Limpieza de datos Cuestionario

Said Muñoz Montero

8/8/2019

Cuestionario

Se realizó un cuestionario en el club de ciencias XAL2.

Para limpiar el dataset lo primero que tenemos que hacer es cargamos las bibliotecas que necesitamos.

```
library("dplyr")
library("tidyverse")
library("ggplot2")
library("broom")
library("reshape2")
library("chron")
library("tibble")
```

Leemos y vemos algunos parámetros.

```
cuestionario<-
  readr::read_csv(
    "https://raw.githubusercontent.com/said3427/XAL2_2019/dev/datos/cuestionario.csv")

head(cuestionario)
```

```
## # A tibble: 6 x 10
##   Timestamp  Edad Estatura  Peso Genero Futbol DeporteOpcional
##   <chr>     <dbl>    <dbl> <dbl> <chr>  <chr>  <chr>
## 1 08/05/19~    16     1.5     55 Mujer  No     <NA>
## 2 08/05/19~    18     1.78    80 Hombre No     Baile
## 3 08/05/19~    20     1.85    85 Hombre No     Baseball
## 4 08/05/19~    21     1.65    65 Mujer  Sí     Básquetbol
## 5 08/05/19~    22     1.73  1000 Mujer  No     Ninguno
## 6 08/05/19~    20     1.64    51 Mujer  Sí     Natación
## # ... with 3 more variables: GradoAcademico <chr>, SignoZodiacal <chr>,
## #   Sentimiento <dbl>
```

```
str(cuestionario)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 21 obs. of  10 variables:
##  $ Timestamp      : chr  "08/05/19 16:01" "08/05/19 16:01" "08/05/19 16:02" "08/05/19 16:03" ...
##  $ Edad           : num  16 18 20 21 22 20 18 16 21 19 ...
##  $ Estatura       : num  1.5 1.78 1.85 1.65 1.73 1.64 1.69 1.55 1.82 1.62 ...
##  $ Peso           : num  55 80 85 65 1000 51 65 60 84 54 ...
##  $ Genero         : chr  "Mujer" "Hombre" "Hombre" "Mujer" ...
##  $ Futbol         : chr  "No" "No" "No" "Sí" ...
##  $ DeporteOpcional: chr  NA "Baile" "Baseball" "Básquetbol" ...
##  $ GradoAcademico : chr  "Bachillerato" "Universidad, tercer semestre" "Estudiante de Lic." "Universi
##  $ SignoZodiacal  : chr  "Escorpio" "Escorpio" "Piscis" "Acuario" ...
##  $ Sentimiento    : num  8 6 10 4 10 8 4 5 9 9 ...
##  - attr(*, "spec")=
##   .. cols(
```

```
##   ..   Timestamp = col_character(),
##   ..   Edad = col_double(),
##   ..   Estatura = col_double(),
##   ..   Peso = col_double(),
##   ..   Genero = col_character(),
##   ..   Futbol = col_character(),
##   ..   DeporteOpcional = col_character(),
##   ..   GradoAcademico = col_character(),
##   ..   SignoZodiacal = col_character(),
##   ..   Sentimiento = col_double()
##   .. )
```

```
summary(cuestionario)
```

```
##   Timestamp              Edad          Estatura          Peso
##  Length:21          Min.   :16.00   Min.   :1.500   Min.   :  46.0
##  Class :character   1st Qu.:17.00   1st Qu.:1.620   1st Qu.:  60.0
##  Mode  :character   Median :20.00   Median :1.690   Median :  65.0
##                     Mean   :19.48   Mean   :1.689   Mean   : 110.6
##                     3rd Qu.:21.00   3rd Qu.:1.780   3rd Qu.:  80.0
##                     Max.   :24.00   Max.   :1.850   Max.   :1000.0
##     Genero             Futbol          DeporteOpcional
##  Length:21          Length:21          Length:21
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  GradoAcademico     SignoZodiacal       Sentimiento
##  Length:21          Length:21          Min.   : 4.000
##  Class :character   Class :character   1st Qu.: 6.000
##  Mode  :character   Mode  :character   Median : 8.000
##                                        Mean   : 7.571
##                                        3rd Qu.: 9.000
##                                        Max.   :10.000
```

Arreglando Timestamp

```
cuestionario <-
  cuestionario %>% mutate(Timestamp= as.chron(Timestamp,format = "%m/%d/%Y %H:%M"))
```

Arreglando Deportes opcionales

```
cuestionario %>% distinct(DeporteOpcional)
```

```
## # A tibble: 15 x 1
##    DeporteOpcional
##    <chr>
##  1 <NA>
##  2 Baile
##  3 Baseball
```

```
##  4 Básquetbol
##  5 Ninguno
##  6 Natación
##  7 Volleybol
##  8 No
##  9 Futbol
## 10 Tocho
## 11 Basquetball
## 12 Basquetbol
## 13 basketball
## 14 Volleyball
## 15 cachibol
```

Contar las opciones

```
cuestionario %>% group_by(DeporteOpcional) %>% summarise(NumeroPosibilidades=n())
```

```
## # A tibble: 15 x 2
##    DeporteOpcional NumeroPosibilidades
##    <chr>                         <int>
##  1 Baile                             1
##  2 Baseball                          2
##  3 basketball                        1
##  4 Basquetball                       1
##  5 Basquetbol                        1
##  6 Básquetbol                        1
##  7 cachibol                          1
##  8 Futbol                            1
##  9 Natación                          1
## 10 Ninguno                           1
## 11 No                                1
## 12 Tocho                             1
## 13 Volleyball                        1
## 14 Volleybol                         1
## 15 <NA>                              6
```

Volleybol

```
cuestionario <-
  cuestionario %>%
    mutate(DeporteOpcional=
             ifelse(
               str_detect(DeporteOpcional, "Volleyb"), "Volleybol", DeporteOpcional))
```

```
cuestionario %>%
  group_by(DeporteOpcional) %>%
  summarise(NumeroPosibilidades=n())
```

```
## # A tibble: 14 x 2
##    DeporteOpcional NumeroPosibilidades
##    <chr>                         <int>
```

```
##  1 Baile                           1
##  2 Baseball                        2
##  3 basketball                      1
##  4 Basquetball                     1
##  5 Basquetbol                      1
##  6 Básquetbol                      1
##  7 cachibol                        1
##  8 Futbol                          1
##  9 Natación                        1
## 10 Ninguno                         1
## 11 No                              1
## 12 Tocho                           1
## 13 Volleybol                       2
## 14 <NA>                            6
```

Basquetball

```
cuestionario <-
  cuestionario %>%
  mutate(DeporteOpcional=
           ifelse(
             DeporteOpcional %in% c("basketball","Basquetball","Basquetbol","Básquetbol"),
             "Basquetbol",
             DeporteOpcional))

cuestionario %>%
  group_by(DeporteOpcional) %>%
  summarise(NumeroPosibilidades=n())
```

```
## # A tibble: 11 x 2
##    DeporteOpcional NumeroPosibilidades
##    <chr>                         <int>
##  1 Baile                             1
##  2 Baseball                          2
##  3 Basquetbol                        4
##  4 cachibol                          1
##  5 Futbol                            1
##  6 Natación                          1
##  7 Ninguno                           1
##  8 No                                1
##  9 Tocho                             1
## 10 Volleybol                         2
## 11 <NA>                              6
```

Ninguno

```
cuestionario <-
  cuestionario %>%
  mutate(DeporteOpcional=
           ifelse(
             DeporteOpcional %in% c("Ninguno","No"),
             "Ninguno",
```

```
            DeporteOpcional))

cuestionario %>%
  group_by(DeporteOpcional) %>%
  summarise(NumeroPosibilidades=n())
```

```
## # A tibble: 10 x 2
##    DeporteOpcional NumeroPosibilidades
##    <chr>                         <int>
##  1 Baile                             1
##  2 Baseball                          2
##  3 Basquetbol                        4
##  4 cachibol                          1
##  5 Futbol                            1
##  6 Natación                          1
##  7 Ninguno                           2
##  8 Tocho                             1
##  9 Volleybol                         2
## 10 <NA>                              6
```

Grado académico

```
cuestionario %>% group_by(GradoAcademico) %>% summarise(NumeroPosibilidades=n())
```

```
## # A tibble: 12 x 2
##    GradoAcademico               NumeroPosibilidades
##    <chr>                                      <int>
##  1 5 semestre preparatoria                        1
##  2 8vo semestre de licenciatura                   1
##  3 Bachillerato                                   2
##  4 Estudiante de Lic.                             1
##  5 Estudiante de licenciatura                     1
##  6 Licenciatura                                   4
##  7 preparatoria                                   1
##  8 Preparatoria                                   5
##  9 universidad                                    1
## 10 Universidad                                    1
## 11 Universidad, tercer semestre                   1
## 12 Universitario                                  2
```

```
cuestionario %>%
  group_by(GradoAcademico) %>%
  summarise(NumeroPosibilidades=n())
```

```
## # A tibble: 12 x 2
##    GradoAcademico               NumeroPosibilidades
##    <chr>                                      <int>
##  1 5 semestre preparatoria                        1
##  2 8vo semestre de licenciatura                   1
##  3 Bachillerato                                   2
##  4 Estudiante de Lic.                             1
##  5 Estudiante de licenciatura                     1
##  6 Licenciatura                                   4
```

```
##  7 preparatoria                                1
##  8 Preparatoria                                5
##  9 universidad                                 1
## 10 Universidad                                 1
## 11 Universidad, tercer semestre                1
## 12 Universitario                               2
```

```
cuestionario <-
  cuestionario %>%
   mutate(GradoAcademico=tolower(GradoAcademico))
```

Preparatoria

```
cuestionario <-
  cuestionario %>%
   mutate(GradoAcademico=
            ifelse(
              str_detect(GradoAcademico, "preparatoria"), "preparatoria", GradoAcademico)) %>%
  mutate(GradoAcademico=
            ifelse(
              str_detect(GradoAcademico, "bachillerato"), "preparatoria", GradoAcademico))

cuestionario %>%
  group_by(GradoAcademico) %>%
  summarise(NumeroPosibilidades=n())
```

```
## # A tibble: 8 x 2
##   GradoAcademico               NumeroPosibilidades
##   <chr>                              <int>
## 1 8vo semestre de licenciatura           1
## 2 estudiante de lic.                     1
## 3 estudiante de licenciatura             1
## 4 licenciatura                           4
## 5 preparatoria                           9
## 6 universidad                            2
## 7 universidad, tercer semestre           1
## 8 universitario                          2
```

Universidad

```
cuestionario <-
  cuestionario %>%
  mutate(GradoAcademico=
            ifelse(
              str_detect(GradoAcademico, "universi"), "universidad", GradoAcademico)) %>%
  mutate(GradoAcademico=
            ifelse(
              str_detect(GradoAcademico, "lic"), "universidad", GradoAcademico))

cuestionario %>%
  group_by(GradoAcademico) %>%
  summarise(NumeroPosibilidades=n())
```

```
## # A tibble: 2 x 2
##   GradoAcademico NumeroPosibilidades
##   <chr>                        <int>
## 1 preparatoria                     9
## 2 universidad                     12
```

Signo zodiacal

```
cuestionario <-
  cuestionario %>%
  mutate(SignoZodiacal=tolower(SignoZodiacal))

cuestionario %>%
  group_by(SignoZodiacal) %>%
  summarise(NumeroPosibilidades=n())
```

```
## # A tibble: 13 x 2
##    SignoZodiacal  NumeroPosibilidades
##    <chr>                       <int>
##  1 acuario                         2
##  2 aries                           3
##  3 cáncer                          1
##  4 capricornio                     3
##  5 escorpio                        2
##  6 escorpion                       1
##  7 geminis                         1
##  8 géminis                         1
##  9 leo                             2
## 10 lol                             1
## 11 no creo en eso                  1
## 12 piscis                          2
## 13 <NA>                            1
```

Cambiar Escorpion

```
cuestionario <-
  cuestionario %>%
  mutate(SignoZodiacal=
           ifelse(
             str_detect(SignoZodiacal, "escorpio"),
             "escorpión",
             SignoZodiacal))
```

Géminis

```
cuestionario <-
  cuestionario %>%
  mutate(SignoZodiacal=
           ifelse(
```

```
            SignoZodiacal %in% c("geminis","géminis"),
            "géminis",
            SignoZodiacal))

cuestionario %>%
  group_by(SignoZodiacal) %>%
  summarise(NumeroPosibilidades=n())
```

```
## # A tibble: 11 x 2
##    SignoZodiacal  NumeroPosibilidades
##    <chr>                       <int>
##  1 acuario                         2
##  2 aries                           3
##  3 cáncer                          1
##  4 capricornio                     3
##  5 escorpión                       3
##  6 géminis                         2
##  7 leo                             2
##  8 lol                             1
##  9 no creo en eso                  1
## 10 piscis                          2
## 11 <NA>                            1
```

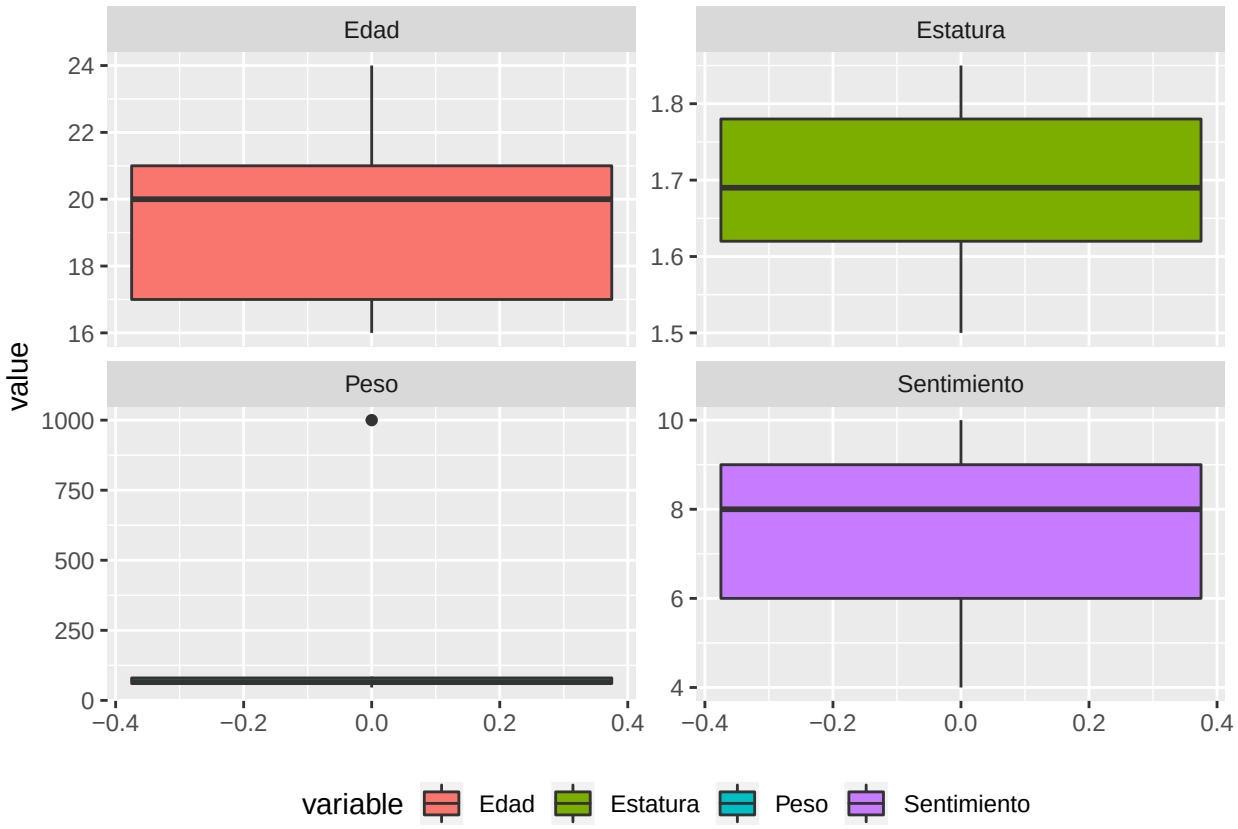
Eliminar los que no especifican alguno

```
cuestionario <-
  cuestionario %>%
  mutate(SignoZodiacal=
           ifelse(
             !SignoZodiacal %in% c("acuario","aries","cáncer","capricornio","escorpión","géminis","leo"
             NA,
             SignoZodiacal))
```

Analizar rápidamente los datos

```
cuestionarioColumnasNumericas<-
  select_if(cuestionario,is.numeric) %>%
  melt %>%
  filter(variable!="Timestamp")

ggplot(cuestionarioColumnasNumericas,aes(y=value,fill=variable)) +
  geom_boxplot() +
  facet_wrap(~variable,scales = "free_y") +
  theme(legend.position = 'bottom')
```

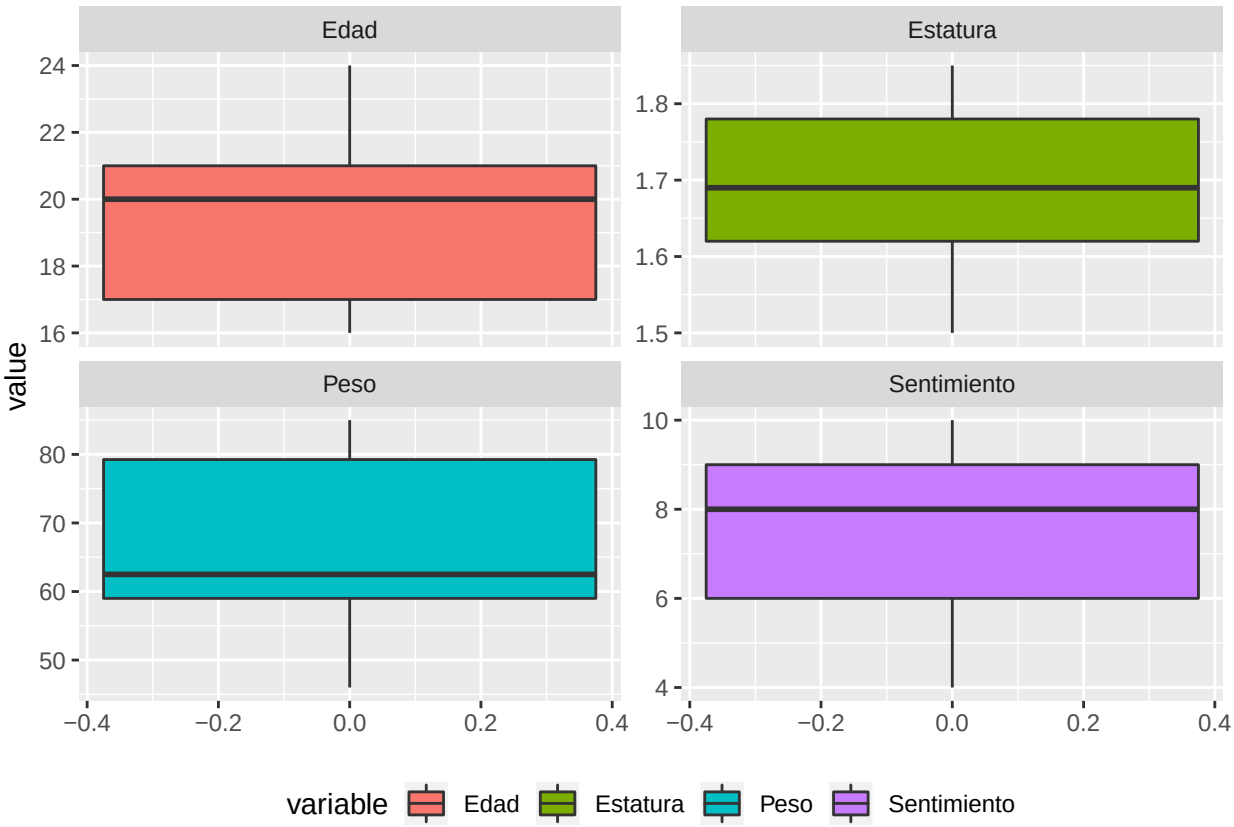


Se observa que hay un problema en peso (es dificil que alguien pese 1000 kg)

```
cuestionario<-
  cuestionario %>% mutate(Peso=ifelse(Peso>200,NA,Peso))

cuestionarioColumnasNumericas<-
  select_if(cuestionario,is.numeric) %>%
  melt %>%
  filter(variable!="Timestamp")

ggplot(cuestionarioColumnasNumericas,aes(y=value,fill=variable)) +
  geom_boxplot() +
  facet_wrap(~variable,scales = "free_y") +
  theme(legend.position = 'bottom')
```



Ahora están limpios los datos

Algunas cosas que pueden hacer después

```
cuestionario<- cuestionario %>% mutate(IMC=Peso/Estatura^2)
```