

PRACTICA - 2

Autor

9 de enero, 2023

Contents

1. Descripció del dataset	1
2. Integració i selecció	2
3. Neteja de dades	4
4. Anàlisi de dades	11
5. Resolució del problema	18

1. Descripció del dataset

Per a la realització d'aquesta segona pràctica s'utilitza el *dataset* generat a la primera pràctica, i es combina amb altres *datasets* que resultin d'interès, per poder realitzar un anàlisi més profund, tenint en compte factors socioeconòmics.

Com a recordatori; el conjunt de dades extret en la PRACTICA_1 conté la informació (general i de preu) de totes les estacions de servei d'Espanya obtinguda en cinc dies consecutius, del 14/11/2022 al 18/11/2022 ambdós inclosos. Per a cada execució de *web scraping* (diària), es van extreure les dades de totes les estacions de servei a totes les províncies d'Espanya i per a tots els tipus de carburants disponibles en la pàgina.

Preguntes que es pretenen respondre <- FALTA XAVI

Fins aquí

Per a la correcta execució del *script* és imprescindible definir la ruta on es troba l'arxiu **Practica_2.Rmd** com a *working directory*.

A continuació es procedeix a la lectura de l'arxiu de dades amb opcions, escollint el separador i el tipus de codificació.

```
#Importar arxiu
fueldata <- read.csv(paste(datadir, "FuelScrapper", "dataset.csv", sep = "/"),
                     encoding="UTF-8", sep=";")
#Visualitzar summary de dades
summary(fueldata)
```

```
## Capture_date      Capture_time      Province      City
## Length:195357      Length:195357      Length:195357      Length:195357
## Class :character    Class :character    Class :character    Class :character
## Mode :character     Mode :character     Mode :character     Mode :character
##
##
##
## Address            Road_side          Update_date          Price
## Length:195357      Length:195357      Length:195357      Min. :0.768
## Class :character    Class :character    Class :character    1st Qu.:1.779
```

```
## Mode :character Mode :character Mode :character Median :1.879
## Mean :1.836
## 3rd Qu.:1.959
## Max. :3.700
## Brand Sale_1 Sale_2 Fuel_type
## Length:195357 Length:195357 Length:195357 Length:195357
## Class :character Class :character Class :character Class :character
## Mode :character Mode :character Mode :character Mode :character
##
##
##
```

El primer anàlisi del *dataset* indica que pot ser interessant canviar algunes dades a tipus factor, així com canviar el format de les variables temporals.

```
#Vector de variables a modificar
t_vector<-c("Province","Road_side","Sale_1", "Sale_2", "Fuel_type")

#Loop
for (i in t_vector){
  #Canvi de tipus a factor
  fueldata[,i]<-as.factor(fueldata[,i])
}

#Canvis en variables temporals
fueldata$Capture_date<-as.Date(fueldata$Capture_date, format = "%Y/%m/%d")
fueldata$Update_date<-as.Date(fueldata$Update_date, format = "%d/%m/%Y")
fueldata$Capture_time<-lubridate::hms(fueldata$Capture_time)
```

Com a darrer pas en el procés de càrrega del *dataset* original es generen variables de *backup* per a *Province* i *City* ja que aquestes posteriorment s'hauran de modificar a través de processos de normalització de noms per tal de fer-les compatibles amb les dades dels altres *datasets* que es volen integrar.

```
#Creació de variables de backup
fueldata$bckup.Province<-fueldata$Province
fueldata$bckup.City<-fueldata$City
```

2. Integració i selecció

Amb l'objectiu d'obtenir un *dataset* amb més informació que permeti generar més valor, es llegeix un arxiu addicional amb el cens de població per municipis. Aquesta informació es extreta de l'Institut Nacional d'Estadística (INE). En aquest cas, la lectura també es realitza amb opcions.

```
#Importar arxiu
pobdata <- read.csv(paste(datadir, "pobmun", "pobmun22.csv", sep = "/"),
  encoding="UTF-8", sep=";")
```

Es canvien els noms de les variables i transformem les dades a majúscules per habilitar posteriors comparacions entre els dos *datasets*.

```
#Canvis de noms
names(pobdata)[names(pobdata) == "PROVINCIA"] <- "Province"
names(pobdata)[names(pobdata) == "NOMBRE"] <- "City"
names(pobdata)[names(pobdata) == "CPRO"] <- "P_code"
names(pobdata)[names(pobdata) == "CMUN"] <- "C_code"
names(pobdata)[names(pobdata) == "POB22"] <- "Population"
names(pobdata)[names(pobdata) == "HOMBRES"] <- "P_Male"
```

```
names(pobdata)[names(pobdata) == "MUJERES"] <- "P_Female"
```

```
#Transformació a majúscules
```

```
pobdata$Province<-toupper(pobdata$Province)
pobdata$City<-toupper(pobdata$City)
```

També, es normalitzen les paraules en les variables *Province* i *City* dels dos *datasets*, eliminant accents i caràcters especials com la ñ. Per fer-ho es canvia el tipus de dades d'aquestes variables de UTF-8 a ASCII.

```
#Conversió de codificació per a normalització de caràcters
```

```
fueldata$Province<-iconv(fueldata$Province, from = 'UTF-8', to = 'ASCII//TRANSLIT')
fueldata$City<-iconv(fueldata$City, from = 'UTF-8', to = 'ASCII//TRANSLIT')
pobdata$Province<-iconv(pobdata$Province, from = 'UTF-8', to = 'ASCII//TRANSLIT')
pobdata$City<-iconv(pobdata$City, from = 'UTF-8', to = 'ASCII//TRANSLIT')
```

Es canvia la denominació de 3 províncies per tal de fer la informació compatible entre els *datasets* de preus de combustibles i de població per municipis.

```
#Canvis específics en variable província
```

```
fueldata[fueldata$Province=="ALICANTE","Province"]<-"ALICANTE/ALACANT"
fueldata[fueldata$Province=="VALENCIA / VALENCIA","Province"]<-"VALENCIA/VALENCIA"
fueldata[fueldata$Province=="CASTELLON / CASTELLO","Province"]<-"CASTELLON/CASTELLO"
```

A continuació es modifica l'ús d'articles en els camps *Province* i *City* utilitzant RegEx, també buscant la compatibilitat entre *datasets*.

```
#Canvis en l'ús d'articles a través de RegEx
```

```
fueldata$Province<-sub("(\\w+) \\((\\w+)\\)", "\\1, \\2", fueldata$Province, fixed=FALSE)
fueldata$City<-sub("(\\w| )+ \\((\\w|')+)", "\\1, \\3", fueldata$City, fixed=FALSE)
```

Finalment es realitzen tot un seguit de canvis individuals (que no es mostren en la memòria, però si en el codi), per tal de maximitzar la informació disponible en el *dataset* resultant.

S'integren els dos *datasets* amb l'objectiu d'obtenir un únic *dataset* resultant que contingui tota la informació combinada. Aquesta integració es realitza de manera completa (all = TRUE), per tal de garantir que les dades que no tenen una parella en l'altre *dataset* es mantenen afegint NA en la informació.

```
#Combinació de datasets
```

```
total<-merge(fueldata, pobdata, by=c("Province", "City"), all = TRUE)
```

De manera adicional a les tasques d'integració s'analitza la variable *Brand* específicament, convertint-la en factor i obtenint les 10 marques amb més representació.

```
#Obtenció de marques mes rellevants
```

```
total$Brand.factor<-as.factor(total$Brand)
Brands<-as.data.frame(head(summary(total$Brand.factor),10))
names(Brands)[1]<-"Stations"
kable(Brands)
```

	Stations
REPSOL	57624
CEPSA	26791
GALP	9540
SHELL	7093
BP	4059
PETRONOR	3541
AVIA	2778

	Stations
CARREFOUR	2635
BALLENOIL	2189
CAMPSA	1511

Un cop obtingudes les 10 marques amb més representació s'analitzen tots els registres per tal de normalitzar el camp marca. També per definir com a “OTROS” aquelles estacions de servei que no són d'una marca de les 10 més representatives.

```
#Noms de les 10 marques més representatives
Brand.names<-row.names(Brands)
total$Brand.factor<-as.character(total$Brand.factor)

#Iteració en les 10 marques, normalitzant nom de la marca si el conté en el string
for (brand in Brand.names){
  total$Brand.factor<-if_else(grepl(brand, total$Brand.factor),
                              brand, total$Brand.factor)
}

#Assignació de camp "OTROS"
total$Brand.factor<-if_else(total$Brand.factor %in% Brand.names,
                             total$Brand.factor, "OTROS")
total$Brand.factor<-as.factor(total$Brand.factor)
```

Finalment, i per acabar amb les tasques de selecció es seleccionen les dades:

- del dia 16 de Novembre
- els combustibles
 - Gasóleo A habitual
 - Gasolina 95 E5

```
#Selecció de dades
data<-total[total$Capture_date == as.Date("2022/11/16", format ="%Y/%m/%d"),]
data<-data[data$Fuel_type == "Gasóleo A habitual" | data$Fuel_type == "Gasolina 95 E5",]
```

3. Neteja de dades

Zeros i elements buits

Com a primer pas en la neteja de dades, es procedeix a eliminar tots aquells registres del *dataset* resultant on el camp *Capture_date* sigui *NA*. Aquests seran municipis que apareixen en el cens de població, però no tenen benzineria. Aquests han aparegut en el *dataset* quan s'ha realitzat l'operació de combinació plena o *FULL JOIN* en el pas anterior.

```
#Neteja de NAs
data<-data[!is.na(data$Capture_date),]
```

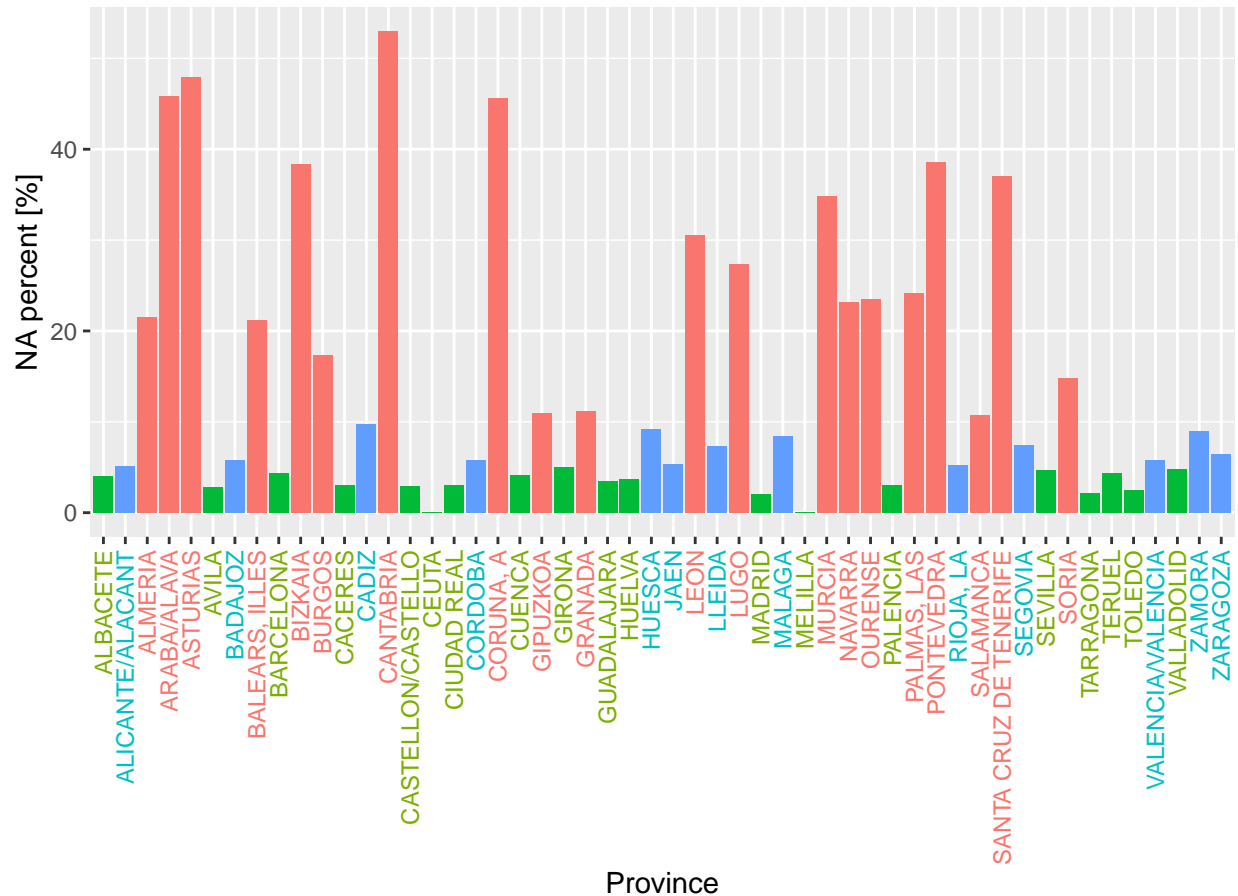
NOTA: Aquest pas seria prescindible si en el moment de realitzar l'operació *merge* anterior haguessin realitzat una *LEFT JOIN* amb les opcions **all.x = TRUE**, **all.y = FALSE**, enlloc del **all = TRUE** utilitzat.

Actualment el *dataset* conté 21329 registres, dels quals 2840 són registres dels quals no se'n coneix el cens. Això representa un 13% del total de registres.

Quan es procedeix a realitzar aquest mateix estudi per a cada una de les províncies amb l'objectiu d'identificar aquelles que tinguin una representació més pobre s'obté el següent gràfic en el que es marquen:

- en verd, aquelles províncies amb un percentatge de NA inferior a 5%

- en blau les províncies amb un percentatge entre 5 i 10%
- en vermell aquelles províncies amb un percentatge de NA superior al 10%



Conseqüentment, per a estudis que no considerin dades de cens per municipis, es pot utilitzar el *dataset* complet ja que aquest prové de l'operació de *scraping* i no conté zeros o valors nuls. Tanmateix, quan l'objectiu de l'estudi requereixi considerar informació del cens; per raons de representativitat, es recomana limitar l'estudi a les províncies anteriorment identificades en verd i en blau.

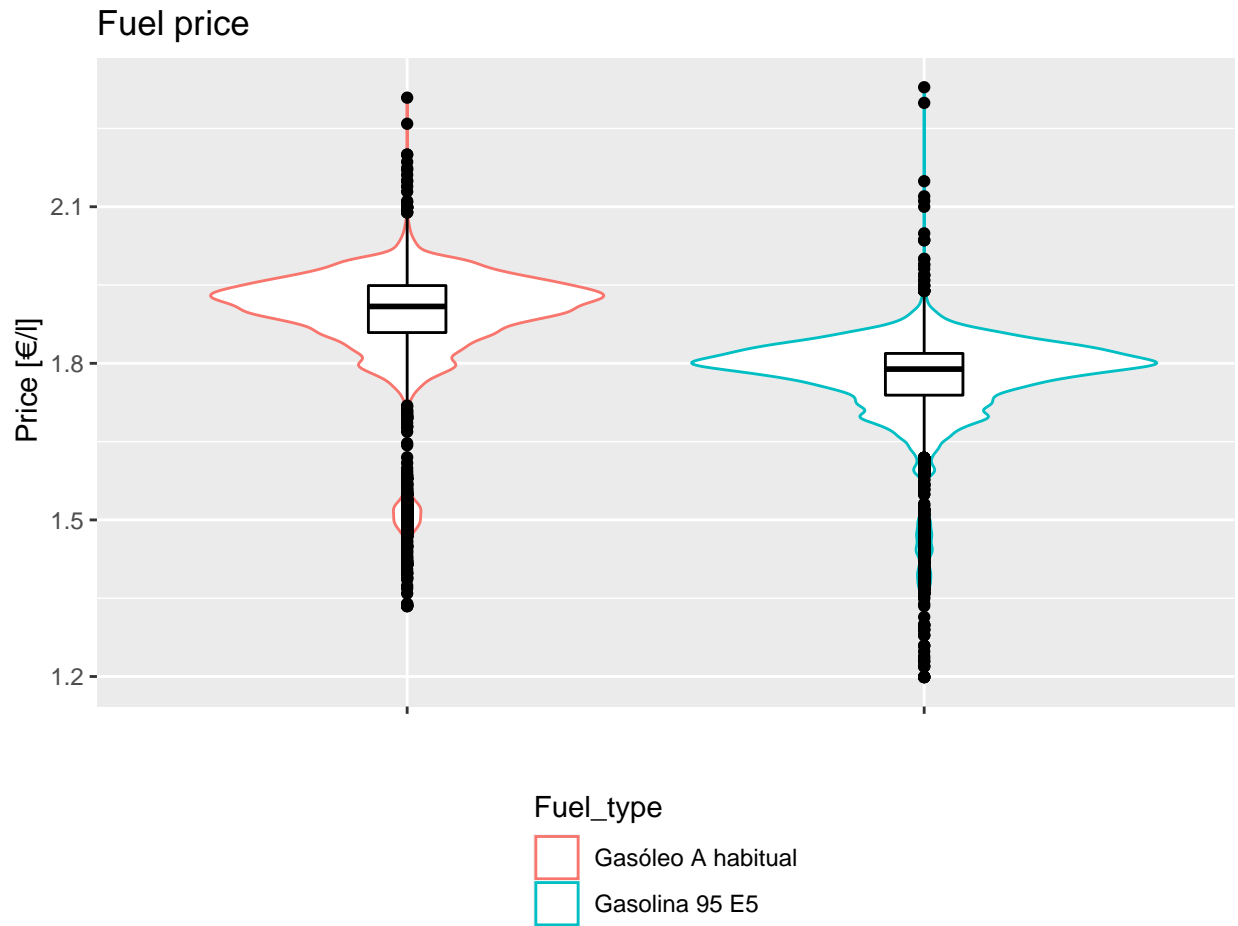
#Agrupament de províncies segons NAs percent

```
green_province<-DT[DT$Population<5,1]
```

```
blue_province<-DT[DT$Population>=5 & DT$Population<10,1]
```

Valors extrems

Per tal d'identificar els valors extrems es genera un *violin plot* amb un *boxplot* al interior per cada tipus de combustible. Aquest tipus de gràfic ampliat ens permet, per una banda analitzar els *outliers* a través de la visualització *boxplot*, alhora que ens permet conèixer la distribució de la població a través de la visualització *violin*. Així doncs permet, d'un cop de vista, veure si la informació aporta el *boxplot* es consistent amb la distribució de les dades.



Observant l'existència de valors extrems s'aprofundeix en l'anàlisi, inicialment obtenint els valors característics del boxplot per a cada un dels combustibles seleccionats i graficats.

	Min	Q1	Med	Q3	Max
Gasóleo A habitual	1.727	1.859	1.909	1.949	2.079
Gasolina 95 E5	1.622	1.739	1.789	1.819	1.933

S'observa que la diferència de preus entre les medianes de la població *Gasóleo A habitual* i *Gasolina 95 E5* és de 0.12€/l, sent el combustible *Gasóleo A habitual* el més car.

Valors extrems superiors

S'obtenen els registres que son valors extrems superiors tant pel cas del combustible *Gasóleo A habitual* com pel *Gasolina 95 E5*.

```
#Valors extrems superiors Gasóleo A habitual
st<-boxplot.stats(data[data$Fuel_type=="Gasóleo A habitual","Price"])
DiesUP<-data[data$Fuel_type=="Gasóleo A habitual" & data$Price>st$stats[5],] %>%
  group_by(Province, Brand) %>%
  as.data.frame()

#Valors extrems superiors Gasolina 95 E5
st<-boxplot.stats(data[data$Fuel_type=="Gasolina 95 E5","Price"])
```

```
GasUP<-data[data$Fuel_type == "Gasolina 95 E5" & data$Price>st$stats[5],] %>%
  group_by(Province, Brand) %>%
  as.data.frame()
```

Combinant la informació anterior, s'obté les dades de les estacions de servei on els dos combustibles es consideren *outliers*. Fet que indica que la estació de servei en general té uns preus més cars que la mitjana.

```
#Combinació de valors extrems
UPs<-merge(DiesUP, GasUP, by=c("Province", "City", "Capture_date", "Address", "Brand"),
  all = FALSE)

#Ordenar valors
UPs<-UPs[c("Province", "City", "Address", "Brand", "Fuel_type.x", "Price.x",
  "Fuel_type.y", "Price.y")] %>%
  arrange(Province, City, Brand, Address)
```

Finalment es comprova quines estacions de servei (de les marcades com a cares) no mantenen la diferència de medianes de preus de combustibles obtinguda anteriorment. Per tal de ser una mica més flexibles, es considerarà com a límit el 80% de la diferència anterior. Per tant es seleccionen els registres on la diferència de preu entre els dos combustibles sigui inferior a 0.1€/l.

Els registres d'aquestes estacions de servei s'eliminen de l'estudi per a ser *outliers* i no mostrar prou consistència en els seus valors, que ens permetin garantir que els preus no continguin errades.

```
#Comprovació incosistència
UPs<-UPs[UPs$Price.x-UPs$Price.y < round(0.8*(stt[1,3]-stt[2,3]),2),]

#Extreure dades inconsistentes
data<-anti_join(data,UPs, by=c("Province", "City", "Address", "Brand"))
```

NOTA: L'estudi d'*outliers* en l'extrem superior es podria allargar força més considerant la consistència de registres en l'horitzó temporal, considerant la consistència de registres d'una mateixa marca en una zona pròxima ...

Valors extrems inferiors

Observant el *violin plot* per al combustible *Gasóleo A habitual*, crida l'atenció la concentració de mostres al voltant del preu 1.5€/l, import que es considera "extrem". Així doncs inicialment, es seleccionen els valors per estudiar-los amb més profunditat.

```
#Selecció de registres en la zona objectiu i summarització
DiesConc<-data[data$Fuel_type == "Gasóleo A habitual" &
  1.55>data$Price &
  data$Price>1.45,] %>%
  group_by(Province) %>%
  summarise(n=n()) %>%
  as.data.frame()
kable(DiesConc)
```

Province	n
CEUTA	9
MELILLA	1
PALMAS, LAS	185
SANTA CRUZ DE TENERIFE	167

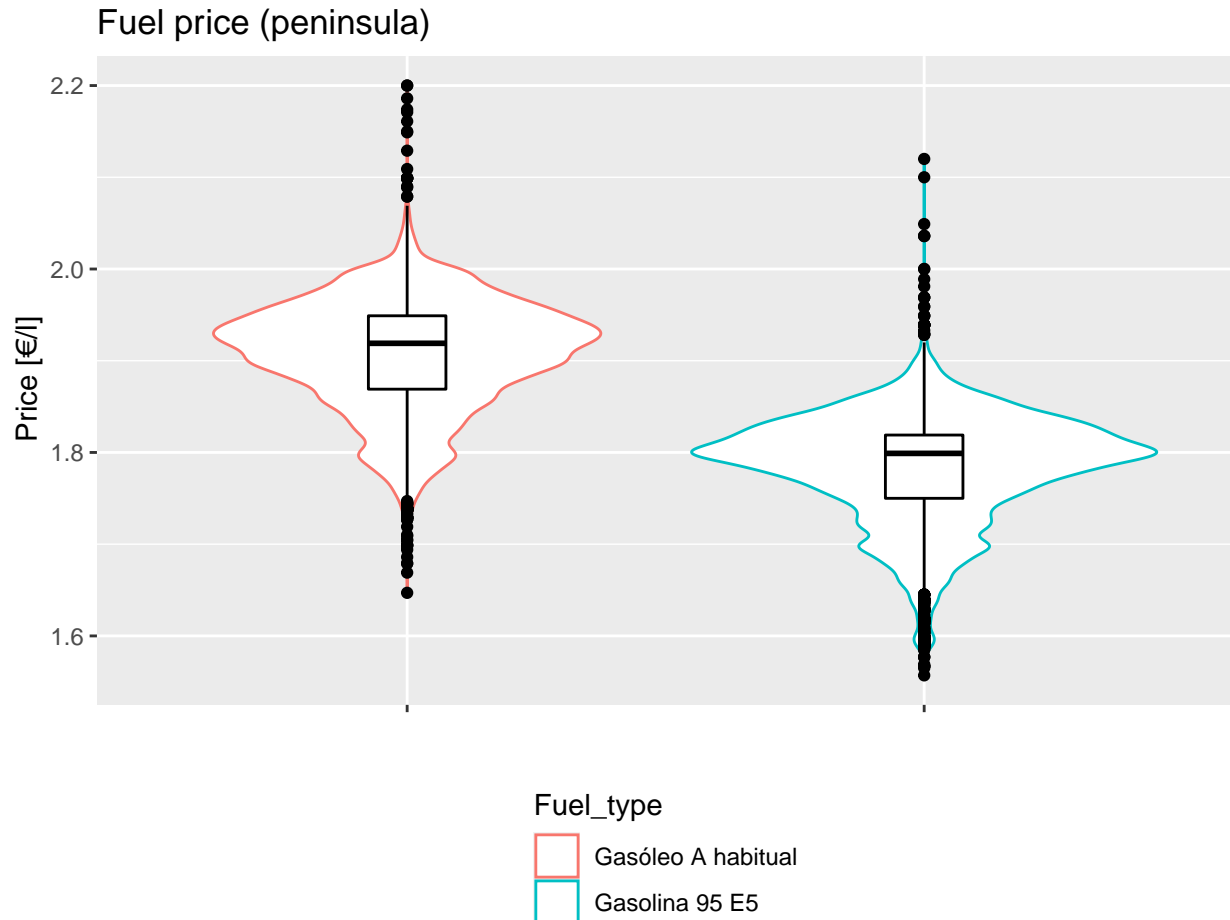
On s'observa que els registres obtinguts pertanyen tots a províncies amb tipus impositius especials. Això

indica, que per tal de fer un estudi coherent del preu dels combustibles s'hauran de considerar els registres de les estacions de servei peninsulars de manera separada dels registres insulars i de ciutats autònomes.

#Selecció dades peninsulars

```
non_peninsula=c("CEUTA", "MELILLA", "PALMAS, LAS", "SANTA CRUZ DE TENERIFE")
data.peninsula<-data[!(data$Province %in% non_peninsula), ]
```

Graficant novament s'observa una clara reducció en el nombre de valors extrems.



Per finalitzar l'anàlisi dels valors extrems inferiors, es procedeix d'acord al mètode de selecció utilitzat prèviament en els valors extrems superiors. Així doncs, s'obtenen els registres que són valors extrems inferiors tant pel cas del combustible *Gasóleo A habitual* com pel *Gasolina 95 E5*.

#Valors extrems superiors Gasóleo A habitual

```
st<-boxplot.stats(data.peninsula[data.peninsula$Fuel_type=="Gasóleo A habitual","Price"])
DiesLow<-data.peninsula[data.peninsula$Fuel_type == "Gasóleo A habitual" &
                        data.peninsula$Price<st$stats[1],] %>%
  group_by(Province, Brand) %>%
  as.data.frame()
```

#Valors extrems superiors Gasolina 95 E5

```
st<-boxplot.stats(data.peninsula[data.peninsula$Fuel_type=="Gasolina 95 E5","Price"])
GasLow<-data.peninsula[data.peninsula$Fuel_type == "Gasolina 95 E5" &
                      data.peninsula$Price<st$stats[1],] %>%
  group_by(Province, Brand) %>%
```



```
as.data.frame()
```

Es vol saber si els registres seleccionats com a valors extrems provenen d'estacions de servei amb una marca entre les 10 més representatives o formen part del grup "OTROS" que inclou estacions de servei de baix cost.

```
#Anàlisis de les marques dels valors extrems inferiors
```

```
st_D<-summary(DiesLow$Brand.factor)
```

```
st_G<-summary(GasLow$Brand.factor)
```

```
stt<-rbind(st_D,st_G)
```

```
rownames(stt)<-c("Gasóleo A habitual","Gasolina 95 E5")
```

```
kable(t(stt))
```

	Gasóleo A habitual	Gasolina 95 E5
AVIA	0	0
BALLENOIL	5	18
BP	1	0
CAMPSA	0	1
CARREFOUR	0	0
CEPSA	0	1
GALP	0	2
OTROS	49	215
PETRONOR	0	0
REPSOL	0	0
SHELL	1	4

En vista dels resultats interessa analitzar amb més detall els registres per a les estacions de servei amb menors freqüència d'aparició (diferent de 0) per tractar-se de possibles errades.

```
#Anàlisis pel combustible Gasóleo A habitual
```

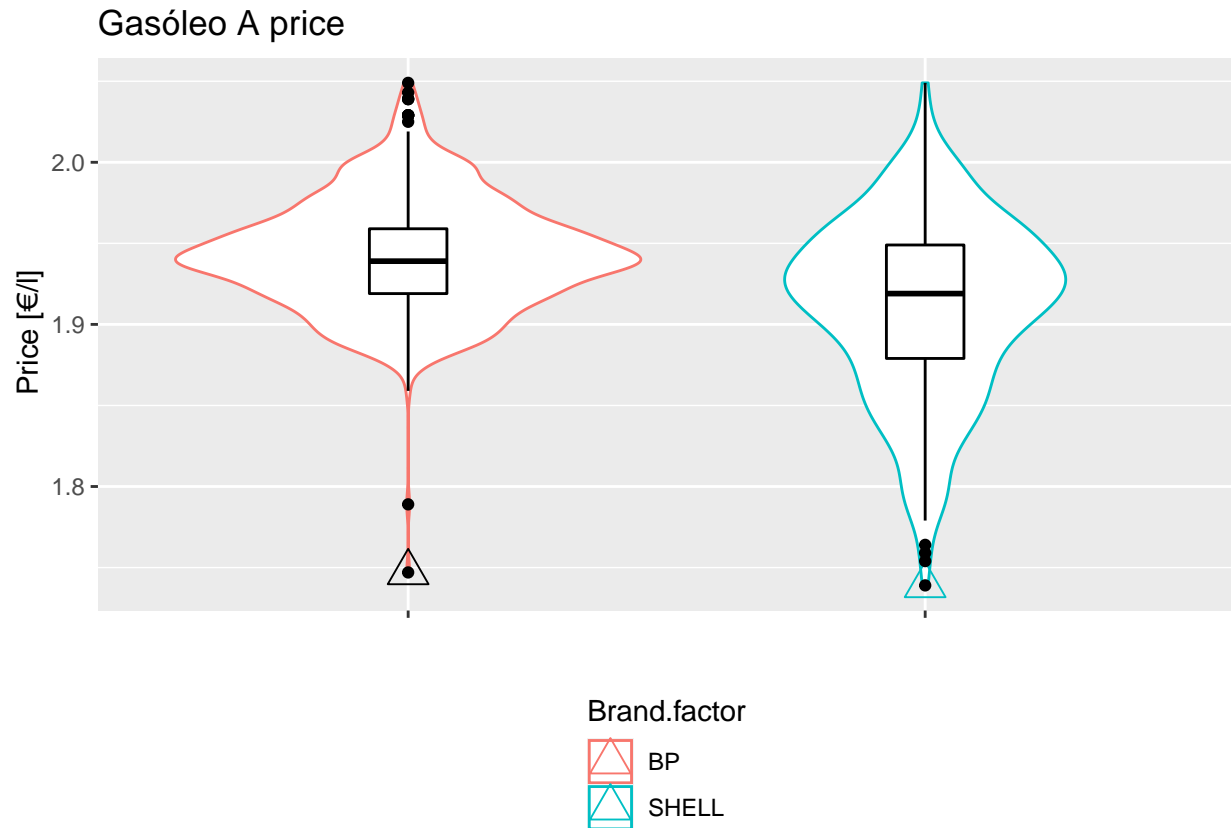
```
D_1<-DiesLow[DiesLow$Brand.factor=="BP",]
```

```
D_2<-DiesLow[DiesLow$Brand.factor=="SHELL",]
```

```
Diesel<-data.peninsula[data.peninsula$Fuel_type=="Gasóleo A habitual" &  
  (data.peninsula$Brand.factor=="BP" |  
    data.peninsula$Brand.factor=="SHELL"),]
```

```
g3<-ggplot(data=Diesel, aes(x=Brand.factor, y=Price, color=Brand.factor)) +  
  geom_violin() +  
  geom_point(data = D_1, size=5, color="black", shape = 2) +  
  geom_point(data = D_2, size=5, shape = 2) +  
  labs(title = "Gasóleo A price", y="Price [€/l]", x="") +  
  theme(axis.text.x = element_blank(), legend.direction = "vertical",  
        legend.position = "bottom")+  
  geom_boxplot(width=0.15, color="black")
```

```
g3
```



En vista dels resultats, es considera el valor extrem de la marca BP poc consistent, mentre que el valor extrem de la marca SHELL es manté.

```
#Element poc consistent a extreure
to_remove_1<-DiesLow[DiesLow$Brand.factor=="BP",]
```

Es continua amb l'estudi pel combustible Gasolina 95 E5

```
#Anàlisis pel combustible Gasolina 95 E5
G_1<-GasLow[GasLow$Brand.factor=="CAMPSA",]
G_2<-GasLow[GasLow$Brand.factor=="CEPSA",]
G_3<-GasLow[GasLow$Brand.factor=="GALP",]
G_4<-GasLow[GasLow$Brand.factor=="SHELL",]
G_4_min<-G_4[which.min(G_4$Price),]
Gasolina<-data.peninsula[data.peninsula$Fuel_type=="Gasolina 95 E5" &
  (data.peninsula$Brand.factor=="CAMPSA" |
   data.peninsula$Brand.factor=="CEPSA" |
   data.peninsula$Brand.factor=="GALP" |
   data.peninsula$Brand.factor=="SHELL"),]

g4<-ggplot(data=Gasolina, aes(x=Brand.factor, y=Price, color=Brand.factor)) +
  geom_violin() +
  geom_point(data = G_1, size=5, shape = 2) +
  geom_point(data = G_2, size=5, color="black", shape = 2) +
  geom_point(data = G_3, size=5, shape = 2) +
  geom_point(data = G_4_min, size=5, color="black", shape = 2) +
  labs(title = "Gasolina 95 E5 price", y="Price [€/l]", x="") +
  theme(axis.text.x = element_blank(), legend.direction = "vertical",
```

```
legend.position = "bottom")+
geom_boxplot(width=0.15, color="black")
```

g4



A la vista dels resultats, es consideren valors poc consistents:

- Valor extrem inferior per la marca “CEPSA”
- Valor extrem mínim per la marca “SHELL”

Finalment es procedeix a eliminar els valors que es consideren poc consistents.

```
#Elements poc consistents
to_remove_2<-GasLow[GasLow$Brand.factor=="CEPSA",]
to_remove_3<-GasLow[GasLow$Brand.factor=="SHELL",]
to_remove_3<-to_remove_3[which.min(to_remove_3$Price),]

#Combinació d'elements
to_remove<-rbind(to_remove_1, to_remove_2, to_remove_3)

#Extracció
data.peninsula<-anti_join(data.peninsula,to_remove,
                           by=c("Province", "City", "Address", "Brand", "Fuel_type"))
```

4. Anàlisi de dades

```
total<-data.peninsula
summary(total$Price)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  1.557   1.789   1.839   1.847   1.919   2.200
```

No hauriem de fer primer els tests de normalitat i de homoscedasticitat?? <-XAVI

En primer lloc, fem un test per comprovar la diferència de mitjanes entre les ciutats de Barcelona i Madrid. Apliquem un filtre utilitzant la funció filter() per seleccionar únicament les ciutats que ens interessin. Seleccionem en primer lloc el carburant Gasoil i, seguidament, seleccionem Gasolina. D'aquesta manera veurem si hi ha diferències entre les ciutats per cada tipus de carburant.

```
capital_filter <- total %>%
  filter(Fuel_type == "Gasóleo A habitual" & (City == "BARCELONA" | City == "MADRID"))

t.test(Price ~ City, data = capital_filter)
```

```
##
## Welch Two Sample t-test
##
## data: Price by City
## t = -2.8273, df = 154.8, p-value = 0.005315
## alternative hypothesis: true difference in means between group BARCELONA and group MADRID is not equal to 0
## 95 percent confidence interval:
## -0.030083327 -0.005335985
## sample estimates:
## mean in group BARCELONA      mean in group MADRID
##           1.906286           1.923995
```

```
capital_filter2 <- total %>%
  filter(Fuel_type == "Gasolina 95 E5" & (City == "BARCELONA" | City == "MADRID"))

t.test(Price ~ City, data = capital_filter2)
```

```
##
## Welch Two Sample t-test
##
## data: Price by City
## t = 2.7598, df = 149.49, p-value = 0.006507
## alternative hypothesis: true difference in means between group BARCELONA and group MADRID is not equal to 0
## 95 percent confidence interval:
##  0.006163916 0.037239523
## sample estimates:
## mean in group BARCELONA      mean in group MADRID
##           1.818845           1.797144
```

Pels dos tests el p-value és inferior a alfa, fet que fa que rebutgem la hipòtesi nul·la. Per tant, arribem a la conclusió que les diferències de preu entre les dues capitals són significatives tant per la gasolina com pel gasoil.

A continuació fem un test ANOVA per veure si hi ha diferències significatives entre la mitjana de preus dels dos combustibles en funció del nombre d'habitants del municipi.

```
pop_filter <- total %>%
  filter(Fuel_type == "Gasóleo A habitual") %>%
  mutate(Pop_size = ifelse(Population < 10000, "S",
```

```

        ifelse(Population > 100000, "L", "M"))))
pop_size_test <- aov(Price ~ Pop_size, data = pop_filter)
summary(pop_size_test)

```

```

##              Df Sum Sq Mean Sq F value Pr(>F)
## Pop_size      2   0.41  0.20669   58.06 <2e-16 ***
## Residuals  9061  32.26  0.00356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1308 observations deleted due to missingness

```

Com veiem, el p-valor és més petit que 0,05 i ens porta a RH0. Arribem a la conclusió que les diferències entre mitjanes són significatives.

```
TukeyHSD(pop_size_test)
```

```

## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = Price ~ Pop_size, data = pop_filter)
##
## $Pop_size
##      diff      lwr      upr      p adj
## M-L -0.003287284 -0.007297198 0.000722631 0.1326108
## S-L  0.011425009  0.007363463 0.015486555 0.0000000
## S-M  0.014712293  0.011428954 0.017995631 0.0000000

```

El resultat del test de Tukey mostra que les mitjanes són diferents, ja que el p-value és inferior al nivell de significació, i RH0.

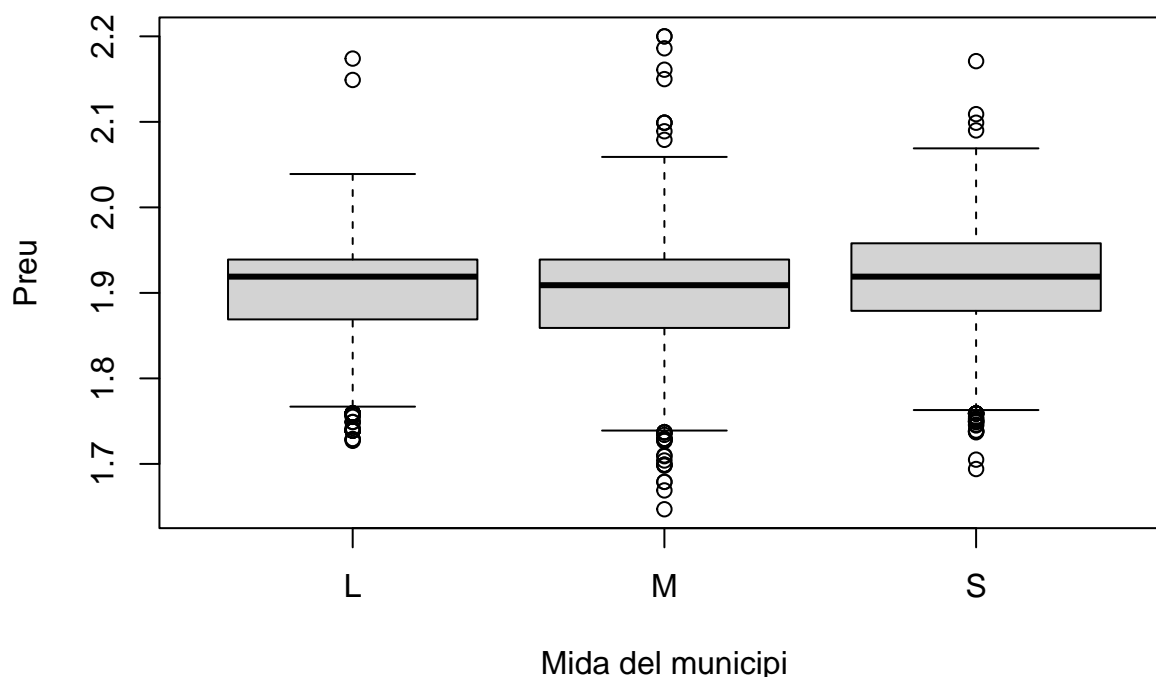
A continuació es mostra el BoxPlot pel preu del gasoil gasolina en funció de la dimensió del municipi.

```

boxplot(Price ~ Pop_size, data = pop_filter, main = "BoxPlot Gasoil",
        xlab = "Mida del municipi", ylab = "Preu")

```

BoxPlot Gasoil



Tornem a repetir el mateix procediment, però pel cas de la gasolina 95.

```
pop_filter_fuel <- total %>%
  filter(Fuel_type == "Gasolina 95 E5") %>%
  mutate(Pop_size = ifelse(Population < 10000, "S",
                           ifelse(Population > 100000, "L", "M")))
pop_size_test_fuel <- aov(Price ~ Pop_size, data = pop_filter_fuel)
summary(pop_size_test_fuel)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Pop_size      2  0.261  0.13048   37.34 <2e-16 ***
## Residuals 8738 30.536  0.00349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1249 observations deleted due to missingness
```

A continuació repetim el test de Tukey, tal com hem fet pel cas del gasoil.

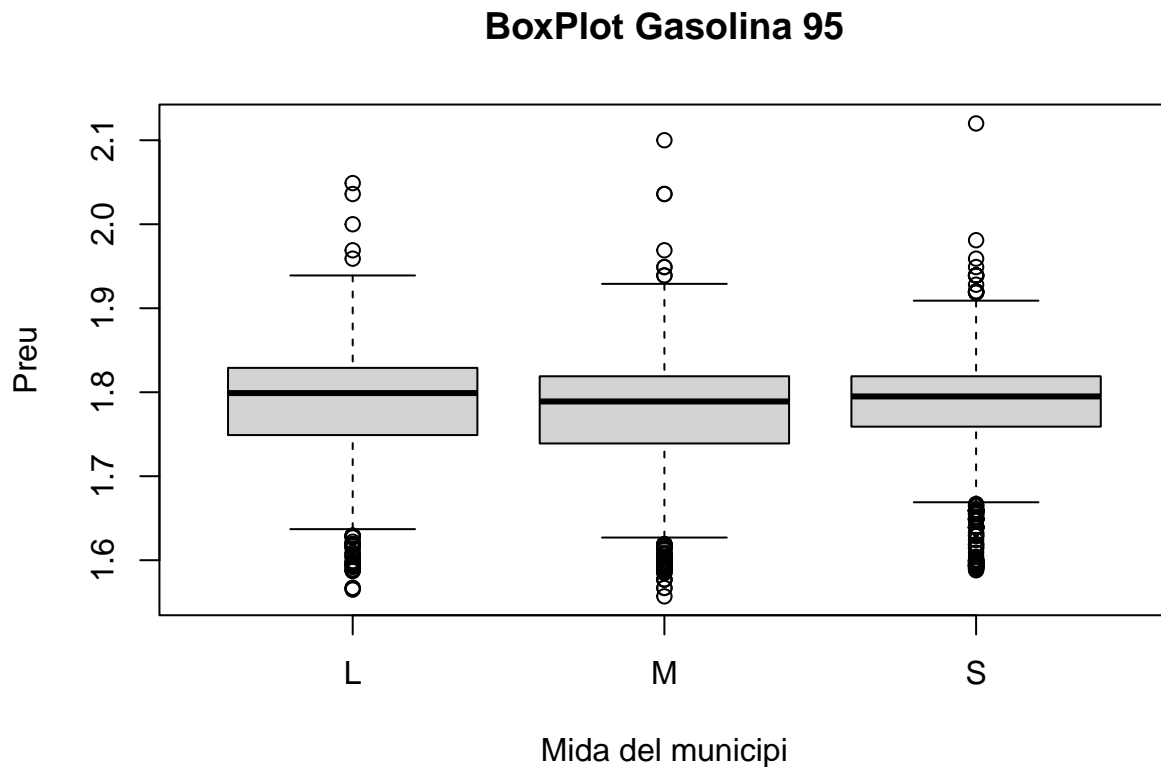
```
TukeyHSD(pop_size_test_fuel)
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = Price ~ Pop_size, data = pop_filter_fuel)
##
## $Pop_size
##          diff          lwr          upr      p adj
## M-L -0.009756182 -0.013767491 -0.005744873 0.0000000
## S-L  0.001900579 -0.002171768  0.005972926 0.5178047
```

```
## S-M 0.011656761 0.008332595 0.014980926 0.0000000
```

Obtenim els mateixos resultats que en el cas del gasoil. Les mitjanes són diferents.

```
boxplot(Price ~ Pop_size, data = pop_filter_fuel, main = "BoxPlot Gasolina 95",  
        xlab = "Mida del municipi", ylab = "Preu")
```



Els resultats d'aquest primer anàlisi mostren que, de mitjana, el gasoil és més car que la gasolina tant per la ciutat de Barcelona com per Madrid. També veiem que mentre el preu del gasoil és més barat a Barcelona, el preu de la gasolina és més barat a Madrid.

En segon lloc volem estudiar l'existència de correlació entre la variable "Price" i la variable "Population". Sospitem que poden estar correlacionades per diferents motius.

Primer crearem 2 dataframes nous per facilitar el tractament i anàlisi. Aquests contenen el preu dels carburants Gasoil i Gasolina, respectivament.

```
gasoil <- subset(data, Fuel_type == "Gasóleo A habitual")  
gasolina <- subset(data, Fuel_type == "Gasolina 95 E5")
```

A continuació fem un test de correlació per les dues variables.

```
cor.test(gasoil$Price, gasoil$Population)
```

```
##  
## Pearson's product-moment correlation  
##  
## data: gasoil$Price and gasoil$Population  
## t = 3.6526, df = 9400, p-value = 0.0002611  
## alternative hypothesis: true correlation is not equal to 0
```

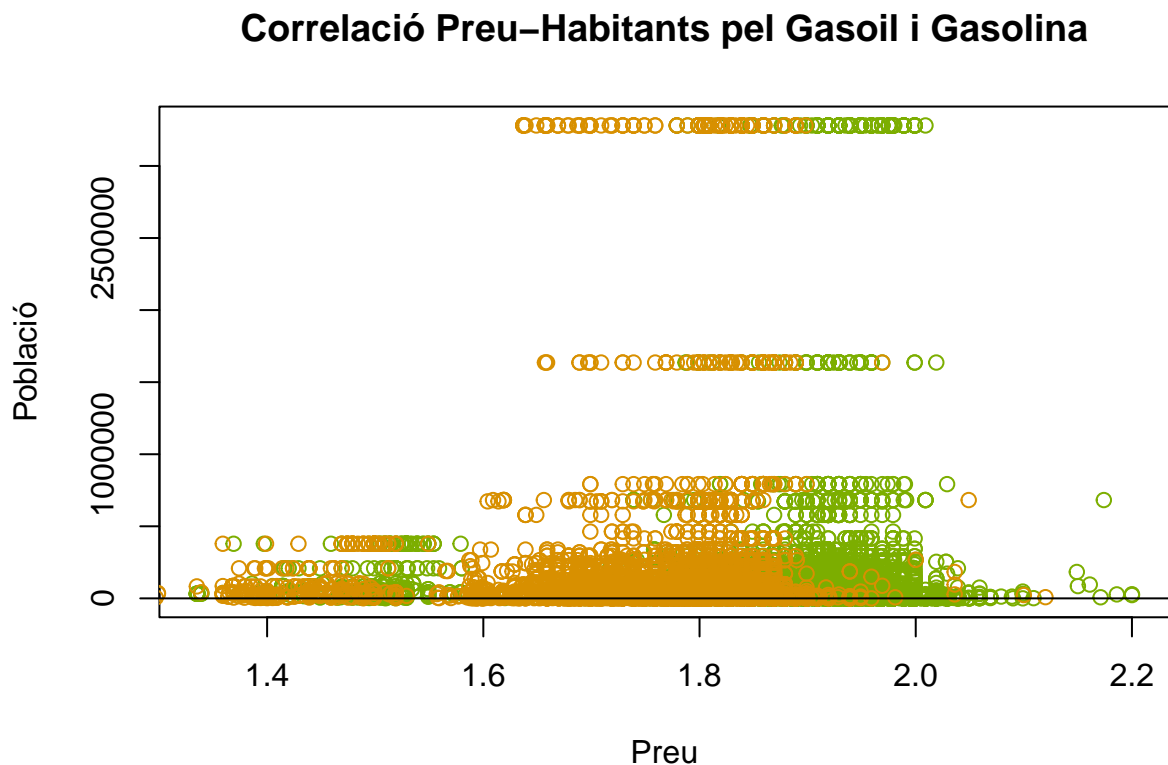
```
## 95 percent confidence interval:
##  0.01744601 0.05781635
## sample estimates:
##          cor
## 0.03764654

cor.test(gasolina$Price, gasolina$Population)

##
## Pearson's product-moment correlation
##
## data: gasolina$Price and gasolina$Population
## t = 5.5132, df = 9079, p-value = 3.62e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.03724077 0.07823954
## sample estimates:
##          cor
## 0.05776451
```

El resultat d'aquest test de correlació mostra que no hi ha correlació entre el preu de la gasolina i el nombre d'habitants del municipi.

```
plot(subset(gasoil, select = c("Price", "Population"))$Price, subset(gasoil, select = c("Price", "Population"))$Population,
points(subset(gasolina, select = c("Price", "Population"))$Price, subset(gasolina, select = c("Price", "Population"))$Population),
abline(lm(gasoil$Price ~ gasoil$Population))
title("Correlació Preu-Habitants pel Gasoil i Gasolina")
```



Per últim analitzarem si el nombre de gasolineres que hi ha al municipi influeix al preu del carburant. El nostre supòsit inicial és que el nombre de gasolineres ha de tenir cert impacte sobre el preu.

```
mean_gasoil_city <- aggregate(gasoil$Price, by=list(gasoil$City), mean)
gasoil_frequency <- table(gasoil$City)
gasoil_frequency <- as.data.frame(gasoil_frequency)

mean_gasolina_city <- aggregate(gasolina$Price, by=list(gasolina$City), mean)
gasolina_frequency <- table(gasolina$City)
gasolina_frequency <- as.data.frame(gasolina_frequency)
```

Un cop creats els dataframes que contenen el preu mitjà dels dos tipus de carburants i la freqüència, és a dir, el nombre de gasolineres, procedim a analitzar si existeix algun tipus de relació.

```
result <- merge(mean_gasoil_city, gasoil_frequency, by.x = "Group.1", by.y = "Var1")
result <- result %>% rename(Municipio = Group.1, Meanp = x)

rg <- lm(Meanp ~ Freq, data = result)
summary(rg)
```

```
##
## Call:
## lm(formula = Meanp ~ Freq, data = result)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51284 -0.02284  0.01563  0.04806  0.21548
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.9027345   0.0016354 1163.49  < 2e-16 ***
## Freq        -0.0008991   0.0002621   -3.43  0.000609 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09254 on 3963 degrees of freedom
## Multiple R-squared:  0.00296,    Adjusted R-squared:  0.002709
## F-statistic: 11.77 on 1 and 3963 DF,  p-value: 0.0006092

cor(result$Meanp, result$Freq)
```

```
## [1] -0.05440874

resultg <- merge(mean_gasolina_city, gasolina_frequency, by.x = "Group.1", by.y = "Var1")
resultg <- resultg %>% rename(Municipio = Group.1, Meanp = x)

lmg <- lm(Meanp ~ Freq, data = resultg)
summary(lmg)
```

```
##
## Call:
## lm(formula = Meanp ~ Freq, data = resultg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47927 -0.01927  0.01211  0.04173  0.22173
##
```

```
## Coefficients:
##           Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  1.7790253  0.0014867 1196.607 < 2e-16 ***
## Freq        -0.0007540  0.0002395   -3.148  0.00165 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08345 on 3881 degrees of freedom
## Multiple R-squared:  0.002548, Adjusted R-squared:  0.002291
## F-statistic: 9.912 on 1 and 3881 DF, p-value: 0.001654
```

```
cor(result$Meanp, result$Freq)
```

```
## [1] -0.05440874
```

En els dos casos el p-value és menor que el nivell de significació i, per tant, podem dir que el nombre de gasolineres té cert impacte en el preu dels dos carburants. D'altra banda, la correlació ens indica que hi ha una relació dèbil entre ambdues variables, és a dir, que a mesura que augmenta el nombre de gasolineres, el preu disminueix lleugerament.

```
pbrand <- lm(Price ~ Brand.factor, data = total)
summary(pbrand)
```

```
##
## Call:
## lm(formula = Price ~ Brand.factor, data = total)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25575 -0.06350 -0.00350  0.06123  0.38725
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.8757343   0.0043786  428.387 < 2e-16 ***
## Brand.factorBALLENOIL -0.1104431   0.0060677 -18.202 < 2e-16 ***
## Brand.factorBP         0.0067608   0.0049010   1.379 0.167762
## Brand.factorCAMPSA     -0.0270574   0.0072186  -3.748 0.000179 ***
## Brand.factorCARREFOUR -0.0042875   0.0064767  -0.662 0.507983
## Brand.factorCEPSA      -0.0001754   0.0046516  -0.038 0.969921
## Brand.factorGALP       -0.0179632   0.0050347  -3.568 0.000361 ***
## Brand.factorOTROS      -0.0629893   0.0044731 -14.082 < 2e-16 ***
## Brand.factorPETRONOR    0.0009142   0.0061256   0.149 0.881357
## Brand.factorREPSOL     -0.0056677   0.0045109  -1.256 0.208971
## Brand.factorSHELL      -0.0283818   0.0052898  -5.365 8.17e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08014 on 20351 degrees of freedom
## Multiple R-squared:  0.127, Adjusted R-squared:  0.1265
## F-statistic: 296 on 10 and 20351 DF, p-value: < 2.2e-16
```

5. Resolució del problema