Milestone 3: Data Preprocessing/Data Cleaning
Deadline: April 4th 11:59 pm EDT

Word from Us Content Creators:

Hello Everyone! We hope you are doing well. During the previous milestone, your team has gotten a chance to explore your datasets. We enjoyed reading your analysis of what was uncovered in your data. This milestone will focus on data preprocessing. A large portion of time in data science is spent in the preprocessing stage.

What it Takes:

When it comes to learning, persistence is key. In a real world project, change is inevitable. Strong teams adapt to change. As long as you keep trying and are willing to ask for help, you will be in good shape :)

Objective:

During Milestone 2, your team explored your chosen datasets for the project. However, in the real world, datasets are typically “dirty” in the sense that you will not always have data in the format you want. Data Preprocessing takes up the bulk of the data science project lifecycle. Data preprocessing allows you to convert raw data into a usable format for modeling. Some techniques used include, but are not limited to, missing value imputation, feature engineering, normalization, and standardization. Your team will be responsible for using techniques and concepts covered from Workshop 3 of bootcamp and beyond to preprocess your datasets and get the data into a usable format. It is **highly recommended** to think critically about why you are performing particular preprocessing steps on your dataset.

Deliverables:

The two major deliverables are the code and a writeup.

1. **Code:** Each team should have already made a **Github Repository titled with your team name** (eg: If your team name is “Team Buzz”, your Github Repository should be named “team-buzz”). Your repository should contain the code you used to preprocess your dataset (preferably an interactive Python notebook such as Jupyter Notebooks or Google Colaboratory since output would be displayed in the same environment as your code). If you would like to put a Google Colaboratory notebook on Github, you can

download the .ipynb file from Google Colaboratory (File → download .ipynb) and upload it to your team's Github repository.

2. **Writeup:** The writeup should consist of an explanation on the preprocessing steps taken to "clean your dataset" (1-2 pages). Some questions to consider in your writeup include (but not limited to):
    a. How was your dataset preprocessed? What were the predominant techniques used and why?
    b. Was there any new insight uncovered from your preprocessing steps that differed from what your team observed from Milestone 2?

The writeup is to be **uploaded as a document** under your **team folder** in Google Drive. Make sure that your team creates this folder **titled with your team name** (eg: if your team name is "Team Buzz", your folder should be named "team-buzz"). Your team should also provide a **link to your Github Repository** in your writeup document and add your assigned mentor as **collaborator**. Teams are also allowed to upload their writeup in their Github Repository.

Link to set of team folders:
https://drive.google.com/drive/folders/1QR-1N5f9JG74hSo1IyfXsUfN27og4rwU?usp=sharing

Tutorial on setting up a Github Repository:
https://drive.google.com/file/d/1GrAFpkprT-6ZftIneSXBHaULSpRHT_Lv/view?usp=sharing

Link to Install Git: https://git-scm.com/downloads

Basic Tutorial on Exploratory Data Analysis with the Kaggle Housing dataset (inspiration for some ideas):
https://www.kaggle.com/spscientist/a-simple-tutorial-on-exploratory-data-analysis

Introduction to Data Preprocessing Article:
https://towardsdatascience.com/the-complete-beginners-guide-to-data-cleaning-and-preprocessing-2070b7d4c6d