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.

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?

The Mann-Whitney U-Test.

The reported p-value is for a one-sided hypothesis in scipy so effectively one-tail is used. But really, to get the two-sided p-value we can just multiply the returned p-value by 2. (reference: http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html#scipy.stats.mannwhitneyu).

The class stated that the null hypothesis asserts that the two populations are the same (in a sense this can mean the mean ranks, or the medians of the two samples are identical).

The calculated one-tail p-value is 0.02499991.

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 does not assume a distribution. Typically if the distributions of the data measured are normal, we can use Welch's Two-Sample t-Test. But by observation (and probably by normality testing), they are not.

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.

Mean for raining: 1105.4463767458733
Mean for non-raining: 1090.278780151855
p-value: 0.024999912793489721
Since p-value is <0.05, we reject the null hypothesis.

1.4 What is the significance and interpretation of these results?

More number of people entries in raining. More people ride NYC Subway then.

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?

Gradient descent was used for calculating theta in exercise 3.5. It is linear regression producing the prediction. I used OLS in exercise 3.8.

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

I simply used the data rain and hour. I originally thought temperature and precipitation might contribute to the prediction but the R square gets worse by adding them. And I tried a couple other combinations as well and it didn't produce good results.

I did use dummy variable "unit". It is quite obvious that utilization of different units (stations) is different.

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."

As explained above, I used unit, rain, and hour only and they automatically get the best results (although I haven't exhausted all combinations). Unit is obvious because some locations are more used than the others. Hour is obvious because there are busy hours and non-busy hours. They explain largely for the Y but these are intuitive and less interesting. The rain makes sense because when it is raining, not only it is less convenient and comfortable to use commute on surface, but it is easier to get into traffic jam. Taking subway in rains is traditional wisdom.

I did some exploration and experimentation, with intuition, say temperature and precipitation, etc. But they did not improve R2.

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

Here are the results:
rain: 64.687935
Hour: 62.284569

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

“Your R^2 value is: 0.48239847426” is shown in scipy.

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?

Unit, hour, rain combined explain 48.2% of the variation in entries. Not too bad for this dataset, because perhaps we might/should have missed something even more important – the fare, the introduction of incentive programs, the convenience of the locations, the competition in the areas (e.g. availability of cab, carparks), etc.

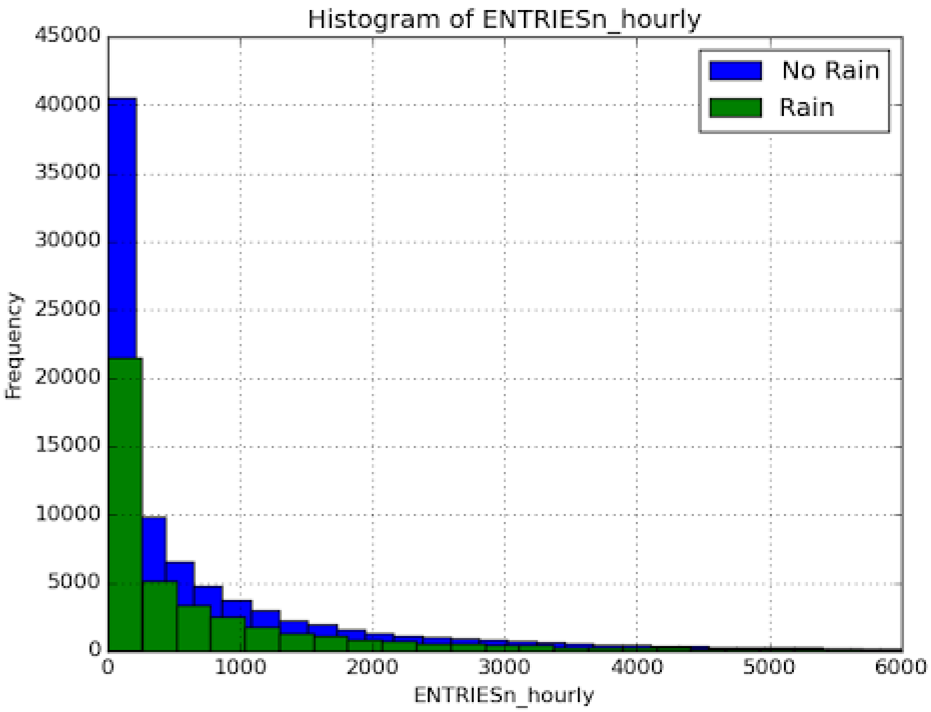
Section 3. Visualization

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.



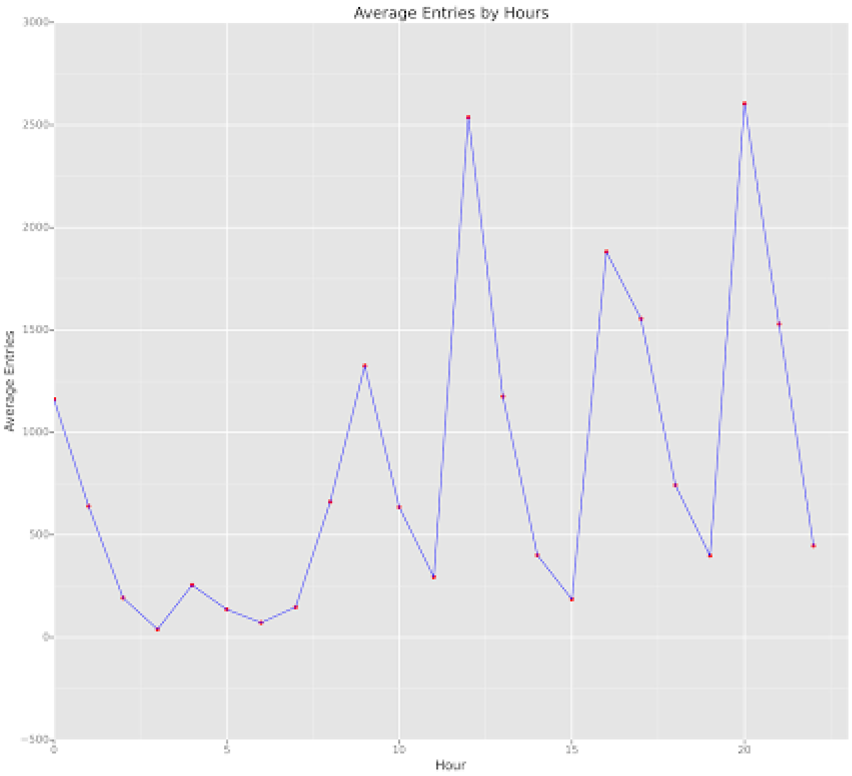
Obviously the distributions are not normal so the Welch’s Two-Sample t-Test cannot be used to compare frequencies under rain and no rain.
Although frequency of no rain seems larger, it is simply because we have more data for no rain hours (no rain hours are more than rain hours). We still need to use statistical testing to determine whether subway is more used in rainy hours.

Code:

```
    plt.figure()
    rain=turnstile_weather[turnstile_weather['rain']==1] # your code here to plot a historgram for hourly entries when it is raining
    notrain=turnstile_weather[turnstile_weather['rain']==0] # your code here to plot a historgram for hourly entries when it is not raining
    notrain['ENTRIESn_hourly'].hist(bins=200, label='No Rain')
    rain['ENTRIESn_hourly'].hist(bins=200, label='Rain')
    plt.xlim(0, 6000)
    plt.title("Histogram of ENTRIESn_hourly")
    plt.xlabel("ENTRIESn_hourly")
    plt.ylabel("Frequency")
    plt.legend()
    return plt
```

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



This is the line-plot for average ridership by time-of-day. It tells us that the average peaked at hour 12 and hour 20, and have low entries from hour 2 to 7. So people start going to office on hour 8 if they take subway.

Code:

```
    avg_hour=[]

    hour=[]

    for i in range(0,23):

        hour_df=turnstile_weather[turnstile_weather['Hour']==i]

        avg_hour.append(numpy.mean(hour_df["ENTRIESn_hourly"]))

        hour.append(i)

    ridership = {'Hour':Series(hour), 'Average Entries': Series(avg_hour)}

    df = DataFrame(ridership)

    plot = ggplot(df, aes("Hour", "Average Entries")) + geom_point(color='red') + geom_line(color='blue') + ggtitle('Average Entries by Hours') + xlab('Hour') + ylab('Average Entries') + xlim(0,23):
```

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?

Yes. 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.

From the Mann-Whitney U-Test, it suggests that the probability of producing the results assuming the null hypothesis is low. That means the median (or mean rank in this exercise's context) for entries when raining should be higher.

From OLS, except the obvious feature like unit and hours, rain is the only feature which improves the R2 value. We should believe that rain does contribute to the prediction model of entries.

So overall, more people do ride the NYC subway when it is raining, according to our predictions.

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,

As discussed, some critical data which might be important can be missing. It is either we do not have the data, or data collection is impossible, or we just bother to collect convenient data. So it leads to another shortcoming for such analysis – we simply analyze whatever we have on hand. This can be misleading.

Also, some qualitative data can be subjective. How to define the hour is raining? Where is the line cutting raining and not raining? Condition is another one. These data might be intuitive for everyone, but there can be one definition in the model but definitions can be slightly different from person to person, which makes the interpretation of the model judgmental to different people.

2. Analysis, such as the linear regression model or statistical test.

The major limitation of statistical testing is obviously that correlation does not imply causation. We found raining drives people use more subway, is it the direct cause? Usually we need deeper understanding of the results. E.g. I mentioned raining might make people to use subway to avoid traffic jam, so maybe in some locations without traffic jam, even in rainy hours, would it increase the subway usage? If indeed it does not increase, we might make a wrong prediction for those locations.

The major limitation of linear regression model can be omitted data, overfitting/underfitting, based on a wrong model (e.g. may not be a linear model), wrong assumptions (e.g. data are not independent), etc.

explained in the course. Specific to our model / method, maybe the data matters more. We do not know if there are obvious outliers too if we did not scan them explicitly beforehand.

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