

PRACTICA - 2

FuelData

Martí Antentas Parés

Xavier Vizcaino Gascon

13 de enero, 2023

Contents

1. Descripció del dataset	1
2. Integració i selecció	2
3. Neteja de dades	5
4. Anàlisi de dades	11
5. Resolució del problema	19
6. Llicència	19
7. Codi	19
8. Vídeo	19
9. Contribucions	20

1. Descripció del dataset

Per a la realització d'aquesta segona pràctica s'utilitza el *dataset* generat a la primera pràctica usant tècniques de *web scraping*. Aquest es combina amb un altre *dataset* amb l'objectiu de realitzar un anàlisi més profund i enriquidor.

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.

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

Les operacions d'aquesta pràctica han de permetre donar resposta a les següents preguntes:

- Les dades d'estudi es poden aplicar a tota la geografia espanyola?
- Hi ha zones amb preus marcadament diferents de la resta?
- La mitjana dels preus dels combustibles són diferents entre Barcelona i Madrid?
- La mitjana dels preus dels combustibles és diferent en les ciutats petites, mitjanes i grans?
- Existeix correlació entre els preus del combustible i el nombre d'habitants d'un municipi?
- El nombre de benzineres en un municipi influeix en els preus del combustible?

El procés s'inicia amb la lectura de l'arxiu de dades; que es realitza amb les següents opcions: escollint el separador i el tipus de codificació. A continuació es fa un primer sumari de les dades.

```
#Importar arxiu
fueldata <- read.csv(file.path(datadir, "FuelScraper", "dataset.csv"),
                     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
##
##
##
```

Aquest primer anàlisi indica que pot ser interessant canviar algunes variables a tipus factor, així com canviar el format de les variables temporals a tipus *date* i *lubridate*.

```
#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 còpies de *backup* per a les variables *Province* i *City*. Aquestes posteriorment s'hauran de modificar a través de processos de normalització de noms per tal de maximitzar la compatibilitat de les dades amb les dels altres *datasets* a 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 a través de l'anàlisi, es llegeix un arxiu addicional que conté el cens de població per municipis. Aquesta informació s'extreu del

web de l'Institut Nacional d'Estadística (INE). En aquest cas, la lectura es realitza amb les mateixes opcions que el *dataset* original.

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

A partir d'aquest moment s'executen un seguit d'operacions d'adaptació, principalment en les variables *Province* i *City* per tal de maximitzar la validesa del resultat de la integració. Així doncs, es canvien els noms de les variables i es transformen 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 els valors en les variables *Province* i *City* dels dos *datasets*, aquesta operació té com a objectiu eliminar 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* **fueldata** i **pobdata**.

```
#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. Inicialment, en el dataset **fueldata** les províncies o municipis amb articles tenen una estructura del tipus: “nom_municipi (article)”, mentre que en el dataset **pobdata** l'estructura d'aquests es del tipus “nom_municipi, article”. Així doncs es realitzen canvis en el primer per a fer-lo compatible amb el segon.

```
#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)
```

També es realitzen tot un seguit de canvis individuals (que no es detallen en la memòria, però si en el codi). Aquests canvis individuals tenen com a objectius maximitzar la informació disponible en el *dataset* resultant.

Després de tots els canvis; 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`), així les dades que no tenen una parella en l'altre *dataset* es mantenen en el *dataset* resultant afegint *NA* en la informació no disponible.

```
#Combinació de datasets
```

```
total<-merge(fuelldata, 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 reduint els seus nivells possibles a 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
CARREFOUR	2635
BALLENOIL	2189
CAMPSA	1511

S'observa que en alguns registres, la marca apareix com a *substring* en la variable *Brand*, per tant, un cop obtingudes les 10 marques amb més representació s'itera sobre tots els registres per tal de normalitzar el camp marca. També, es defineix com a “OTROS” la variable *Brand* de tots aquells registres on la seva marca no és una 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 d'aquest apartat, es realitza una selecció de dades limitada a les següents característiques:

- dades del dia 16 de Novembre
- dades dels 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* integrat on el camp *Capture_date* sigui *NA*. Aquests seran registres que no han estat capturats en la fase de *web scraping* i per tant seran municipis que apareixen en el cens de població, però no tenen estació de servei. Han aparegut en el *dataset* quan s'ha realitzat l'operació de combinació plena o *FULL JOIN* en el pas anterior.

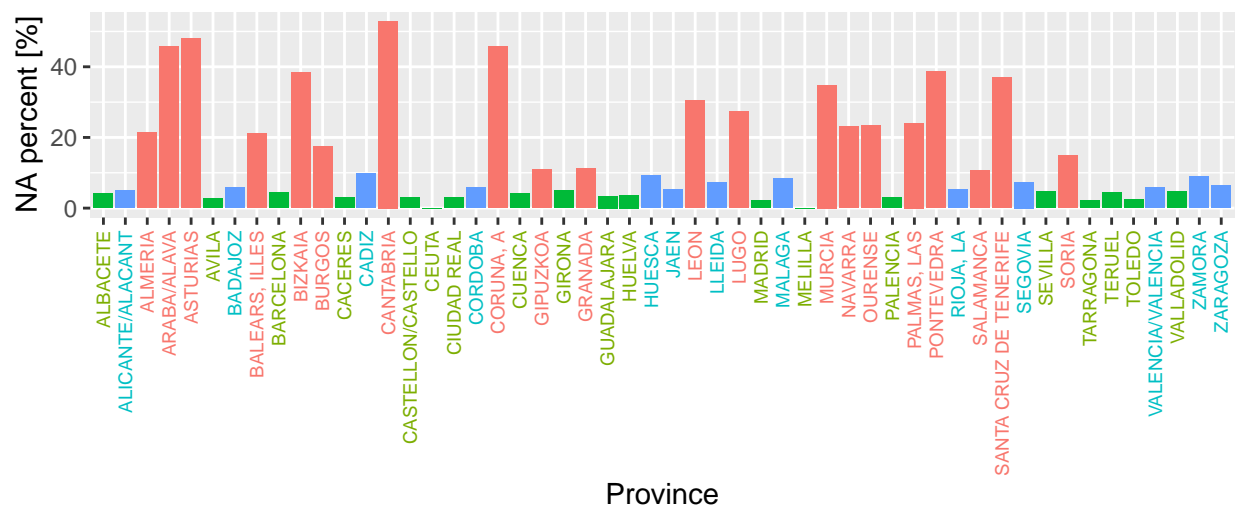
```
#Neteja de NAs
data<-data[!is.na(data$Capture_date),]
```

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.

Amb l'objectiu d'identificar les províncies amb un percentatge de registres on no es coneix el cens, es procedeix a realitzar aquest mateix estudi per a cada província. El resultat d'aquest estudi s'observa en el següent gràfic on s'identifiquen:

- 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%

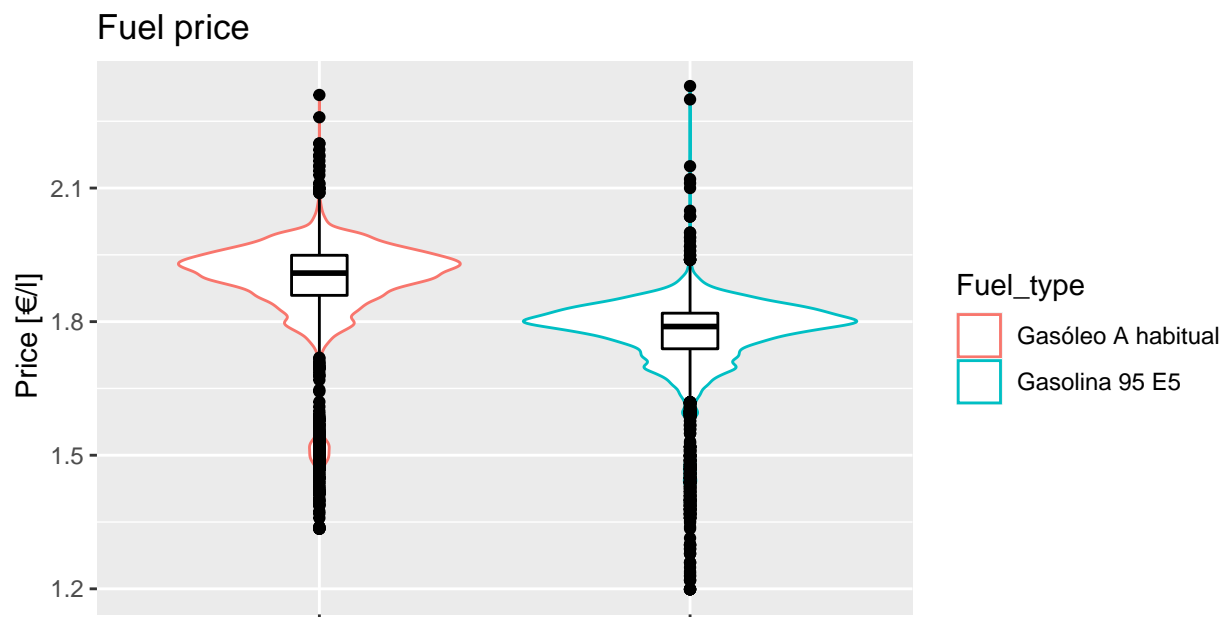


En vista dels resultats, per a estudis que no considerin dades de cens per municipis, es pot utilitzar el *dataset* complet ja que aquest prové de l'operació de *scraping* i no conté zeros o valors nuls. Tanmateix, quan l'objectiu de l'estudi requereixi considerar informació del cens, es recomana limitar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

```
#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 visualment els valors extrems de la variable *price* es genera un *violin plot* amb un *boxplot* al interior per cada tipus de combustible. Aquesta combinació de gràfics permet; per una banda, analitzar els *outliers* a través de la visualització *boxplot*, i per altra, conèixer la distribució de la població a través de la visualització *violin*. Així doncs permet, d'un cop de vista, veure si la informació que 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, pels dos productes.

```

#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 flexibilitzar la condició, es considerarà com a límit el 80% de la diferència de medianes. Per tant, es seleccionen els registres on la diferència de preu entre els dos combustibles sigui inferior a **0.1€**/l per considerar que els seus valors no són prou consistents i que per tant poden contenir errades.

Els registres seleccionats anteriorment s'eliminen de l'estudi per a ser *outliers* i no mostrar prou consistència.

```

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

#Extreure dades inconsistents
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 (diferents dates), considerant la consistència de registres d'una mateixa marca en una zona pròxima, ampliar l'estudi de consistència a tots els productes de l'estació de servei...

Valors extrems inferiors

Observant el *violin plot* anterior 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, es seleccionen els valors d'aquest combustible i en aquest rang per estudiar-los amb més profunditat, obtenint el nombre de registres d'aquestes característiques per província.

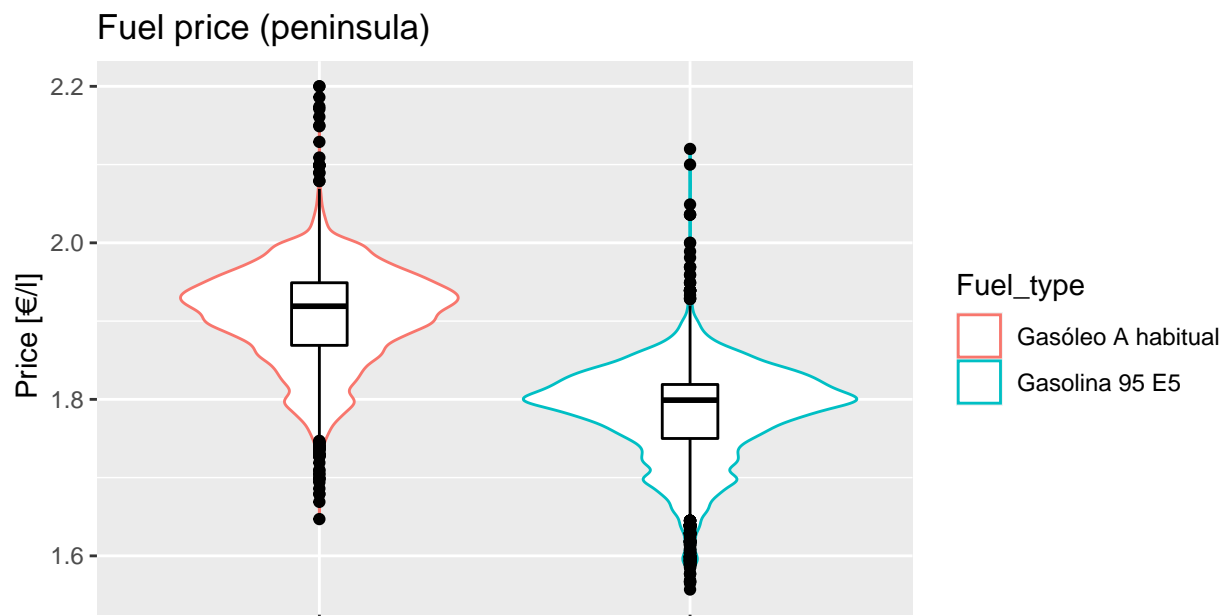
```
#Selecció de registres en la zona objectiu i summarització
DiesConc<-data[data$Fuel_type=="Gasóleo A habitual" &
               1.55>data$Price &
               data$Price>1.45,] %>%
  group_by(Province) %>%
  summarise(n=n()) %>%
  as.data.frame()
kable(DiesConc)
```

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

La taula anterior mostra 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, doncs la diferencia d'impostos, comporta variacions significatives en el preu dels combustibles.

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

Un cop extretes les dades de les ciutats autònomes de Ceuta i Melilla, així com les dades de les províncies de les illes canàries, es grafiquen les dades peninsulars on s'observa una clara reducció en el nombre de valors extrems.



Per tal de seleccionar els valors extrems inferiors, es procedeix d'acord al mètode 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()

```

L'existència d'estacions de servei *low cost* pot explicar l'aparició de valors extrems inferiors, coherents, en la variable *price*. Així doncs, es interessant saber si els registres seleccionats provenen d'estacions de servei d'una marca entre les 10 més representatives o formen part del grup "OTROS" que inclou aquest tipus d'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

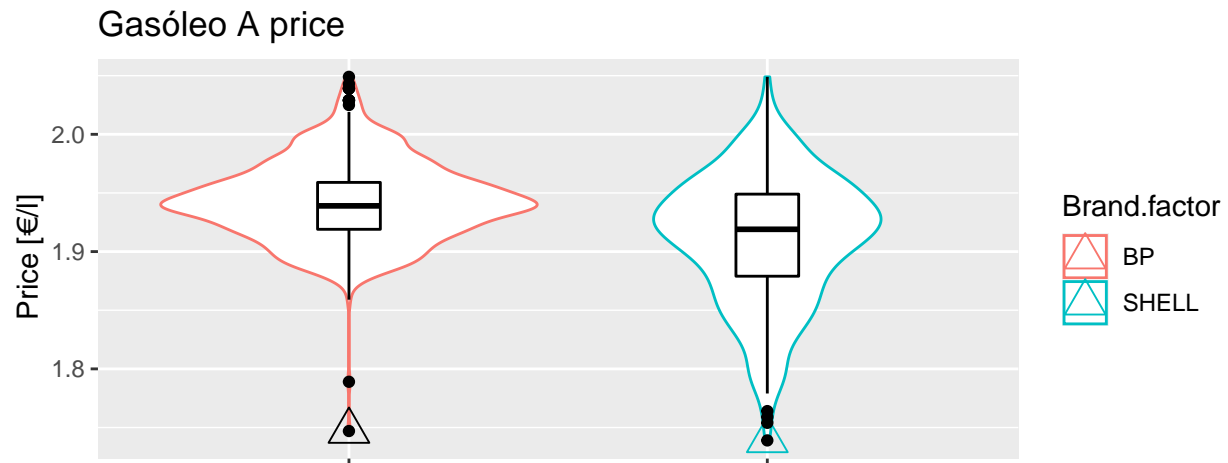
En vista dels resultats, interessa analitzar amb més detall els registres de valors extrems inferiors obtinguts per estacions de servei *no lowcost*, per tractar-se de registres candidats a erronis.

```

#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"),]

```

Graficant la distribució de la variable *price*, es considera el valor extrem de la marca **BP** (marcat amb un triangle negre en el gràfic) poc consistent, per altra banda, es considera que no hi ha prou evidència per decidir si el valor extrem de la marca **SHELL** és erroni i per tant es manté (marcat amb un triangle del mateix color que el gràfic).

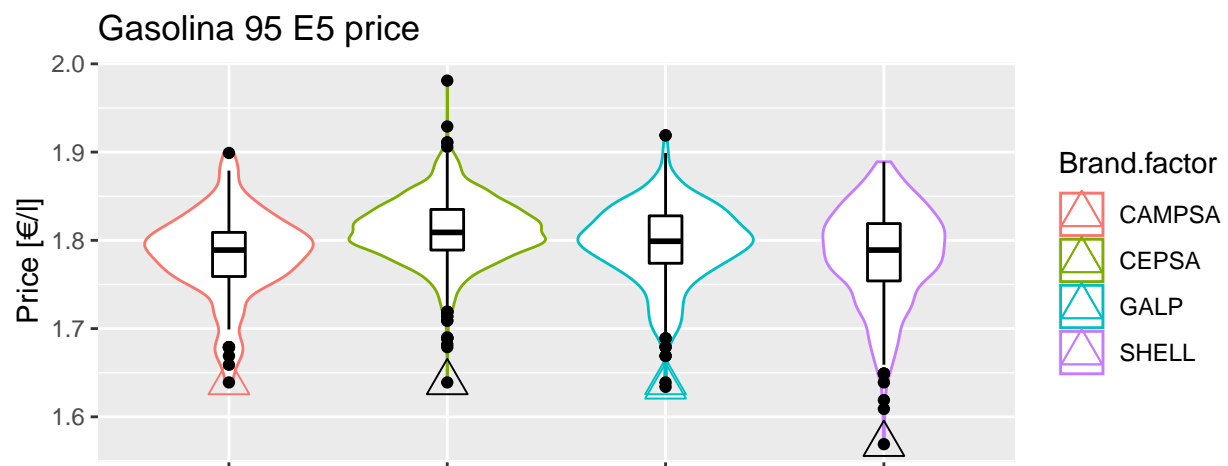


El registre anteriorment esmentat es guarda per a una posterior extracció del *dataset*.

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

Finalitzat l'anàlisi per al combustible *Gasóleo A habitual*, es continua amb l'estudi pel combustible *Gasolina 95 E5*.

```
#Anàlisis pel combustible Gasolina 95 E5
G_1<-GasLow[GasLow$Brand.factor=="CAMPESA",]
G_2<-GasLow[GasLow$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$Brand.factor=="CAMPESA" |
   data.peninsula$Brand.factor=="CEPSA" |
   data.peninsula$Brand.factor=="GALP" |
   data.peninsula$Brand.factor=="SHELL"),]
```



Analitzant els resultats, visualment es consideren valors poc consistents:

- El valor extrem inferior per la marca **CEPSA** (marcat amb un triangle negre).
- El valor extrem mínim per la marca **SHELL** (marcat amb un triangle negre).

Conseqüentment es seleccionen els valors poc consistents i es procedeix a la seva eliminació.

```
#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"))
```

Finalment es genera l'arxiu *clean_dataset.csv* que conté les dades finals, després dels processos d'integració, selecció i neteja de dades.

```
#Exportar a csv
exp_folder<-file.path(datadir, "FuelData")
if (!dir.exists(exp_folder)){
  dir.create(exp_folder)
}
exp_file<-file.path(exp_folder, "clean_dataset.csv")
write.csv2(data.peninsula, file = exp_file, row.names = TRUE)
```

4. Anàlisi de dades

Selecció dels grups de dades

L'anàlisi que durem a terme es divideix en 4 punts. En primer lloc estudiarem si hi ha diferències en els preus dels carburants entre Barcelona i Madrid. En segon lloc analitzarem si el preu es veu influenciat pel nombre d'habitants del municipi. A continuació veurem si el nombre d'estacions de servei per municipi influeixen en el preu i, per últim, analitzarem si la marca influeix en el preu.

```
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
```

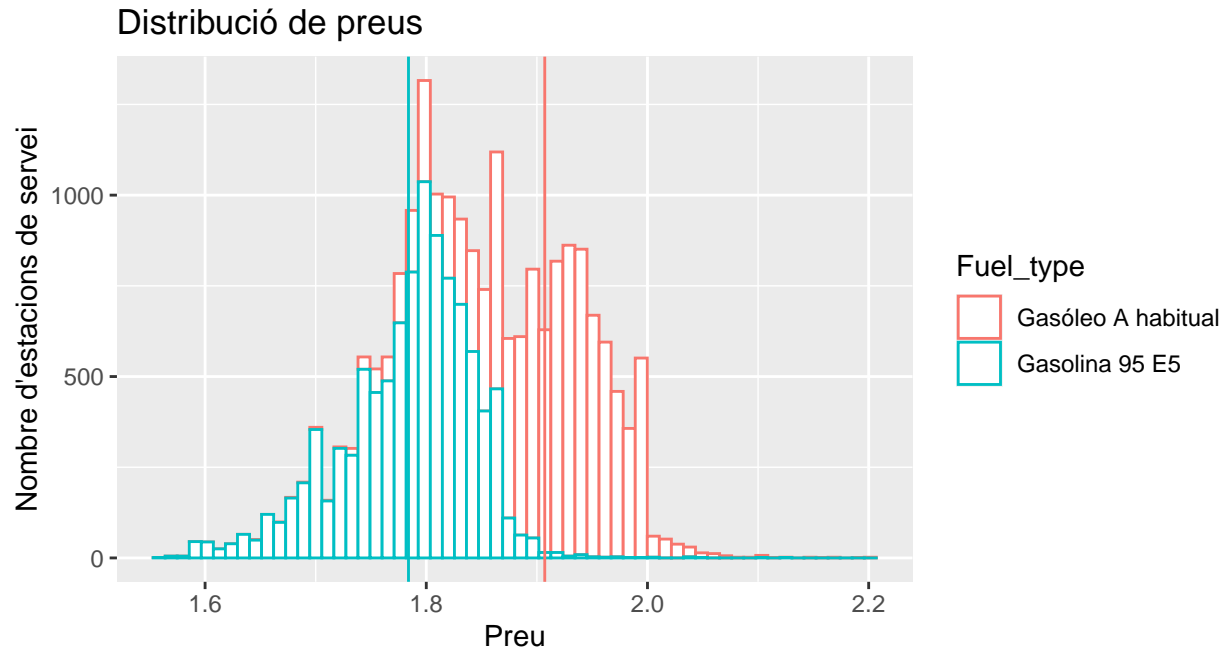
```
summary(total[total$Fuel_type=="Gasóleo A habitual", "Price"])
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  1.647   1.869   1.919   1.907   1.949   2.200
```

```
summary(total[total$Fuel_type=="Gasolina 95 E5", "Price"])
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  1.557   1.750   1.799   1.784   1.819   2.120
```

A continuació es presenta un histograma amb l'objectiu de conèixer la distribució de la variable *Price*.



Comprovació de la normalitat i homoscedasticitat

Tot i que el gràfic anterior ja ens introdueix la **no normalitat**, s'analitza a través de tests de normalitat si la variable *Price* segueix una distribució normal.

```
#Anderson-Darling tests
```

```
ad.test(total$Price)
```

```
##
```

```
## Anderson-Darling normality test
```

```
##
```

```
## data: total$Price
```

```
## A = 76.106, p-value < 2.2e-16
```

```
ad.test(total[total$Fuel_type=="Gasóleo A habitual", "Price"])
```

```
##
```

```
## Anderson-Darling normality test
```

```
##
```

```
## data: total[total$Fuel_type == "Gasóleo A habitual", "Price"]
```

```
## A = 77.715, p-value < 2.2e-16
```

```
ad.test(total[total$Fuel_type=="Gasolina 95 E5", "Price"])
```

```
##
```

```
## Anderson-Darling normality test
```

```
##
```

```
## data: total[total$Fuel_type == "Gasolina 95 E5", "Price"]
```

```
## A = 139.99, p-value < 2.2e-16
```

Valorant el p-valor obtingut, aquest és molt inferior al nivell de significació (0.05) i per tant es concloure que les dades **no** segueixen una distribució normal. Per tant, es crea una nova variable amb els nom *pnorm* amb les dades modificades segons la transformació de BoxCox i s'executa novament el test de normalitat.

```
#Transformacions de BoxCox
```

```
total$pnorm<-BoxCox(total$Price, lambda = BoxCoxLambda(total$Price))
```

```
ad.test(total$pnorm)
```

```
##
```

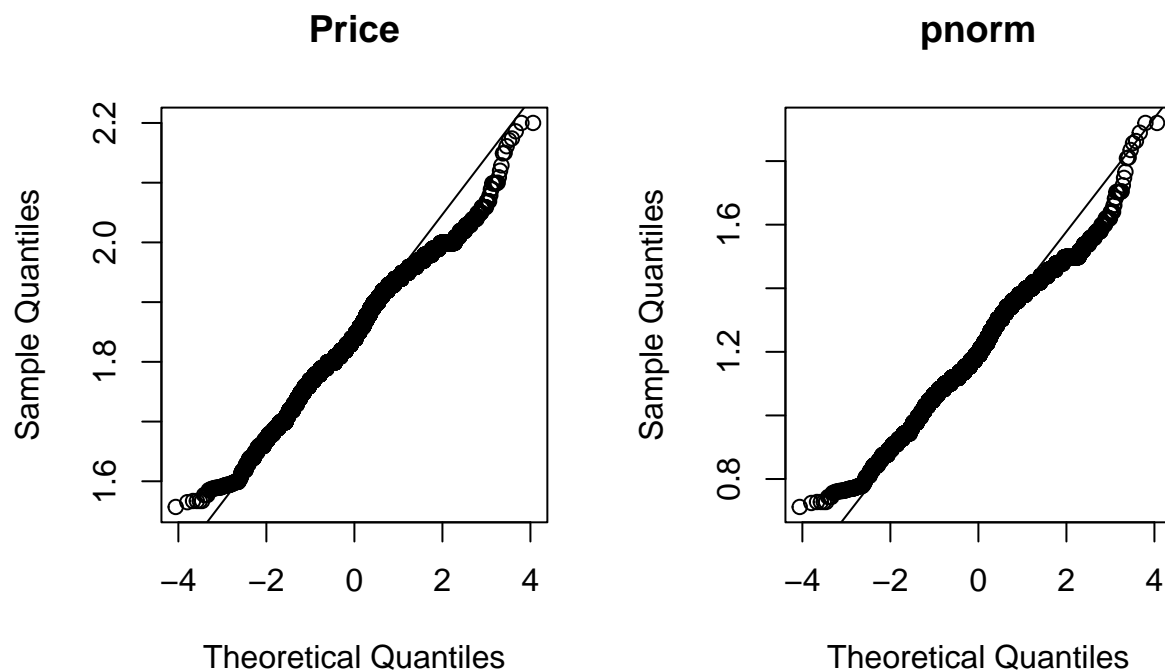
```
## Anderson-Darling normality test
```

```
##
```

```
## data: total$pnorm
```

```
## A = 77.35, p-value < 2.2e-16
```

Donat que les dades de la variable *Price* no segueixen una distribució normal i que la transformació Box-Cox, tampoc en millora la normalitat serà necessari considerar les versions no-paramètriques dels tests a realitzar.



Aplicació de proves estadístiques per comparar els grups de dades

En primer lloc, es realitza un test per comprovar la diferència de mitjanes entre les ciutats de Barcelona i Madrid. Filtrant les dades serà possible analitzar si hi ha diferències entre les ciutats per cada tipus de carburant.

```
#Anàlisis pel carburant Gasóleo A habitual
```

```
capital_filter <- total %>%
```

```
  filter(Fuel_type == "Gasóleo A habitual" & (City == "BARCELONA" | City == "MADRID"))
```

```
wilcox.test(Price ~ City, data = capital_filter, na.rm=TRUE, paired=FALSE,  
            exact=FALSE, conf.int=TRUE)
```

```
##
```

```
## Wilcoxon rank sum test with continuity correction
##
## data: Price by City
## W = 6559, p-value = 0.0001852
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
## -0.029981901 -0.009928359
## sample estimates:
## difference in location
## -0.0199688
```

```
#Anàlisi pel carburant Gasolina 95 E5
```

```
capital_filter2 <- total %>%
  filter(Fuel_type == "Gasolina 95 E5" & (City == "BARCELONA" | City == "MADRID"))
wilcox.test(Price ~ City, data = capital_filter2, na.rm=TRUE, paired=FALSE,
            exact=FALSE, conf.int=TRUE)
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: Price by City
## W = 11690, p-value = 9.983e-05
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
## 0.01003171 0.03002350
## sample estimates:
## difference in location
## 0.02004017
```

En ambdós tests el p-value obtingut és inferior al nivell de significació, així doncs es pot concloure que les diferències de preu entre les dues capitals són significatives pels dos combustibles.

A continuació s'estudia si hi ha diferències significatives entre la mitjana de preus dels dos combustibles en funció del nombre d'habitants del municipi. En aquest cas com que es consideren les dades del cens per a definir la dimensió del municipi, limitarem l'estudi a les províncies amb un baix percentatge de NAs.

```
#Preparació de les dades
```

```
pop_filter <- total %>%
  filter(Province %in% green_province$Province | Province %in% blue_province$Province) %>%
  filter(Fuel_type == "Gasóleo A habitual") %>%
  mutate(Pop_size = ifelse(Population < 10000, "S",
                           ifelse(Population > 100000, "L", "M")))
```

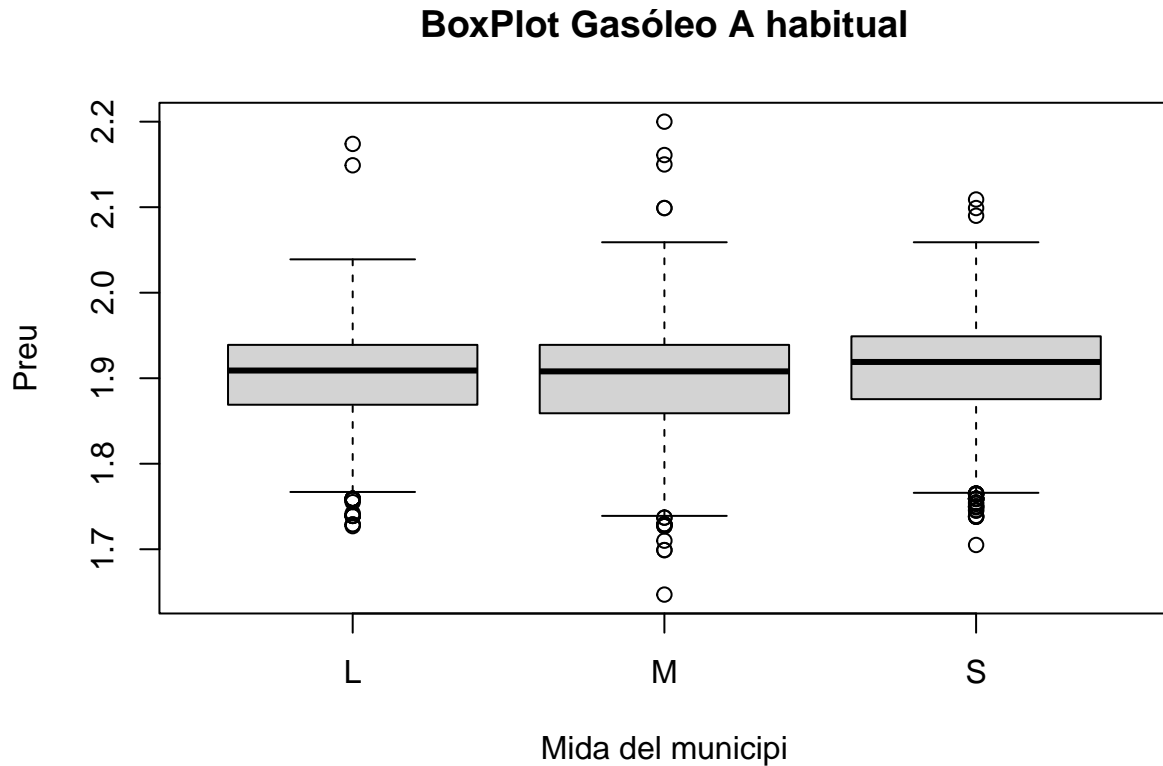
```
#Test
```

```
kruskal.test(Price ~ Pop_size, data = pop_filter)
```

```
##
## Kruskal-Wallis rank sum test
##
## data: Price by Pop_size
## Kruskal-Wallis chi-squared = 59.164, df = 2, p-value = 1.422e-13
```

El p-valor obtingut en el test estadístic és més petit que el nivell de significació (0,05), per tant es pot concloure que hi ha diferències significatives en el preu del combustible *Gasóleo A habitual* en funció del tractament (ciutat S, M o L)

A continuació es mostra el BoxPlot pel preu de *Gasóleo A habitual* en funció de la dimensió del municipi.



Es repeteix el mateix procediment, però pel cas de *Gasolina 95 E5*.

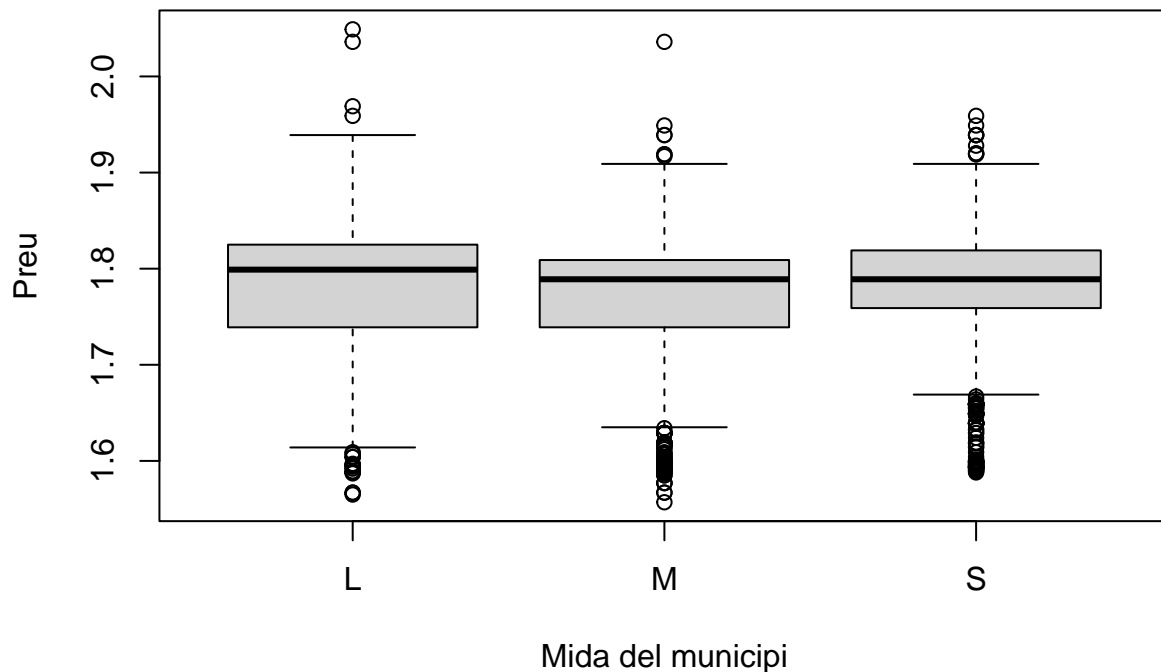
```
#Preparació de les dades
pop_filter_fuel <- total %>%
  filter(Province %in% green_province$Province | Province %in% blue_province$Province) %>%
  filter(Fuel_type == "Gasolina 95 E5") %>%
  mutate(Pop_size = ifelse(Population < 10000, "S",
                           ifelse(Population > 100000, "L", "M")))
```

```
#Test
kruskal.test(pnorm ~ Pop_size, data = pop_filter_fuel)
```

```
##
## Kruskal-Wallis rank sum test
##
## data: pnorm by Pop_size
## Kruskal-Wallis chi-squared = 51.74, df = 2, p-value = 5.819e-12
```

Novament, s'obté un p-valor inferior a 0,05 que permet concloure que la dimensió dels municipis té una influència significativa en el preu de la *Gasolina 95 E5*. Tanmateix, es mostra a continuació un Box-Plot que permet visualitzar gràficament aquestes diferències.

BoxPlot Gasóleo A habitual



Els resultats d'aquests anàlisi mostren que:

- De mitjana, el *Gasóleo A habitual* és més car que la *Gasolina 95 E5*.
- El *Gasóleo A habitual* és més barat a Barcelona que a Madrid, en canvi la *Gasolina 95 E5* es més barata a Madrid que a Barcelona.
- La dimensió del municipi (petit, mitjà o gran) té una influència estadísticament significativa en el preu dels dos combustibles en estudi.

A continuació es vol estudiar si els preus del combustible estan correlacionats linealment amb el nombre d'estacions de servei per municipi.

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

#Aggregats i freqüència
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)
```

A continuació s'analitza si existeix relació entre el preu mitja del combustible de cada municipi i el nombre d'estacions de servei en aquell mateix municipi a través de la creació de models lineals.


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

```
#Linear model pel combustible Gasóleo A habitual
```

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

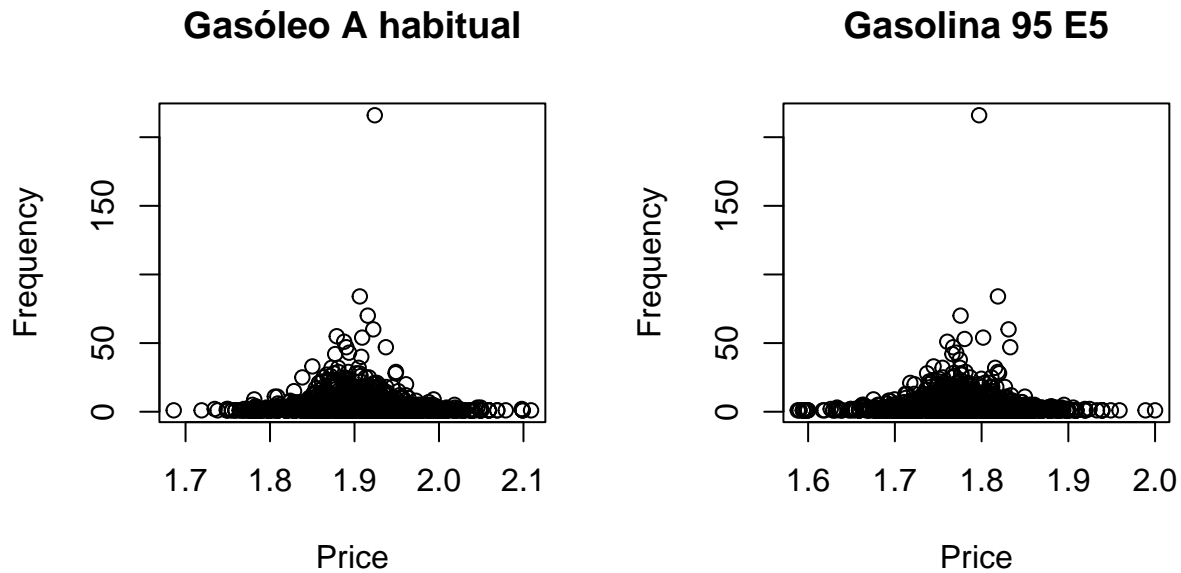
```
##
## Call:
## lm(formula = Meanp ~ Freq, data = result)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.231124 -0.032375  0.001876  0.033876  0.191876
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.9178732  0.0009450 2029.44 < 2e-16 ***
## Freq        -0.0007489  0.0001510   -4.96 7.36e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05243 on 3799 degrees of freedom
## Multiple R-squared:  0.006434, Adjusted R-squared:  0.006172
## F-statistic: 24.6 on 1 and 3799 DF, p-value: 7.365e-07
```

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

```
#Linear model pel combustible Gasolina 95 E5
```

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

```
##
## Call:
## lm(formula = Meanp ~ Freq, data = resultg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.204128 -0.027128  0.002483  0.031427  0.207872
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.7927391  0.0008805 2036.039 < 2e-16 ***
## Freq        -0.0006109  0.0001414   -4.322 1.59e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04844 on 3716 degrees of freedom
## Multiple R-squared:  0.005001, Adjusted R-squared:  0.004733
## F-statistic: 18.68 on 1 and 3716 DF, p-value: 1.589e-05
```



En ambdós casos s'observa que no existeix una relació lineal entre el preu del combustible i el nombre d'estacions de servei en el municipi ja que els coeficients *Adjusted R-squared* són respectivament 0.0061721 i 0.0047333

Per finalitzar, es vol estudiar si hi ha diferències estadísticament significatives en el preu dels combustibles en funció de la marca de l'estació de servei.

```
#Gasóleo A habitual
gasoilb <- total[total$Fuel_type == "Gasóleo A habitual",]
kruskal.test(Price ~ Brand.factor, data = gasoilb)

##
##  Kruskal-Wallis rank sum test
##
## data:  Price by Brand.factor
## Kruskal-Wallis chi-squared = 2553.8, df = 10, p-value < 2.2e-16
```

```
#Gasolina 95 E5
gasolinab <- total[total$Fuel_type == "Gasolina 95 E5",]
kruskal.test(Price ~ Brand.factor, data = gasolinab)

##
##  Kruskal-Wallis rank sum test
##
## data:  Price by Brand.factor
## Kruskal-Wallis chi-squared = 3053.8, df = 10, p-value < 2.2e-16
```

Com que en ambdós casos obtenim p-valors menors que el nivell de significació (0.05), es pot concloure que hi ha, com a mínim, un grup (marca d'estació de servei) estadísticament diferent dels altres grups en per cada combustible.

5. Resolució del problema

A través de l'obtenció de dades, la integració amb altres *datasets* i les operacions de selecció i de neteja; s'ha pogut donar resposta a les preguntes introduïdes a l'inici del document. Concretament es pot concloure que:

- Les dades d'estudi es poden aplicar a tota la geografia espanyola, però s'ha de tenir en compte que hi ha zones amb característiques impositives especials; que generen valors extrems. Tanmateix es important destacar que la integració dels *datasets* realitzada comporta algunes limitacions de compatibilitat. Aquestes limitacions fan que les dades en algunes províncies concretes continguin un nombre elevat (especialment en percentatge sobre el total) de NAs.
- Hi ha zones amb preus marcadament diferents de la resta, aquestes son les ciutats autònomes de Ceuta i Melilla així com les províncies de les illes canaries.
- La mitjana dels preus dels combustibles són estadísticament diferents entre Barcelona i Madrid. El *Gasóleo A habitual* és més barat a Barcelona que a Madrid, en canvi la *Gasolina 95 E5* es més barata a Madrid que a Barcelona
- La mitjana dels preus dels combustibles és estadísticament diferent en les ciutats petites, mitjanes i grans.
- No hi ha evidència d'una relació lineal que vinculi el nombre de benzineres en un municipi (*freq*) i el preu mitja (*meanp*) del combustible en el municipi.
- La mitjana dels preus dels combustibles és estadísticament diferent en diferents marques d'estacions de servei.

6. Llicència

El projecte es distribueix sota llicència CC BY-NC 4.0 (Creative Commons Reconocimiento-No Comercial). Aquesta llicència permet alterar i difondre l'obra original a condició que es faci referència a l'autor, i sempre amb finalitats no comercials.

7. Codi

El codi es pot trobar en el següent repositori de GitHub:


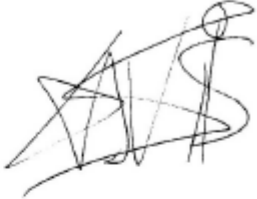

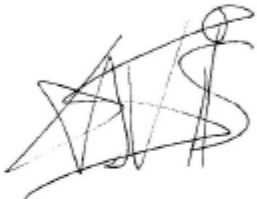

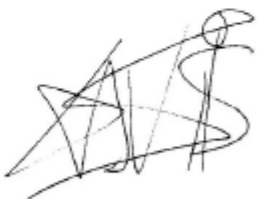
- FuelData

8. Vídeo

El vídeo amb l'explicació del desenvolupament es pot trobar a través del següent enllaç

- vídeo

9. Contribucions

Contribucions	Martí Antentas Paré	Xavier Vizcaino Gascon
Investigació prèvia		
Redacció de les respostes		
Desenvolupament del codi		
Participació en el vídeo	