Data science nanodegree

Introduction to data science project (short questions)

Martin Pons Martinez

INTRODUCTION TO DATA SCIENCE PROJECT: SHORT QUESTIONS

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?

The test used is the Mann-Whitney U test. I used a two-tail p-value since the hypothesis is this

$$H_0: P(x > y) = 0.5$$

$$H_0: P(x > y) \neq 0.5$$

The null hypothesis states that the likelihood of one distribution generating a higher value than the other is 0.5, no matter which distribution is higher. However, the documentation states that the function scipy.stats.mannwhitneyu returns a p-value for a one-sided test. Therefore, we should multiply the p-value returned by two.

With a 5% significance level, my p-value is 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 test assumes the data comes from two independent populations, the distributions can be unknown. This is precisely the problem we had because the distribution of entries for both groups is very skewed. The test also assumes some order (for two values there is a logic in stating which one is the greater), according to wikipedia (http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test#Assumptions_and_formal_statement_of_hypotheses)

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.

Mean entries for rainy days: 1090.28

Mean entries for non-rainy days: 1105.44

statistic: 1924409167

p-value: 0.05 (original is 0.024 which has been multiplied by two)

Taking this into account and assuming that what matters for the exercise is to evaluate if the number of entries is **just different** depending on the weather, the p-value would be 0.05, exactly the critical value for a test hypothesis with a 0.05 significance level.

1.4 What is the significance and interpretation of these results?

The p-value obtain is 0.05, just the critical value, we could reject the null hypothesis with some reserves. That is, rejecting the null hypothesis means that the number of entries is different depending on the weather, specifically if it is raining or not.

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 Gradient Descent

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

The features used to predict entries are: 'rain', 'precipi', 'Hour', 'meantempi', plus other **dummy variables** representing each unit.

The first one is a **dummy variable** indicating whether is raining, the second one measures the amount of precipitation, the third one registers the hour in the day and the forth, the mean temperature.

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 R2 value."

I selected 'Rain' because of the test results on the previous exercise showing that there is probably a difference in the number of entries depending on if it's raining or not, the other three variables seemed the most logical option among the rest.

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

The coefficients are (assuming the first coefficient is the independent term, and the next coefficients in "theta_gradient_descent" are in the same order than the features.

Rain: 1.53443995 (this is actually a dummy variable)

precipi: 2.84499968

Hour: -3.73560088

meantempi: 9.99985952

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

The R2 coefficient is 0.318

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?

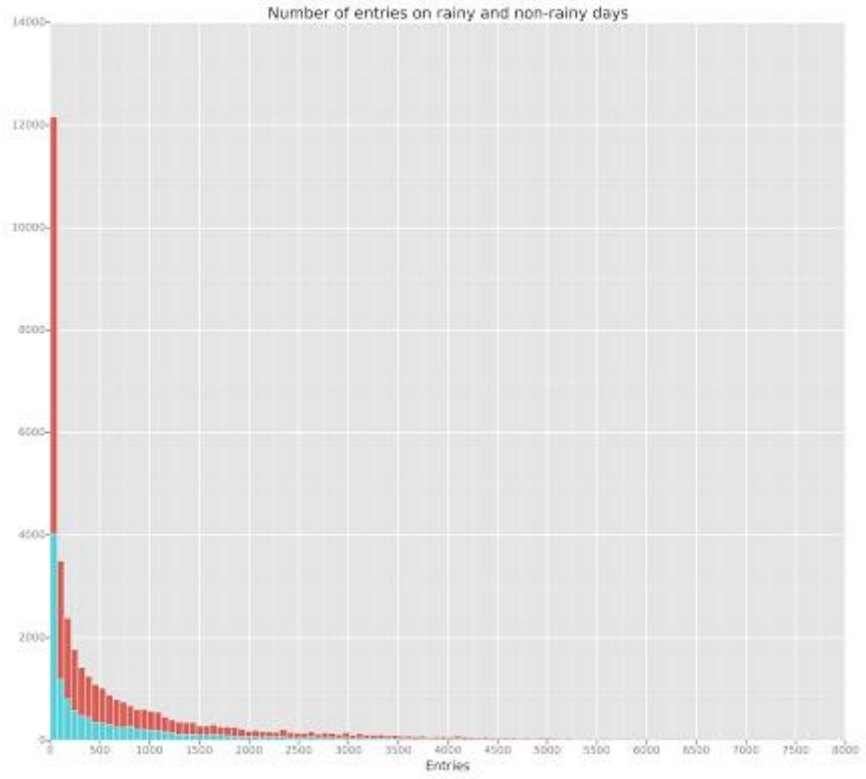
The 0.318 value for R2 means that my model explains the 31.8% of the total variability in the entries variable.

The number entries in a subway station seems to be something routine and repetitive. I guess that given the right features (some of them sure are things like variables related with the weather, like the ones in the data set) it can be can be predicted with a higher accuracy that the one I have obtained, but for a precise answer for this question maybe a higher expertise on the field is needed.

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.

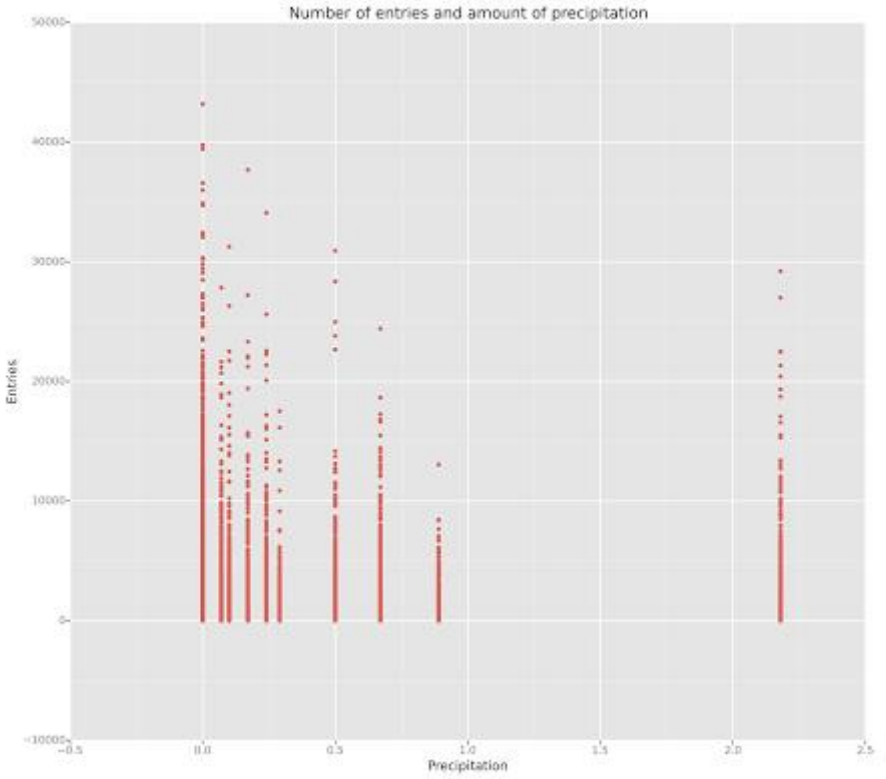
- **You can combine the two histograms in a single plot or you can use two separate plots.**
- **If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.**
- **For the 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.**



In this first plot we can see higher frecuency uf lower number of entries for rainy days.

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**



It can be seen in the scatterplot that the number of entries reach the highest values for low amounts of precipitation, there are a group of extreme values in precipitation (around 2.2), which a not so low level of entries. The data shluld be inspected in order to check wheter these values are legitimate or some kind of errors (measurement, typo, etc.)

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?

It seems that more people ride the NYC subway when it is raining. THe means are not very different, though: the mean entries for rainy days is 1090.28 and for non-rainy days is 105.44.

But From the exploratory analysis, that is, the histogram, it can be seen that the distributions clearly different depending on whether it's raining or not.

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 hypothesis test using the Mann-Whitney U test leads us to the conclusion (with some reserves) that the entries on rainy days are different from the entries on non-rainy days.

On the other hand, the rain variable used in the linear regression shows a positive coefficient for the dummy variable rain: 1.53443 these results seem to be contradictory from the distribution means and the histogram (that is, a positive coefficient means that more entries are expected when it rains).

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.**

Dataset

The number of examples in the data set is 42649, large enough, to extract reliable conclusions when the appropriate test or predictive models are applied. There also seem to be important variables in predicting the number of entries, like those related with the weather.

However maybe the number of variables is not enough and maybe there are missing variables which can be important in explaining the entries.

The period referenced by the data could also be an issue: the data has been collected in the month of May, if there is a seasonal behavior in the number of entries, model predictions could be not extrapolated for other seasons in the year.

Analysis

The statistical test is an appropriate one considering the distributions, but is recommended to report whit some other figures like the means, and in this case, the means are not very different. Besides, the p-value is 0.05, exactly the critical value. This leads us to contemplate the test outcomes with some reserves.

The regression shows a positive relationship between rainy days and number of entries when a negative one is expected, as showed in the histogram, but the regression model can have its limitations: firstly, the R2 obtained is not very high. Maybe we are missing relevant variables (missing values can lead us to biased coefficients), or maybe the relation is not linear.

Url used during the project

ggplot documentation
https://pypi.python.org/pypi/ggplot/

Facets in ggplot
http://blog.yhathq.com/posts/aggregating-and-plotting-time-series-in-python.html

Datetime module
https://docs.python.org/2/library/datetime.html

Strings in python
https://docs.python.org/2/library/datetime.html

Mann-Whitney U test
http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test#Assumptions_and_formal_statement_of_hypotheses