

## Project Overview

In this project, you look at the NYC Subway data and figure out if more people ride the subway when it is raining versus when it is not raining.

You will wrangle the NYC subway data, use statistical methods and data visualization to draw an interesting conclusion about the subway dataset that you've analyzed.

Prepare for this project with: [Intro to Data Science](https://www.udacity.com/course/viewer#!/c-ud359-nd).

## Note

If you have successfully completed the project for the Intro to Data Science course in the past (which entails having graduated from the course and having access to your course certificate), simply email us at dataanalyst-project@udacity.com with your passing evaluation and we'll give you credit for this project.

## What do I Need to Install? (optional)

If you want to complete the programming exercises on your own computer or laptop, you will need to install[Anaconda Scientific Python Distribution](https://store.continuum.io/cshop/anaconda/).

It should contain most of the libraries and packages that you need to work on the assignments.

One caveat is that Anaconda does not include pandasql (needed to complete project #2), but you can easily install pandasql through pip as below:

pip install -U pandasql

Also, if you would like to test the programming exercises on your own computer, we have provided two data sets for you to download.

1. The first version is the version that you have been working with throughout the class.
2. The second version contains extra data points and variables that you can use to improve your linear regression model and visualizations.

## Why this Project?

This project will introduce you to the key concepts of data science, so you will be prepared for subsequent projects in the Data Analyst nanodegree as well as your future career as a data analyst.

In addition, you will be exposed to some of the most popular data science libraries in python, such as Pandas, Numpy, and others.

## What will I learn?

You will be exposed to and learn fundamental data science skills like:

* data wrangling
* applied statistics and machine learning
* effective visualization

## Why is this Important to my Career?

By completing this project, you will have exhibited all of the skills needed to be a data analyst. In addition, you can add this project to your portfolio, which can help you impress recruiters and hiring managers.

## How do I Complete this Project?

This project is connected to the [Intro to Data Science](https://www.udacity.com/course/viewer#!/c-ud359-nd) course, but depending on your background knowledge of data science, you may not need to take the whole thing to complete this project.

If you would like, you can download the data set used for the Intro to Data Science course and explore Problem Sets 2 to 4 independently on your own computer. The download links are below:

* [Original data set](https://www.dropbox.com/s/meyki2wl9xfa7yk/turnstile_data_master_with_weather.csv) - This is the original version of the data set that we've used throughout the Intro to Data Science course.
* [Improved data set](https://www.dropbox.com/s/1lpoeh2w6px4diu/improved-dataset.zip?dl=0) - this version contains a cleaned-up subset of the original dataset with additional variables that you can use to improve your linear regression model and visualizations. The additional variables can be seen in [this document](https://s3.amazonaws.com/uploads.hipchat.com/23756/665149/05bgLZqSsMycnkg/turnstile-weather-variables.pdf).

Here's what you should do:

1. Complete all of the questions in Problem Sets 2 through 4 in the [Intro to Data Science](https://www.udacity.com/course/viewer#!/c-ud359-nd) course.
2. Answer these [short questions](https://docs.google.com/document/d/16T3kirC0IxvtfxlZb7n5kOz5xFF_JTwrG31J2OZj8KM/pub?embedded=True) in a pdf or html document. Please do not use doc or docx format because there are compatibility issues across browsers. If you are using a word processing program such as Microsoft Word or LibreOffice, once you are done, save the file as pdf and include it in your submission.

## Evaluation

A Udacity evaluator will review and check your completion of problem sets 2 through 4 in Intro to Data Science, as well as your answers to the questions listed in the Short Answers document.

Be sure to complete all of questions in the problem sets and in the project document before submitting your project.

## Rubric

Your project will be evaluated by a Udacity reviewer according to this [project rubric](https://docs.google.com/a/knowlabs.com/document/d/1ZWdmlEgtRhreyN7AaiEfoYP70GqxOZqrajWtzIov8HM/pub?embedded=True). Be sure to review it thoroughly before you submit. Your "project meets specifications" if it meets specifications in all the criteria.

# **Analyzing the NYC Subway Dataset Short Answer Rubric**

## Overview

This rubric is here to help you understand the expectations for how your project will be evaluated. It is the same rubric that the person evaluating your project will use. You should look at the rubric **before you begin working** on this project **and before you submit it**.

## 

## Minimum Requirements

Your submission will not meet specifications if you do not answer all of the short questions.

## How Grading Works:

* Your answers to the following short questions will be evaluated against the answer rubric
* Your grade will simply be “pass” or “doesn’t pass.”
* You earn a “pass” by not having **any** criteria items in the “does not meet specifications” column.
* If any criteria item “does not meet specifications,” you will not pass. However, you will be able to make changes and re-submit.

## The Rubric

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Does Not Meet Specifications** | **Meets Specifications** |
| Communication |  |  |
| **Writing is appropriate for the intended audience (Intro DS Classmates).** | Analysis done using methods learned in the course is poorly explained and would not be understandable to a student who has completed the class. | Analysis done using methods learned in the course is explained in a way that would be understandable to a student who has completed the class. |
| **The answers provide a readable summary of the analyses** | The answers are not well-focused (e.g. stream of consciousness) or leave out important information (e.g. not fully answering the question). | The answers are a well-formed summary of the analyses. |
| **Quality of Visualizations** |  |  |
| **Plots shows relationships between two more or variables** | Plots do not depict relationship between two or more variables | Plots depict relationships between two or more variables |
| **Plot type (bar plot, line plot, scatter plots, histograms, etc…) and data represented are appropriate for the analysis.** | Not all plots and data are of the appropriate type. | All plots and data are of the appropriate type |
| **Plots are easy to read.** | 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. | All plots are appropriately labeled and titled. Plot is given an appropriate title. X-axis and y-axis are appropriately labeled. Visual cues (colors, size, etc) are easy to distinguish. It is clear what data are represented. |
| **Quality of Analysis** |  |  |
| **Choice of statistical tests and linear regression models are appropriate** | The choice of statistical test type, features, and linear regression models are sometimes not appropriate based on the characteristics of the data. | When using statistical tests and linear regression models, the choice of test type and features are always well justified based on the characteristics of the data. |
| **Statistical tests and linear regression models are described and justified appropriately.** | Statistical tests or linear regression models are not described thoroughly, or the reasons for choosing them are not clearly articulated. | Statistical tests and linear regression models are described thoroughly, and the reasons for choosing them are articulated clearly. |
| **Analyses are carried out correctly** | Mistakes are made in use or interpretation of statistical techniques | The use and interpretation of statistical techniques are correct |
| **Conclusions are correctly justified with data.** | Some conclusions are not correctly justified with data. | All conclusions are correctly justified with data. |
| **Conclusions are correct.** | Some of the conclusions drawn are incorrect. | No incorrect conclusions are drawn from the data. |
| **Potential shortcomings of the dataset and analysis are addressed.** | Shortcomings of the dataset and statistical tests or regression techniques used are not appropriately acknowledged. | Some shortcomings of the dataset and statistical tests or regression techniques used are appropriately acknowledged. |

## Submission

Ready to submit your project? Collect the following files:

1. Answers to the short questions in a pdf or an html document.
2. Optional: Code used to solve Problem Sets 2-4 or additional code used to answer the short questions.
3. A list of Web sites, books, forums, blog posts, github repositories etc that you referred to or used in this submission (Add N/A if you did not use such resources).

Then go back to your Udacity Home, click on the project, and follow the instructions to submit:

* If you want to submit your files through a "Link to Project", upload your project files onto Github and send us the link.
* If you instead want to submit your files through "Upload a Zip", compress your project directory, and submit that zip file.

It can take us up to 2 weeks to grade the project so keep checking back for updates.

If you are having any problems submitting your project or wish to check on the status of your submission, please email us at dataanalyst-project@udacity.com.

========================================================================

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.

# **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?

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.

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.

1.4 What is the significance and interpretation of these results?

# **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:

* OLS using Statsmodels or Scikit Learn
* Gradient descent using Scikit Learn
* Or something different?

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

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.”

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

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

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?

# **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.

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

# **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?

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.

# 

# **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:

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

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