Short Questions to Analyzing the NYC Subway Dataset

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. Which statistical test did you use to analyse 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?
* We used a Mann-Whitney U test to analyse the NYC subway data. We used a two-tail p-value. The null hypothesis was that there was no difference in ridership on the NYC subway between rainy and non-rainy days. The p-critical value was 0.05 – the one-sided p-value was 0.025 returned by default by the Mann-Whitney U-test in Exercise 3.3, and we multiply this by two to get the final p-value.
  1. 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.

This statistical test was applicable to this dataset because ridership was not normally distributed in either of the two samples we were comparing – ridership on rainy days and ridership on non-rainy days.

* 1. 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 resulting p-value was 0.05, meaning that if both samples were drawn from the same underlying distribution, the probably of us observing the difference in the means that we did would be 5%
* The mean for the first sample (rainy days) was 1,105, and the mean for the second sample was 1,090.
  1. What is the significance and interpretation of these results?

My interpretation of these results is that there was a statistically significant difference in ridership 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:

1. Gradient descent (as implemented in exercise 3.5)
2. OLS using Statsmodels
3. Or something different?

I used gradient descent as implemented in exercise 3.5.

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

I used a categorical variable for whether it was raining as a feature, as well as the mean temperature outside as a numeric variable. I also used dummy variables for each of the turnstiles themselves, as well as dummy variables for each hour of the day.

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 R2 value.”

I was trying to isolate the effect of rain on ridership. I wanted to control for other intrinsic characteristics that could have affected ridership that were unrelated to rain. First of all, some stations simply get more riders than others, necessitating the addition of the dummies for the turnstiles. In addition, some times of day are simply busier than others – again independent of whether it is raining.

I ended up excluding temperature, since it only added about 0.1% to the R^2. By comparison, adding the hour dummies added over 5% to the R^2.

Finally, I intentionally excluded the amount of precipitation as a variable, as it was highly correlated with the rain variable (obviously) and was therefore messing up the coefficient estimate of the rain variable. Indeed, when I excluded the “precipi” variable my R^2 didn’t change at all.

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

The coefficient on the rain variable was 199.7, indicating that controlling for the turnstile itself – i.e. for a given turnstile - and the hour of the day – i.e. for a given time of day – there are about 200 more entries when it is raining than when it isn’t.

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

My model’s R^2 value is 50.4%.

2.6 What does this R2 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 R2value?

My R^2 value means that 50.4% of the variance in ridership can be explained by variance in the “independent” variables: the station itself, the hour of the day, and whether or not it was raining.

Given that NYC subway ridership is influenced by many factors not included in our dataset, I think predicting more than half of the variance with a relatively simple model is impressive, so yes, I think using a linear model is appropriate. However, a couple of caveats:

* If ridership patterns are relatively consistent across different times of day for a given station, then we should expect to see the indicator variables for each station and time of day as explaining much of the variance in ridership. We do indeed see this, as just including those two variables gives a model with an R^2 of about 0.45. However, the fact that these *only* explain 45% of the variance – combined with the “day of the week” visualization below – tells me that adding indicator variables for day of the week would help the model, as my guess is that (based on both the R^2 of 0.45 and intuition about subway ridership) adding these would correct for the fact that at a given station, ridership might look very different on a weekend vs. a weekday, for example.
* Turning to the “linear” part of the question: I would need to look at a plot of actual vs. predicted ridership, and looking for heteroskedasticity, to determine whether a different functional form would be appropriate. Given that the histograms of ridership on rainy vs non-rainy days were of similar shapes, I have no evidence that there is, for example, an exponential, or quadratic relationship between the amount of rain and ridership. More rigorous tests would have to be done to confirm this though.

**Section 3. Visualization**

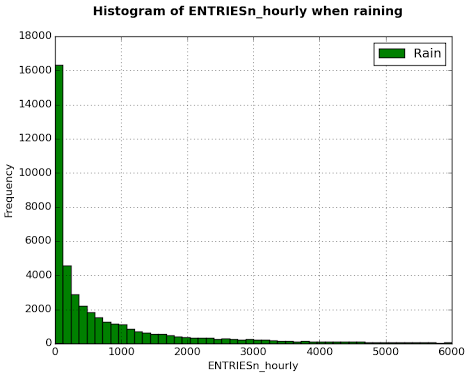
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

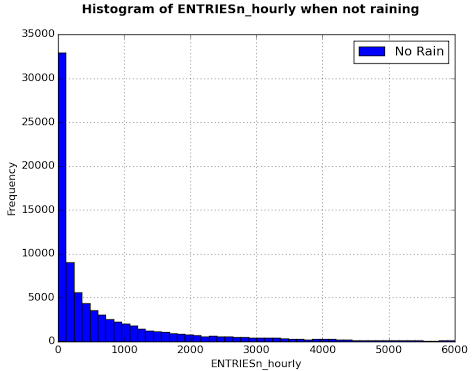
Remember to add appropriate titles and axes labels to your

plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

3.1 One visualization should contain two histograms: one of  ENTRIESn\_hourly for rainy days and one of ENTRIESn\_hourly for non-rainy days.

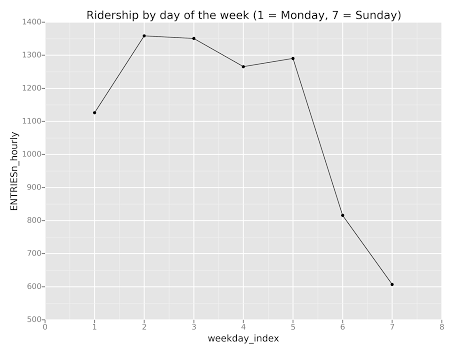
* You can combine the two histograms in a single plot or you can use two separate plots.
* If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
* For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn\_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn\_hourly that falls in this interval.
* Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.





3.2 One visualization can be more freeform. Some suggestions are:

* Ridership by time-of-day or day-of-week
* Which stations have more exits or entries at different times of 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?

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.

Answer to both 4.1 and 4.2:

Yes, more people ride the NYC subway when it is raining than when it is not raining. First, just comparing average ridership when it is raining to when it is not raining shows that average ridership is higher when it is raining. However, we cannot conclude from that alone that this difference is meaningful; to test that we have to look at the statistical significance of the difference. In addition, we can’t use Welch’s T-test since ridership is highly non-normal. Instead, we use the Mann-Whitney U Test – the appropriate test for non-normal data – and find that the difference in ridership between rainy and non-rainy days is just barley significant at the 0.05 level, which is the most common cutoff used.

Still, a significance of 0.05 on such a large sample raises concerns. With a different sample of days on which to test our hypothesis, we might find that the difference in ridership between rainy and non-rainy days was not significant at the 0.05 level. In addition, a more conservative test of the null hypothesis would be to insist on significance at the 0.01 level, thus in our case failing to reject the null hypothesis that there is a difference in ridership between rainy and non-rainy days. Thus, linear regression is required to further determine whether the difference we observed was statistically significant.

The linear regression I used simply controlled for two factors:

* Ridership varies by station.
* Ridership varies by time of day.

I did this by including dummy variables for station and time of day. There are other factors that could affect ridership. However, the other factors included in the dataset that affect ridership did not add much to the model’s R^2. Furthermore, since the other factors included in the dataset were weather-related, and therefore were often correlated with the presence of rain. For example, it might be colder on days when it is raining than when it is not, and less likely to be foggy. If temperature or fog were included in the regression, therefore, the estimate of the coefficient on the rain variable might be off due to multicollinearity. Therefore, since the goal of the model was to find as “pure” as estimate as possible for how much ridership increases (if any) when it is raining, on top of the fact that the other weather variables didn’t add much in terms of predictive power, led me to exclude them.

The coefficient estimate on the “rain” variable in the model, however, led me to conclude that there is significantly more ridership – about 200 more riders (199.8 to be exact) – when it is raining than when it is not.

**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. Linear regression model,
3. Statistical test.

I think the dataset we compiled – merging MTA data with Weather Underground data – was adequate to answer the question of whether ridership is higher when it is raining than when it is not. However, I think a better model could have probably been built by looking at events going on in New York City, rather than just weather. For example: a dummy variable of whether it is within 2 hours of the start or finish of a baseball game would have had a significant impact on ridership.

I also that, given the long right tail of the histogram, both the difference in the means and even the coefficient estimates of the regression could be overly influenced by outliers. If I were to do this over I would consider excluding certain data from my analyses – for example, excluding all hours where ridership was greater than 6,000 – and see what results I got. The reason for this is that these outlier hours would .influence both the estimate of ridership when it is raining out *and* the estimate when it is not raining, so excluding these points wouldn’t necessarily bias the results too much one way or another.

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