Analyzing the NYC Subway Dataset

Overview

This project consists of two parts. In Part 1 of the project, you should have completed the questions in Problem Sets 2, 3, and 4 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.

http://tohtml.com/python/

http://blog.yhathq.com/posts/ggplot-for-python.html

http://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html

http://stackoverflow.com/questions/3674409/numpy-how-to-split-partition-a-dataset-array-into-training-and-test-datasets

https://github.com/paulgb/sklearn-pandas

http://statsmodels.sourceforge.net/stable/examples/notebooks/generated/ols.html

http://www.bertplot.com/visualization/?p=229

http://matplotlib.org/users/legend_guide.html

http://stackoverflow.com/questions/28101623/python-pyplot-histogram-adjusting-bin-width-not-number-of-bins

http://docs.scipy.org/doc/numpy/reference/generated/numpy.logspace.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?

Since the data was not normally distributed as shown using the Shapiro-Wilk-test, the Mann-Whitney-U-test, which as a non-parametric test does not assume normality of the data, was deployed to test, whether there is a difference in passenger numbers between rainy and not rainy days. A two-tailed p-value was used, since in the question was whether there was a difference not regarding whether it was smaller or higher. The null hypothesis was P(*with_rain > without_rain) = 0.5*. Thus a p-value higher than the significance threshold would indicate, that if we randomly draw samples of both populations, we would see no difference in the distribution. The critical p-value was set to 0.05, which is commonly used. The test resulted in a p-value of 0.038, thus there seems to be a significant difference.

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.

As stated in 1.1 the data is not normally distributed as determined by the Shapiro-Wilk-test. Thus a parametric test like a t-Test cannot be applied, since it assumes normality. The Mann-Whitney-U-test however does not assume normality and can be applied in this case.

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.

The p-values of the Shapiro-Wilk-test for both subsets are 0. The means are 1844.1994 and 1997.5024 for the without_rain and with_rain subset respectively. The p-value of the Mann-Whitney-U-test is 0.013, thus under 0.05. In conclusion the mean passenger number is significantly higher on rainy hours than on hours without rain.

1.4 What is the significance and interpretation of these results?

The passenger number in rainy hours is significantly different to non-rainy hours. Looking at the means the mean passenger number is higher in rainy hours. Thus the data can be interpreted such that when it is raining more people are taking the subway in New York City.

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:

OLS using the statsmodel package was used as an approach here.

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

All categorical variables were transformed into dummy variables. But in the final model only the variables 'UNIT' and 'hour' were used as predictors in the linear regression.

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.

At first variables that are redundant time variables were removed ('datetime','TIMEn','DATEn'), since they do not add additional information. Also the variables EXITSn, EXITSn_hourly and ENTRIESn were removed, because they are confounders of ENTRIESn_hourly and thus naturally have a high covariance, interfering with model building. Next a model was build using all variables. It resulted in an R^2-value of 0.49. Next variables clearly being redundant and thus have a high covariance were removed (e.g. tempi, since meantempi exists as well). This resulted in an R^2-value of 0.48. Next one variable after another was removed, checking how big the influence on the R^2-value of the resulting model is every time. Only the variables 'UNIT' and 'hour' resulted in a significant drop of R^2 and therefore were kept in the model.

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

The parameter for the 'hour'-variable is 123.04. The intercept is 602.75.

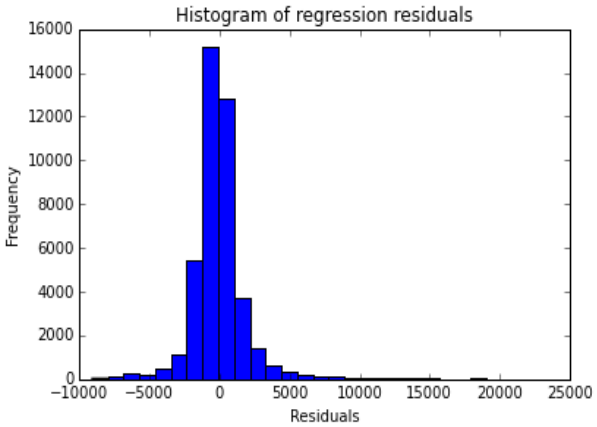
2.5 What is your model's R2 (coefficients of determination) value?

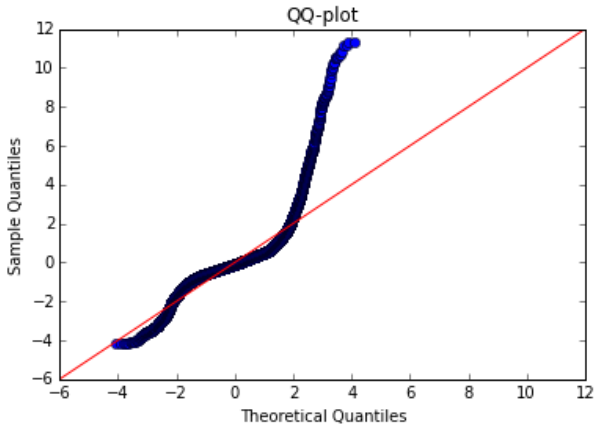
The R^2-value of the final model is 0.458.

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 R^2-value is a measure of the goodness of a regression and is defined by $1 - \frac{Var(Residuals)}{Var(Y)} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$. It can take values between 0 and 1, whereas R^2 is zero if X and Y are unrelated. A R^2-value of 1 appears, when X and Y are perfectly related. The R^2-value of 0.46 that was

obtained in this model indicates that the model is able to predict 46 % of the original variability. Generally a model should explain the variability of the model as good as possible without overfitting. This can be limited by the noise of the data. The R^2-value obtained in the model created for this report is relatively low even when accounting for noisy data. In general R^2 is not a good indicator, whether an appropriate model was used. Thus one can for example get a fairly high R^2-value when using a linear model for a steep exponential increase, since a lot of data points could fall in the approximately linear part of the exponential function. But rather low R^2-values indicate that the regression is not fitting the data well, of which a common reason would be that the model type was ill-chosen. The histogram of the residuals shows long tails, indicating that some values are badly represented by the model. These values could also be caused by outliers in the data, but their relatively high number rather suggests an ill-suited model. This is further confirmed by the QQ-plot that shows, that the residuals are not normally distributed. This usually suggests that the wrong type of model was chosen. Thus a linear model might be rather inappropriate for the given data.

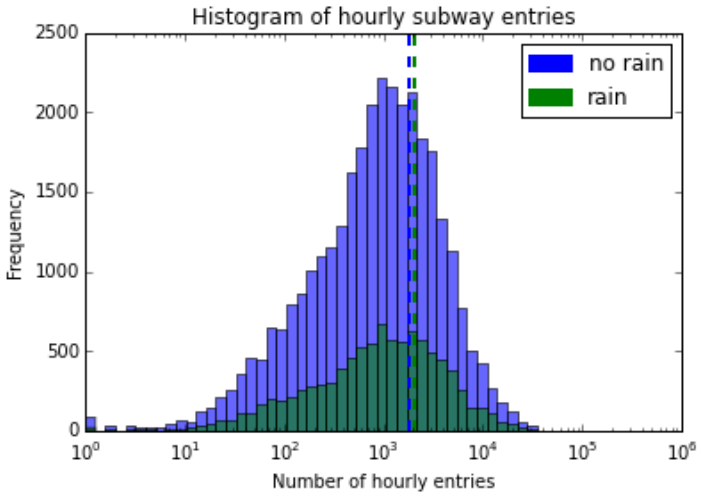




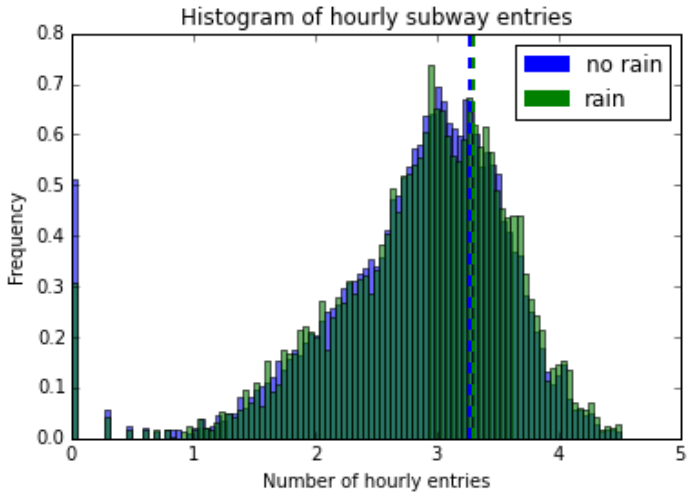
Section 3. Visualization

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

The histogram looks like this:



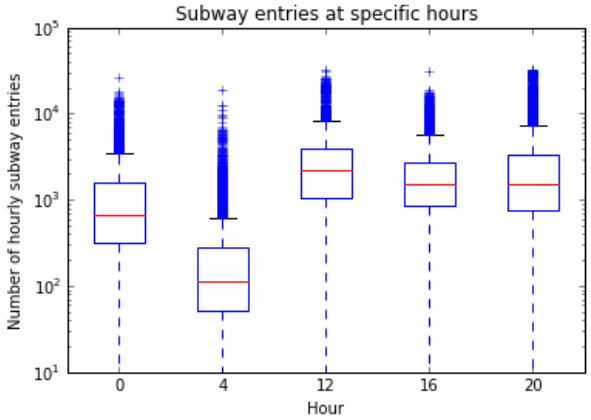
Looking at this plot it does seem like as if there are not a lot more people taking the subway on non-rainy days, since the means are quite close together. The distributions also look quite similar. The generally higher number of hours without rain leads to the higher peak of the distribution, compared to the distribution of rainy days. This can be adjusted for by normalizing the histograms to match the area under the density curve to be 1:



In the normalized histogram it is better visible that the distributions are very similar.

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.

The following graph depicts a bar plot visualizing the subway entries in the given hours:



One can clearly see, that the ridership decreases during the night, but barely differs during the day.

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?

The analyses performed in this study suggest that there may be a minor difference in terms of how many people ride the subway in rainy hours compared to non-rainy hours.

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.

The Mann-Whitney-U-test comparing the means of the ridership-size in rainy and non-rainy hours resulted in a p-value of 0.038, which is lower than the generally accepted significance threshold of 0.05. Thus this result implies that there is a difference in the ridership between rainy and non-rainy hours.

The histograms depicting the subway entries of both populations look rather similar. There seems to be a slight trend of higher frequencies of larger entry numbers in rainy hours, but it seems rather marginal. Thus the differences may be rather small.

On the other side the fact whether it rains or not does not significantly influence the OLS-model, on the first glance hints that rain does not influence the ridership. But since the model itself has a relatively low R^2-value, it is probable that a linear model is not well suited to represent the given data, which is further supported by the analysis of the residuals. Thus the regression should probably not be considered in the answer of the question underlying this study.

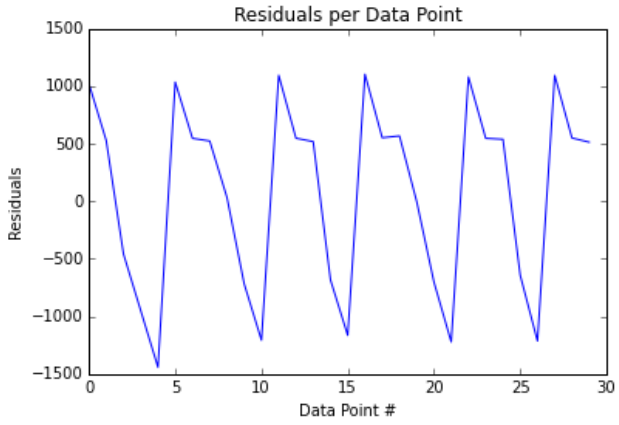
Section 5. Reflection

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.*

The given dataset contains a lot of variables that are redundant as for example the mean temperature, minimal temperature and maximum temperature. Those variables are probably highly correlated and may lead to collinearity, which may lead to false results in some models. The data were taken in one month, which is May thus in late spring, where the weather conditions are generally quite good. It would certainly be interesting to compare this data to data obtained in autumn or winter.

As discussed earlier linear regression is most likely not suited to represent the data, since some data points cannot be explained by the model. Although the variable 'hour' is included into the model, looking at the residuals, there seems to be a dependence of the residuals on the hour of the day, which looks like an oscillatory behavior, indicating non-linearity in at least this aspect.



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