The R Essentials

Part 2 ☺

By the end of this workshop, you’ll know how to …

Identify

- Different variable types
  - Categorical vs. Continuous variable
  - Nominal vs. Ordinal
  - Factor vs. Character

Know

- Advantages to a factor vs. a character

Find

- Outlier in a plot
- Missing values

Utilize

- Tidyverse – decide which packages address your project goals
  - Readr: read_csv, read_tsv
  - Dplyr: filter, select, mutate, group_by/summarize, arrange
  - ggplot: point, boxplot, histogram, barchart

R Packages + Resources

- Hadley Wickham and Jennifer Bryan’s book: https://r-pkgs.org/
- RStudio (Posit) list of packages + use: https://support.posit.co/hc/en-us/articles/201057987-Quick-list-of-useful-R-packages
- General R Tutioral + Packages: https://www.tutorialspoint.com/r/r_packages.htm

Review Packages

ggplot2

- Visualize your data
- Create histograms, barcharts, scatterplots
- Edit the aesthetics of plots

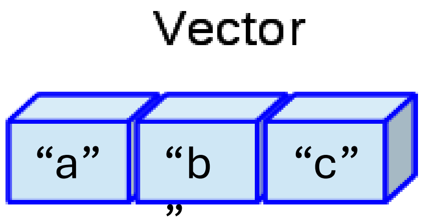
dplyr

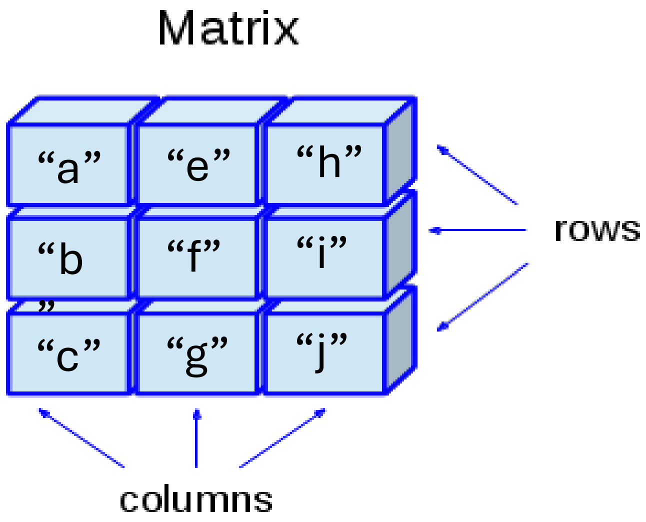
- Manipulate your data
- Select variables
- Filter data frames
- Create new variables

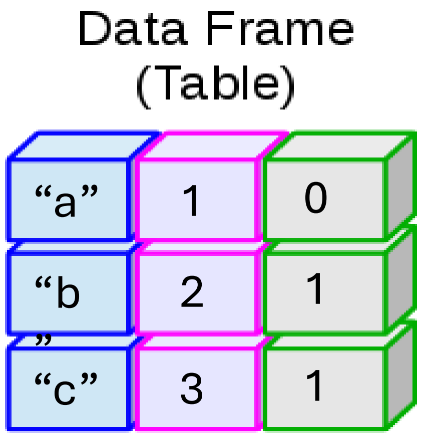
readr

- Read in data from .csv or .tsv files
- Can specify variable type in function call

Review larger data types in R







Simply put, these hold information in different ways…

Image borrowed from: http://venus.ifca.unican.es/Rintro/dataStruct.html

Variable Types

Categorical	Continuous
• Nominal - a variable with groups that have no particular order • Ordinal - a variable with groups that maintain an incremental order • Factor – a variable assigned as a factor will tell R to maintain a specific order of variable groups	• A numeric value that takes on a range of values

Example: A study focuses on adult males defined as 40-60 years old diagnosed with prostrate cancer and records if a patient has a BRCA1 mutation or not (indicated by “1” – yes mutated, “0” – not mutated). A patient may receive one of three available treatment dosages depending on the severity of their diagnosis (“A” – highest dose, “B” – moderate dose, “C” - minimum dose)

Categorical variables

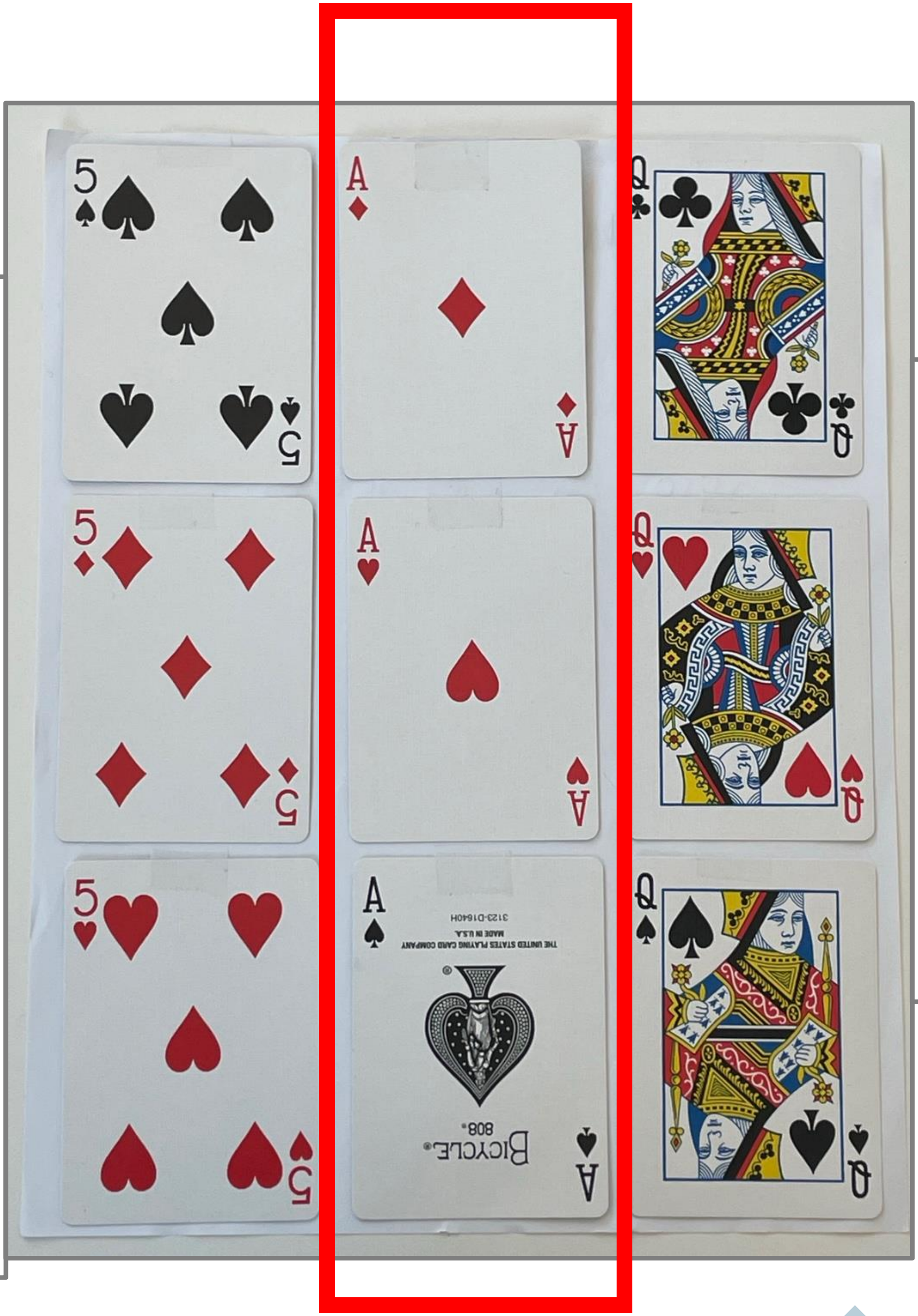
- **Treatment:** Ordinal variable since there is a specific incremental order to understanding this variable
- **BRCA1 mutation indicator:** binary variable since we have two levels (0/1) , nominal variable since there’s no incremental order between binary levels

Continuous variable

- **Age:** can take any numeric value between 40-60 years

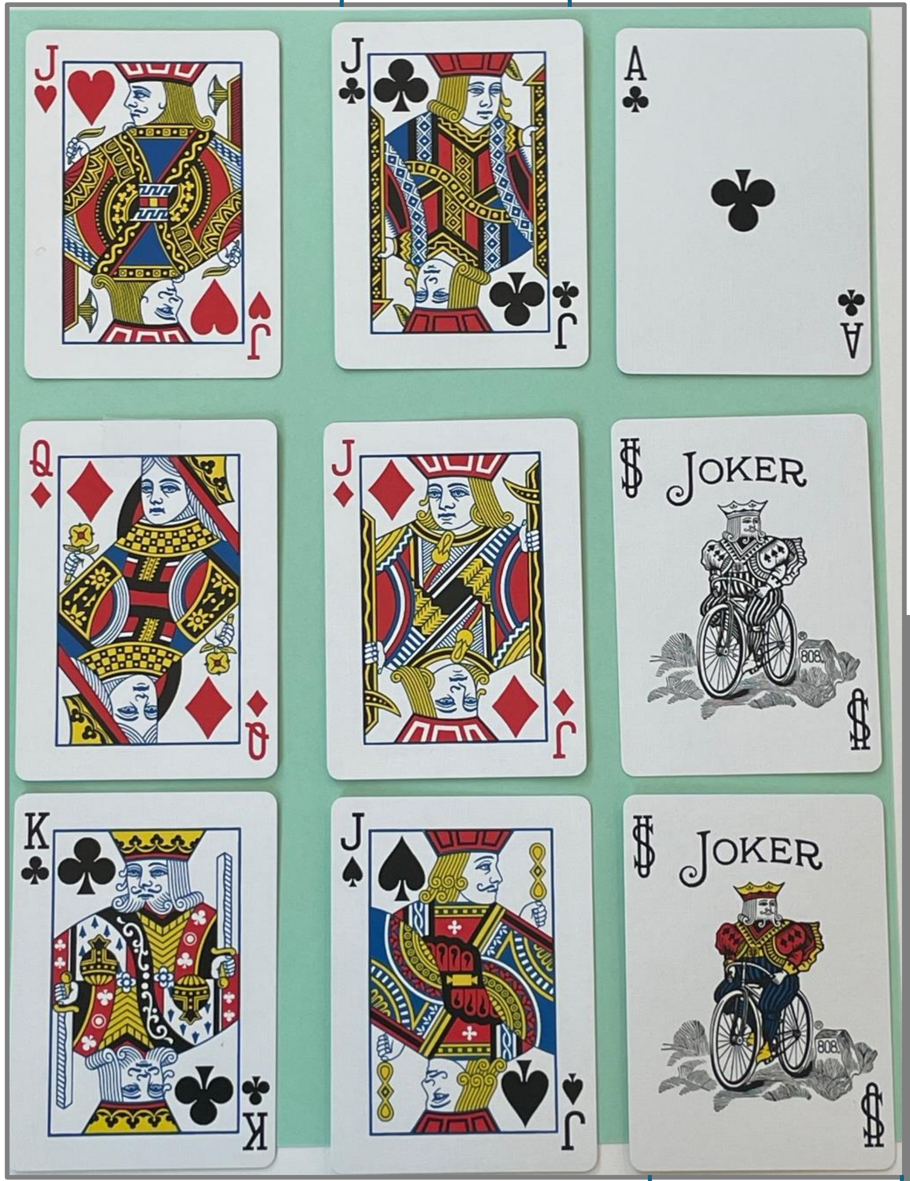
Aces are good examples of a card (or variable) that we understand as a number, but is represented by a letter

Aces are **categorical** variables here with 1 level – "1"



Recall characters or strings of:
"1"
"252"
"444"

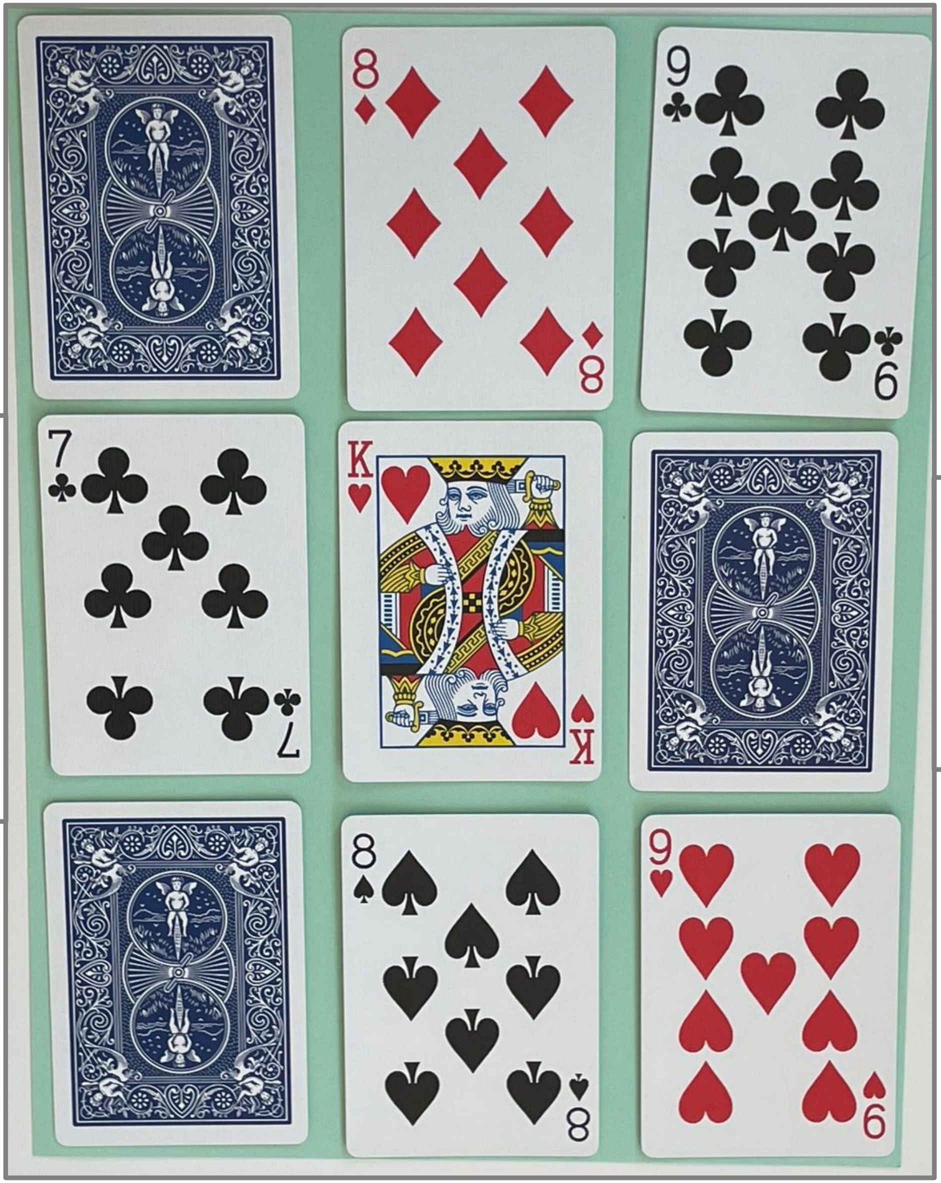
Factors have an order like
”J” → “Q” → “K”



This column is a **categorical** variable with groups “J” or “Jack”

Consider Jokers to be “0” in this example. This column represents a **binary, nominal categorical** variable
- The data would look like “1”/“0”

With these values missing, how can we fully understand the information in this column?



A flipped card here indicates a **missing value**