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, and 4 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. https://www.coursera.org/learn/machine-learning Andrew Ng
2. https://www.udacity.com/course/intro-to-inferential-statistics--ud201
3. http://www.webmd.com/pain-management/features/weather_and_pain Barometric pressure related to joint pain and mood
4. https://en.wikipedia.org/wiki/It_was_a_dark_and_stormy_night
5. http://forums.udacity.com/questions/100154322/significance-of-dummy_units-and-ones-in-features-array-project-3-exercise-5?page=1&focusedAnswerId=100154584#100154584
6. https://www.youtube.com/watch?v=OdZ9c23Qa9g Use of dummy variable in regression
7. https://www.youtube.com/watch?v=gcjw8-dcxFc OLS Tutorial

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?

> Data was analyzed using the Mann-Whitney U Test. Mann-Whitney is a one-tailed test.
>
> The two data sames being compared are data observations when Clear ('rain' ==0) and Rainy ('rain' != 0) which covers any amount of precipitation. The Null Hypothesis for this comparison is that there is no difference between the

'ENTRIESn' values- that is that rain does not effect subway ridership. P-critical for this analysis is .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 Mann-Whitney U Test is appropriate for these datasets because they are not normally distributed, which is required by other test statistics. For Mann Whitney U there is no underling probability distribution assumed.

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 values of the test statistics are as follows:

```
Mean_Rainy = 1105.4463767458733, Mean_NotRainy = 1090.278780151855
U = 1924409167.0
P-Value =  0.024999912793489721
```

1.4 What is the significance and interpretation of these results?

The result of the Mann-Whitney Test shows a p-value of 2.499% chance that the Null Hypothesis explains the difference. This is less than our chosen p-critical and therefor we reject the Null Hypothesis and conclude that there is in fact a difference in ridership in rainy weather.

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. OLS using Statsmodels or Scikit Learn
2. Gradient descent using Scikit Learn
3. Or something different?

I used the gradient descent shell code from the free course.

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

The following features were used:

'rain' - a '0' or '1' indication of if there was any rain

'precipi' – the amount of rain in inches

'Hour' – the wall clock hour variable from 0..23

'meantempi' – the mean temperature

'minpressurei' - minimum barometric pressure

Reluctantly, UNITs were used as a dummy variable, converted to an indicator variable of 0,1 to try to control for the popularity of certain stations and routes.

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."

There is a negative cost to the rider associated with taking the subway both in terms of the fare, time and smell. And they forgo any health benefits and pleasure of a stroll in clement weather. Therefor I assumed in order to overcome these costs to the rider there must be conditions which warrant action. Inclement weather (rain, precipi and mean temp) would all make walking unpleasant. Low barometric pressure ('minpressurei') is associated with joint pain and lower mood, which would tend to make people disinclined to walk. 'Hour' could correlate to personal security and how safe people feel walking alone. 'Hour' might also correlate with optional trips outside the traditional 1st shift workday.

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

Theta_0 (1's) : 1392.77493601
Theta_1 ('rain') : 90.9420720293
Theta_2 ('precipi') : 630.43202429
Theta_3 ('Hour') : 639.819657623
Theta_4 ('meantempi') : 5271.24950251
Theta_5 ('minpressurei') : 3400.40523125

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

The r^2 value is 0.464700476061

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?

46.47% of the variability is explained by the chosen features which is pretty good considering all of the other factors which aren't accounted for in the dataset which could make major differences in the ridership. Were the Rangers playing at Madison Square Garden? How much of a Cinco De Mayo crowd was there out drinking? What about weekends verses weekdays? Or Mother's Day when all good offspring have left the concrete jungle and driven or flown home to the family manse, leaving the subways depleted.

Section 3. Visualization

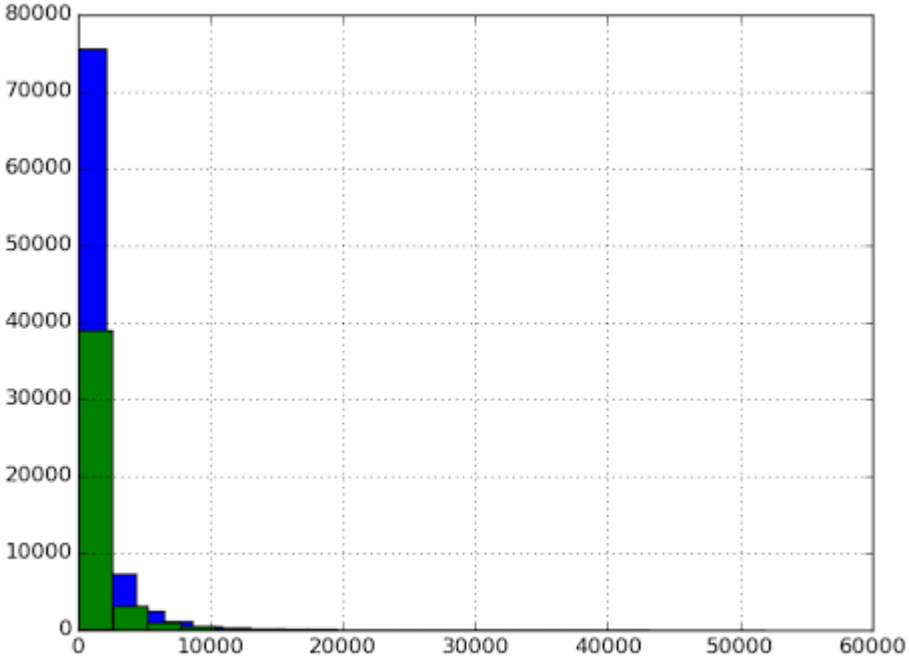
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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.



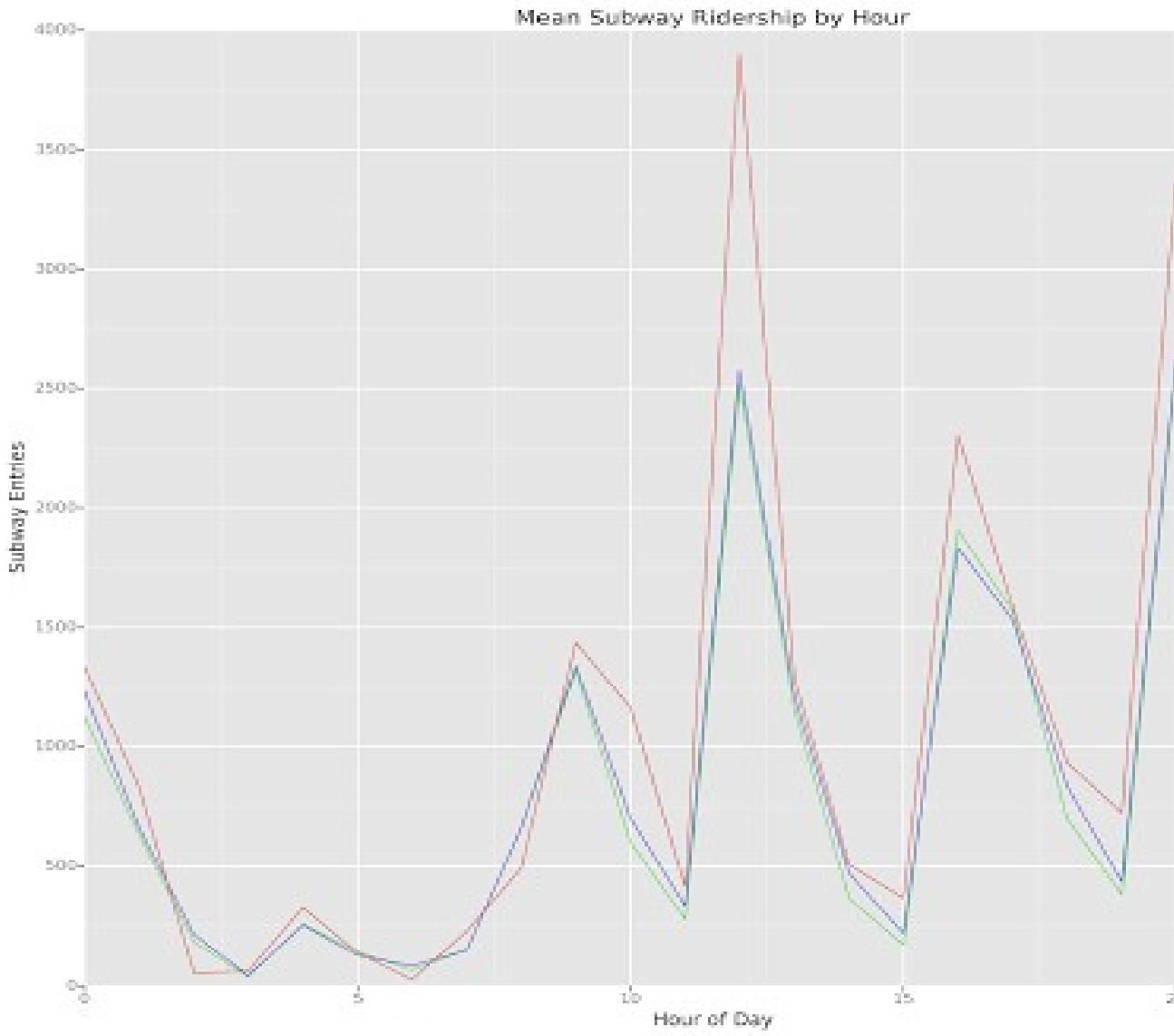
- A histogram of subway ridership when 'rain' ==1 (Green) and when 'rain' ==0 (blue). The y-axis is the ENTRIESn and the bin size is 20.

  There are more rows for clear weather than for rainy.

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

-



- A line graph of subway Mean(ENTRIESn) by normalized by hour and comparing the conditions Clear (Green), Rain (Blue), and Big Rain (Red, 'precipi'>2.0). When comparing the Mean of the entries by hour it is possible to see how the spikes of the normal workday are increased by rain. Doing Mean by Hour is somewhat frustrated by the granularity of most of the Unit observations being every 4 hours.

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?

More people do ride the subway on average when it is raining than when it is clear.

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.

From the Mann-Whitney U Test we have seen that there is in fact a statistically significant difference in the ridership dataset when it is clear and when it is raining. From the means of the two sets we can confirm that more people ride when it is raining.

From the Linear Regression we have seen that it is possible to predict a significant portion, 46%, of this increase from the weather related variables rain, meantemp, minpressi, precipi and the Hour of the day.

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 in this exercise is somewhat problematic. For instance Unit R001 reports every 4 hours, or 6 observations per day (1:00, 5:00, 9:00, 13:00, 17:00, 21:00). With a granularity of 4 hours are we actually capturing the current weather conditions? Since the bad weather presumably doesnt start and end neatly on the reporting interval the ENTRIESn will contain information for both kinds of weather. While there are many observation in the dataset, there are relatively few days, so there is no reason to think this will randomize itself out. For the full dataset, it's a scant N==10 for the rain==1.

Some other Units report with the same frequency but on a different schedule. For instance Unit R004 reports 6 times a day at 0:00, 4:00, 8:00, 12:00, 16:00, 20:00 which is a different schedule and splits morning rush hour into two rows. Worse, Unit R552 reports 125 times a day at seemingly random intervals, some times only a minute or two apart. Get your act together R552! The difference in granularity and timing of observations introduces difficulties in discerning the trend caused by the weather from the other existent trends in subway ridership, such as the start of the work day or bar closing time.

It is quite easy, for me at least, to believe that the effect of precepi on ridership is nonlinear. A little sun shower (percipi = 0.07); thats not much to worry about. But as the rain increase to torrents it is easy to imagine that the growth would

be exponential. Sadly, I didnt have the time or data to test this.

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

It was a dark and stormy night; the rain fell in torrents — except at occasional intervals, when it was checked by a violent gust of wind which swept up the streets (for it is in Manhattan that our scene lies), rattling along the housetops, and fiercely agitating the commuters who struggled against the darkness. Through one of the obscurest quarters of Gotham, and among haunts little loved by the gentlemen of the police, a man, evidently of the lowest orders, was wending his solitary way.

Who is our anonymous traveler? Perhaps off to work in the morning. Work is a trip which can not easily put off, unless the weather is completely forbidding- then he may call in sick. But today is just a normal work day and he must show up for work at National Insurance Company, or as a pizza flipper or at the Water Works. How to get to work- walk and breath lung fulls of fresh air? Or take a cab? Or bus? Or drive? Or maybe the subway. Walking is free, enjoyable and healthy. The other options come with a cost, but could be faster. And also drier- for it is raining today and our anonymous traveler doesnt want to get to work soaked from walking or waiting for a bus. And so R037 ticks over and registers 3525 entries instead of 3524 on 5/18/2011 at 10:00 when 'rain' == 1, 'minpressurei' is at 29.93 millibars, and 'precipi'= 2.18. The Rain has increased 'ENTRIESn'.

Perhaps after work our anonymous commuter will think differently. Traveling home and changing out of wet clothes isnt as big a deal as having to wear them all day. In that case the theta value for percipi will not be as predictive.

'Hour' == 9pm and all is well! Or not so, well as our commuter observes sourly looking out his brownstone's window- the weather isnt fit for a dog out there. He was planning on “making the scene” (as the kids say) at a popular bar where scenes are regularly made. But that is across the city and the weather is so forbidding that our protagonist puts on his slippers and takes to bed early with an improving book. And hence the 'rain' == 1 has suppressed one ENTRIESn for his trip out to the bars and one ENTRIESn on UNIT == R107 for him staggering home. In his case 'rain' == 1 is sometimes negatively correlated with subway usage.

What are all the possible situations? What are all the possible motivations and calculations that happen for every traveler before a turnstile clicks over? It seems a shame to waste 468 thetas on a UNIT dummy variable. There are a million stories in the naked city, this was just one of them.