P1: Short Questions to Analyzing the NYC Subway Dataset

Note. Throughout this project, I used the improved dataset to answer the short questions.

Section 1. Statistical Test

1.1

- **Which statistical test did you use to analyse the NYC subway data?**

  Mann-Whitney U-test

- **Did you use a one-tail or a two-tail p value?**

  a two-tail p value

- **What is the null hypothesis?**

  The distributions of the number of entries for both rainy and non-rainy days are the same.

- **What is your p_{critical} value?**

  $p_{\text{critical}} = 0.05$

1.2

- **Why is this statistical test applicable to the dataset?**

  From the histograms of hourly entries (shown later in Figure 2), we see that the distributions are not normal. Mann-Whitney U-test is still applicable since it does not assume that the datasets follow any particular distribution.

1.3

- **What results did you get from this statistical test?**

  one-tail p value $= 2.74 \times 10^{-6}$
  two-tail p value $= 2 \times$ one-tail p value $= 5.48 \times 10^{-6}$
  the mean of entries with rain $= 2028.2$
  the mean of entries without rain $= 1845.5$

1.4

- **What is the significance and interpretation of these results?**

  The result shows that it is statistically significant, which means that the distributions of the number of entries for rainy and non-rainy days are statistically different because we can reject the null hypothesis as the two-tail p value $< p_{\text{critical}}$.

Section 2. Linear Regression

2.1

- **What approach did you use to compute the coefficients θ and produce prediction for $\text{ENTRIESn}_{\text{hourly}}$ in your regression model?**

  OLS using Statsmodels

2.2

- **What features (input variables) did you use in your model?**

  "meantempi", "hour", "UNIT", "day_week", and "rain".

- **Did you use any dummy variables as part of your features?**

  Yes. I used "UNIT", "day_week" and "rain" as dummy variables.

2.3

- **Why did you select these features in your model?**

  -I used "meantempi" because I thought when it is cold people might to use subway more often.
  -I used "hour" based on an assumption that very few people use subway before dawn while the number of riders increases in the morning and after the evening.
  -I used "UNIT" and "day_week" because by including them in my model, R^2 value drastically improved, and it seems also reasonable that the number of riders strongly depends on the locations of turnstiles and on day of the week.
  -I used "rain" based on an assumption that when it is raining, people use subway more often.

2.4

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

  -15.53 for "meantempi" and 123.4 for "hour"

2.5

- **What is your model's R^2 (coefficients of determination) value?**

  $R^2 = 0.486$

2.6

- **What does this R^2 value mean for the goodness of fit for your regression model?**

  The R^2 value means that the model explains 48.6% of the total variation in the number of entries.

- **Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?**

  Figure 1 shows the residual plot for my model, from which we see the residuals are not distributed randomly but have systematically larger absolute values as the predicted values increase. This indicates that my model is not appropriate for the dataset and there might be some missing variables and/or higher-order terms of variables, or we might need a better model than OLS.

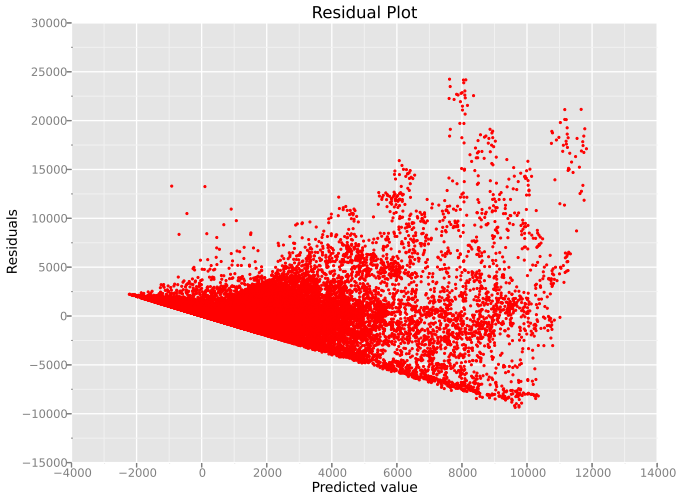

Figure 1: Residual plot for my linear regression model (predicted values vs. residuals).

Section 3. Visualization

3.1

- **Please include two visualizations that show the relationships between two or more variables in the NYC subway data. One visualization should contain two histograms: one of** $\text{ENTRIESn}_{\text{hourly}}$ **for rainy days and one of** $\text{ENTRIESn}_{\text{hourly}}$ **for non-rainy days.**

  The histograms are shown in Figure 2.

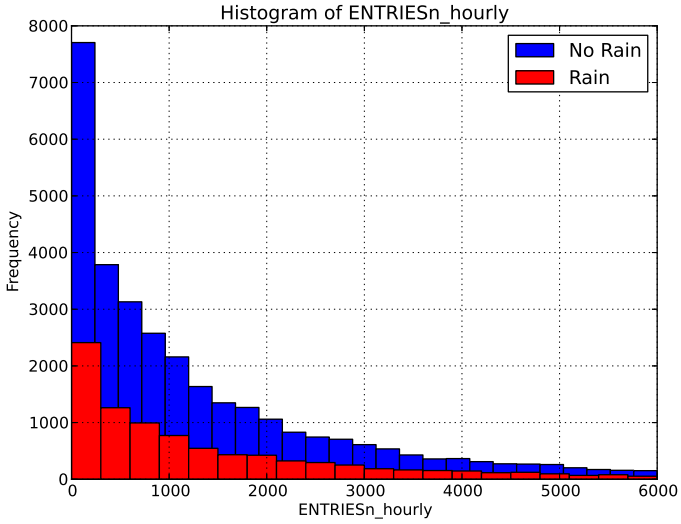

Figure 2: Histograms of ridership for rainy and non-rainy days.

- **One visualization can be more freeform.**

  In Figure 3, a scatter plot is shown for ridership by day-of-week.

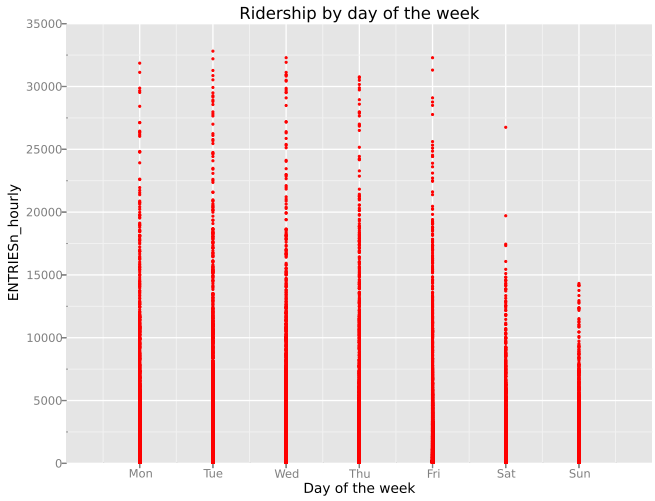

Figure 3: Scatter plot of ridership by day-of-week.

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?**

  More people ride the NYC 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.**

  We can presume that more people give up walking and tend to use subway when it is raining perhaps especially at the central part of the city. As expected, the mean and the median of the number of entries with rain (2028.2 and 939.0 respectively) are both larger than those without rain (1845.5 and 893.0 respectively). When the fact that the one-tail p value (2.74×10^{-6}) is less than p_{critical} (0.05) is taken into account, we can conclude that more people use subway when it is raining.

  The coefficients for the dummy variable "rain" in my linear regression model are 505.3 for non-rain and 456.6 for rain, which indicates more people ride the NYC subway when it is NOT raining. But I would not use this result since as discussed in 2.6, the model is unlikely to be appropriate for the dataset.

Section 5. Reflection

5.1

- **Please discuss potential shortcomings of the methods of your analysis.**

  1. Dataset: There are many variables in the dataset that seem to correlate to each other, such as "precipi", "meanprecipi", "tempi", and "meantempi", which are not very useful. Instead, the following variables might help improve the model:
     -holiday Indicator (0 or 1) if the date is a holiday
     -event Indicator (0 or 1) if there was a special event like a baseball game or a festival near the location at the time.
  2. Analysis: As discussed in 2.6, the linear regression model I used to predict the ridership of NYC subway is not appropriate. The residual plot in Figure 1 is "funnel shaped" with the larger end toward larger predicted values, which suggests that we should transform the values, perhaps by modeling its logarithm or square root, etc[1]. There might be also some missing variables, or higher-order terms, although introducing quadratic or cubic terms of variables did not change the results much for my analysis.

5.2

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

  Figure 4 and 5 are the residual plots with respect to the variables "meantempi" and "hour", respectively. They indicate that my linear regression model tend to predict the ridership systematically smaller than the actual values.

[1] http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

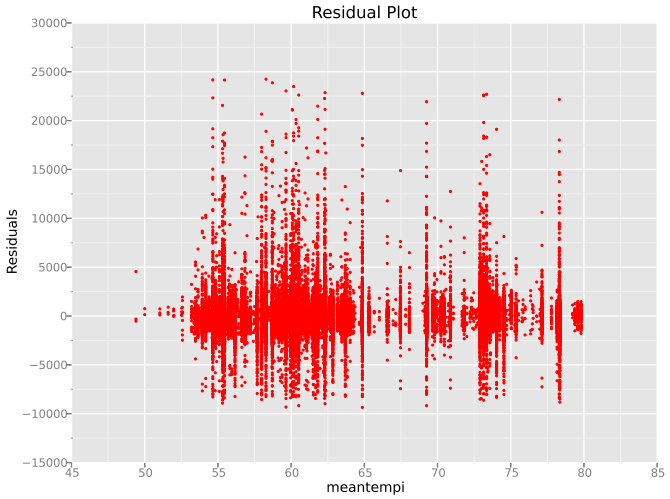

Figure 4: Residual plot (meantempi vs. residuals).

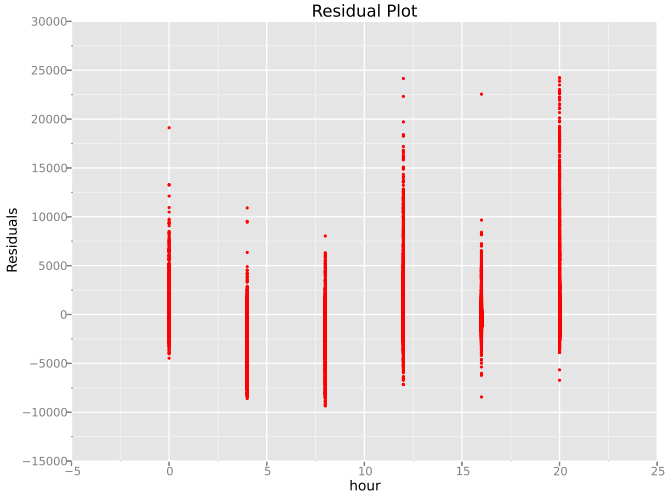

Figure 5: Residual plot (hour vs. residuals).