Business Analytics

Session 4

Ing. Juan José Franklin Uraga, PhD.

MODELO EDUCATIVO TEC 21

Session 4

Tecnológico de Monterrey

Data Cleaning & Data Wrangling

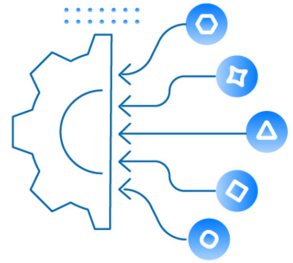
Class evidence # 4

Homework #02

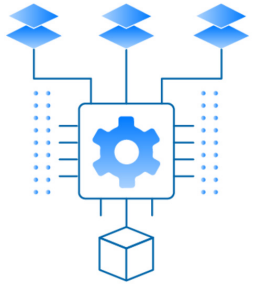
Data
Collection

Data
Cleaning

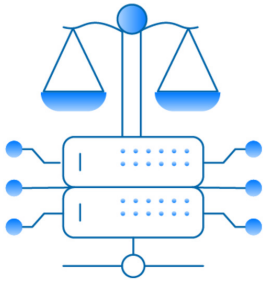
Data
Wrangling



Data
Transformation

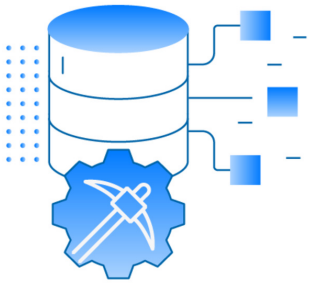


Data
Modeling



Data
Ethics

Data
Storytelling



Data
Mining

Tecnológico de Monterrey

Categorical variables inconsistencies

RStudio®

Task 4.1

Tecnológico de Monterrey

Data set: **Session4**

1. Explore the data set, in particular, determine how many different favorite fruits there are.

```
s4 <- read_excel("Session4.xlsx")
unique(s4$Favourite_fruit)
```

```
##  [1] "Watermelon" "apples"     "Apple"      "strawberry" "Banana"
##  [6] "Kiwi"       "Mango"      "Lichee"     "Strawberry" "Fig"
## [11] "Grape"      "grapes"     "grape"      "mango"      "MANGO"
## [16] "apple"      NA
```

```
table(s4$Favourite_fruit)
```

```
##
##      apple      Apple     apples     Banana        Fig      grape      Grape
##          1          4          1          1          2          1          1
##     grapes       Kiwi     Lichee      mango      Mango      MANGO strawberry
##          1          2          4          1          2          1          1
## Strawberry Watermelon
##          1          2
```

RStudio®

Task 4.1

Tecnológico de Monterrey

1. Use group_by to explore the data set, in particular, determine how many different favourite fruits there are.

```
s4 %>%
  group_by(Favourite_fruit) %>%
  count() %>%
  arrange(-n)
```

```
## # A tibble: 17 × 2
## # Groups:   Favourite_fruit [17]
##    Favourite_fruit     n
##    <chr>           <int>
##  1 Apple               4
##  2 Lichee              4
##  3 Fig                 2
##  4 Kiwi                2
##  5 Mango               2
##  6 Watermelon          2
##  7 <NA>                2
##  8 Banana              1
##  9 Grape               1
## 10 MANGO               1
## 11 Strawberry          1
## 12 apple               1
## 13 apples              1
## 14 grape               1
## 15 grapes              1
## 16 mango               1
## 17 strawberry          1
```

RStudio®

Task 4.1

Tecnológico de Monterrey

Data set: **Session4**

2. Change to lower case all of the fruits' names.

```
s4 %>%
  mutate(Fav_Fruit=str_to_lower(Favourite_fruit)) %>%
  group_by(Fav_Fruit) %>%
  count() %>%
  arrange(Fav_Fruit)
```

```
## # A tibble: 12 × 2
## # Groups:   Fav_Fruit [12]
##    Fav_Fruit      n
##    <chr>      <int>
##  1 apple          5
##  2 apples         1
##  3 banana         1
##  4 fig            2
##  5 grape          2
##  6 grapes         1
##  7 kiwi           2
##  8 lichee         4
##  9 mango          4
## 10 strawberry     2
## 11 watermelon     2
## 12 <NA>           2
```

RStudio®

Task 4.1

Tecnológico de Monterrey

3. Fix the apples and grapes issue.

```
s4 %>%
  mutate(Fav_Fruit=str_to_lower(Favourite_fruit)) %>%
  mutate(Fav_Fruit=if_else(Fav_Fruit=="apples","apple",Fav_Fruit)) %>%
  mutate(Fav_Fruit=if_else(Fav_Fruit=="grapes","grape",Fav_Fruit)) %>%
  group_by(Fav_Fruit) %>%
  count() %>%
  arrange(Fav_Fruit)
```

```
## # A tibble: 10 × 2
## # Groups:   Fav_Fruit [10]
##    Fav_Fruit      n
##    <chr>      <int>
##  1 apple          6
##  2 banana         1
##  3 fig            2
##  4 grape          3
##  5 kiwi           2
##  6 lichee         4
##  7 mango          4
##  8 strawberry     2
##  9 watermelon     2
## 10 <NA>           2
```

RStudio®

Task 4.1

Tecnológico de Monterrey

4. Drop the NA's and sort by descending n.

```
s4 %>%
  mutate(Fav_Fruit=str_to_lower(Favourite_fruit)) %>%
  mutate(Fav_Fruit=if_else(Fav_Fruit=="apples","apple",Fav_Fruit)) %>%
  mutate(Fav_Fruit=if_else(Fav_Fruit=="grapes","grape",Fav_Fruit)) %>%
  group_by(Fav_Fruit) %>%
  drop_na() %>%
  count() %>%
  arrange(-n)
```

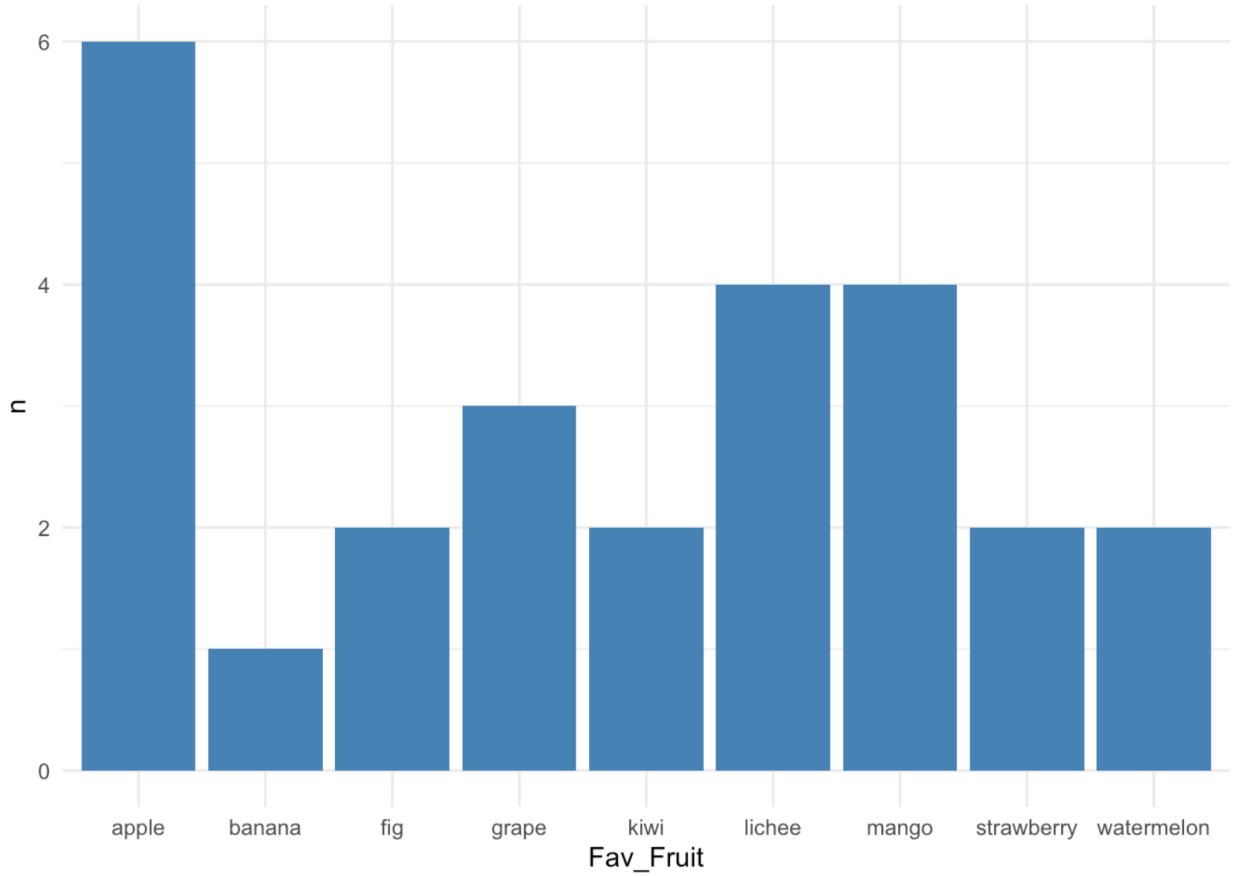
```
## # A tibble: 9 × 2
## # Groups:   Fav_Fruit [9]
##   Fav_Fruit      n
##   <chr>      <int>
## 1 apple          6
## 2 lichee         4
## 3 mango          4
## 4 grape          3
## 5 fig            2
## 6 kiwi           2
## 7 strawberry     2
## 8 watermelon     2
## 9 banana         1
```

RStudio®

Task 4.1

5. Create a new object (s4a) with the result and build a barplot to describe the dataset.

```
s4a %>%
  ggplot(aes(x=Fav_Fruit,y=n))+geom_col(fill="steelblue")+theme_minimal()
```

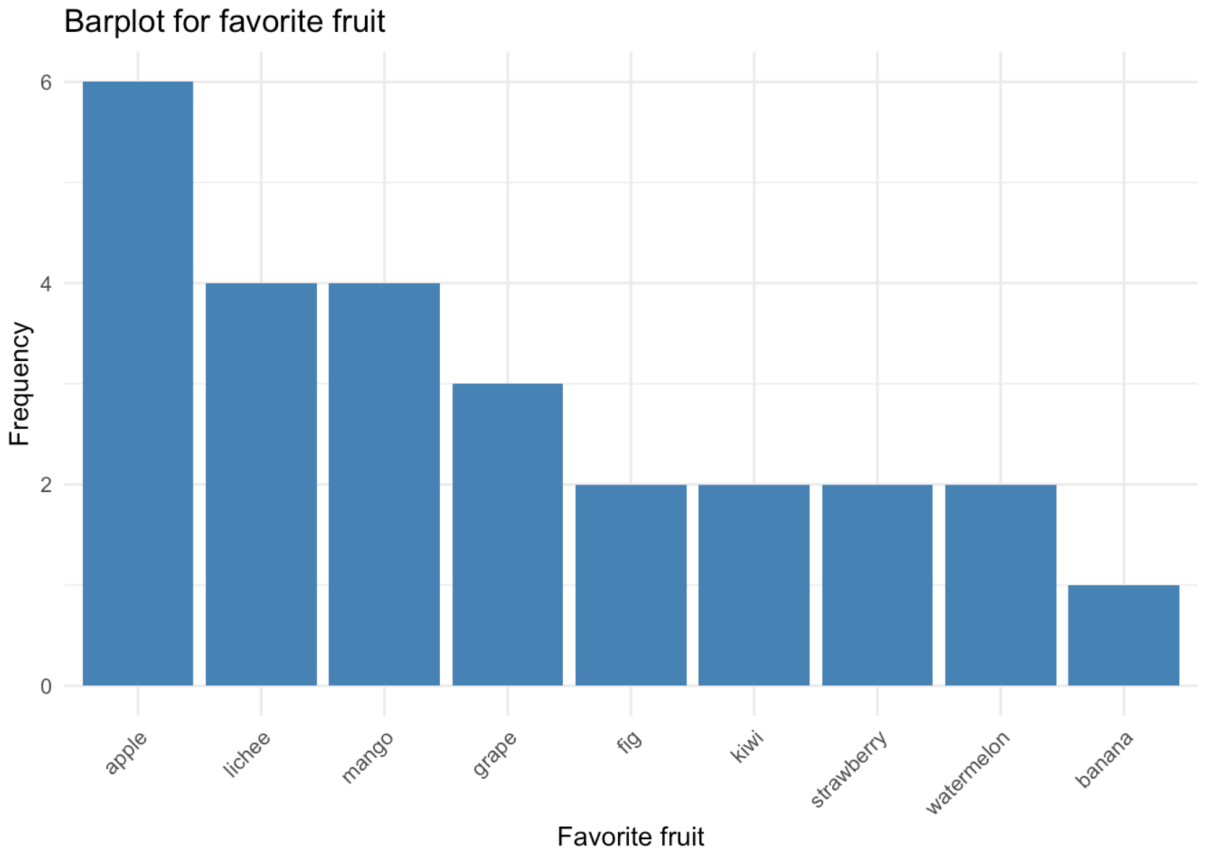


Task 4.1

Tecnológico de Monterrey

6. Sort the bars in descending order and add labels.

```
s4a %>%
  ggplot(aes(x=forcats::fct_reorder(Fav_Fruit,desc(n)),y=n))+geom_col(fill="steelblue")+theme_minimal()+labs(x="F
avorite fruit",y="Frequency",title="Barplot for favorite fruit")+theme(axis.text.x=element_text(angle=45,hjust=1,
vjust=1))
```



RStudio®

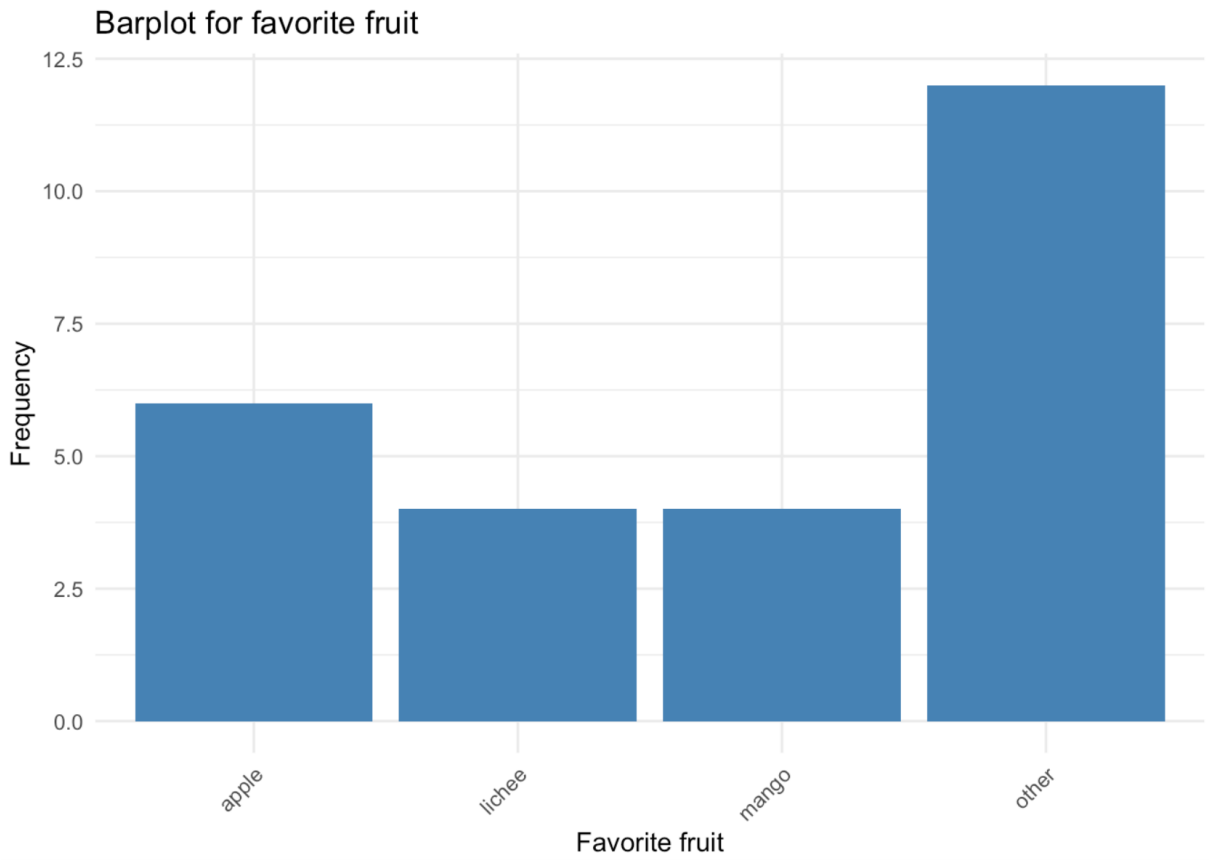
Task 4.1

Tecnológico de Monterrey

Data set: **session1**

7. Collapse the dataset: we only want to see apple, liche and mango vs all of the other fruits.

```
s4a_collapsed %>%
ggplot(aes(x=forcats::fct_reorder(Fav_Fruit_Collapsed,desc(n)),y=n))+geom_col(fill="steelblue")+theme_minimal()+l
abs(x="Favorite fruit",y="Frequency",title="Barplot for favorite fruit")+theme(axis.text.x=element_text(angle=45,
hjust=1,vjust=1))
```



RStudio®


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