Udacity Data Analyst Nanodegree: Cohort February 2015
Class:Intro to Data Science
Project 1: Short Questions
Student: Gordana Neskovic

Analyzing the NYC Subway DataSet
(Does more people ride the subway when it rains vs. doesn't rain?)

0. References:

Data: turnstile_data_master_weather.csv for 1. & 3.
turnstile_weather_v2.csv for 2.
Udacity Intro to Data Science Notes
http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
http://graphpad.com/guides/prism/6/statistics/index.htm?one-tail_vs__two-tail_p_values.htm
Intuitive Biostatistics by Harvey Motulsky
http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm
itoledoc ipNotes

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?

Since entries for rainy and non-rainy days do not fit Gaussian probability distribution, **Mann-Whitney U non-parametric test** is used for data analysis. Also, even if the samples were Gaussian, since samples are large, U test produces a non-significantly larger P values than parametric test.
Null Hypothesis: Rainy and non rainy days turnstiles' hourly entries means are **equal** and any difference between them is due to chance.
Since direction of the rainy and non rainy days turnstiles' hourly entries means difference was not specified before collecting data, a **two-tail P** value is selected. Also we might be intrigued if data goes in “wrong direction”, i.e. if less people ride the subway when it rains vs. when it doesn't.
P-critical is %5 or 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.

Assumptions that Mann-Whitney U test is making about the distribution of ridership in the two samples are the following:

- rainy and non rainy days turnstiles' hourly entries are **randomly selected from larger population.**

- rainy and non rainy days turnstiles' hourly entries **(between and within samples) were obtained independently**.

- rainy and non rainy days turnstiles' hourly entries **don't follow particular distribution but they have the same shape**.

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	= 1924409167.00	(Umax = 1937202044)
p-value (1-tailed)	= 0.02	
p-value (2-tailed)	= 0.04	
Rain_Raidership_mean	= 1105.45	(Rain_count = 44104)
No-Rain_Raidership_mean	= 1090.28	(No-Rain_count = 87847)

1.4 What is the significance and interpretation of these results?

According to U test, **Null hypothesis is rejected** since p-value(2-tailed) = 0.04 is smaller than p-critical = 0.05 and it could be concluded that **there is statistical significance** and that more people rides NYC subway when it rains.

However, if Cohen's d is calculated d = 0.0065, it tells us that difference between rain and no-rain data is **practically insignificant**.

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?

Gradient Descent (GD) with normalization features scaling, learning rate (alpha = 0.1) and 75 iterations, as well as OLS with the same set of features was used for Ridership prediction.

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

Features for both models were UNIT, hour and weekday. UNIT was dummy variable.

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."

OLS with each separate variable in input file turnstile_weather_v2.csv was used for exploration and experimentation and it turned out that *UNIT* has by far the largest coefficient of determination ($R^2 = 0.3752$), then *hour* ($R^2 = 0.0823$) and *weekday* ($R^2 = 0.0212$), while the other variables i.e. potential features contributed with each $R^2 < 0.009$.

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

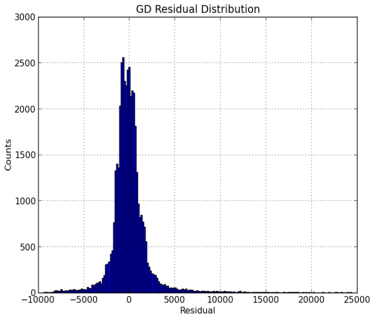
The linear regression model coefficients (or weights) of the non-dummy features start with 3rd value of Θ .

Θ =

```
[  8.55820574e+02   4.43060853e+02  -1.05248074e+02  -8.60918283e+01
  -8.53396463e+01  -7.96885324e+01  -9.54963214e+01  -9.46762049e+01
  -9.96812883e+01   3.59080869e+02   4.49534405e+02   4.76104497e+01
  -7.13595831e+01   1.53993919e+02   3.89404843e+02   9.16716054e+01
   2.97338497e+02   1.85736379e+02   5.04466976e+02   2.82818251e+02
   8.92223009e+01   2.29798812e+02   7.29776739e+01   3.53723800e+02
   8.16838010e+01   1.64140737e+02   1.70207455e+02   4.19923728e+02
  -4.27278137e+01   6.19540025e+01  -6.97026031e+01  -6.52724192e+01
  -1.04505764e+02  -6.64041642e+01  -3.91507129e+01   8.16714056e+01
  -8.07739652e+01   6.76494518e+01   1.85653722e+02   4.27171455e+02
   6.01184022e+01   1.42022171e+02   2.15691861e+02  -3.77599343e+01
   8.69294692e+01  -2.56207328e+01   4.27061314e+02  -2.66238008e+01
   1.99117531e+02  -7.97291612e+01  -4.14755089e+01  -6.74372975e+01
  -7.53484135e+01   5.77462898e+01  -4.03946972e+01  -6.38612796e+01
  -6.12532793e+01  -9.77975548e+01  -5.85649120e+01  -8.29713940e+01
  -5.52906981e+01  -4.76797397e+00   1.15404771e+02   1.11642433e+02
  -2.27361269e+01   8.34177309e+01   5.38218402e+02   4.97593136e+01
   4.87419664e+01  -4.00528980e+01  -8.74910853e+01  -8.66081830e+01
  -4.30135256e+01   1.11409865e+01   1.29607216e+01  -4.79859581e+00
   2.21977532e+01   3.38528234e+01   7.46587969e+01  -1.67941136e+00
   3.50132507e+01  -8.30363675e+01   6.36276978e+01   1.22177580e+02
  -2.53405939e+01  -3.17260842e+01   9.91024299e+01  -4.23540379e+01
  -7.89517315e+01   2.23546252e+02   9.16121078e+01  -9.39560316e+00
  -6.16187952e+01  -3.69448395e+01   9.02383509e+01  -5.59367507e+01
   3.70740945e+00  -1.78751000e+01  -2.07501129e+01   4.92747800e+01
  -1.17281609e+01  -7.09820927e+01   2.09514434e+00   1.95806561e+02
   4.14386367e+01   4.67888671e+01   9.93521074e+01   4.89365617e+00
   3.26401655e+02  -1.97538562e+00  -6.10470217e+01  -4.76815734e+01
  -4.68627107e+01   3.36720712e+01  -2.48843854e+01   1.28958209e+01
  -3.13310150e+01   1.97513649e+01  -6.95437346e+01  -4.86301291e+01
   2.71492262e+01  -1.26549267e+00  -2.39913633e+01  -1.90431154e+01
   1.24729275e+01   4.67484459e+01  -5.98741723e+01  -7.98979796e+01
   3.90767947e+01  -7.93856721e+00  -4.03246320e+01  -6.75052657e+01
  -1.36738750e+01  -6.80359503e+01  -4.73161414e+01   1.18120633e+01
  -3.51081338e+01  -2.18731781e+01  -2.17316021e+01   2.09728974e+01
  -6.68608004e+01  -7.96875164e+01  -6.43528058e+01  -4.69531848e+01
  -4.17331292e+01  -7.51957508e+01  -7.97223384e+01  -5.36157265e+01
  -4.66115513e+01  -3.95939564e+01  -8.96843336e+01   5.02238593e+01
  -1.72865831e+01  -6.92465331e+01   1.98365378e+01  -6.03146445e+01
   5.61498400e+01  -7.93985325e+01  -2.49161318e+01  -1.25301773e+01
  -6.77652418e+01  -9.86038819e+01   8.50454790e+01  -2.79407205e+01
  -5.22138789e+01  -3.42499274e+01  -5.54015020e+01  -6.87591805e+01
   5.12107064e+01  -6.68060058e+01  -4.92395441e+01   4.71268433e+00
  -1.86583747e+01  -5.57960403e+01  -4.17942581e+01  -1.36931158e+01
  -7.86266077e+01  -1.07805982e+02  -8.84557031e+01  -5.59834332e+01
  -5.90243339e+01  -6.05429240e+01  -8.10091525e+01  -8.86336802e+01
  -2.87974582e+01  -5.30864694e+01  -5.74667335e+01  -2.67388772e+01
  -7.90537510e+01  -8.86024848e+01  -6.60513495e+01  -6.94389940e+01
  -3.38070571e+01  -1.48806412e+01  -6.79551101e+01  -7.49024689e+01
  -7.46272373e+01  -8.35820473e-01  -5.46263581e+01  -6.88615880e+01
   2.61102149e+01  -3.69404769e+01  -4.83735570e+01  -8.86881618e+01
  -6.30665908e+01  -6.33828853e+01  -3.01207735e+01  -8.97759627e+01
  -9.68472049e+01  -1.06319334e+02  -8.27570240e+01  -2.95982378e+01
  -4.88131929e+01  -2.73326245e+00  -3.32104175e+01  -9.30182096e+01
  -5.29456800e+01  -8.64260869e+01  -1.09368963e+02  -1.08060986e+02
  -1.19256900e+02  -8.63133705e+01  -8.23465456e+01  -8.52892062e+01
  -3.56697012e+01  -1.06391472e+02  -1.00548229e+02  -4.75611479e+01
  -9.84606990e+01  -8.54941294e+01  -7.04032883e+01  -6.91264205e+01
  -7.40201078e+01  -5.85182106e+01  -8.89316994e+01  -5.40436923e+01
  -8.35006458e+01  -1.19939150e+02   1.88589194e+03]
```

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

Both GD and OLS produced R^2 = 0.4814 and Residual Distribution (should be roughly normal and ~ independently distributed with 0 mean and constant variance) as on Figure below.



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 R^2 = 0.4814 means that this linear model can predict NYC subway ridership with ~48% accuracy. Further analysis (f.e. bias/variance) as well as application of more complex models (f.e. polynomial regression) should be applied on this dataset in order to verify if GD/OLS model is appropriate for this dataset.

3. Visualization:

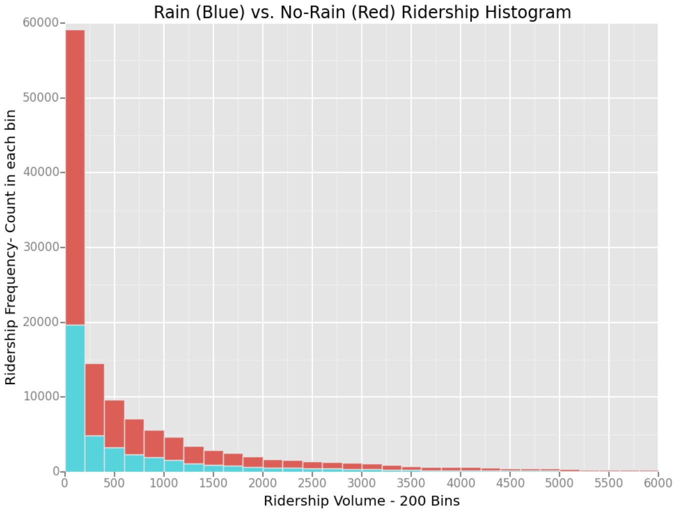
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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.

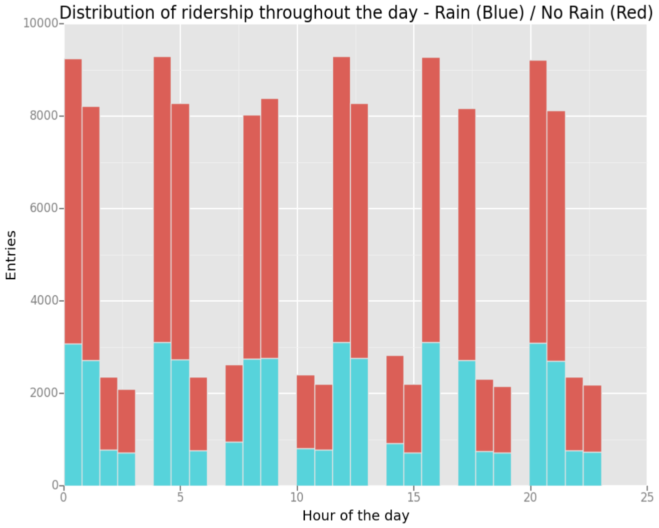
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) 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.



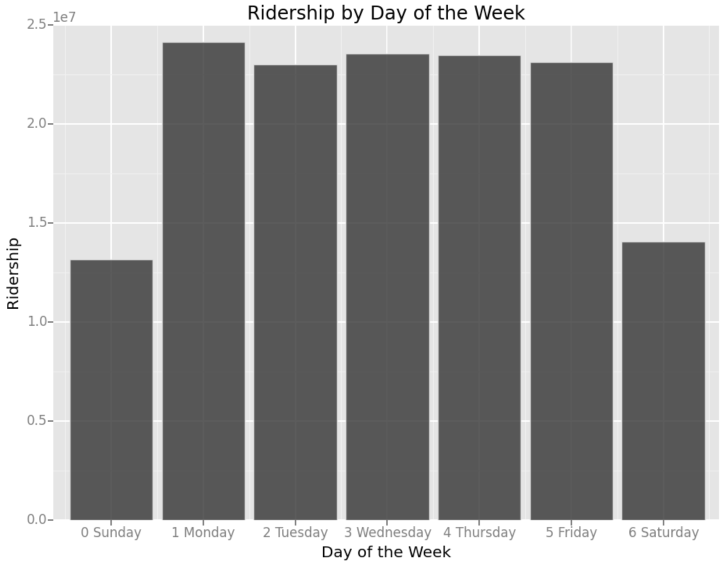
Rain vs. No-rain ridership distributions show that both distributions are non-normal, right - skewed and that their shape is the same.

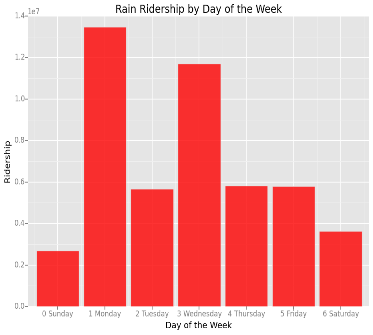
3.2 One visualization can be more free form. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

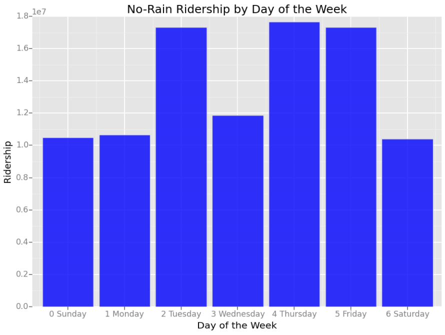
- Ridership by time-of-day
- Ridership by day-of-week



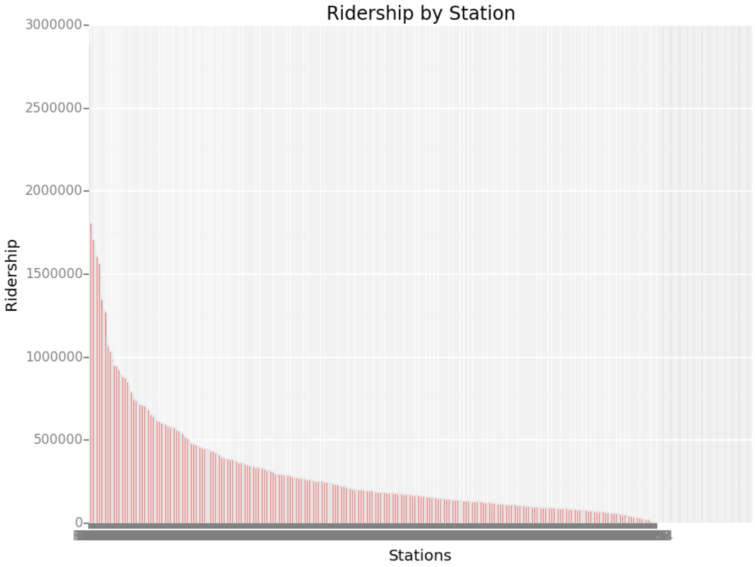
Rain vs. No-rain ridership distributions by hour of the day are of similar shape.







Rain vs. No-rain ridership distributions by day of the week have some similarities (the lowest ridership is on the weekend) and some differences (Mon & Wed have the most ridership for Rain data and Tue,Thur & Fri for No-Rain data).



There is large difference in Ridership among stations. Part of the difference is due to station data availability.

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?

From performed Statistical Analysis as well as from Linear Regression analysis it could be concluded that rain does not practically significantly influence ridership.

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.

According to U test, since p-value(2-tailed) = 0.04 is smaller than p-critical = 0.5, it could be concluded that there is statistical significance and that more people rides NYC subway when it rains. However, Cohen's d = 0.0065 tells that difference between rain and no-rain data is practically insignificant.

Also, OLS with only rain as feature gives coefficient of determination $R^2 = 0.0007$ i.e. according to applied linear models, rain influences << 0.07% NYC ridership prediction accuracy.

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.

Potential shortcomings of applied methods due to dataset are:

- Dataset is focused only on one month of one year. This might not be good dataset for making rain vs. no-rain influence on NYC ridership.

- Different variables not included in available dataset might have more influence on NYC ridership.

- More detailed and diverse dataset analysis could be performed that would led to applying different methods:

    - More data visualization type of dataset analysis

    - Finding transformation that would make rain and no-rain data distribution Gaussian and led to parametric statistical tests application (however, nonparametric tests are slightly less powerful than parametric in case of large samples).

Potential shortcomings of applied methods due to analysis are:

- Statistical tests could be done on more subsets of given dataset and led to improved insight of rain influence on NYC ridership.

- Further analysis of applied linear regression models as testing for over/under fitting could be performed.

- Application of more complex models (f.e. polynomial regression) could be applied on this dataset in order to obtain better NYC ridership prediction and rain influence on it.

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

No.