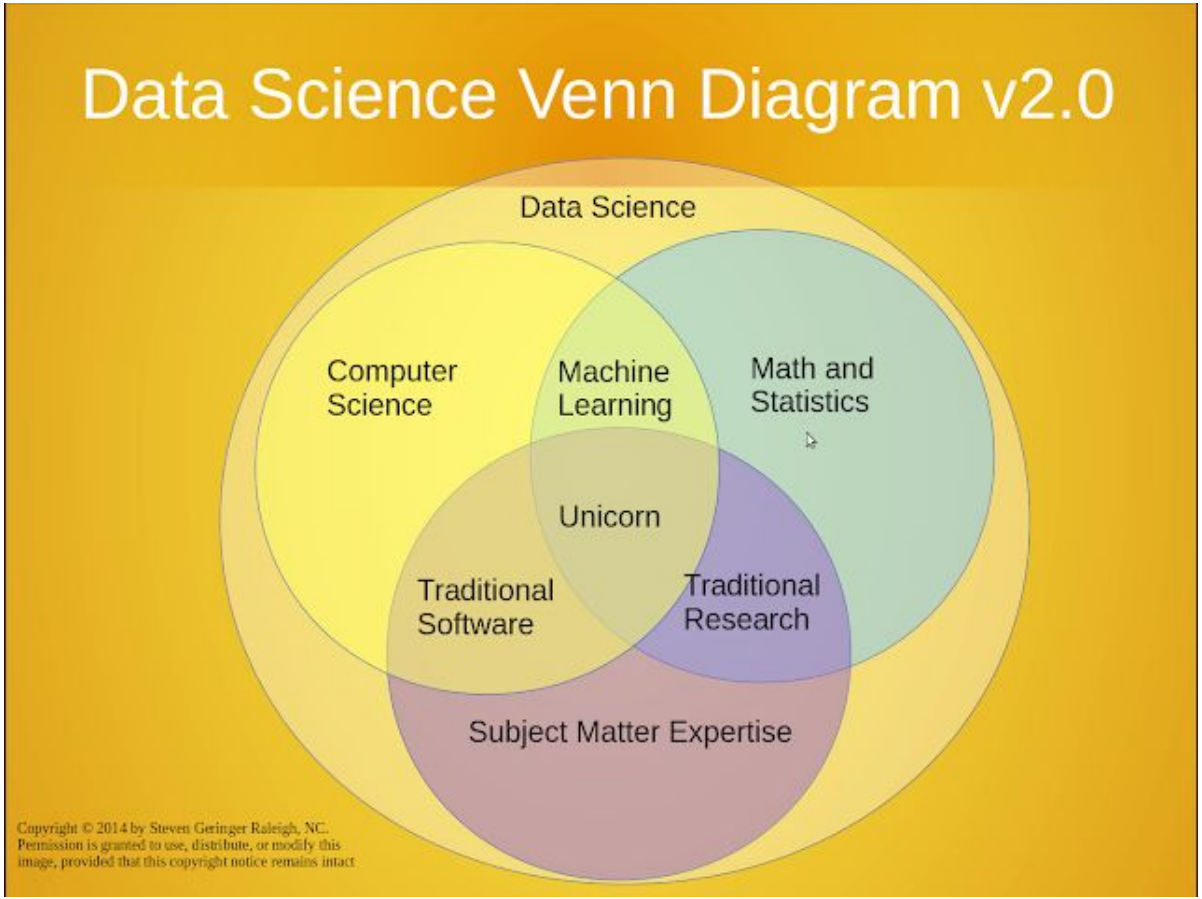
# Introduction to Statistical Modeling
# MSDS 598

(Exploratory) Data Science

Michael Ruddy

# Data Science: What? Who?



# Data Science: What? Who?

- “Fields of research are defined by the people who participate”

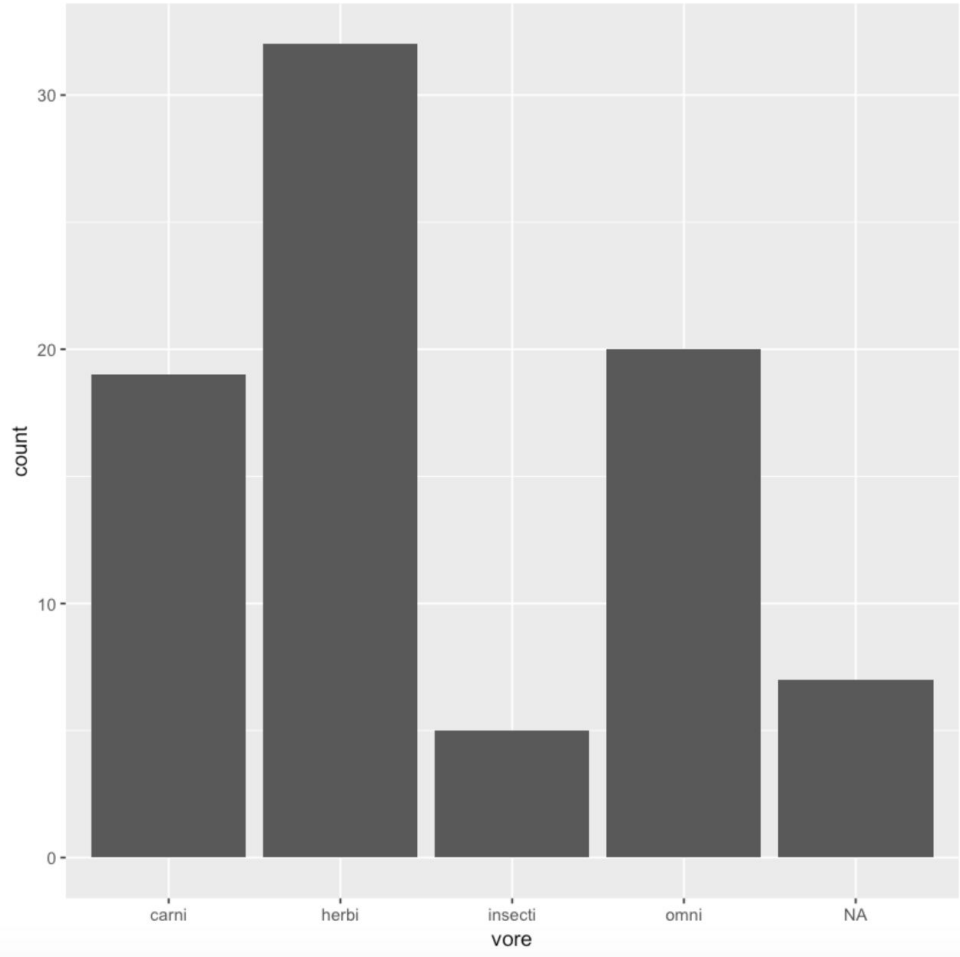
- Things Data Scientists sometimes do:
  - Acquire data
  - Clean/Pre-process data
  - Manage data storage
  - Exploratory analysis of data
  - Make predictions/inferences from data
  - Tell a story with data
    - Summarize, Visualize

# Exploratory Data Analysis

- **Informal** exploration of the data
- Summarize/Visualize properties of the data
    - Be careful of making predictions/inferences off of pure exploration/visualization!

- Iterative process
    1. Generate questions, hypotheses
    2. Visualize, transform, model your data
    3. Refine your questions, repeat

| | name | genus | vore | order | conservation | sleep_total | sleep_rem | sleep_cycle | awake |
|---|---|---|---|---|---|---|---|---|---|
| | <chr> | <chr> | <chr> | <chr> | <chr> | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | Chee… | Acin… | carni | Carn… | lc | 12.1 | NA | NA | 11.9 |
| 2 | Owl … | Aotus | omni | Prim… | <NA> | 17 | 1.8 | NA | 7 |
| 3 | Moun… | Aplo… | herbi | Rode… | nt | 14.4 | 2.4 | NA | 9.6 |
| 4 | Grea… | Blar… | omni | Sori… | lc | 14.9 | 2.3 | 0.133 | 9.1 |
| 5 | Cow | Bos | herbi | Arti… | domesticated | 4 | 0.7 | 0.667 | 20 |
| 6 | Thre… | Brad… | herbi | Pilo… | <NA> | 14.4 | 2.2 | 0.767 | 9.6 |
| 7 | Nort… | Call… | carni | Carn… | vu | 8.7 | 1.4 | 0.383 | 15.3 |
| 8 | Vesp… | Calo… | <NA> | Rode… | <NA> | 7 | NA | NA | 17 |
| 9 | Dog | Canis | carni | Carn… | domesticated | 10.1 | 2.9 | 0.333 | 13.9 |
| 10 | Roe … | Capr… | herbi | Arti… | lc | 3 | NA | NA | 21 |
| 11 | Goat | Capri | herbi | Arti… | lc | 5.3 | 0.6 | NA | 18.7 |
| 12 | Guin… | Cavis | herbi | Rode… | domesticated | 9.4 | 0.8 | 0.217 | 14.6 |
| 13 | Griv… | Cerc… | omni | Prim… | lc | 10 | 0.7 | NA | 14 |
| 14 | Chin… | Chin… | herbi | Rode… | domesticated | 12.5 | 1.5 | 0.117 | 11.5 |
| 15 | Star… | Cond… | omni | Sori… | lc | 10.3 | 2.2 | NA | 13.7 |
| 16 | Afri… | Cric… | omni | Rode… | <NA> | 8.3 | 2 | NA | 15.7 |
| 17 | Less… | Cryp… | omni | Sori… | lc | 9.1 | 1.4 | 0.15 | 14.9 |
| 18 | Long… | Dasy… | carni | Cing… | lc | 17.4 | 3.1 | 0.383 | 6.6 |
| 19 | Tree… | Dend… | herbi | Hyra… | lc | 5.3 | 0.5 | NA | 18.7 |
| 20 | Nort… | Dide… | omni | Dide… | lc | 18 | 4.9 | 0.333 | 6 |
| 21 | Asia… | Elep… | herbi | Prob… | en | 3.9 | NA | NA | 20.1 |
| 22 | Big … | Epte… | inse… | Chir… | lc | 19.7 | 3.9 | 0.117 | 4.3 |
| 23 | Horse | Equus | herbi | Peri… | domesticated | 2.9 | 0.6 | 1 | 21.1 |
| 24 | Donk… | Equus | herbi | Peri… | domesticated | 3.1 | 0.4 | NA | 20.9 |
| 25 | Euro… | Erin… | omni | Erin… | lc | 10.1 | 3.5 | 0.283 | 13.9 |
| 26 | Pata… | Eryt… | omni | Prim… | lc | 10.9 | 1.1 | NA | 13.1 |
| 27 | West… | Euta… | herbi | Rode… | <NA> | 14.9 | NA | NA | 9.1 |
| 28 | Dome… | Felis | carni | Carn… | domesticated | 12.5 | 3.2 | 0.417 | 11.5 |
| 29 | Gala… | Gala… | omni | Prim… | <NA> | 9.8 | 1.1 | 0.55 | 14.2 |
| 30 | Gira… | Gira… | herbi | Arti… | cd | 1.9 | 0.4 | NA | 22.1 |
| 31 | Pilo… | Glob… | carni | Ceta… | cd | 2.7 | 0.1 | NA | 21.4 |
| 32 | Gray… | Hali… | carni | Carn… | lc | 6.2 | 1.5 | NA | 17.8 |
| 33 | Gray… | Hete… | herbi | Hyra… | lc | 6.3 | 0.6 | NA | 17.7 |
| 34 | Human | Homo | omni | Prim… | <NA> | 8 | 1.9 | 1.5 | 16 |
| 35 | Mong… | Lemur | herbi | Prim… | vu | 9.5 | 0.9 | NA | 14.5 |
| 36 | Afri… | Loxo… | herbi | Prob… | vu | 3.3 | NA | NA | 20.7 |
| 37 | Thic… | Lutr… | carni | Dide… | lc | 19.4 | 6.6 | NA | 4.6 |
| 38 | Maca… | Maca… | omni | Prim… | <NA> | 10.1 | 1.2 | 0.75 | 13.9 |
| 39 | Mong… | Meri… | herbi | Rode… | lc | 14.2 | 1.9 | NA | 9.8 |
| 40 | Gold… | Meso… | herbi | Rode… | en | 14.3 | 3.1 | 0.2 | 9.7 |
| 41 | "Vol… | Micr… | herbi | Rode… | <NA> | 12.8 | NA | NA | 11.2 |
| 42 | Hous… | Mus | herbi | Rode… | nt | 12.5 | 1.4 | 0.183 | 11.5 |
| 43 | Litt… | Myot… | inse… | Chir… | <NA> | 19.9 | 2 | 0.2 | 4.1 |
| 44 | Roun… | Neof… | herbi | Rode… | nt | 14.6 | NA | NA | 9.4 |
| 45 | Slow… | Nyct… | carni | Prim… | <NA> | 11 | NA | NA | 13 |
| 46 | Degu | Octo… | herbi | Rode… | lc | 7.7 | 0.9 | NA | 16.3 |
| 47 | Nort… | Onyc… | carni | Rode… | lc | 14.5 | NA | NA | 9.5 |
| 48 | Rabb… | Oryc… | herbi | Lago… | domesticated | 8.4 | 0.9 | 0.417 | 15.6 |
| 49 | Sheep | Ovis | herbi | Arti… | domesticated | 3.8 | 0.6 | NA | 20.2 |
| 50 | Chim… | Pan | omni | Prim… | <NA> | 9.7 | 1.4 | 1.42 | 14.3 |
| 51 | Tiger | Pant… | carni | Carn… | en | 15.8 | NA | NA | 8.2 |
| 52 | Jagu… | Pant… | carni | Carn… | nt | 10.4 | NA | NA | 13.6 |
| 53 | Lion | Pant… | carni | Carn… | vu | 13.5 | NA | NA | 10.5 |
| 54 | Babo… | Papio | omni | Prim… | <NA> | 9.4 | 1 | 0.667 | 14.6 |
| 55 | Dese… | Para… | <NA> | Erin… | lc | 10.3 | 2.7 | NA | 13.7 |
| 56 | Potto | Pero… | omni | Prim… | lc | 11 | NA | NA | 13 |
| 57 | Deer… | Pero… | <NA> | Rode… | <NA> | 11.5 | NA | NA | 12.5 |
| 58 | Phal… | Phal… | <NA> | Dipr… | <NA> | 13.7 | 1.8 | NA | 10.3 |
| 59 | Casp… | Phoca | carni | Carn… | vu | 3.5 | 0.4 | NA | 20.5 |
| 60 | Comm… | Phoc… | carni | Ceta… | vu | 5.6 | NA | NA | 18.4 |
| 61 | Poto… | Poto… | herbi | Dipr… | <NA> | 11.1 | 1.5 | NA | 12.9 |
| 62 | Gian… | Prio… | inse… | Cing… | en | 18.1 | 6.1 | NA | 5.9 |
| 63 | Rock… | Proc… | <NA> | Hyra… | lc | 5.4 | 0.5 | NA | 18.6 |
| 64 | Labo… | Ratt… | herbi | Rode… | lc | 13 | 2.4 | 0.183 | 11 |

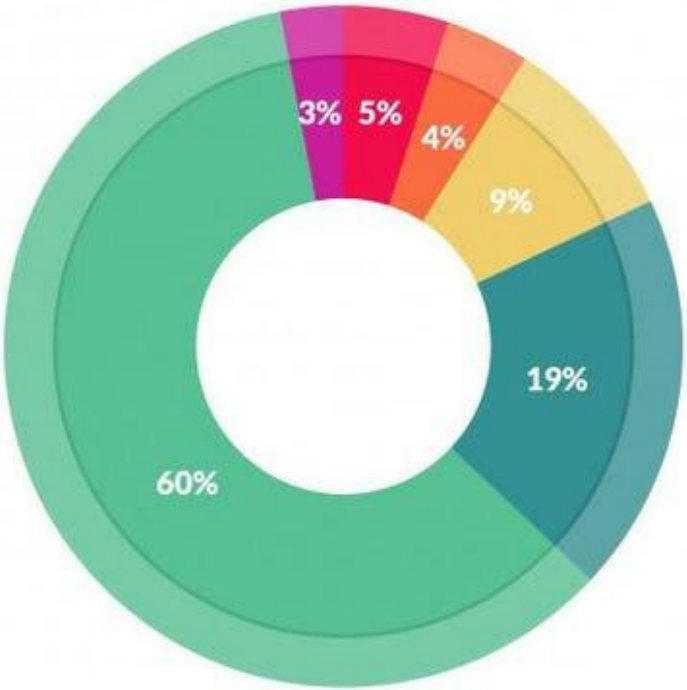

30
20
10
0
count
carni
herbi
insecti
omni
NA
vore

# Why Exploratory Data Analysis?

- But we are learning modeling!
  - What questions are you trying to answer?
  - Are there issues with your data?

- Narratives
  - Easier to remember big picture ideas
  - Communicating your results using summarization/visualization

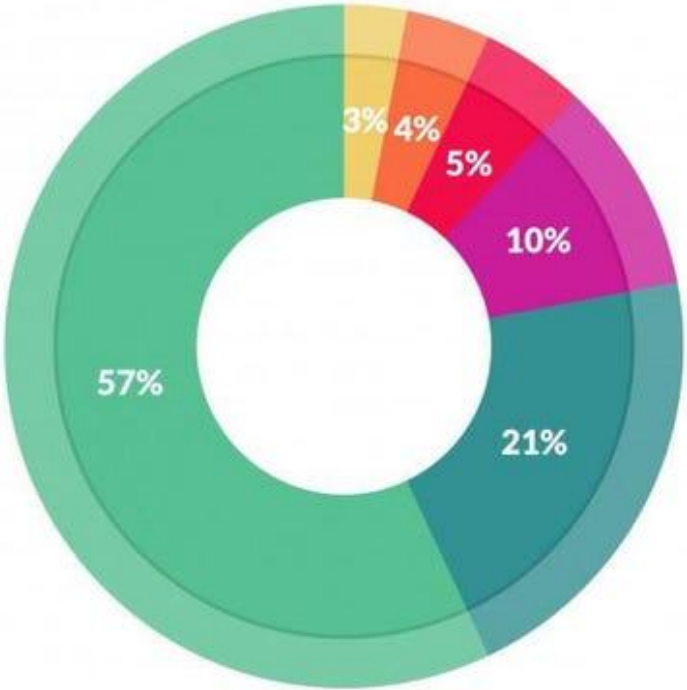
- Data needs to be clean (tidy) before modeling




## What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Forbes Article




What's the least enjoyable part of data science?

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*
- *Other: 5%*

Forbes Article

# Terminology

- We will mostly be working tabular data
  - Most common data type
  - Sometimes referred to as *structured* data
  - (*unstructured* data usually refers to images, audio, text, etc.)

- **Tabular Data:** Set of values, each associated with a variable and an observation.
- **Variable:** A quantity, quality, or property that you can measure
- **Value:** The state of a variable when you measure it
- **Observation**: Set of measurements under similar conditions (time, object, etc.), i.e. several values for different variables.

# Terminology

**Tabular Data**

Variables → name, genus, vore, order, conservation, sleep_total, sleep_rem, sleep_cycle, awake, brainwt, bodywt

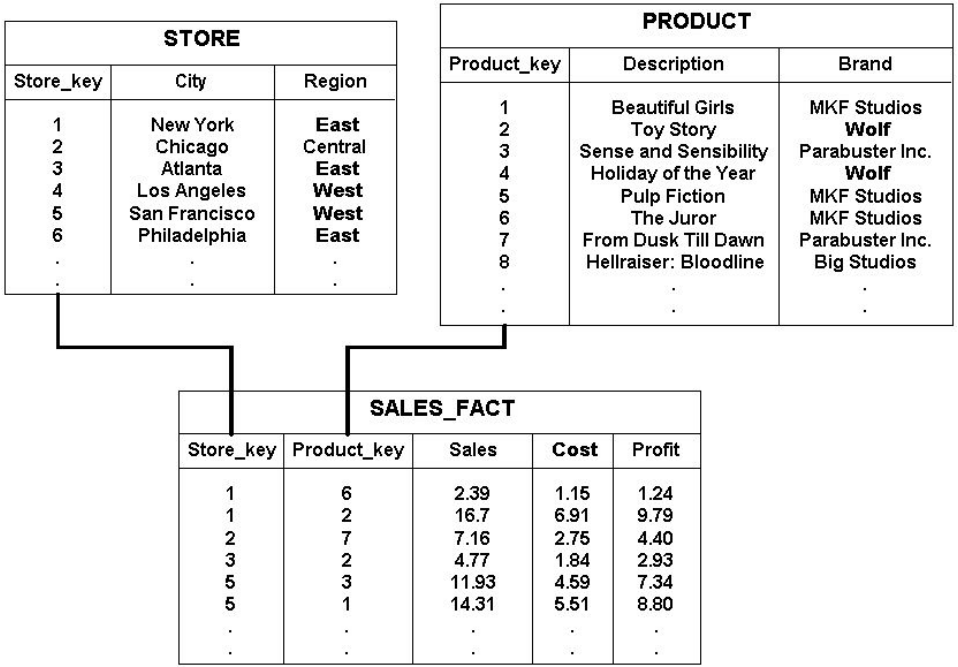
| name | genus | vore | order | conservation | sleep_total | sleep_rem | sleep_cycle | awake | brainwt | bodywt |
|---|---|---|---|---|---|---|---|---|---|---|
| Cheetah | Acinonyx | carni | Carnivora | lc | 12.1 | NA | NA | 11.9 | NA | 50.000 |
| Owl monkey | Aotus | omni | Primates | NA | 17.0 | 1.8 | NA | 7.0 | 0.01550 | 0.480 |
| Mountain beaver | Aplodontia | herbi | Rodentia | nt | 14.4 | 2.4 | NA | 9.6 | NA | 1.350 |
| Greater short-tailed shrew | Blarina | omni | Soricomorpha | lc | 14.9 | 2.3 | 0.1333333 | 9.1 | 0.00029 | 0.019 |
| Cow | Bos | herbi | Artiodactyla | domesticated | 4.0 | 0.7 | 0.6666667 | 20.0 | 0.42300 | 600.000 |
| Three-toed sloth | Bradypus | herbi | Pilosa | NA | 14.4 | 2.2 | 0.7666667 | 9.6 | NA | 3.850 |

Observation → Cow, Bos, herbi, Artiodactyla, domesticated, 4.0, 0.7, 0.6666667, 20.0, 0.42300, 600.000

Values

# Relational Data

- Data is often arranged into many tables with *relations*
- Organized using a **R**elational **D**ata**B**ase **M**anagement **S**ystem (RDBMS)



**STORE**

| Store_key | City | Region |
|---|---|---|
| 1 | New York | **East** |
| 2 | Chicago | Central |
| 3 | Atlanta | **East** |
| 4 | Los Angeles | **West** |
| 5 | San Francisco | **West** |
| 6 | Philadelphia | **East** |
| . | . | . |
| . | . | . |

**PRODUCT**

| Product_key | Description | Brand |
|---|---|---|
| 1 | Beautiful Girls | MKF Studios |
| 2 | Toy Story | **Wolf** |
| 3 | Sense and Sensibility | Parabuster Inc. |
| 4 | Holiday of the Year | **Wolf** |
| 5 | Pulp Fiction | MKF Studios |
| 6 | The Juror | MKF Studios |
| 7 | From Dusk Till Dawn | Parabuster Inc. |
| 8 | Hellraiser: Bloodline | Big Studios |
| . | . | . |
| . | . | . |

**SALES_FACT**

| Store_key | Product_key | Sales | **Cost** | Profit |
|---|---|---|---|---|
| 1 | 6 | 2.39 | 1.15 | 1.24 |
| 1 | 2 | 16.7 | 6.91 | 9.79 |
| 2 | 7 | 7.16 | 2.75 | 4.40 |
| 3 | 2 | 4.77 | 1.84 | 2.93 |
| 5 | 3 | 11.93 | 4.59 | 7.34 |
| 5 | 1 | 14.31 | 5.51 | 8.80 |
| . | . | . | . | . |
| . | . | . | . | . |

[From Oracle Documentation]

# Types of Variables (non-exhaustive)

- **Quantitative**: variables representing numerical values you can perform arithmetic operations with.
  - *Discrete*: integer numbers (nothing "in-between")
  - *Continuous*: real numbers

- **Qualitative (categorical)**: variables representing non-numeric properties.
  - *Nominal*: No rank or ordering
  - *Ordered*: Clear rank or ordering

# Types of Variables (non-exhaustive)

Nominal Discrete Ordered Continuous

| name | year | month | day | hour | lat | long | status | category | wind | pressure | ts_diameter | hu_diameter |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Amy | 1975 | 6 | 27 | 0 | 27.5 | -79.0 | tropical depression | -1 | 25 | 1013 | NA | NA |
| Amy | 1975 | 6 | 27 | 6 | 28.5 | -79.0 | tropical depression | -1 | 25 | 1013 | NA | NA |
| Amy | 1975 | 6 | 27 | 12 | 29.5 | -79.0 | tropical depression | -1 | 25 | 1013 | NA | NA |
| Amy | 1975 | 6 | 27 | 18 | 30.5 | -79.0 | tropical depression | -1 | 25 | 1013 | NA | NA |
| Amy | 1975 | 6 | 28 | 0 | 31.5 | -78.8 | tropical depression | -1 | 25 | 1012 | NA | NA |
| Amy | 1975 | 6 | 28 | 6 | 32.4 | -78.7 | tropical depression | -1 | 25 | 1012 | NA | NA |
| Amy | 1975 | 6 | 28 | 12 | 33.3 | -78.0 | tropical depression | -1 | 25 | 1011 | NA | NA |
| Amy | 1975 | 6 | 28 | 18 | 34.0 | -77.0 | tropical depression | -1 | 30 | 1006 | NA | NA |
| Amy | 1975 | 6 | 29 | 0 | 34.4 | -75.8 | tropical storm | 0 | 35 | 1004 | NA | NA |
| Amy | 1975 | 6 | 29 | 6 | 34.0 | -74.8 | tropical storm | 0 | 40 | 1002 | NA | NA |
| Amy | 1975 | 6 | 29 | 12 | 33.8 | -73.8 | tropical storm | 0 | 45 | 1000 | NA | NA |

# Example EDA Questions

- What type of variation occurs within a variable?
  - Typical, unusual values?

- What type of covariation occurs between variables?
  - How are two or more variables related?

- Are there missing or corrupted values?
  - How bad is the issue?

- Are there outliers?
  - Real or corrupted?

# Example EDA Questions

- What type of variation occurs within a variable?
  - Typical, unusual values?

Important for Modeling!

- What type of covariation occurs between variables?
  - How are two or more variables related?

- Are there missing or corrupted values?
  - How bad is the issue?

- Are there outliers?
  - Real or corrupted?

# Example EDA Questions

- What type of variation occurs within a variable?
  - Typical, unusual values?

- What type of covariation occurs between variables?
  - How are two or more variables related?

Also Important for Modeling!

- Are there missing or corrupted values?
  - How bad is the issue?

- Are there outliers?
  - Real or corrupted?

# Tools

- Categorical Variables
  - Count, proportion, percentage, bar graphs

- Numerical Variables
  - Mean,  standard deviation, histograms, box plots.

- Covariation
  - Correlation coefficients, scatter plots, cross tables

# Tools

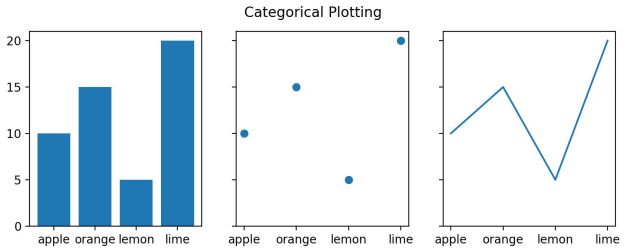
- Categorical Variables
  - Count, proportion, percentage, bar graphs

- Numerical Variables
  - Mean,  standard deviation, histograms, box plots.

- Covariation
  - Correlation coefficients, scatter plots, cross tables
  - Models!

# Python Libraries

- Pandas
  - Very common tool for manipulating/exploring/joining/transforming tabular data in Python
  - Uses the *DataFrame* object




From matplotlib.org

- Matplotlib
  - Built on NumPy (arrays, matrices)
  - Provides MATLAB-like visualizations