DS 1.1: Data Analysis & Visualization
Project Guidelines: Summer Academy NPS Data

Project Timeline

- Description & data given: Thursday, April 9
- Data cleaning check-in: Tue, April 14
- In-class presentations: Tuesday, April 21

Project Summary

For your first major project of DS 1.1, you'll investigate real-world data from feedback surveys completed during Make School's very own Summer Academy program! Completing this project will you'll strengthen your understanding of:

- The overall Data Science process (define, measure, analyze, improve, control)
- Aggregating datasets from multiple files, locations, and types
- The importance of scripting and automating data preprocessing
- Transforming data so that it has the same scale and data type
- Best practices for investigating data and asking interesting questions
- Data Visualization strategies
- Distilling findings down into small, understandable, non-technical (!) presentations

Description of Problem

Clean and investigate Make School NPS data to find interesting and actionable trends that help inform decision-makers. Create a presentation in a Jupyter Notebook using data visualizations and other techniques that allow non-technical team members to understand your findings.

Background on NPS

Every summer, Make School welcomes hundreds of students into the Summer Academy to study software development and build cool stuff. The management wants to make sure that students continue to be satisfied with their experience as the program scales. The main way we measure this is through **Net Promoter Score (NPS)**, which is a tool commonly used to measure customer loyalty and promotion. You've seen NPS before if you've been asked a question like:

"On a scale of 1 to 10, how likely are you to recommend [X] to a friend or colleague?"

NPS segments all responses between 1 and 10 into three categories based on their sentiment:

- Promoter (9 – 10)
- Passive (7 – 8)
- Detractor (1 – 6)

To calculate NPS, companies follow these steps:

1. Segment all responses into Promoter, Passive, and Detractor categories.

2. Calculate the percentage of responses in each category out of the total number of responses to the survey.
3. Subtract the Detractors percentage from the Promoters percentage. This is the NPS.

In other words, NPS can be calculated with this equation:

$$NPS = (Promoters - Detractors) \div (Promoters + Passives + Detractors)$$

NPS can range from –100 (if everyone is a detractor) to +100 (if everyone is a promoter).

For more detailed information on NPS, read this article.

Background on Dataset

Make School's Summer Academy typically lasts for 8 weeks, although this can vary by location. Every week, students are given a survey asking their satisfaction with the program in the form of an NPS question (see example question in "Background on NPS" section).

You have been given data from Summer Academy in 2016 (optional) and 2017 (mandatory). Download the data here. We will do an initial data investigation during class to get a feel for how the data is structured.

You should create a **data dictionary** that describes what each column's possible values mean.

Questions to Consider Answering

In this scenario, you've just been given access to this data from your boss, with the instructions to *"See if you can find anything in here that can help the business."* – This is a very broad set of instructions. In order to complete this task well, you may want to consider finding answers to the following questions:

- How many more promoters are there than detractors across our 2017 data?
- Which track boasts the best promoter-to-detractor ratio?
- Does the student experience get better the longer that they are enrolled at the Summer Academy?
- Does student satisfaction vary by location?
- What are things we could find here that could "help the business"?
- What sorts of information does this dataset contain?
- What kinds of questions might we be able to answer with this data?
- What kinds of questions *can't* we answer with this data?
- What sorts of information might be *actionable?*
- How can you present your findings in a way that non-technical employees can understand and use to make decisions?

Data Wrangling Issues to Consider

- CSV files may have header rows
- Collating data from multiple sources
- Introducing new columns/attributes
- Converting data types (string values to integers)
- Converting categorical values (either to integers or one-hot encoding)
- Normalizing values based on different units