# W1

## Intro of DS

* What is data science?
* What compare with traditional algorithms how ML different?
* List four V of big data
* What is the size of small data? Size of big data?
  + Why we focusing on small first?
* Why use R (4 reasons)?
* What is rectangular data?
  + What is non rectangular data?
* What Is Hypothesis?
  + Hypothesis generation?
  + Hypothesis confirmation?
* Why we focus on hypothesis generation?
  + How it helps in comprehending business problem?(4 reasons)

## Data Science life cycle

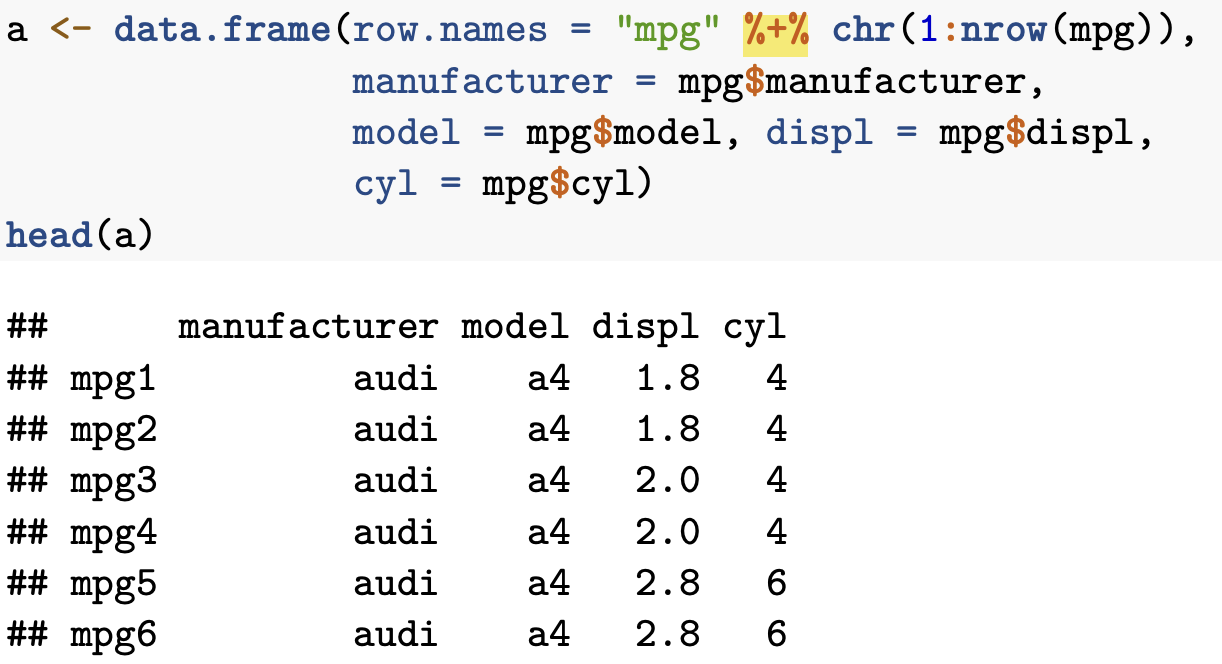
* List 6 stages and question for each stage
  + For each stage why we need them?

## Intro of R

* Different between arrow (<-) and equal?
  + Answer: (<-)use for apply value, = use in function parameter passing
* What is naming restriction?(4)
* Definition of?
  + Vectors
  + Matrices
  + Arrays (what is different between vector and array?)
  + List( different between it and vector?)
  + Data Frames

# W2

## Basic R

* Subsetting (page 4)
* 
  + Explain logic and what it will return result
    - a[-c(2:nrow(mpg)), ]
    - a[c(T,F,T),]
    - a$manufacturer == "audi"
  + Operators?
    - %/%
    - %%
    - %\*%
    - %in%
  + Ifelse()?
  + Next vs break?

## exploratory data analysis

* What Is EDA
* Why it is iterative cycle?
* What is the goal abt it?
* Term of EDA
  + What is
    - Variable
    - Value
    - Observation
    - Tabular data
* Variation? Covariation? Different between them

# W3

* (you need to instant has graph picture in your mind)
* When should we use logarithmic scale
* Histogram vs bar chart?
* Density (you forgot what is )
* Boxplot
  + Median
  + Inter-quartile range
    - What is the range?
  + Upper quartile
  + Lower quartile
  + Whiskers
* Why median and IQR are better than mean and standard deviation?
* Bar chart
  + What it records?
  + What is the different between it and histogram?
* Dot
  + But why bar is misleading?

# W4

* What is stat, what stat short for?
* What is stat = "identity" actually do in

# W5

* mutate()?
* na\_if()?
  + How to use them?
* %>%
* What is NA representing about?
* What boxplot.stats() return about?
* What strategies we treat missing value?(2)
* 2 methods dealing with missing values
  + When face large portion of data is NA spread throught data what should we consider (2)
    - One
    - Two (2; this one can be randomly or systematically) and explain why
* What is vtreat
  + How to use?(page 37)
* For date function
  + julian()?

# W6

* What is normalization?

# Mock mid

* method of create list/matrix/array?
* What is tile plot?
* hexbin plot?
* what is mutable joins? Vs full joins?
* identity vs dodge vs fill vs stack?
* As.factor() is what?
* Floor()?
* After exam:
  + Dataframe vs list vs array
  + Barplot vs histogram
  + Density plot