Short Questions to Analyzing the NYC Subway Dataset

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 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 Mann–Whitney U test for dataset (turnstile_data_master_with_weather.csv) testing `ENTRIESn_hourly` against `rain` and `fog`.

As I don't assume how would rain or fog impact ridership, so I need two-tail P value. Null hypothesis is that samples when there is rain or fog aren't different from samples when there is no rain or fog. My p-critical value equals 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.

This statistical test applicable to the dataset because it doesn't make any assumptions about the distribution of ridership in two samples as it is essentially **non-parametric test**. It can be compared to general **t-test, which** assumes that distribution of outcome is normal.

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.

```
mean ENTRIESn_hourly when rain == 0
[1] 1090.279
mean ENTRIESn_hourly when rain == 1
[2] 1105.446
ENTRIESn_hourly by rain Mann-Whitney U test
W = 1924409167, p-value = 0.04988
mean ENTRIESn_hourly when fog == 0
[3] 1083.449
mean ENTRIESn_hourly when fog == 1
[4] 1154.659
ENTRIESn_hourly by fog Mann-Whitney U test
data:  ENTRIESn_hourly by fog
W = 1189034718, p-value = 2.4e-05 (two tailed)
```

1.4 What is the significance and interpretation of these results?

This values show that difference between samples is quite significant. As we can see both rain and fog seems to increase ridership.

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:

a. **Gradient descent (as implemented in exercise 3.5)**
b. **OLS using Statsmodels**
c. **Or something different?**

c. I used built-in linear regression (lm()) model in R which uses OLS method to fit model.

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

```
UNIT + day_week + Hour + rain + fog
```

All variables used as dummy variables except fog and rain.

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."**

Normal dataset: I decided to include `UNIT` variable, as it is the variable that capture most difference between locations, which, I think, is the biggest predictor of ridership at all. It is intuitively obvious for me – difference between some big, central station (like Grand Central) and some remote dead-end station will be the biggest no matter time, weather or something else. Second is `day_week` which is also obvious for me – people tend to use subway quite differently on weekdays and weekends or at Fridays (won't use car to be able to drink for example). Third is time, which is best represent is this dataset by `Hour` variable – there is big change in ridership in the peak hours, for example. `rain` and `fog` I included after some experimentation and building couple of models. They showed some increase in R^2 and have low p-value so I decided to leave them in the model to capture weather impact on ridership.

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

	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	3509.61	122.61	28.625	< 2e-16	***
rain	-64.39	11.31	-5.693	1.25e-08	***
fog	70.81	14.11	5.020	5.17e-07	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

Multiple R-squared: 0.514, Adjusted R-squared: 0.5121

F-statistic: 280.8 on 495 and 131455 DF, p-value: < 2.2e-16

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?

Such quite big values or R^2 mean that we have quite good fit of our regression model. Linear model is sufficient for this datasets as we can explain 51% of variability in outcome variable despite very big range of changes in it.

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.

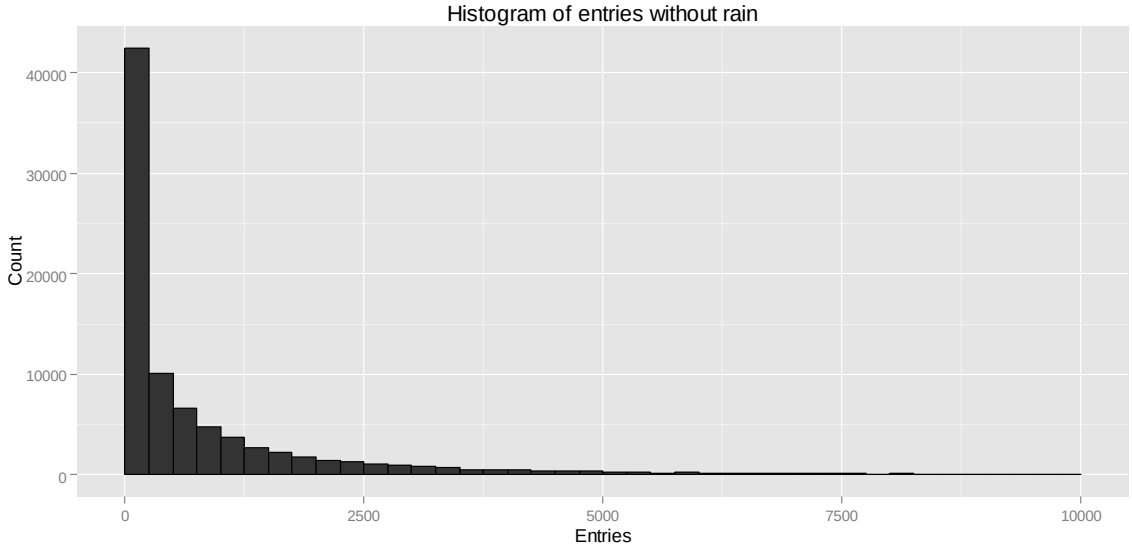
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

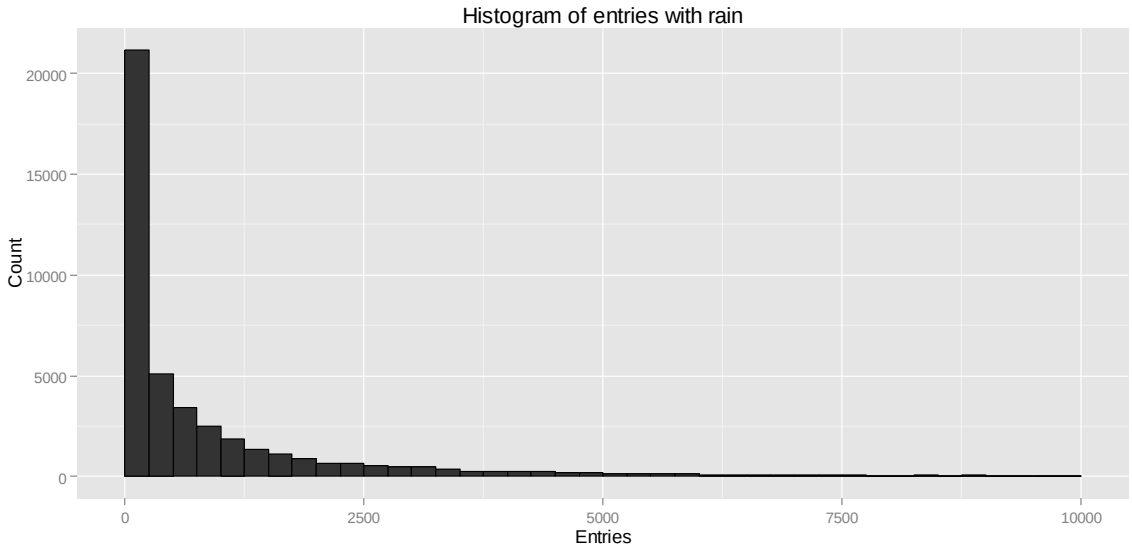
● You can combine the two histograms in a single plot or you can use two separate plots.

● 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.

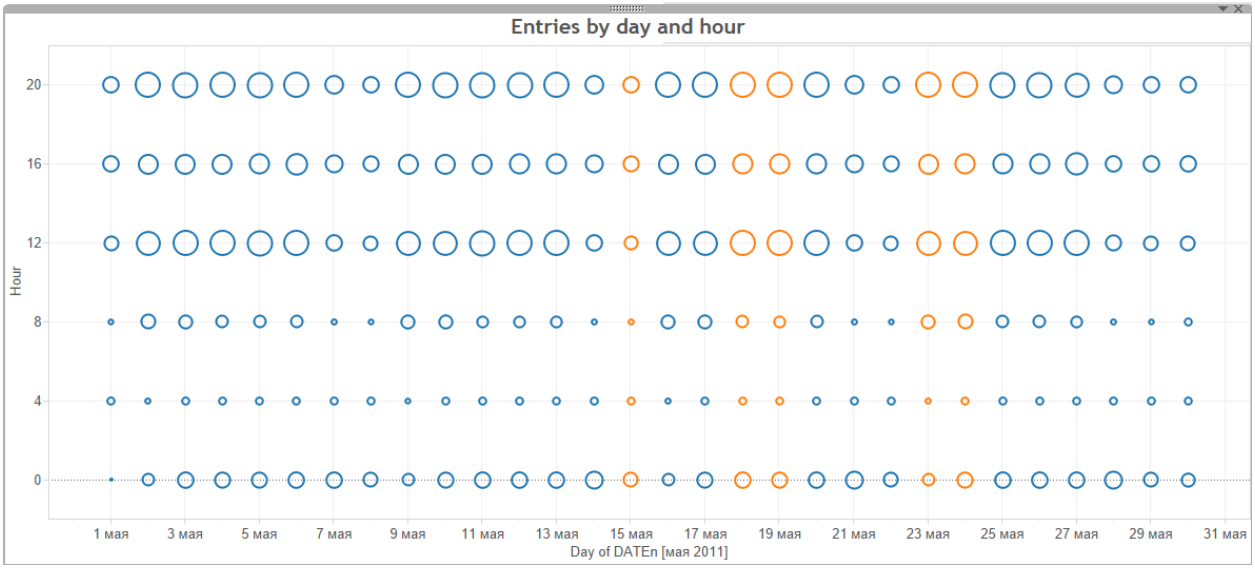




As we can see on this histograms, distribution is highly skewed to the left and have very long tail (some values going up to 60000 entries, so x-axis was truncated to the same value on both histograms for

comparison and to emphasize differences in distribution which otherwise unnoticeable). We can see that overall less entries made at rainy days (y-axis have much less values).

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:



Orange circles represent foggy days. Size of each circle represent cumulative ridership at this time and date. On this graph, we can clearly see effect of day of the week and time of the day.

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?

From my analisys we can see that **less** people ride NYC subway when it is raining if we account other factors like exact UNIT and day_week and Hour when measurements where made. What we can say about fog is quite opposite – at foggy days there are more people

-64.39	11.31	-5.693	1.25e-08
70.81	14.11	5.020	5.17e-07

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.

If we look at only statictical test we can think that more people ride subway at rainy days as Mann–Whitney U test shows us very strongly that difference between means in samples at rainy days and not rainy days is significant and mean number of ridership when it is raining is higher. However, if we account for other factors, like time of the day, day of the week and exact turnstile (Hour, day_of_week and UNIT variables respectively) we can see that ridership actually **decreases,** as coefficient of rain variable become negative and is very significant(-64.39 to be exact and p-value almost zero, see table in section 2.4).

For fog we can see that Mann-Whitney U test and linear regression show the same picture – both methods predict increase in ridership when it is foggy.

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. Potential shortcomings of datasets are clear for me: dataset include very fine resolution data on location (data represent each turnstile and differences between some of them located at the same station may be pure random) but very coarse weather data – one entry for each day averaging all the changes during the day.
   One suggestion is to take longer period, ideally couple of years to be able to account for seasonally changes in ridership. For one-month dataset, we don't know whether changes in ridership between days are from our predictors or it is just some changes from beginning of May and end of May. Using multi-year dataset, we can adjust for these changes and be more assured in our conclusions. Also, we can ask – what if people will got used to rain and fog in the next month and its effect on ridership will fade away?
2. About analysis. Our models have quite high R^2 values. It can suggest that we can predict ridership quite well. But most of the predictive power of this model based on obvious fact – most of riders use subway the same way every weekday (and quite different on weekends). If we want more "real" predictive power to see some subtle changes like those from weather, we should first "smooth" our data – adjust for daily and hourly changes and for seasonal changes. Then maybe we should use some tools from time series analysis and forecasting. For example, we can look not on absolute values of entries but differences in time on different conditions. This analysis can be sensitive to outliers, i.e. some very big and rare values. Maybe I could use some robust method like Theil–Sen estimator.

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

Generally, I explained all of my insights in section 5.1.