setwd("~/CLASES DE MACHINE LEARNING")

library(readxl)

data <- read\_excel("datasetalquilereschamberi.xlsx")

View(data)

# Inspeccionamos el dataset para ver su extructura

str(data) # ver si se cambia la opci�n de exteror como factor

summary(data)

row.names(data)

names(data)

# vemos si hay alg�n missing value y seleccionamos los registros que contengan NA

sapply(data, anyNA) # vemos que hay NAs en la variable Superficie

missing<-data[is.na(data$Superficie)==TRUE,]

# Todas las casas con NA en la superficie tienen un precio de 1000

# �Qu� tal si construimos un data frame imputando un valor a los NA?

# Pongamos que imputamos la superficie media de todos las viviendas con el mismo precio.

# Calculamos la media de la superficie de aquellas viviendas con un precio igual a 1000 euros

mileuros<-data[data$Precio==1000,]

meanSup<-mean(mileuros$Superficie, na.rm=TRUE) # precauci�n: poner na.rm=TRUE para que prescinda de los NA en el c�lculo

data$Superficie[is.na(data$Superficie)]<-meanSup

subset(data, Precio==1000) # para comprobar que la imputaci�n ha sido correcta, ya no hay ninguna vivienda de 1000 eruos con Na en la superficie

# Guardamos en un fichero .csv el nuevo data frame (habiendo imputado los nuevos valores a los NAS)

write.csv(data, file="viviendasChamberi\_limpio.csv")

# �Cu�les son las caracter�sticas de la casa m�s cercana/lejana a la Puerta del Sol?

data[data$Distanciasol==max(data$Distanciasol),]

data[data$Distanciasol==min(data$Distanciasol),]

# Vamos a ver c�mo se distribuyen los precios de estas casas dibujando un histograma de la variable Precio

hist(data$Precio, main="Distribuci�n de frecuencias de los precios", xlab="Precio (euros)", ylab="N� de viviendas")

# Ahora vamos a hacer una comparativa de precios seg�n la variable Altura utilizando gr�ficos de caja y bigotes (boxplot)

boxplot(Precio ~ Altura, data, xlab="Altura de piso", ylab="Precio (euros)", main="Comparativa de precios seg�n altura")

# me parece curioso lo de la altura 7 y lo de la altura 0. Vamos a ver qu� pasa

subset(data, Altura==7) # vale, s�lo hay uno

subset(data, Altura==0) # vale, hay muchos pisos de precio igual a 1000 y altura 0

# No tienen bigotes ni el -1 ni el 5

# Vamos a crear una variable nueva que diga "SI" "NO" en funci�n de si la consideramos candidata a ser alquilada o no

# Ser�n candidatas aquellas viviendas que cuesten menor o igual a 1200 euros y tengan una superficie mayor de 62 m2

data$Candidata<-ifelse(data$Precio<=1200 & data$Superficie>=62, "SI", "NO" )

subset(data, Candidata=="SI")

dim(subset(data, Candidata=="SI")) # para ver cu�ntas candidatas tenemos

nrow(subset(data, Candidata=="SI"))

# Vamos a trabajar con las correlaciones entre variables (quitamos Exterior y Candidata)

cor(data[,-c(5,7)])

pairs(data[,-c(5,7)], main="Matriz de correlaciones") # la mayor correlaci�n aparece entre Precio y superficie

# Hagamos un scatter plot para las variables Precio y Superficie. Dibujaremos la superficie en el eje x y el precio en el eje Y

# vamos a marcar tambi�n con una l�nea nuestros l�mites de precio y de superficie (nuestros requisitos a la hora de seleccionar candidatas)

# pintaremos en azul la viviendas que son candidatas y en rojo las no candidatas

plot(data[,c(2,1)], main="Superficie vs Precio", xlab="Superficie en m2", ylab= "Precio (euros)", col=c("red","blue")[as.factor(data$Candidata)])

abline(v=62, h=1200)

# a�adimos leyenda

legend("topright", legend=c("Candidata", "No candidta"),

col=c("blue", "red"), cex=0.8, pch=1)

# text(data[,c(2,1)],labels=row.names(data)) esto tengo que mirar

# Ahora vamos a sacar muchas muestras (n) con reemplanzamiento. con estas muestras haremos la media (media muesltral).

# Las muestras seleccionadas tendr�n el mismo tama�o que la muestra original

# Guardaremos estas medias muestrales en un vector, y haremos un histograma con los valores de estas medias muestrales

n=100

size<-length(data$Precio) # detectamos el tama�o de la muestra original

media\_muestra<-vector(mode="numeric", length=n)

for( i in 1:n){

mm<-mean(sample(data$Precio,size, replace=TRUE))

media\_muestra[i]<-mm

}

hist(media\_muestra, main="Distribuci�n de las medias muestrales", xlab="Precio (euros)", ylab="frecuencia")

# vemos los valores que dejan por debajo el 2,5% de los datos y el 97,5% respectivamente

q<-quantile(media\_muestra,c(0.025,0.975))

cat("el intervalo de confianza (95%) es:", q[1], "-", q[2])

# Este proceso se llama bootstrapping. Vamos a crear una funci�n que realice el bootstrapping, tomando un n�mero n de muestras

# A la funci�n se le pasar�n los siguientes argumentos: n�mero de muestras (n), vector con los valores sobre los que se har� el muestreo (variable),

bootstrapping<-function(n, variable){

# n n�mero de muestras, variable son los valores sobre los que se hace el muestreo

size<-length(variable)

media\_muestra<-vector(mode="numeric", length=n)

for( i in 1:n){

mm<-mean(sample(variable,size, replace=TRUE))

media\_muestra[i]<-mm

}

q<-quantile(media\_muestra,c(0.025,0.975))

cat("el intervalo de confianza (95%) es:", q[1], "-", q[2])

hist(media\_muestra)

abline(v=q[1])

abline(v=q[2])

}

bootstrapping(4000,data$Precio)