Louw Otto (ID: 3836408615)

Part 2: Analyzing the NYC Subway Dataset

Udacity Data Analyst Nanodegree - Introduction to Data Science

23 May 2015

TABLE OF CONTENTS

Section 0. References	2
Section 1. Statistical Test	3
Section 2. Linear Regression	5
Section 3. Visualization	9
Section 4. Conclusion	15
Section 5. Reflection	16

Section 0. References

Problem Set	Reference	Information Gleaned
Problem 2.5	https://docs.python.org/2/library/csv.html#examples	Reading and writing CSV files
Problem 2.5	Udacity Discussion Forum	(This was a hard problem for someone with limited python experience.) The forum provided the idea of reconstructing each row separately using a for statement and row element references
Problem 2.6	Udacity Discussion Forum	Getting to the count of rows by looking at other examples. Use of \n to go to a new line
Problem 2.11	https://docs.python.org/2/library/datetime.html#datetime.datetime.strptime	“datetime objects”, “strftime()” and “strptime()” behavior
Question 1.1	Understanding the Mann-Whitney U Test - supplement from Udacity	Understanding the basic theory around the Mann-Whitney U Test
Question 1.2	https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php	Assumptions for Mann-Whitney U test
Question 2.6	http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit	Understanding and explaining goodness of fit for linear regression
Question 3.1	http://discussions.udacity.com/t/project-requirement/16731 and https://github.com/yhat/ggplot/issues/382	Solve the problem of the dataframe for days without rain not having an index 0 by reseting the index
Question 3.1	http://ggplot.yhathq.com/docs/index.html	Familiarisation with ggplot functionality
Question 3.1	http://stackoverflow.com/questions/23964236/python-ggplot-rotate-axis-labels	Rotation of axis labels for visualisations
Question 3.1	http://stackoverflow.com/questions/25061822/ggplot-geom-text-font-size-control	Adjust all text sizes on gglot
Question 3.2	https://docs.python.org/2/library/datetime.html#datetime.datetime.strptime	Extract day of the week for plotting
Question 3.2	http://stackoverflow.com/questions/16729483/converting-strings-to-floats-in-a-dataframe	Convert string to floats to split turnstiles dataset into two, turnstiles with ±6 observation periods and those with 24 observation periods

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?

- A **Mann-Whitney U-Test** was used to analyse the NYC subway data.
- **Two-tailed P value:** The two-tailed test is the more conservative option in favour of not rejecting the null hypothesis.
- The **null hypothesis**: **H_0:** $P\ (\ X_{(\text{with rain})} > Y_{(\text{without rain})}\ ) = 0.5$
- The **p-critical value**: $p_{critical} = 0.05$

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** was selected based on the following:

- The datasets are samples of ridership over the period of month of May 2011 and there is not a full understanding of the population.
- The datasets distribution is non-normal resulting in settling for a non-parametric test, namely the Mann-Whitney U Test.

In using the **Mann-Whitney U Test** the following assumptions was made regarding the datasets:

- There is **independence of observations** and the hourly ridership entries observations was randomly made.
- The **independent variables** consists of **two categorical, independent groups.** (‘rain’ is 0 or 1)
- Distributions of datasets are unknown.

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.

- **p-value:** 0.049999826 (2 x *scipy* returned p-value for two tailed test)
- **mean**$_{\text{(with rain)}}$**:** 1105.4463767458733
- **mean**$_{\text{(without rain)}}$: 1090.278780151855

1.4 WHAT IS THE SIGNIFICANCE AND INTERPRETATION OF THESE RESULTS?

The returned p-value is smaller than that of the chosen critical p-value, and therefor **the null hypothesis can be rejected**. The probability that the sampled value of $X_{\text{(with rain)}}$ is higher than $Y_{\text{(without rain)}}$ is statistically significant. Therefor it can be assumed that there is a difference in ridership, measured by hourly entries at turnstiles, between days with rain and days without rain.

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 STATS MODELS**
- **OR SOMETHING DIFFERENT?**

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?

The following fields were used in the last version of the model (Table 1):

Fields in Dataframe	Used in model as:
(Index)	
UNIT	Dummy Variable
DATEn	
TIMEn	
Hour	Input Variable
DESCn	
ENTRIESn_hourly	
EXITSn_hourly	
maxpressurei	
maxdewpti	
mindewpti	
minpressurei	
meandewpti	
meanpressurei	
fog	Input Variable
rain	Input Variable
meanwindspdi	
mintempi	Input Variable
meantempi	Input Variable
maxtempi	Input Variable
precipi	Input Variable
thunder	

Table 1: Input and Dummy Variables used in 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.

The initial reasoning, before looking at the data, was based on **intuition**. My initial hypothesis was: People will use the subway to avoid delays in travelling as a result of poor visibility and slippery road conditions. This assumption immediately raised **fog and rain** as key features to influence ridership.

The second step was to **look at the data** to determine what data there was to work with in the model. I noticed temperature and it immediately occurred to me, having used the London Subway in the middle of summer, that temperature can play an important role in the comfort of travelling using subways. Analysis of the data, however, highlighted that the data was only for the month of May which had quite a mild **mean temperature** but I was considering using the extremes in my model, namely **maximum temperature** and **minimum temperature**. It occurred to me that the **hour** of travel will influence overall ridership volumes. I also realised that the amount of **precipitation** in inches would influence ridership and could potentially be more influential than the rain indicator because the precipitation is provided for the time and location.

The third step was to experiment and explore of how adding and removing features influenced R^2. The table below shows how my exploration unfolded (Table 2):

Description of Action	Features	R^2	Difference
Initial test with four features: Rain, Hour, Average Temperature and Precipitation	rain, Hour, meantempi, precipi	**0.463968815**	Starting position
Add Fog	rain, Hour, meantempi, precipi, fog	**0.46449236**	Improve
Remove Rain	Hour, meantempi, precipi, fog	**0.464469607**	Reduced, therefor I added Rain back
Remove Precipitation	rain, Hour, meantempi, fog	**0.464482928**	Reduced, therefor I added Precipitation back
Add Maximum and Minimum Temperatures and remove Average Temperature	rain, Hour, precipi, fog, maxtempi, mintempi	**0.465023535**	Improve
Add Average Temperature	rain, Hour, meantempi, precipi, fog, maxtempi, mintempi	**0.465033995**	Improve
Remove Maximum T Temperature	rain, Hour, meantempi, precipi, fog, mintempi	**0.464905027**	Reduced, therefor I added Maximum Temperature back
Increase cycles to determine theta to 100 (from 75)	num_iterations = 100	**0.465051837**	Improve
Further experimentation could improve model but this is where my investigation concluded			

Table 2: Exploration and Experimentation to improve R^2

2.4 WHAT ARE THE COEFFICIENTS (OR WEIGHTS) OF THE NON-DUMMY FEATURES IN YOUR LINEAR REGRESSION MODEL?

Fields in Dataframe	Weights:
UNIT	N/A (Dummy)
Hour	468.61442904
fog	71.40277148
rain	-0.93503146
mintempi	-99.76451712
meantempi	-29.05598109
maxtempi	52.30845128
precipi	-7.75373038

Table 3: Coefficients for features used in linear regression model

2.5 WHAT IS YOUR MODEL'S R2 (COEFFICIENTS OF DETERMINATION) VALUE?

Coefficients of determination: $R^2 = 0.465051837$

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 R2 VALUE?

The coefficient of determination (R^2) is a statistical measure that indicates how closely a fitted regression represents the observed data. Our value of ±0.465 indicates that 46.5% of the variability in the observed data is explained by the model. The appropriateness of this value depends on what data is modelled, in this case human decisions to choose the subway on the account of varying weather conditions. Human behaviour is unpredictable and an R^2 below 0.5 is common, and therefor 0.47 is reasonable. It should be considered that a non-linear model could provide more flexibility to get a better fit.

Section 3. Visualization

3.1 ONE VISUALIZATION SHOULD CONTAIN TWO HISTOGRAMS: ONE OF ENTRIESN_HOURLY FOR RAINY DAYS AND ONE OF ENTRIESN_HOURLY FOR NON-RAINY DAYS.

Two histograms below show the distribution of hourly entries for days with rain (Figure 1) and days without rain (Figure 2):

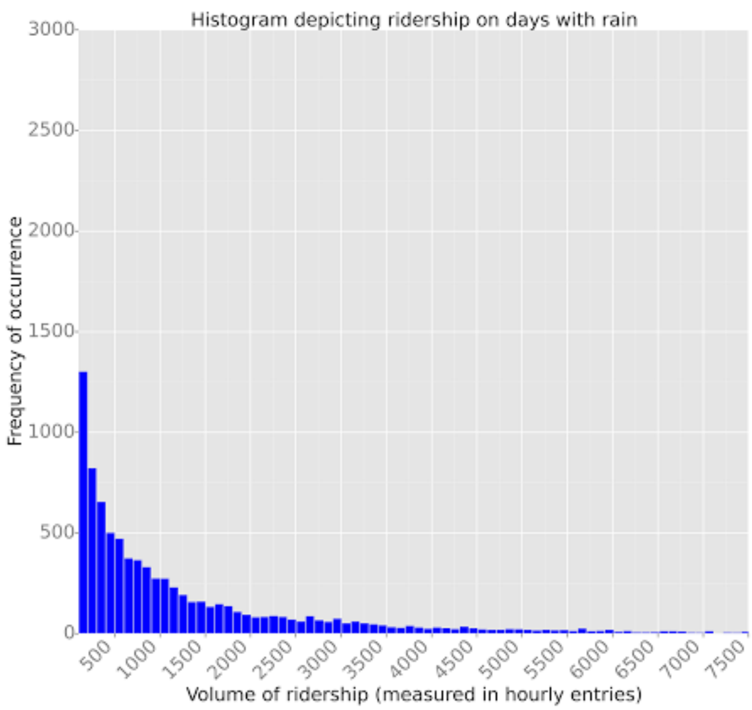

Figure 1: Histogram showing ridership on days with rain

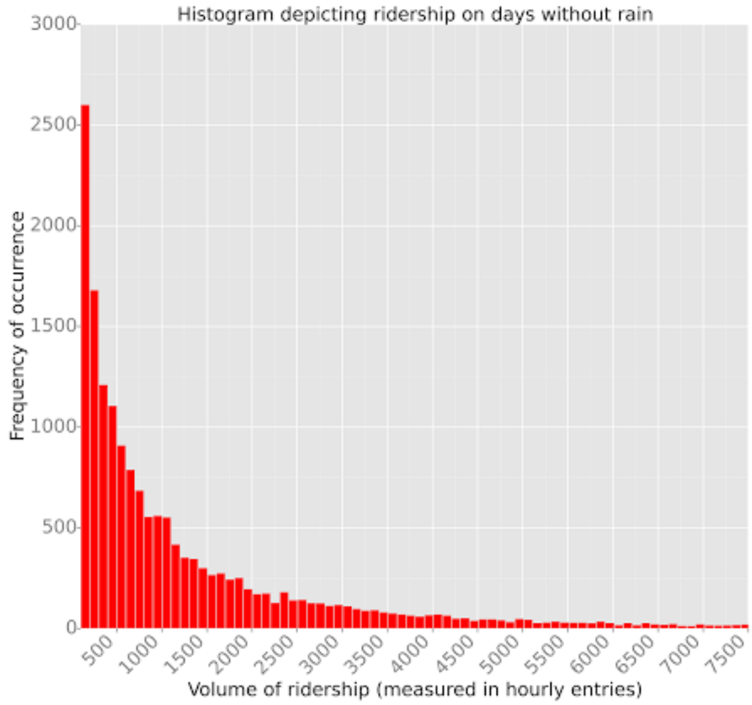

Figure 2: Histogram showing ridership on days without rain

Adjustments to Graphs

- To make the graphs more readable for comparison entries per hour below 100 has been assumed as inactive periods and has been truncated.
- Entries per hour above 7500 was truncated to make the two figures easier to compare.

Observations:

- Distributions for both samples are similar and positively skewed. This type of non-normal distribution is common in natural circumstances where low volumes are common or the default.
- As expected the total frequency of observations for the rainy days are lower based on there having been less rainy days. The histograms present the absolute entries and not the entry averages.

3.2 ONE VISUALIZATION CAN BE MORE FREEFORM. 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**
- **RIDERSHIP BY DAY-OF-WEEK**

Ridership by time-of-day: Figure 3 is the initial graph produced for average ridership by time-of-day using the entire dataset.

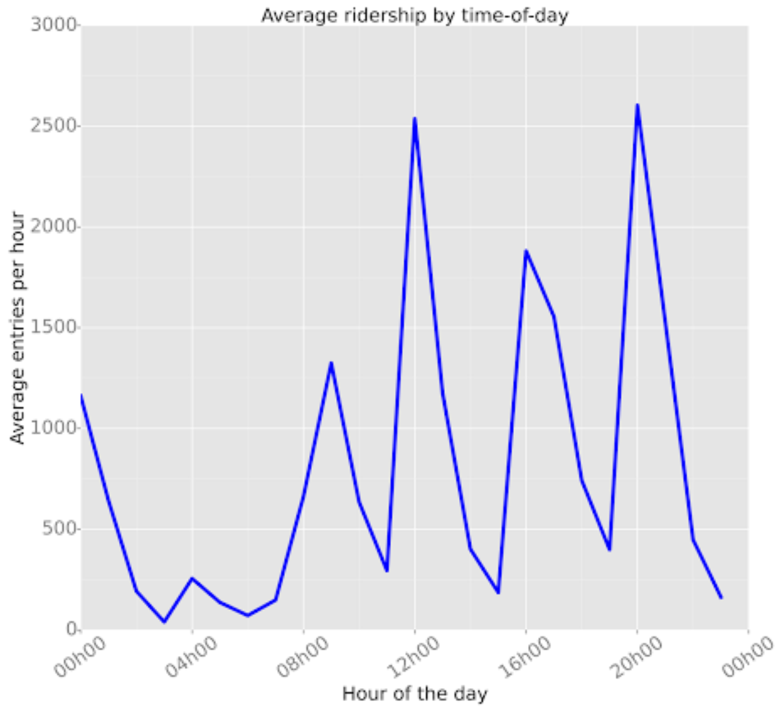

Figure 3: Average ridership by time-of-day using the entire dataset

Initial observations:

- The expectation was that the profile of ridership will be smoother and gradual. Upon doing the above graph I went back to study the data. The problem with the dataset is that the observations was not consistently taken across hours, skewing the plot of the aggregated data on the continuum . For Example:
  - Many turnstiles has 6 observations with ±30 observations in total for the month

  - The hours when these 6 observations were taken are inconsistent between turnstiles. Many observations was taken at 00h00, 04h00, 08h00, 12h00, 16h00 and 20h00 but others were observed at 01h00, 05h00, 09h00, 13h00, 17h00 and 21h00.
- In reviewing the data, it was observed that turnstiles R540 to R552 have observations covering all 24 hours. For this data however there seems to have been multiple observations for a given hour, inconsistent with the dataset of R001 to R536 where a single aggregation of observations was included for each of the observed hours.
- To see whether this data provide the expected profile two graphs was plotted:
  1. R540 to R552 (green) which looks to have observations covering every hour of the day.
  2. R001 to R536 (orange) where observations was included per hour but for inconsistent hours between turnstiles.

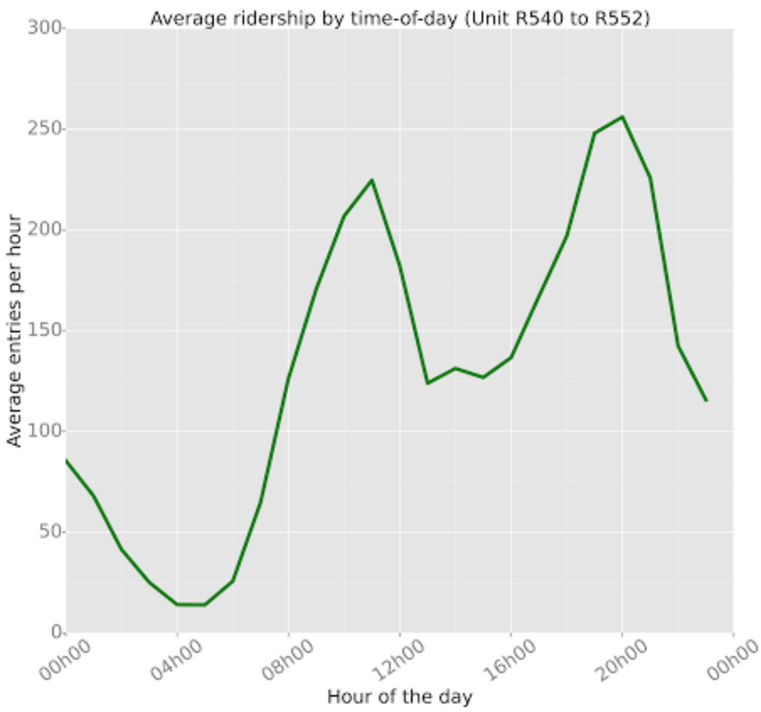

Figure 4: Average ridership by time-of-day for turnstiles with observations across 24 hours.

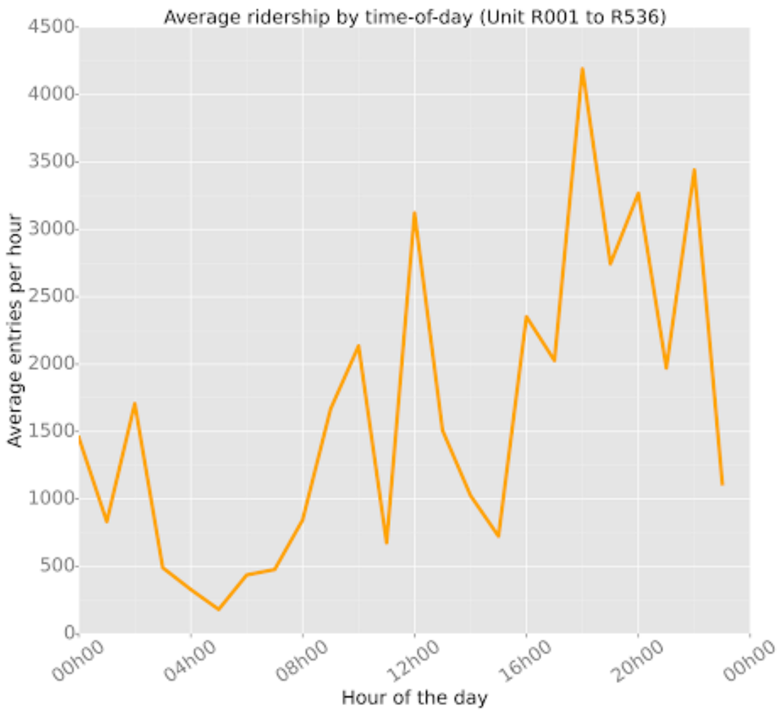

Figure 5: Average ridership by time-of-day for turnstiles with inconsistent observations.

Second set of observations:

- It appears that the observations captured for turnstiles R540 to R552 displays the smooth profile expected showing two rush hour periods:
    - One centred around 11h00 ± 2hours and
    - The second centred around 20h00 ± 2hours
- Despite the irregularity of the second graph covering turnstiles R001 to R536 there seems to be a hint of a similar profile displaying two rush hour periods. The recommendation would however be to attempt understand how the data was observed and aggregated for this dataset.
- The average entries for the turnstiles with the “more granular” observations (R540 to R552) is lower. Investigation is required whether these stations really have less foot flow or whether the data set for turnstiles R001 to R536 was correctly aggregated.

Ridership by day-of-weekday: Figure 6 shows the total ridership (indicated by hourly entries) by day-of-week using all data.

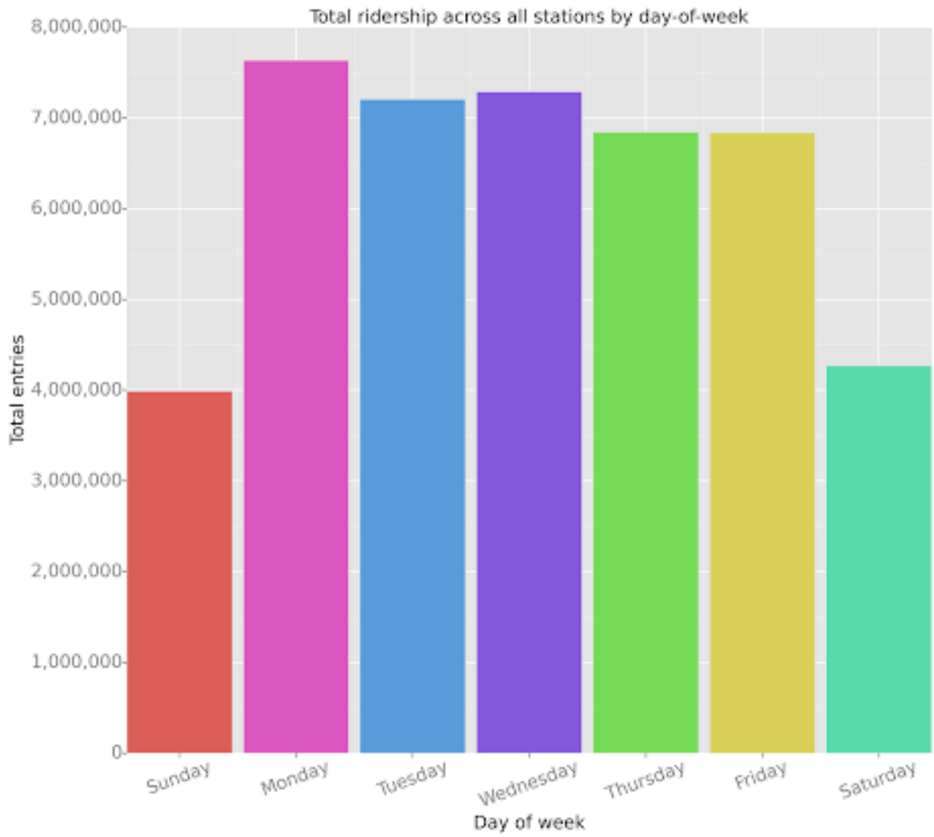

Figure 6: Total ridership by day-of-week

Observations:

- As expected ridership volumes are lower during the weekend than during the week.
- Saturday seems to be slightly busier than Sunday.
- Monday appears to be the day with the highest total ridership but this conclusion cannot be drawn from the totals displayed, because the dataset includes one more Sunday and Monday, in total five of each, compared to four days of observations for each of the other days of the week.

Section 4. Conclusion

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?

For the month of May 2011 ridership was observed for the first 30 days of the month. During this period there was 10 days with rain. Using entries per hour as a proxy for ridership, observations for hours on days with rain was compared to the observations for hours on days without rain. The average turnstile entries for hours on days with rain across all observed turnstiles was 1105 where the average for hours on days without rain was 1090. The Mann-Whitney U Test showed that the probability of observations on days with rain being higher than observations in hours on days without rain was statistically significant. Using the dataset as it is provided, it is concluded that more people ride the subway when it rains compared to when it does not rain. As will be discussed in Section 5 this conclusion is made reluctantly before a more detailed investigation is done to address inconsistencies in the dataset.

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** was used to test whether there was a statistically significant difference in ridership between days with rain and days without rain. Using a critical p-value of 0.05 allowed us to reject the null hypothesis that the probability of observing a higher or lower volume of ridership between days with rain and days without rain was equal. As mentioned the average ridership was 15 riders higher for days with rain compared to days without rain. In the **linear regression analysis** it was shown that rain was one of the factors that influenced ridership. Other factors influencing ridership was the presence of fog, the time of day (hour), the amount of precipitation, the minimum temperature, the maximum temperature and the average temperature.

Section 5. Reflection

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.**

1.1 Variability of dataset: As a start I will, when possible, use the complete dataset available for my analysis. I completed my analysis on the Udacity Portal which meant that the dataset for analysis ranged from 18,000 to 131,951 data points. For the reduced datasets little information was provided on how the data was randomly rationalised. Understanding the dataset is critical for drawing the right conclusions.

1.2 Inconsistency in observations: Applying consistent methods to measure or obtain observations is critical to ensure that the collective dataset provides an accurate representation of reality. As mentioned in Question 3.2, on closer inspection of the data, it became evident that observations wasn't consistently measured or aggregated across turnstiles. This created skewed views of ridership across NYC when totalled.

1.3 More data for a better understanding: It was determined that more riders choose the subway on days with rain but which mode of transport do these subjects use when it does not rain. Understanding both sides of the equation will improve the accuracy and reasoning behind the conclusions.

2.1 Linear Regression: Human behaviour is obviously difficult to model using linear regression based on humans' unpredictability, but I used it based on its introduction in this course. I need to spend time to explore and familiarise myself with more sophisticated modelling options that will suite my future analyses more accurately.

2.2 Statistical Test: The tip of the iceberg, in terms of statistical methods, was discussed in this course. I need to expand my statistical toolbox beyond the Welch's T Test and the Mann-Whitney U Test.

5.2 (OPTIONAL) DO YOU HAVE ANY OTHER INSIGHT ABOUT THE DATASET THAT YOU WOULD LIKE TO SHARE WITH US?

One of the key lessons learnt was that a deep understanding of the dataset is critical to draw the correct conclusions, I gained a lot of insights by slice-and-dicing the data to understand slight nuances like, for example, how many Mondays was included in the observed data or is observations made across the same hours. In retrospect I would have set up my own instance of Python to have more control over the dataset and analysis.

This was a great introduction course and really wetted my appetite for the possibilities and applications of Data Science. Thank you.