```r
################################################################
# TITANIC                                                      #
# Pepa Montero                                                 #
################################################################

# PAQUETES UTILIZADOS
install.packages("xtable")
library(xtable)
install.packages("e1071")
library(e1071)

#             Análisis de los datos (ap. 3 del pdf)

setwd("") # directorio donde se encuentran los datos

# Primero guardamos los datos en una variable
data <- load(file = 'titanic_train.rda')
attach(titanic_train)

# Visualizamos el dataset
View(titanic_train)

# LIMPIEZA DEL DATASET
# Eliminar columnas innecesarias
titanic_train$Ticket <- NULL
titanic_train$PassengerId <- NULL
titanic_train$Name <- NULL

# Ver huecos vacíos
# 1. Cabin
table(Cabin)
# vemos que 687 pasajeros toman el valor " "
# eliminamos la variable Cabin
titanic_train$Cabin <- NULL
# 2. Age
colSums(is.na(titanic_train))
# vemos que 177 pasajeros no tienen valor de edad
# lo solucionaremos más adelante

# Variable Embarked
table(Embarked)
# Vemos que 644 pasajeros son de Southampton
titanic_train$Embarked <- NULL

# Volvemos a visualizar los datos, ahora solo con las variables a estudiar
View(titanic_train)


#             Análisis unidimensional


# SUPERVIVIENTES
# Tablas de frecuencias
table_survived <- table(Survived)
frec_survived <- prop.table(table_survived)
percent_survived <- frec_survived*100
# Resultado: 61.62% de muertos y 38.38% de supervivientes
# Exportar tablas para latex
table_survived_latex <- xtable(table_survived)
frec_survived_latex <- xtable(frec_survived)
# Pie chart
percent_0 <- toString(round(percent_survived[1], 2))
percent_1 <- toString(round(percent_survived[2], 2))
pie(percent_survived, labels=c(paste("Fallecidos:", percent_0, "%"),
                               paste("Supervivientes:", percent_1, "%")),
    col=c("red", "lightblue"), main="Survived")

# SEXO
# Tablas de frecuencias
table_sex <- table(Sex)
frec_sex <- prop.table(table_sex)
percent_sex <- frec_sex*100
# Resultado: 35.24% de mujeres y 64.76% de hombres
```

```r
# Exportar tablas para latex
table_sex_latex <- xtable(table_sex)
frec_sex_latex <- xtable(frec_sex)
# Pie chart
percent_female <- toString(round(percent_sex[1], 2))
percent_male <- toString(round(percent_sex[2], 2))
pie(percent_sex, labels=c(paste("Mujeres:", percent_female, "%"),
                          paste("Hombres:", percent_male, "%")),
    col=c("yellow", "magenta"), main="Sex")

# CLASE
# Tablas de frecuencias
table_class <- table(Pclass)
frec_class <- prop.table(table_class)
percent_class <- frec_class*100
# Resultado: 24.24% Clase 1, 20.65% Clase 2, 55.10% Clase 3
# Exportar tablas para latex
table_class_latex <- xtable(table_class)
frec_class_latex <- xtable(frec_class)
# Diagrama de barras
percent_class1 <- toString(round(percent_class[1], 2))
percent_class2 <- toString(round(percent_class[2], 2))
percent_class3 <- toString(round(percent_class[3], 2))
barplot(percent_class, legend.text=c("Clase 1", "Clase 2", "Clase 3"),
        col=c("yellow", "green", "blue"), main="Clase")

# EDAD
# ..................................................
# Resolver datos NA
# Sobreescribimos los NA con el valor de la media
summary(Age)
# Nos devuelve que la media es 29.70
titanic_train$Age[is.na(titanic_train$Age)] <- 29.70 # sobreescribimos
Age = titanic_train$Age
# De esta forma la media se mantiene
# ..................................................
# Convertir en franjas de edad
Age_bands <- cut(Age, 10)
# ..................................................
# Tablas de frecuencias
table_age <- table(Age_bands)
frec_age <- prop.table(table_age)
# Exportar tablas para latex
table_age_latex <- xtable(table_age)
frec_age_latex <- xtable(frec_age)
# Histograma
breaks <- c(0.34, 8.38, 16.3, 24.3, 32.3, 40.2, 48.2, 56.1, 64.1, 72, 80)
hist(Age, breaks=breaks, freq = T, col="lightgreen", include.lowest = T,
     right = T, main = "Age", xlab=NULL, ylab = "Frecuencia")
# Datos
summary(Age)
mean_age <- mean(Age)
var_age <- var(Age)
dt_age <- sd(Age)
#   La media de edad de los pasajeros es 29.70
#   El rango intercuartílico nos dice que el 50% de los pasajeros estaban
#   entre los 22 y los 35 años
#   La varianza es 169.05 y la desviación típica es 13
# ..................................................
# Coeficiente de variación
cv_age <- dt_age / mean_age * 100 # = 43.78
# Coeficiente de asimetría de Fisher
b3_age <- moment(Age, order = 3, center = TRUE)
af_age <- b3_age / dt_age**3 # = 0.43

# FAMILIARES A BORDO
Familiares = SibSp + Parch
# Tablas de frecuencias
table_fam = table(Familiares)
frec_fam = prop.table(table_fam)
# Observamos que el 60% de los pasajeros viajaban solos
# Exportar tablas para latex
table_fam_latex <- xtable(table_fam)
```

```r
frec_fam_latex <- xtable(frec_fam)
# Datos
median(Familiares) # = 0
# Diagrama de puntos
plot(frec_fam, type = "o", main = "Familiares",
     xlab = "n° de familiares a bordo", ylab = "f. relativa",
     col = "darkred")

# PRECIO DEL TICKET
summary(Fare)
# Diagrama de cajas y bigotes
boxplot(Fare, col="lightpink", main="Fare")
# Medidas de dispersión
mean_fare = mean(Fare)
dt_fare = sd(Fare) # = 49.69
b3_fare = moment(Fare, order = 3, center = TRUE)
cv_fare = dt_fare / mean_fare * 100 # = 154.3
af_fare = b3_fare / dt_fare**3 # = 4.77


#             Análisis bidimensional


# SEXO VS. SUPERVIVENCIA
# Tablas de frecuencias
table2_sex_surv <- table(Sex, Survived)
frec2_sex_surv <- prop.table(table2_sex_surv)
percent2_sex_surv <- prop.table(table2_sex_surv, 1)*100
# Exportar tablas para latex
# xtable(table2_sex_surv)
# xtable(percent2_sex_surv)
# Barras apiladas
barplot(table2_sex_surv, names.arg = c("Fallecidos", "Supervivientes"),
        col = c("lightpink", "lightblue"), main = "Sex vs. Survived",
        legend.text = c("Mujeres", "Hombres"))
# Test chi-cuadrado
chi_SS <- chisq.test(table2_sex_surv)

# EDAD VS. SUPERVIVENCIA
# Tablas de frecuencias
table2_age_surv <- table(Survived, Age_bands)
frec2_age_surv <- prop.table(table2_age_surv)
percent2_age_surv <- prop.table(table2_age_surv, 1)*100
# Exportar tablas para latex
# xtable(table(Age_bands, Survived))
# xtable(prop.table(table(Age_bands, Survived), 1)*100)
# Barras apiladas
barplot(table2_age_surv, names.arg = breaks[1:10],
        xlab = "Edad mínima de cada intervalo", main = "Age vs. Survived",
        legend.text = c("Fallecidos", "Supervivientes"))
# Test chi-cuadrado
chi_AS <- chisq.test(table2_age_surv)
# Al ejecutarlo recibimos un error,
# por lo que podría no ser concluyente

# FAMILIARES VS. SUPERVIVENCIA
# Tablas de frecuencias
table2_fam_surv <- table(Survived, Familiares)
frec2_fam_surv <- prop.table(table2_fam_surv)
percent2_fam_surv <- prop.table(table2_fam_surv, 1)*100
# Exportar tablas para latex
# xtable(table(Familiares, Survived))
# xtable(prop.table(table(Familiares, Survived), 1)*100)
# Barras agrupadas
barplot(table2_fam_surv, xlab = "n° de familiares a bordo",
        main = "Familiares vs. Survived", beside = T,
        legend.text = c("Fallecidos", "Supervivientes"))
# Convertir Familiares en una variable binaria
Familiares_bin <- Familiares
Familiares_bin[Familiares_bin > 0] <- 1
table2_famb_surv <- table(Survived, Familiares_bin)
barplot(table2_famb_surv, names.arg = c("No", "Sí"),xlab = "Tiene familiares a bordo",
        main = "Familiares vs. Survived", beside = T,
```

```r
        legend.text = c("Fallecidos", "Supervivientes"))
# Test chi-cuadrado
chi_FS <- chisq.test(table2_famb_surv)

# CLASS VS. SUPERVIVENCIA
# CLass vs. Fare
boxplot(Fare ~ Pclass)
# Vemos que claramente están relacionadas
# Luego podemos continuar sólo fijándonos en la clase
# Tablas de frecuencias
table2_class_surv <- table(Pclass, Survived)
frec2_class_surv <- prop.table(table2_class_surv)
percent2_class_surv <- prop.table(table2_class_surv, 1)*100
# Exportar tablas para latex
# xtable(table2_class_surv)
# xtable(percent2_class_surv)
# Barras apiladas
barplot(table2_class_surv, names.arg = c("Fallecidos", "Supervivientes"),
        col = c("lightpink", "lightblue", "lightgreen"), main = "Pclass vs. Survived",
        legend.text = c("Clase 1", "Clase 2", "Clase 3"))
# Test chi-cuadrado
chi_CS <- chisq.test(table2_class_surv)

# FARE VS. AGE
# Diagrama de dispersión
plot(Age, Fare)
# Covarianza
cov(Age, Fare) # = 59.16
# Correlación
cor(Age, Fare) # = 0.09
```