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.
http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
http://en.wikipedia.org/wiki/Coefficient_of_determination

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?
Statistic test: Mann–Whitney U test
Use two-tail P value
Null hypothesis: the distribution of number of people riding the subway when it is raining is same as the distribution of number of people riding the subway when it is not raining.
p-critical value: 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.
Both of the two samples of the data are not normally distributed as shown in problem 3.1. But they are independent and have enough number of samples (>20). Mann–Whitney U test does not assume the data is drawn from any particular underlying probability distribution. Mann–Whitney U test does assume number of observation in each sample is > 20 and you have 2 independent samples. In sum, the Mann–Whitney U test is appropriate for this dataset.

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.
U statistics = 1924409167.0
p-value = 0.0249*2 = 0.0498
$\mu_1 = 1105.4463767458733$ (rainy days)
$\mu_2 = 1090.278780151855$ (non-rainy days)

1.4 What is the significance and interpretation of these results?
The P critical is 0.05. The significant level is 95%.
Since the P value (0.0498) is smaller than P critical (0.05), we can reject the null hypothesis and conclude that the distributions of two groups are different. That is to say, the distribution of the number of entries is statistically different between rainy and non-rainy days.

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?

Both Gradient Descent and OLS using Statsmodels have been applied. I would like to share the analysis using OLS in the following questions.

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 'EXITSn_hourly' , 'fog', 'rain', 'meantempi', 'Hour' in my model. 'fog' and 'rain' here were used as dummy variable since they have boolean values of 0 and 1.

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."

'EXITSn_hourly' is based on my intuition. If a station were major interception of two subway lines or a busy station in the city center, I would expect the number of entries is as many as number of exits.
'fog' and 'rain' are based on my intuition. If it were a foggy or rainy day, it would be un-safe to drive. So more people would like to take subway.
'Hour' is based on my intuition. It is expected that more people would take subway at busy hours of the day.
'meantempi' is based on data exploration. When I added that feature to my model, the overall R squared did not change. But its p value is much smaller than the p critical. So it is statistically significant to include this feature in this model.
2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
non-dummy:
const 342.1632
EXITSn_hourly 0.7665
meantempi -7.4734

Hour 19.5112

Here are the results I got in problem 3.8:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        ENTRIESn_hourly   R-squared:                       0.554
Model:                            OLS   Adj. R-squared:                  0.554
Method:                 Least Squares   F-statistic:                     2484.
Date:                Mon, 04 May 2015   Prob (F-statistic):               0.00
Time:                        06:05:06   Log-Likelihood:                -87425.
No. Observations:               10000   AIC:                         1.749e+05
Df Residuals:                    9994   BIC:                         1.749e+05
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                    coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const          342.1632     78.524      4.357      0.000       188.240   496.087
EXITSn_hourly    0.7665      0.007    108.282      0.000         0.753     0.780
Hour            19.5112      2.220      8.787      0.000        15.159    23.864
meantempi       -7.4734      2.372     -3.151      0.002       -12.123    -2.824
rain_0.0       182.3361     46.769      3.899      0.000        90.660   274.012
rain_1.0       159.8271     39.572      4.039      0.000        82.259   237.395
fog_0.0        113.0362     40.188      2.813      0.005        34.260   191.812
fog_1.0        229.1270     50.219      4.563      0.000       130.687   327.566
==============================================================================
Omnibus:                     5907.593   Durbin-Watson:                   1.770
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1859605.166
Skew:                           1.634   Prob(JB):                         0.00
Kurtosis:                      69.726   Cond. No.                     5.08e+19
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 2.13e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
Your R^2 value is: 0.55409305141
Can you beat the 0.4 R^2 value that we achieved with gradient descent?
```

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

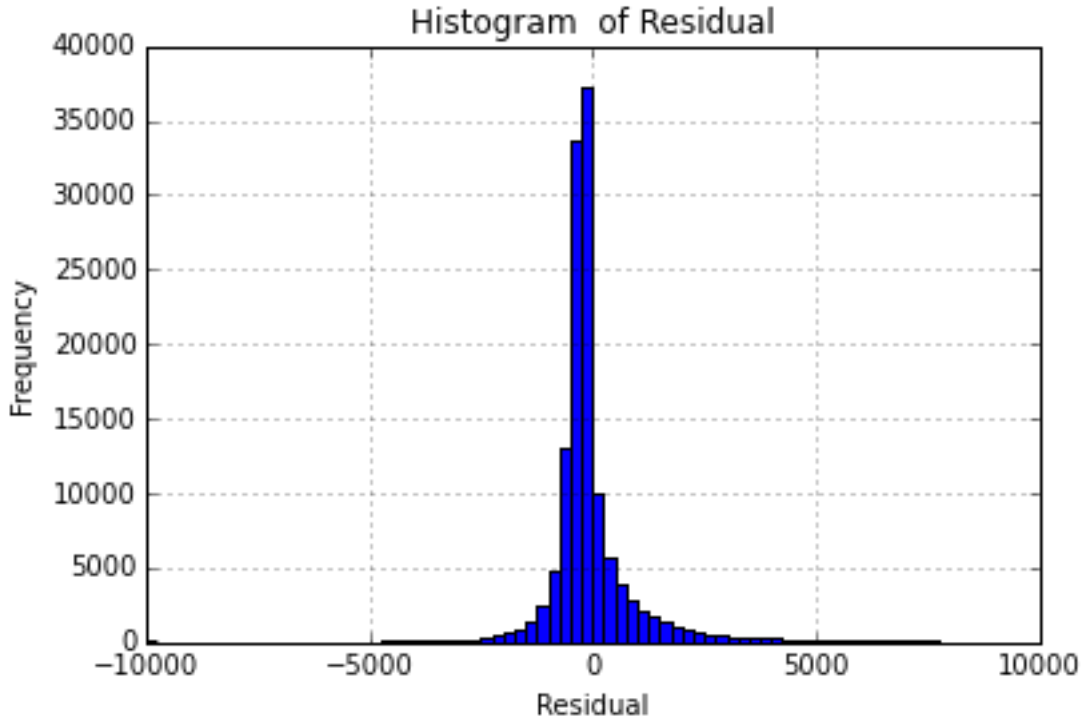
$R^2 = 0.554$

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 could be interpreted as the percentage of variability that could be explained by the regression model. The closer R^2 is to 1, the better goodness of fit of the regression model.

In this case 0.554 is relatively not good enough to predict the ridership. We need to look at the residuals to determine how good the model is. The following plot is a

histogram of the residual. We can see that the residuals are largely normally distributed. We can conclude that the variance of predicted values is normally distributed. To sum up, even the R^2 is not that high, variance could be as large as 5000, this linear regression model is valid to be used to predict the subway ridership.



Section 3. Visualization

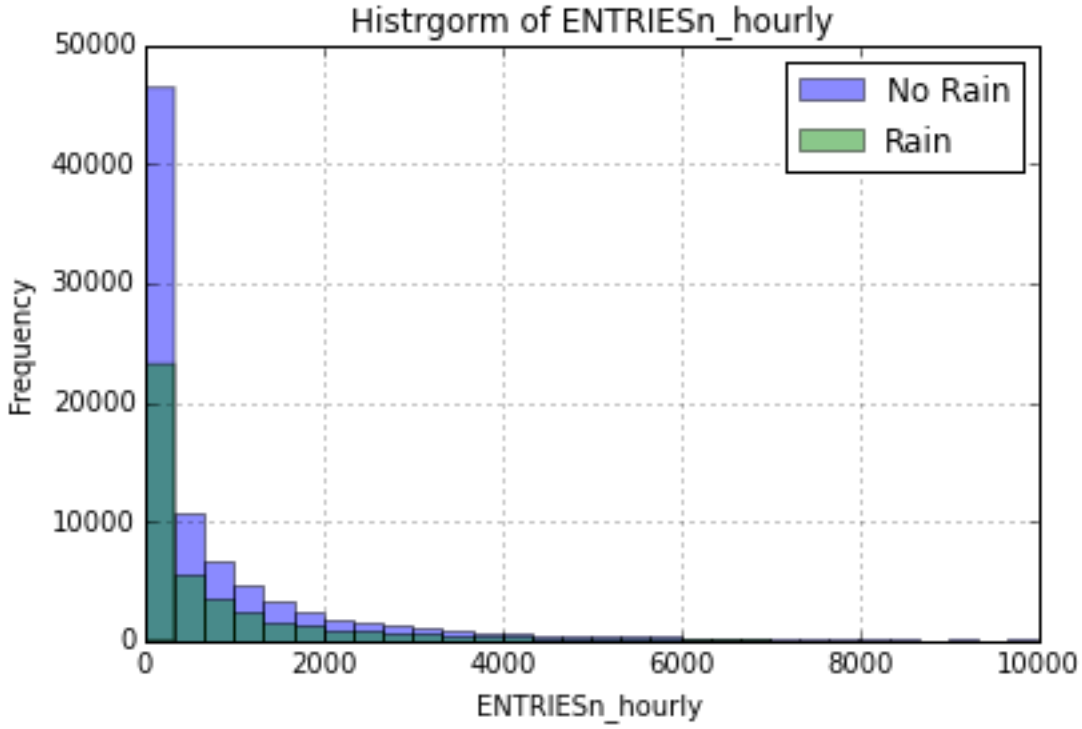
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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.

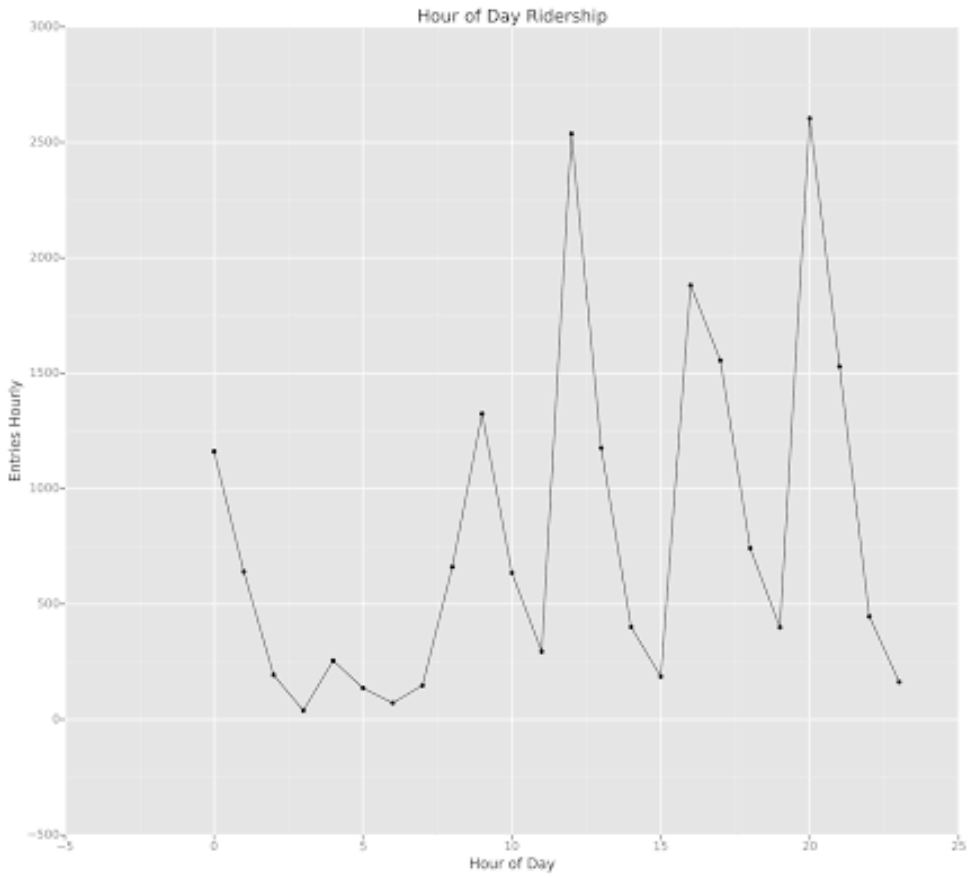
- Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.



From the histogram plots above, we can clearly see that both ENTRIESn_hourly when rainy and ENTRIESn_hourly when not rainy are not normally distributed. The majority of two samples were both at the first bin, which indicates most of the samples have a relatively small number of ENTRIESn_hourly. Also, there are more samples at not rainy group.

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



From the line chart above, we can see the average ENTRIESn_hourly of each hour of the day. Some peaks indicates that more ENTRIESn_hourly happened at busy hours around 9am, 12pm, 16pm, and 20pm.

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?

I conclude that more people ride the NYC subway when it is 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 Mann–Whitney U test indicated that the distribution of number of entries is statistically different between rainy and non-rainy days. And the mean of ENTRIESn_hourly at rainy days (1105) is larger than the mean of ENTRIESn_hourly at non-rainy days (1090). So I can conclude that more people ride the NYC subway when it is raining.

The OLS regression analysis used “rain” as a dummy variable. Both rain_0.0 and rain_1.0 (non-rainy or rainy) have p value < p critical. They are significant in the

linear regression model. The coefficient of rain_1.0 = 159.8271, which is a positive number, indicates that when it is rainy, the ENTRIESn_hourly will be increase. So I can conclude that more people ride the NYC subway when it is raining.

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.

The dataset only consist of one month data of May 2011. The ridership could vary a lot in different month/season of the year due to temperature and weather variations. It will be great to have a better understanding of the ridership if we can have more data of longer duration, say one year's data. Also more samples would be able to perform cross validations.

The linear regression model I build have a R suqared of 0.554, which is not very high. I would like to seek other models, such as polynomial or Logestic to see if better fit can be achieved.

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

I think in city like NYC, the subway ridership could vary due to special events such as sport games, shows, concerts, etc. I would be curious to see the correlations between ridership and special events.