Short Questions to Analyzing the NYC Subway Dataset

Version History

0.1	02/01/2015	Initial Version
0.2	17/01/2015	Amendments following failed evaluation. Please note, all changes are written in blue for easier identification by reviewers

Analyzing the NYC Subway Dataset

Short Questions

Overview

For 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.

For 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.

Section 1. Statistical Test

1.1 Which statistical test did you use?

Mann Whitney U Test

1.2 Why is this statistical test appropriate or applicable to the dataset?

The histogram of ENTRIESn_hourly does not show the typical bell like shape of a normal/Gaussian distribution (not even skewed), instead it shows a pattern of consistent reduction in frequency from one bucket to the next.

So we shouldn't perform a test such as the Students T test which assumes a normal distribution. The Mann Whitney U test (a non-parametric test) does not assume any particular distribution and can tell you whether there is a statistically significant difference between two sample means

More accurately (Paraphrased from - http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm):

“The null hypothesis is that the distributions of both groups (rainy, non rainy) are identical, so that there is a 50% probability that an observation from a value randomly selected from one population exceeds an observation randomly selected from the other population”.

“The P value (returned from the test) answers this question:

If the groups are sampled from populations with identical distributions, what is the chance that random sampling would result in the mean ranks being as far apart (or more so) as observed in this experiment?”

We choose a 2 sided test because rainy day ridership may be higher or lower than non-rainy – we have no prior knowledge to tell us it can be only be either higher or lower

1.3 What results did you get from this statistical test?

Pre chosen p value Significance Level = 0.05

Pre chosen test – 2 tail (see reason above)

Rainy Day Sample Size = 44104

Non-Rainy Day Sample Size = 87847

Rainy Day Mean = 1105.4463767458733

Non Rainy Day Mean = 1090.278780151855

Test Statistic = 1924409167.0

(P Value from scipy = 0.024999912793489721)

2 Tail P value = 0.049999825587

(note scipy performs a one sided test so we have to multiply by 2 for a 2 sided test)

(Rainy Day Median = 282

Non Rainy Day Median = 278)

1.4 What is the significance of these results?

Here we had a null hypothesis that the mean of the two groups is equal (see 1.2 above for a better description) but we received a 2 tail test p value of a fraction under 0.05, this is of course our pre chosen significance level, so do we decide to reject the null hypothesis or not?

Tentatively I'm going to say yes because the sample sets are quite large (and thus we should be getting a more trustworthy answer).

Rainy days had the higher mean (and median) so we expect that more people (on average) go through the turnstiles on rainy days.

(In practice I think a better answer might be that it warrants further analysis, perhaps using different

tests. I have read some articles that seem to suggest that for large sample sizes you can perform parametric tests (even when the distribution is non-normal) but I will confess that I haven't done enough research to understand whether this is true.

Another way to go might be to report confidence intervals.)

Section 2. Linear Regression

2.1 What approach did you use to compute the coefficients theta and produce prediction in your regression model:

a. Gradient descent (as implemented in exercise 3.5) - YES

b. OLS using Statsmodel – YES

c. Or something different? - NO

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

rain

precipi

hour

meantempi

UNIT

ones

weekday - whether or not it was a weekday or weekend

UNIT was dummy coded but this was provided.

For weekday I used 1 for weekday and 0 for weekend. I also experimented with dummy coding day of the week (0 - 6) but this proved to be less effective in practice

2.3 Why are these features appropriate?

- rain, precipi - intuitively rain and the degree of rain will affect whether people choose to walk, cycle, ... or use the subway
- hour - density of travel fluctuates with 'typical working patterns' and 'typical sleeping patterns' for instance
- meantempi - again intuitively this will affect peoples choice of transport (& whether they go out at all)

- UNIT - unit is tightly linked with station, clearly different stations (based upon geography, attractions, population density etc.) will have there own typical patterns of entry
- weekday - a reflection of the working week, more people work Mon - Fri than on Sat/Sun. Calculating simple averages on the dataset backs this intuition up.

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

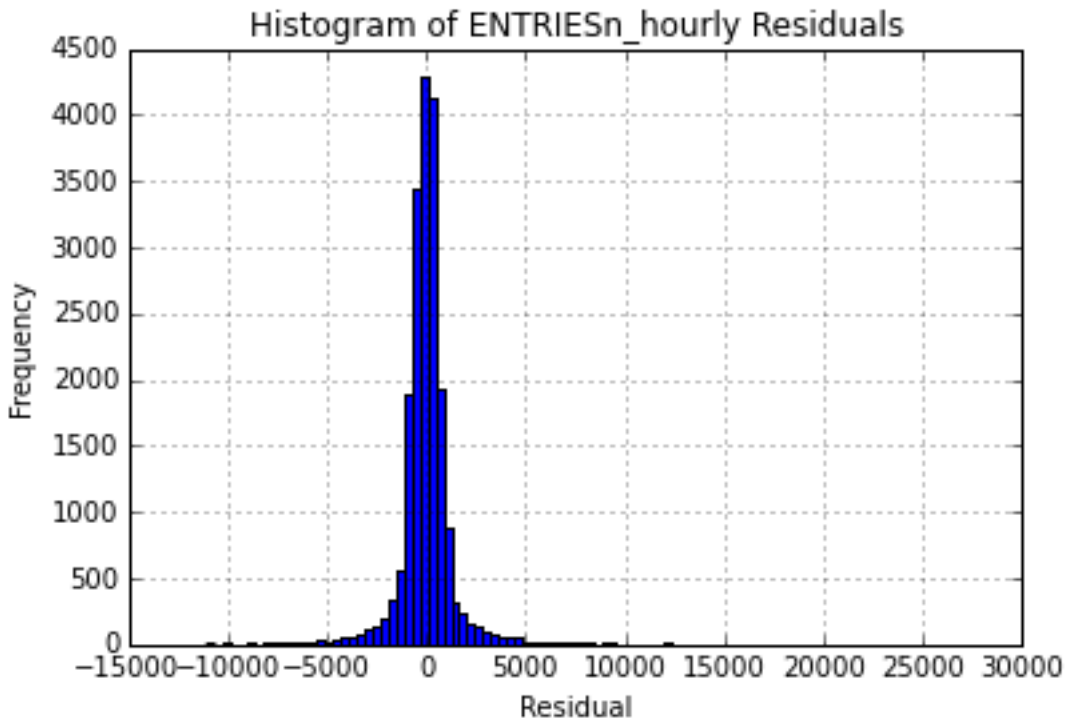
- Gradient Descent: 0.474350279
- OLS: 0.493958188507
- OLS with Regularization: 0.474436652446

2.5 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model is appropriate for this dataset, given this R2 value?

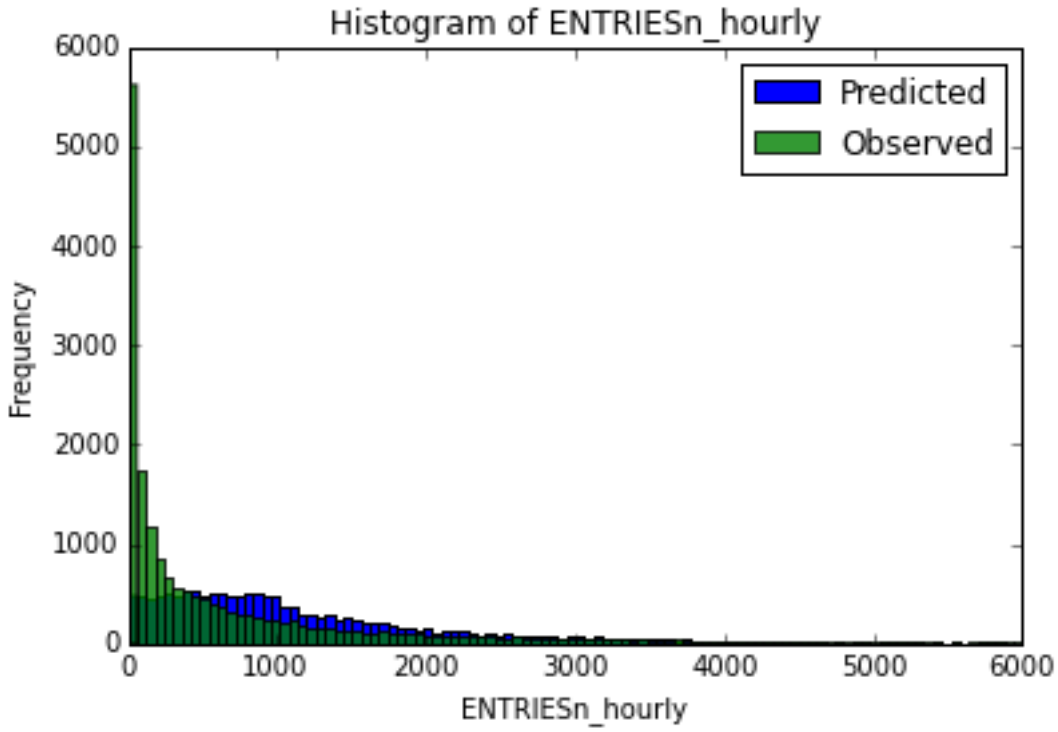
We know 0 is bad (explains none of the error) and 1 is perfect prediction. So is my value of ~ 0.47 good or bad? Hard to say

It's significantly better than the ~ -0.2 we got with the given model (including parameters) but ...

For a random 15% set of the data I first plotted residuals which does give a nicely normal distribution peaking around 0 error, so at first glance it looks pretty good.



I then plotted (bins of) observed values alongside predicted values as follows:



You could conclude that the model does quite well when ENTRIESn_hourly > 300 (ish), for less than that the model is very poor. Looking at the observed values there is a conspicuously high frequency of observations for ENTRIESn_hourly < 10 (ish). Could this be situations stations are closed, either routinely or for maintenance etc.? If so and we can characterise these occurrences into features (either by acquiring extra data or 'better understanding the data we have got' then perhaps we can dramatically improve our model for these cases too.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatterplots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

[A short moan about ggplot:

I found this project particularly frustrating due in no small measure to the python ggplot library. I wasted a lot of time trying to do (arguably) more interesting visualisations combining the turnstile/weather data with station reference data only to be thwarted by missing functionality such as:

- rotate axis labels

- displaying legends

- ordering bar charts by y val

- fill/colour by categorical values

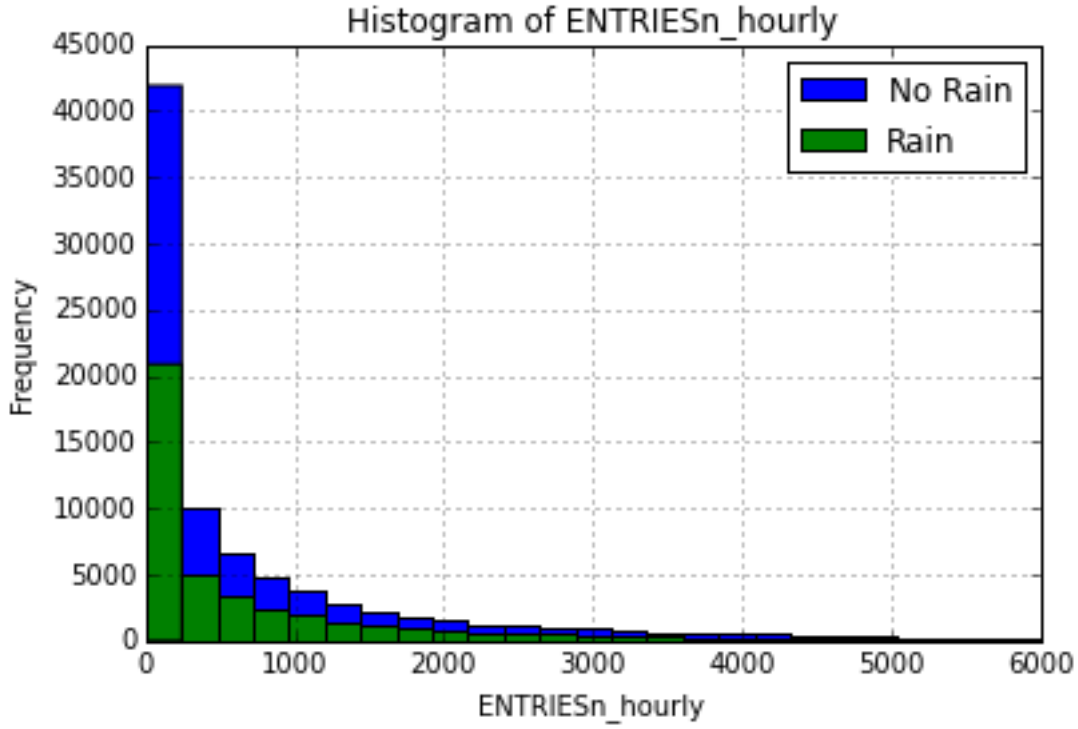
- stack bars by categorical values

- scatter plots with categorical x values

- ...

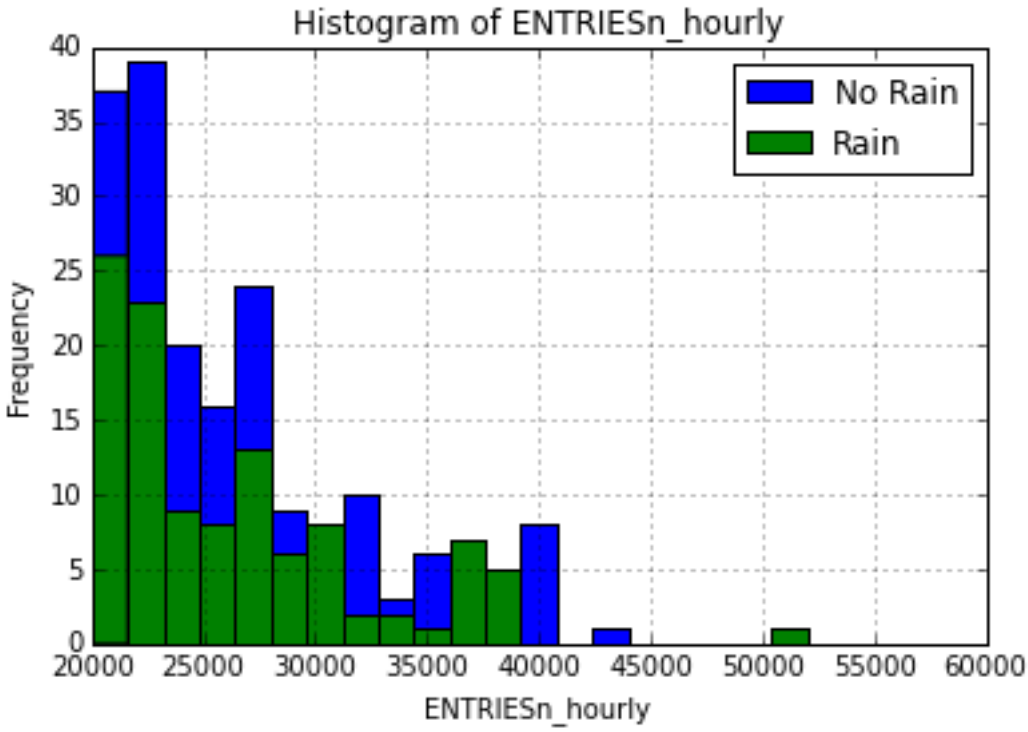
Several of these are known bugs/missing features and either in or scheduled for future releases. At this point in time though I would say that the stable release is 'more trouble than its worth']

3.1 One visualization should be two histograms of ENTRIESn_hourly for rainy days and non-rainy days



Shows, as we might expect that there are more data points for dry weather (roughly twice as many) but that the distributions for rainy vs non-rainy days are similar.

For fun I thought I'd also take a look at the more extreme cases (i.e. very high number of entries) to see if the pattern remained the same. That is to say, was it still the case that the frequency of dry day observations was roughly double that of rainy day cases or was the frequency of rainy day cases now higher

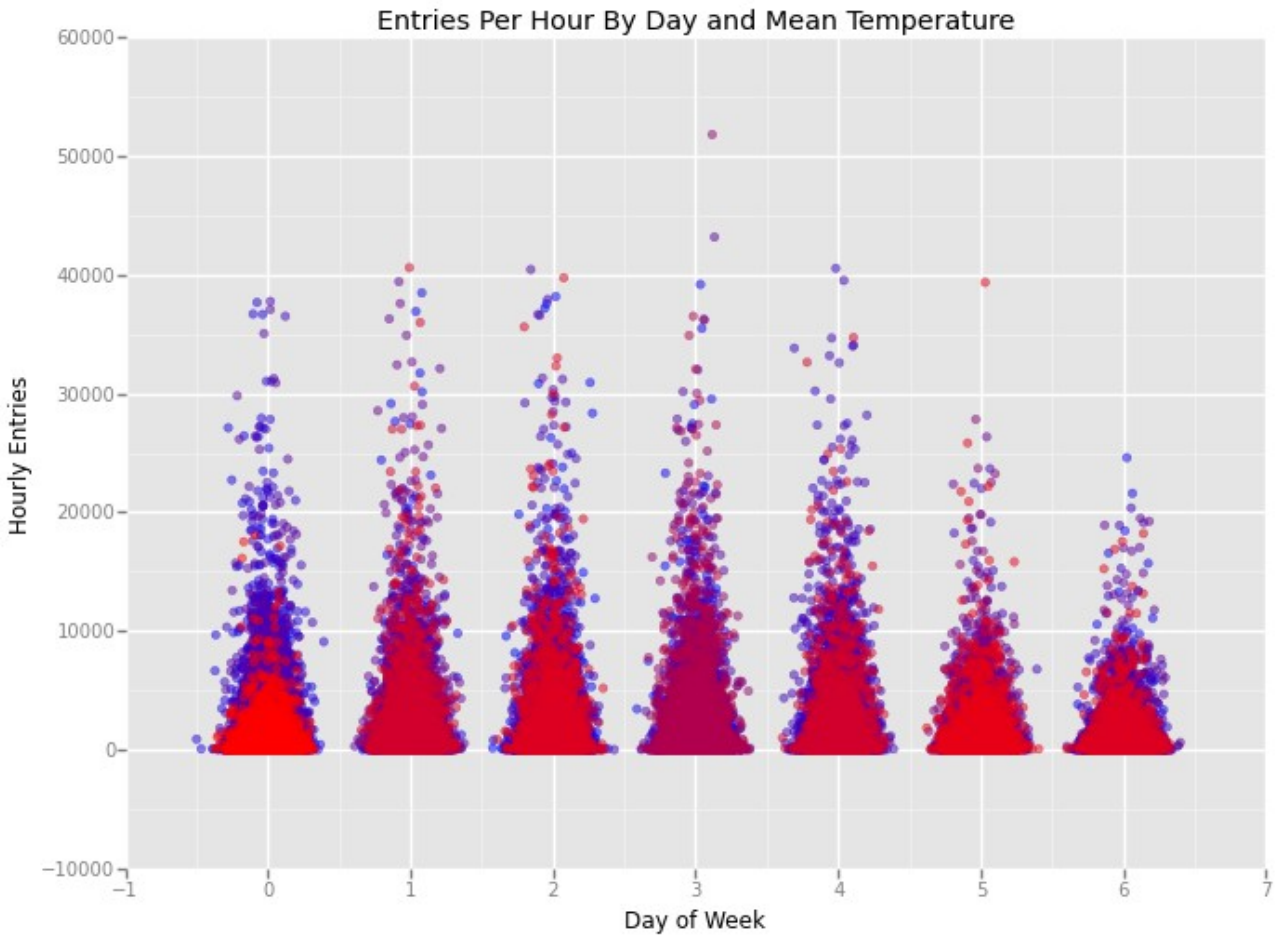


Inconclusive (to the eye)

3.2 One visualization can be more freeform, some suggestions are:

- Rider-ship by time-of-day or day-of-week - YES
- Which stations have more exits or entries at different times of day - NO

Ridership by day-of-week:



I additionally used a colour gradient scale to highlight mean temperature (low = blue, high = red).

You can see that rider-ship is clearly higher during weekdays compared to weekends and indeed it appears to peak towards the middle of the working week. In truth if this was all I wanted to show it would have been clearer to just plot daily averages (or bar chart) rather than every data point.

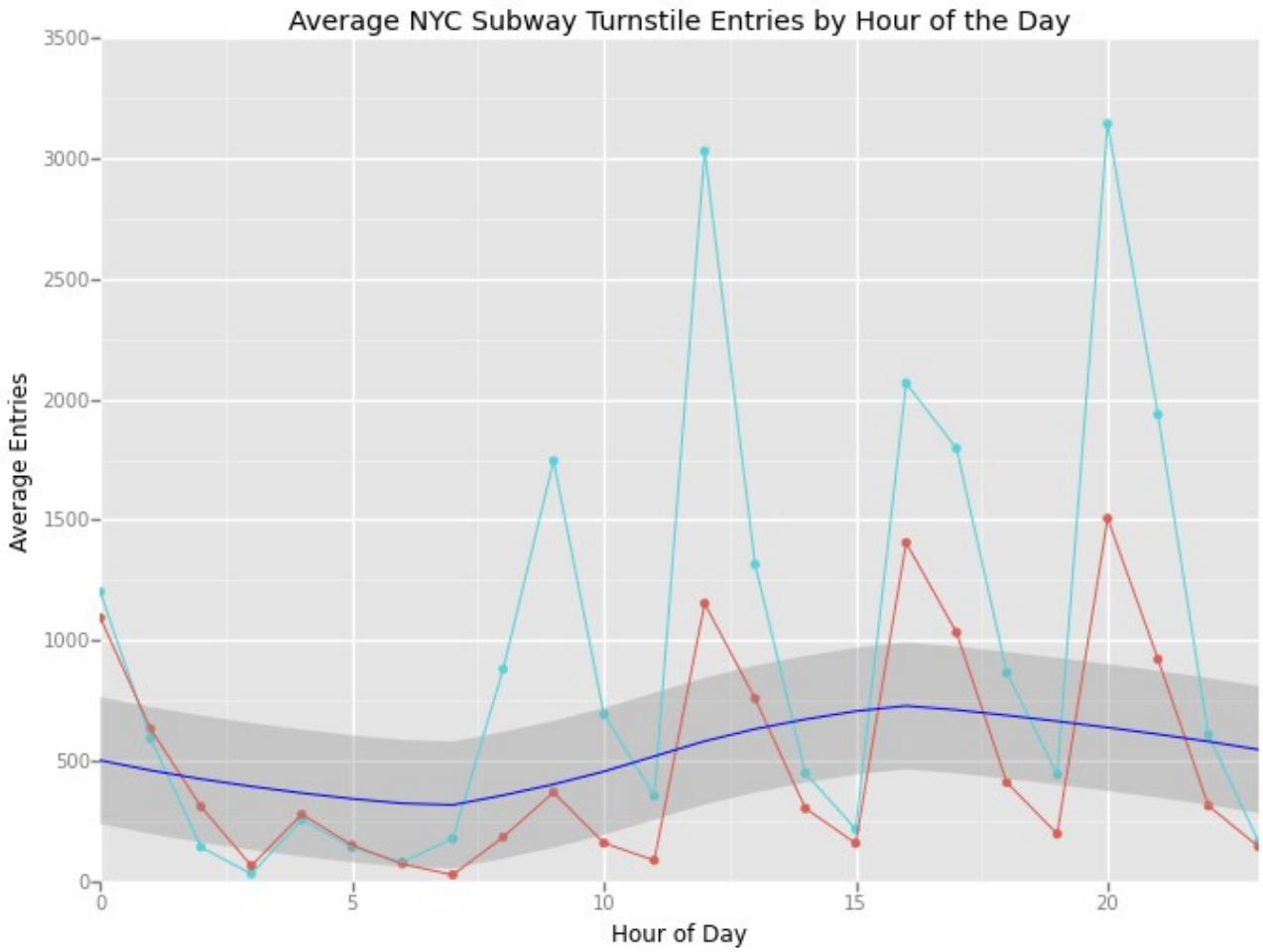
At a glance it looks like there is a tendency for the higher data points to be on days with a lower mean temperature, so do more people use the subway on colder days? Maybe but by way of a cautionary tale, simply changing the low and high colours can lead you to think the exact opposite!!!

Note that for this visualisation I had hoped to utilise the geom_jitter layer to spread the densely populated data points but the spread was proportionate to the x axis value (so day 6 was markedly more spread than day 0 for example). The R ggplot2 library has an api to control the spread, it doesn't look like the python implementation does yet. As a consequence I implemented my own jitter, tailored from - http://nbviewer.ipython.org/gist/fonnesbeck/5850463

Addendum:

I tried to illustrate the 'pattern of subway rider-ship throughout the course of a day'. That is to say plotting the average frequency of subway entries by hour over the course of the data period.

I expected to see slightly different patterns for weekends vs week days (perhaps right shifted on a weekend & without such obvious peaks) so I categorised the data accordingly and plotted:



Legend:

NOTE: I used ggplot (as requested in the problem brief) but there appears to be a bug in the latest stable release that means legends are not displayed (https://github.com/yhat/ggplot/issues/376). Please forgive the absence, in lieu of which below is a description

Light Blue – Weekday Average Entries

Red – Weekend Average Entries

Dark Blue – stat_smooth layer with 0.95 confidence interval (light gray)

(Purpose of the stat_smooth layer is to help visually discern the overall pattern by smoothing natural variation in the plotted data. Here, all the data rainy and non-rainy are used by ggplot to calculate the smooth line)

To my surprise whilst the plot gave a sense of the peaks and troughs you might expect during the day (if you squint a bit), closer inspection showed very sharp peaks and troughs from one hour to the next that 'intuitively did not make sense'.

I consulted the key to the original dataset (http://web.mta.info/developers/resources/nyct/turnstile/ts_Field_Description_pre-10-18-2014.txt), which suggested that audit events occurred every 4 hours (not hourly). The peaks/troughs started to make sense.

I then performed some cursory analysis of the given dataset and determined that most 'units' were indeed audited at 4 hourly intervals - some at 0,4,8,12,16,20 and others at 1,5,9,13,17,21. To confuse things further some units were audited every hour. Cross referencing with the unit - station mapping data (http://web.mta.info/developers/resources/nyct/turnstile/Remote-Booth-Station.xls), it became clear that the 'PTH' division alone was performing hourly audits.

I abandoned this question at this point because I couldn't easily reconcile the data to provide a trustworthy output but it occurred to me that analysis would be much easier if they were to standardise audit periodicity and timing.

It also occurred to me that it would be an interesting exercise to try and smooth/standardise the data. Maybe I'll revisit in a later project....

Section 4. Conclusion

4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining versus when it is not raining?

YES

4.2 What analyses lead you to this conclusion?

2 Tail Mann Whitney U Test suggesting that there was a statistically significant difference between the means of entries whilst raining vs not raining, coupled with a higher mean (and median) for the raining case.

However, in-spite of this conclusion because the p value was the merest fraction under the chosen significance level and the difference between the means is so small (~ 0.015%) I retain a healthy level of suspicion!

Section 5. Reflection

5.1 Please discuss potential shortcomings of the data set and the methods of your analysis.

- **More data**: Here we have considered data from only one month and for only one year. It would therefore be dangerous to make any generalised statements about rider-ship (particularly relating to weather) about other times of the year AND even for that month, the weather (for example) may have been atypical for the time of year.
- **Audit periodicity mismatch**: I've already discussed the 'audit periodicity mismatch' between units (more accurately divisions and stations). This makes drawing conclusions about patterns of use over the course of a day difficult. Recommend - audits occur at a consistent time and frequency across the subway network (preferably hourly)
- **Binary rainy vs non-rainy**: For a given day we are told whether it was rainy or not, this coupled with the amount of rain can enable us to derive models for day by day analysis BUT it would be useful to have a more fine grained time record (e.g. hourly, 4 hourly) of when it was raining.
- **Dirty data**: I have performed little to no checking for (and thus cleansing of) false, missing, inaccurate data points. In this instance the data is pretty well controlled and clean (I've mostly got away with it)
- **Outliers**: I didn't look for and handle outliers that might be skewing my models and conclusions. I have already discussed what appears to be a special case where (maybe) stations are closed and thus reporting 0 entries. Recommend - Get supplementary data detailing station closures whether due to opening hours, maintenance etc. we could then make a feature of this to improve our machine learning algorithm for example
- **Over fitting**: I gave little consideration of over fitting in my 'linear regression model using gradient descent'. Recommend - More testing, I could have chosen more random samples for training and test sets and plotted the effectiveness (using techniques already covered - e.g. plotting residuals) to help determine whether over-fit was a significant problem. Is so then I could introduce a Regularisation term to the model.
- **Higher Order Polynomials**: To be honest its a bit of a 'punt' but it occurs to me that a simple linear model might not be sufficient so I did a little experimenting with higher order polynomials (powers 2 and 3). These both appeared to improve the predictions for low ENTRIESn_hourly BUT both had worse R squared values. Perhaps a more 'reasoned' exploration of higher order polynomials would improve the model.
- **OLS Pre-Requisites**: I didn't properly test the OLS pre-requisites (running out of time)
- & many more ....