Métodos Multivariados: Tarea 1

Aldo, Diego, Mateo, Victor

31/01/2024

library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(lattice)  
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────────────────────────── tidyverse 2.0.0 ──  
✔ forcats 1.0.0 ✔ readr 2.1.4  
✔ ggplot2 3.4.4 ✔ stringr 1.5.0  
✔ lubridate 1.9.2 ✔ tibble 3.2.1  
✔ purrr 1.0.1 ✔ tidyr 1.3.0

── Conflicts ────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library(ggplot2)  
library(ggExtra)  
library(plotly)

Attaching package: 'plotly'  
  
The following object is masked from 'package:ggplot2':  
  
 last\_plot  
  
The following object is masked from 'package:stats':  
  
 filter  
  
The following object is masked from 'package:graphics':  
  
 layout

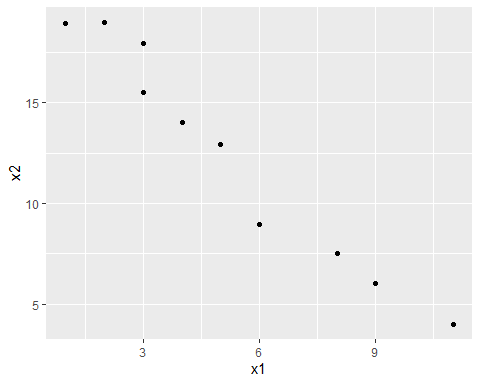
library(aplpack)  
library(pander)

# Ejercicio 1.2)

## a)

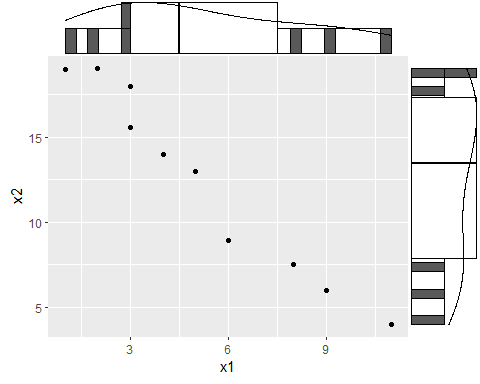
scatterplot A morning newspaper lists the following used-car prices for a foreign compact with age measurred in years and selling price measurred in thousands of dollars.

x1 <- c(1, 2, 3, 3, 4, 5, 6, 8, 9, 11)  
x2 <- c(18.95, 19, 17.95, 15.54, 14, 12.95, 8.94, 7.49, 6, 3.99)  
datos <- data.frame(x1, x2)  
  
(plot <- ggplot(datos, aes(x1, x2)) + geom\_point())



marginal dot diagram

plot1 <- ggMarginal(plot, type="histogram")  
plot2 <- ggMarginal(plot, type="boxplot")  
plot3 <- ggMarginal(plot, type="density")  
  
plot1  
plot2  
plot3



## b)

Infiero que la covarianza es negativa porque hay una tendencia: entre más años tiene el carro, en menos precio se vende.

## c)

m1 <- mean(x1)  
m2 <- mean(x2)  
  
s11 <- var(x1)  
s22 <- var(x2)  
  
s12 <- cov(x1, x2)  
r12 <- cor(x1, x2)  
  
cat("media x1: ", m1, "\n",   
 "media x2: ", m2, "\n",  
 "varianza x1: ", s11, "\n",  
 "varianza x2: ", s22, "\n",  
 "covarianza: ", s12, "\n",  
 "correlación: ", r12, "\n")

media x1: 5.2   
 media x2: 12.481   
 varianza x1: 10.62222   
 varianza x2: 30.85437   
 covarianza: -17.71022   
 correlación: -0.9782684

Interpretación:

Media x1: 5.2 años. Representa el valor central de los años de antigüedad de los autos.

Media x2: $12,481. Indica el valor central de los precios de los autos.

Varianza x1: 10.62222. Refleja la dispersión de los años de antigüedad de los autos alrededor de su valor central.

Varianza x2: 30.85437. Muestra la dispersión de los precios de los autos alrededor de su valor central.

Covarianza: -17.71022. Indica cómo varían conjuntamente los años de antigüedad y los precios de los autos. Una covarianza negativa sugiere que los autos más antiguos tienden a tener precios más bajos, y viceversa.

Correlación: -0.9782684. Representa la fuerza y la dirección de la relación entre los años de antigüedad y los precios de los autos. Una correlación negativa cercana a -1 indica una fuerte relación inversa: a medida que los años de antigüedad aumentan, los precios tienden a disminuir

## d)

colMeans(datos)

x1 x2   
 5.200 12.481

var(datos)

x1 x2  
x1 10.62222 -17.71022  
x2 -17.71022 30.85437

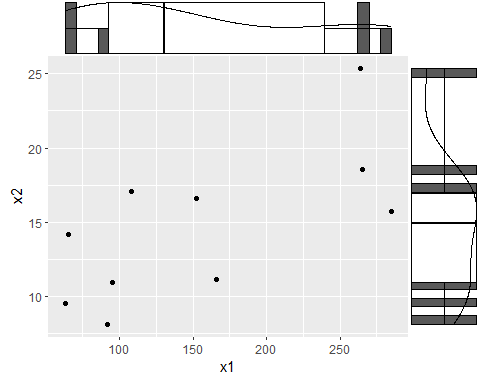
cor(datos)

x1 x2  
x1 1.0000000 -0.9782684  
x2 -0.9782684 1.0000000

# Ejercicio 1.4)

## a)

df <- data.frame(  
 x1 = c(108.28, 152.36, 95.04, 65.45, 62.97, 263.99, 265.19, 285.06, 92.01, 165.68),  
 x2 = c(17.05, 16.59, 10.91, 14.14, 9.52, 25.33, 18.54, 15.73, 8.10, 11.13),  
 x3 = c(1484.10, 750.33, 766.42, 1110.46, 1031.29, 195.26, 193.83, 191.11, 1175.16, 211.15))  
  
plot <- ggplot(df, aes(x1, x2)) + geom\_point()  
plot1 <- ggMarginal(plot, type="histogram")  
plot2 <- ggMarginal(plot, type="boxplot")  
plot3 <- ggMarginal(plot, type="density")  
  
plot1  
plot2  
plot3

 Las ventas tienen mucha variabilidad y no parece seguir una distribución normal porque se cargan mucho los datos a la izquierda (es difícil tener ventas grandes). La ganancia parece estar también muy cargada, en general no se dan ganancias grandes, pero sí aumenta conforme aumentan las ventas.

## b)

m1 <- mean(df$x1)  
m2 <- mean(df$x2)  
  
s11 <- var(df$x1)  
s22 <- var(df$x2)  
  
s12 <- cov(df$x1, df$x2)  
r12 <- cor(df$x1, df$x2)  
  
cat("media x1: ", m1, "\n",   
 "media x2: ", m2, "\n",  
 "varianza x1: ", s11, "\n",  
 "varianza x2: ", s22, "\n",  
 "covarianza: ", s12, "\n",  
 "correlación: ", r12, "\n")

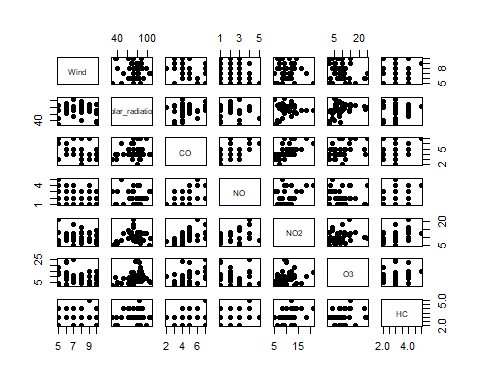
media x1: 155.603   
 media x2: 14.704   
 varianza x1: 7476.453   
 varianza x2: 26.19032   
 covarianza: 303.6186   
 correlación: 0.686136

La correlación no llega a ser fuerte, pero sí hay cierta tendencia entre ambas variables que nos permite decir que ante un aumento de ventas, hay un aumento en el profit.

# Ejercicio 1.6)

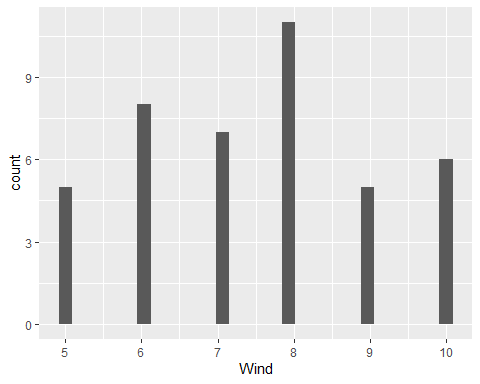
## a)

datos <- read.table("data/T1-5.DAT", header = FALSE)  
colnames(datos) <- c("Wind", "Solar\_radiation", "CO", "NO", "NO2", "O3", "HC")  
  
pairs(datos, pch = 19)



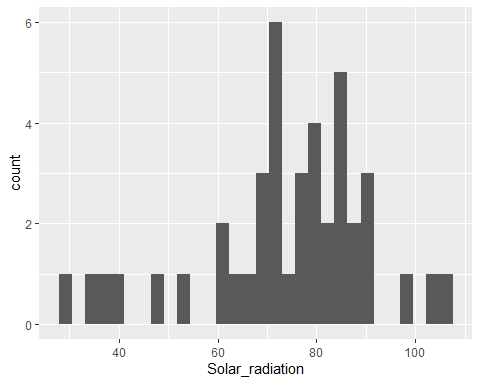
g1 <- ggplot(datos, aes(x = Wind)) +  
 geom\_histogram()  
  
g2 <- ggplot(datos, aes(x = Solar\_radiation)) +  
 geom\_histogram()  
  
g3 <- ggplot(datos, aes(x = CO)) +  
 geom\_histogram()  
  
g4 <- ggplot(datos, aes(x = NO)) +  
 geom\_histogram()  
  
g5 <- ggplot(datos, aes(x = NO2)) +  
 geom\_histogram()  
  
g6 <- ggplot(datos, aes(x = O3)) +  
 geom\_histogram()  
  
g7 <- ggplot(datos, aes(x = HC)) +  
 geom\_histogram()  
  
  
g1

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



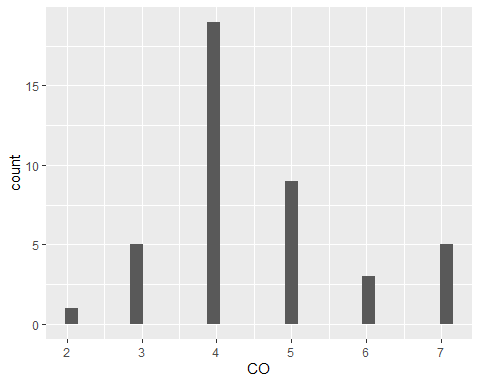
g2

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



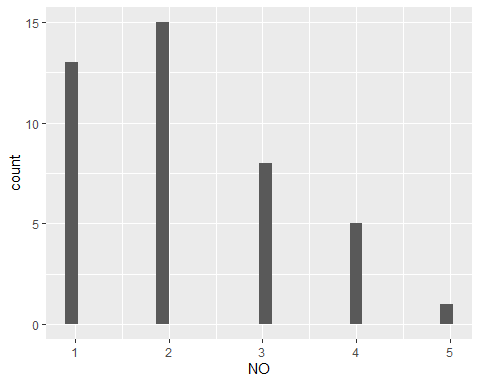
g3

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



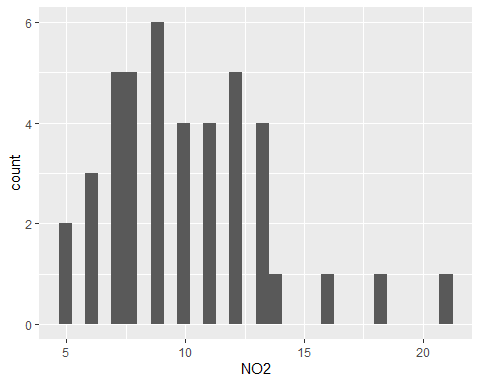
g4

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



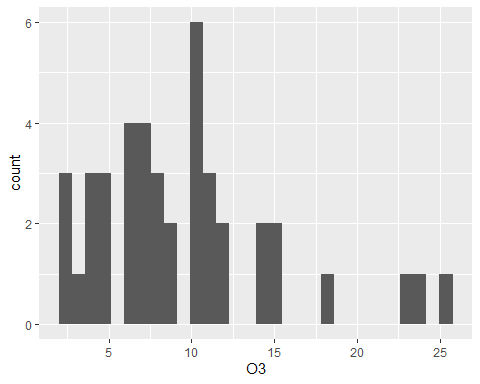
g5

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



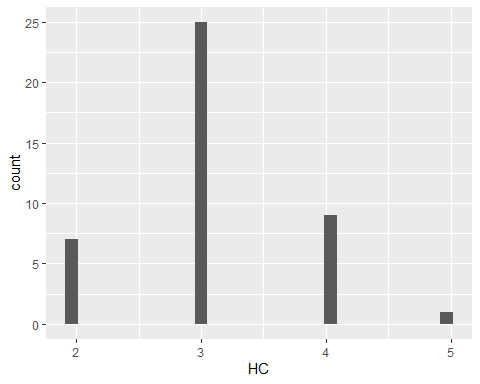
g6

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



g7

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## b)

x <- colMeans(datos)  
S <- var(datos)  
R <- cor(datos)  
  
x

Wind Solar\_radiation CO NO NO2 O3   
 7.500000 73.857143 4.547619 2.190476 10.047619 9.404762   
 HC   
 3.095238

S

Wind Solar\_radiation CO NO NO2 O3 HC  
Wind 2.5000000 -2.7804878 -0.3780488 -0.4634146 -0.5853659 -2.2317073 0.1707317  
Solar\_radiation -2.7804878 300.5156794 3.9094077 -1.3867596 6.7630662 30.7909408 0.6236934  
CO -0.3780488 3.9094077 1.5220674 0.6736353 2.3147503 2.8217189 0.1416957  
NO -0.4634146 -1.3867596 0.6736353 1.1823461 1.0882695 -0.8106852 0.1765389  
NO2 -0.5853659 6.7630662 2.3147503 1.0882695 11.3635308 3.1265970 1.0441347  
O3 -2.2317073 30.7909408 2.8217189 -0.8106852 3.1265970 30.9785134 0.5946574  
HC 0.1707317 0.6236934 0.1416957 0.1765389 1.0441347 0.5946574 0.4785134

R

Wind Solar\_radiation CO NO NO2 O3 HC  
Wind 1.0000000 -0.10144191 -0.1938032 -0.26954261 -0.1098249 -0.2535928 0.15609793  
Solar\_radiation -0.1014419 1.00000000 0.1827934 -0.07356907 0.1157320 0.3191237 0.05201044  
CO -0.1938032 0.18279338 1.0000000 0.50215246 0.5565838 0.4109288 0.16603235  
NO -0.2695426 -0.07356907 0.5021525 1.00000000 0.2968981 -0.1339521 0.23470432  
NO2 -0.1098249 0.11573199 0.5565838 0.29689814 1.0000000 0.1666422 0.44776780  
O3 -0.2535928 0.31912373 0.4109288 -0.13395214 0.1666422 1.0000000 0.15445056  
HC 0.1560979 0.05201044 0.1660323 0.23470432 0.4477678 0.1544506 1.00000000

Interpretación:

el vector de medias x nos dice el nivel que podemos esperar de cada medida tomada por el estudio en Los Ángeles, de tal forma que podemos esperar niveles de CO de 4.54 en un día cualquiera en Los Ángeles.

La matriz S nos permite ver un poco sobre la proporcionalidad entre pares de mediciones. Si el signo del elemento $ s\_{2,1} $ es negativo, como es el caso, podemos observar que las variables Solar\_radiation y Wind tienen tendencias a ser inversamente proporcionales. Al contrario si el signo es positivo, podemos argumentar que hay cierta proporcionalidad o tendencias a ello.

La matriz R nos permite ver la magnitud con la que covarían cierto par de variables, con esto podemos observar que la relación que tienen las variables entre ellas no es para nada fuerte, la correlación más fuerte es 0.55, que pertenece al par de variables (NO2, CO).

# Ejercicio 1.8

## a)

## b)

## c)

# Ejercicio 1.10

## a)

## b)

# Ejercicio 1.12

## a)

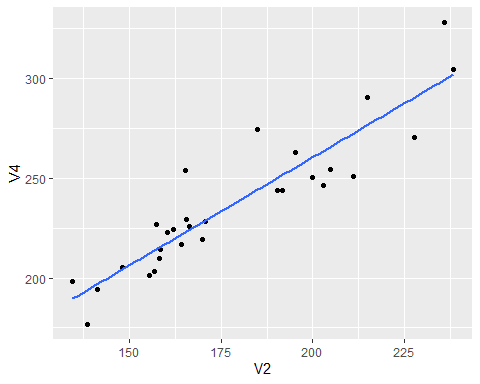
## b)

## c)

# Ejercicio 1.14

## a)

`geom\_smooth()` using formula = 'y ~ x'



Paceriera seguir un comportamiento lineal con correalación positiva, ya que al ajustar una regresión lineal a este conjunto de datos podemos ver de mejor manera esta posible relación.

## b)

Medias para ambas clases

| Group.1 | V1 | V2 | V3 | V4 | V5 |
| --- | --- | --- | --- | --- | --- |
| Non-Positive | 37.99 | 147.3 | 1.562 | 195.6 | 1.62 |
| Positive | 42.07 | 178.3 | 12.28 | 236.9 | 13.08 |

Para esclerosis esto resulta

|  | V1 | V2 | V3 | V4 | V5 |
| --- | --- | --- | --- | --- | --- |
| **V1** | 117 | 50.97 | -19.52 | 65.78 | -28.79 |
| **V2** | 50.97 | 815.6 | 236 | 881 | 103.1 |
| **V3** | -19.52 | 236 | 306.3 | 224.4 | 287.1 |
| **V4** | 65.78 | 881 | 224.4 | 1139 | 78.3 |
| **V5** | -28.79 | 103.1 | 287.1 | 78.3 | 338.9 |

Para no esclerosis esto resulta

|  | V1 | V2 | V3 | V4 | V5 |
| --- | --- | --- | --- | --- | --- |
| **V1** | 273.6 | 94.02 | 5.284 | 102.2 | 3.194 |
| **V2** | 94.02 | 110.7 | 1.741 | 105.2 | 2.013 |
| **V3** | 5.284 | 1.741 | 1.779 | 2.202 | 0.4941 |
| **V4** | 102.2 | 105.2 | 2.202 | 182.5 | 2.317 |
| **V5** | 3.194 | 2.013 | 0.4941 | 2.317 | 2.321 |

Para esclerosis esta es la matriz R

|  | V1 | V2 | V3 | V4 | V5 |
| --- | --- | --- | --- | --- | --- |
| **V1** | 1 | 0.165 | -0.1031 | 0.1802 | -0.1446 |
| **V2** | 0.165 | 1 | 0.4722 | 0.9139 | 0.1961 |
| **V3** | -0.1031 | 0.4722 | 1 | 0.3798 | 0.8909 |
| **V4** | 0.1802 | 0.9139 | 0.3798 | 1 | 0.126 |
| **V5** | -0.1446 | 0.1961 | 0.8909 | 0.126 | 1 |

Para no esclerosis esta es la matriz R

|  | V1 | V2 | V3 | V4 | V5 |
| --- | --- | --- | --- | --- | --- |
| **V1** | 1 | 0.5403 | 0.2395 | 0.4574 | 0.1268 |
| **V2** | 0.5403 | 1 | 0.1241 | 0.7404 | 0.1256 |
| **V3** | 0.2395 | 0.1241 | 1 | 0.1222 | 0.2431 |
| **V4** | 0.4574 | 0.7404 | 0.1222 | 1 | 0.1126 |
| **V5** | 0.1268 | 0.1256 | 0.2431 | 0.1126 | 1 |

# Ejercicio 1.16

Falta

# Ejercicio 1.18

Convertir los datos de la tabla a rapidez medida en m/s. Calcular ,, . Interpretar las correlaciones a pares.

datos <- read.table("data/T1-9.DAT")  
  
datos <- datos %>% mutate(speed\_100 = 100 / V1, speed\_200 = 200 / V2,  
speed\_400 = 400 / V3, speed\_800 = 800 / (V4 \* 60),  
speed\_1500 = 1500 / (V5 \* 60), speed\_3000 = 3000 / (V6 \* 60),  
speed\_marathon = 42195 / (V7 \* 60))  
X <- datos %>% select(speed\_100, speed\_200, speed\_400,  
speed\_800, speed\_1500, speed\_3000, speed\_marathon)

Vector de medias

x\_bar <- colMeans(X)  
x\_bar

speed\_100 speed\_200 speed\_400 speed\_800 speed\_1500 speed\_3000   
 8.619563 8.477682 7.508260 6.438315 5.809894 5.327651   
speed\_marathon   
 4.154344

Matriz de covarianza

n <- nrow(X)  
S <- var(X)  
Sn <- (n-1)/n \* S  
Sn

speed\_100 speed\_200 speed\_400 speed\_800 speed\_1500 speed\_3000 speed\_marathon  
speed\_100 0.10760676 0.12153268 0.1020016 0.07809972 0.09734084 0.1013165 0.1323820  
speed\_200 0.12153268 0.15053074 0.1241922 0.09227774 0.11161663 0.1152740 0.1553819  
speed\_400 0.10200163 0.12419217 0.1382760 0.10913176 0.11949475 0.1200003 0.1490451  
speed\_800 0.07809972 0.09227774 0.1091318 0.10656842 0.11982963 0.1177373 0.1441559  
speed\_1500 0.09734084 0.11161663 0.1194947 0.11982963 0.15951532 0.1588241 0.1928020  
speed\_3000 0.10131654 0.11527402 0.1200003 0.11773734 0.15882406 0.1702817 0.2059045  
speed\_marathon 0.13238201 0.15538191 0.1490451 0.14415595 0.19280198 0.2059045 0.3157511

Matriz de correlaciones

R <- cor(X)  
R

speed\_100 speed\_200 speed\_400 speed\_800 speed\_1500 speed\_3000 speed\_marathon  
speed\_100 1.0000000 0.9549062 0.8362072 0.7293155 0.7429749 0.7484738 0.7181856  
speed\_200 0.9549062 1.0000000 0.8608117 0.7285680 0.7203028 0.7200039 0.7127142  
speed\_400 0.8362072 0.8608117 1.0000000 0.8990078 0.8045891 0.7820324 0.7132994  
speed\_800 0.7293155 0.7285680 0.8990078 1.0000000 0.9190704 0.8740091 0.7858612  
speed\_1500 0.7429749 0.7203028 0.8045891 0.9190704 1.0000000 0.9636761 0.8590883  
speed\_3000 0.7484738 0.7200039 0.7820324 0.8740091 0.9636761 1.0000000 0.8879928  
speed\_marathon 0.7181856 0.7127142 0.7132994 0.7858612 0.8590883 0.8879928 1.0000000

Se puede observar que todas las correlaciones entre los valores de rapidez son positivas. Se puede observar que después del valor 1 de la diagonal principal, las correlaciones tienden a disminuir y antes del valor tienden a aumentar. Cuando la disntancia aumenta el tiempo en completarlo también aumenta, pero naturalmente, la rapidez promedio para completar un maratón es menor que la de un circuito de 100 m.

# Ejercicio 1.20

No trace type specified:  
 Based on info supplied, a 'scatter3d' trace seems appropriate.  
 Read more about this trace type -> https://plotly.com/r/reference/#scatter3d

No scatter3d mode specifed:  
 Setting the mode to markers  
 Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode

PhantomJS not found. You can install it with webshot::install\_phantomjs(). If it is installed, please make sure the phantomjs executable can be found via the PATH variable.

NULL

Se puede observar que una gran cantidad de puntos se concentran en un cúmulo alrededor del (-0.2, 0.4) en el eje x, (-0.2, 0.2) en el eje y, (1, 3.27) en el eje z. El punto con las coordenadas (0.58, 0.04, 5.06) parece ser un dato atípico. b) Colorear los puntos de acuerdo a los que están en bancarrota ¿Hay alguna orientación en la que se pueden distinguir las compañías en bancarrota de las que no lo están? ¿Existen observaciones que pueden llegar a tener un impacto significativo en alguna regla para clasicar nuevas empresas?

datos$V5 <- as.factor(datos$V5)  
  
#fig <- plot\_ly(datos, x = ~V1, y = ~V2, z = ~V3, color = ~V5,  
#colors = c("#FF0000", "#0000FF"))  
#fig

Si orientamos el eje X con x2, el eje y con x3 y el eje z con x1 se puede disntiguir a una gran mayoría de las compañías en bancarrota, específicamente, las que están en bancarrota tienden a tener menor x2, menor x3 y menor x1. En las empresas que no están en bancarrota un punto que puede tener un gran impacto puede ser el (0.14, -0.03, 0.46) debido a que presenta un valor de x3 muy bajo. En el caso de las que sí están en bancarrota podría ser el (0.37, 0.11, 1.99) debido a su valor mayor de x1.

# Ejercicio 1.22

datos <- read.table("data/T6-12.DAT") %>% select(V1, V2, V3, V4)

x\_bar <- colMeans(datos)  
x\_bar

V1 V2 V3 V4   
 0.3554 5.2542 3.0014 43.7876

## a) Graficar el dataset en 3 dimensiones.

#fig <- plot\_ly(datos, x = ~V1, y = ~V2, z = ~V3)  
#fig

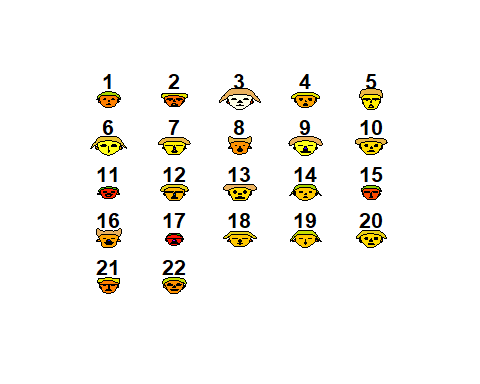
## b) Checar outliers.

Un valor que parece ser un outlier es el (-0.45, -0.41, 1.09) ya que sus valores en las tres coordendas son menores que la media.

# Ejercicio 1.24

Representar el dataset con caras de Chernoff ¿Existen diferentes grupos?

datos <- read.table("data/T12-4.DAT") %>% select(V1, V2, V3, V4, V5, V6, V7, V8)  
  
faces(datos, face.type = 1)



effect of variables:  
 modified item Var   
 "height of face " "V1"  
 "width of face " "V2"  
 "structure of face" "V3"  
 "height of mouth " "V4"  
 "width of mouth " "V5"  
 "smiling " "V6"  
 "height of eyes " "V7"  
 "width of eyes " "V8"  
 "height of hair " "V1"  
 "width of hair " "V2"  
 "style of hair " "V3"  
 "height of nose " "V4"  
 "width of nose " "V5"  
 "width of ear " "V6"  
 "height of ear " "V7"

Existen algunas compañías que sus representaciones se parecen, por ejemplo, la 7, 9, 12 y la 11, 15, 17.

# Ejercicio 1.26