All questions must be answered in order for your project to meet specifications. ​Please include full answers to the following questions.

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.

Sorry, I did neglect the p-value, but the means were in the original submission. I called them out more clearly in this version.

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model? ​These are the ‘theta’ values from your gradient descent.

I added these.

Section 3: Please add a short description below each figure commenting on the key insights depicted in the figure.

I added these

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?

As stated in the original version and again in this one, my analysis does not support the claim that subway ridership changes when it rains.

The communication at the level of sentences and explanations is very good! However, there have been adjustments made to the dataset that are not revealed to the reader, making it difficult to follow. For example, according to the code, the histogram plotted refers to only a small subset of the UNITs - this should be explained in the text.

I explained in words what I had done to the data in the original submission. To clarify further in this submission, I inserted screen captures of the data before and after the change. I corrected the histogram to use all of the UNITs.

● The answers are not well-focused (e.g. stream of consciousness) or leave out important information (e.g. not fully answering the question). Comment: The project is very long: 21 pages including code. In most industries, a data analyst will be required to write short summaries of a detailed analysis. A long report might simply be discarded by busy colleagues so sensible editing is an important skill to learn. Please resubmit a project that fits within the Short Question format. Many successful submissions are less than 5 pages long.

This is not a valid comment. The text of my submission (not including code and plots) was approximately five pages and is about the same in this submission. Perhaps I should have made a clearer separation between the document’s answers and the associated code and figures, but this was not clearly stated as a requirement for the project. Udacity needs to do a better job explaining what is expected in the project submission.

Quality of Visualizations Does Not Meet Specifications ● Plots depict relationships between two or more variables. ● All plots are of the appropriate type. ● Some plots are not appropriately labeled and titled or visual cues are not always easy to distinguish. It is not clear what data are represented. Comment: Please make sure that the details in your titles and labels clearly describe what is shown in each plot. In particular, it seems that some histogram plots show total ENTRIESn\_hourly per day for the whole subway system whereas others look at ENTRIESn\_hourly per hour, but for only a few UNITs. I recommend that in the rain/no-rain histogram you plot the frequency of ENTRIESn\_hourly values for all of the UNITs, rather than a small subset. You should also highlight your use of a logarithmic vertical scale on one of the plots. Please improve the way that the date is represented in the 'scatter' type plot of "Total Riders" by day of the week. This plot also requires a title. Please include a short description below each figure commenting on the key insights depicted in the figure.

Thanks, I fixed most of these issues. The dates on the scatter plot are correct, and now it has a caption.

Quality of Analysis Does Not Meet Specifications ● The choice of statistical test type, features, and linear regression models are sometimes not appropriate based on the characteristics of the data. Comment: Well done for choosing a two-tailed test Mann-Whitney U-test! Removing 'outlier UNITs', Saturdays, Sundays and Mondays and some other readings from the linear regression model is not appropriate here. The purpose of the linear regression model is to predict the value of ENTRIESn\_hourly for and UNIT on any day of the week. Please construct a model that can be used to predict the value of ENTRIESn\_hourly across the whole dataset. The two things that we want to see from a linear regression model are explanation (‘What drives ENTRIESn\_hourly?’) and prediction (‘What will ENTRIESn\_hourly be on 27th May 2015?’). It seems to me that by excluding outliers you’ve selected the sorts of points that will be easiest to predict. Doing a good job at predicting those is definitely a useful aspect of a model, but it does not do well at explaining the overall patterns in ENTRIESn\_hourly. Perhaps these ‘outliers’ are actually the most important, popular stations? If so, it might be especially important to understand what drives ENTRIESn\_hourly at these UNITs. It is a nice idea to use the value of ENTRIESn\_hourly from a previous reading, this is not very useful for prediction in the future. (One could consider a model that takes multiple 'steps forward' into the future, but this would not work with the model you constructed, as several days of the week are omitted.) It's not clear from your report how useful this feature actually is: perhaps a lot of this information could be found from more general features such as the UNIT, the time of day or the day of the week. This would make the model far more usable, but may require some sacrifice in terms of R^2 value in order to produce a more general model.

I removed the outlier detection code, and use the full dataset. I have retained the 24-hour-lagged turnstile entries as a feature of my model as I have shown it is effective and demonstrates exactly the type of insightful data analysis Udacity should be encouraging in the Data Science Nanodegree program. The grader’s comment that “this is not very useful for prediction” is unsupported- I have shown how it increases the R-sq value. I also added illustrations of the dataset before and after my change in case it was not clear enough initially. If the grader does not agree with my approach, I respectfully request that my project be assessed by another grader.

● Statistical tests and linear regression models are not described thoroughly, or the reasons for choosing them are articulated clearly. Comment: You should state that you are using daily UNIT totals of entries to perform your statistical test, particularly because this differs from the analysis in the course. Why did you decide to do this?

I explained this in more detail within the report

● The use and interpretation of statistical techniques are correct. Comment: Well done - you’ve correctly used your statistical test. Your detailed analysis of the linear regression model is very good; you did well to identify that the residuals are not normally distributed here. ● Some conclusions are not correctly justified with data. Comment: Please summarise the results from your statistical test and descriptive statistics in Section 4 in order to back up your conclusion. You should note that the existence of an effective regression model without rain does not imply that rain would not be a useful feature in a linear regression model. I recommend that you test a model including a ‘rain’ feature in order to investigate this further. ● No incorrect conclusions are drawn from the data. ● Shortcomings of the statistical tests or regression techniques used are not appropriately acknowledged.