Business Analytics

Session 3

Ing. Juan José Franklin Uraga, PhD.

MODELO EDUCATIVO TEC 21

Session 3

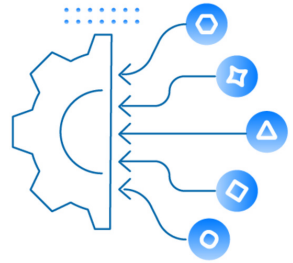
Tecnológico de Monterrey

- **Data Cleaning & Data Wrangling**
- **Class evidence # 3**
- **Homework review**

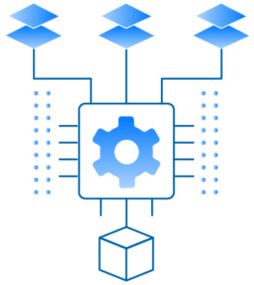
Data
Collection

Data
Cleaning

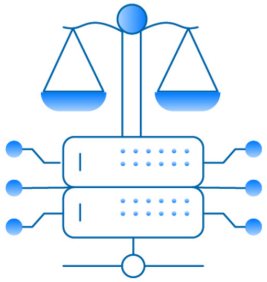
Data
Wrangling



Data
Transformation

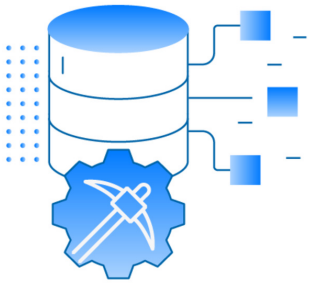


Data
Modeling



Data
Ethics

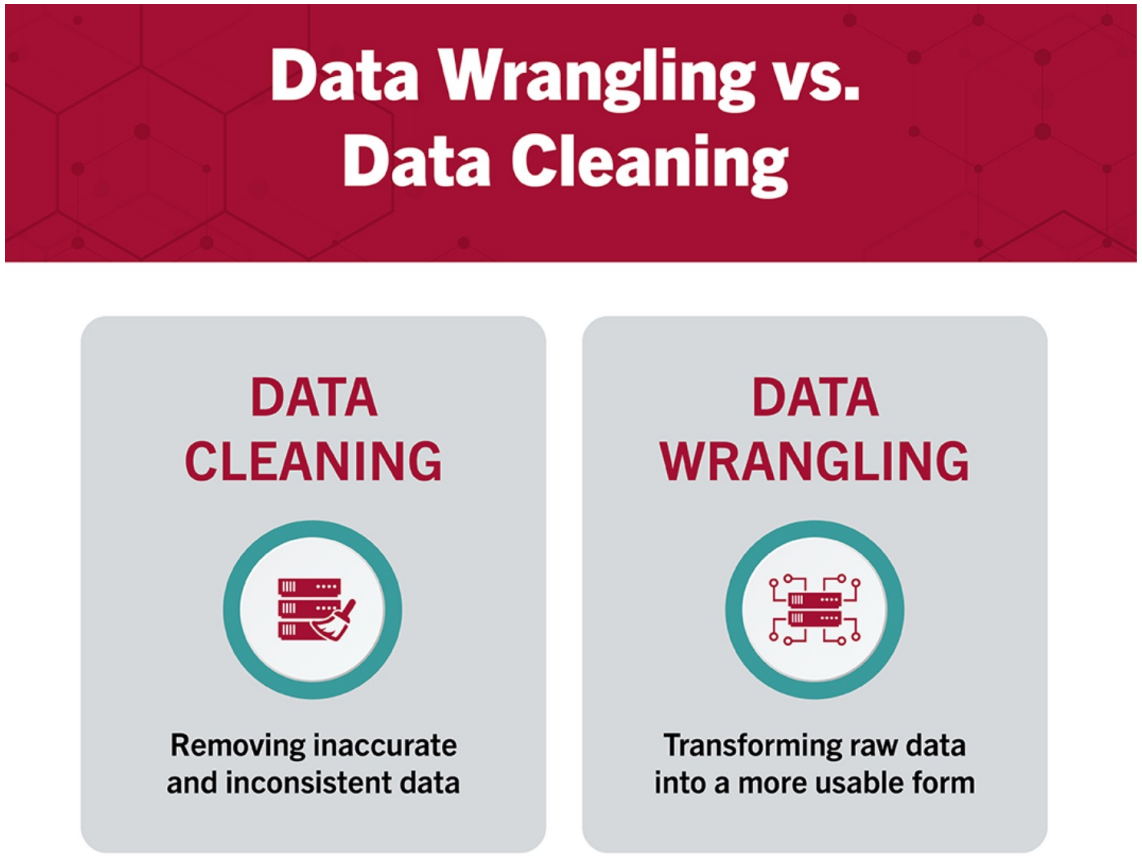
Data
Storytelling



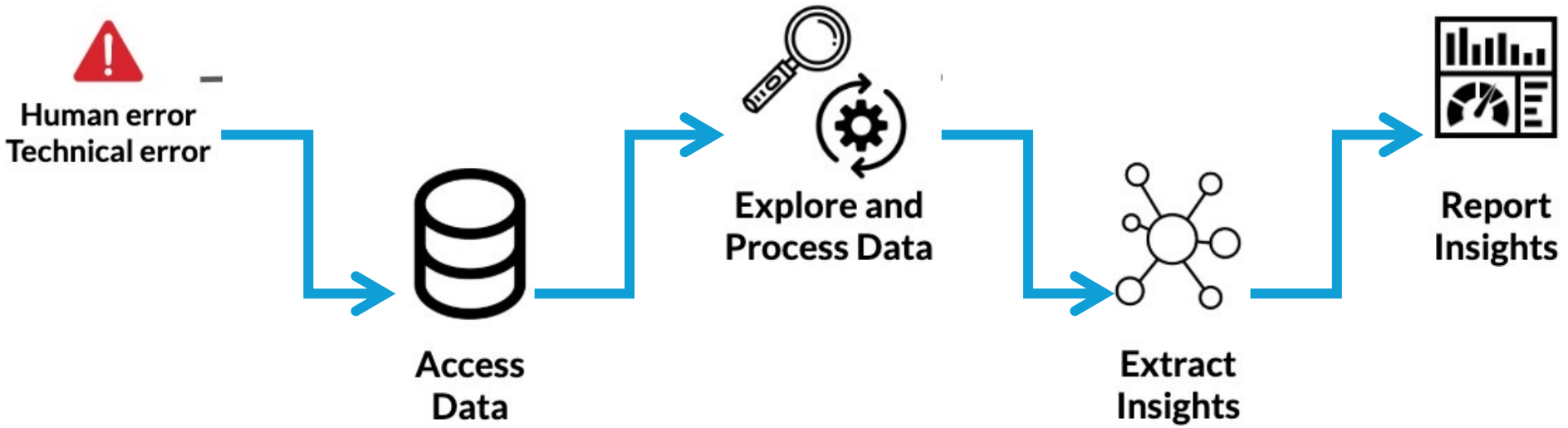
Data
Mining

Finding missing values

Tecnológico de Monterrey



Importance of cleaning data



Data Analysts spend **80%** of their time cleaning and manipulating data and only **20%** of their time analyzing it. The time spent cleaning is vital since analyzing dirty data can lead you to **draw inaccurate conclusions.**

Tecnológico de Monterrey

Finding missing values

RStudio®

Task 1- description

Tecnológico de Monterrey

Data set: starwars

1. Determine how many different hair colors are available.
2. Select only the variables name, gender, hair_color and height.
3. For those variables, omit any NA.

```
starwars %>%
  select(name,gender,hair_color,height) %>%
  na.omit()
```

Not the best idea, **first**, find out where your NA's are.

RStudio®

Task 1

Tecnológico de Monterrey

```
starwars %>%
  select(name,gender,hair_color,height) %>%
  filter(!complete.cases(.)) %>%
  View()
```

Now:

1. Drop NA for height.
2. Change the NA's of hair_color to none.

```
# A tibble: 15 × 4
   name                gender    hair_color height
   <chr>               <chr>     <chr>       <int>
 1 C-3PO               masculine NA            167
 2 R2-D2               masculine NA             96
 3 R5-D4               masculine NA             97
 4 Greedo              masculine NA            173
 5 Jabba Desilijic Tiure masculine NA          175
 6 Jek Tono Porkins    NA        brown         180
 7 Arvel Crynyd        masculine brown          NA
 8 Gregar Typho        NA        black         185
 9 Cordé               NA        brown         157
10 Sly Moore           NA        none          178
11 Finn                masculine black          NA
12 Rey                 feminine  brown          NA
13 Poe Dameron         masculine brown          NA
14 BB8                 masculine none           NA
15 Captain Phasma      feminine  none           NA
```

RStudio®

Task 1

Tecnológico de Monterrey

```
starwars %>%
  select(name,gender,hair_color,height) %>%
  filter(!complete.cases(.)) %>%
  drop_na(height) %>%
  mutate(hair_color=replace_na(hair_color,"none"))
```

```
# A tibble: 9 × 4
  name                 gender    hair_color height
  <chr>                <chr>     <chr>       <int>
1 C-3PO                masculine none          167
2 R2-D2                masculine none           96
3 R5-D4                masculine none           97
4 Greedo               masculine none          173
5 Jabba Desilijic Tiure masculine none         175
6 Jek Tono Porkins     NA        brown         180
7 Gregar Typho         NA        black         185
8 Cordé                NA        brown         157
9 Sly Moore            NA        none          178
```

RStudio®

Task 2

Tecnológico de Monterrey

Finally, create an object (sw2) that will not show any NA's, except for gender. Create a boxplot that shows height related to gender.

```
sw2 <- starwars %>%
  select(name,gender,hair_color,height) %>%
  drop_na(height) %>%
  mutate(hair_color=replace_na(hair_color,"none"))
```

	name	gender	hair_color	height
13	Chewbacca	masculine	brown	228
14	Han Solo	masculine	brown	180
15	Greedo	masculine	none	173
16	Jabba Desilijic Tiure	masculine	none	175
17	Wedge Antilles	masculine	brown	170
18	Jek Tono Porkins	*NA*	brown	180
19	Yoda	masculine	white	66
20	Palpatine	masculine	grey	170
21	Boba Fett	masculine	black	183
22	IG-88	masculine	none	200
23	Bossk	masculine	none	190
24	Lando Calrissian	masculine	black	177
25	Lobot	masculine	none	175

RStudio®

Tecnológico de Monterrey

Removing patterns and changing types of variables

Data Wrangling

RStudio®

Data types in R

Data type	**Example**	**Rdata type**
Text	First name, last name, address	Character
Integer	Subscriber count, # products sold	Integer
Decimal	Temperature, exchange rate,	Numeric
Binary	Is married, new customer, yes/no,	Logical True False / Boolean
Category	Marriage status, color, …	factor
Date	Order dates, date of birth, …	Date

Task 3 - description

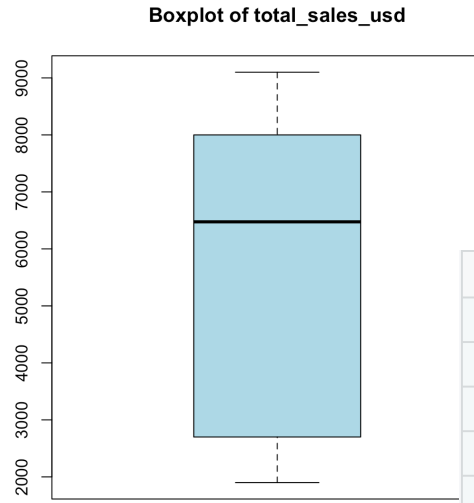
Tecnológico de Monterrey

Compute the mean for the variables **total_sales** that belongs to the **sales.df** data frame.

(First, rename the column **sales** to **Total_sales**).

Functions: **str_remove** & **as.numeric**

```
sales.df2 <- sales.df %>%
  mutate(Total_sales_usd=as.numeric(str_remove(Total_sales,",")))
class(sales.df2$Total_sales_usd)
mean(sales.df2$Total_sales_usd)
```




```
> summary(sales.df2)
 Total_sales        Total_sales_usd
 Length:10          Min.   :1900
 Class :character   1st Qu.:2900
 Mode  :character   Median :6475
                    Mean   :5745
                    3rd Qu.:7950
                    Max.   :9100
```

	Total_sales_usd
1	2500
2	2700
3	5200
4	7800
5	9100
6	1900
7	7750
8	8000
9	3500
10	9000

RStudio®

Tecnológico de Monterrey

Finding and removing duplicate values

RStudio®

Task 4 - description

Tecnológico de Monterrey

For the **Session1S24** data set, find and drop any full duplicate. Also, find any partial duplicate.

```
s1part2 <- read_excel("Session1S24.xlsx")
s1part2 <- s1part2 %>%
  select(-c(`Name of the variable`,Commments))


duplicated(s1part2)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE
```

```
s1part2[duplicated(s1part2),]
```

```
## # A tibble: 2 × 13
##   Name    Age Program Semester Place_of_birth Workout_per_week Height Cat_or_dog
##   <chr> <dbl> <chr>      <dbl> <chr>                    <dbl>  <dbl> <chr>
## 1 Cris…    20 LCPF           6 Texas                       48   1.53 Dog
## 2 Vict…    23 LIT            6 Durango                    175   1.75 Dog
## # ℹ 5 more variables: Favourite_fruit <chr>, Pet_not_get <chr>,
## #   least_favourite_food <chr>, last_birthday_present <chr>,
## #   time_social_media <dbl>
```

```
s1part2a <- s1part2[!duplicated(s1part2),]

s1part2 %>%
  distinct()
```

RStudio®

Task 4

Tecnológico de Monterrey

Create a new object (s1part2a) that keeps only non duplicated values.

```
s1part2a <- s1part2[!duplicated(s1part2),]

s1part2 %>%
  distinct()
```

```
## # A tibble: 17 × 13
##    Name              Age Program Semester Place_of_birth Workout_per_week Height
##    <chr>           <dbl> <chr>      <dbl> <chr>                      <dbl>  <dbl>
##  1 Juan               41 IIS            9 Hidalgo                      180   1.72
##  2 Claudia            41 LIN            9 Hidalgo                      150   1.61
##  3 Tony               31 INA            9 Chiapas                      450   1.61
##  4 Mia                 7 NA             0 Hidalgo                       60   1.19
##  5 Miguel             21 LEM            6 Monterrey                    360   1.79
##  6 Luis               23 LIT            6 CDMX                         600   1.9
##  7 Cristina Arrie…    20 LCPF           6 Texas                         48   1.53
##  8 Luis Arturo        19 LIT            4 Cuernavaca                   300   1.81
##  9 Juan Cepeda        20 LIT            4 Edo. Mex                     180   1.82
## 10 Kathia R           20 LIT            4 Monterrey                    180   1.66
## 11 Manuel             19 LIT            4 Leon                         840   1.74
## 12 Silvia             22 LIT            4 Guasave                       60   1.55
## 13 Victor             23 LIT            6 Durango                      175   1.75
## 14 Alma               19 LIT            4 Oaxaca                        80   1.61
## 15 Kathia R           20 LIT            4 Monterrey                    180   1.66
## 16 Kathia R           24 LEM            6 Monterrey                    180   1.66
## 17 Juan               35 LAET           9 Hidalgo                       NA  NA
## # ℹ 6 more variables: Cat_or_dog <chr>, Favourite_fruit <chr>,
## #   Pet_not_get <chr>, least_favourite_food <chr>, last_birthday_present <chr>,
## #   time_social_media <dbl>
```

RStudio®

Task 4 – partial duplicates

Tecnológico de Monterrey

Find people with the same name and age

```
s1part2a %>%
  count(Name,Age) %>%
  filter(n>1)
```

```
## # A tibble: 1 × 3
##   Name       Age     n
##   <chr>    <dbl> <int>
## 1 Kathia R    20     2
```

RStudio®


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