Analyzing the NYC Subway 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.

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.

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.

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
- How to work with big data using MapReduce.

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 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 Set 2 to 5 independently on your own computer. The download links are below:

- Original data set - This is the original version of the data set that we've used throughout the Intro to Data Science course.
- Improved data set- this version contains extra data points and variables that you can use to improve your linear regression model and visualizations. The additional data points and variables can be seen in this document.

Here's what you should do:

1. Complete all of the questions in Problem Sets 2 through 5 in the Intro to Data Science course

2. 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.

Evaluation

A Udacity evaluator will review and check your completion of problem sets 2 through 5 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. Be sure to review it thoroughly before you submit. Your "project meets specifications" if it meets specifications in all the criteria

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 2-5 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 the portal, 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.