Project 2v1.0: Analyzing the NYC Subway Dataset

Is this project for you?

This document is intended for use by Data Analyst Nanodegree Program students who submit the Project 2v1.0, Analyzing the NYC Subway Dataset, **before February 1, 2016.** *If it is February 1, 2016, or later, you must complete the Project 2v2.0 project: Investigate a Dataset.

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.

How is this project different than the new one?

Checkout this doc to see what's new. You can decide which project you'd like to complete as long as you submit Analyzing the NYC Subway Dataset before February 1, 2016 if you choose to complete it.

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.

Supporting Course

Prepare for this project with: Intro to Data Science.

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.

How Do I Complete This Project?

Step One:

Complete all of the questions in Problem Sets 2 through 4 in the Intro to Data Science course.

Optional Installation

If you want to complete the programming exercises on your own computer or laptop, you will need to install Anaconda Scientific Python Distribution. 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
```

We have provided two data sets for you to download and explore Problem Sets 2 to 4 independently. The download links are below:

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.
    a. The additional variables can be seen in this document.

Step Two:

Answer these short questions 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.*

Rubric

Your project will be evaluated by a Udacity reviewer according to this project rubric. Be sure to review it thoroughly before you submit. Your "project meets specifications" if it meets specifications in each section of the rubric.

Ready to Submit?

Step One:

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).

Step Two:

Submit your work HERE.

- 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.

All new submissions of Project 2v1.0, Analyzing the NYC Subway Dataset must be before February 1, 2016. **If you have submitted by this date, you can resubmit your project using the link above until you meet expectations**. If it is February 1, 2016, or later and you have not yet submitted this project for the first time, *you must complete* Project 2v2.0: Investigate a Dataset.

Troubleshooting

Can't manage to submit your project? No worries! If you have any questions, please email us at dataanalyst-project@udacity.com or visit us in the discussion forums.

What's Next?

You will get an email as soon as your reviewer has feedback for you. This usually happens in about 24 hours, but can sometimes be up to 1 week. In the meantime, connect with your learning community and consider sharing your work in the discussion forum and keep this momentum by moving on to the next project!