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? Did you use a one-tail or a two-tail P value? What is the null hypothesis?

I used a two-tailed Mann-Whitney U test. The null hypothesis is that the two samples (hourly entries “with rain” and “without rain”) come from the same population, that is, there is the distributions of the two groups are identical. The alternative hypothesis is that the two groups have statistically different distributions. In other words, I am testing if the “rain” and “no rain” groups have the same distribution and therefore whether I can conclude if there’s a statistical difference in the hourly entries based on if it is raining or not.

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.

I chose to run a Mann-Whitney U test because the distribution of ridership in the two samples (rain vs. no rain) is not normally distributed. Therefore, a nonparametric statistical test is needed.

I came to this conclusion in two ways: graphically and numerically.

1. Histogram Plots

The two histograms in Problem 3.1 show the hourly entries when it is raining versus not raining. While the “rain” distribution has fewer samples, it is still clear from the chart that both distributions are skewed right, creating a long right tail, instead of a bell-shaped curve.

While useful and effective, a visual inspection is subjective. To support my conclusion from the graphical representation of the data, I used a statistical test to more objectively test for normality.

2. Shapiro-Wilk test

A Shapiro-Wilk test can determine whether the distribution of the two samples is normal and therefore whether a parametric or non-parametric test would be more appropriate.

Thus, the null hypothesis is that the distribution is normal. I ran the test using an alpha of 0.05 for both rain and no rain samples with these results:

|  |  |  |
| --- | --- | --- |
| Sample condition | W Statistic | p-value |
| With rain | 0.59 | 0.0 |
| Without rain | 0.60 | 0.0 |

Because the p-value for both samples is less than 0.05, I reject the null hypothesis (explanation of how to interpret Shapiro-Wilk test [here](http://stackoverflow.com/questions/15427692/perform-a-shapiro-wilk-normality-test), [here](https://statistics.laerd.com/spss-tutorials/testing-for-normality-using-spss-statistics.php), [here](http://geography.unt.edu/~wolverton/Normality%20Tests%20in%20SPSS.pdf)). Therefore, I reject the assumption that the distributions are normal.

Based on the graphical representation of the distributions and statistical testing for normality, I conclude that the two distributions of ridership with rain and without rain are not normal. Therefore, a nonparametric statistical test, such as the Mann-Whitney U test, is appropriate.

The Mann Whitney U test is used to compare differences between two independent groups when the dependent variable is continuous (the frequency of hourly entries is continuous) and the observations are independent. Thus, this test an appropriate choice for understanding whether the hourly entries differ based on whether there is rain or no rain.

Furthermore, I chose a two-tailed Mann Whitney U test because I don’t have insight on which direction the distributions of the two groups would differ. Therefore a two-tail test is a safer bet as it includes both tails.

One reason not to use a t-test is because the t-test assumes data conforms to a normal distribution. Because the two samples are not normally distributed, a parametric test is not appropriate.

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.

|  |  |
| --- | --- |
| U test statistic | 153635120.5 |
| Two-tailed p-value | 5.48 e-06 |

|  |  |  |  |
| --- | --- | --- | --- |
| Descriptive Statistics | ENTRIESn\_hourly  With Rain | ENTRIESn\_hourly  Without Rain | ENTRIESn\_hourly  (total) |
| count | 9585.00 | 33064.00 | 42649.00 |
| mean | 2028.20 | 1845.54 | 1886.59 |
| std | 3189.43 | 2878.77 | 2952.39 |
| min | 0.00 | 0.00 | 0.00 |
| 25% | 295.00 | 269.00 | 274.00 |
| 50% | 939.00 | 893.00 | 905.00 |
| 75% | 2424.25 | 2197.00 | 2255.00 |
| max | 32289.00 | 32814.00 | 32814.00 |

1.4 What is the significance and interpretation of these results?

Based on a p-value of less than 0.05, I reject the null hypothesis of the Mann-Whitney U test in favor of the alternative hypothesis; i.e., there is a statistically significant difference between the distribution of hourly entries “with rain” and “without rain”. But this conclusion doesn’t tell me how the distributions are different, and the U statistic in isolation has little meaning.

Therefore, we need to look at descriptive statistics to make sense of the distributions of the two groups. For example, the “no rain” group is almost 4 times larger than the “rain” group. This suggests the distributions are quite different. However, consider the mean and standard deviation; it does not look like the data has as great of a spread as the difference in count would suggest. Furthermore, the difference in the median is only 5%.

**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:

Gradient descent (as implemented in exercise 3.5)

OLS using Statsmodels

Or something different?

I built a linear regression model using the ordinary least squares approach from Statsmodels.

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 the following features:

hour, rain, precipi, unit, day\_week.

UNIT and day\_week were dummy variables in my linear regression model.

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 improved my model by testing different features over several iterations. I picked hour, rain, and precipi based on intuition. First, I hypothesized that subway ridership would vary at different times of the day, such as increasing during morning traffic when commutes were traveling to work. Second, I guessed that people are more likely to ride the subway when it is raining to avoid getting wet.

The features unit and day\_week were added because they increased the R-squared value. Before I added unit, the regression model had an R-squared value of 0.083. When I added unit as a dummy variable, the R-squared value jumped up to 0.46. Based on this large increase in the R-squared value, I decided to keep unit as a feature in the model.

Once the regression model reached an R-squared of 0.46, I experimented with adding different features, including tempi, meantempi, fog, wspdi, and meanwspdi. These weather conditions made less than 0.002 different in the R-squared value. I decided to drop them from the features because there was almost no additional benefit of adding them (the increase in R-squared was trivial), and I wanted to avoid the risk of superficially inflating my R-squared value to increase my goodness of fit by adding more variables.

Finally, I added day\_week as a dummy variable (day of the week is categorical data), which increased the R-squared value to 0.49. This 3% increase was more tangible than the effect of the variables related to the weather, and intuitively it made sense that subway ridership would vary based on the day of the week. Based on these two factors, I decided to keep day\_week as a feature in the regression model.

2.4 What is your model’s R^2 (coefficients of determination) value?

My mode’s R-squared value is

0.49 (or 49%)

2.5 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?

The R-squared value means that the features in my linear regression model account for approximately 49% of the variance in the original subway ridership data. In other words, I can explain about 49% of the variance in the labels using the features I chose.

There is no fixed rule for what a good value for R-squared is. The importance of the R-squared value depends on how much variance my features need to account for. This linear model explains about half of the variance, which means it has a one out of two chance of correctly predicting subway ridership. This isn’t a good predictive model because the rate of accurately predicting is the same as the chances of flipping heads on a coin. Therefore, the model could be improved.

I don’t think a linear model is appropriate for this dataset because many of the variables in the data set pose multicollinearity problems. For example, the likelihood that the temperature drops on a rainy day is high. Without teasing apart which predictor variables are correlated, it is a problem because it makes it hard to have confidence that the predictions will be accurate and stable.

**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.

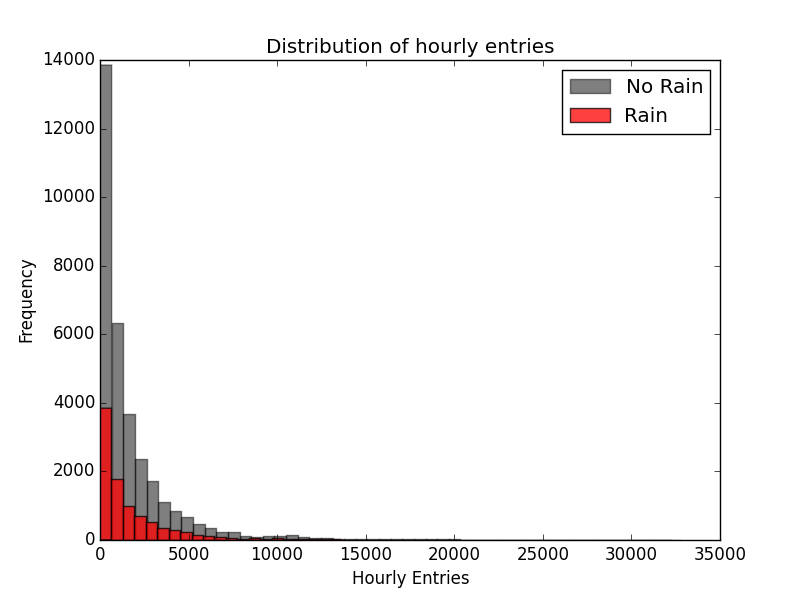
3.1 One visualization should contain two histograms: one of ENTRIESn\_hourly for rainy days and one of ENTRIESn\_hourly for non-rainy days.

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, you might have one interval (along the x-axis) with values from 0 to 1000. The height of the bar for this interval will then represent the number of records (rows in our data) that have ENTRIESn\_hourly that fall into 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.



Description:

The plot above with two histograms shows:

1. The pattern of hourly entries is consistent whether there is rain or no rain because the histograms follow the same shape. This means that there is not a noticeable drop off in the number of subway riders when it is not raining, otherwise we would see the No Rain histogram skew close to zero, and it would not tail off like the Rain histogram
2. The number of observations for hourly entries with rain is much greater (5 times) than the number of observations for hourly entries without rain.

See code for visualization (Subway2Matplotlib.py).

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



N.B. the plot above doesn’t contain a legend because the x-axis labels should make it clear what the bars mean

Description:

The plot above with the bar chart shows:

1. The average hourly entries for subway turnstiles vary based on the day of the week. Wednesday, Thursday, and Friday have the highest averages with Thursday having the greatest average hourly entries.
2. There is a big drop off in the amount of subway ridership during the weekends, with Sunday having the lowest average hourly entries, less than half of the average hourly entries on Thursday. So the day with the least traffic for riding the subway is Sunday (though this plot doesn’t account for different stations).
3. The range of average hourly entries from Monday to Sunday is approximately 1050 (Sunday) to 2300 (Thursday).

See code for visualization (Subway2MatplotlibFreeform.py).

**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 versus when it is not raining?

The findings from the above analyses suggest there is a slight correlation between subway ridership and rain but other features have a significantly larger impact on subway ridership.

Based on the linear model and the R-squared value, which is very small, I concluded there is a relationship between the variables rain and subway ridership but I don’t think it is strong enough to conclude that more people ride the subway when it is raining compared to when it is not.

4.2 What analyses lead you to this conclusion?

To summarize the findings in previous sections, I performed analyses including the following: Welch’s t-test, linear modeling with consideration of R-squared values, visual “eyeballing” test with a histogram plot. The Welch’s t-test suggests that there is a statistically significant difference between the mean of the hourly entries with rain versus without rain. This conclusion isn’t robust enough to predict whether more people ride the subway when it is raining. The linear model that used the Hour, rain, meantempi, meanwindspdi, and precipi features suggests that these factors explained almost 50% of the variance in the labels, which means that a combination of time, rain, mean temperature, wind speed, and how much rain potentially predict whether more people ride the subway. But the linear regression model that uses only rain and a constant as the features suggests there isn’t enough predictive power to conclude that rain causes more people to ride the subway. Additionally, there are some shortcomings in the data discussed below that cause me to hesitate to conclude that rain predicts subway ridership.

**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 data set and the methods   
of your analysis.

To answer the question of whether more people ride the subway when it is raining versus when it is not, there needs to be more data in the dataset. The turnstile data only covers May 2011. I plotted on a line chart the number of hours in a day when rain was recorded at a certain unit (R003). It looks like there were only three rain storms whereas most of the month remained rain free.



(See ConclusionLineDateOnlyOneUnit.py)

Therefore, it is inaccurate to conclude whether more people ride the subway when it is raining considering there is not enough data to control for seasonal weather.

The shortcomings of the analyses revolve around the models and assumptions. The linear regression model assumes that each feature is acting independently. My models don’t take into account how the features are correlated with each other and I didn’t perform any analysis to determine covariance or to control for it. It’s intuitive to think that when it is raining, the mean temperature might be lower. But which factor is really the one causing more people to ride the subway (if either of them have an effect)?

Another shortcoming of the way I performed linear regression is the more features I add the closer my model will fit with the dataset. If I used all the features in the data to create the model and calculate the R-squared value, the R-squared value would be very strong. However, this isn’t a good way to perform predictions: adding more features may result in a higher R-squared value, but the model would be a good fit only for my data set because I have created an spurious fit. This artificially tricks me into thinking I have a good model but my data may not be representative of the true population. A better way is to have two separate datasets: a training and a test set to determine the robustness and predictive power in my model.

Additionally because there are only three rain events, there is a random chance that one or more of those rain storms coincided with higher subway ridership. For example, it could have rained during commuting hours, and thus produce a false positive correlation between rain and higher subway ridership.

One way I could improve my linear model is to make Hour a dummy variable. Intuitively, the time of day should have an effect on subway ridership – with more people riding during the day/waking hours. However, using the value of Hour as a feature doesn’t really help my model, i.e. because Hour doubles (from 4 pm to 8 pm) does not mean that the hourly entries will double. But adding it as a feature suggests this kind of linear relationship. Therefore adding Hour as a dummy variable would likely strengthen my model and increase my R-squared value.

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

The ENTRIESn\_hourly column was not calculating the entries per hour but the difference between the number of entries at the current point in time and last time the entries was measured. Usually the common interval was 4 hours between measurements. Therefore, it would be more accurate to divide the ENTRIESn\_hourly column by the number of hours that has elapsed so that conclusions drawn from this number more accurately reflect hourly entries .