Limpieza de datos Apple 1

Luis Ángel Ramírez Franco A01363601

31/8/2020

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.1
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Paso 1: Leer datos

```
Apple1 <- read.csv("E:/The Last Dance/Lab Diseño y Opt/equipo_5_apple_1_datos_sucios.csv")
```

Paso 2: Hacer análisis general de los datos

```
dim(Apple1)
```

```
## [1] 19890    14
```

```
names(Apple1)
```

```
##  [1] "punto_de_venta" "fecha"          "mes"            "anio"
##  [5] "num_ventas"     "sku"            "marca"          "gamma"
##  [9] "costo_promedio" "zona"           "estado"         "ciudad"
## [13] "latitud"        "longitud"
```

```
summary(Apple1)
```

```
##  punto_de_venta        fecha               mes                anio
##  Length:19890       Length:19890       Length:19890       Min.   :  19
##  Class :character   Class :character   Class :character   1st Qu.:2018
##  Mode  :character   Mode  :character   Mode  :character   Median :2018
##                                                           Mean   :2016
```

```
##                                                        3rd Qu.:2018
##                                                        Max.   :2019
##    num_ventas     sku                marca              gamma
##  Min.   :1    Length:19890       Length:19890       Length:19890
##  1st Qu.:1    Class :character   Class :character   Class :character
##  Median :1    Mode  :character   Mode  :character   Mode  :character
##  Mean   :1
##  3rd Qu.:1
##  Max.   :1
##  costo_promedio     zona               estado             ciudad
##  Min.   :   0   Length:19890       Length:19890       Length:19890
##  1st Qu.:4977   Class :character   Class :character   Class :character
##  Median :8856   Mode  :character   Mode  :character   Mode  :character
##  Mean   :6920
##  3rd Qu.:8856
##  Max.   :9800
##     latitud             longitud
##  Min.   :     14.9   Min.   :-8684938
##  1st Qu.:     19.3   1st Qu.:    -101
##  Median :     19.7   Median :     -99
##  Mean   :    127.3   Mean   :    -537
##  3rd Qu.:     21.1   3rd Qu.:     -99
##  Max.   :2115143.0   Max.   :     -87
```

```
#head(Apple1,20)
```

```
#Apple1%>%select(punto_de_venta)%>%unique()
```

5 errores en los puntos de venta

```
Apple1$punto_de_venta <- str_replace(Apple1$punto_de_venta,"5 de mayo zm","5 de mayo zmm")
Apple1$punto_de_venta <- str_replace(Apple1$punto_de_venta,"5 de mayo zmmm","5 de mayo zmm")
```

```
Apple1$punto_de_venta <- str_replace(Apple1$punto_de_venta,"arsa cty shops dl valle","arsa city shops d
Apple1$punto_de_venta <- str_replace(Apple1$punto_de_venta,"ARSA periSUR","arsa perisur")
Apple1$punto_de_venta <- str_replace(Apple1$punto_de_venta,"acr ATLIXCOCENTROPUE","acr atlixcocentropue
Apple1$punto_de_venta <- str_replace(Apple1$punto_de_venta,"cotzacoalcos","coatzacoalcos")
```

```
#head(Apple1%>%select(punto_de_venta)%>%unique())
```

```
Apple1$punto_de_venta<- as.factor(Apple1$punto_de_venta)
str(Apple1)
```

```
## 'data.frame':    19890 obs. of  14 variables:
##  $ punto_de_venta: Factor w/ 1293 levels "5 de mayo zmm",..: 1 1 1 1 2 2 2 2 2 3 ...
##  $ fecha         : chr  "09/06/2018" "20/12/2018" "01/02/2019" "21/02/2019" ...
##  $ mes           : chr  "6" "12" "2" "2" ...
##  $ anio          : int  2018 2018 2019 2019 2018 2018 2018 2018 2018 2019 ...
##  $ num_ventas    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ sku           : chr  "N.ISE32GR" "N.IP732B" "N.IP732B" "N.ISE32GR" ...
##  $ marca         : chr  "apple" "apple" "apple" "apple" ...
```

```
##  $ gamma         : chr  "baja" "media" "media" "baja" ...
##  $ costo_promedio: num  4641 8856 8856 4641 4983 ...
##  $ zona          : chr  "centro occidente" "centro occidente" "centro occidente" "centro occidente"
##  $ estado        : chr  "michoacan" "michoacan" "michoacan" "michoacan" ...
##  $ ciudad        : chr  "zamora" "zamora" "zamora" "zamora" ...
##  $ latitud       : num  20 20 20 20 17.9 ...
##  $ longitud      : num  -102.3 -102.3 -102.3 -102.3 -94.9 ...
```

```
#Apple1%>%select(mes)%>%unique()
```

```
Apple1$mes <- str_replace(Apple1$mes,"ENERO","1")
Apple1$mes <- str_replace(Apple1$mes,"JUL","7")
Apple1$mes <- str_replace(Apple1$mes,"OCT","10")
Apple1$mes <- str_replace(Apple1$mes,"JUN","6")
Apple1$mes <- str_replace(Apple1$mes,"MAR","3")
```

Los meses han sido corregidos, observaciones: no hay mes ni 4 ni 5

```
#Apple1%>%select(mes)
```

```
Apple1%>%select(anio)%>%unique()
```

```
##    anio
## 1  2018
## 3  2019
## 20   19
```

```
Apple1$anio <- str_replace(Apple1$anio,"202019","2019")
Apple1$anio <- str_replace(Apple1$anio,"19","2019")
```

```
Apple1$anio <- str_replace(Apple1$anio,"202019","2019")
```

El único valor “19” fue remplazado a 2019

```
Apple1%>%select(marca)%>%unique()
```

```
##            marca
## 1          apple
## 2555       Apple
## 2679 Apple-apple
## 9292       APPLE
## 9307      aApple
```

```
#Apple1%>%select(marca)
```

verificamos que sean minúsculas la mayoría

```
Apple1$marca <- str_replace(Apple1$marca,"Apple","apple")
Apple1$marca <- str_replace(Apple1$marca,"Apple-apple","apple")
Apple1$marca <- str_replace(Apple1$marca,"APPLE","apple")
Apple1$marca <- str_replace(Apple1$marca,"aApple","apple")
Apple1$marca <- str_replace(Apple1$marca,"apple-apple","apple")
Apple1$marca <- str_replace(Apple1$marca,"aapple","apple")
```

Todos los errores en marca corregidos

```
#Apple1%>%select(zona)%>%unique()
```

```
Apple1$zona <- str_replace(Apple1$zona,"NRTE","norte")
```

Hemos corregido todas las zonas.

```
#Apple1%>%select(estado)%>%unique()
```

```
Apple1$estado <- str_replace(Apple1$estado,"cancun","quintana roo")
Apple1$estado <- str_replace(Apple1$estado,"tepic","nayarit")
Apple1$estado <- str_replace(Apple1$estado,"cancun","quintana roo")
```

Sólo encontramos 3 errores y podemos ver al final sólo 32 estados, como debe ser

```
#Apple1%>%select(latitud)%>%unique()
```

```
max(Apple1$latitud)
```

```
## [1] 2115143
```

```
which.max(Apple1$latitud)
```

```
## [1] 2553
```

```
Apple1$latitud<-as.numeric(Apple1$latitud)
Apple1[2553,13]<-"21.15143"
max(Apple1$latitud)
```

```
## [1] "32.66578"
```

```
which.max(Apple1$latitud)
```

```
## [1] 12454
```

```
min(Apple1$longitud)
```

```
## [1] -8684938
```

```
which.min(Apple1$longitud)
```

```
## [1] 2553
```

```
Apple1$longitud<-as.numeric(Apple1$longitud)
Apple1[2553,14]<-"-86.84938"
max(Apple1$longitud)
```

```
## [1] "-99.9945"
```

```r
which.max(Apple1$longitud)
```

```
## [1] 2970
```

```r
str(Apple1)
```

```
## 'data.frame':    19890 obs. of  14 variables:
##  $ punto_de_venta: Factor w/ 1293 levels "5 de mayo zmm",..: 1 1 1 1 2 2 2 2 2 3 ...
##  $ fecha         : chr  "09/06/2018" "20/12/2018" "01/02/2019" "21/02/2019" ...
##  $ mes           : chr  "6" "12" "2" "2" ...
##  $ anio          : chr  "2018" "2018" "2019" "2019" ...
##  $ num_ventas    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ sku           : chr  "N.ISE32GR" "N.IP732B" "N.IP732B" "N.ISE32GR" ...
##  $ marca         : chr  "apple" "apple" "apple" "apple" ...
##  $ gamma         : chr  "baja" "media" "media" "baja" ...
##  $ costo_promedio: num  4641 8856 8856 4641 4983 ...
##  $ zona          : chr  "centro occidente" "centro occidente" "centro occidente" "centro occidente" ...
##  $ estado        : chr  "michoacan" "michoacan" "michoacan" "michoacan" ...
##  $ ciudad        : chr  "zamora" "zamora" "zamora" "zamora" ...
##  $ latitud       : chr  "19.98131" "19.98131" "19.98131" "19.98131" ...
##  $ longitud      : chr  "-102.28329" "-102.28329" "-102.28329" "-102.28329" ...
```

```r
Apple1$anio<- as.factor(Apple1$anio)
Apple1$marca<- as.factor(Apple1$marca)
Apple1$zona<- as.factor(Apple1$zona)
Apple1$estado<- as.factor(Apple1$estado)
Apple1$latitud<- as.numeric(Apple1$latitud)
Apple1$longitud<- as.numeric(Apple1$longitud)
```

```r
summary(Apple1)
```

```
##                           punto_de_venta     fecha
##  tda cdmx forum buenavista        :  233   Length:19890
##  tda puebla huexotitla            :  150   Class :character
##  red celular barrio de coaxustenco:  132   Mode  :character
##  tda leon centro max bajio        :  122
##  tda cdmx tintoreto               :  121
##  tda monterrey humberto lobo      :  120
##  (Other)                          :19012
##      mes              anio         num_ventas     sku              marca
##  Length:19890       2018:15454   Min.   :1    Length:19890       apple:19890
##  Class :character   2019: 4436   1st Qu.:1    Class :character
##  Mode  :character                Median :1    Mode  :character
##                                  Mean   :1
##                                  3rd Qu.:1
##                                  Max.   :1
##
##     gamma           costo_promedio               zona
##  Length:19890       Min.   :   0   centro sur      :8851
##  Class :character   1st Qu.:4977   centro occidente:3707
##  Mode  :character   Median :8856   noreste         :1480
```

```
##                     Mean   :6920   norte           :1449
##                     3rd Qu.:8856   noroeste        :1432
##                     Max.   :9800   golfo de mexico :1189
##                                    (Other)         :1782
##                estado        ciudad             latitud         longitud
##  cdmx             :3586   Length:19890       Min.   :14.87   Min.   :-117.11
##  estado de mexico:2550   Class :character   1st Qu.:19.30   1st Qu.:-101.38
##  guanajuato       :1510   Mode  :character   Median :19.67   Median : -99.27
##  jalisco          :1411                      Mean   :20.97   Mean   :-100.07
##  nuevo leon       :1057                      3rd Qu.:21.15   3rd Qu.: -99.03
##  veracruz         : 938                      Max.   :32.67   Max.   : -86.81
##  (Other)          :8838
```

```
write.csv(Apple1, file="E:/The Last Dance/Lab Diseño y Opt/equipo_5_apple_1_datos_limpios.csv", row.nam
```

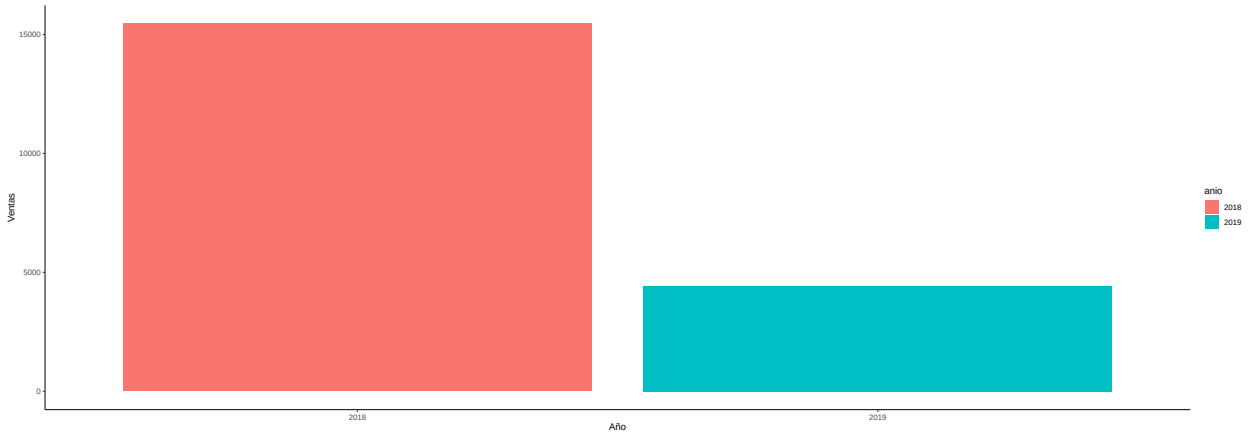
```
library(ggplot2)
```

```
library(babynames)
```

```
data()
```

¿Qué año tuvo más ventas? En las gráficas correspondientes a cada año podemos observar que el año 2018 tuvo la mayoría de ventas de Apple pasando un número de 15,000 en los diferentes puntos de venta .

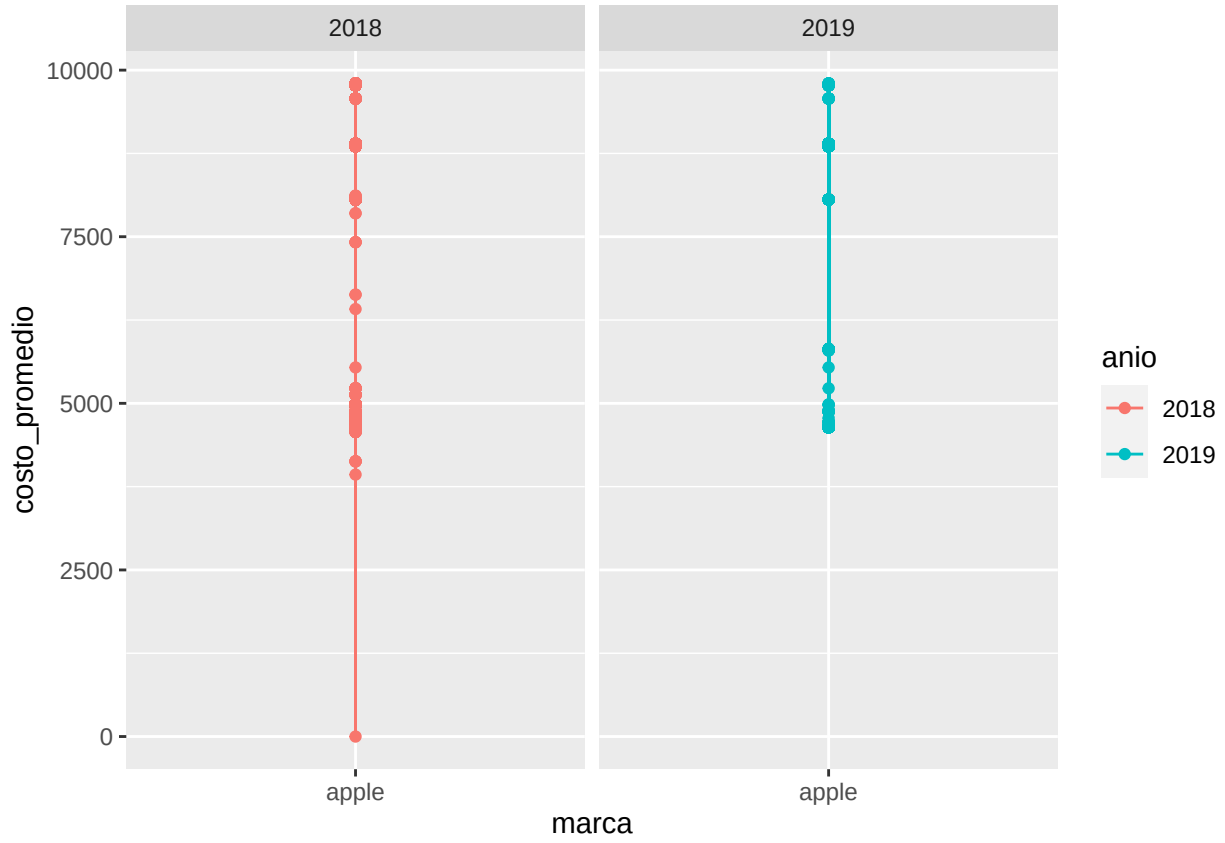
```
grafica1 <- ggplot(Apple1, aes(x=forcats::fct_infreq(anio), fill=anio))+
  geom_bar()+
  theme_classic()+
  xlab("Año")+
  ylab("Ventas")
grafica1
```




```
graficas <- babynames
```

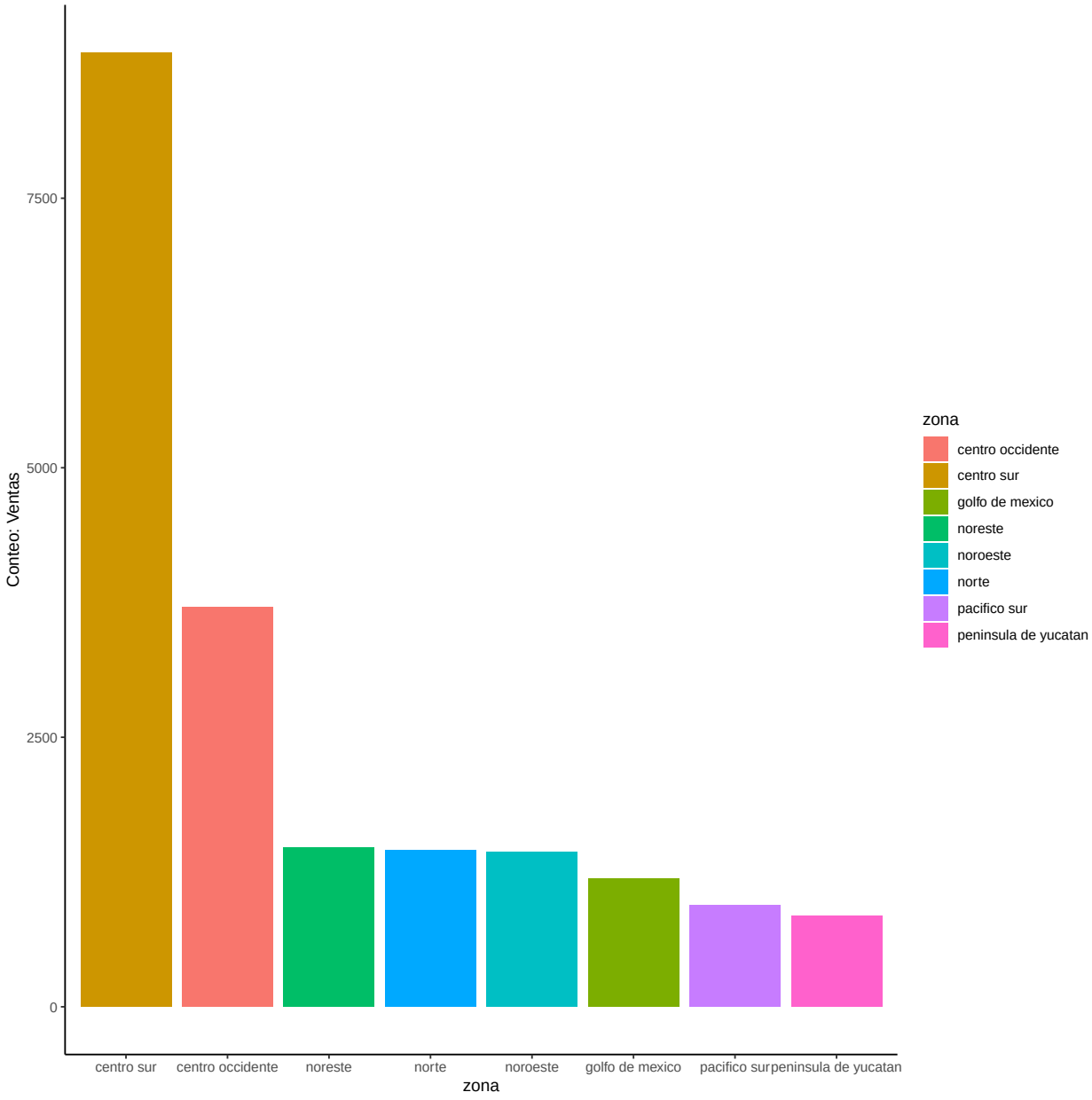
¿Cuál fue el costo promedio del celular marca Apple en los dos años registrados? En la gráfica se puede observar que hubo un costo promedio de 5,000 pesos.

```
grafica2 <- ggplot(Apple1, aes(x=marca, y=costo_promedio, color=anio))+
  geom_line()+
  geom_point()+
  facet_wrap(~anio)
grafica2
```



¿Cuál es la zona donde se vende más Apple? Como nos muestra la gráfica, la zona en donde se vende más la marca Apple es Centro Sur de la República Mexicana, lo que nos da una idea de que se podría concentrar el proyecto en esa zona o investigar por qué en la zona de la península se están teniendo menores ventas.

```
grafica3<-ggplot(Apple1, aes(x=forcats::fct_infreq(zona), fill=zona))+
  geom_bar()+
  theme_classic()+
  xlab("zona")+
  ylab("Conteo: Ventas")
grafica3
```




```
#install.packages("lubridate")
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
Apple1$fecha <- as.Date(Apple1$fecha, format ="%d/%m/%Y")
#Apple1
```

```
HIST <- Apple1 %>%
  group_by(month=floor_date(fecha, "month")) %>%
  summarize(amount=sum(num_ventas))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
HIST
```

```
## # A tibble: 10 x 2
##    month      amount
##    <date>      <int>
##  1 2018-06-01   3270
##  2 2018-07-01   3217
##  3 2018-08-01   3126
##  4 2018-09-01   1145
##  5 2018-10-01    698
##  6 2018-11-01   1535
##  7 2018-12-01   2463
##  8 2019-01-01   2010
##  9 2019-02-01   1689
## 10 2019-03-01    737
```

Historico de ventas de los años registrados Por lo que se puede observar en la gráfica, las ventas a lo largo de los 12 meses es constante, en ambos años

```
ggplot(HIST, aes(x=month, y=amount))+
  geom_line(color="dodgerblue4", size=3)+
  theme_classic()+
  geom_smooth(method="lm")+
  xlab("Mes")+
  ylab("Ventas")
```

```
## `geom_smooth()` using formula 'y ~ x'
```