Curso : MACHINE LEARNING CON R

Tema : Visualización de datos.

Docente : Victor Henostroza

**GUIA DE LABORATORIO 2**

**Objetivos del laboratorio:**

1. **Desarrollar un Scatter Plot.**
2. **Desarrollar un Gráfico de Barras.**
3. **Desarrollar un Gráfico de Distribuciones.**
4. **Desarrollar un Gráfico de Mozaico.**
5. **Desarrollar un Gráfico de Matriz de Correlaciones.**
6. **Desarrollar un Box Plot.**
7. **Desarrollar Gráficos Multivariantes.**
8. **Desarrollar Gráficos 3D.**

**Objetivo 1:**

Un Scatter Plot sirve para ver tendencias entre las variables independientes.

**#Scatter Plots**

install.packages("ggplot2")

library(ggplot2)

auto <- read.csv("auto-mpg.csv", stringsAsFactors = F)

auto$cylinders <- factor(auto$cylinders,

labels = c("3C", "4C", "5C", "6C", "8C"))

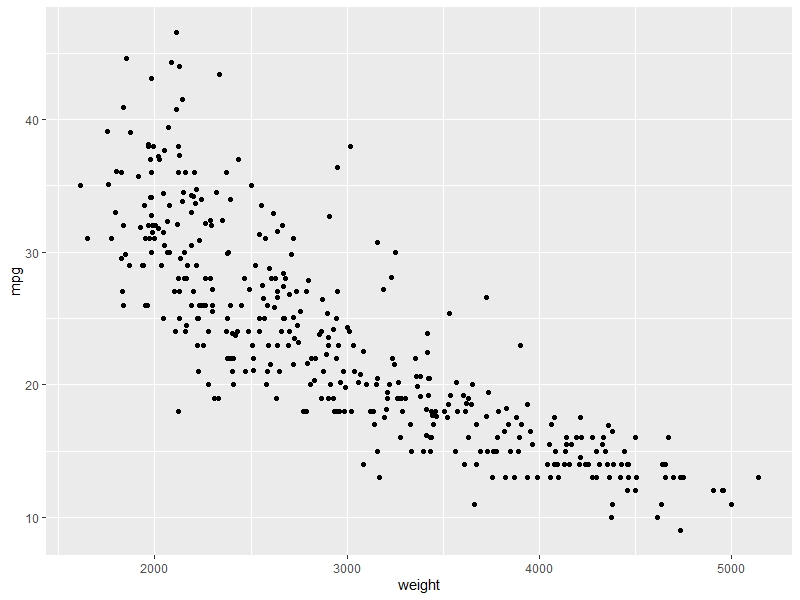
head(auto)

**# Ploteo Básico de un Scater Plot**

# plot = ggplot(dataset, aes(var\_x, var\_y))

plot <- ggplot(auto, aes(weight, mpg))

plot + geom\_point()



**# Ploteo con más detalles de un Scater Plot**

plot + geom\_point(alpha = 1/2, size = 5, aes(color = factor(cylinders))) +

geom\_smooth(method = "lm", se = TRUE, col ="green") +

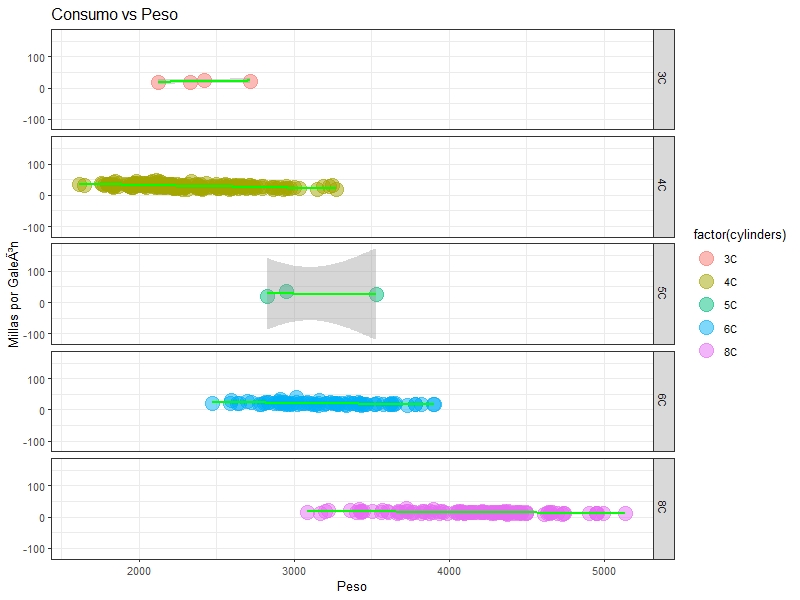
facet\_grid(cylinders~.) +

theme\_bw(base\_family = "Arial", base\_size = 10)+

labs(x = "Peso") + labs(y = "Millas por GaleÃ³n")+

labs(title = "Consumo vs Peso")





**Objetivo 2:**

**#Bar Graphs**

library(ggplot2)

bike <- read.csv("daily-bike-rentals.csv")

bike$season <- factor(bike$season,

levels = c(1,2,3,4),

labels = c("Invierno", "Primavera", "Verano", "Otoño"))

bike$workingday <- factor(bike$workingday,

levels = c(0,1),

labels = c("Día libre", "Día de trabajo"))

bike$weathersit <- factor(bike$weathersit,

levels = c(1,2,3),

labels = c("Buen tiempo", "Día nublado", "Mal tiempo"))

**#Generar tabla para ploteo**

library(dplyr)

bike.sum = bike %>%

group\_by(season, workingday) %>%

summarize(rental = sum(cnt))

bike.sum

**#Diseñar gráfico de barras**

ggplot(bike.sum, aes(x = season, y = rental,

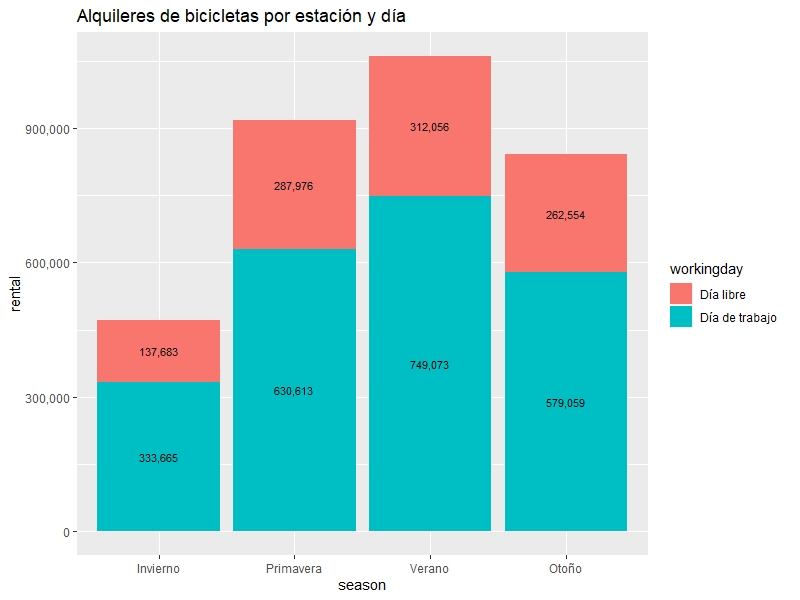
fill = workingday, label = scales::comma(rental))) +

geom\_bar(show.legend = T, stat = "identity") +

labs(title = "Alquileres de bicicletas por estación y día") +

scale\_y\_continuous(labels = scales::comma) +

geom\_text(size = 3, position = position\_stack(vjust = 0.5))



**Objetivo 3:**

Los gráficos de distribuciones sirven para ver cómo se distribuye una variable en particular respecto a los valores que toma en todo su dominio.

**#Gráficos de Distribución**

library(ggplot2)

geiser <- read.csv("geiser.csv")

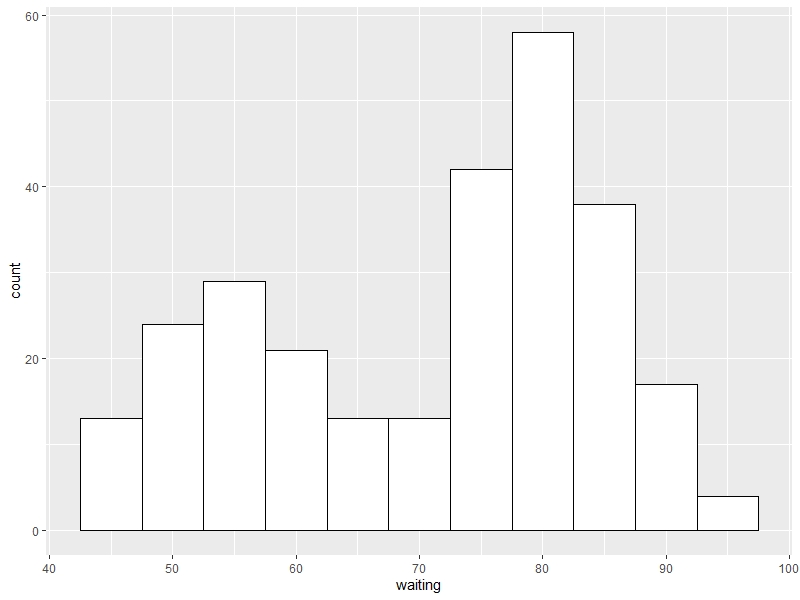
head(geiser)

**#Diseñar Histograma Básico**

plot <- ggplot(geiser, aes(x=waiting))

plot + geom\_histogram(binwidth = 5,

fill="white", colour = "black")

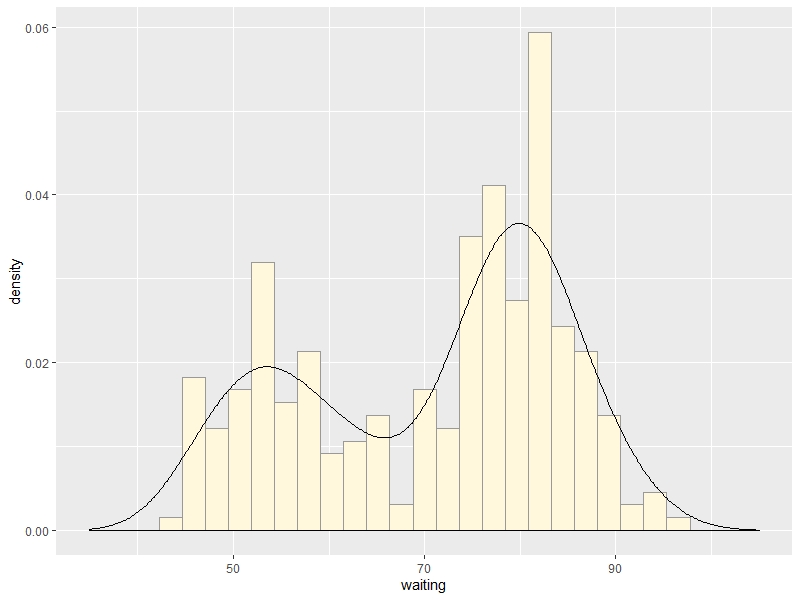


**#Diseñar Histograma con mayores especificaciones**

ggplot(geiser, aes(x=waiting, y = ..density..))+

geom\_histogram(fill="cornsilk", color = "grey60", size=.2)+

geom\_density() + xlim(35, 105)



**Objetivo 4:**

El gráfico de mosaic sirve para ver cantidades de dos variables de tipo factor en vs con el que se pueden ver según archuras del ploteo qué variable es más representativa en cada grupo cruzado.

**# Gráfico de Mosaico**

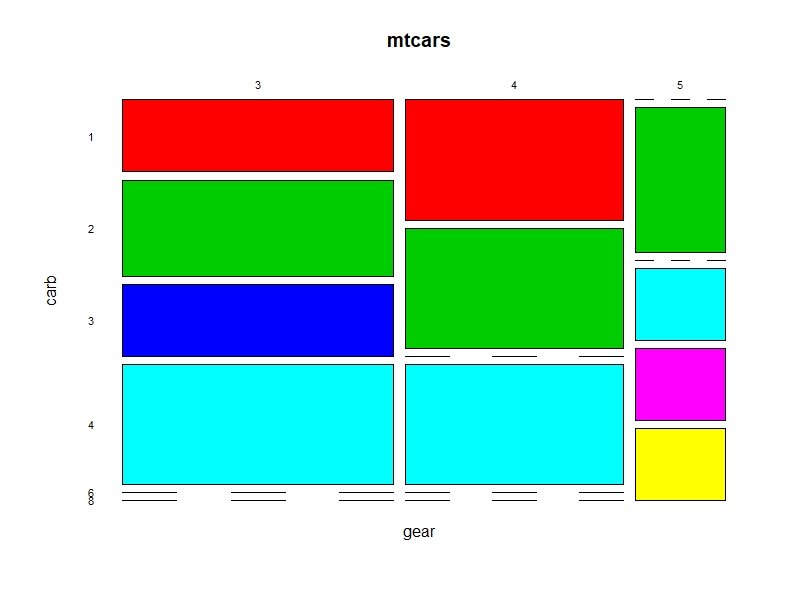
library(stats)

mtcars <- read.csv("../data/tema7/mtcars.csv")

head(mtcars)

mosaicplot(~ gear + carb, data = mtcars,

color=2:7, las = 1)



**Objetivo 5:**

Un gráfico de matriz de correlaciones sirve para ver qué tanta relación existe entre las variables, siendo más relacionadas mientras estén con una correlación cercana a 1 o -1, siendo 1 una correlación positiva y -1 una negativa, en cambio, si se acerca dicha cantidad a 0, existirá baja correlación entre las variables.

**# Matriz de Correlaciones**

library(ggplot2)

library(corrplot)

mtcars <- read.csv("mtcars.csv")

head(mtcars)

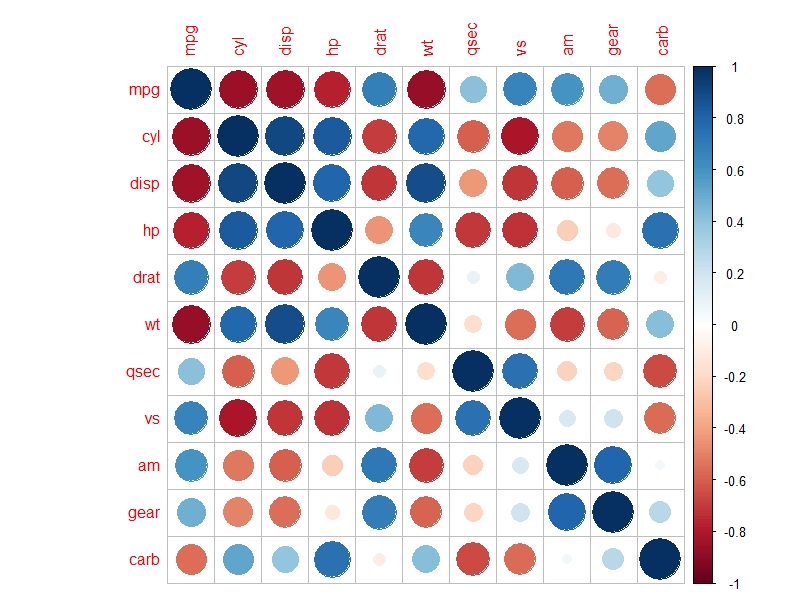
mtcars$X = NULL

mtcars.cor <- cor(mtcars, method = "pearson")

**# Pintar Matriz de Correlación Básica**

round(mtcars.cor, digits = 2)

corrplot(mtcars.cor)



**# Pintar Matriz de Correlación con más parámetros**

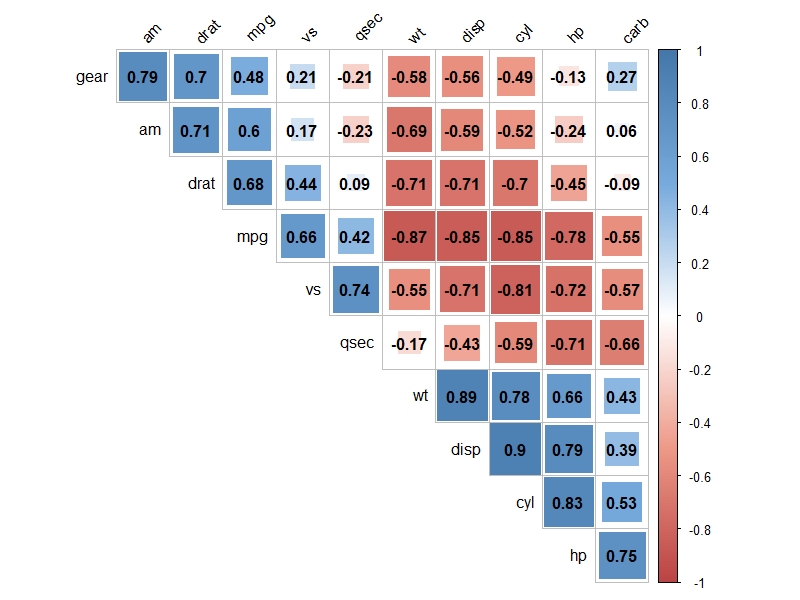
# Se eligen colores para la matriz

col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))

corrplot(mtcars.cor, method = "square", tl.col = "black", tl.srt = 45,

col = col(200), addCoef.col = "black", order = "AOE", type = "upper",

diag = F, addshade = "all")



**Objetivo 6:**

Un box-plot (diagrama de caja y bigotes), es un método estandarizado para representar gráficamente una serie de datos numéricos a través de sus cuartiles. De esta manera, el diagrama de caja muestra a simple vista la mediana y los cuartiles de los datos,1​ pudiendo también representar los valores atípicos de estos.

**#Diseñar Boxplot**

tooth <- read.csv("ToothGrowth.csv")

head(tooth)

> head(tooth)

X len supp dose

1 1 4.2 VC 0.5

2 2 11.5 VC 0.5

3 3 7.3 VC 0.5

4 4 5.8 VC 0.5

5 5 6.4 VC 0.5

6 6 10.0 VC 0.5

ggplot(tooth, aes(x=dose, y=len, fill = as.factor(dose)))+

geom\_boxplot() +

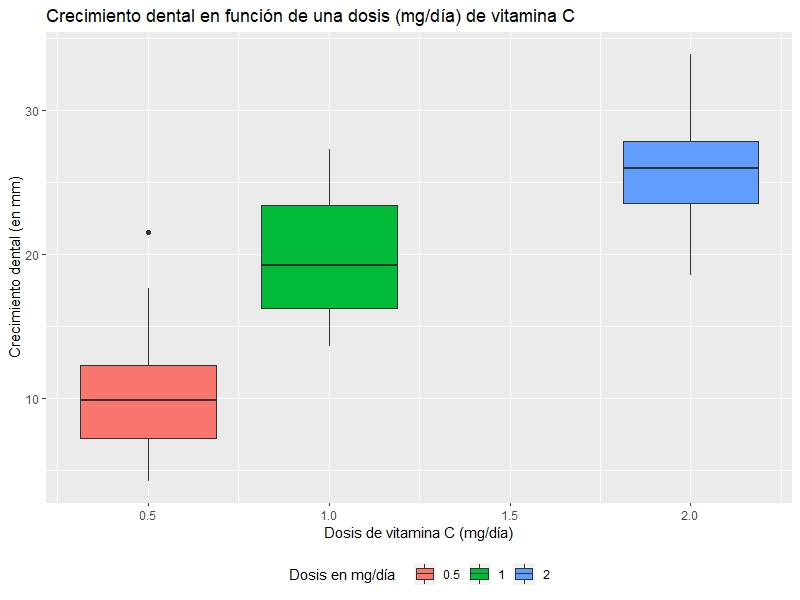
ggtitle("Crecimiento dental en función de una dosis (mg/día) de vitamina C")+

xlab("Dosis de vitamina C (mg/día)") +

ylab("Crecimiento dental (en mm)") +

labs(fill = "Dosis en mg/día") +

theme(legend.position = "bottom")



**Objetivo 7:**

Realizaremos un gráfico de varias variables a la vez, este tipo de gráfico es muy útil para ver relaciones entre variables a groso modo al momento del análisis exploratorio, ayudando así a prever que correlaciones o que variables pueden ser explicativas.

**#Plot Multivariante 1**

library(ggplot2)

library(GGally)

**# Importar Dataset**

bike <- read.csv("daily-bike-rentals.csv")

head(bike)

str(bike)

> head(bike)

instant dteday season yr mnth holiday weekday workingday weathersit temp atemp

1 1 2011-01-01 1 0 1 0 6 0 2 0.344167 0.363625

2 2 2011-01-02 1 0 1 0 0 0 2 0.363478 0.353739

3 3 2011-01-03 1 0 1 0 1 1 1 0.196364 0.189405

4 4 2011-01-04 1 0 1 0 2 1 1 0.200000 0.212122

5 5 2011-01-05 1 0 1 0 3 1 1 0.226957 0.229270

6 6 2011-01-06 1 0 1 0 4 1 1 0.204348 0.233209

hum windspeed casual registered cnt

1 0.805833 0.1604460 331 654 985

2 0.696087 0.2485390 131 670 801

3 0.437273 0.2483090 120 1229 1349

4 0.590435 0.1602960 108 1454 1562

5 0.436957 0.1869000 82 1518 1600

6 0.518261 0.0895652 88 1518 1606

**#Convertir a factor las variables correspondientes**

bike$season = factor(bike$season,

levels = c(1,2,3,4),

labels = c("Invierno", "Primavera", "Verano", "Otoño"))

bike$weathersit <- factor(bike$weathersit,

levels = c(1,2,3),

labels = c("Despejado", "Nublado", "Lluvia"))

bike$weekday <- factor(bike$weekday,

levels = 0:6,

labels = c("D", "L", "M", "X", "J", "V", "S"))

bike$windspeed.fac <- cut(bike$windspeed, breaks = 3,

labels = c("Poco", "Medio", "Elevado"))

head(bike)

> head(bike)

instant dteday season yr mnth holiday weekday workingday weathersit temp atemp

1 1 2011-01-01 Invierno 0 1 0 S 0 Nublado 0.344167 0.363625

2 2 2011-01-02 Invierno 0 1 0 D 0 Nublado 0.363478 0.353739

3 3 2011-01-03 Invierno 0 1 0 L 1 Despejado 0.196364 0.189405

4 4 2011-01-04 Invierno 0 1 0 M 1 Despejado 0.200000 0.212122

5 5 2011-01-05 Invierno 0 1 0 X 1 Despejado 0.226957 0.229270

6 6 2011-01-06 Invierno 0 1 0 J 1 Despejado 0.204348 0.233209

hum windspeed casual registered cnt windspeed.fac

1 0.805833 0.1604460 331 654 985 Poco

2 0.696087 0.2485390 131 670 801 Medio

3 0.437273 0.2483090 120 1229 1349 Medio

4 0.590435 0.1602960 108 1454 1562 Poco

5 0.436957 0.1869000 82 1518 1600 Medio

6 0.518261 0.0895652 88 1518 1606 Poco

**# Dibujar los puntos de "windspeed" en los ejes "temp" vs "cnt"**

**# tomando como factores a "weekday" y "season"**

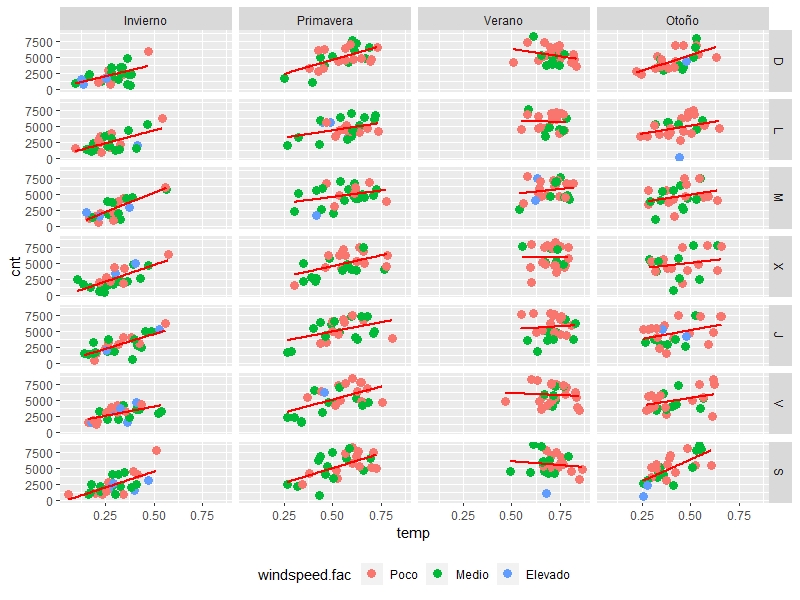
ggplot(bike, aes(x=temp, y = cnt))+

geom\_point(size=3, aes(color=windspeed.fac))+

theme(legend.position = "bottom")+

geom\_smooth(method="lm", se=F, col="red")+

facet\_grid(weekday ~ season)



**# Plot Multivariante 2**

# Importar Dataset

auto <- read.csv("auto-mpg.csv", stringsAsFactors = F)

auto$cylinders <- factor(auto$cylinders,

labels = c("3C", "4C", "5C", "6C", "8C"))

ggpairs(auto[,2:7],

aes(colour = cylinders,

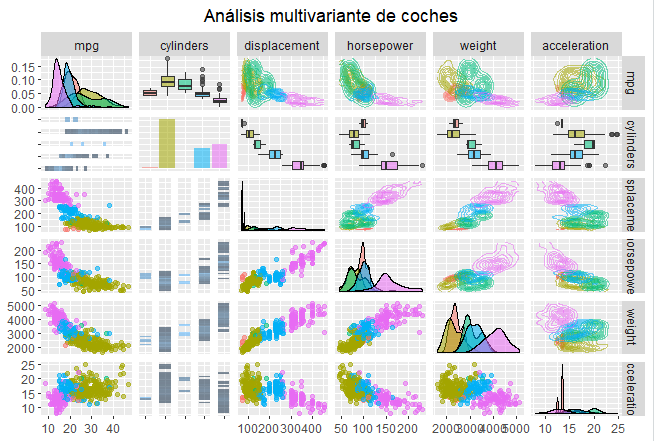
alpha = 0.4),

title = "Análisis multivariante de coches",

upper = list(continuous = "density"),

lower = list(combo = "denstrip"))+

theme(plot.title = element\_text(hjust = 0.5))



**Objetivo 8:**

Realizaremos gráficos en 3 dimensiones, muy útiles cuando se desea visualizar scatter plots con dos variables cuantitativas (ejes X e Y) y una tercera (eje Z)

**#Gráficos en 3D**

library(plot3D)

mtcars <- read.csv("mtcars.csv")

**# Cambiar nombre a las filas con las de la columna X**

rownames(mtcars) = mtcars$X

mtcars$X = NULL

head(mtcars)

> head(mtcars)

mpg cyl disp hp drat wt qsec vs am gear carb

Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4

Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4

Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1

Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1

Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2

Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

**# Generar un Scatter Plot 3D**

scatter3D(x=mtcars$disp,

y=mtcars$wt,

z=mtcars$mpg,

clab = c("Millas/Galeón"),

pch=19,

cex = 0.5,

theta = 18, #azimutal

phi=20, # colatitud

main="Coches de los 70'",

xlab = "Desplazamiento (cu.in.)",

ylab = "Peso (x1000lb)",

zlab = "Millas por galeón",

bty = "g")

text3D(x=mtcars$disp,

y=mtcars$wt,

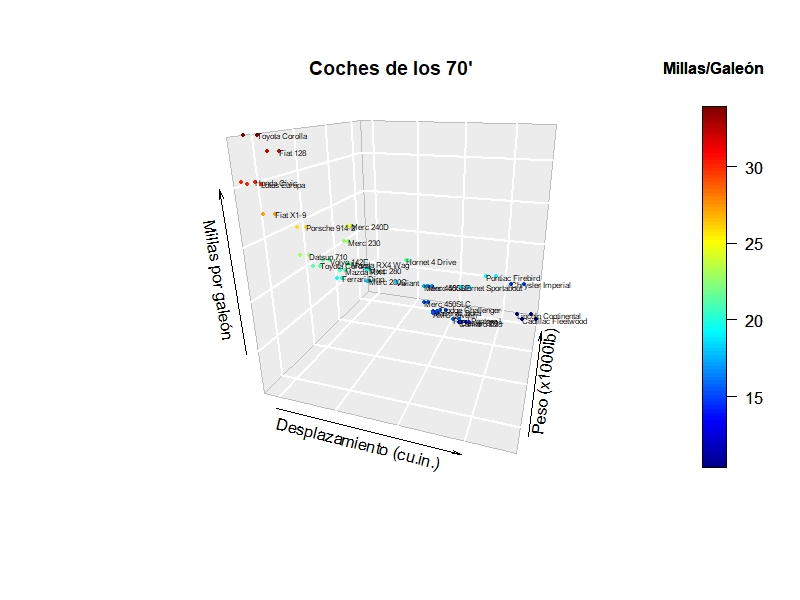
z=mtcars$mpg,

labels = rownames(mtcars),

add = T,

colkey = F,

cex = 0.5)



**# Desarrollar un Histograma 3D**

data(VADeaths)

head(VADeaths)

hist3D(z = VADeaths,

scale = F, expand = 0.01,

bty = "g", phi = 30,

col = "#1188CC", border = "black",

shade = 0.2, ltheta = 80,

space = 0.3,

ticktype = "detailed")

