NYC Subway Dataset – Questionnaire (Vijay Balakrishnan)

Section 1. Statistical Test

Which statistical test did you use to analyse 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)
I used the Mann-Whitney U test as the data doesn't assume normal distribution
I used a two-tail p-value – multiplying the p-value by 2
Null Hypothesis: There's no significant variation in NYC subway ridership based on weather (rainy/dry day)
The p-critical value is 0.05 to evaluate the hypothesis with 95% confidence

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.
A)

Mann-Whitney U test is most applicable for this dataset because,

- NYC subway ridership need not be a normal distribution
- The difference between the datasets can be either +ve or -ve
- Also, the readings recorded are totally independent of each other
- The dataset isn't a continuous one

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.
A)

```
Good job! Your calculations are correct.

Mean ridership on rainy days: 1105.4463767458733

Mean ridership on non-rainy days: 1090.278780151855

One-tailed p-value: 0.024999912793489721

Two-tailed p-value: 0.048
```

What is the significance and interpretation of these results?
A)
The null hypothesis can be rejected with 95% confidence – ie: with 95% confidence we can state that ridership does vary based on weather (rain/no-rain).

Also,

The given dataset is just a random collection of ridership data and some additional metrics in a random month of the year. The day/time/station when the snapshot is recorded, is not uniform and hence cannot be compared/correlated. The seasonal variations in ridership are not taken care of with given data. The geographical position of the station isn't factored in.

What would provide better insight is some geographical data about the position of the stations. Also, instead of a month's data alone, it would be interesting to collect last 3-4 years' data and compare similar months/fiscal weeks and weekends. Further, it would be useful if the measurement timings/dimensions are normalized – data recorded at same time each day across the years.

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?
A)
I used the Gradient descent approach to produce predictions of NYC subway ridership.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?
A)
Input Variables: Rain, Precipi, Hour and meantempi
Dummy Variable: I used the Unit as the dummy variable

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."

Variables - Rationale
Rain – When it rains more people will use subways – just logical to use this feature
Precipi – higher precipitation means higher chances of rain and therefore I used it as one of my features
Hour – The time of the day decides the amount of ridership – peak vs. non-peak
Meantempi – On a hot day, many would use subways – it was again just logical to use this feature

Dummy Variables - Rationale
Units – Added the extra dimension of subway stations because the actual route is also a factor of ridership.

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

The coefficients (theta) Rain, Precipitation, Hours and Meantempi respectively are as follows

```
[ 128.65369174  128.65369174  607.55207572  108.06458824]
```

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

```
Your r^2 value is 0.461129068126
```

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?

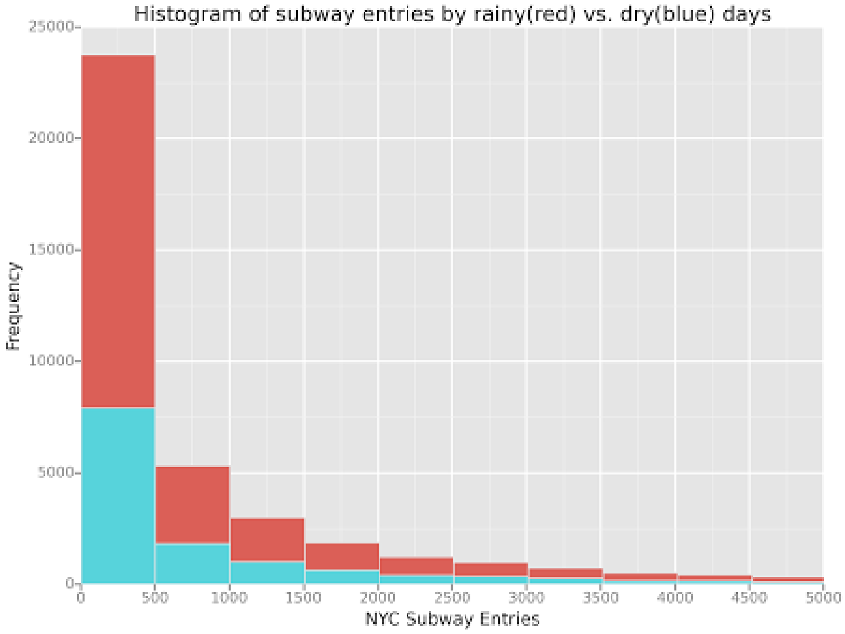
This R^2 value means that the predictions are accurate ~46%. I think this is appropriate for the given dataset. This means that 46% of the variations in the dataset has been explained and 54% is still unknown.

I believe that a polynomial regression would make more sense to predict ridership of the dataset as I am sure it would increase the R^2 value. Also, a higher % of R^2 can be achieved with more complex dataset that includes further dimensions

Section 3. Visualization

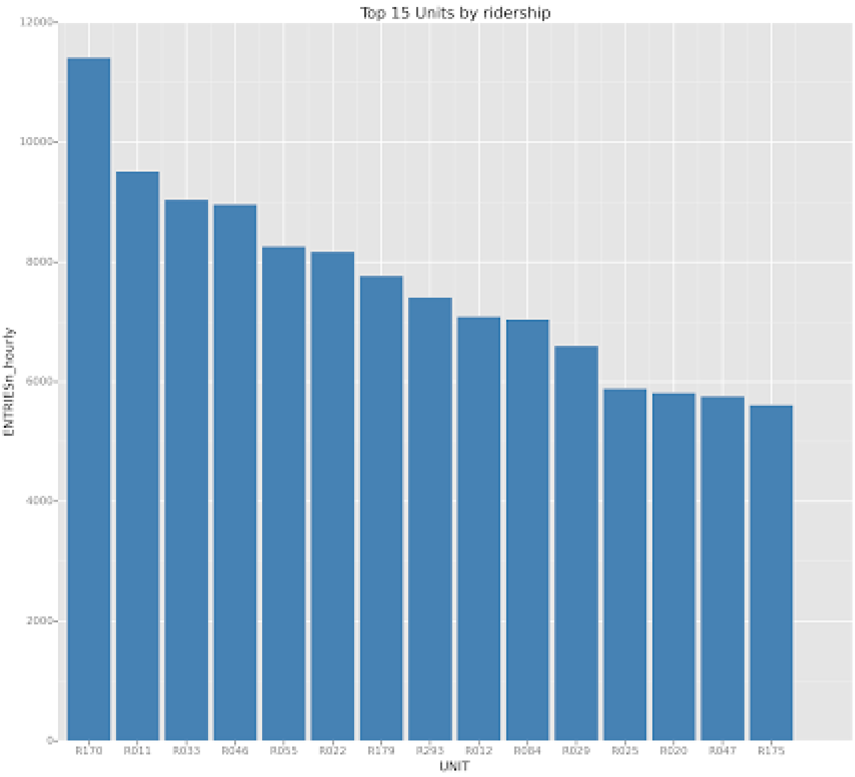
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.
Remember 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.



Note: Color legend isn't possible as it is a limitation in ggplot2.

3.2 One visualization can be more freeform. Some suggestions are:
Ridership by time-of-day or day-of-week
Which stations have more exits or entries at different times of 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?
A)
Yes, from my analyses based on given data, I statistically conclude that more people ride the NYC subway on rainy days. But the relationship looks very marginal and hence we would need more data points to be more confident regarding the same.

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.

From below analysis, the mean ridership on rainy days is more and with 95% confidence we can conclude that they are not part of the same population.

```
Mean ridership on rainy days: 1105.4463767458733

Mean ridership on non-rainy days: 1090.278780151855

One-tailed p-value: 0.024999912793489721

Two-tailed p-value: 0.048
```

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, Linear regression model, Statistical test.

The given dataset is just a random collection of ridership data and some additional metrics in a random month of the year. The day/time/station when the snapshot is recorded, is not uniform and hence cannot be compared/correlated. The seasonal variations in ridership are not taken care of with given data. The geographical position of the station isn't factored in.

What would provide better insight is some geographical data about the position of the stations. Also, instead of a month's data alone, it would be interesting to collect last 3-4 years' data and compare similar months/fiscal weeks and weekends. Further, it would be useful if the measurement timings/dimensions are normalized – data recorded at same time each day across the years.

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