Introduction to Data Science and Analytics

Exploratory Data Analysis w/ Descriptive Statistics

Beginning Thoughts

Many people can be intimidated by statistics and/or probability theory especially as an academic subject

However, statistics and probability are widely used throughout our technology-driven society and provide foundational elements for much of the predictive capabilities in Data Science

Branches of Statistics

Descriptive statistics allows data scientists to describe data using numerical and visual summaries

Inferential statistics allows data scientists to draw conclusions and make predictions about full data population using sampled information.

Additionally, predictions (inferences) can be made about future observations!

Useful Insights from Statistics

- How and why did an event happen?
- What is the current state of a system? Are we using appropriate techniques to measure it?
- Can we model data to predict the outcomes of processes, situations, and events?
- What is the inherent or expected variability / uncertainty in our measurements and outputs?
- Can we enact corrective measures that are not postmortem? Can we make adjustments to avert a failure or negative events?

Statistics and Data Science

- Statistics allows us to understand the data
  - its characteristics
  - its attributes
  - its meaning
- Exploratory Data Analysis (EDA)
  - Blends statistics with visualization
  - Understand the shapes of the data
  - Understand the trends in the data
  - Understand the patterns in the data

DATA SCIENCE
& ANALYTICS

Statistics and Data Science

- Interrelations of the data
  - Correlations
  - Regressions
- Characteristic sub-groupings of the data
  - Classes
  - Separability
  - Comparisons of means, distributions, etc.

Descriptive Statistics

- Descriptive statistics give a “first look” at the data
- What are the measurements?
- What are the typical values, range of values, and likely values within each measurement?
  - Typical values → “Central Tendency”
  - Range of values → “Dispersion” or “Variability”
- What is the basic shape of the data, i.e. its frequency distribution?
  - Histogram, Probability Mass Function (PMF), etc.

DATA SCIENCE
& ANALYTICS

Data Science – *From Data to Decisions*

- The entire process should be enumerated so that others (and you!) can repeat the process
- In the context of data science, this means you must document your analytical procedures
- In data science this will usually be in the form of version-controlled scripts

  "...non-reproducible single occurrences are of no significance to science"

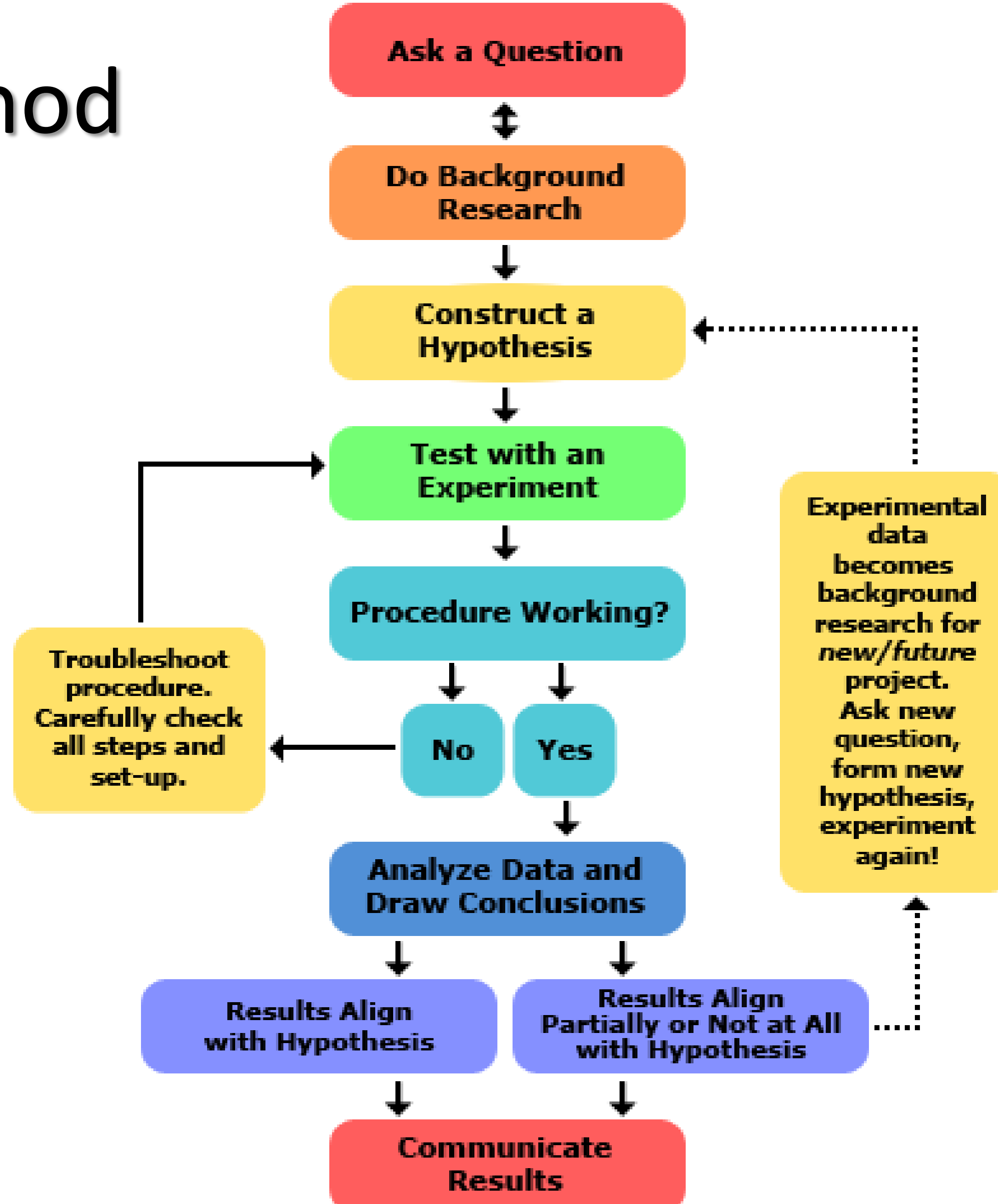
  - *Karl Popper, The Logic of Scientific Discovery*

- This concept is important no matter what field, business, or purpose for which you are applying data science practices

DATA SCIENCE & ANALYTICS

Scientific Method

A reproducible experiment is a series of steps that are [repeatable by others]() such that results can be verified



DATA SCIENCE
& ANALYTICS

Exploratory Data Analysis

- Exploratory Data Analysis (EDA) activities should be repeatable and well documented
  - Visualization
  - Statistical Analysis
  - Preliminary Modelling
- Can a collaborator achieve the same analysis and reach similar conclusions about the characteristic properties of the data?
  - Sometimes the “collaborator” is your future self!!
  - Can you reproduce your work in the future when / if needed?

DATA SCIENCE
& ANALYTICS

Results

- Obtaining useful and enlightening results is the most rewarding part of data science
- Processes for obtaining results must have provenance
- Generation of the results must be repeatable and verifiable by others
- Collaboration through VCS provides provenance of the tools that are developed and/or applied to data collections to achieve results

DATA SCIENCE
& ANALYTICS

Reproducibility

- Scripted, repeatable data collection
- Scripted, repeatable data cleaning
- Scripts or notebooks with comments/notes and data analysis steps
- Scripts or notebooks of preliminary data modelling
- Reproducible research <u>must</u> have provenance of all the steps, scripts and notebooks in VCS

DATA SCIENCE
& ANALYTICS

Common EDA Objectives

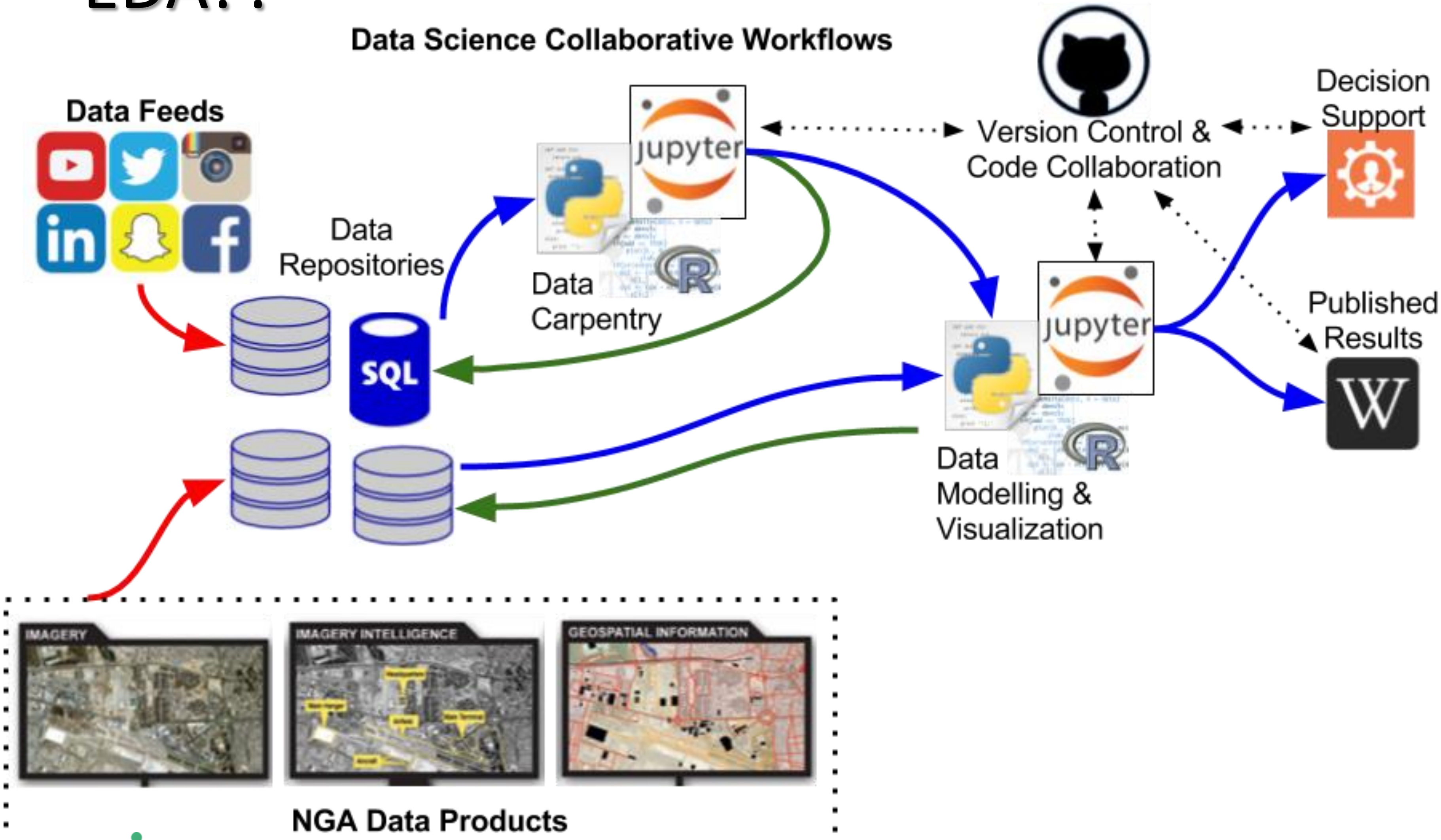
- Selection and formatting of appropriate input data elements
  - Extraction, transformation, and loading (ETL)
- Detection and elimination of bad (e.g. outliers) or missing data
  - Cleaning, pruning, etc.
- Description and summarization of data
- Validation of basic assumptions about the data
- Evaluating/interpreting relationships between explanatory variables
- Sometimes (optional) selection of <u>preliminary</u> models (explanatory vs. outcome variables)

DATA SCIENCE
& ANALYTICS

Exploratory Data Analysis or EDA

- Four basic EDA types
  - Univariate quantitative
    - e.g. mean, median, variance, etc.
  - Univariate graphical
    - e.g. histogram visualization
  - Multivariate (usually bivariate) quantitative
    - e.g. correlation
  - Multivariate graphical
    - e.g. regression plot, scatter plot, etc.

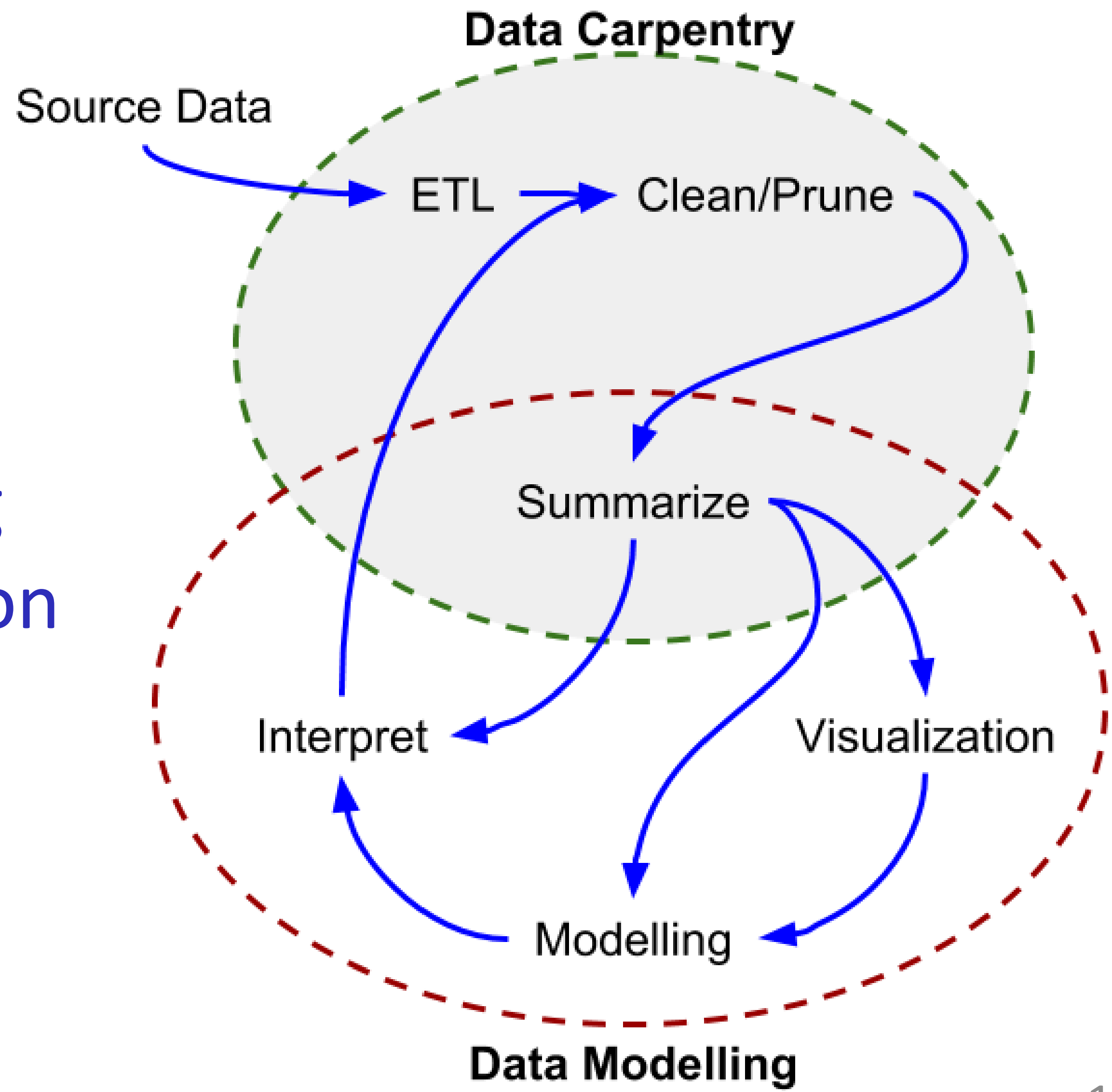
The "Science" in Data Science – Where is EDA??



DATA SCIENCE
& ANALYTICS

EDA in Data Science Workflow

EDA links data carpentry to data modelling and interpretation



EDA in Data Science Workflow

- Descriptive statistics provide a "first look" at the data
- Central tendency and dispersion

Age (yrs)		Weight (kg)		Height (cm)	
Min.	:18.00	Min.	: 42.00	Min.	:147.2
1st Qu.	:23.00	1st Qu.	: 58.40	1st Qu.	:163.8
Median	:27.00	Median	: 68.20	Median	:170.3
Mean	:30.18	Mean	: 69.15	Mean	:171.1
3rd Qu.	:36.00	3rd Qu.	: 78.85	3rd Qu.	:177.8
Max.	:67.00	Max.	:116.40	Max.	:198.1
St.Dev.	: 9.61	St.Dev.	: 13.35	St.Dev.	: 9.41
Var.	:92.32	Var.	:178.11	Var.	:88.50

DATA SCIENCE
& ANALYTICS

Central Tendency

- Given a particular variable within a data set, the measurement of central tendency is conceptually an average or other similar value of *likelihood*
- If a data measurement is collected along a numbered line, it is conceptually the center of the data
- There are a variety of ways to measure the "center" with arithmetic mean being the most common

DATA SCIENCE & ANALYTICS

Central Tendency

- Multiple methods to measure central tendency
- Various methods accommodate the different characteristics of the data set
- Some measures may be heavily influenced by outliers (extremal / abnormal measurements)
- Different types of measurements require different assessments
  - e.g., what is the <u>most likely</u> label within a data set?

DATA SCIENCE
& ANALYTICS

Central Tendency – Mean (μ)

- Example:

  [1, 2, 3, 4, 5, 6, 7, 8, 9]

- The arithmetic average, or <u>mean</u>, is the sum of all the numerical values divided by the number of values

- Mean = 5

  1+2+3+4+5+6+7+8+9 = 45; 45/9 = 5

Central Tendency - Mode

- Example:

[1, 2, 4, 4, 4, 6, 6, 9, 9]

- The most likely value determined by the most commonly occurring value

- Given the table of value counts we see the most common value is 4

Value	Count
1	1
2	1
4	3
6	2
9	2

Mode = 4

DATA SCIENCE & ANALYTICS

Central Tendency - Mode

- Example, non-numeric:

  [A, B, D, D, D, F, F, H, H]

- The most likely value as determined by the most commonly occurring value

- Given the table of value counts

we see the most common value

is D

Value	Count
A	1
B	1
D	3
F	2
H	2

Mode = D

DATA SCIENCE
& ANALYTICS

Central Tendency - Median

- The middle value (odd numbered list), or average of two middle values (even-numbered list), after the data is sorted
- Example:

  [1, 1, 1, 2, 2, 9, 9, 9, 9] $\rightarrow$ Median = 2

- Median is preferred in the presence of outliers or extremal measurements

  [1, 2, 3, 4, 5, 6, 7, 8, 99] $\rightarrow$ Median = 5

  $\rightarrow$ Mean = 15

outlier

Which value is more "typical" of the middle values of the data set?

DATA SCIENCE & ANALYTICS

Other "means"

- There are various other measures of central tendency

- Geometric and harmonic means

  - See Chpt 5.2 and 5.3 in CK-12 Probability and Stats course reference book for more information

- Additional measures of data set mean can incorporate dispersion measures

Central Tendency – Mean, Mode, Median

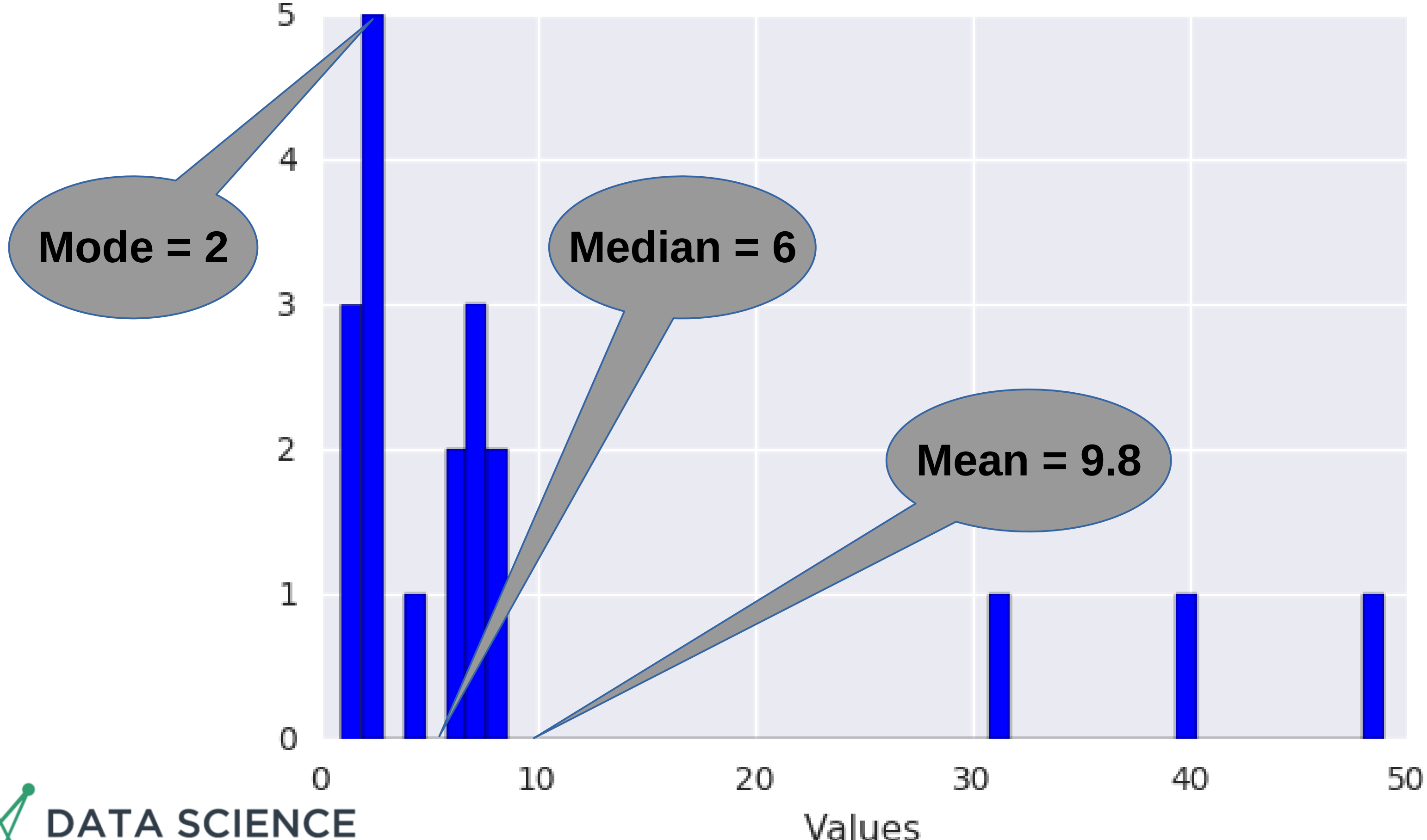
- Example: [1, 1, 1, 2, 2, 9, 9, 9, 9]
  - Mean = 4.8
  - Mode = 9
  - Median = 2
- Differences between the measurements can provide indication of outlier/extremal data
- The "best" measure of central tendency depends on:
  - Shape of the data (skewness, extremals, etc.)
  - Type of measurement (next lesson)

DATA SCIENCE
& ANALYTICS

Central Tendency – Mean, Mode, Median

Example Data: [1,1,1,2,2,2,2,2,4,6,6,7,7,7,8,8,31,40,49]

Histogram: Visualization of a count of the occurrence of values

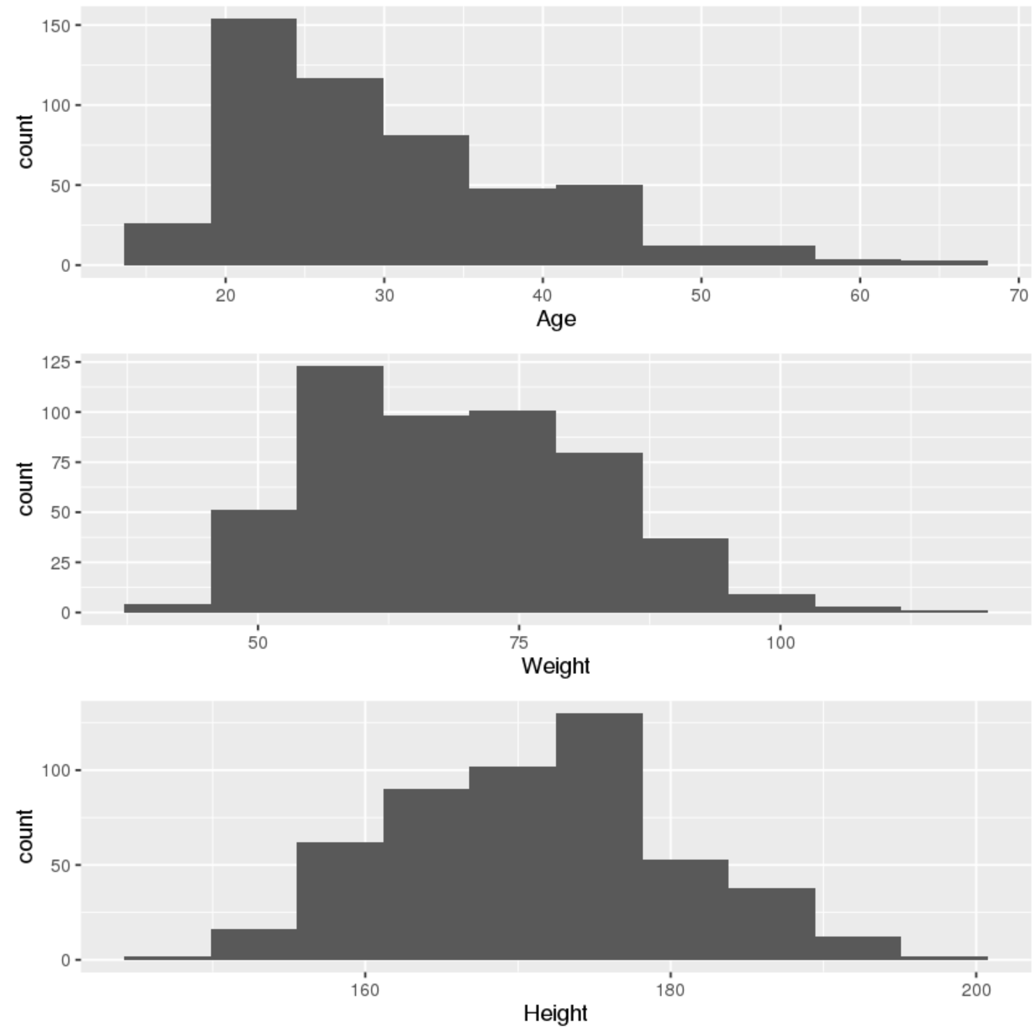


DATA SCIENCE
& ANALYTICS

EDA in Data Science Workflow

Visual exploration of the “shapes” of data

Example: histogram of data values



DATA SCIENCE
& ANALYTICS

Dispersion – Variance (σ^2)

- Squared expected deviation from mean
- Where mean (μ) is defined as

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Where x_i is a series of measurement values then

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

DATA SCIENCE
& ANALYTICS

Dispersion – Standard Deviation (σ)

- Standard deviation is the square root of variance
- More commonly used to quantify the dispersion than variance
  - Why? $\rightarrow$ same units as the measurement variable
- Lower standard deviation indicates less dispersion of the data
- Higher standard deviation indicates greater dispersion

Exploratory Data Analysis

- These techniques are the foundational first steps towards
  - Inferential statistics
  - Decision Support
  - Predictive Analytics

  … and more!

DATA SCIENCE
& ANALYTICS

Introduction to Data Science and Analytics

Exploratory Data Analysis w/ Descriptive Statistics