Cheat sheet

3ikakke

Outline

- Objectives
- Data types
- Data type mapping to python primitives
- Exploring Data - single variable (univariate)
- Exploring accross two variables (bivariate)
- Visualizing Single variables
- Visualizing Two Variables
- Hypothesis (AB) Testing
- Supervised Learning

Objectives

- One stop to remembering the core parts of statistics in datascience leading up to machine learning

Data types

- Numeric
  - Continuous (with no clear delineation between any two levels eg weight, height) => 2.2, 11.5, 18.2, 68.5
  - Discrete AKA Count AKA Interval (distinct numbers eg counts) => 1, 3, 7, 12
- Categorical
  - Binary (2 levels only) => male or female, dead or alive, healthy or ill, soldier or civilian
  - Nominal (Named with no inherent order eg location, firstname)
  - Ordinal (Named with inherent order eg) => small, medium, large, extra large
- Data may be missing

Data type mapping to python primitives

- Numeric
  - Continuous: Float

  - Discrete: Integer

- Categorical
  - Binary: Boolean (True or False)
  - Nominal: String
  - Ordinal: String

- Missing: None

Exploring Data - single variable (univariate)

- Numeric
  - Report minimum, maximum, and range
  - Check for distribution by looking at a histogram (decide what to report as center and spread)
    * if normally or approximately normally distributed report Mean (as center) and Standard Deviation (as spread)
    * if not normally distributed report Median (as center) and Inter Quartile Range (IQR) as spread
  - Remember the five-number summary
    * min, Q1, median, Q3, max
- Categorical
  - Report the frequencies
  - Repoet Proportions or percentages

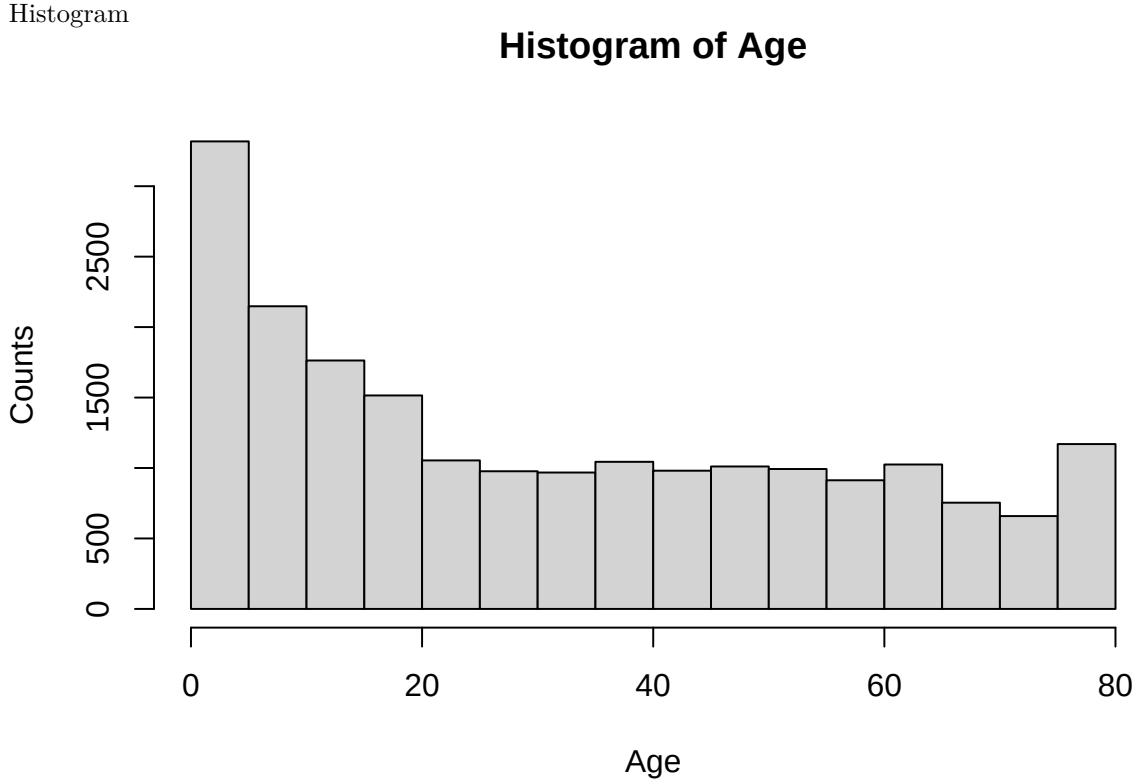
Exploring accross two variables (bivariate)

- Numeric vs Numeric
  - Correlation (using Pearsons Correlation Coefficient)
  - Correalation ranges from -1 through 0 to +1
  - They correspond to strong negative correlation, no correlation and strong positive correlation respectively

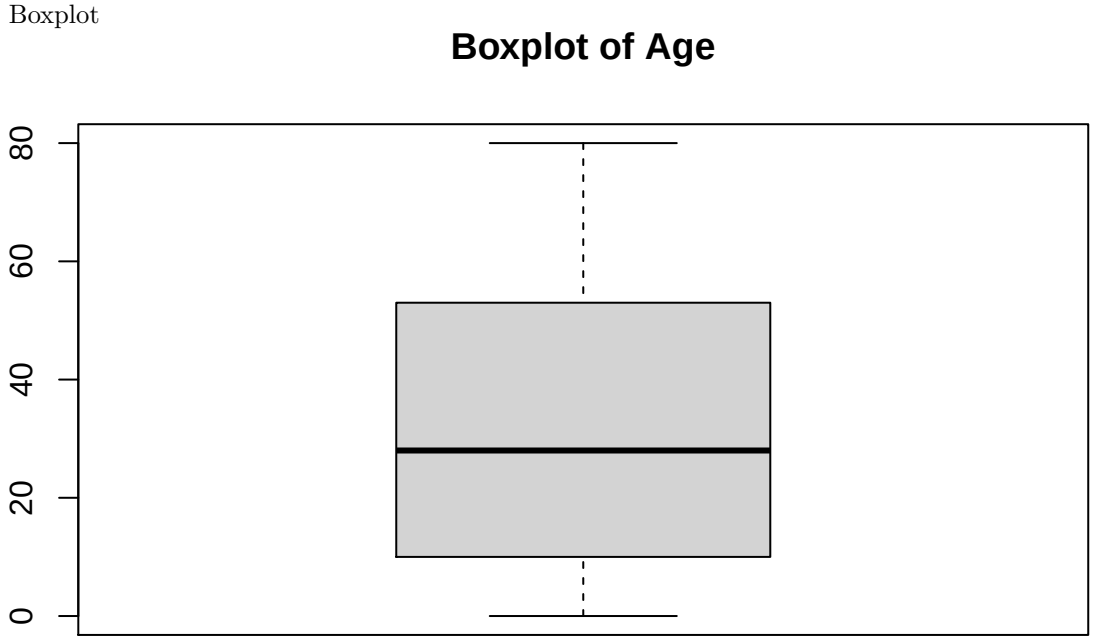
- Categorical vs Categorical
  - Report a two way table showing:
    * Frequencies
    * Proportions or Percentages
- Numeric vs Categorical
  - Report the numeric summaries at each level of the categorical variable
  - eg Age vs Sex
    * Mean and standard deviation for males
    * Mean and standard deviation for females

Visualizing Single variables

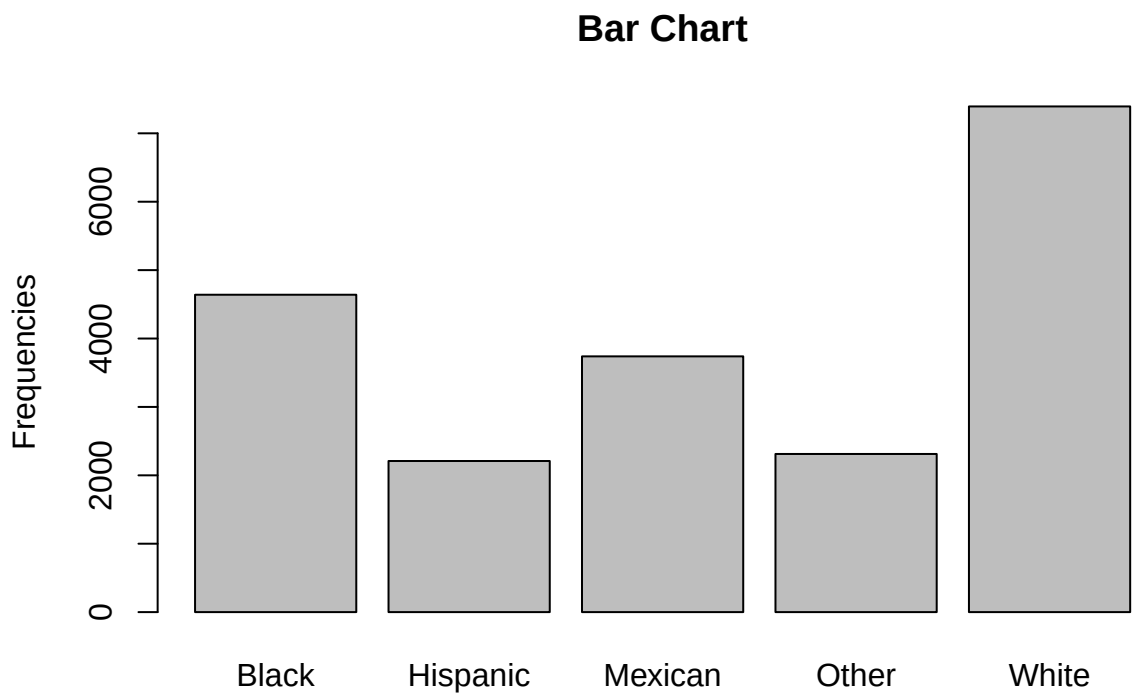
- Numeric
  - Histogram



  - Boxplot



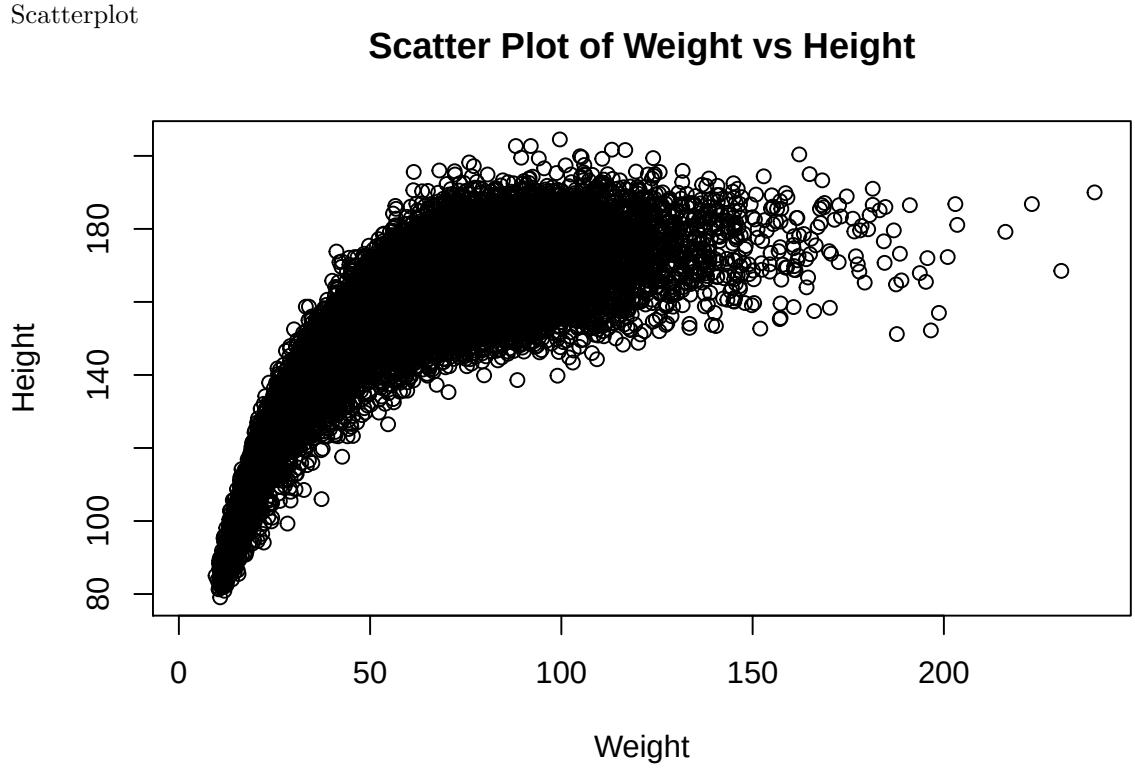
- Categorical
  - Bar Charts



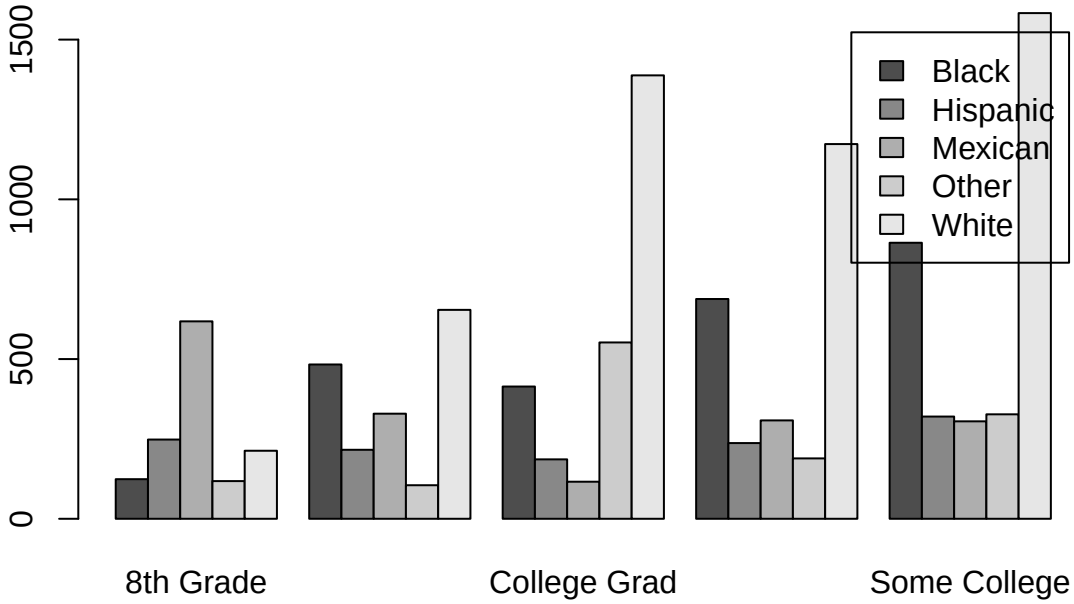
-

Visualizing Two Variables

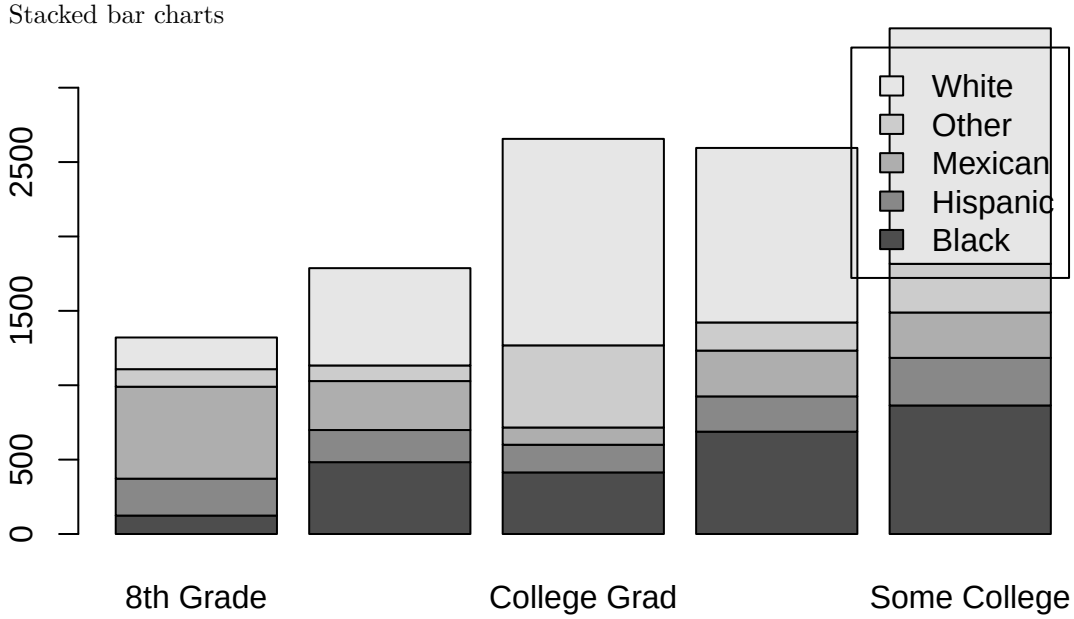
- Numeric vs Numeric
  - Scatterplot



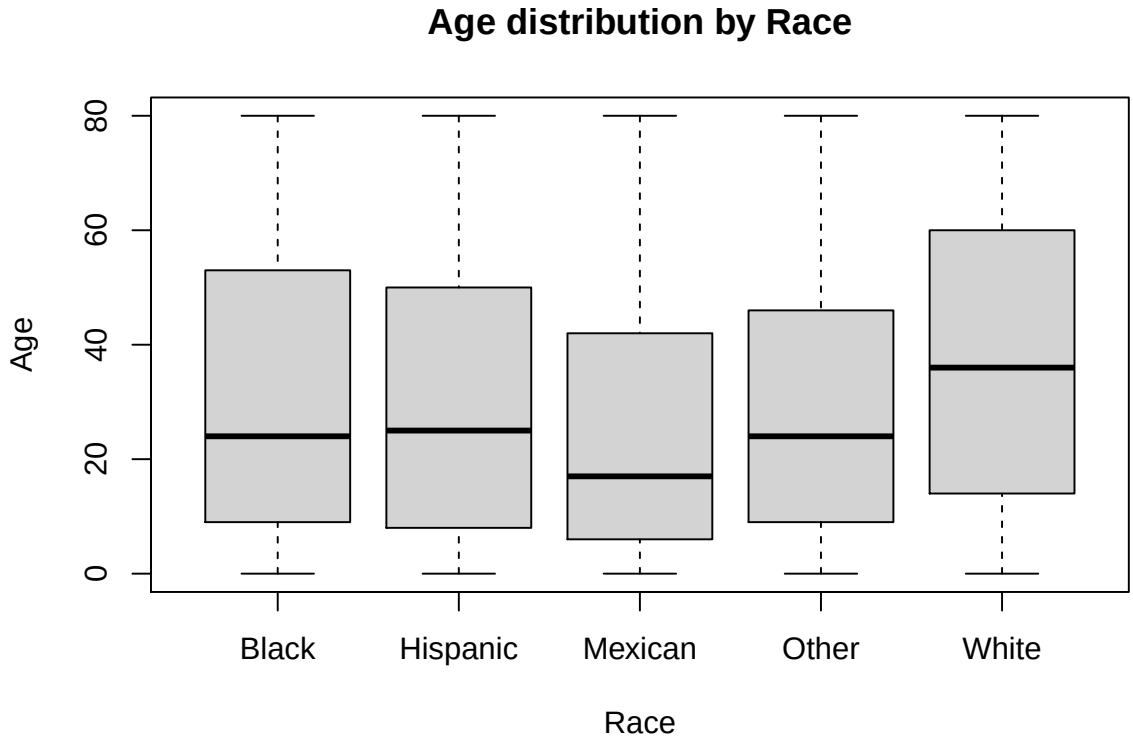
- Categorical vs Categorical
  - Grouped bar charts



– Stacked bar charts



- Numeric vs Categorical
  - Grouped box plots



Hypothesis AB Testing

- Numeric vs Numeric
  - Regression (tests for association)
- Categorical vs Categorical
  - Chi Squared Test (tests for eveness of distribution)
- Binary vs Numeric
  - T-Test (test for differences in mean)
- Categorical vs Numeric
  - Analysis of variance (ANOVA) (tests for difference in variance)

Supervised Learning

- Numeric label
  - Regression algorithms
- Categorical label
  - Classification algorithms

End