Analyzing the NYC Subway Dataset

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 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 stackoverflow.com, try to include a specific topic from Stackoverflow that you have found useful.

Normal Probability Plot: http://en.wikipedia.org/wiki/Normal_probability_plot
Two-Sided Mann-Whitney U Test: http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
Time Series Introduction: http://www.maths.bris.ac.uk/~guy/Research/LSTS/STSIntro.html

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?
Mann-Whitney U test.
I used two tailed P value.
Null hypothesis: the distributions of the number of entries between rainy and non-rainy days are not statistically different.
p-critical value: 5%.

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 two samples (rainy and non-rainy) are not normally distributed and unequal sized population (the non-rainy samples, 33,064, is much larger than 9,585 of rainy samples). Therefore, it's better to use Mann-Whitney U test.

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.
Using Mann-Whitney U test.
Two-sided p-value: 0.000548%
mean values:
- rainy: 2028.196
- non-rainy: 1845.539

1.4 What is the significance and interpretation of these results?

Given that the two-sided p-value of 0.000548%, which is much lower than the critical p-value (5%), the null hypothesis is rejected, i.e. the mean value of rainy day ridership is statistically larger than the non-rainy day ridership mean value.

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 used OLS from statsmodels.formula.api and fitted the linear model by formula.

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 'unit', 'hour', 'rain', 'day_week', 'meantempi' as the input features in my model. I used 'unit', 'hour' and 'day_week' as the dummy variables. I chose 'unit' as a dummy variable because it is naturally categorical. 'hour' and 'day_week' are included as dummy variables because they don't necessarily follow the linear relationship as we see later.

The statsmodels OLS formula is as below:
ENTRIESn_hourly ~ C(UNIT) + C(hour) + rain + C(day_week) + meantempi

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."

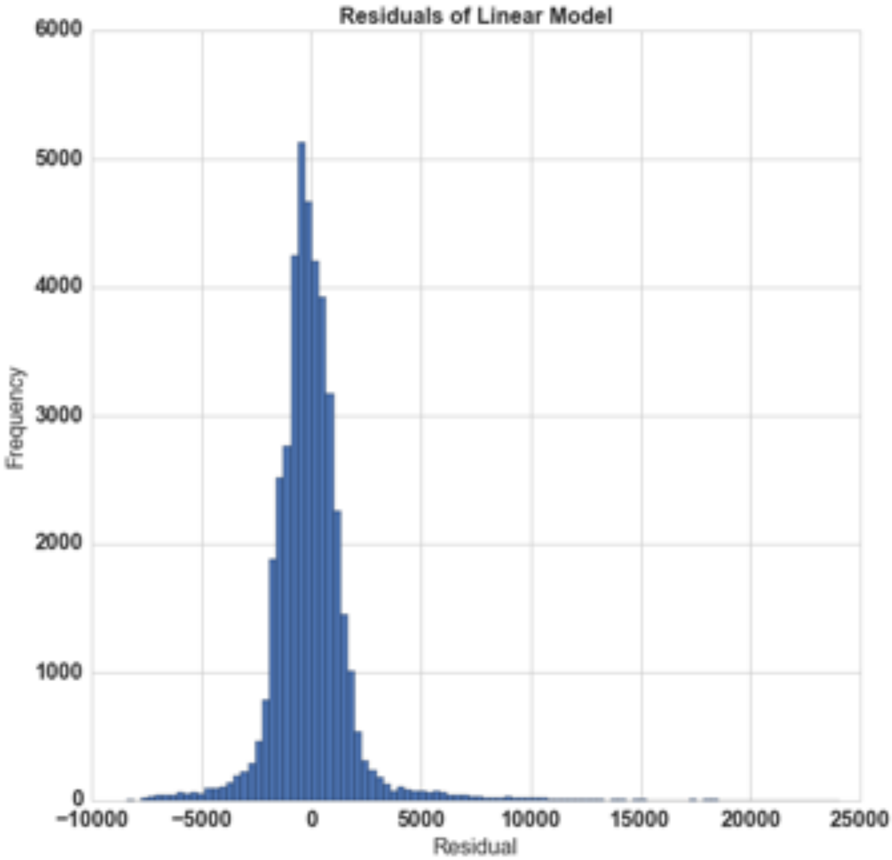
I used 'unit' because it is believed that the ridership follows different patterns from station to station and I also found it with better R2 when I included it in my model. I used 'rain' because it is intuitively true that people tend to take trains to avoid rain when it's a rainy day. I used 'meantempi' because it is intuitive that people tend to take trains to avoid high temperature. It is also intuitively correct that ridership correlates with date or time, for example people tend to take the subway during weekdays and peak hours.

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
rain: -42.9492
meantempi: -15.6128

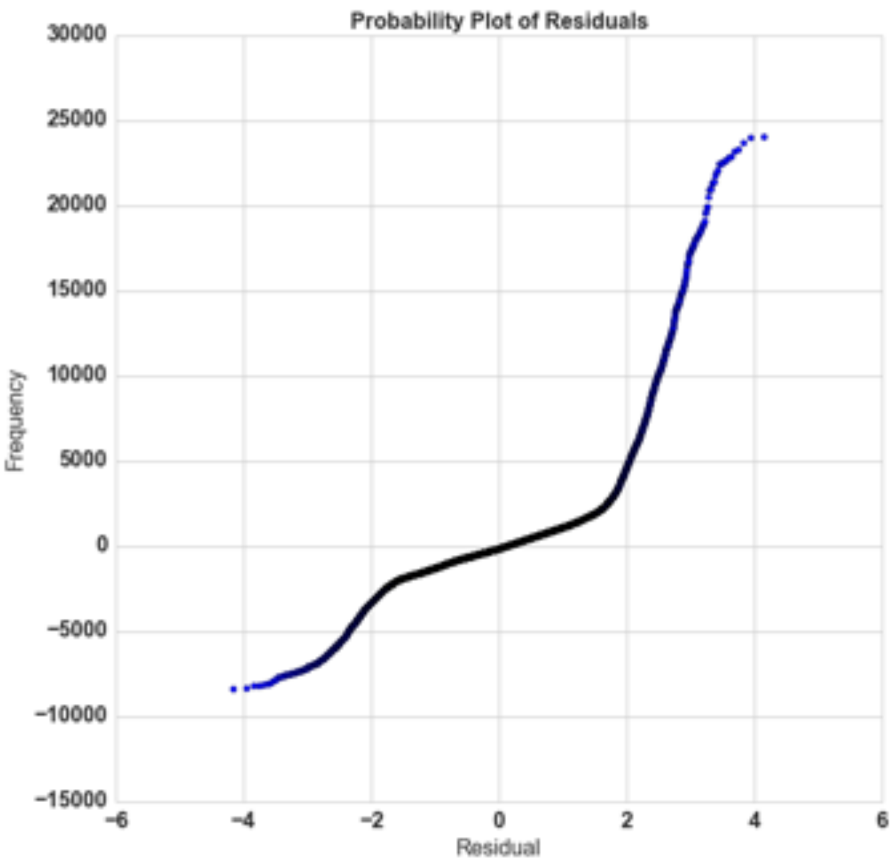
2.5 What is your model's R2 (coefficients of determination) value?
R2: 0.545

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 linear model explains 54.5% of the variability of the response data round its mean.

In order to evaluate the suitability of the linear model, it is proper to to check the normality of the residuals. Below are the plots of the histogram and QQ plot of residuals.



As shown from the residual histogram plot, the majority of the residuals fall within +/- 5000. However, it is noticed that there are some outlier residuals tailed over range of 10,000, which indicates there is skewness of the data, i.e. the normality of the model is not ensured.

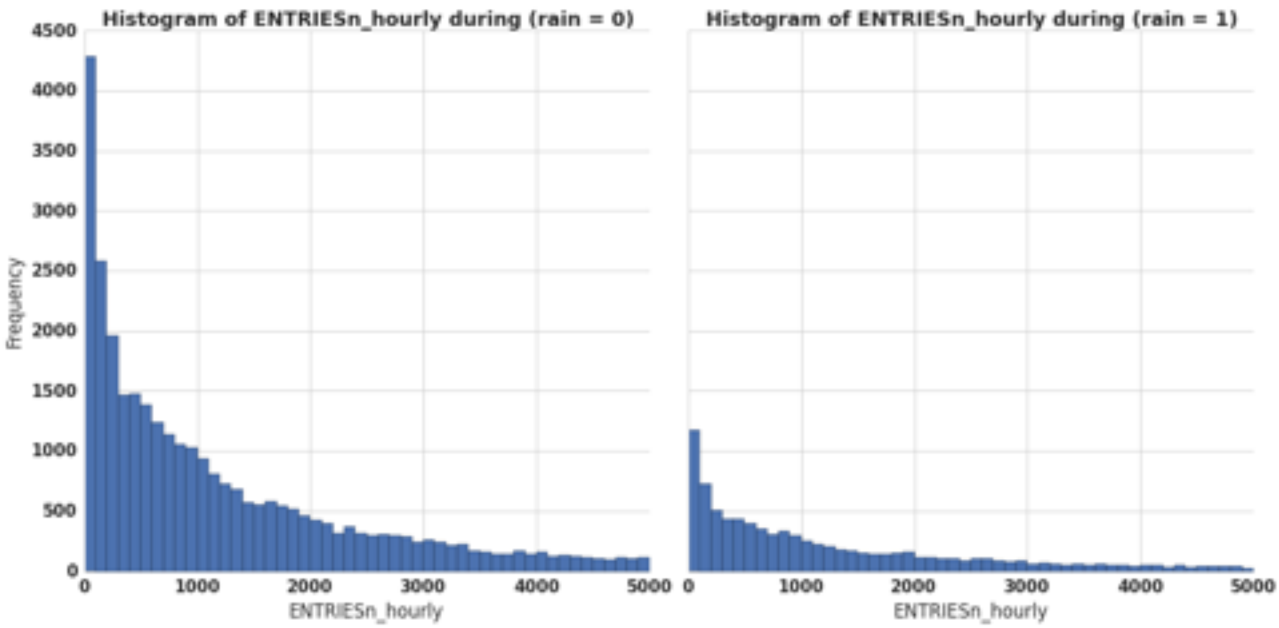


Above is the probability plot against normal distribution, from the probability plot above, it is clear that the distribution is right-skewed, which verified the previous doubt that there are heavy outliers over 10000.

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.

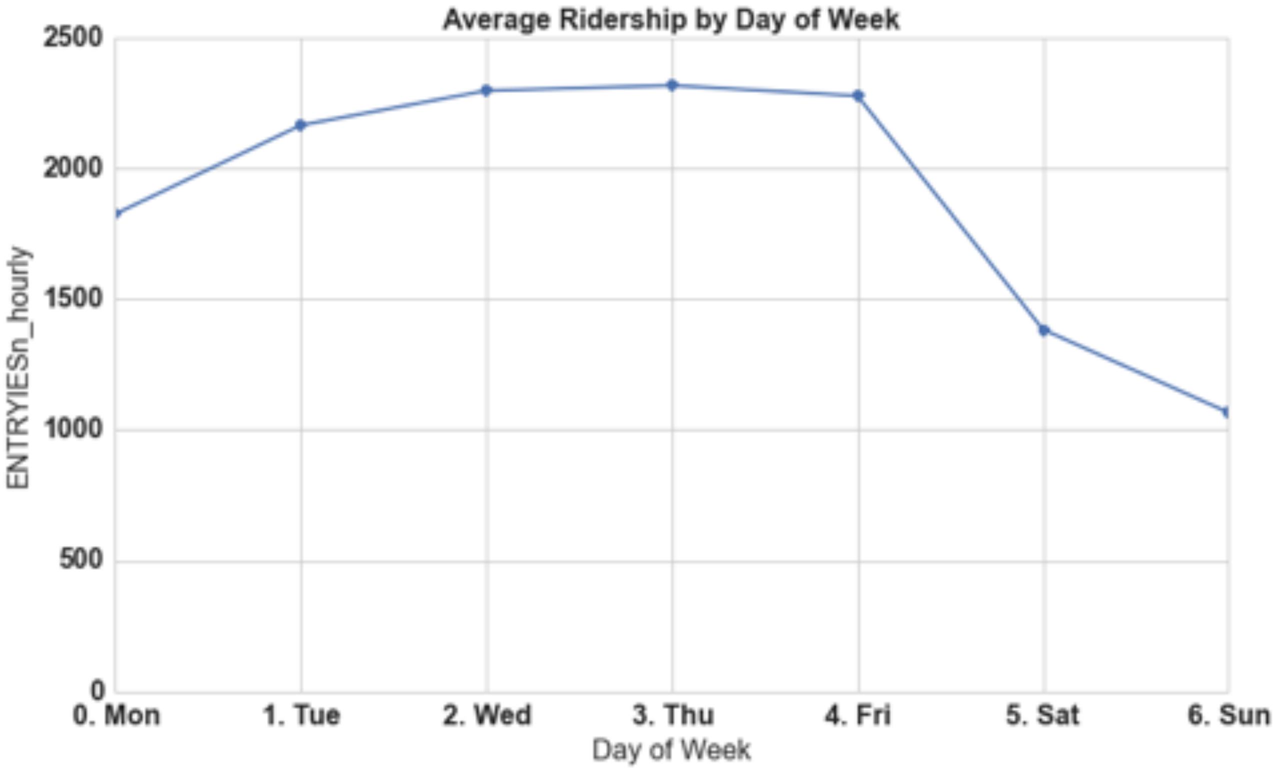
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.
You can combine the two histograms in a single plot or you can use two separate plots.
If you decide to use to two separate plots for the two histograms, please ensure 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.



From the above histograms of both rainy and non-rainy days' samples of ridership, it is evident that the ridership doesn't follow a normal distribution, but rather long-tail. It cannot be concluded in this plot that which, rainy day has more ridership than non-rainy days as there are more samples in non-rainy days than rainy days. Therefore, taking a Mann-Whitney U test is necessary to determine this hypothesis.

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



It is shown from the plot that the ridership during the weekend is much lower than that during weekdays. This makes sense as people tend to commute to work during the weekdays rather than weekends. Interestingly, the average ridership increases with the increase of day of week during the weekdays and peaked on Thursday. Monday ridership is particularly lower than the other days of week. Is it because of the fact that people tend to travel less at the beginning of the week after the weekend?

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?

The mean value of rainy days, 1,459, is 9.4% higher than that of non-rainy days, 1,333, indicating on average people tend to ride NYC subway more on rainy days than when it is non-rainy days. This is statistically significant. From the results of Mann-Whitney U test, the p value, 0.0042%, which is way below the critical p-value (5%), it is therefore safely to say the distribution of the two samples are statistically different.

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.

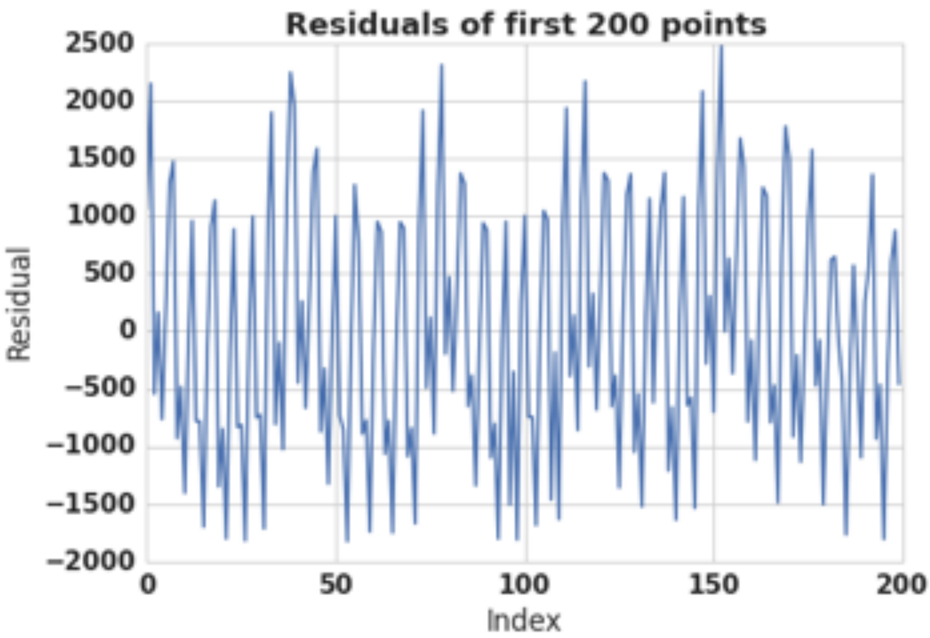
As stated in previous question, the Mann-Whitney U test gives very low p-value which indicates a statistically difference between the rainy-day and non-rainy-day samples. The regression analysis shows that ‘rain’ is positively correlated to ENTRIESn_hourly which should be interpreted as that ridership increases when it’s rainy.

Section 5. Reflection

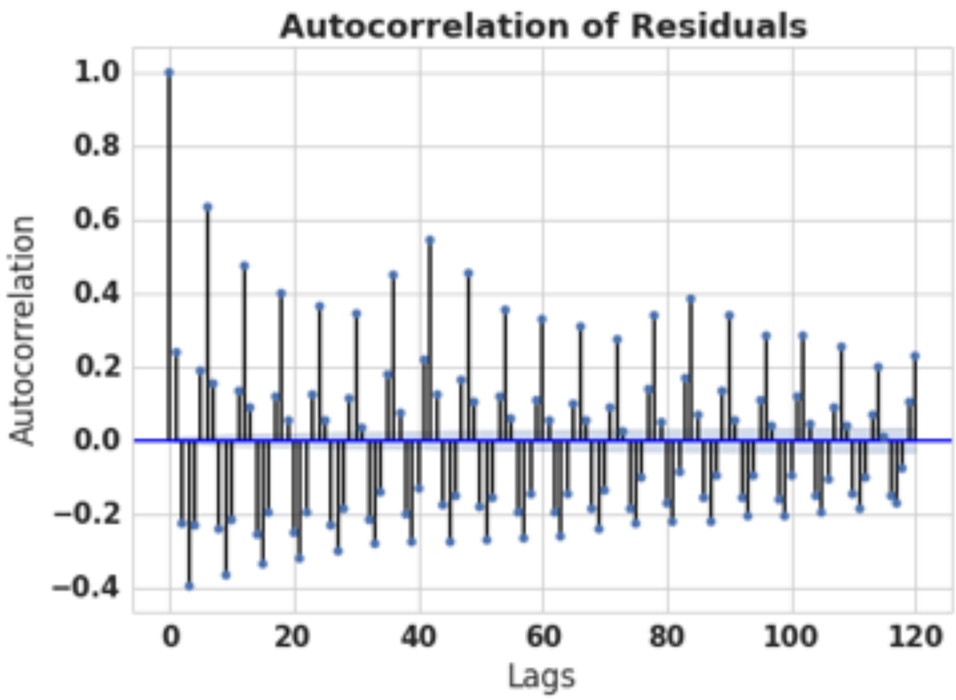
Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

5.1 Please discuss potential shortcomings of the methods of your analysis, including:
Dataset, Analysis, such as the linear regression model or statistical test.

The linear regression model only explains 54% of the variance, left with nearly half of the unexplained part.



Above chart shows the residuals of the first 200 points. It is clear that the line follows cyclic pattern which indicates some autocorrelation or seasonal factors. If we plot the autocorrelation of the residuals, we can find that the residuals' autocorrelation is high, therefore there are some time series factor that may contribute to the unexplained part by the linear regression model. It might be



better to design a non-linear model to address the variance that cannot be explained by the linear model.

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