Analyzing the NYC Subway Dataset

Short Questions

Section 1

Statistical Test

1.1 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? What is your p-critical value?

Given that the data is not normally distributed (it is right-skewed), I decided to use the Mann-Whitney U Test to test whether it is more likely that more people take the subway in a rainy day or not.

The data has been analyzed using a two-tail test because I don't want to rule out any possibility (a one-tail test would test in one direction only, not both).

The null hypothesis is:

$H_0 : P(x_A > x_B) = 0.5$,
$H_1 : P(x_A > x_B) \neq 0.5$,
where x_A and x_B are the *ENTRIESn_hourly* for rainy days and non-rainy days respectively.

The test has been done at a 95% significance level (p-critical 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.

This is a good choice because it doesn't make any assumptions on the probability distribution of the data: there are two populations (rainy and non-rainy) for which the distribution is unknown. Besides, we are not making any assumptions on the direction of the result, so the test checks on both ends (it's two-tailed).

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.
The mean for non-rainy days is 1845.5 and for rainy days is 2028.2.

The Mann-Whitney U test returned a p-value of 2.74e-06, doubling the value gives 5.48e-06.

(results obtained from aremeansequal.py script, run under python2.7)

Since the p-value returned is much smaller than p-critical, the null hypotheses is rejected. It can't be concluded that there will be more people using the subway in a rainy day.

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:

A. Gradient descent (as implemented in exercise 3.5)
B. OLS using Statsmodels
C. Or something different?

I've used Gradient descent to predict entries hourly for rainy days (option A).

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

The input variables are the entriesn_hourly for rainy days (rain=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.

I've did some tests using a meantempi value of 0.016 or more, foggy days and cold days but I didn't find any significant difference, so I sticked to using rainy vs non-rainy days.

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

67.9744157264 -71.3248609458 786.551373465 -77.6455823304 -97.8945995818 -80.0161392302 -79.8162538385
-75.8952576547 -89.3012428696 -87.7830968907 -93.2672981784 334.249609999 417.818783457 44.2684156945
-65.7878079176 143.141805558 361.032045795 83.9508065004 276.367056951 172.820217904 468.87342095
262.871859109 81.5930999407 212.327002903 67.8448554747 328.77187253 75.9363786097 152.57231564 158.145468314
390.298454396 -37.8384773671 58.096950241 -65.0139480381 -61.7143420255 -97.9550239631 -60.8098891214
-37.3224725631 75.9248583188 -74.8839144987 62.8927761692 172.566603271 397.034533026 55.8933766058
131.882140378 200.484217791 -35.4635802384 80.1483775212 -23.559168788 395.783600289 -24.2229157338
185.079943181 -74.25047108 -38.3469276477 -62.1169791559 -69.0212197024 53.6887220901 -37.296592406
-58.9589648197 -56.6116568116 -89.7823689524 -55.1521849957 -77.9254678891 -51.5874924185 -4.41238358244
107.276836114 103.342718444 -20.949472277 77.5479027227 500.242185454 46.1836642328 45.3200536726
-36.5184590167 -80.7803677436 -81.4780603615 -40.4248816852 9.06510121803 11.8030932743 -5.67507439534
19.9845984072 30.2684477142 68.5067611291 -1.54185625622 32.5605086891 -78.7537929001 59.1549355183
113.571523029 -25.0309636784 -30.3786637149 92.1253496007 -38.8697271475 -73.4181857542 207.784132369
85.1638024216 -9.92668249589 -58.6431386511 -35.5860353163 83.8870250453 -51.9635123244 3.13539827069
-17.6175898938 -18.7169876008 44.2532681789 -10.3536487695 -65.3031260608 1.96623688022 182.002708922
37.3837315859 43.7628831916 92.3574011734 4.56718940079 303.378518308 -1.76717616851 -56.64165671
-43.0622285035 -43.5175476327 31.2493485091 -22.1122647702 12.1578991743 -28.9343936918 18.5233903976
-64.2081140969 -46.1712361022 23.7812735914 -0.43471373561 -22.2787089199 -18.3768647011 11.5468434335
42.3591318539 -55.1392221241 -73.488145924 36.3371891777 -7.10424394093 -37.0727754338 -61.6696743712
-12.4593348705 -63.2943912486 -43.2792785147 9.78312615054 -33.8531348191 -20.3100557908 -18.0698994074
18.2978317855 -61.3576951021 -73.840129311 -59.8560944807 -43.6195539747 -38.5201371702 -69.2042401749
-73.5460099417 -49.4439472253 -41.731475993 -36.5601020527 -82.2317051161 46.5279940008 -16.4903696815
-65.6241376721 17.2355237163 -56.0377691981 52.3332775409 -72.3627849529 -23.8129380386 -11.4643018727
-62.2146593718 -90.2665016167 79.3186158801 -25.1433525344 -47.8167745898 -31.0880633066 -51.3203168401
-66.489374352 46.2910578131 -63.3185932121 -45.6323400942 4.39899315452 -16.7442809308 -51.5642701368
-40.4049088692 -13.0270526862 -72.3957245025 -100.079246188 -82.3198410124 -50.5249245393 -56.0062443896
-55.7240888583 -73.6529881661 -81.9660813608 -24.4807379221 -49.8545022611 -54.2130429919 -24.8322636724
-72.0751723985 -81.7526884662 -61.7732044858 -63.5569261332 -30.705286154 -13.6235382939 -62.9875948717
-69.7423028258 -67.9640139217 -0.757818176748 -50.6359051183 -62.2825084387 24.2859775053 -34.2125440617
-44.8274948191 -82.4762236087 -58.9489621374 -60.3842293461 -28.4619887731 -82.3090399312 -89.5926150918
-98.4141799098 -76.5512681396 -27.4511657251 -45.3482559031 -2.31266696425 -28.7428658597 -87.2507560454
-48.9858706935 -80.7145151474 -102.418489191 -100.935364613 -111.137569879 -81.1672655972 -77.1183135218
-78.7155292824 -32.5654708193 -99.1541081438 -94.3368211704 -44.0479551661 -92.040722183 -78.5126642071
-64.6456563809 -61.9447332424 -68.8103166939 -55.0252353535 -82.4868866464 -50.8599787412 -2.78265863793
-97.5665291597 -101.368327067 -94.5867166402 -76.0054970283 -111.599874719 1751.15204197

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

The value for R2 is 0.456

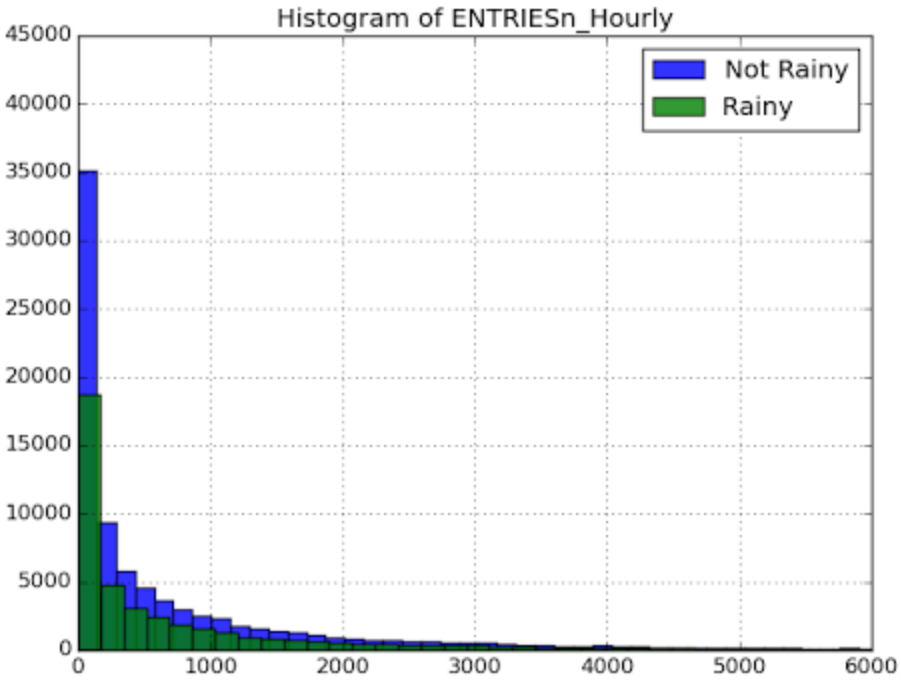
2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

The value for R2 is 0.456, this means that the model accounts for about 46% of variability in the values. Given this value, the model can give a general idea of the difference between rainy and non rainy days, but it might be possible to make better predictions with extra data (holidays, sports or other social events that might influence ridership).

Section 3

Visualization

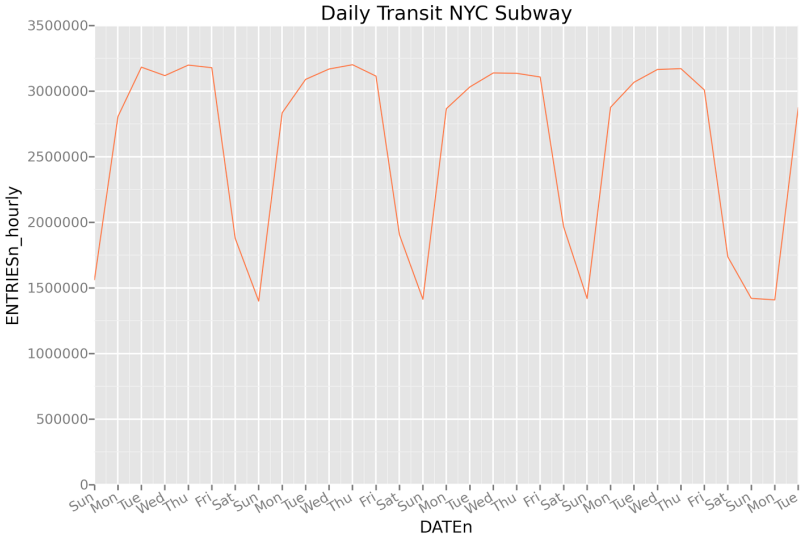
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.



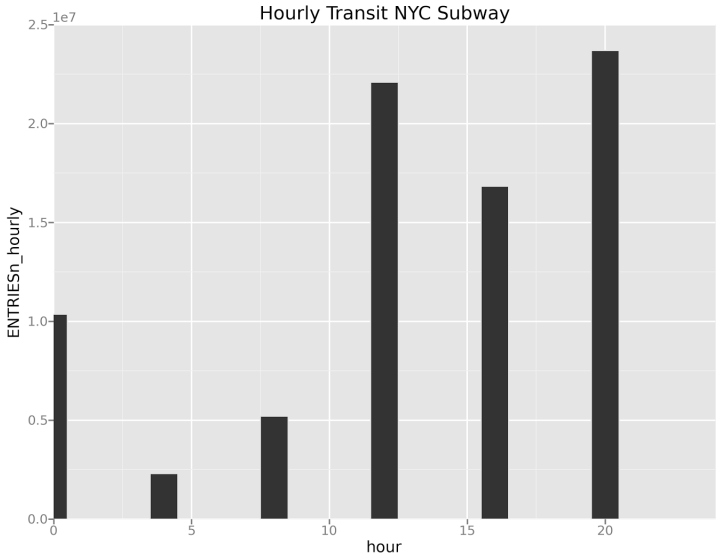
3.2 One visualization can be more freeform. Some suggestions are:

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

Plotting the number of entries per hour for a 3 week period helps in having an idea about the weekly pattern of subway usage.



Hourly ridership bar plot



Section 4

Conclusion

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?

Rain doesn't seem to 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.

The results of the tests indicate that a rainy day doesn't imply a statistical difference in ridership, since the null hypothesis has been rejected and the value for r squared indicates too much variability in the model.

Section 5

Reflection

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 can give an insight in to what is more likely to happen according to a set of meteorological data, but it doesn't take into account social events like sports, demonstrations, holiday season, tourist destinations etc. which probably influences ridership. For instance in the weekly plot it can be seen that new yorkers use the subway primarily during the week.

Analyzing the data using MW U-test to see how likely it is that the number of passengers will be larger in rainy days, is a simple but reasonable way of getting an idea about the behaviour of NYC subway users. A better score of the r_squared value is needed to support any conclusions.

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