# Analyzing the NYC Subway Dataset

## Short Questions

## Overview

This project consists of two parts. In Part 1 of the project, you should have 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. Please use this document as a template and answer the following questions to explain your reasoning and conclusion behind your work in the problem sets. You will attach a document with your answers to these questions as part of your final project submission.

## Section 1. Statistical Test

1.1

- **Which statistical test did you use to analyse the NYC subway data?**
  - I used a Mann—Whitney *U* test comparing the distributions of ENTRIESn_hourly variable between `rain == 1` and `rain == 0` data of the `turnstile_weather.csv` dataset.
  - I also conducted a *T* test; dealing with the non-normal distribution of rain/norain by adjusting the data <- log(data + 1).
    (a) + 1 was to prevent +-inf values from values of 0
    (b) The data looked more like a normal distribution, and the *T* test yielded an even more significant result (below – Section 1.3)

- **Did you use a one-tail or a two-tail P value?**
  - I used a two-tail P value because we are not defining a difference in any particular direction (less/greater than), instead, we are looking to see if there are differences at all.

*(continued)*

❖ **What is the null hypothesis?**
- The null hypothesis is that both populations (rain and no-rain) are the same.

❖ **What is your p-critical value?**
- I used a p-critical value of 0.05

1.2

❖ **Why is this statistical test applicable to the dataset?**
- After creating histograms for both distributions, it was apparent that the distribution is non-normal; skewed towards zero. The *U* test is ideal on non-normal distributions over the *T* test by using the sum of the ranks on both sets of observations and comparing them, rather than taking the observations at face-value.

❖ **In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.**
- The observations from both groups are independent of each other
- The responses are ordinal
- Under the null hypothesis, the distributions are equal
- Under the alternative hypothesis, the probability of one group's observation exceeding the others is not equal to 0.5

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.**
- mean(rain) : 1105.446
- mean(norain) : 1090.279
- *U* test – two-sided p-value (R) : 0.04988
- *U test* – two-sided p-value (Python) : 0.03862
- *T test* – two-sided p-value (R, Python; log(data + 1) adjusted) : 0.002519

1.4

❖ **What is the significance and interpretation of these results?**
- The critical value is set to be the point at which you can reject the null hypothesis that the two distributions are the same
- The critical value was set to 0.05 and both results (R, Python) yields results that reject the null hypothesis.
- The distributions are not the same.

# 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 Ordinary Least Squares to make the model

2.2

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

- UNIT, DATEn, Hour
- All of my variables were dummy variables

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 removed variables that had no variance (DESCn, thunder), that were redundant (TIMEn is represented by Hour), and the indexing variable (X).
- We are left with [UNIT, DATEn, Hour, maxpressurei, maxdewpti, mindewpti, minpressurei, meandewpti, meanpressurei, fog, rain, meanwindspdi, mintempi]
- Studying the coefficients of the OLS model on the variables reveal multicollinearity issues
  a) While [UNIT, DATEn, Hour] had coefficients, the rest returned NA for coefficients
  b) This means…
  - that the these variables can be linearly predicted from the other variables in the model
  - these variables are highly correlated to other variables and provide no significant improvement to the accuracy of prediction

2.4

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

- I did not use non-dummy features

2.5

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

- 0.5155325

2.6

❖ **What does this $R^2$ value mean for the goodness of fit for your regression model?**

- The model explains ~51% of the total variation in the data

❖ **Do you think this linear model to predict ridership is appropriate for this dataset, given this $R^2$ value?**

- Depends on the level of accuracy required
- I believe this is the maximum $R^2$ achievable considering none of the other variables in the set adds to the predictability of the model
- If the required model only needed to predict a range of entries, i.e.3000-4000, I would break those values in to categories and run a classification tree based model which would most definitely increase the accuracy because only a ‘ballpark’ is needed.

# Section 3. Visualization

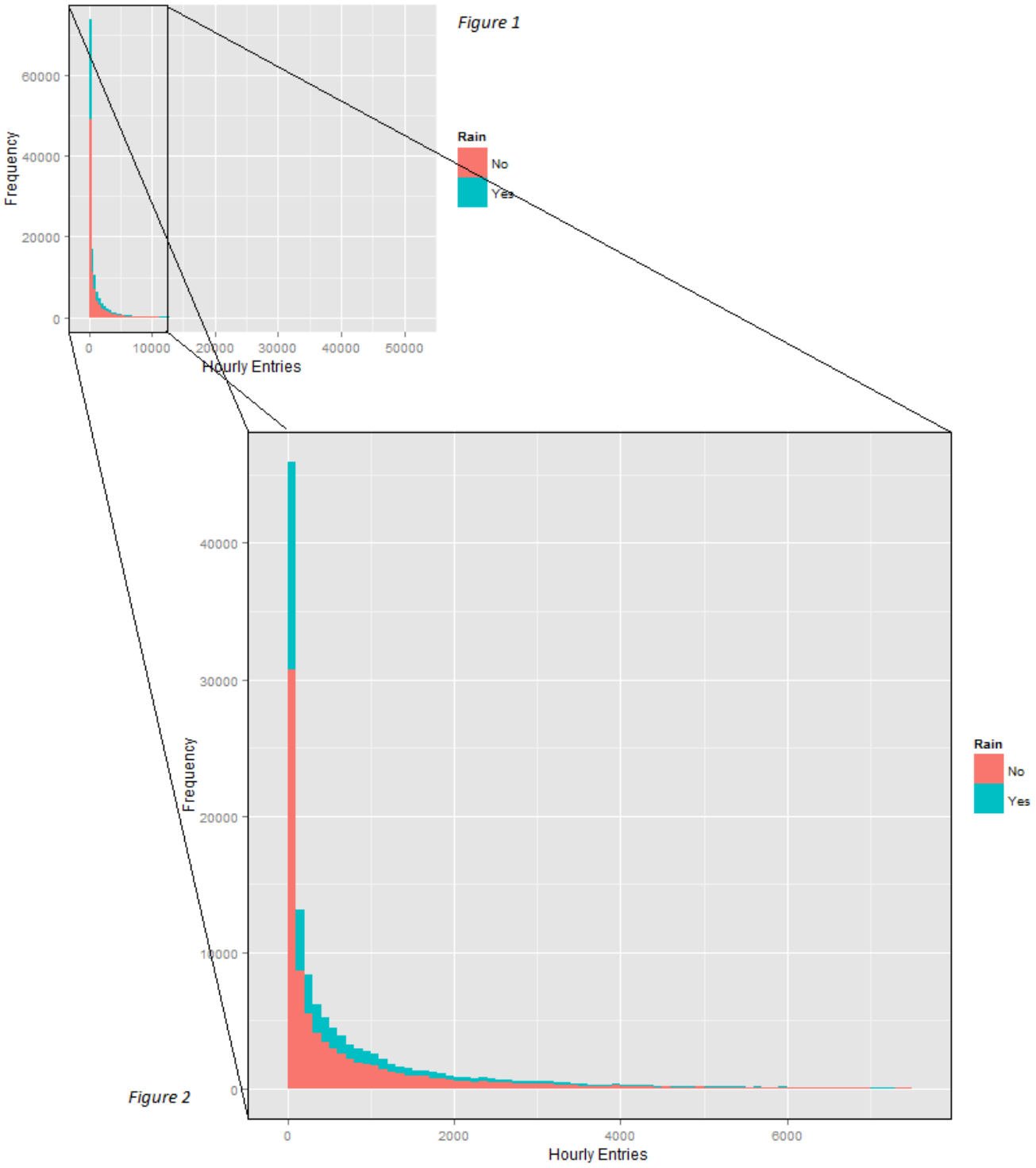


*Figure 1* shows a histogram of the dataset colored by 'Rain'.

*Figure 2* zooms in on *Figure 1* to show how skewed the distribution is.

***This key feature of the data led me to use Mann-Whitney U test over the T test.***

*(continued)*


*Figure 3*

*Figure 3 shows…*

- Stations with lower R# (left of plot) had more entries
- Weekend riders rode later in the evening than Weekday riders
- More people rode in the weekday on average than the weekends
- Ridership categorically shifts between the days, not linearly

***Although this visualization offered many great insights into patterns in ridership, entering Weekends as a categorical variable did not offer any increase in the model's Rsquared, so it was not included in the final model.***

***However, entering the Hour variable as a dummy increased accuracy considerably.***

# 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?**

If you are considering the means of both categories (rain = 1105.446, no-rain = 1090.279), it would seem that more people on average use the subway when it is raining on average. If the question has to do with absolute values, 95.7 million rode when there was no rain in the month of May vs. 48.7 million when there was rain. This information alone should be used to draw any conclusions since there is close to double the observations of no-rain days over rain days.

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.**

The Mann—Whitney *U* Test with a p-critical value of 0.05 rejected the null hypothesis that the distributions are the same with a p-value of 0.04988. I also adjusted for the skewness of the ENTRIESn_hourly frequency table by using the log(data + 1). This created a normal distribution without +-inf values. This allowed me to confidently apply a *T* test which also rejected the null hypothesis that the distributions are the same with a p-value of 0.002519. A *U* test p-value for mean(norain) < mean(rain) = 0.02494, meaning with a critical value of 0.05, mean(norain) < mean(rain) is True.

The linear regression model pointed out that if the end goal is to simply predict the hourly entries, utilizing data pertaining to rain/norain did not offer any substantial accuracy to the data, in fact, the rain/norain variable showed strong colinearity with the categorical variables. Meaning, the values in rain/norain could be explained linearly by the other variables. In conclusion, although the occurrence of rain does increase the use of NYC subways, utilizing rain a variable will not help the predictability of the model.

# 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**
    - The majority of the variables were not used to find the best prediction model
    - Intuitively, weather should affect the use of the subway, but according to the model, weather variables offered no increase in accuracy
  - **2. Analysis, such as the linear regression model or statistical test.**
    - Rsquared only accounted for ~51% of the variance
      - This would indicate that there are missing variables that would increase accuracy
    - I would have liked to have used a classification tree on this model
    - A tree was out of the question; required too much computer resource
    - In real world applications, a range for the entries predicted would be sufficient versus an exact value.
    - A classification tree would not only allow for a higher accuracy, but it would point key variables at different levels of the tree so not only would one be able to predict the entries, but potential city planners could control the entries themselves, preventing 'bottlenecks' and creating more efficient flow through the subway

5.2

- **(Optional) Do you have any other insight about the dataset that you would like to share with us?**
  - Maybe I missed a way to manipulate the data, but I would have loved to have a dataset where the majority of the variables were relevant to the final prediction model. After utilizing the dummy-variables, and identifying the rest of the variables as collinear, there seemed to be nothing left to do. At least nothing that the class taught us to look out for.
  - I would have also liked a dataset where there existed multiple relationships where multiple outcome variables could have been predicted. For example, this dataset could have added a *traffic_incidents* variable. Multiple relationships could be drawn between the existing variables and the new variable. Weather conditions could have causal effects on *traffic_incidents* and then weather becomes more relevant to the dataset rather than weather's effect on subway entry. *traffic_incidents* could mean traffic blockage causing more people to use the subway, or conversely, higher subway entries may mean less people are driving therefore there are less *traffic_incidents*. This would be a fun dataset I could probably get lost in for a while. Overall, this exercise was great and I hope there are more fun projects like this along the way.

**Helpful webpages**

This site, offers a lot of 'hows' of running non-parametric and parametric tests after applying linear regression
http://connor-johnson.com/2014/02/18/linear-regression-with-python/
This site goes over the basics of Ordinary Least Squares
http://www.datarobot.com/blog/ordinary-least-squares-in-python/
Everything you need to know about SQL queries
http://www.w3schools.com/sql/