P2: Analyzing the NYC Subway Dataset

Supporting Course: Intro to Data Science
Nanodegree: Data Analyst
Andri Rizhakov

Revision history:
Rev. 0. 08/31/2015. Initial submission.
Rev. 1. 09/20/2015. First revision submission.

Section 0. References

References used in answering above questions:

1. Notes from Udacity's Intro to Data Science course.
2. Intro to Data Science course, problem statement [P2: Analyzing the NYC Subway Dataset].
3. https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test.
4. https://en.wikipedia.org/wiki/Welch%27s_t_test.
5. https://en.wikipedia.org/wiki/Level_of_measurement#Ordinal_scale.
6. http://www.statisticssolutions.com/mann-whitney-u-test/.
7. https://en.wikipedia.org/wiki/P-value.
8. http://blog.minitab.com/blog/adventures-in-statistics/how-to-correctly-interpret-p-values.
9. https://en.wikipedia.org/wiki/Coefficient_of_determination.
10. http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit.
11. http://matplotlib.org/api/pyplot_api.html#matplotlib.pyplot.boxplot.
12. https://en.wikipedia.org/wiki/Box_plot.
13. http://matplotlib.org/examples/statistics/boxplot_demo.html.
14. http://www.physics.csbsju.edu/stats/box2.html.
15. http://matplotlib.org/examples/pylab_examples/boxplot_demo2.html.

Section 1. Statistical Test

1.1

The Mann–Whitney U test was used because exploratory data analysis via histogram revealed that the distributions were not normal. Although there may be other statistical tests that also allow for non-normal distributions, the Mann–Whitney U test was used because it was covered during class.

One-tail P value was used because the histogram distributions from exploratory data analysis revealed their non-normal, asymmetric distribution to have only one tail.

The experiment involves 2 conditions (rain condition, or non-rain condition). The objective is to understand if there is a statistically significant difference between the conditions, i.e., that two samples come from the same population. Therefore, the null hypothesis is that two samples come from the same population and that there is no statistical difference between the sample means, and the alternative hypothesis that two samples come from different populations and that there is a statistical difference between the sample means. More succinctly:

Appropriate set of hypotheses:

- $H_0 : \mu_R - \mu_N \leq 0$
- $H_A : \mu_R - \mu_N > 0$

where: H_0 is null hypothesis,
H_A is alternative hypothesis,
μ_R is population mean of the rain condition, and
μ_N is population mean of the non-rain condition

The p-critical (p*) value is 0.05.

1.2

The Mann–Whitney U test is applicable to the dataset because no restrictions on the distribution type are made. Additionally, the below assumptions are made before using the test:

1) All the observations from both groups are independent of each other; samples are not paired, so it is reasonable that this assumption is valid.
2) The responses are ordinal (i.e. one can at least say, of any two observations, which is the greater). The samples represent dichotomous, ordinal data, so it is reasonable that this assumption is valid.
3) Under the null hypothesis H_0, the distributions of both populations are equal.
4) Under the alternative hypothesis H_A, the distributions of both populations are not equal.

1.3

The results from the test are:

with_rain_mean = 1105.4463767458733
without_rain_mean = 1090.278780151855
p = 0.024999912793489721

1.4

The null hypothesis, H_0, can be rejected because actual (p = 0.024999912793489721) < (p* = 0.05), suggesting that the observed event is statistically unlikely to occur from pure randomness in population alone; there is a 2.5% probability of obtaining an effect at least as extreme as the one in the sample data, assuming the truth of the null hypothesis.

An interpretation of the results can be that there appears to be an increase in subway ridership when it is raining over when it is not raining, given the small probability of the observed effect occurring by chance. These results match up with my expectations because it is logical that people would seek shelter in the subways, rather than choose to walk in the rain. Because the data was collected in May, it is possible that there are more tourists visiting NYC, some would choose to travel via subways system (due to probable saturation of taxis) than stroll around the city.

Section 2. Linear Regression

2.1

OLS using Statsmodels was used to compute coefficients theta and thus predict ENTRIESn_hourly.

2.2

The follow features were used in the regression model: 'rain', 'precipi', 'Hour', 'meantempi', 'fog','meanwindspdi','mintempi','maxtempi'. Additionally, the dummy variable for 'UNIT' was used as part of the features.

2.3

The features were used for the following reasons:

1) 'rain': As discovered from the statistical testing of Section 3, this turned out to be an important feature. Intuitively, if it is raining, I expect people to take cover in the subway to make progress on their transportation, as taxis are taken up.
2) 'precipi': Similar analogy to the 'rain' feature; if it is precipitating, I expect people to take cover in the subway to make progress on their transportation, as taxis are taken up. Although the dataset is taken in May, it is possible that there was a non-rain type of precipitation such as snow (It is New York after all :)).
3) 'Hour': Intuitively, the hour of the day is vital because people's behavior will correlate with the time of day. For example, subway usage is expected to peak during rush hour.

4) 'meantempi': Intuitively, if the average temperature for the day was either too high or too low, I expect people to take cover in the subway to make progress on their transportation, since otherwise it would be uncomfortable and other modes are not available.
5) 'fog': Intuitively, if there is fog, I expect interference to people's visibility. Thusly, cautious bikers and drivers may err on the side of caution and take cover in the subway to make progress on their transportation.
6) 'meanwindspdi': Intuitively, if there is high winds, I expect interference to people's comfort. Thusly, annoyed pedestrians may take cover in the subway to make progress on their transportation.
7) 'mintempi': Related to the mean temp argument above. Intuitively, if the minimum temperature for the day was too low, I expect people to take cover in the subway to make progress on their transportation, since otherwise it would be uncomfortable and other modes are not available.
8) 'maxtempi': Related to the mean temp argument above. Intuitively, if the maximum temperature for the day was too high, I expect people to take cover in the subway to make progress on their transportation, since otherwise it would be uncomfortable and other modes are not available.
9) 'UNIT' (dummy variable): Commenting out the code to remove the dummy variable made my R^2 decrease drastically to (R^2 = 0.0333) for the same features listed above. Thusly, I chose to retain the 'UNIT' dummy variable.

2.4

The parameters of the non-dummy features in the linear regression model are:

rain	22.835280
precipi	-62.418887
Hour	65.420604
meantempi	-117.419472
fog	244.405729
meanwindspdi	28.274234
mintempi	40.208117
maxtempi	65.054290

2.5

The model's R^2 (coefficients of determination) value is 0.4812.

2.6

The value of 0.4812 means that the goodness of fit for the regression model is average because ~48% of the total variability around the sample mean is explained by the model.
Yes, this linear model to predict ridership is appropriate for this dataset, given R^2 of 0.4812. Because ~48% of the total variability around the sample mean is explained by the model, ~52% is explained by other variables. When used for prediction purposes, the predicted value should

be reported along with an uncertainty of ±~52% around the predicted value. Depending on what the needs of the client are or the impact of how the predicted value is to be used, and as long as all assumptions and results are explicitly stated, there are no technical issues with using the model for predictive purposes.

Section 3. Visualization

3.1

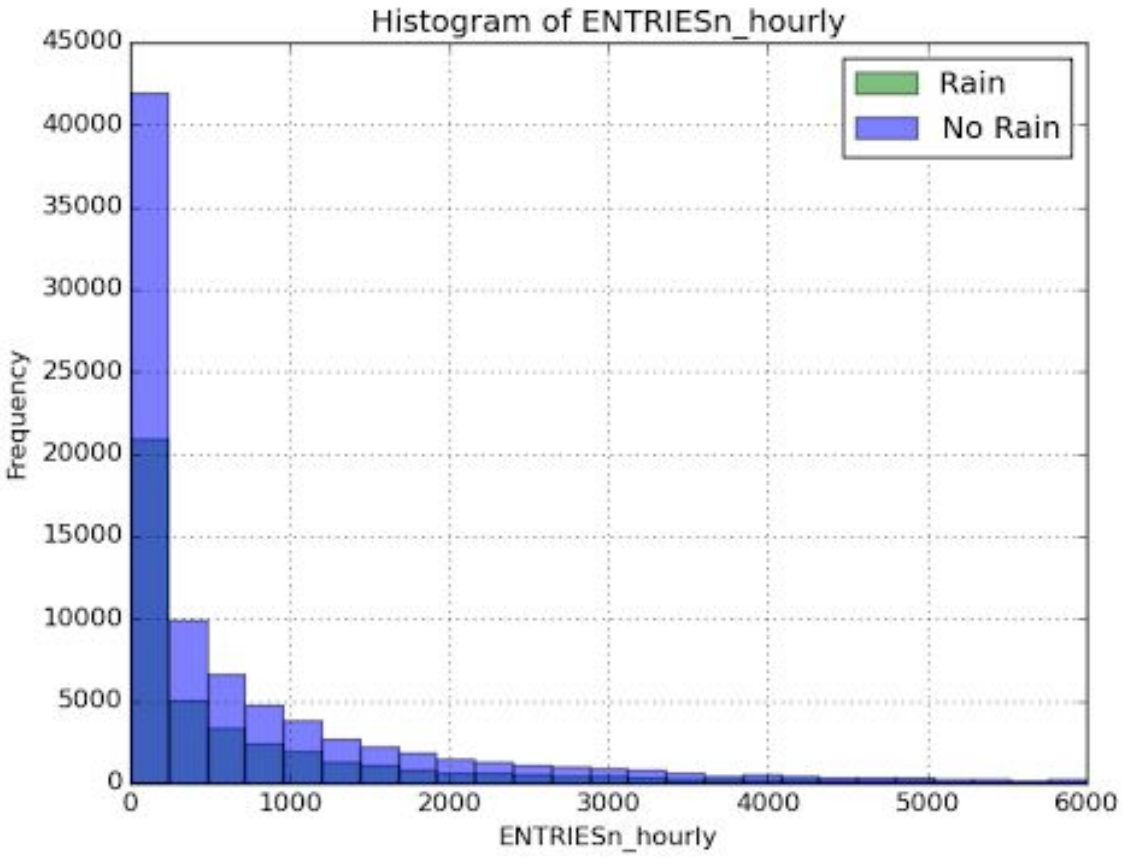
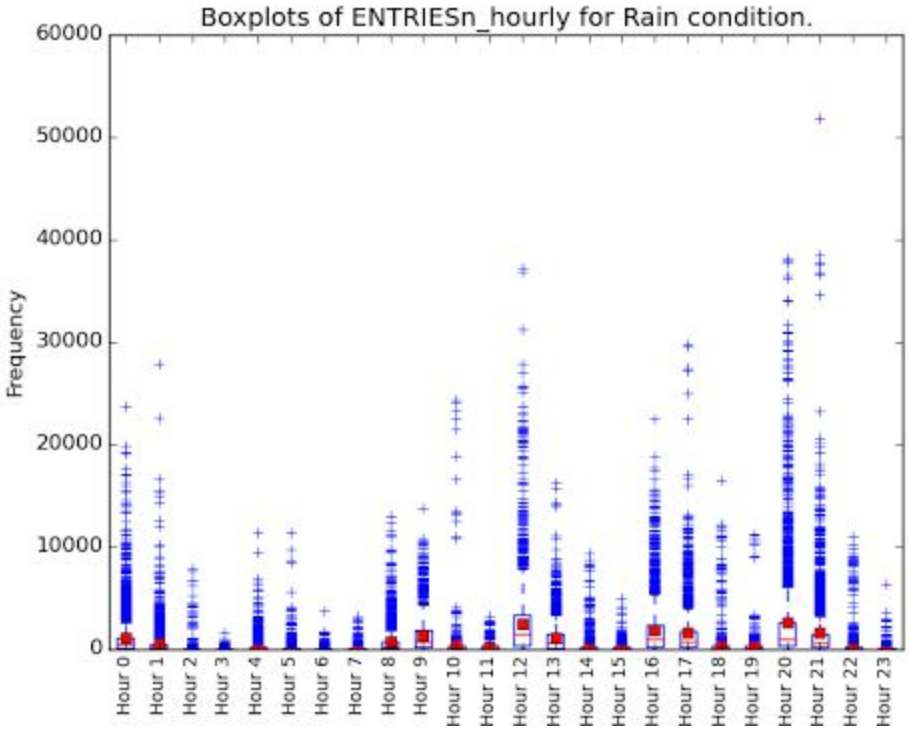
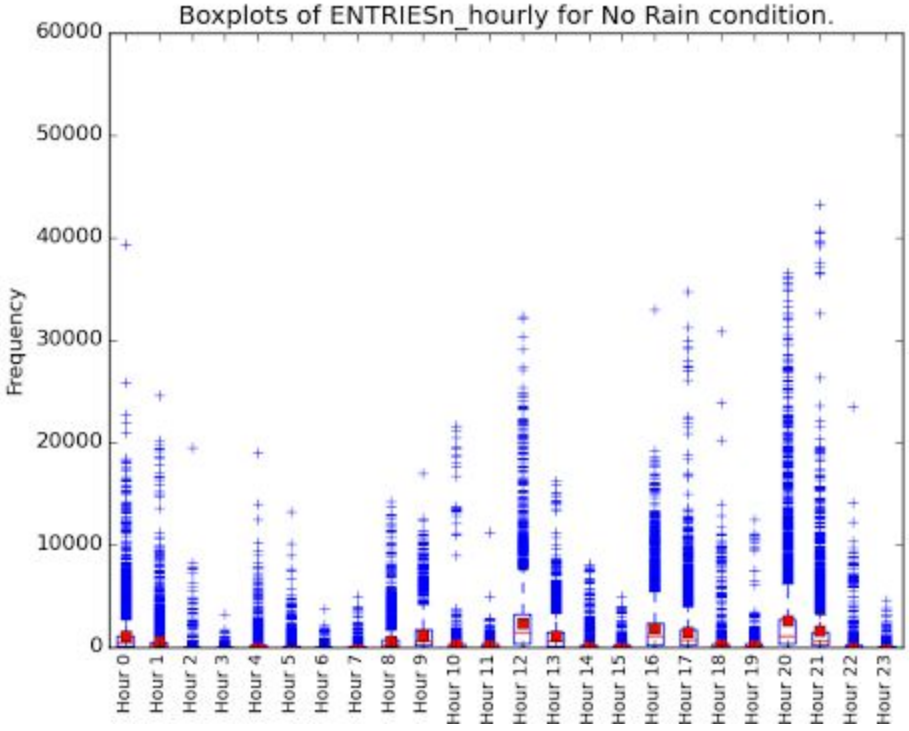
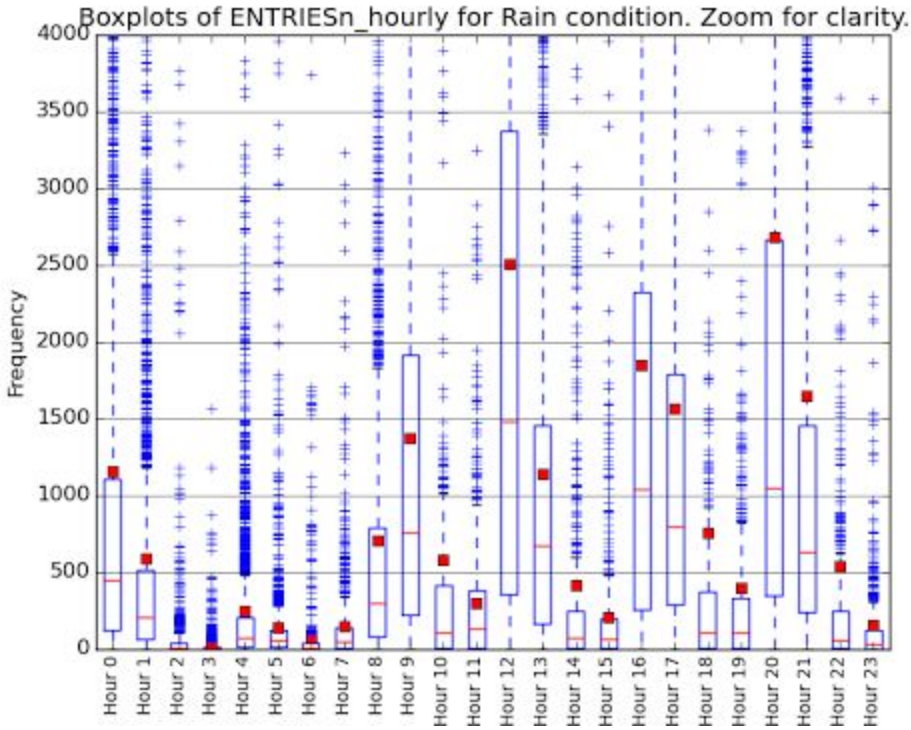


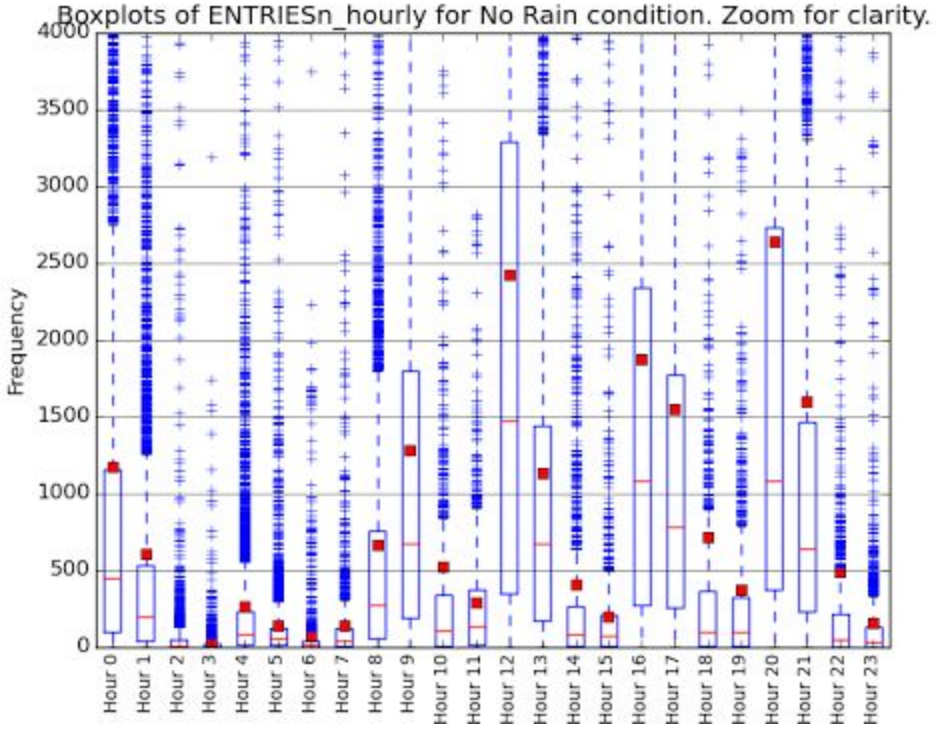
Figure “Histogram of ENTRIESn_hourly” above shows the frequency of ENTRIESn_hourly, binned into equal bin sizes, ranging from 0 to 6000 in each ENTRIESn_hourly row. It appears that there are more no-rain data (blue) than rain data (green, hidden behind the translucent blue), which makes intuitive sense for a typical month for NYC. Furthermore, relative to the total amount of entries in each group (rain or no-rain), the frequency of rain values seems to be higher than no-rain. This indicates that the tail for the rain data is distributed with a thicker tail than the no-rain data. The thicker tail should also shift the mean value of the data set in the direction of the tail.

3.2


Boxplots of ENTRIESn_hourly for Rain condition.
Frequency
60000
50000
40000
30000
20000
10000
0
Hour 0
Hour 1
Hour 2
Hour 3
Hour 4
Hour 5
Hour 6
Hour 7
Hour 8
Hour 9
Hour 10
Hour 11
Hour 12
Hour 13
Hour 14
Hour 15
Hour 16
Hour 17
Hour 18
Hour 19
Hour 20
Hour 21
Hour 22
Hour 23


Boxplots of ENTRIESn_hourly for No Rain condition.
Frequency
60000
50000
40000
30000
20000
10000
0
Hour 0
Hour 1
Hour 2
Hour 3
Hour 4
Hour 5
Hour 6
Hour 7
Hour 8
Hour 9
Hour 10
Hour 11
Hour 12
Hour 13
Hour 14
Hour 15
Hour 16
Hour 17
Hour 18
Hour 19
Hour 20
Hour 21
Hour 22
Hour 23


Boxplots of ENTRIESn_hourly for Rain condition. Zoom for clarity.
Frequency
0
500
1000
1500
2000
2500
3000
3500
4000
Hour 0
Hour 1
Hour 2
Hour 3
Hour 4
Hour 5
Hour 6
Hour 7
Hour 8
Hour 9
Hour 10
Hour 11
Hour 12
Hour 13
Hour 14
Hour 15
Hour 16
Hour 17
Hour 18
Hour 19
Hour 20
Hour 21
Hour 22
Hour 23


Boxplots of ENTRIESn_hourly for No Rain condition. Zoom for clarity.
Frequency
0
500
1000
1500
2000
2500
3000
3500
4000
Hour 0
Hour 1
Hour 2
Hour 3
Hour 4
Hour 5
Hour 6
Hour 7
Hour 8
Hour 9
Hour 10
Hour 11
Hour 12
Hour 13
Hour 14
Hour 15
Hour 16
Hour 17
Hour 18
Hour 19
Hour 20
Hour 21
Hour 22
Hour 23

Figures below shows subway ridership vs. hour of the day in form of box plot. The box contains and lower first quartile and upper third quartile. The red line in each box is the median. The red dot is the mean. The dotted blue line is the fence, 1.5 IQR (interquartile range) lengths from the third quartile. The blue crosses ('+') represent outliers, per common box plot plotting.

It appears that there is roughly the same shape of distribution and same amount of high values for people taking the subway when it is raining and when it is not; because there may be a lower total amount of riders when it rains, this results in a higher relative ridership when it rains. Slightly greater hour averages for the Rain condition vs. No Rain condition (e.g., Hour 12, 20, 22) indicate a possible small increase in ridership.

Section 4. Conclusion

4.1

From the analysis and interpretation of the data, more people ride the NYC subway when it is raining. The results from statistical tests, linear regression, and data visualizations support this conclusion.

4.2

statistical tests

The null hypothesis, H_0, can be rejected because actual (p = 0.024999912793489721) < (p* = 0.05), suggesting that the observed event is statistically unlikely to occur from pure randomness in population alone; there is a 2.5% probability of obtaining an effect at least as extreme as the one in the sample data, assuming the truth of the null hypothesis. An interpretation of the results can be that there appears to be an increase in subway ridership when it is raining over when it is not raining, given the small probability of the observed effect occurring by chance.

linear regression

The positive value on the parameter associated with 'rain' (value of 22.8) indicates the increase in ridership when it rains.

data visualizations

- From the histogram in the exploratory data analysis portion, it appears that there are more no-rain data than rain data, which makes intuitive sense for a typical month for NYC. Furthermore, relative to the total amount of entries in each group (rain or no-rain), the frequency of rain values seems to be higher than no-rain. This indicates that the tail

for the rain data is distributed with a thicker tail than the no-rain data. The thicker tail should also shift the mean value of the data set in the direction of the tail.

- From the box plot, it appears that there is roughly the same shape of distribution and same amount of high values for people taking the subway when it is raining and when it is not; because there may be a lower total amount of riders when it rains, this results in a higher relative ridership when it rains. Slightly greater hour averages for the Rain condition vs. No Rain condition (e.g., Hour 12, 20, 22) indicate a possible small increase in ridership.

Section 5. Reflection

5.1

Potential shortcomings of the methods of your analysis include the dataset and the analysis. Nevertheless, for the purposes of this lesson, a preliminary investigation into rain vs. no-rain subway ridership is appropriate with the dataset and analysis used.

Dataset

There are a few shortcomings with regards to the dataset:

- the time resolution seems to be every 4 hours for each meter station in the csv, and the same resolution is preserved for the same meter. For example, “R001” is sampled at hours 1,5,9,13,17,21. I believe this filter is too high and may blend some important trends.
- a truncated set of May 2011 may have outlier data due to anomalous behavior during the month. Conclusions drawn here may not be transferable to other months.

Analysis

There are a few shortcomings with regards to the dataset:

- the parameters for the meters are orders of magnitude higher than the rain parameter.

```
params:

rain               22.835280
precipi           -62.418887
Hour               65.420604
meantempi        -117.419472
fog               244.405729
```

```
meanwindspdi      28.274234
mintempi          40.208117
maxtempi          65.054290
unit_R001       4092.398250
unit_R002       -947.002253
unit_R003      -1251.229903
unit_R004      -1077.055372
unit_R005       -435.720410
unit_R010       4521.043861
unit_R011       7036.225456
unit_R015       -343.372102
unit_R016       -600.074037
unit_R022       8513.180630
unit_R451       -626.720846
unit_R452       4908.084887
unit_R453         -57.409925
unit_R454      -1218.303871
unit_R462         19.441434
unit_R463       1528.945370
unit_R464      -1459.927690
unit_R550      -1565.297794
unit_R551      -1481.950375
unit_R552      -1462.390844
```

This indicates a stronger relationship for the various meter units in comparison to the rain parameter. Thus, knowing the meter unit is of greater importance and leads to a better prediction. This was also tested in Section 3, and the R^2 fell to 0.0333 from 0.48 when the unit were not part of the parameters. There appears to be a geographic dependency that is vital for the regression model.

- The statistical analysis showed that there is a ~2.5% probability of obtaining an effect at least as extreme as the one in the sample data, assuming the truth of the null hypothesis (no difference in population means for subway ridership in rain or no-rain conditions). If true, this effect is not exceedingly small: 2.5% is small but still possible. Reducing the

p-critical value to 0.01 would have resulted in a conclusion of failing to reject the null hypothesis.

- Furthermore, the magnitude of the effect (rain condition on subway ridership), even if true by the statistical test (i.e., shows statistically significant increase in rain sample mean over no-rain, indicating 2 separate population distributions), the relative magnitude is small. Specifically, 1105.4463767458733 / 1090.278780151855 = 1.014, or ~1.4% increase in subway ridership. Such a small relative increase may be in the random noise of the dataset, with other hidden variables unaccounted for that may have contributed to the 1.4% increase.

Appendix A. Code summary and Calculations

Calculations were performed in the following files. These are attached to the report submission folder.

1. Titanic exercises, pandas intro
2. Data wrangling exercises, manipulate files/rearrange
3. Analysis
    - 3.1. histogram (exploratory data analysis)
    - 3.2. N/A
    - 3.3. Mann Whitney U-test
    - 3.4. N/A
    - 3.5. Linear Regression
    - 3.6. plot of residuals (that is, the difference between the original hourly entry data and the predicted values)
    - 3.7. compute r^2
    - 3.8. gradient descent
4. ggplot [pandas + ggplot]
    - 4.1. plot 1
    - 4.2. plot 2

Appendix B. Reviewer Comments and Disposition

Reviewer Comment	Disposition in latest version
“Quality of Visualizations”. SPECIFICATION All plots and data are of the appropriate type. DOES NOT MEET SPECIFICATION Reviewer Comments A scatterplot is not an appropriate type to summarise the distributions of entries for different time periods. It suffers strongly from overplotting, so information about the density of points is lost for most values of hourly entries. A different type of plot might be a better way to visualise this information. A simple bar plot showing average or total entries for each time period, or a box plot or violin plot to describe the distributions for each time period are possible alternatives.	Section 3.2 of current document revised to reflect changes.