

Analyzing the NYC Subway Dataset

Questions

Overview

This project consists of two parts. In Part 1 of the project, you should have completed the questions in Problem Sets 2, 3, 4, and 5 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.

Section 1: Statistical Test

I used this book as a general reference on statistical methodology:

Herbert Basler: Grundbegriffe der Wahrscheinlichkeitsrechnung und Statistischen Methodenlehre, Physica Verlag, 11. Auflage

Mainly Chapter 3, especially 3.4.2 (t-test) and 3.4.5 (Mann Whitney U-test)

On z-transformation:

<http://de.wikipedia.org/wiki/Studentisierung>

why chose a significance level of 0.05:

http://www.p-value.info/2013/01/whats-significance-of-005-significance_6.html

http://www.radford.edu/~jaspelme/611/Spring-2007/Cowles-n-Davis_Am-Psyc_origins-of-05-level.pdf

Interpretation of p-value:

<http://stats.stackexchange.com/questions/46856/interpretation-of-p-value-in-hypothesis-testing/46858#46858>

Section 2: Linear Regression

Plotting residuals

<http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm>

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

http://matplotlib.org/api/pyplot_api.html#matplotlib.pyplot.hist

computing R^2

<http://docs.scipy.org/doc/numpy/reference/generated/numpy.mean.html>

<http://docs.scipy.org/doc/numpy/reference/generated/numpy.sum.html>

http://en.wikipedia.org/wiki/Gradient_descent

http://en.wikipedia.org/wiki/Coefficient_of_determination

Correlation matrix to find out the best candidates for linear regression:

<http://pythonprogramming.net/pandas-statistics-correlation-tables-how-to/>

<http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.corr.html>

Interpretation of R^2

<http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit>

<http://people.duke.edu/~rnau/rsquared.htm>

Section 3: Visualization

Introduction and overview with good examples:

<http://opr.princeton.edu/workshops/201501/ggplot2Jan2014DawnKoffman.pdf>

overview on existing geoms:

<http://docs.ggplot2.org/current/>

<http://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf>

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?

I used the Mann Whitney U-test to find out if there is a significant difference between the number of riders with rain versus the number of riders without rain.

The Mann Whitney U-test assumes that two input variables x and y follow the same distribution with the only difference that there is a lag between them named a : $F_y(x) = F_x(x-a)$. The null-hypothesis states that $a = 0$, i.e. there is no difference in the variables.

Transferred to our question here: the null-hypothesis says that there is no difference between the number of riders with rain and without rain.

I used the two-tail p-value, because the question in test aims at differences in both directions, there could as well be more riders or less riders when it rains.

I choose a significance value of 0.05. There is no deterministic reason to pick exactly 0.05 but this has established itself as an agreed level of significance following the idea that an event which occurs 5% of the time or less is rare enough to assume that there is a different cause but mere chance behind it.

- 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.

Welch's T-Test postulates that the data set follows a normal distribution. But the histogram has clearly shown that this is not the case. Therefore Welch's T-Test is not applicable on this set.

However, the histogram has shown that the Entries at Rain and the Entries at no Rain follow a very similar distribution. This makes it a candidate for the Mann Whitney U-test (see assumption above in 1.1., second paragraph).

It has to be considered that the number of without rain occurrences is much higher than the number of with rain occurrences. Therefore the test to choose has to be robust against different sample size. This is true for the Mann Whitney U-test. This is why I picked this test for this data set.

- 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 mean values between with rain and without rain differ by $1105.4 - 1090.3 = 15.1$, which means that the average number of riders with rain is about 15.1 higher than the number of riders without rain.

The question is: can this difference be explained by chance or is it a significant difference? This question shall be answered by a statistical test, in this case the Mann-Whitney U-Test.

Executing this test produces the result of $p = 0.024999912793489721$, which is the value for the one sided test.

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

The result of our test shows a p-value of 0.024999912793489721 , which must be doubled for the two sided test and leads to a value of 0.04999982558697944 . Compared with our alpha-Value of 0.05 we see that $p < \alpha$.

Based on our significance level of 0.05 we have to reject the null-hypothesis that there is no difference between the number of riders with rain and without rain. Thinking of 'Fisher's disjunction': Either a rare event has happened or the null hypothesis is false. Following our approach we think that the event is rare enough to reject the null hypothesis.

The result is that significantly more people use the Subway when it rains.

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:

1. Gradient descent (as implemented in exercise 3.5)
2. OLS using Statsmodels
3. Or something different?

I used the the Gradient descent approach.

- 2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features? I used the features 'Hour', 'meanwindspd' and 'mintempi'. The dummy variable "UNIT" was also added.

- 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 R^2 value."

I calculated the correlation coefficient for each variable in relation to the variable `ENTRIESn_hourly`:

```
(print(turnstile_weather.corr()))
```

The result:

	correlation coefficient to ENTRIESn_hourly
Hour	0.175430
ENTRIESn_hourly	1.000.000
EXITSn_hourly	0.744316
maxpressurei	-0.017084
maxdewpti	-0.009893
mindewpti	-0.020135
minpressurei	-0.020517
meandewpti	-0.016198
meanpressurei	-0.016128
fog	0.011368
rain	0.003062
meanwindspdi	0.026627
mintempi	-0.029034
meantempi	-0.022796
maxtempi	-0.014303
precipi	0.009665
thunder	NaN

Taking only the absolute values into account it shows that the two highest values are ENTRIESn_hourly and EXITSn_hourly. This is trivially true and self evident. It proves that the approach produces correct results but does not reveal any insight.

The next three variables are hour, meanwindspdi and mintempi. Before picking this values one should ask if they seem to make sense judging by intuition: Hour makes sense, otherwise there would be no such thing as a rush hour. Mintempi (negatively correlated) makes sense, because people should be likely to prefer the train instead of walking when temperature decreases (at least in New York, there might be a different situation in the desert). What about meanwindspdi? It is positively correlated to the number of entries. The more wind, the more people riding the train? When I think of women being afraid to ruin their hair-do: It makes sense for me.

So the three variables which showed the highest correlation to the entries (without being trivial) made sense according to my intuition. This is why I picked them for my model.

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
-1.57248270e+02, -1.55110529e+02, 1.21866703e+03

2.5 What is your model's R^2 (coefficients of determination) value?
I calculated an R^2 Value of 0.464610132035.

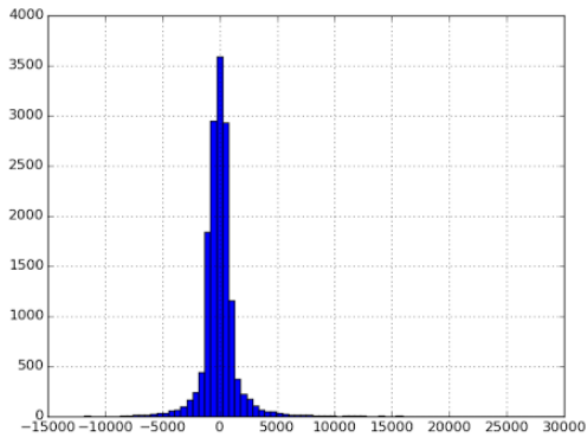
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

R^2 is a measure indicating how well a a statistical model fits to the actual data, where 0 is the worst and 1 is the best possible value. So our value of 0.46 looks like a pretty bad result.

Most authors state that a high R^2 does not necessarily mean that the model is good. But a low R^2 value is never an indicator for a good fitting model, even if in some cases the model might still be helpful to get some insight into the data.

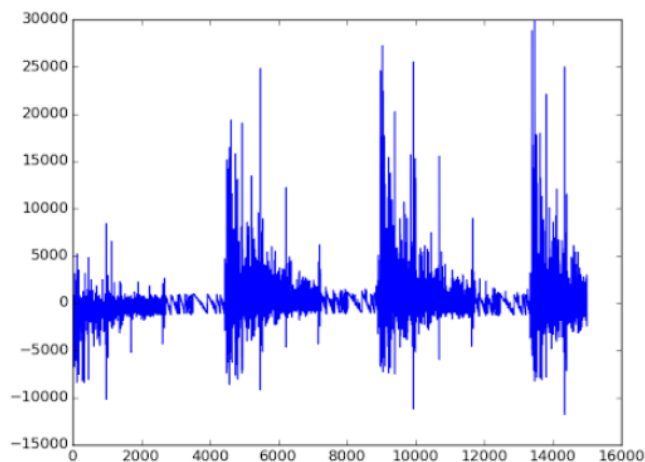
This is the time to have a look at the residuals distribution. The residuals as the difference between calculated and actual value should be distributed randomly, meaning that they follow a normal distribution. The analysis of the residuals tells a lot about how well the model fits to the data.

First let us have a look at the histogram to get an idea of the distribution:



Even if this curve looks symmetric around the 0-Value and to some degree reminds of a bell-curve, it strikes that this curve shows long tails, so our linear model unfortunately seems to produce a number of very large residuals.

So let us have a deeper look at that residuals and see if there is any pattern we can identify. The following plot shows the residuals on y-axis and the values by order of occurrence (which means ordered by unit, date, and time ascending) on the x-axis.



We see very clearly that there is a strong cyclic pattern, showing three big and one smaller peak. With this we have identified a basic relationship in the data which is non-linear. This is a structure which a linear model will not be able to reflect.

We can conclude that a linear model is not the appropriate type to represent the number of riders while a non-linear model would be more promising.

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.

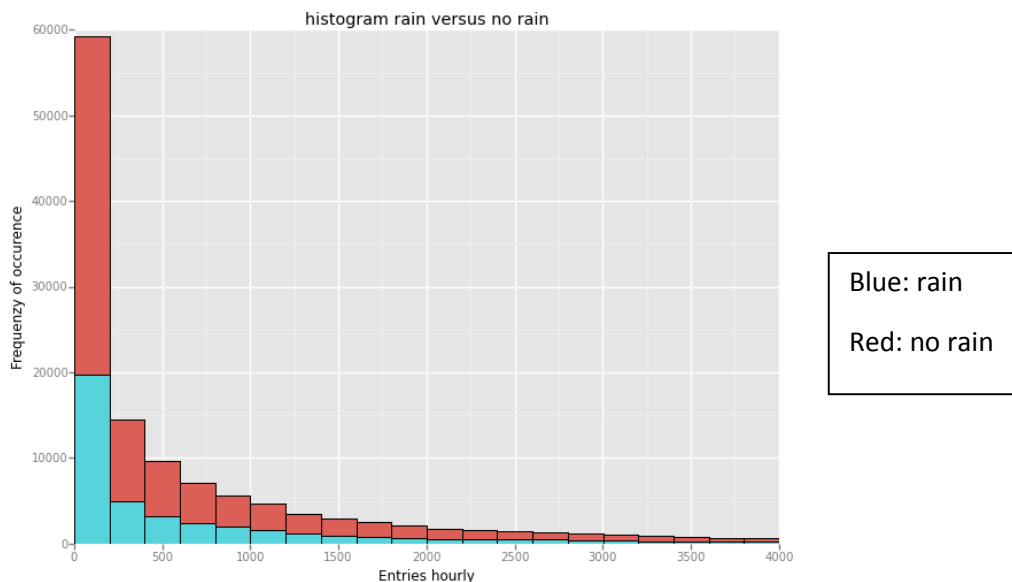
3.1 One visualization should contain two histograms: one of `ENTRIESn_hourly` for rainy days and one of `ENTRIESn_hourly` for non-rainy days.

- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use 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.

My solution:

```
plot = ggplot (aes(x='ENTRIESn_hourly', fill = 'rain'), data = df )
plot + stat_bin(binwidth = 200, color = 'black') \
+ xlim(0, 4000) \
+ ggtitle ("histogram rain versus no rain") \
+ xlab('Entries hourly') \
+ ylab('Frequency of occurrence')
```

.. creates this plot:



(I could not add an automatic legend in the plot, it seems this feature is not fully implemented in Python.)

This plot shows the hourly entries and their frequency of occurrence. It can be seen that the frequency without rain exceeds the frequency with rain for every number of hourly entries, which is not very surprising.

We can further see that the distribution is extremely skewed. The frequency of low values with ≤ 200 entries per hour exceeds the accumulated rest of all entries. This tells us that far most of the overall traffic volume takes place as a „base load“ with a low number of passengers entering the train at a time.

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. Some suggestions are:

- Ridership by time-of-day
- Ridership by day-of-week

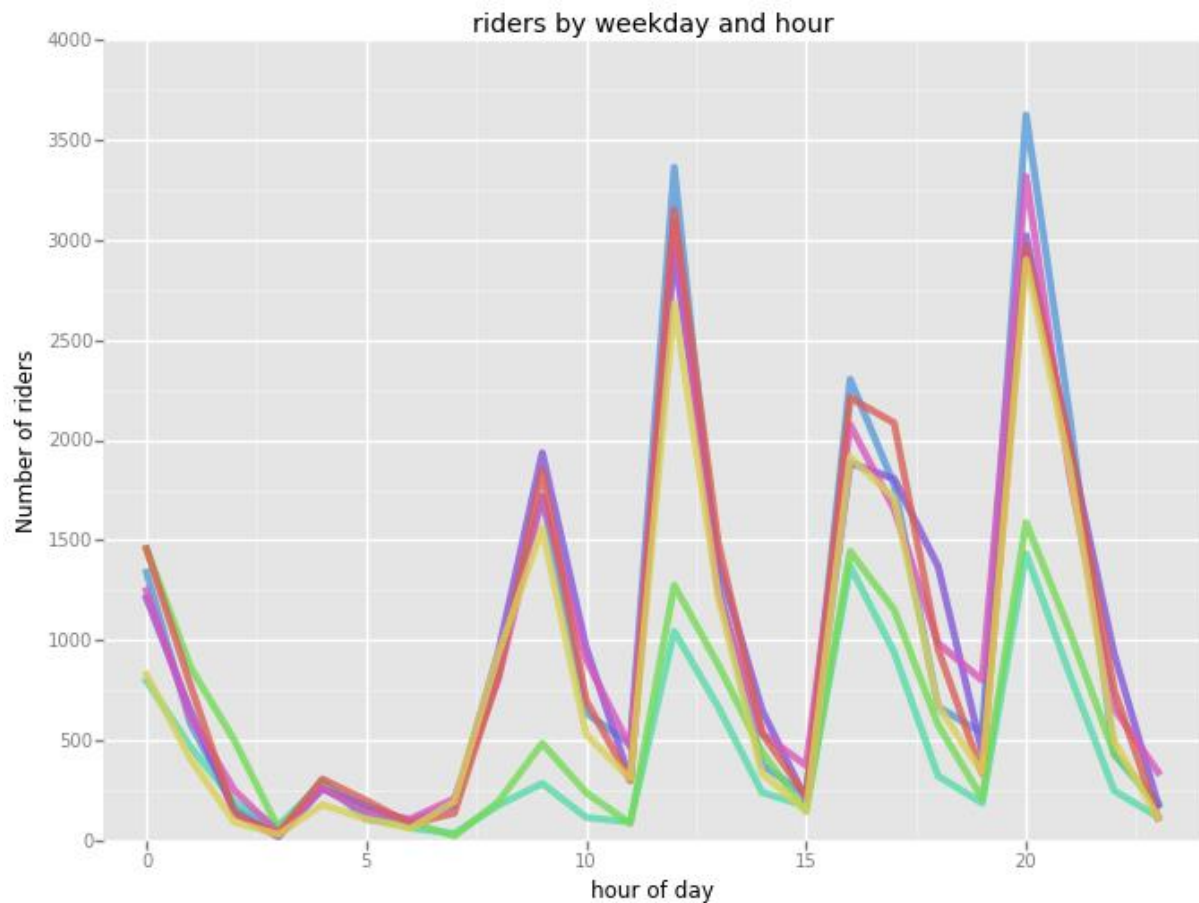
After adding the weekday to the data set:

```
from datetime import datetime
df['Weekday'] = map(lambda dat: datetime.strptime(datetime.strptime(dat, '%Y-%m-%d'), '%A'), df['DATEn'])
```

.. I used the following code:

```
p = ggplot(aes(x='Hour', y='ENTRIESn_hourly', color = 'Weekday'), data=df)
p + stat_summary(geom='line', size = 4, alpha = 0.8) \
+ xlim(-1, 24) \
+ ggtitle ("riders by weekday and hour") \
+ xlab('hour of day') \
+ ylab('Number of riders')
```

.. to generate this plot:



The outlying green lines represent Saturday and Sunday, the other colors stand for the rest of the week.

This visualization allows us to compare the distribution by time of day for the different weekdays. The idea is not to look at the exact values but to get an idea of the patterns in the data.

With a rough visual examination we find that there seem to be two types of curves: weekend versus working days. Again we might ask ourselves: does this result make sense? Of course it does, human behaviour obviously differs a lot between weekend and working days. It confirms our approach to find this expected behaviour also reflected in the curves.

We see that there are four peaks each day around 8h, 12h, 16h and 20 h. The pattern is similar on weekday and weekend while the weekend days have a much smaller amplitude.

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?

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 linear regression I conducted did not provide any valid result. The R^2 value of 0.46 showed clearly that this linear model was not able to explain the number of riders based on linear relationships based on the chosen features.

The analysis of the residuals showed the reason for this. There are strong cyclic processes at work which make a linear analysis inappropriate to reflect what is going on.

The statistical test was much more use in this case. Having detected a higher average number of riders with rain compared to no rain, I used the Mann-Whitney U-Test to find out if this difference is statistically significant. The result was that significantly more people use the train when it rains.

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:

1. Dataset,
2. Analysis, such as the linear regression model or statistical test.

Practical knowledge tells us that people tend to have the same driving behaviour during their workdays. They go to work each day at the same time, have lunch at the same time, go back home at the same time. Friday might be a difference, but by and large this should apply for the working days. At the weekend there will be different behaviour but again it will be quite stable.

This means there are two cyclic patterns at work: day of week and hour of day. This cyclicity impeded the linear approach, which can not be a surprise: The very basic assumption for the use of a linear model is that there actually is a linear relationship in the data which can be detected. Important to notice: this is not a question of bad adjustment of the model. The problem is that the model is just not capable to reflect the behaviour of the riders.

This does not affect the statistical test result, which is valid. More people choose the train when it rains.

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