Business Analytics

Session 2

Ing. Juan José Franklin Uraga, PhD.

MODELO EDUCATIVO TEC 21

Session 1

Tecnológico de Monterrey

- **Data Wrangling**
- **Class evidence # 2**
- **Homework review**

Statistics Review

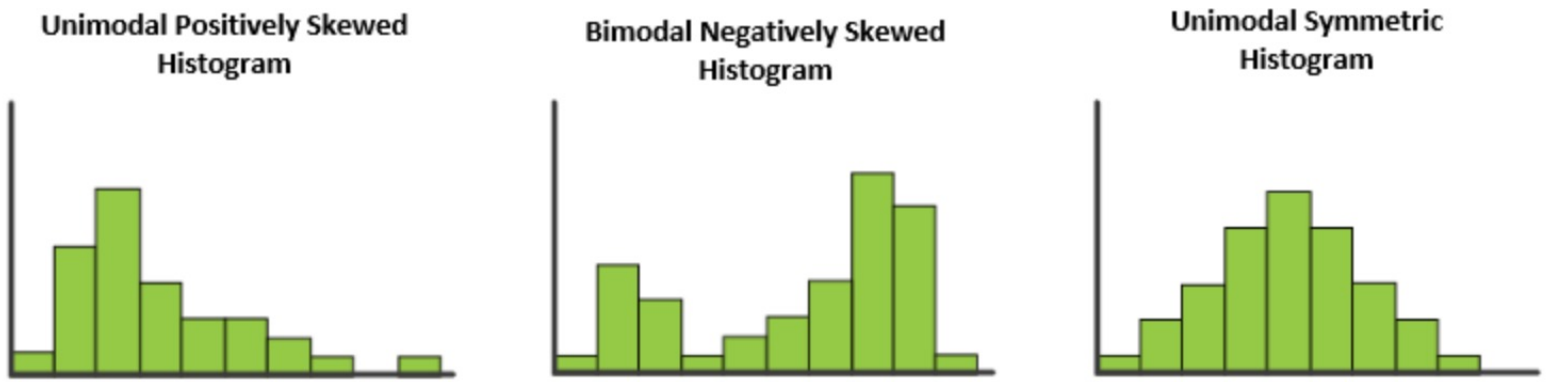
Statistics

- It is the art and science of collecting data, analyzing it, presenting it and interpreting it.

Histogram

Tecnológico de Monterrey

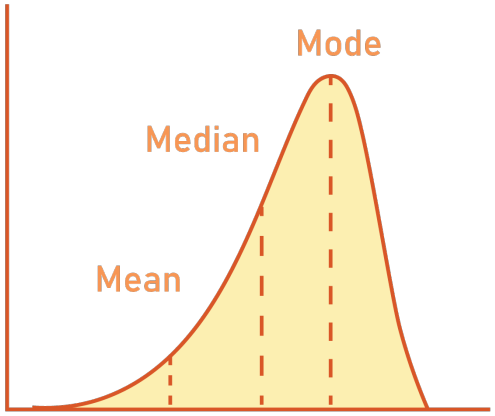
- It is used to represent the frequencies of the observed values of a quantitative variable.
- It constitutes a graphical way of visualizing the behavior and trends of the variable, detecting atypical data and forms of a frequency distribution.



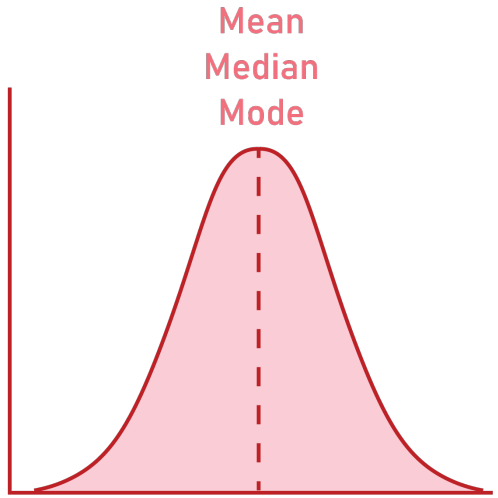
Shapes of numerical distributions

Tecnológico de Monterrey

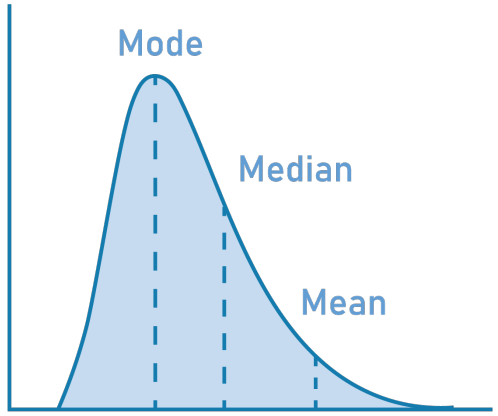
- It describes how data are distributed.



Left skew



Normal distribution
Symmetric



Right skew

Central tendency measures

Tecnológico de Monterrey

MEAN OR AVERAGE

the sum of the numbers divided by the amount of numbers

$5+5+5+6+7+7+14$

$49/7 = 7$

MEDIAN

the number in the middle

5 5 5 **6** 7 7 14

(numbers must be in ascending order)

MODE

the number that appears the most

5 5 5 6 7 7 14

Mean

- Arithmetic mean or average.

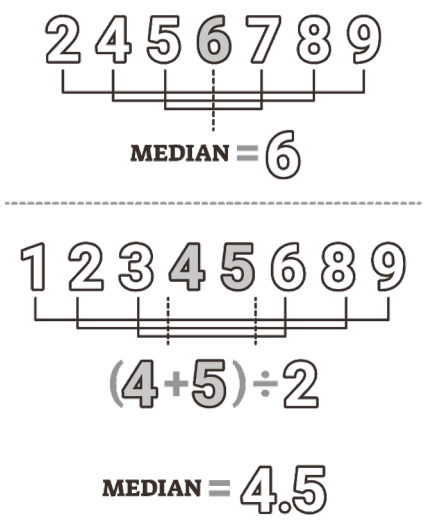
$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

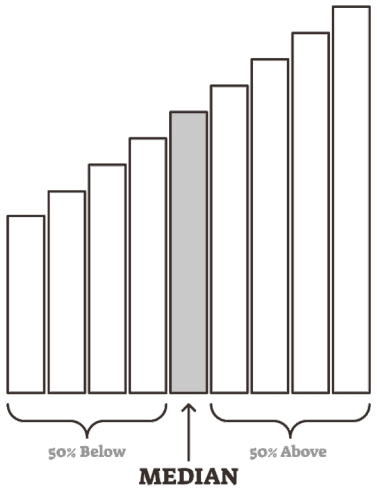
- Each one of the observed values
- sample size

=PROMEDIO(dataset)

Median

- Finding the median value: position (n+1)/2 of the **sorted dataset**.
- If the number of values is odd, the median is the middle number.
- If the number of values is even, the median is the average of the two middle numbers.





=MEDIANA(dataset)

Mean vs Median

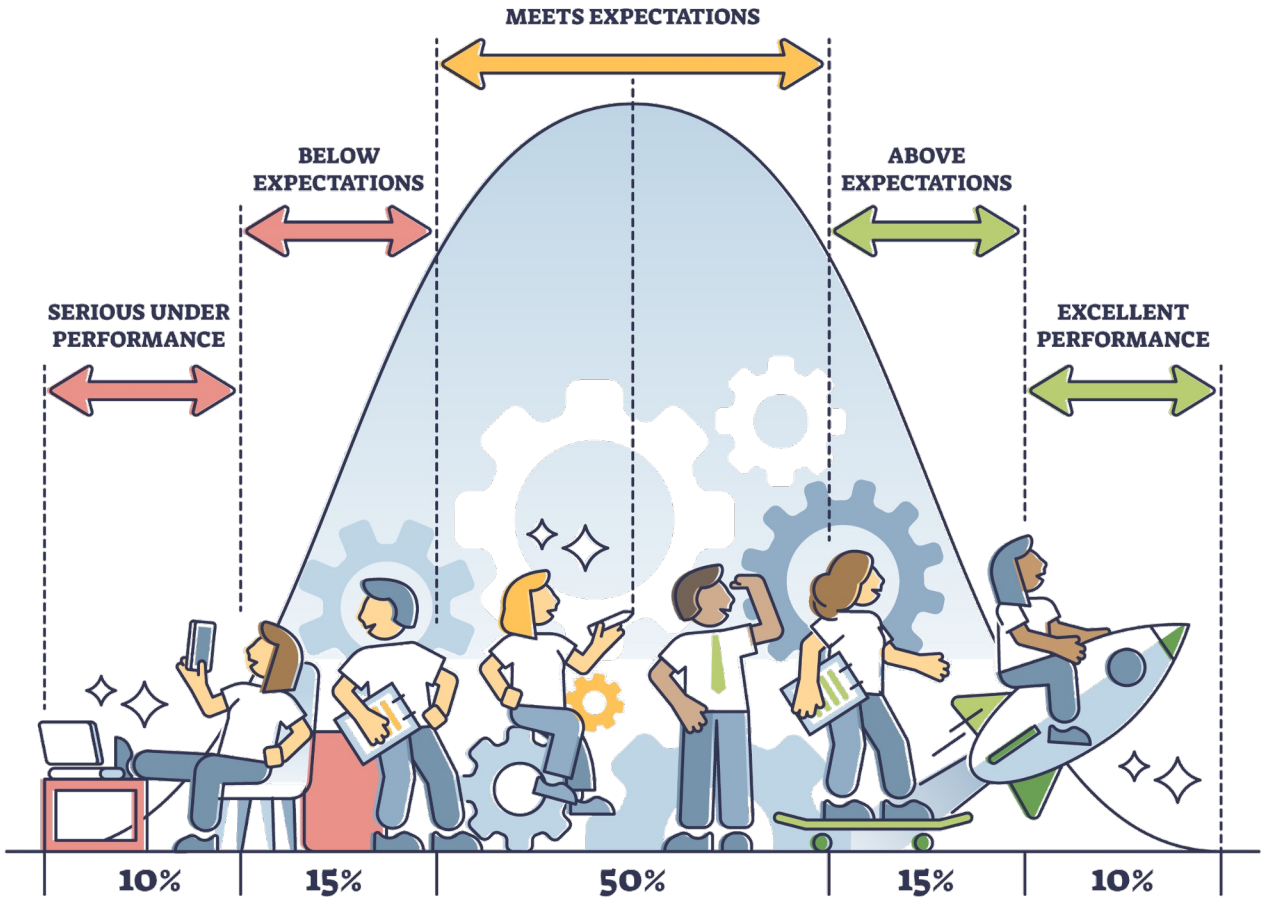
Tecnológico de Monterrey

- Month income (USD) is analyzed for a group of friends:
  - $1000, $1250, $1300, $1500, $1600, $2000, $70,000
  - Mean: $11,236
  - Median: $1,500
- Which one is **better** to represent this dataset?
- The **mean is heavily influenced** by extreme values.
- **Median** will be a better option when dataset is **skewed** or under the presence of **outliers**.

Probability distributions

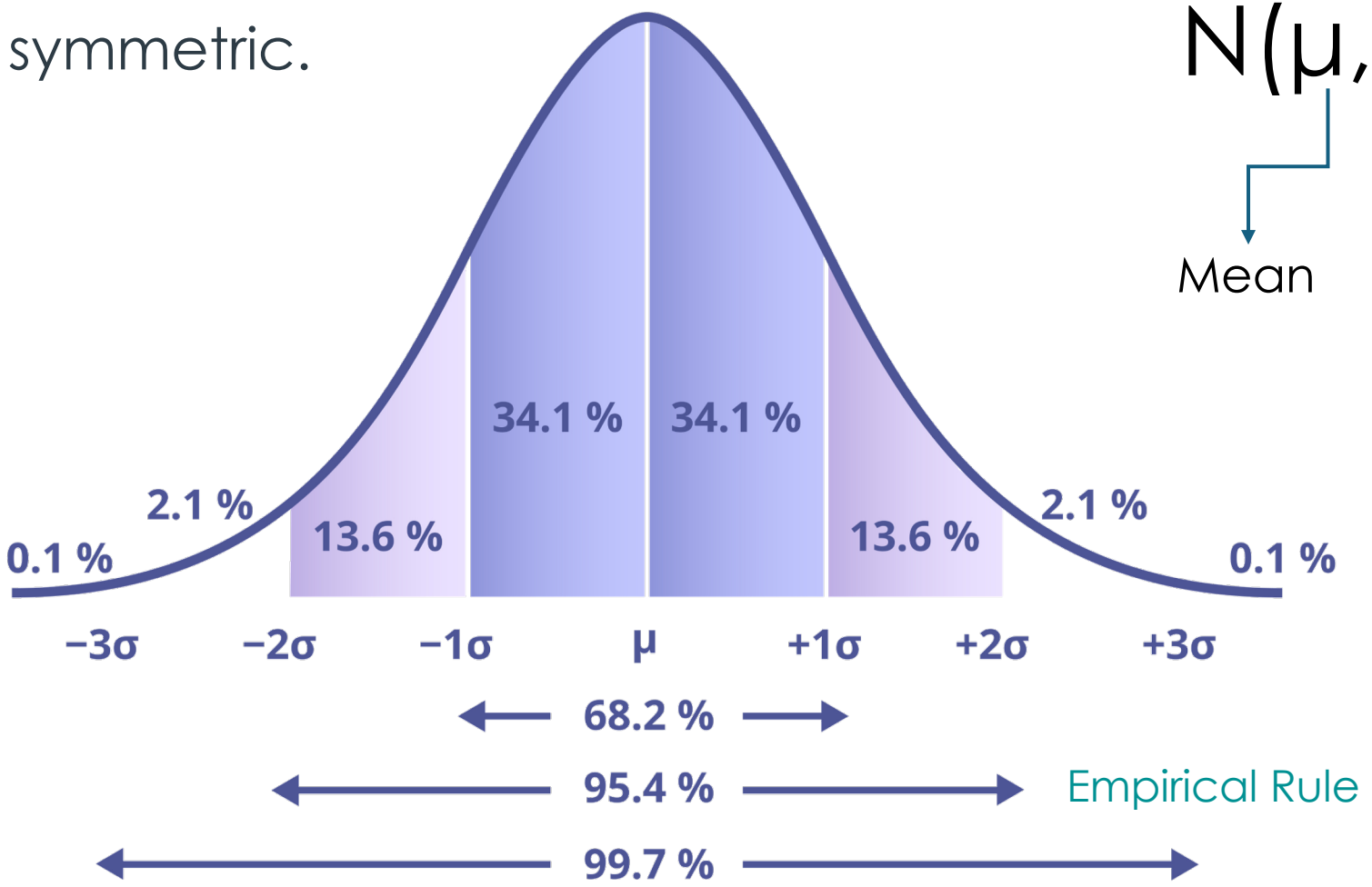
Normal distribution

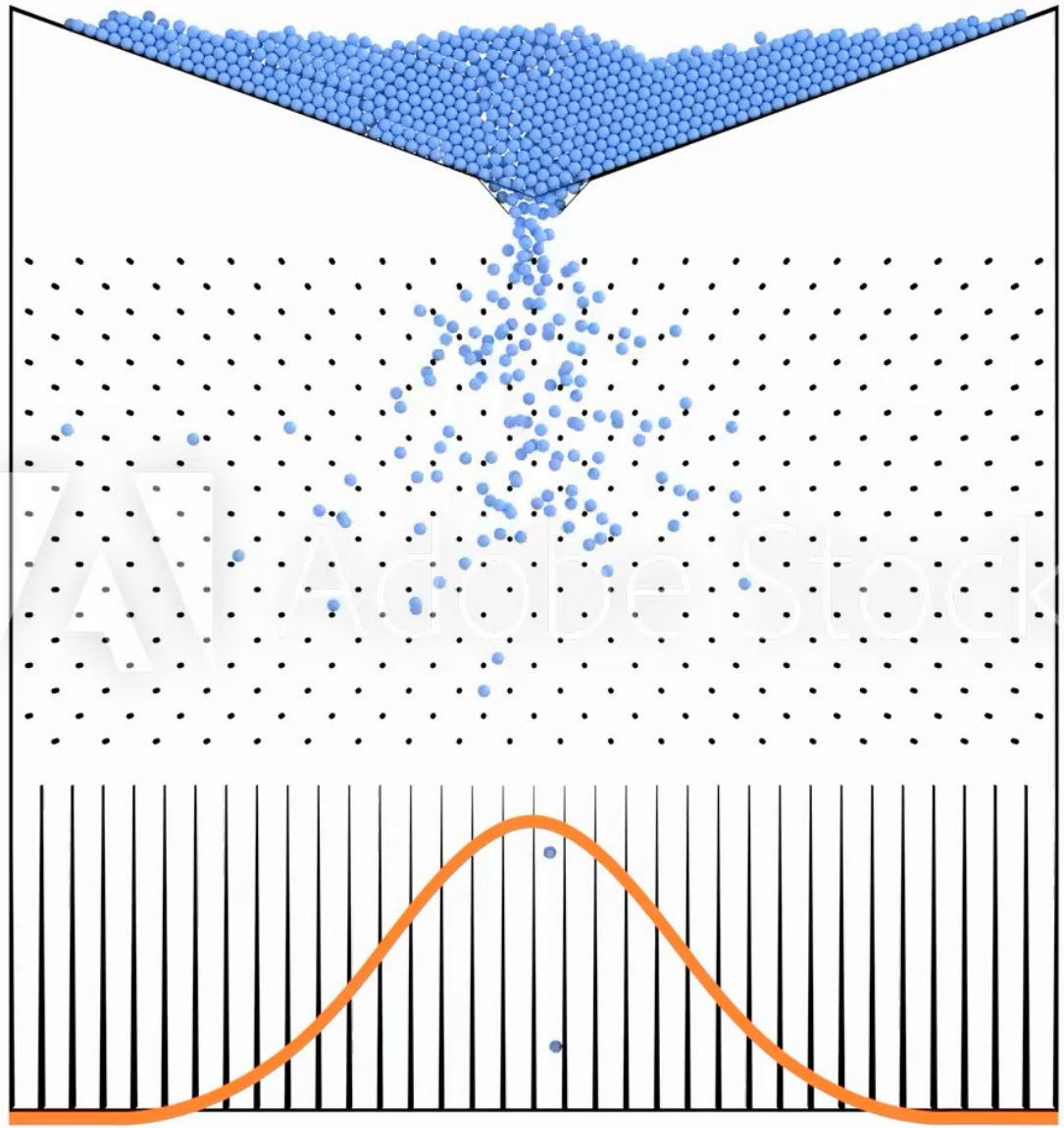
- Many variables in nature are normally distributed.



Normal distribution

- Unimodal and symmetric.
- Bell-shaped.





Chebysev Rule

Tecnológico de Monterrey

- Regardless of how the data are distributed, at least $(1 - 1/k^2)$ x 100% of the values will fall within k standard deviations of the mean (for k > 1)
  - Examples:

At least		within
$(1 - 1/1^2) \times 100\% =$	0%	k=1 $(\mu \pm 1\sigma)$
$(1 - 1/2^2) \times 100\% =$	75%	k=2 $(\mu \pm 2\sigma)$
$(1 - 1/3^2) \times 100\% =$	89%	k=3 $(\mu \pm 3\sigma)$

Business Analytics

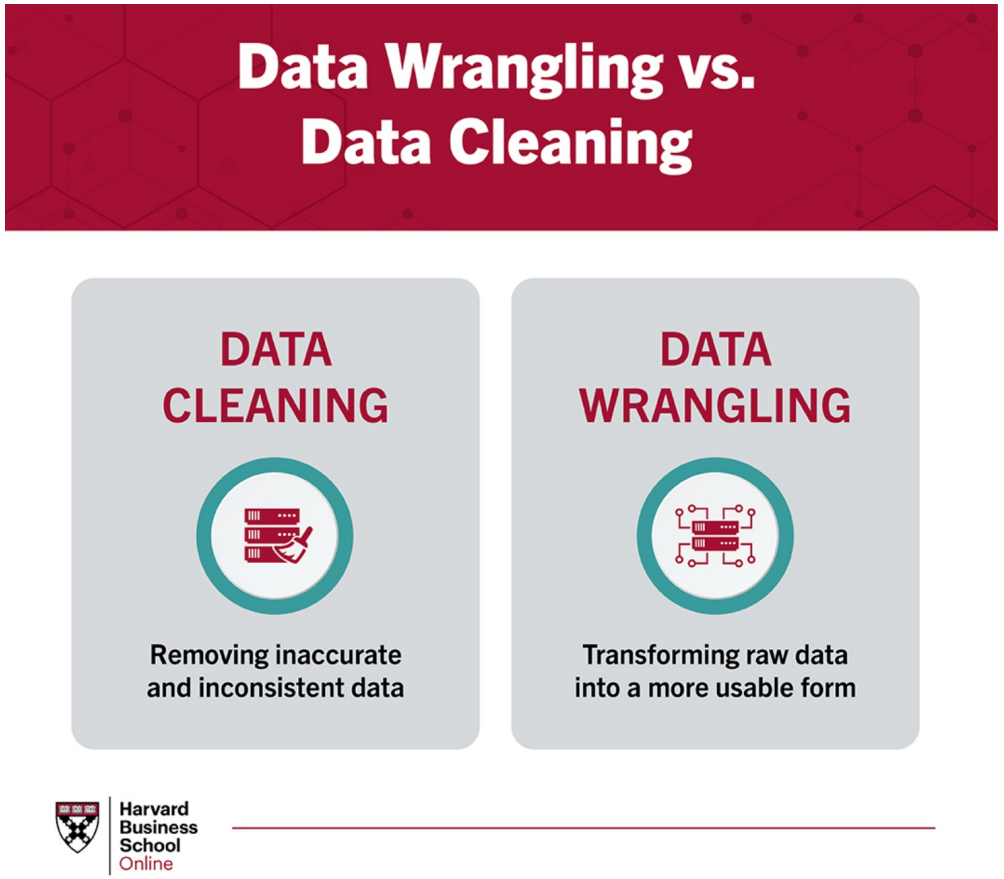
Data wrangling With R Studio

Ing. Juan José Franklin Uraga, PhD.

Data Wrangling

Tecnológico de Monterrey

- Data wrangling—also called data cleaning, data remediation, or data munging—refers to a variety of processes designed to transform raw data into more readily used formats.



Objectives for today

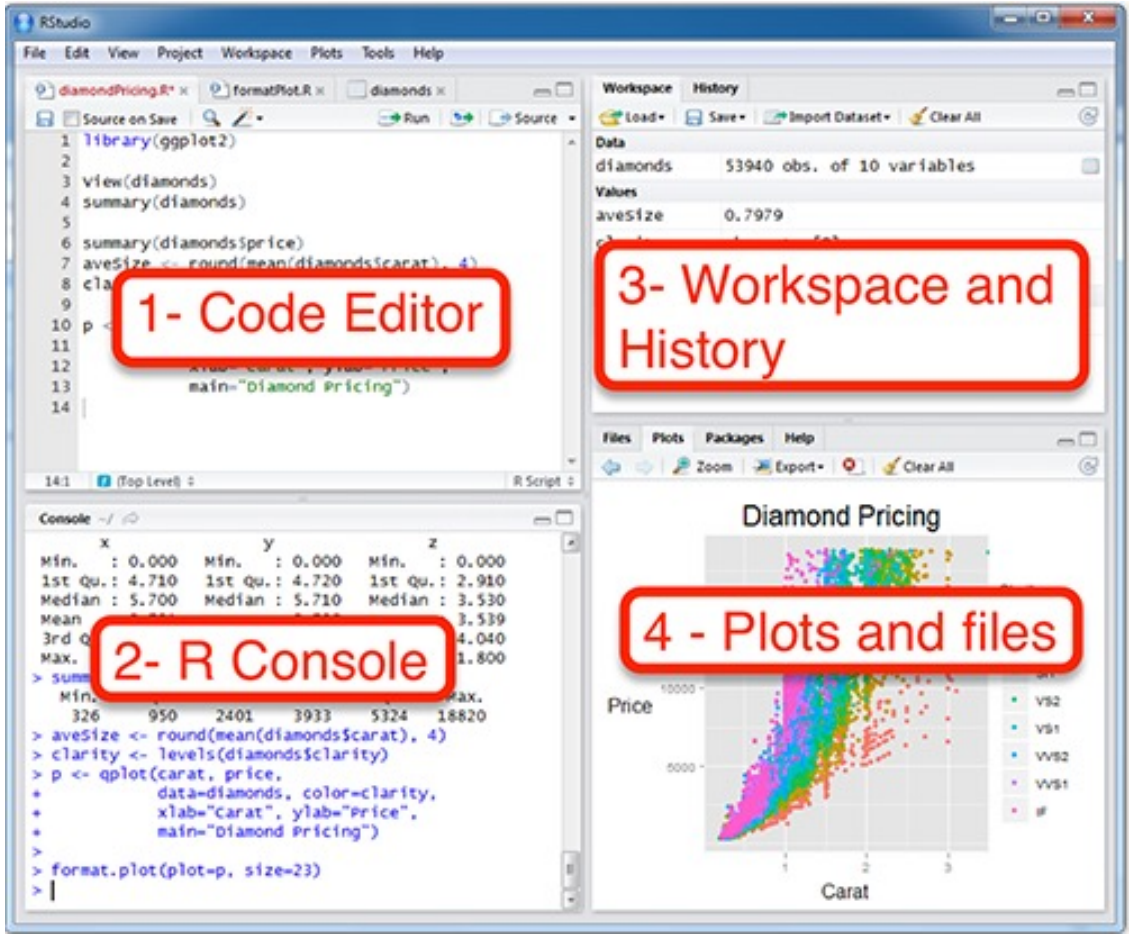
- Pipe operator %>% (and then…)
- Filter
- Select
- Group_by
- Count
- Arrange
- Summarise
- Mutate
- Ggplot (histograms with facets)

Objectives for today

Start working with R Markdown

- R Markdown is a special type of document that mixes together regular text with R code.
- In an R Markdown document, you can write explanations, make lists, or add images just like in any other document.
- But what's special is that you can also include pieces of R code directly in the document.
- When you "knit" or compile an R Markdown document, it runs the R code and inserts the results back into the document.

The R studio 4 windows



Filter

Tecnológico de Monterrey

```r
library(tidyverse)
starwars %>%
  filter(homeworld=="Tatooine") %>%
  filter(height>170)
```

A tibble: 6 × 14

name <chr>	height <int>	mass <dbl>	hair_color <chr>
Luke Skywalker	172	77	blond
Darth Vader	202	136	none
Owen Lars	178	120	brown, grey
Biggs Darklighter	183	84	black
Anakin Skywalker	188	84	blond
Cliegg Lars	183	*NA*	brown

Select

Tecnológico de Monterrey

```r
starwars %>%
  select(height,mass,homeworld,species,ends_with("color"))
```

A tibble: 87 × 7

height <int>	mass <dbl>	homeworld <chr>	species <chr>	hair_color <chr>
172	77.0	Tatooine	Human	blond
167	75.0	Tatooine	Droid	*NA*
96	32.0	Naboo	Droid	*NA*

```r
starwars1 <- starwars %>%
  select(height,mass,homeworld,species,ends_with("color"))
```

<- Creates an object with the selection.

Group_by

Tecnológico de Monterrey

```r
starwars %>%
  group_by(species) %>%
  count() %>%
  arrange(desc(n))
```

A tibble: 38 × 2 | Groups: species [38]

species <chr>	n <int>
Human	35
Droid	6
NA	4
Gungan	3
Kaminoan	2
Mirialan	2
Twi'lek	2

Dealing with NA values

Tecnológico de Monterrey

```
is.na(starwars$species)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
## [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
## [85] FALSE FALSE FALSE
```

```
starwars[is.na(starwars$species),]
```

```
## # A tibble: 4 × 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>           <dbl> <chr> <chr>
## 1 Jek Tono…    180   110 brown      fair       blue               NA <NA>  <NA>
## 2 Gregar T…    185    85 black      dark       brown              NA <NA>  <NA>
## 3 Cordé        157    NA brown      light      brown              NA <NA>  <NA>
## 4 Sly Moore    178    48 none       pale       white              NA <NA>  <NA>
## # ℹ 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

Group_by, not considering NA

Tecnológico de Monterrey

```r
starwars %>%
  drop_na(species) %>%
  group_by(species) %>%
  count() %>%
  arrange(desc(n))
```

A tibble: 37 × 2 | Groups: species [37]

species <chr>	n <int>
Human	35
Droid	6
Gungan	3
Kaminoan	2
Mirialan	2
Twi'lek	2

Group_by & summarise

Tecnológico de Monterrey

Task: find the mean heigth by species

```r
starwars %>%
  drop_na(species) %>%
  group_by(species) %>%
  summarise("Mean height"=mean(height))
```

A tibble: 37 × 2

species <chr>	Mean height <dbl>
Aleena	79.0000
Besalisk	198.0000
Cerean	198.0000
Chagrian	196.0000
Clawdite	168.0000
Droid	*NA*
Dug	112.0000
Ewok	88.0000

Group_by & summarise & arrange

Tecnológico de Monterrey

Task: find the mean heigth by species

```r
starwars %>%
  drop_na(species) %>%
  group_by(species) %>%
  drop_na(height) %>%
  summarise(Mean_height=mean(height)) %>%
  arrange(desc(Mean_height))
```

A tibble: 37 × 2

species <chr>	Mean_height <dbl>
Quermian	264.0000
Wookiee	231.0000
Kaminoan	221.0000
Kaleesh	216.0000
Gungan	208.6667
Pau'an	206.0000

Alternative

```r
starwars %>%
  drop_na(species) %>%
  group_by(species) %>%
  summarise(Mean_height=mean(height,na.rm=TRUE)) %>%
  arrange(desc(Mean_height))
```

Mutate & case_when

Tecnológico de Monterrey

Task: create a new variable in starwars data set. The name is "size". Get a summary of mass, first.
If mass < 40, classify it as small, if <100 it should be regular, more than 100: big.

Notice the customized order

```r
starwars %>%
  drop_na(mass) %>%
  mutate(Size=case_when(mass<40~"Small",
                        mass<100~"Regular",
                        mass>100~"Big")) %>%
  select(name,species,Size) %>%
  arrange(match(Size,c("Small","Regular","Big")))
```

A tibble: 59 × 3

name <chr>	species <chr>	Size <chr>
R2-D2	Droid	Small
R5-D4	Droid	Small
Yoda	Yoda's species	Small
Wicket Systri Warrick	Ewok	Small
Ratts Tyerel	Aleena	Small
Luke Skywalker	Human	Regular

Task

1. Create a new df of starwars with only human and droids.
2. Classify them as small or tall, depending on their height (<180 small, >= 180 tall).
3. Create a contingency table that shows how many humans and droids are short and tall. In absolute frequency and relative frequency.
4. Create a mosaic plot to present the information.

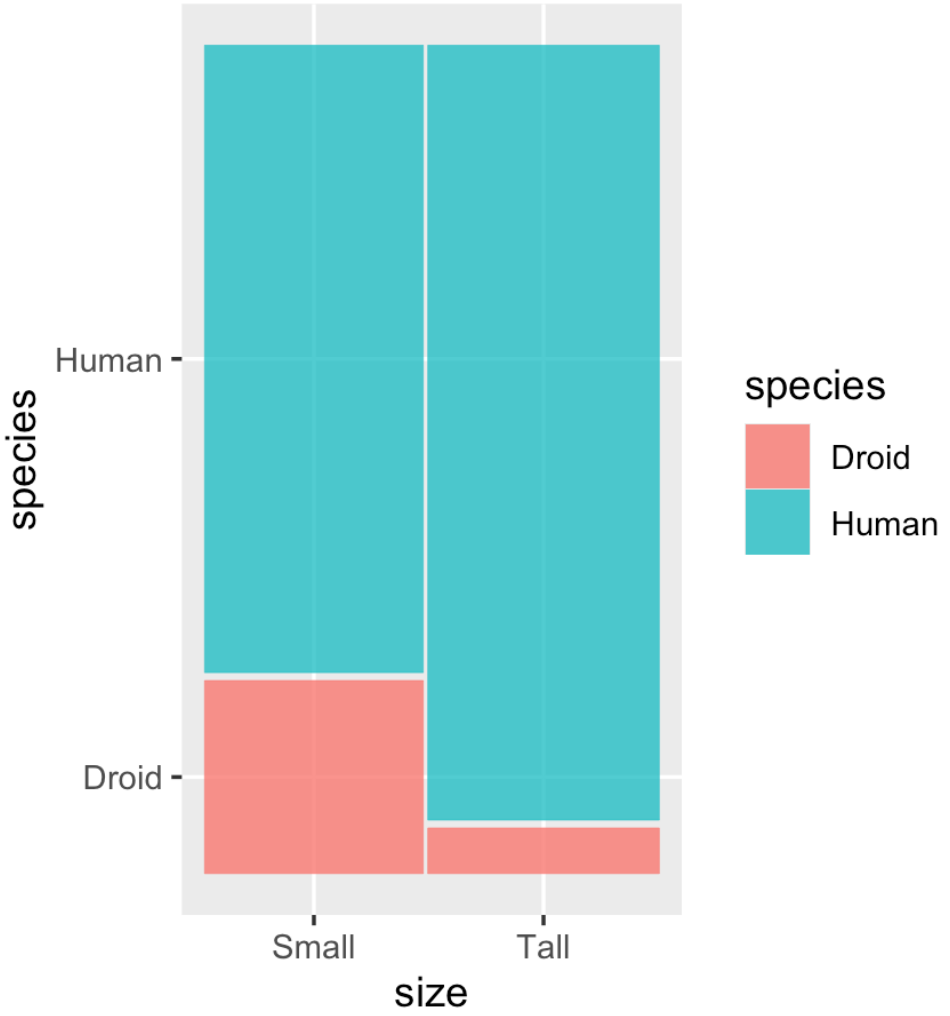
Task

```
starwars3 <- starwars %>%
  filter(species=="Human" | species=="Droid") %>%
  mutate(size=case_when(height<180~"Small",
                        height>=180~"Tall"))

table(starwars3$species,starwars3$size) %>%
  prop.table() %>%
  addmargins()

starwars3 <- starwars3 %>%  drop_na(size)

ggplot(starwars3) + geom_mosaic(aes(x = product(species, size), fill = species))
```



Of the total population:

14% are droids

86% are humans

- 11% are small droids
- 3% are tall droids
- 37% are small humans
- 49% are tall humans.


Statistics
organization
scientific
descriptive
analysis
samples
collection
interpretation
data
statistical
Inference
theory
mathematical
called
experiments
science
population
prediction
probability
discipline
pertaining
including
referring
used
statistic
total
mathematics
parameters
provides
inferential
consider
also
concerned
word
wide
design
branch
well
often
ways
improve
process
working
useful
social
considered
deduce
part
problems
soundness
accounts
experience
application
terms
empirical
summarize
successful
element
necessary
using
logically
modeled
expertise
people
thinking
gained
models direct
planning
companies
usually
sciences
comprise
applications
aspects
inferences
applied
difference
variety
studied
interpreting
randomness
based
starts
natural
guiding
art
organizations
moves
others
versed
someone
help
statistician
vital
draw
prove
theoretical
collecting
since
larger
roots
wrong
given
survey
confused
one
provide
grouped
inductive
observations
set
hypothesis
disciplines
mean
presentation
methods
consultants
focus
probabilities
misleading
may
new
academic
way
opposite
particular
sampling
particularly
increase
forecasting
Statisticians
explanation
subject
predictions
basis
holds
roughly
whose
distinct
closely
research
seems
singular
true
results
government
direction
deals
number
together
method
patterns
communicating
uncertainty
calculated
quality
plural
rather
fields
related
available
without
advance
leads
a.k.a
applicable
business
access
median
quantity
tools
addition
relevant
predictive
describe
tested
surveys