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.

1. http://pandas.pydata.org/pandas-docs/stable/generated/pandas.get_dummies.html - to understand how to create dummy data
2. http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.mannwhitneyu.html - for the MannWhitneyU test syntax
3. https://storage.googleapis.com/supplemental_media/udacityu/649959144/MannWhitneyUTest.pdf?GoogleAccessId=1069728276824-fdhtlb98k1m9qrmdgj4jgc7gjp2l1lsm@developer.gserviceaccount.com&Expires=1432263445&Signature=lR2iiPm347gZ080ezCRaSracD/Wi2PhKd2uXzIORN2TeimihvXRdBTBFLBqXqCmDJdy7aqeRtJActr7fLkGHTjukG/NcOuaE3XPlmZ58RAar06DsDB9OuL5rbQBQG%2BaPrsH9pivB6mMaDAlo%2Ba6m58MK3l/k/nkg5JOWbqLkM9U%3D – To understand the MannWhitney U test
4.

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 determine whether there would be significant difference in the subway ridership on the days it rained versus the days without rain. I used the two-tailed P value because we are only concerned with whether or not the two samples come from the same population. The null hypothesis is that the two samples (with and without rain) would be identical that is they come from the same population. That is:

H-nought: x = y (where x and y are the two samples)

H-Alternate: x < y or x > y

The P critical value was 0.05 (alpha level of 0.05).

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 Mann Whitney U test is applicable because we can't make any assumptions about the distribution of ridership in the two samples. Also the sample sizes could be significantly different.

1.2 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 of the two samples, the MannWhitneyU value and the P-value were as follows:

Mean of entries with rain:	1105.4463767458733,
Mean of entries without rain:	1090.278780151855,
MannWhitneyU:	1924409167.0,
P-value:	0.024999912793489721

1.4 What is the significance and interpretation of these results?

The P-value of 0.0499 (we multiple the P-value by 2 for a two-sided test) indicates that we should reject the null at a P-Critical value of 0.05. This indicates that there IS a statistically significant difference in the subway ridership when it is raining versus when it is not raining. It is worth noting that the P-value was very close to the threshold that I had set to determine whether or not the value is statistically significant.

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 OLS using Statsmodels. This model gave me a better R^2 than the Linear Gradient Descent (which was .4662) for the same set of features.

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

I first tried R^2 with just 'UNIT' in the features list. I got an R^2 value of 0.4482

Next I tried adding 'rain' and 'fog' to the features. The R^2 values changed only slightly to 0.4490.

Next I tried with only 'Hour' and 'UNIT' in the features. The R^2 went to 0.4822 which tells me that in addition to the station, the hour of the day is a significant indicator or the ridership.

I then added ‘rain’ and ‘fog’ back in along with ‘Hour’ and ‘UNIT’ to the features. The R^2 changed slightly to 0.4830. This tell me that the rain and fog have impact (albeit small) on ridership.

I then tried ‘UNIT’, ‘Hour’, ‘rain’ and ‘meantempi’ as the features. I got a R^2 of 0.4833, which tells me that the temperature has a slightly more significant impact than fog on the ridership

Next I tried ‘UNIT’, ‘Hour’, ‘rain’, ‘meantempi’ and 'meanwindspdi' as the features. I got a R^2 of 0.4838

Next I tried ‘UNIT’, ‘Hour’, ‘rain’, ‘fog’, ‘meantempi’ and 'meanwindspdi' as the features. I got a R^2 of 0.4849 (making progress!)

Adding ‘thunder’ to the above did not change the R^2 score at all, which tells me that the ridership was not impacted by whether there was thunder or not.

I added 'meandewpti', 'meanpressurei' to the above (excluded ‘thunder’) and got an R^2 of .4850.

I then tried ‘UNIT’, 'Hour', 'rain', 'fog', 'mintempi', 'maxtempi','meanwindspdi', 'maxdewpti', 'mindewpti','maxpressurei', 'minpressurei' and got a R^2 score of .4865.

I then normalized the non-dummy features and then removed 1 of the dummy variables from the features data frame. I found that the R^2 value jumped to .578

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 tried various combinations of features (see above) to see how adding or removing them would change the R^2 value. I found that UNIT and Hour had the most impact but the other features also contributed slightly (about 5%) to the R^2 score. I also found that by normalizing the non-dummy features, the error regarding a potential multi-collinearity in the model went away. Also removing just one of the dummy variables from the features seemed to help improve the value of R^2.

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

When I exclude the dummy variable, and don’t normalize the features I get the following:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        ENTRIESn_hourly   R-squared:                       0.208
Model:                            OLS   Adj. R-squared:                  0.208
Method:                 Least Squares   F-statistic:                     263.0
Date:                Wed, 27 May 2015   Prob (F-statistic):               0.00
```

```
Time:                        18:53:36   Log-Likelihood:                -91290.
No. Observations:               10000   AIC:                         1.826e+05
Df Residuals:                    9990   BIC:                         1.827e+05
Df Model:                          10
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------
Hour            55.7828      3.232     17.259      0.000        49.447    62.118
rain          -108.0636     72.866     -1.483      0.138      -250.895    34.768
fog            179.6255     76.512      2.348      0.019        29.647   329.604
mintempi       -26.6995     11.292     -2.365      0.018       -48.833    -4.566
maxtempi         7.0879      5.617      1.262      0.207        -3.923    18.099
meanwindspdi    42.4883     12.127      3.503      0.000        18.716    66.260
maxdewpti       20.2353      9.896      2.045      0.041         0.837    39.634
mindewpti       -7.9576      7.173     -1.109      0.267       -22.017     6.102
maxpressurei   744.7955    438.160      1.700      0.089      -114.086  1603.677
minpressurei  -733.2743    439.772     -1.667      0.095     -1595.316   128.768
==============================================================================
Omnibus:                    11157.795   Durbin-Watson:                   1.345
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1224831.396
Skew:                           5.702   Prob(JB):                         0.00
Kurtosis:                      56.005   Cond. No.                     3.53e+03
==============================================================================


Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spec
ified.
[2] The condition number is large, 3.53e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
Your R^2 value is: 0.0339755953001
```

```
Can you beat the 0.4 R^2 value that we achieved with gradient descent?
```

When I normalize the non-dummy variables and exclude the dummy variables, the get the following values:

```
                           OLS Regression Results
==============================================================================
Dep. Variable:        ENTRIESn_hourly   R-squared:                       0.028
Model:                            OLS   Adj. R-squared:                  0.027
Method:                 Least Squares   F-statistic:                     28.83
Date:                Thu, 28 May 2015   Prob (F-statistic):           3.21e-55
Time:                        05:14:10   Log-Likelihood:                -92316.
No. Observations:               10000   AIC:                         1.847e+05
Df Residuals:                    9990   BIC:                         1.847e+05
Df Model:                          10
Covariance Type:            nonrobust
================================================================================
                  coef    std err          t      P>|t|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------
Hour          385.2343     24.735     15.574      0.000       336.749   433.720
rain          -56.4811     38.306     -1.474      0.140      -131.569    18.607
fog            60.9746     31.956      1.908      0.056        -1.665   123.615
mintempi     -201.3315     82.172     -2.450      0.014      -362.405   -40.258
maxtempi       68.2454     48.662      1.402      0.161       -27.142   163.633
meanwindspdi   78.3658     26.870      2.917      0.004        25.696   131.036
maxdewpti     144.2691     98.577      1.464      0.143       -48.963   337.501
mindewpti     -41.2067     95.101     -0.433      0.665      -227.624   145.211
maxpressurei   63.5229     65.621      0.968      0.333       -65.108   192.154
minpressurei -119.3004     72.802     -1.639      0.101      -262.007    23.406
==============================================================================
Omnibus:                    11162.053   Durbin-Watson:                   1.095
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1227438.269
Skew:                           5.705   Prob(JB):                         0.00
```

```
Kurtosis:                      56.063   Cond. No.                         10.5
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spec
ified.
```

Note that there is no warning about potential multicollinearity in the above, however, the R^2 value has now dropped to .028.

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

The highest R^2 score I got was .571. Here is the summary:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        ENTRIESn_hourly   R-squared:                       0.571
Model:                            OLS   Adj. R-squared:                  0.549
Method:                 Least Squares   F-statistic:                     26.71
Date:                Thu, 28 May 2015   Prob (F-statistic):               0.00
Time:                        05:18:03   Log-Likelihood:                -88231.
No. Observations:               10000   AIC:                         1.774e+05
Df Residuals:                    9526   BIC:                         1.808e+05
Df Model:                         474
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------
Hour           431.6455     17.312     24.933      0.000       397.710   465.581
rain           -45.4625     26.647     -1.706      0.088       -97.697     6.772
fog             52.3587     22.241      2.354      0.019         8.762    95.955
mintempi      -160.1259     57.215     -2.799      0.005      -272.279   -47.973
maxtempi        16.1255     33.819      0.477      0.634       -50.167    82.418
meanwindspdi    58.5070     18.730      3.124      0.002        21.791    95.223
maxdewpti      254.0944     68.498      3.710      0.000       119.825   388.364
```

```
mindewpti     -164.7691     66.022     -2.496      0.013      -294.187    -35.352
maxpressurei    14.2347     45.672      0.312      0.755       -75.292    103.761
minpressurei   -38.5183     50.683     -0.760      0.447      -137.867     60.831
unit_R002      497.4693    486.004      1.024      0.306      -455.201   1450.140
unit_R003      272.2407    507.980      0.536      0.592      -723.509   1267.991
unit_R004      641.2801    507.683      1.263      0.207      -353.887   1636.447
unit_R005      980.0455    467.147      2.098      0.036        64.338   1895.753
unit_R006      678.7972    595.481      1.140      0.254      -488.472   1846.067
unit_R007      285.5227    421.027      0.678      0.498      -539.780   1110.825
~
unit_R548      150.2592    231.269      0.650      0.516      -303.078    603.596
unit_R549       12.6980     54.381      0.234      0.815       -93.900    119.296
unit_R550       15.0537     72.058      0.209      0.835      -126.194    156.302
unit_R551       92.2285    112.249      0.822      0.411      -127.803    312.260
unit_R552      127.6401    107.391      1.189      0.235       -82.869    338.150
==============================================================================
Omnibus:                     9113.198   Durbin-Watson:                   1.798
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1167818.916
Skew:                           3.911   Prob(JB):                         0.00
Kurtosis:                      55.360   Cond. No.                         94.7
==============================================================================


Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spec
ified.
Your R^2 value is: 0.476014497048
Can you beat the 0.4 R^2 value that we achieved with gradient descent?
```

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?

The R^2 score of .571 indicates that 57.10% of the variability in the data can be attributed to the features chosen. While there is still a lot of variability in the predicted values, a value over 57% is a good fit. The model shows that much of the variation can be attributed to the station rather than any of the weather related data, although the weather does contribute to variability in ridership to a small extent.

Section 3. Visualization

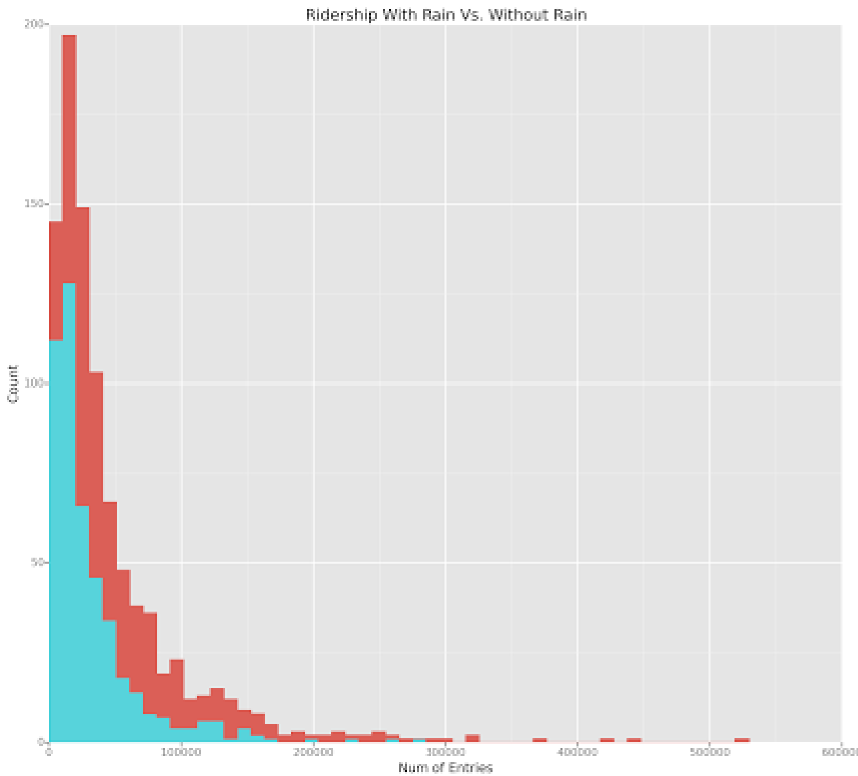
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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 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.

The histogram produced was as follows:

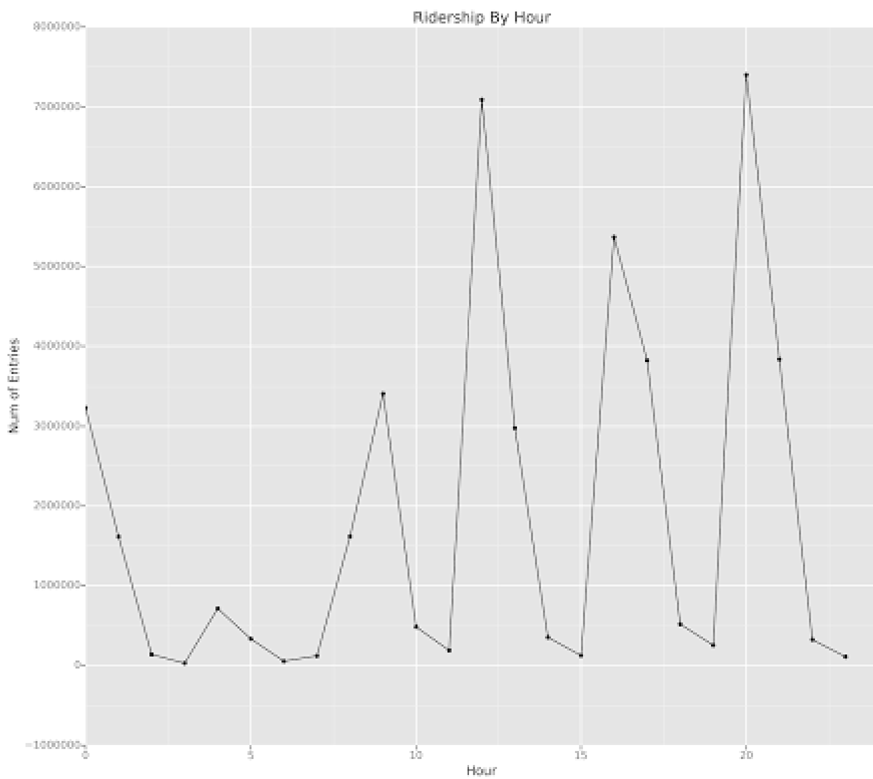


The blue bar represent Number of Entries when it was raining and the red represents Number of Entries without rain. The Number of Entries for each Unit were grouped together and summed for with and without rain. We see from the histogram that there is a noticeable difference in the ridership when it rains vs. when it does not rain across all stations. It appears that the ridership decreases when it is raining.

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

I plotted the graph of ridership by time of day. The result was as follows:



The plot above shows that when mapped across all stations, the ridership peaks around 1 PM and 8 PM with the next significant number being around 4-5 PM and 9 PM.

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?

The histogram clearly shows that there are more riders when it is not raining than when it is raining when the data is aggregated over all the stations. This appears to be in conflict with the mean values of entries when it rains vs. when it does not rain in the sample provided, as was seen in 3.3 (mean of entries with rain = 1105.45 and mean of entries without rain = 1090.28), but that might be due to outliers in the sample data set. The data in the histogram was aggregated over all stations and mapped by frequency of occurrence of a given number of entries, thus removing any station specific variability.

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 MannWhiteyU statistic was used to determine that there *IS* a statistically significant difference in the ridership when it is raining or when it is not raining. Although the P-value was almost equal to the P-critical value that I used for my threshold, it still proved that the difference was significant.

When I used the OLS method for linear regression, I used rain as one of the features to predict the value. The coefficient for 'rain' was a -45.46 and the P value was 0.088. The confidence interval was towards the negative as well, which I think indicates that there was an inverse relationship between rain and ridership (or entries).

Hence both the statistics support the conclusion that the ridership decreases 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.

I have thus far only run the code in Udacity's test environment (IDE provided) and hence the models were likely only run against a subset of the data (~10%). I would like to set up an environment on my local machine and rerun some of the code locally to see if I get different results. That would indicate how consistent my model was against different datasets.

Also, in the OLS model used to predict ridership, a large part of the variability is due to the dummy variables (or stations). Hence the effect of weather related features is overshadowed and cannot be interpreted properly.

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