P2 - Analyze the NYC Subway Dataset – Laura Hoyte

Section 0. References

https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test#Normal_approximation

http://www.isixsigma.com/tools-templates/hypothesis-testing/making-sense-mann-whitney-test-median-comparison/

http://stats.stackexchange.com/questions/116315/problem-with-mann-whitney-u-test-in-scipy

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.levene.html#scipy.stats.levene

http://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm

http://ggplot.yhathq.com/docs/index.html

https://docs.python.org/2/tutorial/datastructures.html

http://greenteapress.com/thinkpython/thinkpython.pdf

http://stackoverflow.com/questions/19482970/get-list-from-pandas-dataframe-column-headers

https://docs.python.org/2/tutorial/datastructures.html

http://docs.scipy.org/doc/numpy/reference/generated/numpy.corrcoef.html

http://pandas.pydata.org/pandas-docs/stable/genindex.htmls

http://www.dotnetperls.com/string-list-python

https://triangleinequality.wordpress.com/2013/12/04/creating-dummy-variables-in-pandas/

https://www.safaribooksonline.com/library/view/matplotlib-plotting-cookbook/9781849513265/ch04s04.html

http://matplotlib.org/examples/pylab_examples/scatter_star_poly.html

http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

http://stat.ethz.ch/~mmarloes/teaching/fall08/5-LinearRegression.pdf

https://plot.ly/matplotlib/bar-charts/

http://people.duke.edu/~ccc14/pcfb/numpympl/MatplotlibBarPlots.html

http://www.wunderground.com

Section 1. Statistical Test - *rain_histogram.py*

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?

The Mann-Whitney U statistical test was used to analyze the NYC subway data.
The task is to analyze whether more people ride the subway when it is raining versus when it is not raining. Therefore a two-tailed test was performed.
The null hypothesis for this task is as follows:
Is the population mean/median of people riding the subway when it is raining greater than the population mean/median of people riding the subway when it is not raining?
If we draw randomly from the two populations, one where it is raining (r) and the other where it is not (nr) whether one distribution is likely to have higher a value than the next.
The null hypothesis is the probability that r > nr is equally as likely as the probability that nr > r.
Therefore a two-tailed test is used.
H_o: P(r > nr) = 0.5
H_1: P(r > nr) ≠ 0.5
p-critical value = 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 test is used when data is not normally distributed as shown in Figure 1 below. The data must also come from independent samples, have equal variances and the same distribution shape. The sample sizes do not need to be equal.
No. of samples (no rain) = 33,064
No. of sample (rain) = 9,585
Hence the Mann-Whitney U test is used as it does not assume that the data is drawn from any particular distribution, the distributions have the same shape or that the sample sizes are equal.
The Sharp-Wilcox test was also performed to test if the data is a normal distribution.

rain = 0
Shapiro-Wilcox statistic = 0.595618069172
p-value = 0.0

rain = 1
Shapiro-Wilcox statistic = 0.593882083893
p-value = 0.0

The p-value is less than 0.05, hence the null hypothesis that the data is from a normal distribution can be rejected.

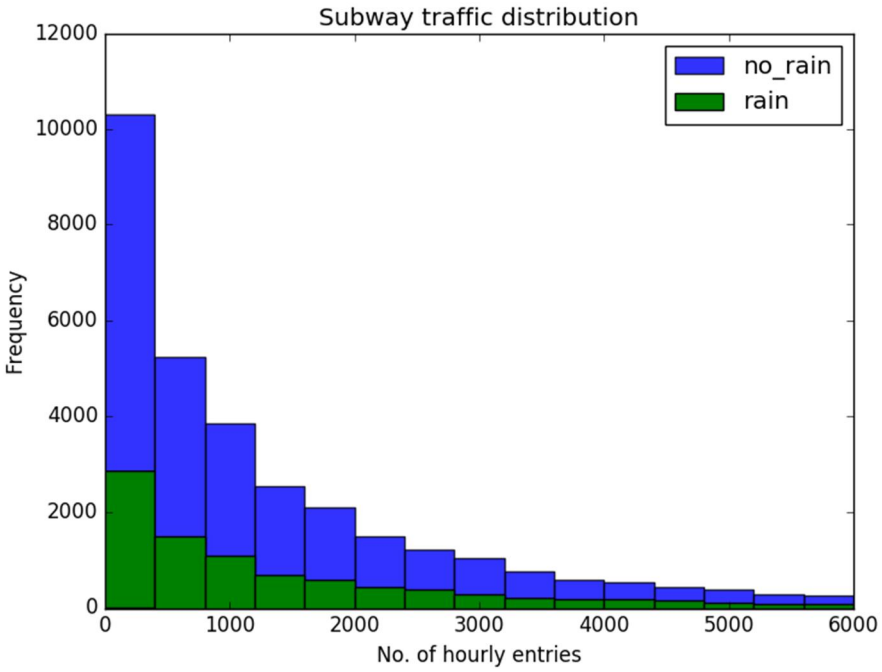

Figure 1: *Subway traffic left skewed non-normal distribution*

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.

μ_{nr} = 1,845.54
μ_r = 2,028.20
U = 153,635,120.5
p-value = 5.483e-06

Note:- p-value returned a value of nan. Based on the recommendations in MannWhitneyU.pdf p-value was found manually. For large sample sizes U approximates a normal distribution. *See mannwhitneyu.py*

1.4 What is the significance and interpretation of these results?

The p-value at 5.483e-06 is less than the p-critical value at 0.05, hence we reject the null hypothesis. If a particular station is chosen at random on average more people will ride the subway when it is raining. This can also be seen from μ_r and μ_{nr} as on average 183 more people ride the subway when it is raining.

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?s

Gradient Descent using Scikit Learn

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

Features used in the model are precipi, wspdi and rain, along with the dummy variables UNIT, conds, hour and day_week.

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.

Features were chosen using a combination of factors.

1. UNIT, station, latitude, longitude, weather_lat and weathr_lon are all location-based features and give a pretty good indication of where the subway traffic occurs, hence nothing would be gain by using multiple inputs from this list. As a result only UNIT was chosen. The ENTRIESn_hourly also has a wide range of daily entries across all station as shown in Table 1 and Figure 9 (Section 3). Due to this broad variation in the hourly traffic at each location, predicting subway traffic has to be done with location in mind. This is also illustrated by the model as performance dropped to $R^2 = 0.168423$ when UNIT was excluded from the input features.

ENTRIESn_ hourly	count	mean	std	min	25%	50%	75%	max
	42,649	1,886.59	2,952.39	0	274	905	2,255	32,814

Table 1: *Summary statistics for hourly subway traffic at every UNIT*

2. Numerical continuous values such as pressure, temperature, precipitation, and windspeed might be a better indication of weather conditions. This differs from the rain and fog variables that are binary. The fact that it is raining does not mean that much as it could have been light rainfall, but this still qualifies as rain (rain = 1). Therefore rain and fog variables are not good indicators of how traffic will vary. The variable conds gives a more explicit description of the weather at any period of time and is used in the model.

Model	rain only	fog only	rain + fog	no rain + no fog
R^2	0.54275	0.54246	0.54377	0.54184

Table 2: *Effect of rain and fog on model performance*

3. The subway traffic will have peak times. For example during morning and evening rush hour traffic. Likewise the stations will have more entries on Sundays to Thursdays compared to Fridays and Saturdays. Hour and day of the week are both time-based variables and were used in the model. However hour and day_week are not measuring any quantitative value and are in fact categorical, hence both were treated as dummy variables.
   Hour was selected as a feature based on the average hourly trend seen in Figure 3 below. There is a clear relationship between hour and certain peak traffic times of the day, e.g.12pm and 8pm. There is also a lot of traffic at 4pm when evening rush hour begins.
   These relationships are also seen by the R^2 values in Table 3.
   weekday is not used as a feature as it is redundant with the inclusion of day_week and the performance of the model is better with day_week.

Model	hour only	day_week only	hour + day_week	no hour + no day_week	weekday only	hour + weekday
R^2	0.518808	0.418275	0.543789	0.399222	0.414005	0.541600

Table 3: *Effect of time series on model performance*

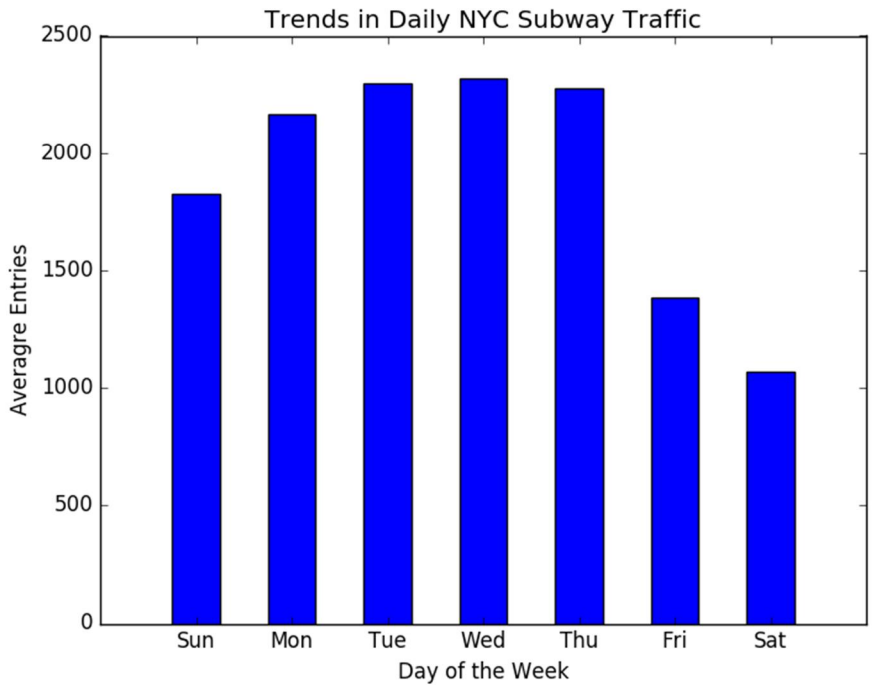

Figure 2: *Average daily NYC traffic*

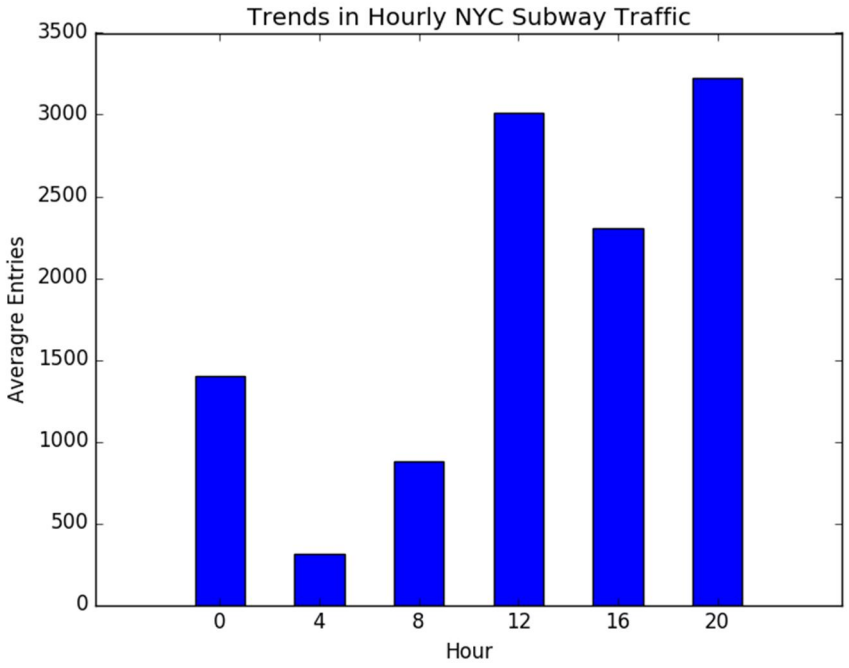

Figure 3: *Average hourly NYC subway traffic (with peaks at 12pm and 8 pm))*

4. Numerical variables precipi, and pressurei were not chosen due to the underlying data values.

	count	mean	std	min	25%	50%	75%	max
precipi	42649	0.004618	0.025832	0.0	0.0	0.0	0.0	0.3
pressurei	42649	29.971	0.137942	29.55	29.89	29.96	30.06	30.32
wspdi	42649	6.9279	4.5102	0.0	4.6	6.9	9.2	23.0
tempi	42649	63.1038	8.4556	46.9	57.0	61.0	69.1	86.0

Table 4: *Summary statistics for the four numerical weather variables*

- <u>Precipitation - precipi</u>
  Insufficient non-zero values as shown in Figure 4 and the mean, minimum, 75% quartile and maximum values in Table 4. This feature is not used in the model.

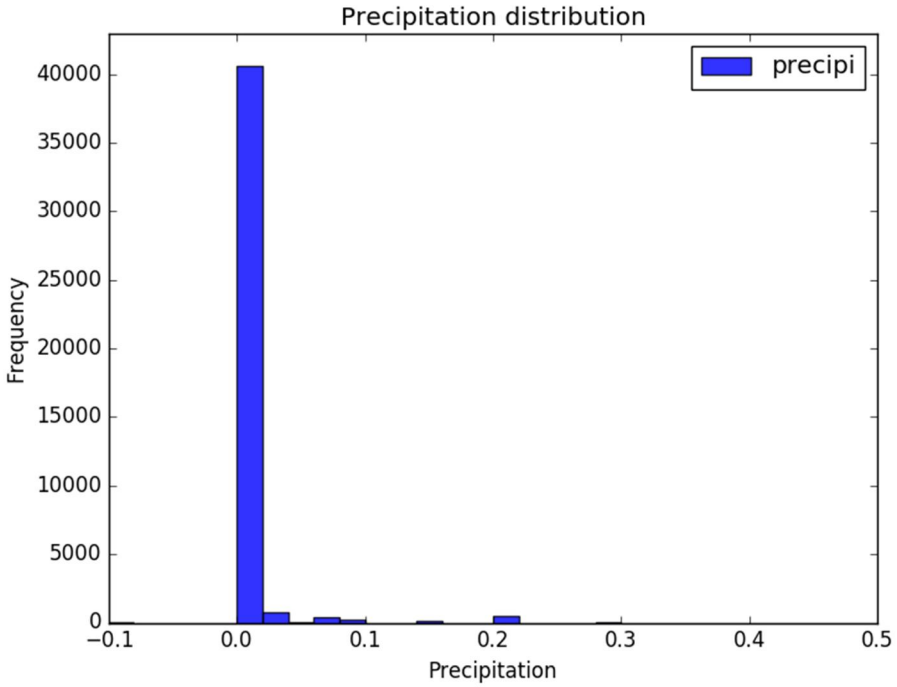

Figure 4: *Most (over 40,000 out of 42,649) precipi values are close to zero*

- Pressure - pressurei
  This variable has a very small range of 0.77, hence it is not a good selection for an input feature.

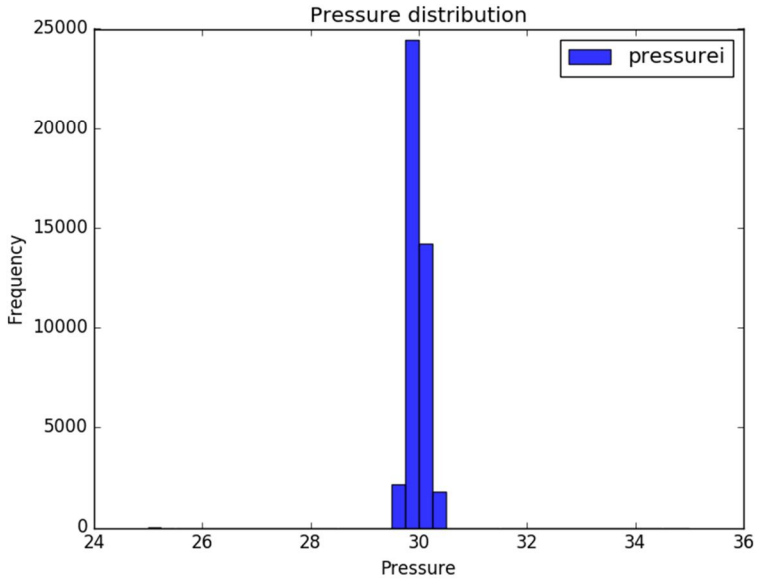

Figure 5: *pressurei has a very narrow distribution*

- Wind speed – wspdi and Temperature - tempi

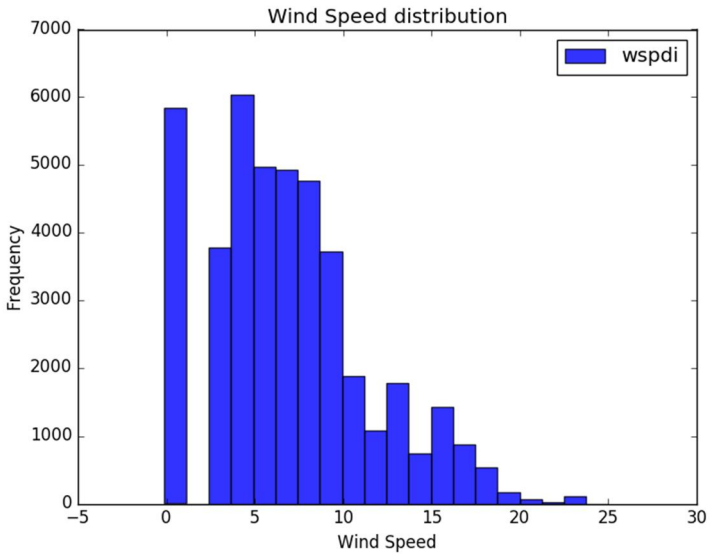

Figure 6: *wspdi has a great distribution of values*

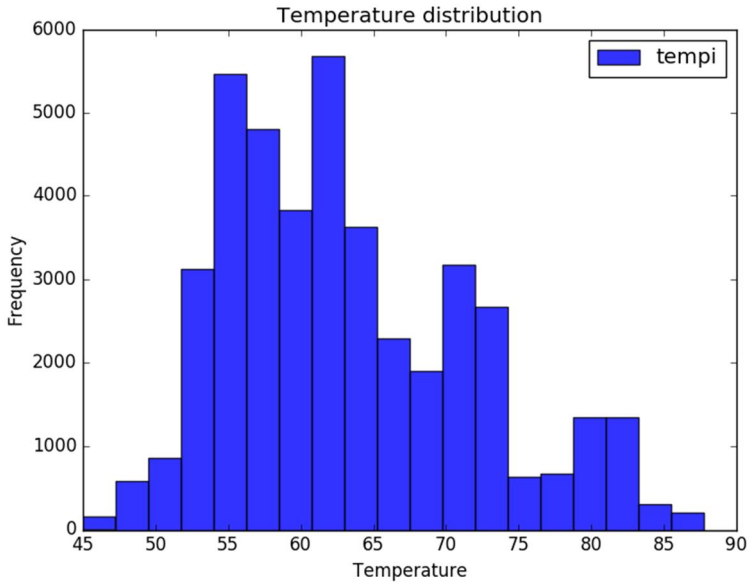

Figure 7: *tempi has a great distribution of values*

wspdi and tempi are chosen as numerical inputs to the model as these two features have better variability in the data.
Combined these features also give the very important “feels like” temperature rather than the actual temperature.

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

intercept = 2,953.2849
wspdi = -2.1991
tempi = -16.7065
rain = 84.0548

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

	Features	alpha	iterations	R^2
1.	UNIT, conds, hour, day_week, wspdi, tempi	0.0001	1,000	0.54184
2.	UNIT, conds, hour, day_week, wspdi, tempi, pressurei	0.0001	1,000	0.54288
3.	UNIT, conds, hour, day_week, wspdi, tempi, pressure, precipi	0.0001	1,000	0.54143
4.	UNIT, conds, hour, day_week, wspdi, tempi	0.0001	3,000	0.54404
5.	UNIT, conds, hour, day_week, wspdi, tempi	0.0001	5,000	0.54494
6.	UNIT, conds, hour, day_week, wspdi, tempi	0.005	1,000	0.54256
7.	UNIT, conds, hour, day_week, wspdi, tempi	0.001	1,000	0.54222
8.	UNIT, conds, hour, day_week, wspdi, tempi, pressurei	0.001	3,000	0.54388
9.	**UNIT, conds, hour, day_week, wspdi, tempi, rain**	**0.0001**	**3,000**	**0.54453**

Table 5: *Variations in model performance*

The final model chosen has a R^2 value of 0.54453 (no. 9.). There is no benefit in increasing iterations to 5,000 from 3,000 as the improvements in R^2 is only by 0.0009 (0.17%), yet it takes longer to find the local optima. Rain was included in the final model, however the improvement in performance is marginal with R^2 increasing by 0.09%.

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 R^2 value means that 54.45% of the relationship between weather conditions and subway traffic can be explained by the model. The remaining 45.6% is due to some other criteria. The other criteria can be additional features not in the underlying dataset or it can be that the relationship between weather and subway traffic is non-linear.

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.

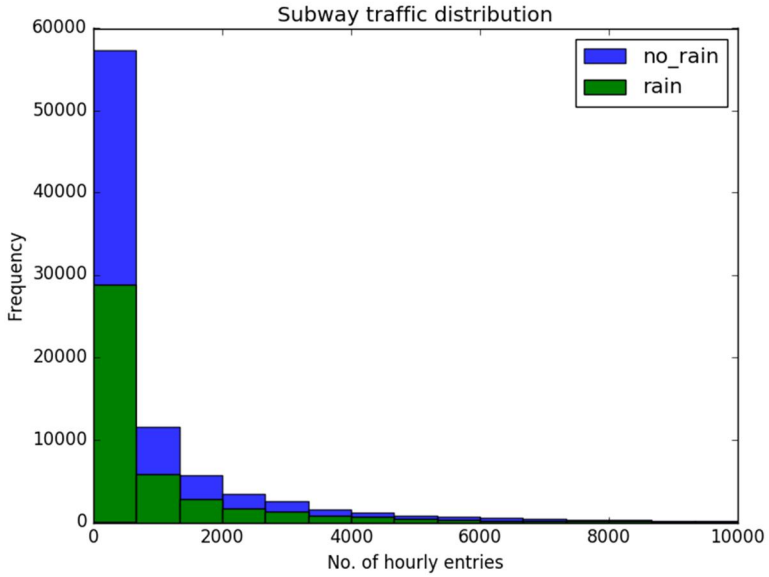

Figure 8: *How the subway traffic distribution varies with rain*

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)

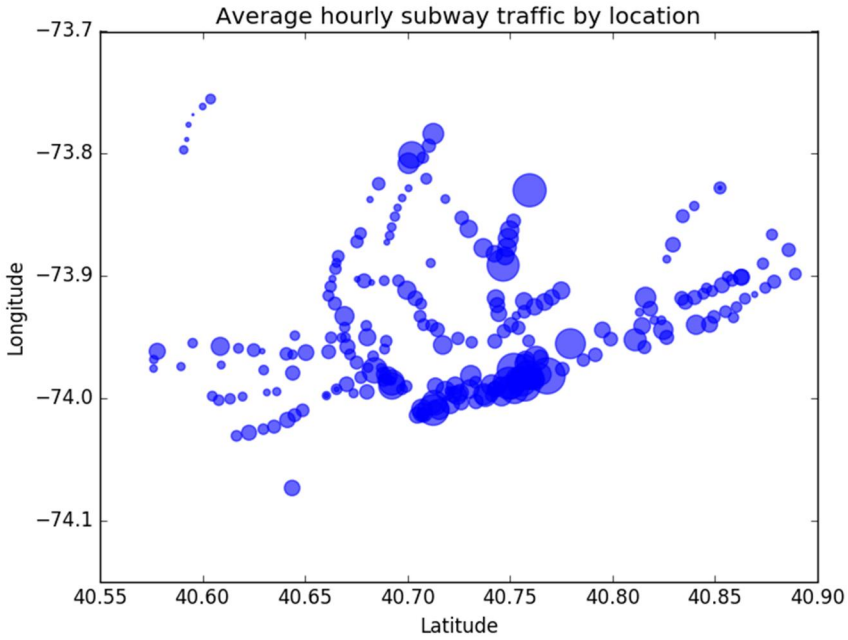

Figure 9: *Average hourly NYC traffic at each station*

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?

More people ride the subway when it is 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

Based on the Mann-Whitney statistical test and the returned p-value of 5.483e-06, μ_{nr} = 1,845.54 and μ_r = 2,028.20, more people ride the subway when it is raining. However the linear regression model returned a R^2 value of 0.54404 without the rain variable as an input and 0.54453 when rain is included. In the overall scheme of things rain is only very marginally responsible for changes in subway traffic with a 0.09 % increase in model performance. Unit, hour, day_week, and conds have a greater impact on changes in subway traffic.

Section 5 Reflections

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. **Dataset**

   Data is for one month only the month of May. This may be insufficient data points to build a good model.

   Slightly more people ride the subway when it rains, however the time series data has a greater impact on model than rain or weather in general. More data is needed to get a better assessment of weather impact.

   Some stations have very little traffic and more than likely add very little to the performance of the model. If these stations were removed from the input data, then the model might give better results.

   Data was reported every four hours, hence there might be some information loss compared to if the data given every hour.

   The data was not divided into two separate training and testing samples. This might have improved model performance.

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

   Linear regression using gradient descent only gives results for the local optima and not the global optima. There can be possible improvements to the performance of the model if the global optima was found OLS method.

   Linear regression assumes an underlying linear model, but the relationship between subway traffic and weather is not completely linear.
   The model gives an R^2 value 0.5445, hence only 54.45% of the relationship between the features and the output can be attributed to a linearity.

Residuals	count	mean	std	Min	25%	50%	75%	max
	42,649	-8.60	1,994.59	-8,758.15	-888.44	-128.64	688.79	22,889.41

Table 6: *Summary statistics for model's residuals*

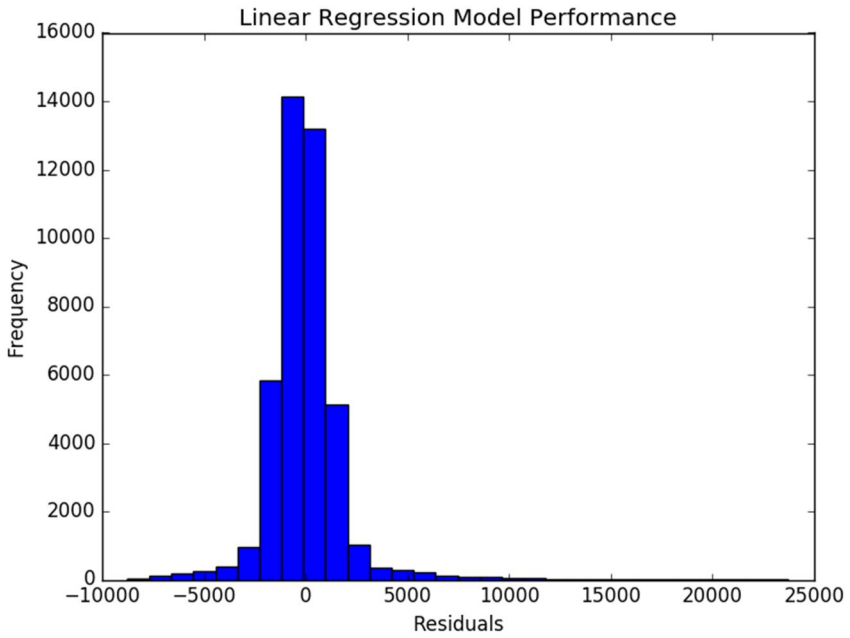

Figure 10: *Mode performance with both positive and negative residuals and standard error of 1,995*

The residuals approximate a normal distribution, hence the standard deviation is a good estimate of the residual standard error = 1,994.59. On average the linear regression model to predict hourly subway traffic results in an error of almost 1,995 entries. 2/3 of the residuals fall in the range ±1,994.59 and 95% fall in the range ±3,989.18.

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

There is not a lot of variation in weather data for May and months with more extreme weather conditions will give a better results.

New York City (NYC) has four seasons. May is just one month in spring when NYC experiences pretty moderate conditions. Winter is known for low temperatures, lots of precipitation (snow) and wind chill (wind speed); extreme conditions from December to February. Likewise summer is known for hot and humid conditions; extremes from June to August. The seasonality of weather needs to be taken into account to get a better idea about how subway traffic varies with changing weather conditions.

One years worth of data encompassing the four seasons will lead to a far better dataset. Using one year of daily gives the following distributions for the two numerical features (wind speed and temperature) used in the model.

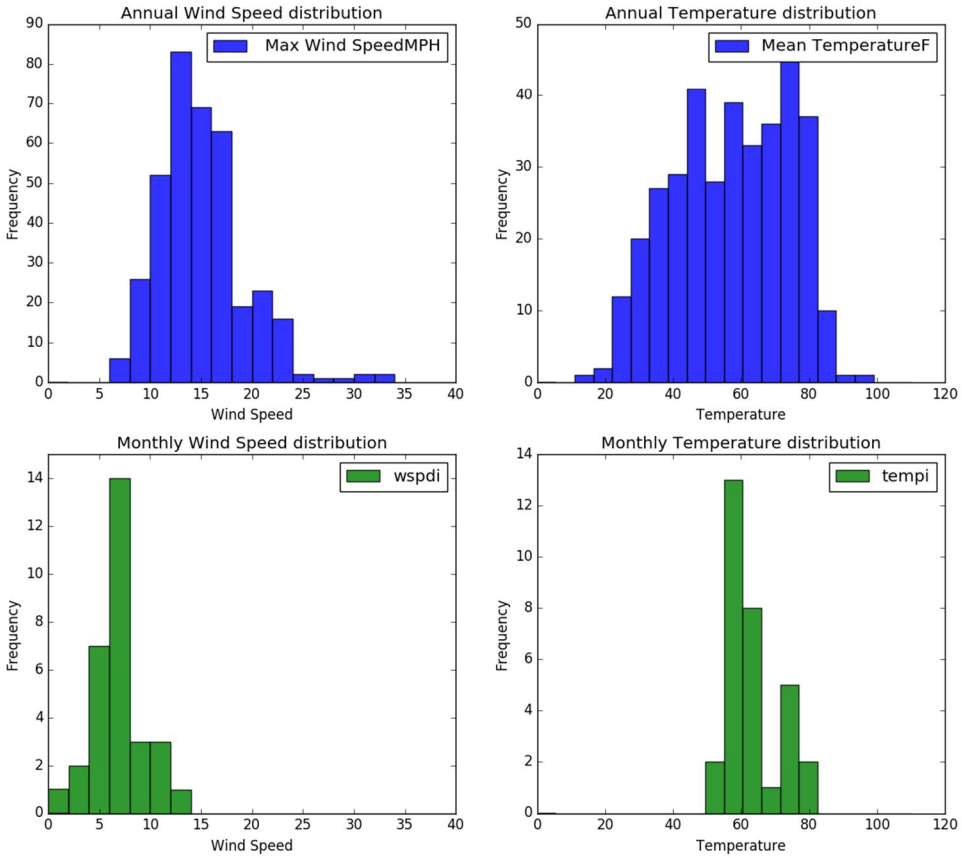

Figure 11: *Comparison of daily wind speeds and temperature using one month and one year datasets. (Note mean temperature and maximum wind speeds were used for annual dataset)*