Laura Julià Melis
NIUB: 16810883

Memòria

Exercici 1

Apartat a (Introducció)

Per a la realització d' aquesta pràctica ha estat necessària l'obtenció d'una base de dades, la qual s'ha trobat a la pàgina web https://support.spatialkey.com/spatialkey-sample-csv-data/.

El fitxer, descarregat en format .csv, consta de 985 registres i 12 variables sobre transaccions de vivendes reals que s'han portat a terme durant 5 dies a l'àrea de Sacramento, una ciutat de Califòrnia. L'objectiu és estudiar les característiques més habituals de les vivendes que es compren avui dia de manera que les posteriors vendes podran ser previsibles.

Per al present estudi, només ens han fet falta nou variables, que són les següents:

Nom variable	Tipus de variable	Descripció
street	Factor	Identificació de cada registre (carrer)
city	Categòrica polinòmica	Ciutat en la que es troba la vivenda.
beds	Numèrica (integer)	Nombre de llits
baths	Numèrica (integer)	Nombre de banys
sq__ft	Numèrica (integer)	Superfície (peus quadrats)
type	Categòrica binària	Tipus de vivenda (condomini o residencial)
price	Numèrica (integer)	Preu (en dòlars)
latitude	Numèrica	Localització en la terra (paral·lels)
longitude	Numèrica	Localització en la terra (longitud)

Apartat b

```
#  Importació de les dades.
df  <-  read.table("SALES.csv", header = TRUE, sep = ';', nrows = 50)
df  <-  transform(df, zip = NULL, state = NULL, sale_date = NULL)
head(df)
class(df)

#  Substitució del nom de cada registre per la variable identificadora street.
df <- transform(df, row.names = street , street = NULL)
str(df)

#  Introducció de dades faltants aleatòriament.
df$beds[sample(1:50, 1)] <- NA
df$baths[sample(1:50, 2)] <- NA
```

Apartat c

A partir del resum numèric es pot observar que Sacramento és la ciutat on es troben la majoria de les vivendes venudes, i que predomina el tipus residencial. També es veu que la superfície mitjana de les cases és de 1146.8 ft^2 (uns 106.5m^2) i que el 75% tenen menys de 1300.2 ft^2 (120.8m^2). El preu mitjà és d'uns 124 573$, només un 25% de les vivendes té un preu inferior a 107 014$ i els preus màxim i mínim són 168 000$ i 59 222$, respectivament. A més, se sap que el 50% dels immobles tenen 3 habitacions i/o 2 banys.

```
#  Resum numèric de totes les variables del data frame.
summary(df)
```

De la primera finestra de gràfics es poden comentar diverses coses. Primer, com ja s'havia observat en el resum, que la ciutat més habitual és la de Sacramento i del tipus residencial, amb una gran diferència sobre tota la resta en ambdós casos. La majoria de cases té dos o tres dormitoris i en tot cas aquest nombre es troba entre 1 i 4; a més, el nombre d'habitatges amb dos banys és més del doble dels que només en tenen 1. Es podria dir que la superfície segueix una distribució normal ja que sembla haver-hi una variabilitat natural en l'histograma i, finalment, que la variable preu té asimetria negativa (si es torna a observar el resum, aquesta asimetria es confirma ja que la mediana es troba per sobre de la mitjana).

```
#  Obrim una finestra per als gràfics i establim els seus paràmetres.
windows (title = "Gràfiques de la descriptiva univariant", width = 10, height = 5)
par(mfrow = c(2, 3), lwd = 1, font = 2, font.lab = 2, font.axis = 2, las = 1)

#  Creació d'un gràfic de barres de les freqüències de les ciutats
#  on es troben les vivendes.
barplot(table(df$city), col = 3, las = 3, cex.names = 0.6,
        ylab = "Freqüències absolutes",
        main = "Diagrama de barres de la variable city")

#  Creació d'un gràfic de barres de les freqüències del nombre d'habitacions.
barplot(table(df$beds), col = 4, xlab = "Nombre d'habitacions",
        ylab = "Freqüències absolutes",
        main = "Diagrama de barres de la variable beds")

#  Creació d'un gràfic de barres de les freqüències del nombre de banys.
barplot(table(df$baths), col = 5, xlab = "Nombre de banys",
        ylab = "Freqüències absolutes",
        main = "Diagrama de barres de la variable baths")

#  Creació d'un histograma de freqüències de la superfície.
hist(df$sq__ft, col = 2, freq = F, breaks = 15,
     main = "Histograma de la variable sq__ft",
     xlab = "Peus quadrats", ylab = "Freqüències relatives")
lines(density(df$sq__ft), lwd = 2)
```

```
#  Creació d'un diagrama de sectors del tipus de vivenda.
pie(table(df$type), col = heat.colors(4) ,
      main = "Diagrama de sectors de type", border = NA)

#  Creació d'un histograma de freqüències del preu.
hist(df$price, col=gray.colors(20),
       seq(40000, 180000, by = 10000),
       xlab = "Preu", ylab = "Freqüència",
       main = "Histograma de la variable price")
```

Per acabar amb l'anàlisi descriptiva univariant, s'han realitzat dos gràfics de densitat de la latitud i la longitud dels quals es pot comentar que la major part de les vivendes es troben en una longitud entre -121.5 i -121.4, i una latitud de 38.5 o 38.7.

```
#  Obrim una nova finestra per als dos darrers gràfics
#  i establim uns nous paràmetres.
quartz (title = "Gràfiques de la descriptiva univariant", width = 6, height = 4)
par(mfrow = c(1, 2), lwd = 1, font = 2, font.lab = 2, font.axis = 2, las = 1)

# Creació d'un gràfic de densitat de la latitud.
plot(density(df$latitude), xlab = "Latitud", ylab = "Densitat",
    main = "Gràfic de densitat de la latitud")
polygon(density(df$latitude), col = "red", border = "blue")

# Creació d'un gràfic de densitat de la longitud.
plot(density(df$longitude), xlab = "Longitud", ylab = "Densitat",
    main = "Gràfic de densitat de la longitud")
polygon(density(df$longitude), col = "blue", border = "red")

#  Tanquem les dues finestres.
graphics.off()
```

Apartat d

La variable *matR* conté una matriu de correlacions de totes les variables numèriques; en la diagonal, òbviament, apareixen uns ja que el coeficient de correlació lineal d'una variable amb sí mateixa és 1. També podem comentar que la matriu obtinguda és simètrica perquè el coeficient de correlació també ho és, que el coeficient entre les variables latitud i longitud amb totes les altres és molt baix o quasi cero indicant així que el grau de relació lineal es pràcticament nul. El coeficient entre les variables de la resta de la matriu són lleugerament positives, perquè podrien tenir cert grau de relació lineal; les més correlacionades són la superfície amb el nombre de dormitoris (0.7065444) i el preu amb la superfície (0.53821697).

Ara el que es vol és contrastar si aquests valors mostrals dels coeficients són capaços de demostrar que els coeficients de correlació poblacional són significativament distints de zero. Per tant, la hipòtesi alternativa és que el coeficient es diferent de 0 i la hipòtesi nul·la que el coeficient és 0. En els dos casos, el p-valor (1.406e-08 per a les dades relatives al preu i les habitacions, i 5.544e-05 per a les de la superfície i el preu) és molt inferior a 0.05, així que es tenen evidències suficients per rebutjar la hipòtesi nul·la i afirmar que existeix una relació estadísticament significativa entre els dos parells de variables.

```
#  Matriu de correlacions:
matR <- cor(df[c(2:4, 6:8)], use = "pairwise.complete.obs")
matR

#  Test de correlació entre superfície i nombre d'habitacions de la vivenda:
cor.test(df$sq__ft, df$beds)

#  Test de correlació entre superfície i preu de la vivenda:
cor.test(df$sq__ft, df$price)
```

A continuació, es realitza una anàlisi descriptiva entre les dues variables categòriques: primer, s'ha fet una taula de contingència a partir de la qual s'observa com 29 de les 50 vivendes són residencials i estan situades a Sacramento; en segon lloc, una ordenació de forma descendent segons el preu, d'on es por concloure que els immobles més cars són els de les ciutats de Antelope i Elk Grove i que el tipus de vivenda no afecta al preu d'aquesta.

```
#  Taula de contingència entre la ciutat i el tipus de vivenda.
with(df, table(city, type))

#  Descriptiva del preu mitjà de les cases segons la ciutat i el tipus
#  de vivenda, ordenades de major a menor preu.
sort(tapply(df$price, df$city, mean), decreasing = TRUE)
sort(tapply(df$price, df$type, mean), decreasing = TRUE)
```

Finalment, s'ha dut a terme la creació de tres diagrames de dispersió. Entre les variables *price* i *sq__ft* es veu clarament la relació lineal que existeix, confirmant que hi ha una certa tendència a que les vivendes amb més superfície siguin més cares. El mateix ocorre entre el nombre de dormitoris i el preu, encara que es tracti d'una variable que només pot prendre valors enters. En darrer lloc, com si d'un mapa de la terra es tractàs, es pot veure les zones de Califòrnia on es localitzen la majoria de les cases de la base de dades, les que tenen una latitud de 38.5 i una longitud de -121.45, i les d'una latitud de 38.7 i una longitud de -121.35.

```
# Creació d'un gràfic bivariant entre el preu i la superfície.
plot(sq__ft~price, data = df, pch = 20, las = 1,
    xlab = "Preu de la vivenda", ylab = "Superfície (peus quadrats)")
title("Diagrama de punts entre el preu i la superfície")
abline(lm(sq__ft~price, data = df), col = 2)

# Creació d'un gràfic bivariant entre el preu i el nombre de llits.
plot(beds~sq__ft, data = df, pch = 19, col = 2, yaxt = 'n',
    xlab = "Preu de la vivenda", ylab = "Nombre de llits")
title("Diagrama de punts entre el preu i el nombre de llits")
abline(lm(beds~sq__ft, data = df))
axis(2, at = c(0, 1, 2, 3, 4))
```

```
# Creació d'un gràfic bivariant entre la latitud i la longitud.
plot(df$longitude~df$latitude, pch = 4, las = 1, xlab = "Latitud",
     ylab = "Longitud")
title("Diagrama de punts entre la latitud i la longitud")

#  Tanquem els gràfics.
dev.off()
```

Apartat e

Tota aquesta informació extreta a partir de la base de dades permet arribar a la conclusió que l'immoble més fàcil d'aconseguir vendre és aquell residencial situat a la ciutat de Sacramento que tingui 3 habitacions i 2 banys, d'uns 1146.8 ft^2 i un preu de 124 573$.

Per acabar, els compradors hauran de tenir en compte que, quan més dormitoris o més peus quadrats vulguin per a la casa, més doblers hauran de pagar. Si volen estalviar-se una mica de diners, una bona idea seria considerar comprar una casa a les ciutats de Rio Linda o Rancho Cordova, ja que són les més econòmiques.