# CS418 – Lab 2 – Exploratory Data Analysis

This lab is an individual assignment. You should NOT work with your project teammate nor any other classmate. If you have questions, please ask us on Piazza.

For this Lab you will analyze a superheroes dataset to summarize their main characteristics and information.

## The Superhero Characteristics and Powers Dataset

These datasets include basic information for over 700 superheroes (and villains). The first dataset, heroes_information.csv, provides demographic characteristics such as gender, race, comic publisher, etc., while the second dataset, super_hero_powers.csv, maps out the powers for each superhero by assigning Boolean (true/false) values for 168 different superpowers.[1]

- heroes_information.csv (45 KB)

Columns in the heroes_information dataset are as follows:

| **Column** | **Description** |
|---|---|
| Name | The name or alias of the superhero |
| Gender | The gender of the superhero |
| Race | The superhero's race (such as Human, Amazon, Vampire, etc.) |
| Eye Color | The superhero's eye color |
| Hair Color | The superhero's hair color |
| Skin Color | The superhero's skin color |
| Height | The superhero's height (in centimeters)<br>**Note:** Many of the listed superheroes are given a height and weight of -99. I am not exactly sure what this means, but I suspect it indicates that this information is unknown. |

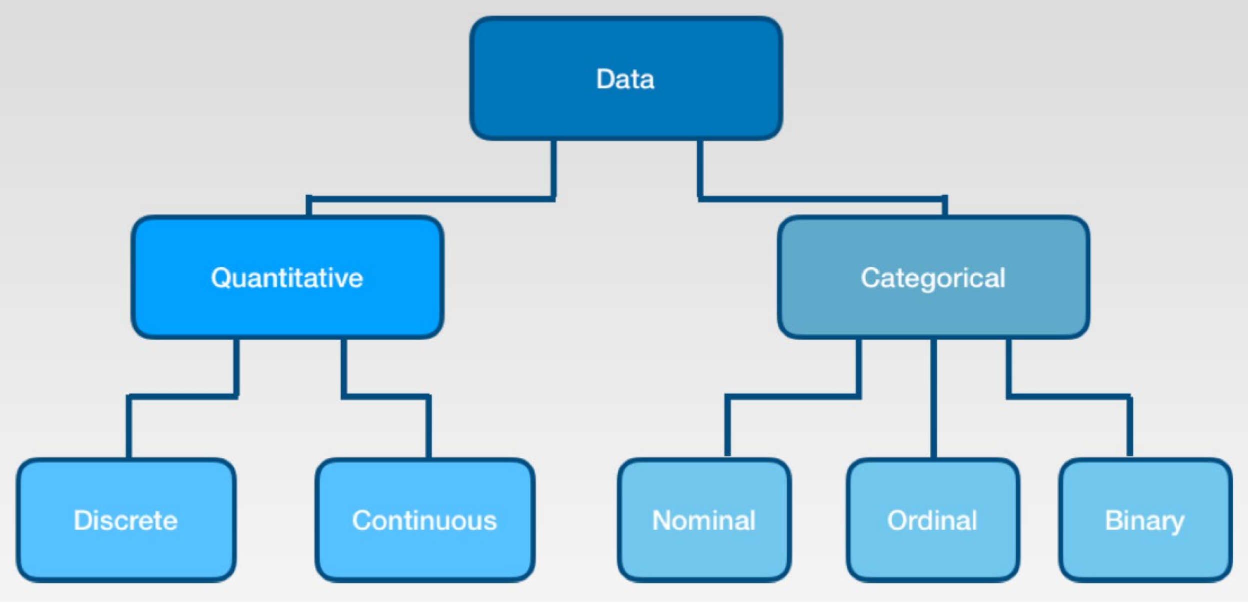
[1] These datasets were compiled by Kaggle user ClaudioDavi. For more information, see https://www.kaggle.com/claudiodavi/superhero-set/home.

| | |
|---|---|
| Weight | The superhero's weight (in kilograms)<br>**Note:** Many of the listed superheroes are given a height and weight of -99. I am not exactly sure what this means, but I suspect it indicates that this information is unknown. |
| Publisher | The comics company that created this superhero (such as Marvel, D.C., etc.) |
| Alignment | The superhero's overall alignment (good, bad, or neutral) |

## Descriptive statistics

Write a Jupyter Notebook with a code cell for each of the following tasks. Add text cells to separate and organize your code.

1. Loading the data into a DataFrame from the CSV File
2. In a text cell, describe the data type of each column



3. Compute the following measures for each column in the data set. If you cannot compute them for a particular column, you should write a text cell explaining why you cannot do it and propose and implement a solution.
    a. Measures of Central Tendency
        i. Mean
        ii. Median
        iii. Mode
    b. Measures of Dispersion
        i. Standard Deviation
        ii. Variance
        iii. Interquartile Range (IQR)

iv. Skewness

4. Clean the dataset. In the text cell, explain how was the cleaning was done and present your thought process for every decision you made
5. Show a scatter plot for each column versus the character ID. For each column/plot, reply: What can you learn from this plot? What would be a more appropriate plot to use? Plot the column using your proposed plot and discuss what observe.
6. Show a boxplot for each column. Describe what you observe from the box plot.

## Exploratory Data Analysis

On the same Jupyter Notebook, write a code cell for each of the following tasks. Use chapter 6 of our runestone book[2] as a guide for these steps and do not forget to add text cells to separate and organize your code.

After each code cell, use a text cell to describe your analysis of the visualizations.

1. Describing the Data¶
2. Visualizing Distribution with Histograms
3. Select four columns and plot a scatter matrix for discovering relationships
4. Show a scatter plot for a couple of variables that seem to have no relationship and compute:
    a. Covariance
    b. Correlation
5. Show a scatter plot for a couple of variables that seem to have a relationship and compute:
    a. Covariance
    b. Correlation

## Lab submission

1. Print your Jupyter Notebook as a PDF and upload it to Gradescope.
2. Export your Jupyter Notebook as a python file (.py) and upload it to Gradescope.

[2] https://runestone.academy/ns/books/published/CS418/WorldFacts/cs1_exploratory_data_analysis.html