Project 2
Analyzing the NYC Subway Dataset

Udacity
Data Analyst Nanodegree Program

Marc Wu
udacity@marcwu.de

Submission date: October 18, 2015

Section 1

Statistical Test

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?**

   We want to investigate if there is a significant difference in NYC subway ridership on rainy days compared to non-rainy days. For that we perform the Mann-Whitney U test with a two-tail P value. Ridership is measured by the ENTRIESn_hourly variable, thus our null hypothesis is: There is no difference in the population distribution of ENTRIESn_hourly with regards to rainy and non-rainy days. We choose a p-critical value of 0.05.

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.**

   We examined the distribution of the ENTRIESn_hourly variable when raining vs. when not raining and concluded that both are not normally distributed. Therefore, we cannot perform Welch's T test, because it requires normally distributed populations. Instead, we perform the Mann-Whitney U test, because it is a non-parametric test, i.e. it has no requirements regarding the probability distributions of the populations.

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.**

   The results are as follows:

   mean(rain) = 1105.4463767458733
   mean(no rain) = 1090.278780151855
   U = 1924409167.0
   p-value (one-tailed) = 0.024999912793489721
   p-value (two-tailed) = 0.04999982558697944

   Since scipy.stats.mannwhitneyu() returns the p-value for a one-sided hypothesis, we multiply it by 2 to get the corresponding p-value for our two-tailed hypothesis.

4. **What is the significance and interpretation of these results?**

   The reported two-tailed p-value is just below our p-critical value of 0.05. Thus, we reject our null hypothesis and accept the alternative hypothesis that there is a statistically significant difference between both population distributions. We state that the NYC subway ridership is different on rainy days compared to non-rainy days.

Section 2

Linear Regression

1. **What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:**

   (a) OLS using Statsmodels or Scikit Learn

   (b) Gradient descent using Scikit Learn

   (c) Or something different?

   We use (a) ordinary least squares from the StatsModels library.

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

   The model consists of only two dummy variables: UNIT and Hour.

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 R^2 value."**

   We select the features with a bottom-up approach: Starting with an empty feature set, we add one feature at a time, i.e. the feature which increases the R^2 value the most. If we cannot improve the model's R^2 value noticeable we stop.

   In the first step we add UNIT, followed by Hour in the next iteration. Surprisingly, our model with these two features cannot be improved significantly by adding one of the remaining features. Thus, we end up with a selection of UNIT and Hour for our model. Note, that we choose Hour as a dummy variable, because of its "modulo" nature. The table below summarizes our computations of different feature combinations and their corresponding R^2 values.

UNIT	Hour	meandewpti	meanpressurei	fog	rain	meanwindspdi	meantempi	precipi	R^2
✓									0.4183
✓	✓								0.5008
✓	✓	✓							0.5011
✓	✓		✓						0.5010
✓	✓			✓					0.5010
✓	✓				✓				0.5008
✓	✓					✓			0.5015
✓	✓						✓		0.5013
✓	✓							✓	0.5009

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

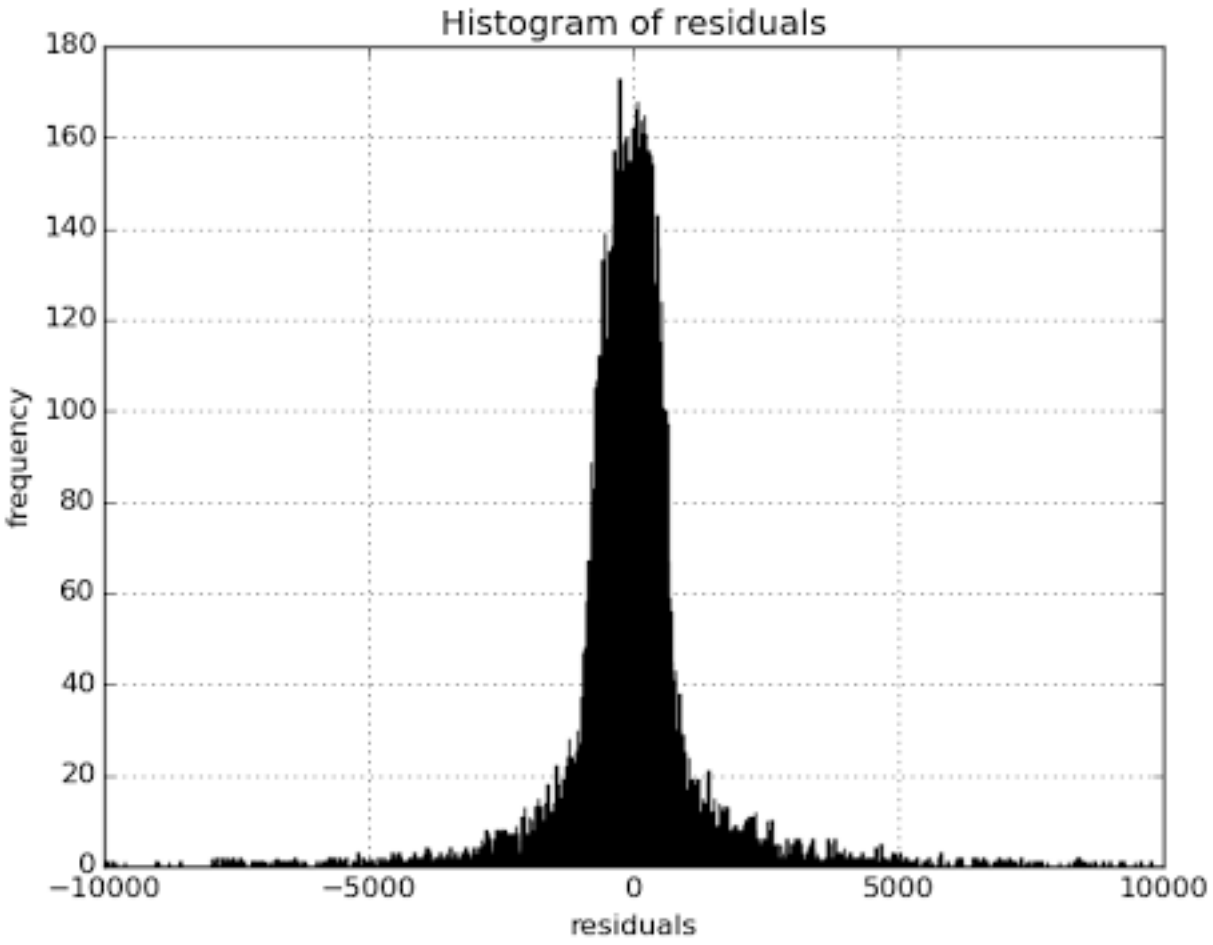
   We do not use any non-dummy features for our model.

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

   The R^2 value of our model is 0.500867953114.

6. **What does this R^2 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 R^2 value?**

   R^2 can be interpreted as how well the model explains the variability within the response variable. Our model explains just 50% of variation, which indicates a poor fit. However, an assessment of the residual plot shows that there is no bias in our predictions. The residuals are normally distributed with a mean around 0.



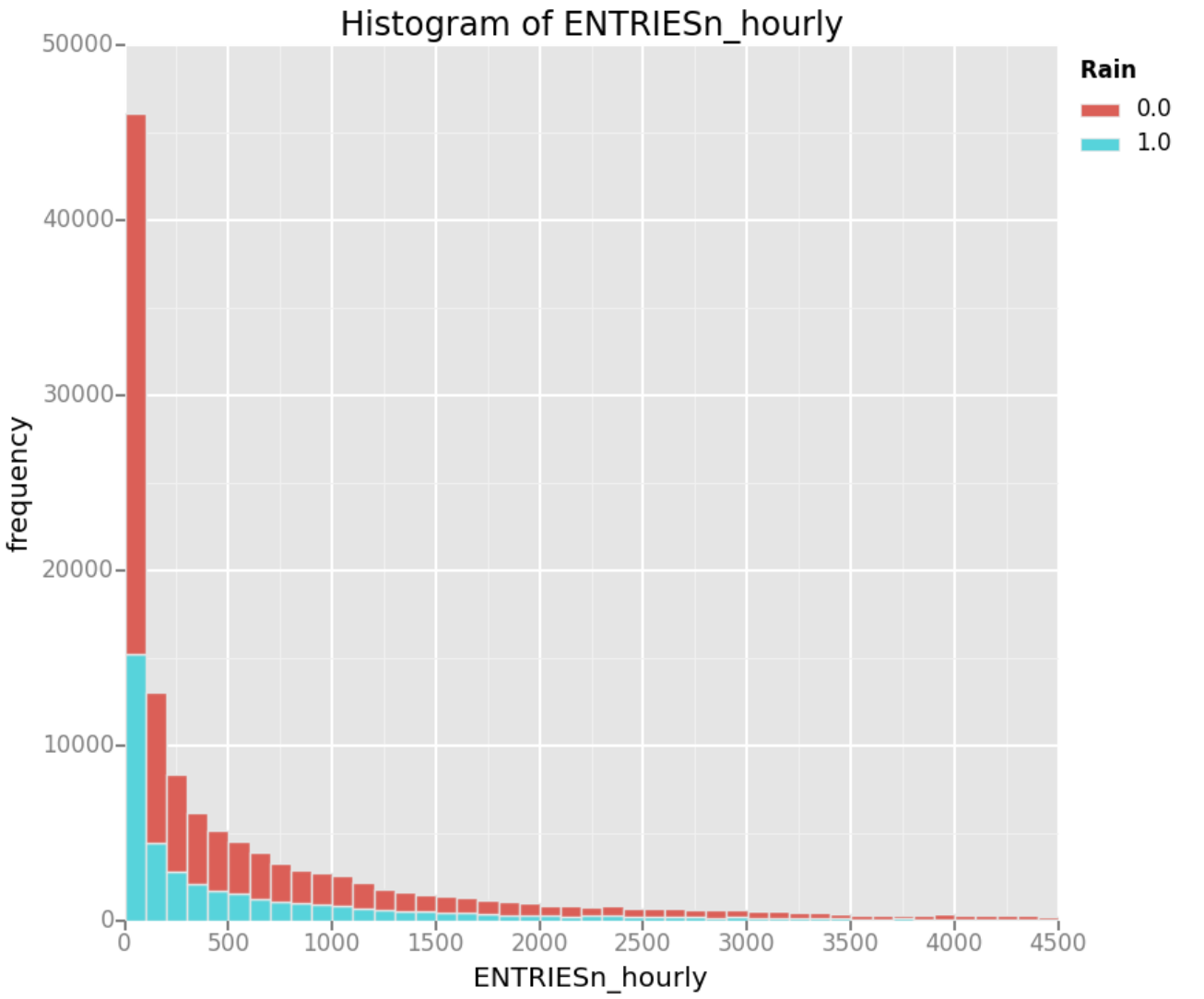
   While our model is not suitable to predict precise ridership because of its low R^2 value, the residual plot demonstrates that it is good enough for rough predictions.

Section 3

Visualization

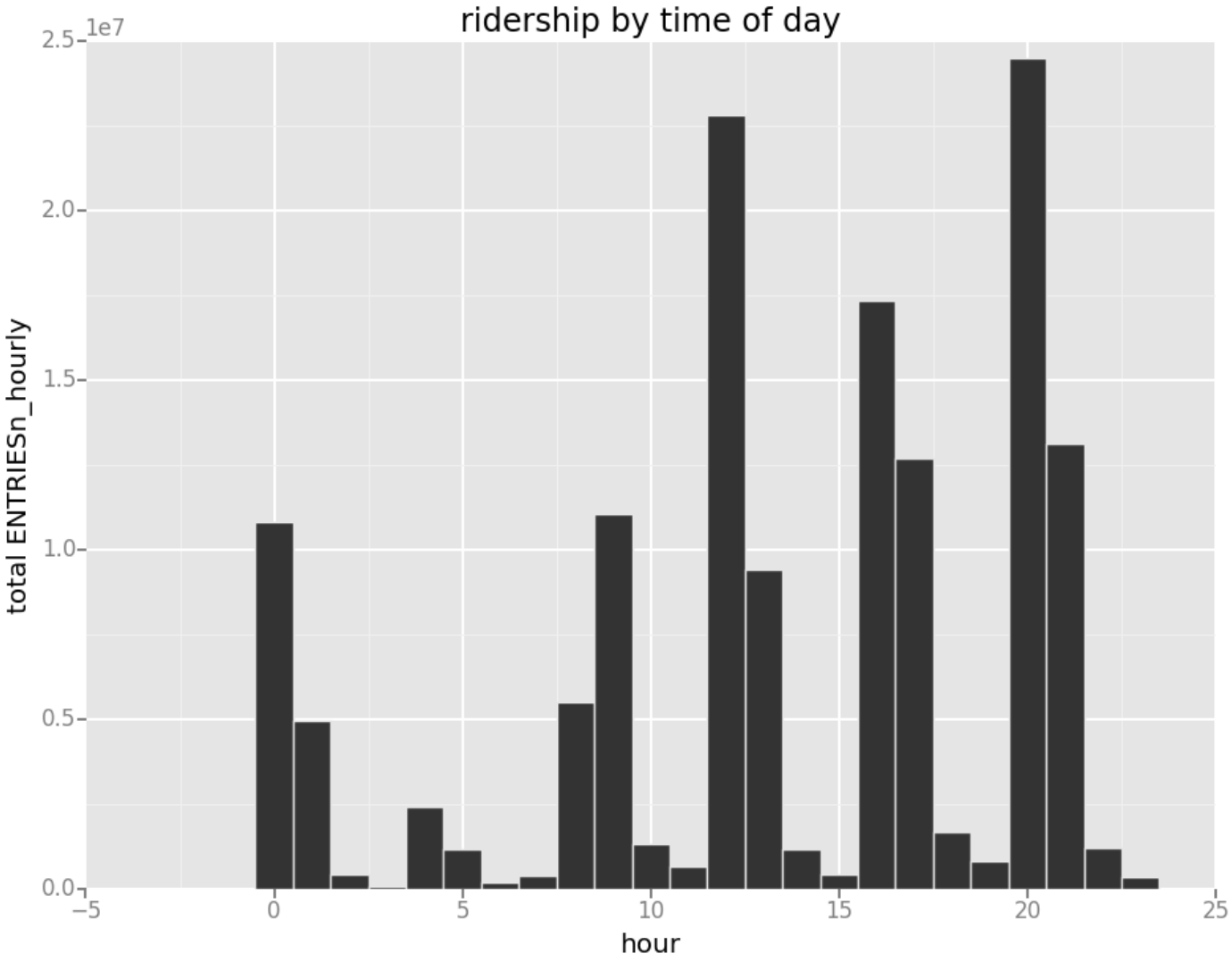
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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



Whether it is on rainy (cyan) or non-rainy (light-red) days, the histogram shows that ENTRIESn_hourly is not normally distributed. Instead, we observe for both cases right-skewed distributions. Moreover, we have much more observations for non-rainy days compared to rainy days across the whole ENTRIESn_hourly range.

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.**



This bar chart shows for each hour the total ridership as measured by the ENTRIESn_hourly variable. Interestingly, ridership's global peak is at 20:00, closely followed by 12:00. Moreover, there is a cycle of roughly four hours where ridership locally peaks.

Section 4

Conclusion

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?**

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 performed the Mann-Whitney U test with a confidence level of 95%. As a result we rejected our null hypothesis and concluded that there is a difference in ridership when raining compared to when it is not raining. When we look at the mean for ENTRIESn_hourly, we see that the mean for rainy days is higher with 1105.44 compared to 1090.27 for non-rainy days. Thus, slightly more people ride the NYC subway when it is raining.

   However, our linear model does not support this claim, because it does not incorporate the rain variable. The rain feature as well as other weather features were not able to improve our model's R^2 value substantially. If we consider again both means from our statistical test, we see that the relative difference between them is very small. This could be an explanation why rain is not a suitable feature, i.e. it is too indistinct to predict ridership.

   Conclusion: There is statistical significant more ridership on rainy days, yet the difference is so small that it is practical irrelevant.

Section 5

Reflection

1. **Please discuss potential shortcomings of the methods of your analysis, including:**

   (a) Dataset,

   (b) Analysis, such as the linear regression model or statistical test.

   (a) The present dataset has several shortcomings. First, we notice that a total sum of 144,532,327 entries are registered in contrast to only 117,026,133 exits. This huge disparity indicates some form of error regarding the counting process. Naturally, everybody who enters the subway should also exit it at some point.

   Moreover, the dataset includes only records from May 2011. Holiday season or other special events have an impact on ridership and can skew the data. A better dataset would include entries from a whole or even several years.

   Another drawback for our regression analysis is the given feature set. For instance, the thunder feature has only entries with 0s and thus it is useless for linear regression. In addition to that, other features are highly correlated such as mintempi, maxtempi, and meantempi. Highly correlated features should not be included together in a linear model.

   (b) The Mann-Whitney U test resulted in a statistical significant difference between ridership on rainy versus non-rainy days. Yet, the difference between both means is so tiny, that the rain feature was

not included in our linear model. This case shows nicely that there is also a practical significance to consider. For tiny differences tests can show statistical significance, even when it is not practical relevant. This tends to happen when the sample size becomes very large.

If for some reason we needed precise ridership predictions, our linear regression model would not be sufficient, because it can predict the ridership only roughly. One explanation could be that ridership does not inherently behave linear to certain features. In this case we would need to do some form of feature engineering or apply a different machine learning algorithm like logistic regression or decision trees.

References

1. https://github.com/jdavis/latex-homework-template
2. https://en.wikibooks.org/wiki/LaTeX
3. http://ggplot.yhathq.com/docs/geom_histogram.html
4. https://en.wikipedia.org/wiki/Nonparametric_statistics
5. http://stackoverflow.com/questions/3121979/how-to-sort-list-tuple-of-lists-tuples
6. https://discussions.udacity.com/t/hours-vs-entries-histogram-ggplot/33243
7. http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm
8. http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit