TD Data Synthesis Challenge

Introduction#

The world is full of data which can be found everywhere and anywhere you look. Fortunately, the internet makes it very easy to access all kinds of data that span across a plethora of different industries. Your objective for this challenge is to synthesize data into a large and unique dataset that can solve a real world problem!

Description#

This challenge involves you to use your data gathering skills to create your very own dataset. There are essentially no boundaries or limits as to what you can choose to include in your dataset. Supplemented with your dataset will be a data visualization as well as a write-up. There are a few different ways you can achieve gathering your data: web scraping, web crawling, and surveying.

Web scraping means extracting data from websites/webpages. The data you extract can be formatted into a new file like an excel spreadsheet. An example of web scraping could be a company extracting information about televisions on Amazon in order to figure out how to position their new product in the market. A key aspect of web scraping is that its done in a focused approach; that is, you'll usually be trying to extract specific information from the website. Web crawling means using bots to read and store all of the content on a website for indexing purposes. These web crawlers will usually go through every single page on a website, as opposed to web scrapers which will focus on a specific set of data on a website. Lastly, we have surverying which is as simple as it sounds! It is the act of examining a process or questioning a sample of individuals to obtain data. Feel free to extract your data by asking your peers in the Datathon!

There will also be a data visualization that accompanies your dataset, which should be informative and aesthetically organized. You can get as creative as you want with the data visualization, but more details about what we're looking for specifically can be found in the judging section below.

Finally, the write-up should showcase your journey as well as your visualization. Your journey should include how you approached the problem, what you decided to gather, your team's thought process, etc. You should include what your data visualization means, how it could be used, how it could be interpreted, etc. At the end, you could include a section talking about how you would alter/change your approach to this problem if your team was given more than 24 hours.

Submission#

In your submission, you should include a link to the GitHub repository that contains all the code used to generate the dataset.

Judging#

The judging criteria for this challenge will be broken down into 3 main components:

The Dataset:#

- Size
  - Number of rows and columns in the dataset.
  - The cleanliness of the dataset.
- Uniqueness
  - Make each feature as unique as possible.
  - Try to have unique data points as well.
  - Fresh collection that hasn't been seen already on Kaggle or other websites.
- Usefulness
  - Applicability to different industries.
  - Ability to solve real world problems.

The Visualization:#

- Insight
  - The visualization should be informative and interesting.
  - Showcases connections between variables that are difficult to describe with words alone.
- Presentation
  - Everything on the visualization should be clear and not cluttered.
  - Appropiate coloring.

Write-Up:#

- Process
  - Describes team's approach to problem.
  - Includes

- Visualization Interpretability
  - Explains what it means/portrays.
- Reflection (optional)
  - What gave your team the most problems during this challenge?
  - What would your team do differently if you were given more than 24 hours?

Prizes#

1st Place: Apple iPad

2nd Place: Nintendo Switch Lite

3rd Place: Apple AirPods