Intro to Data Science Project

Jay Baker

Resources

- A variety of wikipedia articles of course.
- Doing Data Science, O'Neil, Schutt
- Predictive Modelling Applications in Actuarial Science: Volume 1 Predictive Modelling Techniques, Frees, Derrig, Meyers
- R Graphics Cookbook, Chang
- Mastering Machine Learning with scikit-learn, Hackeling
- Applied Predictive Analytics, Abbott
- StackOverflow for some pandas questions
- A number of youtube videos on everything from some statistics oriented things to pandas and ggplot.
- I also used a number of the references that were pointed out in the instructor notes.

Section 1. Statistical Test

Which statistical test did you use to analyse the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis?

I used the Mann-Whitney-Wilcoxon test to determine if the difference in ridership when raining versus not raining is statistically significant. The two-tail P value is applicable here as we are determining whether there is any difference, lesser or greater ridership.

The Null hypothesis is that ridership does not vary between when it is raining and when it is not raining. Or, put another way, the likelihood of a random sample from one population being higher than the other is 0.5.

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.

This test works well with this dataset because it works whether the samples are normally distributed or not. The assumptions of the Mann-Whitney test hold for this population as well. For example, observations are independent, and under the null hypothesis the distributions of both groups are equal.

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.

Rain Mean = 1105.4463767458733

No Rain Mean = 1090.278780151855

U = 1924409167.0

p = 0.024999912793489721 * 2 which is just barely under .05

What is the significance and interpretation of these results?

This test tells us that we can reject the null hypothesis. In other words, there is a statistically significant difference in ridership of the NY subway between rainy times and times when it is not raining.

Section 2. Linear Regression

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)

OLS using Statsmodels

Or something different?

I chose to use gradient descent. This is an iterative approach where the values for the thetas are updated by small amounts on each iteration. In other contexts (outside the course) I have utilized an analytical solution.

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

I used rain, precipi, Hour, meantempi, and meanwindspdi as features. Yes, I added UNIT as a dummy feature as that was specifically indicated in the instructions. The unit would seem to be a significant feature as it essentially represents location and obviously ridership is going to tend to vary by location (among other factors).

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 excluded thunder because a quick summary showed it was all zeroes. There were other features I chose to exclude through a combination of “domain expertise” and a quick check to see if including the feature made any difference in R2. For example, I thought that maxpressurei would not make any difference because people don’t make decisions about riding the subway based on what the max barometric pressure for the day is. And, sure enough, including that feature did not help my model’s R2. Other features that fell into this category were things like maxdewpti, meandewpti, etc.

I included Hour as a feature because, obviously, the time of day is going to have a big impact on ridership. A quick bar plot confirmed that.

I also included meantempi as, again, the temperature seems likely to have an impact. And adding the feature did in fact improve the model’s R2.

2.4

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

```
[   7.64069332    4.01365831  463.84049187  -43.17987472   55.19689051]
```

Once again, I used rain, precipi, Hour, meantempi, and meanwindspdi as features, in that order. I used gradient descent for the regression which does not give a measure of statistical significance along with the weights, but in trying different features, this set gave me the best R2 value.

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

```
0.468637530418
```

What does this R^2 value mean for the goodness of fit for your regression model?

I would like to have some other models with which to compare, but 0.468 does not seem all that good. If we were trying to actually predict ridership, we would not be that happy with the results. The results would be better than guessing, but I think we could do better (see next question).

Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

No, I think a different model would be better. I think we got about the best linear fit we can, and the results are not that great. That suggests that something other than a linear model might be better.
Furthermore, I plotted the ridership as a time series and of course this showed that there is some periodicity to the data. Common sense and personal experience tells you this also. So a linear fit might do OK with a short term (in time) prediction, but it is not going to do well across the data.

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., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.
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.

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 different plots.

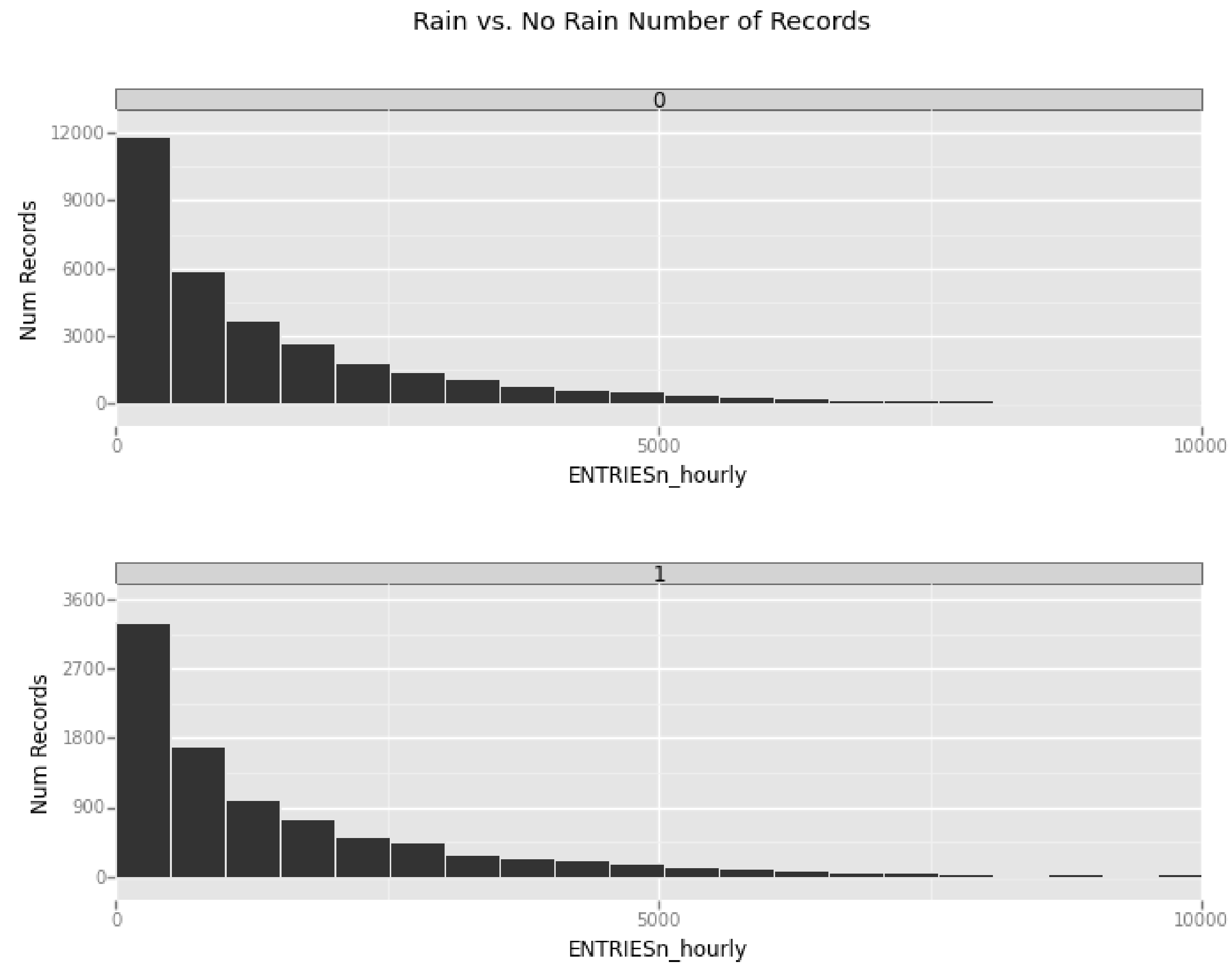
For the histogram, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, you might have one interval (along the x-axis) with values from 0 to 1000. The height of the bar for this interval will then represent the number of records (rows in our data) that have ENTRIESn_hourly that fall into 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.

One visualization can be more freeform, some suggestions are:

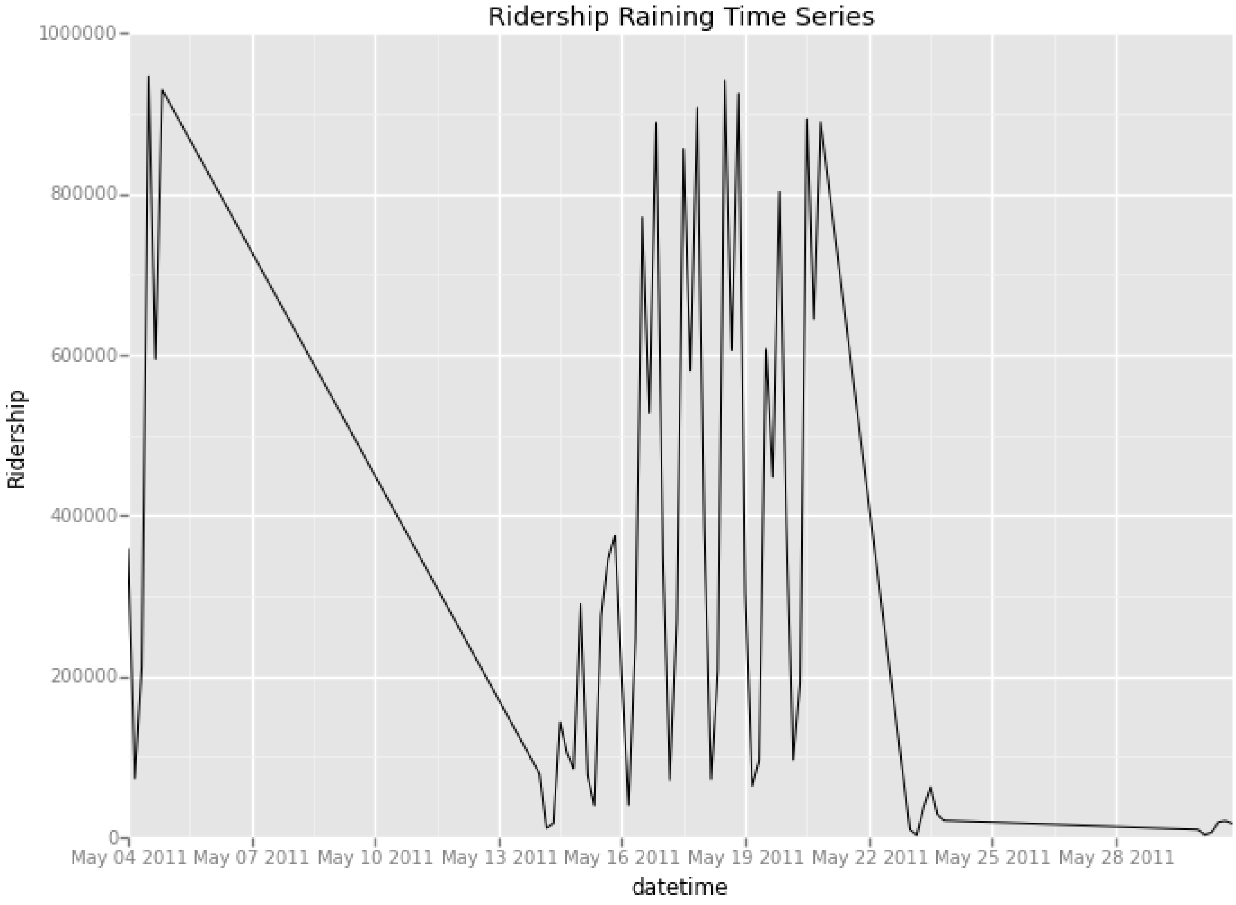
Ridership by time-of-day or day-of-week

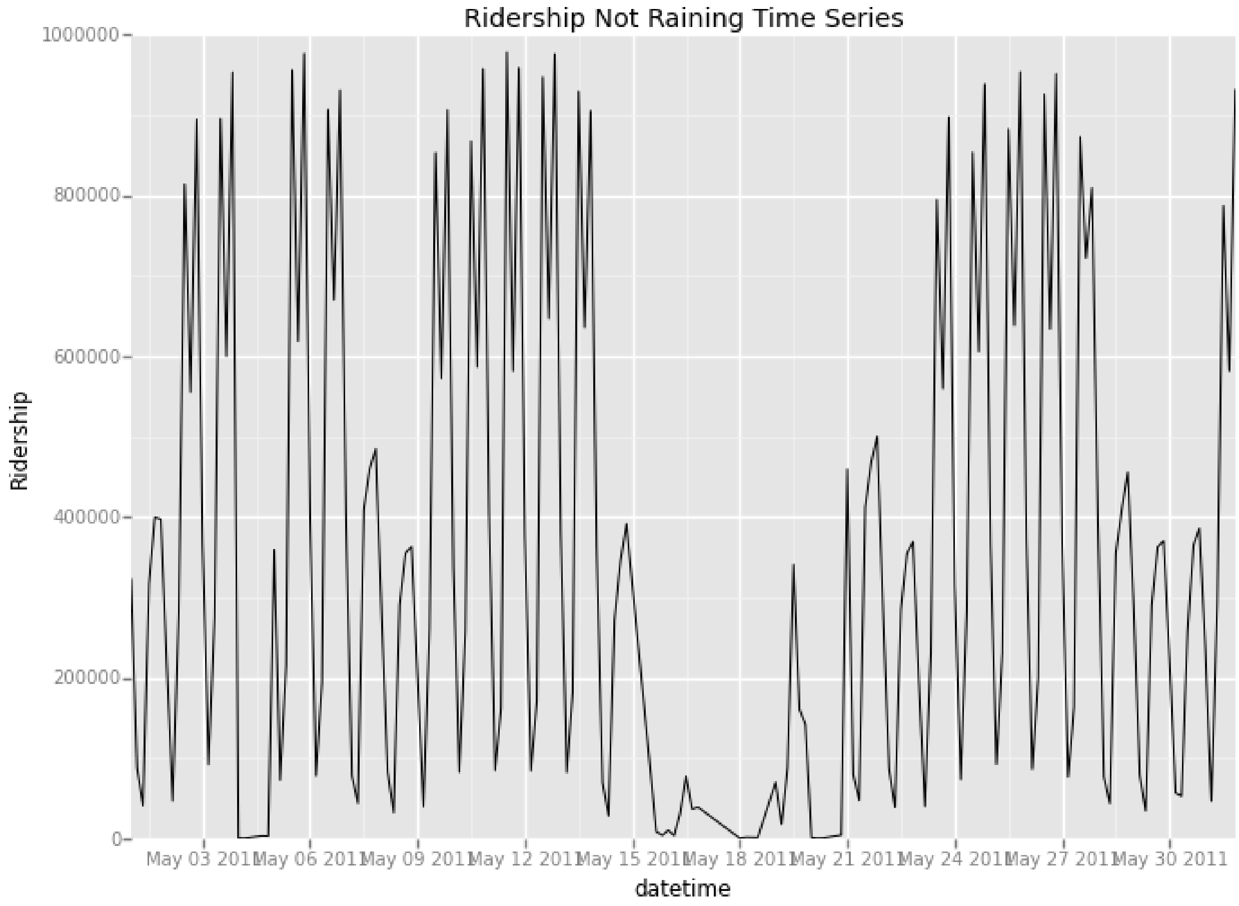
Which stations have more exits or entries at different times of day

I am including the visualizations in the zip file for the project. I also have some visualizations in the iPython notebook that I am including with the project submission. I am pasting two visualizations into this document as well.



This plot is looking at the number of observations. The top plot is no rain (0) and the bottom plot is with rain (1). It makes sense that there would be more observations without rain since it is not raining most of the time. We see the same sort of "spike" in both data sets around 500 entries per hour. And both data sets also exhibit a long tail. The tail on the with rain plot may be more pronounced relative to rest of the with rain data.


Ridership Raining Time Series
Ridership
1000000
800000
600000
400000
200000
0
May 04 2011
May 07 2011
May 10 2011
May 13 2011
May 16 2011
May 19 2011
May 22 2011
May 25 2011
May 28 2011
datetime



These two time series graphs show ridership when raining and not raining. The ridership is expressed as the sum of entries per datetime observed. We easily see from these two graphs that the observations taken when raining essentially fill a period of a few days during this one month of observed data. As discussed above, this fact may tend to weaken confidence in general conclusion drawn from this particular set of observations. A more broad data set over many months might yield more general results / answers.

Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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?
What analyses lead you to this conclusion?

I conclude that more people ride the subway when it is raining versus when it is not raining. However, the difference is not very great and the confidence with which we can say this just barely passes the typical alpha value of 0.05%.

To come to this conclusion, I used a statistical test on the data, the Mann-Whitney U-test, and compared the mean and median of each sample of data (raining versus not raining). This test had a p-value of .05 which means we can reject the null hypothesis that there is no difference in ridership between times when it is raining versus not raining. I used the Mann-Whitney test because the data is not normally distributed. To determine if the data are normally distributed, I plotted a histogram and it was clear that the data is obviously not normally distributed. There are also statistical tests that can be applied.

I also plotted the two samples as time series using the sum of ridership per unit time. This is the volume of ridership, not the count of data points. Visually inspecting this also suggests that the values are pretty close.

Furthermore, a linear regression was performed on the data using rain, precipi, Hour, meantempi, and meanwindspdi as features. The coefficients obtained from this regression were 7.64069332 4.01365831 463.84049187 -43.17987472 55.19689051. Because the coefficient for rain is positive, this further indicates that the presence of rain has the effect of increasing ridership.

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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

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

There are several features of the data set that are less than ideal for determining whether there is a difference in ridership when raining versus not raining.

I think the most significant is that the data only cover one month of time (May 2011). To really look for a general result, I think we need to analyze more data over a greater time horizon. Related to this is the fact that during this one month, there was only one significant period of rain (it looks like there was some rain at the beginning of the month and a little bit after this period of rain in the middle of the month). This is a small sample of data with rain from which to derive general conclusions.

The fact that the data is collected every four hours presents a bit of a challenge as well. The count of data points per hour also reveals that there is some missing data for 08:00 at least. We could potentially add data to the sample through linear interpolation to minimize the impact of this. The count of data points by day of week is also not uniform. And the difference is not due to what day the month starts and ends with.

Finally, with regards to prediction, the data is periodic to some extent, so we need to go beyond a linear model if we want to try and predict ridership outside a short time horizon of a few hours.

And finally, in analyzing this data we probably need to factor in exits per hour in some way. Our model is incomplete without taking this into account.