Analyzing the NYC Subway Dataset

Ovidiu E Icreverzi

May 29, 2015

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 0. References

Please include a list of references you have used for this project. Please be specific - for example, instead of including a general website such as `http://www.stackoverflow.com/`, try to include a specific topic from Stackoverflow that you have found useful.

Some of most frequented sources of information I utilized are as follows:

- The prerequisite Udacity courses in the Data Science category. Courses such as Introduction to Statistics, Introduction to Descriptive & Inferential Statistics, Introduction to Git & GitHub. These Udacity courses were very helpful in getting me started on the coursework of this project.
- `https://www.python.org/`. I used the Python 2.7 documentation extensively for this class.
- `scipy.org`. I used this site for searching the documentation on the `matplolib` and `pandas` modules. These modules are very well documented and the tutorials, examples and HOWTO's are extremely helpful.
- `http://ggplot.yhathq.com/` This site seems to be the official online documentation for the python `ggplot` module.

- *Learning Python* 5^{th} *Edition* by Mark Lutz © 2013. This is an introductory python book which I often referenced for basic python tasks such as how to turn the string '2011-05-11' into the ISO weekday value of 3. I found this book to be a great desk reference for common python programming tasks.

- `http : //www.epa.gov/ttn/airs/airsaqs/training/` This EPA training materials website contains a series of presentations titled "SQL Basics" and "SQL Basics II" by Jonathan Miller, which I found to be a very quick and practical introduction to using SQL for the problem set in lesson 2.

- *Fundamentals in Statistical Pattern Recognition, Lab 1: Linear Regression with Python* by Dr. Andre Anjos & Dr. Sebastien Marcel
  This is a detailed presentation which covers gradient descent using the linear algebra notation and python programming. I used this set of slides for supplementary reading and learning in Lesson 3.

- `http://www.socscistatistics.com/` I found this to be a great resource for learning the basics of the Mann-Whitney $U-$test, more intuitively known as the rank-sum test.

- *Introduction to Statistics and Data Analysis* 2^{nd} *Edition* by Peck, Olsen and Devore © 2004. An introductory level textbook on data analysis concepts and techniques. My go-to resource for a first look at a statistics or data analysis concept.

- `http://blog.minitab.com/` A blog about using the Minitab statistical software that includes many useful posts, especially about R^2 and its interpretation.

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?

- I used the Mann-Whitney u-test to test whether or not the values in the `ENTRIESn_hourly` column are significantly different when we compare rainy and non-rainy days. More formally, the u-test is trying to quantify the difference the `ENTRIESn_hourly` values between rainy and non-rainy days in order to see if we can consider the two categories as statistically distinct populations.
- The `scipy.stats.mannwhitneyu()` function produces a one-tail $p-$value but I multiplied it by 2 to get the two-tail $p -$ value.
- The null hypothesis is that we have equal probability of getting a larger value of `ENTRIESn_hourly` on either rainy or non-rainy days. In mathematical notation this looks like the following:

  $$H_0 : P\left(\texttt{ENTRIESn\_hourly}_{\text{rainy}} > \texttt{ENTRIESn\_hourly}_{\text{non-rainy}}\right) = 0.5$$

  Another way of stating it is to say that this is testing whether or not the median of the two samples is equal.
- I used the two-tailed $p_{\text{critical}} = 0.05$, which is a standard value for testing the null hypothesis.

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 data set because it does not assume or require a particular type of distribution. In our case graphing the data of the `ENTRIESn_hourly` values on rainy and non-rainy days showed, qualitatively, that the distributions of the samples are not normal. Welch's t-test for example assumes normal distributions so that makes it inappropriate in our case.

- More formally the Mann-Whitney u-test requires the following: Two random and independent samples, scale of measurements is a ratio which clearly distinguishes the sizes of measurements, no particular assumptions are made about the distribution of the sample values.

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.

The results of the $u-$test without Memorial Day removed from the dataset are as follows:

$$\begin{aligned}
u &= 1,924,409,167 \\
n_{\text{rainy}} &= 44,104 \\
n_{\text{non-rainy}} &= 87,847 \\
p_{\text{one-tail}} &= 0.024999 \\
p_{\text{two-tail}} &\approx 0.05 \\
\bar{x}_{\text{rainy}} &= 1105 \\
\bar{x}_{\text{non-rainy}} &= 1090
\end{aligned}$$

The results of the $u-$test *with* Memorial Day removed from the dataset are as follows:

$$\begin{aligned}
u &= 1,711,616,498 \\
n_{\text{rainy}} &= 39,754 \\
n_{\text{non-rainy}} &= 87,847 \\
p_{\text{one-tail}} &= 7.37 \times 10^{-9} \\
p_{\text{two-tail}} &\approx 1.5 \times 10^{-8} \\
\bar{x}_{\text{rainy}} &= 1159 \\
\bar{x}_{\text{non-rainy}} &= 1090
\end{aligned}$$

Removing what I consider to be an out-lier, the Memorial Day holiday, only makes the $u-$test more extreme in rejecting the null hypothesis. Memorial Day, a United States national holiday celebrated on the last Monday of May, was a rainy day in 2011. The dramatically lower ridership on that single rainy holiday actually masks just how much higher the average ridership of typical rainy days tends to be.

1.4 What is the significance and interpretation of these results?

The two samples have a numeric $u-$ statistic value which has a probability of occurring approximately 5% of the time if two samples of the same sizes as above, namely $n_{\text{rainy}} = 44,104$ and $n_{\text{non-rainy}} = 87,847$, were to be selected from a statistically homogeneous population. That is, if the median of the two sample source populations was equivalent we would get this same result 5% of the time.. In our particular case, we are testing whether for the purpose of ridership numbers we can consider rainy and non-rainy days as part of the same homogeneous population of days. The answer in our case is no, in the case of including Memorial Day as a data day, and **no way**! in the case we take Memorial Day out of our dataset. For the purposes of subway ridership we can consider rainy and non-rainy days as significantly different populations.

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?

Two methods were used to compute the elements of $\boldsymbol{\theta}$. First I used the gradient descent algorithm. Then I used a more automated approach and called upon the `statsmodels` module and computed the elements of $\boldsymbol{\theta}$ using ordinary least squares (OLS) .

2.2

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

- The features used in both modeling techniques are; `rain`, `precipi`, `Hour` and `meantempi`. The values of `precipi` and `meantempi` are the imperial unit values for precipitation and temperature namely inches and Fahrenheit.
- Both modeling techniques utilized the same dummy variables, or indicator variables, which are the `UNIT`, `weekday` and `holiday` values. This was critical in making the model somewhat realistic. The `UNIT` dummy variable allowed the model to consider each turnstile as a distinct individual and create a linear function for that unit alone. The `weekday` dummy variable allowed the model to consider each day of the week as distinct. The `holiday` dummy variable allowed the model to consider holidays as a special category, in our dataset this only affected $\texttt{2011} - \texttt{05} - \texttt{30}$ or Memorial 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 R^2 value.”

In addition to the default features already included by Udacity which are `rain`, `precipi`, `Hour` and `meantempi`, I included the weekday values and holidays. Our society runs on a weekly schedule and although New York is “the city that never sleeps”, it still abides by a seven day cycle. Intuition is what first led me to include these features and the improvement of the R^2 value seems to be quantitative proof that it’s worth including them. However, I noticed in the Statistics and Data Analysis by Peck, Olsen and Devore 2nd edition page 772-773,

> Using R^2 to choose between models is not so straightforward, because adding a predictor to a model can never decrease the value of R^2. When statisticians base model selection on R^2, the objective is not simply to find the model with the largest R^2 value … Instead we should look for a model that contains relatively few predictors but has a large R^2 value and is such that no other model containing more predictors gives much improvement in R^2. … A small increase in R^2 resulting from the addition of a predictor to a model can be offset by the increased complexity of the new model …

The statement above is important for our case since I went from initially using `rain`, `precipi`, `Hour` and `meantempi`, to compute the $\boldsymbol{\theta}$ vector to adding the weekdays and holidays as features. Initially I was glad to see the R^2 value increasing, but that was tempered by reading the above quote.

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

The first four elements of $\boldsymbol{\theta}$ computed using linear regression are as follows:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_{\texttt{rain}} \\ \theta_{\texttt{precipi}} \\ \theta_{\texttt{Hour}} \\ \theta_{\texttt{meantempi}} \\ \vdots \end{bmatrix} = \begin{bmatrix} 16.328 \\ -5.562 \\ 468.801 \\ -9.699 \\ \vdots \end{bmatrix}$$

It's important to keep in mind that these values are for use with the normalized, also known as standardized, values of the features matrix. This means that without normalization we would think a positive value indicates a positive correlation and negative value a negative correlation. However in this case that correlation only holds if we consider the positive or negative deviation from the mean as our variable.

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

The two coefficient of determination, or R^2, values are shown as computed by the Udacity python interpreter.

- For linear regression with $\alpha = 0.1$ and $n_{\text{iterations}} = 100$ we have, rounded to three decimal places:
$$\mathrm{R}^2 = 0.478$$
- Using the method of ordinary least squares as implemented by the `statsmodels` module we have
$$\mathrm{R}^2 = 0.496$$

Comparing the two R^2 values above is appropriate since the same exact features were used and we simply have $\boldsymbol{\theta}$ using two different methods.

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?

- Informally the R^2 value represents the following ratio
$$\mathrm{R}^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$
We can take this to mean that only 50% of the variation in ridership can be accounted for by our linear model. But now let us combine this R^2 value with the ***estimated standard deviation***, or
$$s_e = \frac{\text{SSResid}}{\text{degrees of freedom}} = \frac{\sum (y - \hat{y})^2}{n-2} \approx 1,845$$

This means that typically the amount by which an actual observation of `ENTRIESn_hourly` differs from that predicted by our least squares linear model is $1,845$ which is quite significant if we consider that

$$\begin{aligned}\bar{x}_{\text{rainy}} &= 1159\\ \bar{x}_{\text{non}-\text{rainy}} &= 1090\end{aligned}$$

- As stated in Peck et al. on page 748, " a desirable model is one that results in both a large R^2 and a small s_e value. In our case it seems we actually have the opposite on both fronts. Let's take a look at the normal probability plot of the standardized residuals as shown below.

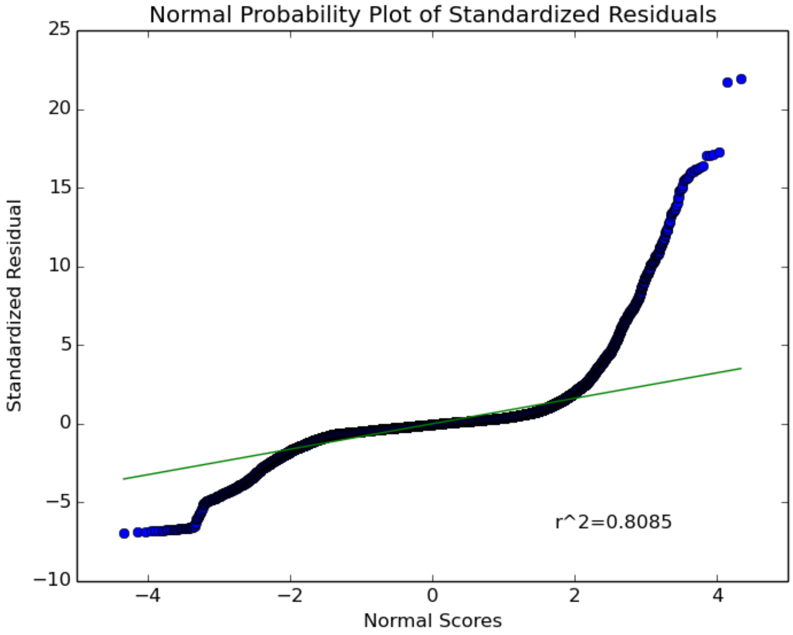

Figure 1: Residuals Normal Probability Plot

Please note that the r^2 value shown in Figure 1 is not that of our linear model but of the residual values being fitted to a *normal* distribution. Furthermore, even this value is overstating how poorly this model performs since it was computed with a reduced data set on my own machine. On Udacity's python interpreter $r^2 \approx 0.679$.
This linear approach simply cannot account for the six spike and decay cycles of ridership throughout the day, which are shown in Figure 4. For example we can make the following qualitative observation, $\theta_{\texttt{Hour}} = 468.801$

which implies that all other factors being equal, entries should increase between the hours of 20 and 23. However if we look at the actual data, ridership decays during that time frame from a maximum close to $1,000,000$ passengers to nearly zero during the 11pm hour.
Ultimately this result show us that a linear approach to understanding the NYC subway entries pattern is inadequate.

Section 3. Visualization

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

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 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.

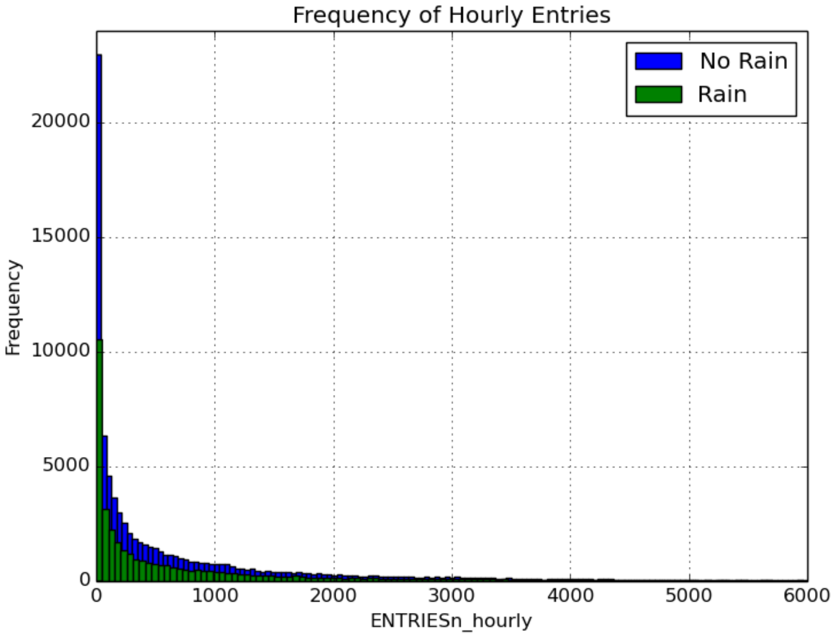

Figure 2: Exploratory Data Analysis Plot 1

The image above gives us a qualitative look at the data points. On the $x-$axis the cutoff is at $6,000$ entries per hour (1.67 entries per second) since this seems to be a reasonable limit to what is realistically possible for a single turnstile. The $y-$axis on the other hand is related to the number of actual data points analyzed.

The rapidly decaying pattern of the entries values is clearly not normal and is similar for both rainy and non-rainy days. In our case the $y-$axis magnitude values are not nearly as important as the pattern of their decay and the comparison between the rainy and non-rainy magnitudes is simply a comparison of the number of data points in each category.
Another method of showing just how non-normal the data is through the normal probability plots of the standardized `ENTRIESn_hourly` data on rainy and non-rainy days.

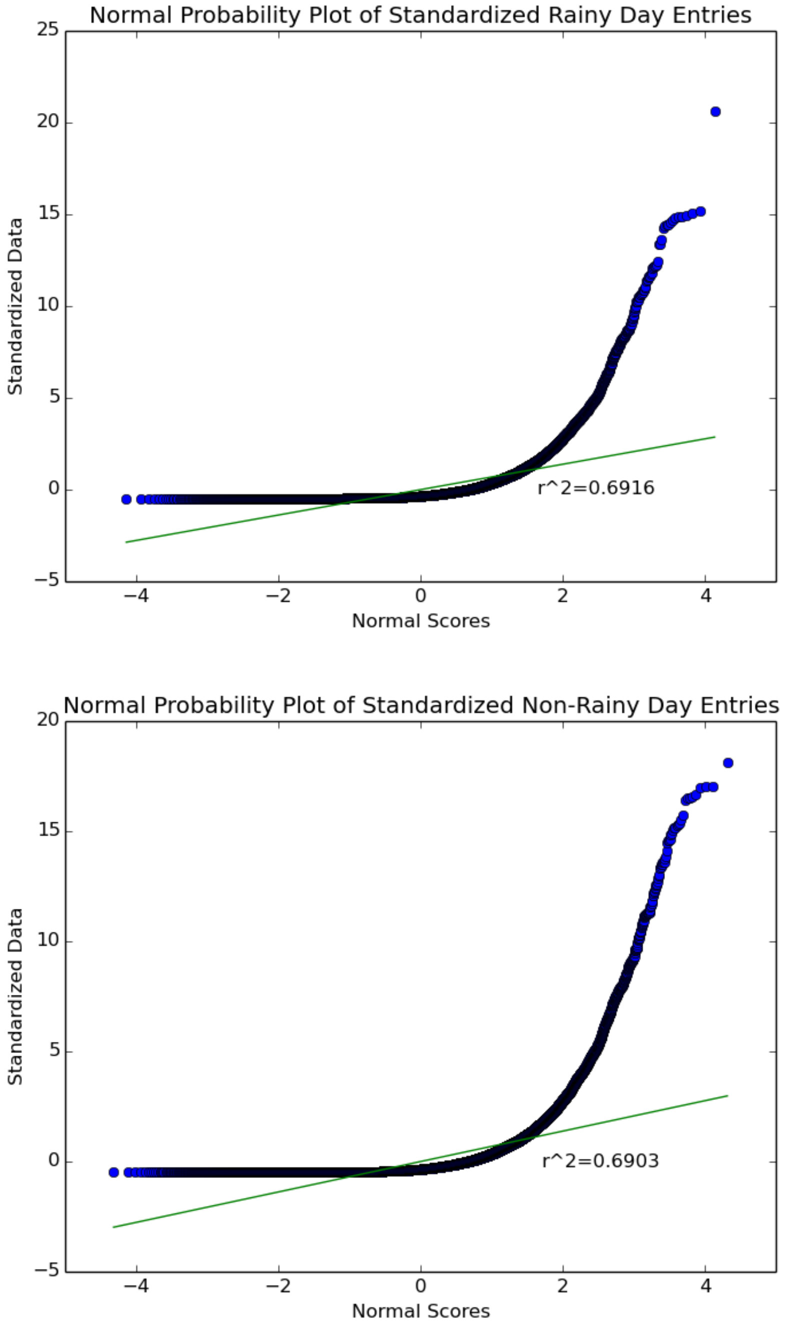

Figure 3: Exploratory Data Analysis Plot 2

Figure 3 is convincing proof that our given data is not a normal distribution and therefore not eligible for analysis using Welch's $t-$test.

3.2 One visualization can be more free form. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

- Ridership by time-of-day

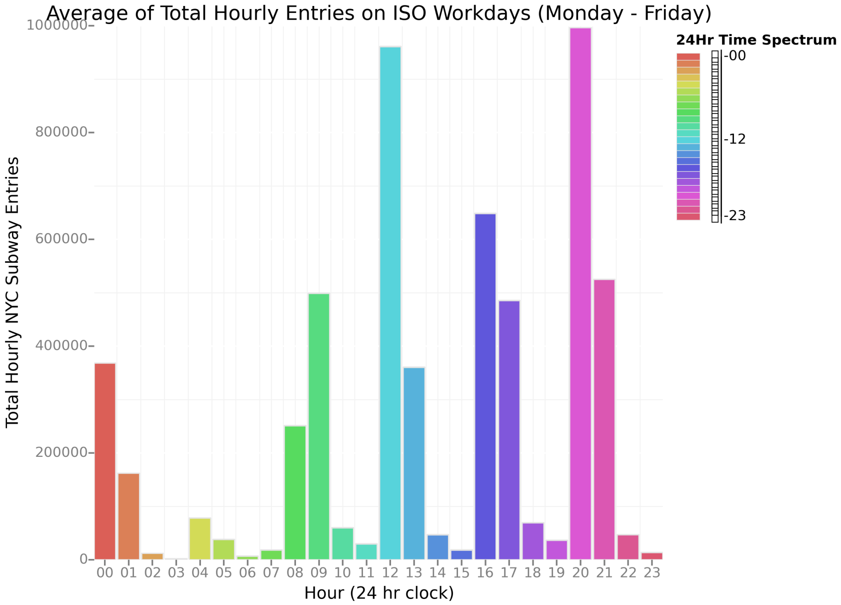

Figure 4: Ridership by Hour of Day

The image above gives us another qualitative look at the ridership based on the hour of the workday. For this plot Saturday, Sunday and holiday data was removed. This shows interesting oscillatory behavior. Ridership seems to be characterized by spikes with rapid decays in four hour increments, except for the 8am to 11am time frame which is a slight variation on that. For example between the hours of 20 and 23, we see a spike to almost 1 million entries decay to practically zero. This set of six spikes and their accompanying decays could be useful data for use in a variety of settings. If advertising is sold on the NY subway system this pattern could be used to price electronic banners based on hour and day of the week. If a delivery or courier business uses the subway to transport people or product, they could also use this schedule in a similar way to price based on when deliveries are made. Preventive maintenance schedules can be synchronized with this pattern and so on.

If we integrate the ridership plot above the average number of total entries on workdays is $5,732,329$, based on the data file that I downloaded from Udacity.

- Ridership by day-of-week

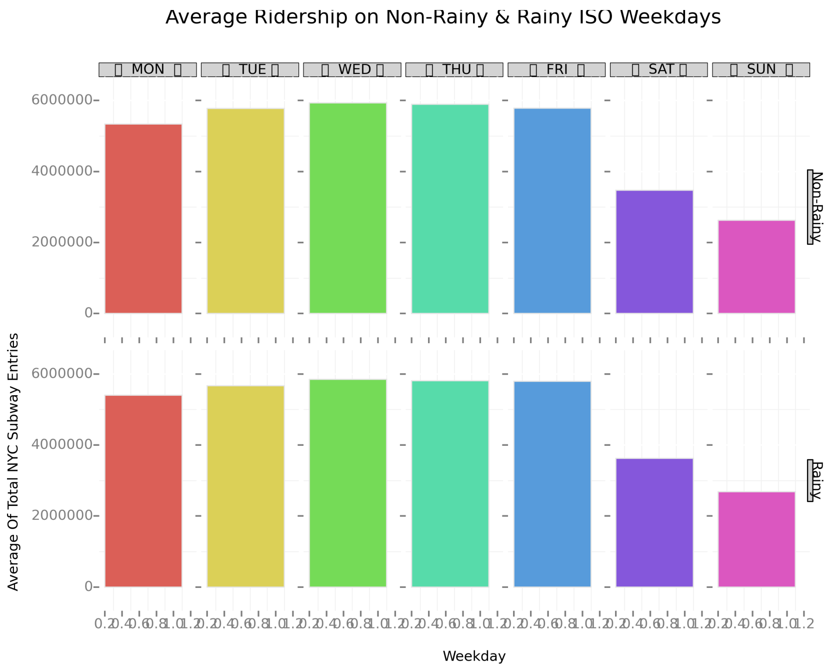

Figure 5: Average of Total NYC Subway Entries by Day of Week and Rain Status

The image above gives us another qualitative look at the ridership based on the rain status and day of week. On the top is a histogram with rain value 0, or sunny days, and on the bottom is the histogram for rainy days. Both histograms are on the same scale so a 1:1 comparison of magnitudes is appropriate. Overall the data is something we would intuit from the working schedule of an average professional employed in New York City. It also shows that it would be hard to draw a rigorous quantitative conclusion from simply looking at the data in this way so I computed the numeric values and listed them in the table below for another look at the data. The numbers in parenthesis are the number of unique days which were averaged to get the net values listed next to them. For example, there were four (4) non-rainy Sundays in our dataset and only one (1) rainy Sunday. To me this is an important detail to keep in mind as we try to understand the significance of our data. The table is sideways in order to fit on a single page.

Table 1: Summary of Average Daily Entries

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
$\text{avg}_{\text{non\_rainy}}$ (20)	$5,331,900$ (2)	$5,776,038$ (3)	$5,930,089$ (2)	$5,891,724$ (3)	$5,779,588$ (3)	$3,472,327$ (3)	$2,623,677$ (4)
$\text{avg}_{\text{rainy}}$ (9)	$5,402,385$ (2)	$5,669,840$ (1)	$5,851,229$ (2)	$5,811,475$ (1)	$5,792,009$ (1)	$3,627,123$ (1)	$2,685,406$ (1)
$\text{net}\Delta$	$-70,485$	$106,198$	$78,860$	$80,249$	$-12,421$	$-154,796$	$-61,729$
$\%\Delta = \frac{\text{net}\Delta}{\text{avg}_{\text{non\_rainy}}}$	-1.32	1.84	1.33	1.36	-0.21	-4.46	-2.35

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?

Yes, *if* we consider our dataset as representative of general ridership, on average the number of hourly turnstile entries are 1105 (if we include a rainy Memorial Day) and 1159 (if we exclude a rainy Memorial Day) on rainy days and 1090 on non-rainy days. From the same data we can say that on average the net number of entries during an entire rainy week is $34,124$ more than during a non-rainy week if we exclude a rainy Memorial Day.

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 the Mann-Whitney $u-$test we can see that the probability of picking two sample groups with such differently ordered values is 5% or below, $p_{\text{two-tail}} \leq 0.05$, depending on what we choose to call out-lier data. This means that we can safely reject the null hypothesis and state that rainy and non-rainy days are significantly different. However, if we look at the computed a linear models for the subway entries we get the following results for the value and change in value of R^2,

$$\begin{aligned} \text{rain as a feature } R^2 &= 0.477668664672 \\ \text{rain not a feature } R^2 &= 0.477643344627 \\ \Delta R^2 &\approx 0.000025 \end{aligned}$$

These changes are barely noticeable and therefore rain is not very important in terms of our predictive linear model. Far more importance is placed on the Mann-Whitney $u-$test in rejecting the null hypothesis, the average `ENTRIESn_hourly` values for rainy and non-rainy days and the removal of out-lier data such as Memorial Day.

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. Analysis, such as the linear regression model or statistical test.

---

1. One of the most glaring inadequacies of our given dataset is the fact that it only spans one month. This month may not be adequate, no pun intended, to get a realistic snapshot of NYC subway ridership. Another aspect on which this dataset falls short is the fact that there are only nine rainy days of data and 20 of non-rainy day data. Although the actual number of sample data points is in the tens of thousands for each category, humans function on a daily, weekly, monthly and yearly schedule system. In this way it could be argued that these are not independent and randomly selected samples.

2. The Mann-Whitney $u-$test, especially with Memorial Day removed, was a very important step in showing the statistical difference between rainy and non rainy entries numbers. However, as discussed earlier, the values of R^2 barely changes when the `rain` feature is utilized. To me this shows that while rain is important in explaining some of the variation, it is dwarfed in magnitude by the hour of the day variable. As the qualitative argument mentioned before with the elements of the $\boldsymbol{\theta}$ vector, we have

   $$\begin{aligned} |\theta_{\texttt{rain}}| &\ll |\theta_{\texttt{Hour}}| \\ 16.328 &\ll 468.801 \end{aligned}$$

   The linear regression falls short because the vast majority of variation occurs with the hours of the day, and that variation is a series of six spike and decay patterns which are not linear in nature.
   In summary, the Mann-Whitney $u-$test is great for showing that rain matters in terms of ridership but the linear regression falls short as a predictive formula because the vast majority of the variation is *non-linear* and occurs with the hours of the day.

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

I enjoyed having a national holiday in the dataset which allowed us to analyze it both with and without this out-lier included. This process of being discriminating about what information to include in our models is part of real world data analysis.