ((turnstile['ENTRIESn_hourly']-numpy.mean(turnstile['ENTRIESn_hourly']))**2).sum()

r_squared = 1 - top / bottom
print ("r squared is %s" % r_squared)

pandas.options.mode.chained_assignment = None
turnstileweather['UNIT'] = turnstileweather['UNIT'].map(lambda x: float(x.lstrip('R')))

water = ggplot(turnstileweather, aes(x = 'ENTRIESn_hourly', fill =
'rain'))+geom_histogram(binwidth = 1000)+ggtitle("Ridership on rain vs non-
raindays")+labs("Hourly Entries","# of Recorded Occurances")+xlim(0,15000)+ylim(0,100000)

fire = ggplot(rain,aes(y ='Hour', x ='ENTRIESn_hourly')) + geom_boxplot()+labs(x= 'Hourly
Entries', y= 'Hour in the day')+ggtitle('Hourly Entries by hour in the day when raining')
print (water)
```