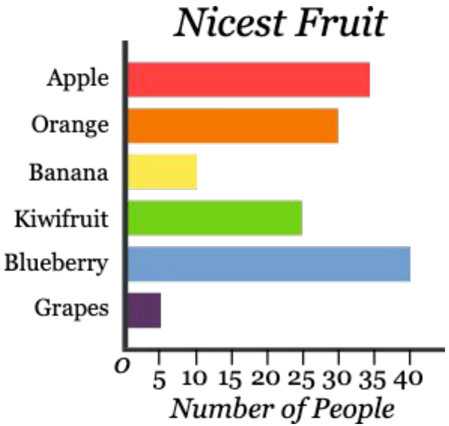
MSDS610 Fall 2022 - Exploratory Data Analysis

Ensun Pak
Abhradeep Mukherjee
Stephen Louie

What is EDA?

Technique used in Data Science used to analyze and investigate data sets and summarize their main characteristics, to draw insights from the data often employing data visualization methods and basic statistical measures.



Why EDA?

Due to the high volume of data that can be contained in a single file, EDA acts as an intermediary between asking a question about a business or operation and the development of a hypothesis.

- What is the distribution of 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
   a. Scatterplots, Box Plots, Bar Charts
3. Treatment of missing/null values
   a. Use statistical measure as a placeholder or drop altogether?