Analysing the NYC Subway dataset

Udacity Data Analyst Nanodegree – Project 1

Zhanzhan He

March 2015

Contents

0. References

duckworthd. (2012, July). Retrieved from pandas: complex filter on rows of DataFrame: http://stackoverflow.com/questions/11418192/pandas-complex-filter-on-rows-of-dataframe

flow. (2011). *Plot 2 histograms with matplotlib*. Retrieved from Stack overflow: http://stackoverflow.com/questions/6871201/plot-two-histograms-at-the-same-time-with-matplotlib

matplotlib. (2015). *Matplotlib Text introduction*. Retrieved from http://matplotlib.org/users/text_intro.html

NIST. (n.d.). Retrieved from http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

wikipedia. (2015). *Nonparametric statistics*. Retrieved from http://en.wikipedia.org/wiki/Nonparametric_statistics

wikipedia. (2015). *Simpson's paradox*. Retrieved from http://en.wikipedia.org/wiki/Simpson%27s_paradox

YHat. (2013). *ggplot for python*. Retrieved from http://blog.yhathq.com/posts/ggplot-for-python.html

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?

A one-tailed Mann-Whitney U-Test was applied to the data.

μ_{rain}: Mean number of turnstile entries on rainy days

μ_{dry}: Mean number of turnstile entries on dry days

H0: $\mu_{rain} = \mu_{dry}$

H1: $\mu_{rain} > \mu_{dry}$

p-critical: 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.

The data appears non-normal from the histograms of turnstile entries. The distribution as plotted looks like an exponential distribution (FIGURE 2).

Welch's two sample t-test could be appropriate as the sample size is large, so by the central limit theorem the sampling distribution of the mean entries will be approximately normal. The Mann-Whitney U-test is used instead as it makes no assumptions about the distribution of the data.

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.

$\bar{x}_{rain} = 1105.4$
$\bar{x}_{dry} = 1090.3$

U= 1924409167

p = 0.025

1.4 What is the significance and interpretation of these results?

We can reject H0 at the 5% level as $p < 0.05$. There is a statistically significant difference between the mean number of turnstile entries on rainy days and wet days.

From the calculated sample means we also get $\bar{x}_{rain} > \bar{x}_{dry}$. So we conclude from the data that on average more people ride the subway on rainy days than on dry days.

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

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

- Hour
- meantempi
- precipi
- meanwindspdi
- fog
- UNIT – was transformed into a set of dummy variables, one variable for each unit with value 1 if the data corresponds to that unit and 0 otherwise

2.3 Why did you select these features in your model?

From the visualizations in Section 3.2, we can see how ridership varies over time of day (Figure 3) and we also observe significantly different usage patterns across the day at different turnstiles (FIGURE 4, FIGURE 5). Time of day is accounted for by using hour as an input. The UNIT dummy variable (already set up in the sample code) accounts for the different usage patterns at different turnstiles. These two features increased the R^2 of the model by the greatest amount.

The weather variables meantempi, precipi and meanwindspdi and fog were included as intuitively, poorer weather like fog, fast winds, low temperatures and rain would mean people would prefer to ride the subway than walk or cycle outdoors. We showed in Section 1 that there is a statistically significant difference between the number of turnstile entries on wet days and dry days. Adding each of these weather variables made a relatively small contribution (~1%) to R^2. precipi covers the amount of rain at a particular time and location, so it was chosen over the less specific rain variable which only indicates whether it rained on that calendar day.

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

Hour	468.4
meantempi	-52.7
precipi	-29.1
meanwindspdi	64.5
fog	70.6

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

0.465

2.6 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?

An R^2 of 0.465 means that our model explains approximately 46.5% of the total variability in the number of entries. This seems reasonable given the large number of variables that could affect ridership (public holidays, seasonal cycles, extraordinary events, etc.) which we haven't accounted for in the model.

However the influence of weather seems disproportionately small – perhaps the relationships between weather and ridership are non-linear and such a model would be more suitable. In fact, the weight of precipi is negative, which means our model will predict fewer riders when there is more rain which is at odds with our result from Section 1. When we plot the histogram of residuals (Figure 1), we see the distribution has a very long right tail, which is another indication that a linear model is not necessarily appropriate.

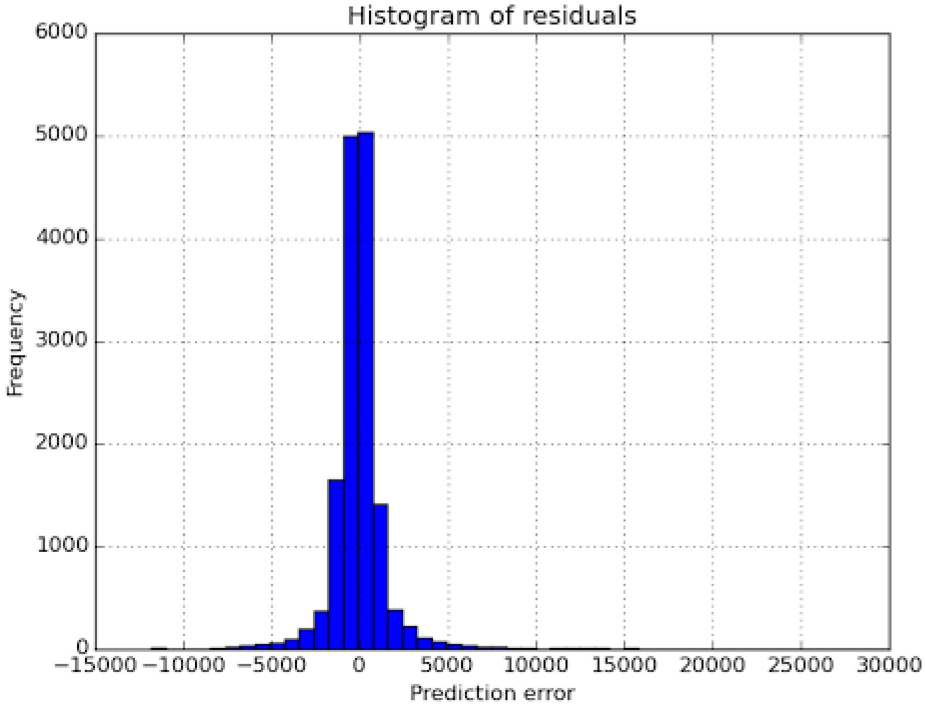

Figure 1 - histogram of regression residuals

3. Visualization

3.1 Histogram of ENTRIESn_hourly for rainy/non-rainy days

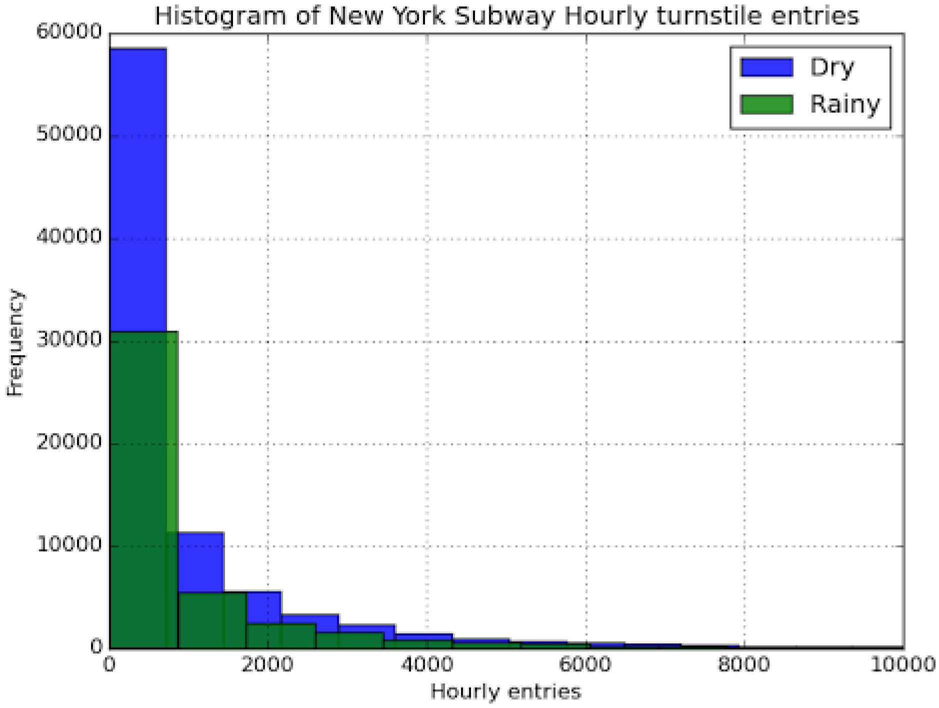

Figure 2 - Frequency of hourly entries on the NYC subway for rainy and dry days. x-axis truncated at 10000 to exclude outliers and give a clearer picture of the shape

Insights

- The distribution of entries is the same shape for both rainy and dry days
- The distribution of turnstile entries does not have the bell shape characteristic of normal distributions. It looks more like an exponential distribution

3.2 Turnstile entries by hour of day

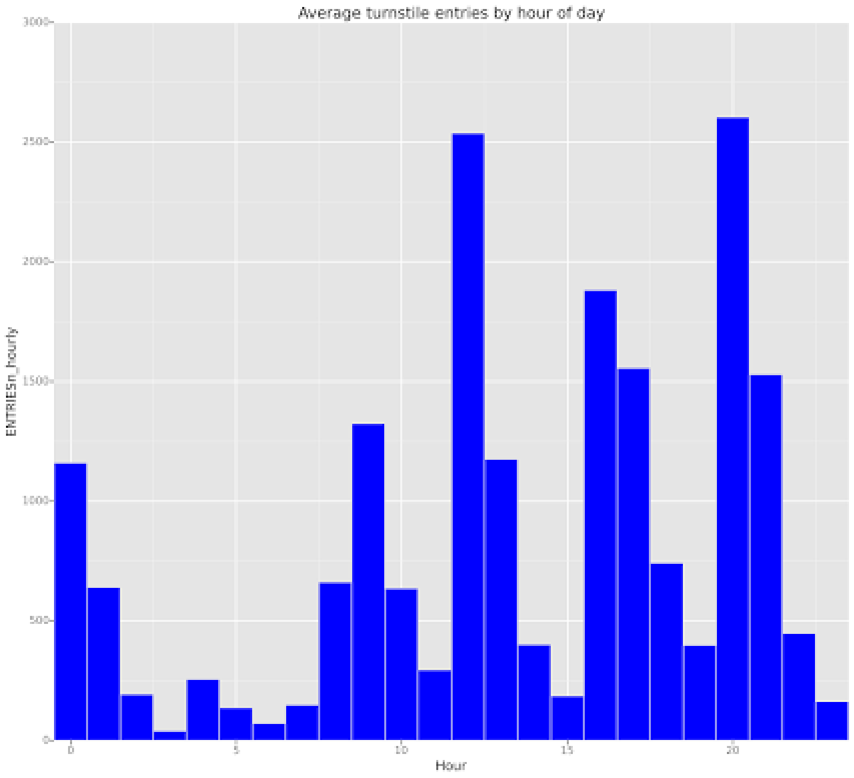

Figure 3 - Average turnstile entries by hour of day calculated over all the data

Intuitively, we expect turnstile entries to peak in the morning and evening rush hours. However, when the mean turnstile entries were calculated over all the data (FIGURE 3), we get a diagram with multiple peaks. These may correspond to the influence of other variables, such as different usage patterns between weekdays and weekends, as well as differences in usage between different turnstiles at different locations.

The data was separated into weekdays and weekend data and a plot made for each turnstile. When we do this a more discernible pattern emerges. We can see a distinct peak at around 9am, 7pm or both for the weekday turnstiles plotted in FIGURE 4.

For the same turnstiles at weekends (FIGURE 5), the average turnstile entries has much less pronounced peaks, looking closer to uniform which is expected due to far fewer people commuting to work.

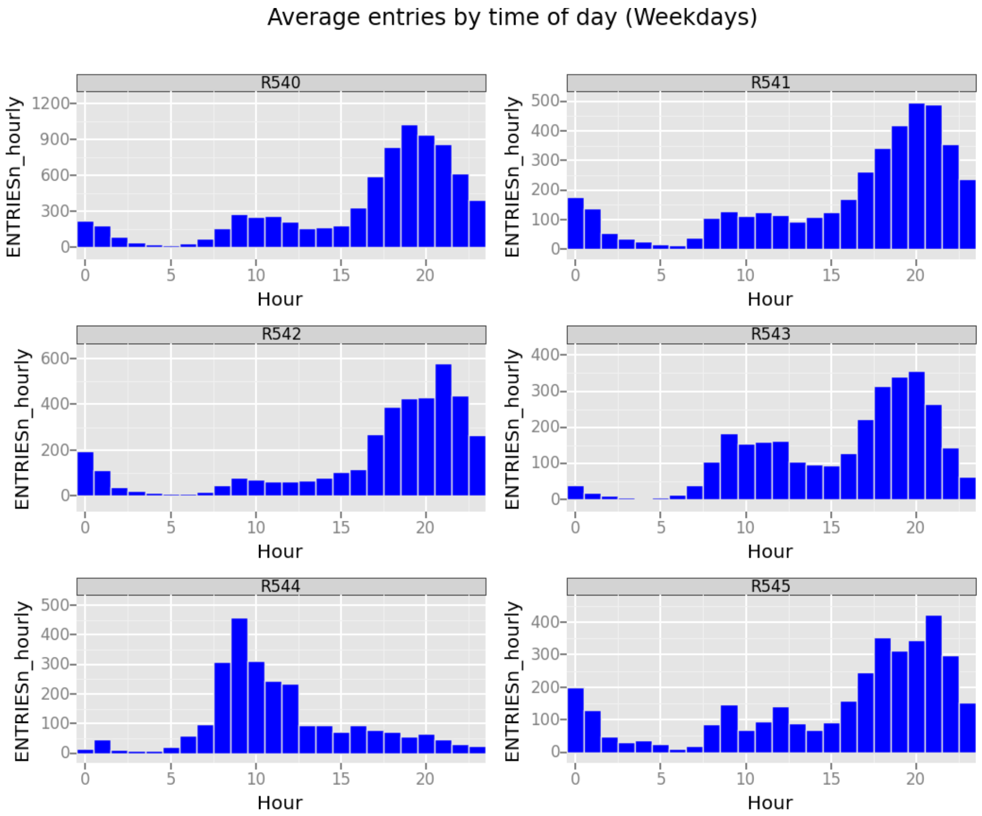

Figure 4 - Weekday average entries for a small sample of turnstiles

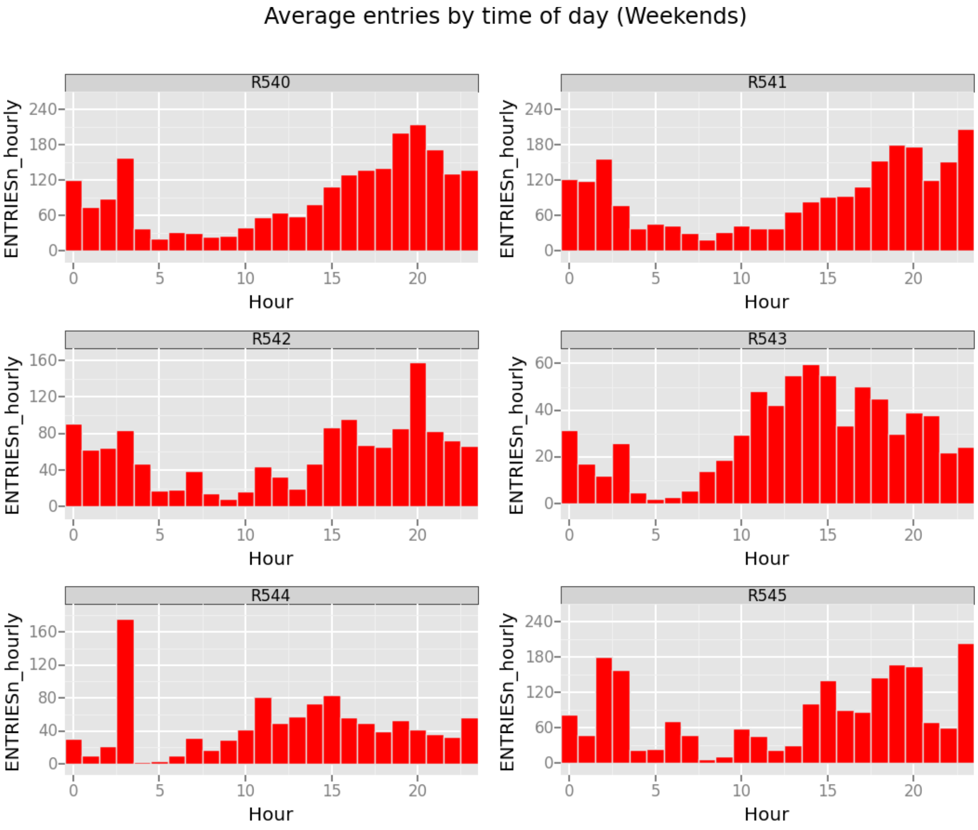

Figure 5 - Weekend average entries for a small sample of turnstiles

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 subway when it is raining. The difference is relatively small – on average ~1.4% more people ride the subway when it is raining, but is statistically significant.

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.

A one tailed Mann-Whitney U test was performed on the data, split into rainy days and dry days. It was found there was a statistically significant difference at the 5% level (though not at the 1% level). The sample mean number of riders was greater on rainy days than dry days.

Taking the amount of rain into account improved the quality of the prediction by a small amount in our linear regression model, which we wouldn't expect if there was no difference between rainy days and dry days.

5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis

5.1.1 Dataset

For the regression, the online code editor was used which runs the regression on a random subset of the data due to resource limitations. Running the regression offline with the full dataset would have given better results.

Some turnstiles appear to have much less associated data than others. The sample of turnstiles plotted in FIGURE 6 show average entries of zero for many hours of the day. The usage pattern is not plausible as the hours with zero entries seem to be randomly interspersed with hours with many entries. It is more likely data is missing than this being a genuine turnstile entry pattern. This will reduce the quality of the regression predictions, which has the turnstile unit as an input variable.

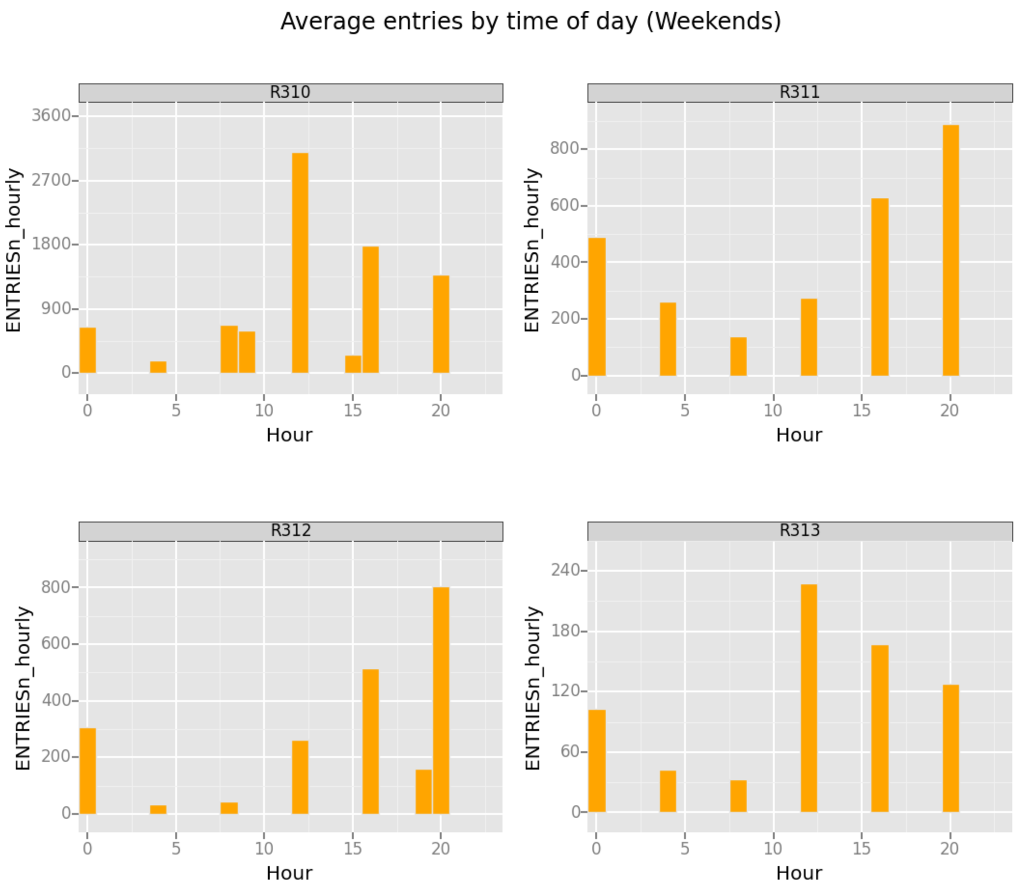

Figure 6 - sample stations with missing data for hourly entry analysis

5.1.2 Analysis, such as the linear regression model or statistical test.

We did a visual check for whether the turnstile entry data was normally distributed by plotting a histogram. A more principled way to check this would be to use a Shapiro-Wilk test.

Gradient descent was used to find the parameters for our regression model. However gradient descent isn’t guaranteed to find a global optimum. It would be valuable to do more experimentation with increasing the number of iterations and changing the learning rate. We should have done multiple iterations of the gradient descent, starting with different random values of the parameters to increase the chance of converging on a global optimum.

We have considered R^2 which measures the quality of the model fit, but have not considered whether the model will generalize well. It could be that the model fits our specific dataset but will not generalize in the real world (overfitting). To check this, we could have split up the data into training and test data and evaluated R^2 only on the test data.

Simpson’s paradox is a potential problem in our dataset. We observe very different usage patterns at different turnstiles, but the linear regression is calculated over the data for all the turnstiles. It is possible the trend will disappear or reverse when the regression is performed over the data for each turnstile separately (wikipedia, 2015).