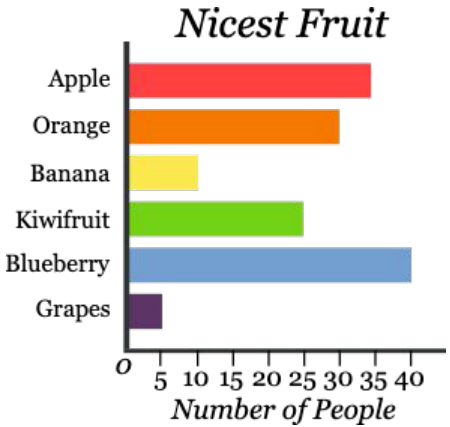
MSDS610 Fall 2022 - Exploratory Data Analysis

Ensun Pak
Abhradeep Mukherjee
Stephen Louie

What is EDA?

A technique used in Data Science to analyze and investigate data sets, summarize their main characteristics and draw insights from them. Often, data visualization methods and basic statistical measures are employed.



Why EDA?

Due to the possibility of a high volume of information found in a single file, EDA acts as an intermediary between asking a question about a business or operational process and constructing a hypothesis.

- What is the distribution of the data?
- Does the median align with the average?
- Does the data have any outliers?

How is EDA done?

1. Check summary statistics
    a. Mean, Median, Min, Max, etc.
    b. Distributions of variables (e.g. Histogram)
2. Develop visualizations aids
    a. Scatterplots, Box Plots, Bar Charts
3. Treatment of missing/null values
    a. Use statistical measure as a placeholder or drop altogether?