Tipologia y Ciclo de Vida de los Datos, Practica 2

Imanol Miguez Quintela/Ivan Cuevas Ortin

Contents

1. Descripcion del dataset

En esta practica utilizaremos el dataset que obtuvimos en la Practica 1, "Precio de los carburantes por estación de servicio en España".

El dataset incluye datos de precios de los diferentes carburantes para cada una de las estaciones de servicios localizadas en España. También incluye información detallada sobre cada una de las gasolineras con el fin de que estas puedan ser localizadas fácilmente por el usuario final.

El dataset se puede encontrar en el repositorio de Github creado para la Practica 1 (https://github.com/icuevort/web_scraping_gas_prices). A continuacion cargarmos los datos desde Github.

provincia	latitud	longitud	precios	empresa
ALBACETE	39.21142	-1.539167	['SP-95 1,749', 'G-A 1,869', 'Gasóleo B 1,370']	Nº 10.935
ALBACETE	39.10039	-1.346083	['SP-95 1,909', 'SP-98 2,200', 'G-A 2,040', 'G-A+ 2,300']	REPSOL
ALBACETE	38.99822	-1.869889	['SP-95 1,769', 'G-A 1,839', 'G-A+ 1,889']	INPEALSA
ALBACETE	39.00964	-1.878111	['SP-95 1,736', 'SP-98 1,874', 'G-A 1,837', 'G-A+ 1,874']	ALCAMPO
ALBACETE	38.98217	-1.853306	['SP-95 1,769', 'G-A 1,839']	INPEALSA
ALBACETE	38.98925	-1.849028	['SP-95 1,799', 'SP-98 1,899', 'G-A 1,919', 'G-A+ 1,999']	CEPSA

Los datos se han cargado correctamente. Para explorar los datos en mas detalle, utilizamos ahora la funcion str.

```
## 'data.frame':    11909 obs. of  12 variables:
##  $ provincia    : chr  "ALBACETE" "ALBACETE" "ALBACETE" "ALBACETE" ...
##  $ localidad    : chr  "ABENGIBRE" "ALATOZ" "ALBACETE" "ALBACETE" ...
##  $ direccion    : chr  "AVENIDA CASTILLA LA MANCHA, 26" "CR CM-332, 46,4" "AVDA. DE LOS TOREROS, S/N" "SECTOR PARCELA T-3 LOCAL52, 14" ...
##  $ margen       : chr  "Derecho" "Izquierdo" "N" "N" ...
##  $ horario      : chr  "L-D: 07:00-22:00" "L-D: 7:00-23:00" "L-D: 07:00-23:00" "L-S: 06:00-23:00; D: 08:00-23:00" ...
##  $ latitud      : num  39.2 39.1 39 39 39 ...
##  $ longitud     : num  -1.54 -1.35 -1.87 -1.88 -1.85 ...
##  $ empresa      : chr  "Nº 10.935" "REPSOL" "INPEALSA" "ALCAMPO" ...
##  $ precios      : chr  "['SP-95 1,749', 'G-A 1,869', 'Gasóleo B 1,370']" "['SP-95 1,909', 'SP-98 2,200', 'G-A 2,040', 'G-A+ 2,300']" "['SP-95 1
##  $ fechaRevision: chr  "18/11/2022" "18/11/2022" "18/11/2022" "18/11/2022" ...
##  $ fechaUltima  : chr  "17/11/2022" "01/11/2022" "18/11/2022" "18/11/2022" ...
##  $ ventaPublico : chr  "Venta al Público" "Venta al Público" "Venta al Público" "Venta al Público" ...
```

Tenemos 12 variables, 10 de ellas son del tipo Character, y dos de ellas (latitud y longitud) son numericas.

2. Limpieza de datos

2.1 Restructuracion de la variable precios

La variable precios incluye ,para cada estacion de servicio, los combustibles disponibles, y su precio. Para que estos datos sean utliles durante nuestro analisis, queremos crear nuevas columnas, una por cada tipo de carburante, que nos den el precio para cada gasolinera.

El primer paso sera crear una nueva columna con el indice, y despues crear una nueva tabla con solo 2 variables: index y precios

A continuacion realizamos algunas tareas de limpieza sobre la variable precio, que nos ayudaran a separarla posteriormente.

Hay 9 tipos de carburantes, asi que separaremos la columna precios en 8 nuevas columnas. Cada una de ellas incluira el tipo de carburante y el precio

Seguidamente moveremos cada uno de las nuevas columnas a filas, manteniendo los valores de index y precio. De esta manera tendremos una fila por cada estacion de servicio y carburante disponible.

Como no todas las gasolineras disponen de todos los productos, muchas filas en la columna value estan vacias. A continuacion eliminamos dichas filas

Ahora separamos la columna value en dos nuevas columnas: tipo de carburante y precio.

Antes de continuar, tenemos que cambiar el tipo de la variable precio a numeric.

Creamos nuevas columnas, cada una para un tipo de carburante, y como valor el precio de cada combustible.

index	Biodiésel	G-A	G-A+	Gas_N._Compr	Gas_N._Licuado	Gasóleo_B	Gasóleo_C	GLP	SP-95	SP-98
1	NA	1.869	NA	NA	NA	1.37	NA	NA	1.749	NA
2	NA	2.040	2.300	NA	NA	NA	NA	NA	1.909	2.200
3	NA	1.839	1.889	NA	NA	NA	NA	NA	1.769	NA
4	NA	1.837	1.874	NA	NA	NA	NA	NA	1.736	1.874
5	NA	1.839	NA	NA	NA	NA	NA	NA	1.769	NA
6	NA	1.919	1.999	NA	NA	NA	NA	NA	1.799	1.899

Finalmente unimos la tabla inicial con esta ultima.

Como no necesitamos la columna de precios, la quitamos de la tabla.Tambien cambiaremos el nombre de alguna de las nuevas columnas, ya que los simbolos incluidos pueden darnos problemas posteriormente

provincia	latitud	longitud	empresa	Gasoleo_A
ALBACETE	39.21142	-1.539167	Nº 10.935	1.869
ALBACETE	39.10039	-1.346083	REPSOL	2.040
ALBACETE	38.99822	-1.869889	INPEALSA	1.839
ALBACETE	39.00964	-1.878111	ALCAMPO	1.837
ALBACETE	38.98217	-1.853306	INPEALSA	1.839
ALBACETE	38.98925	-1.849028	CEPSA	1.919

2.2 Datos estadisticos basicos

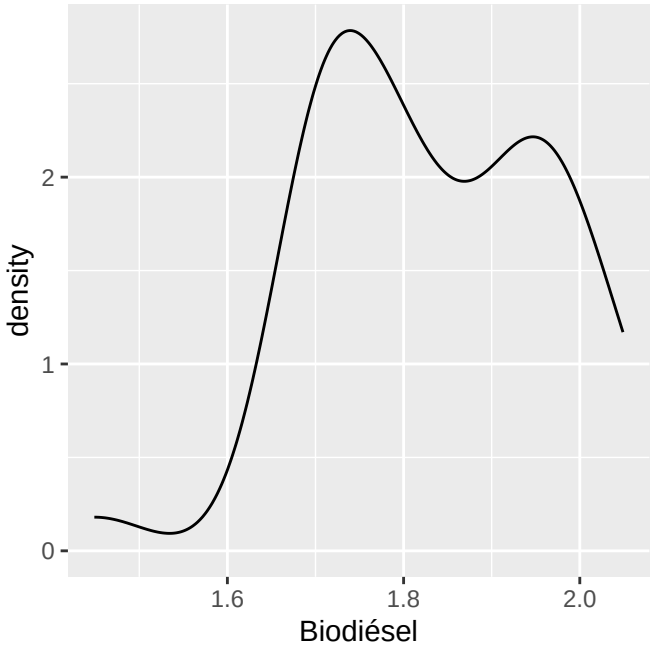
El siguiente paso sera la limpieza de cada uno de los datos. Veamos los datos estadisticos basicos.

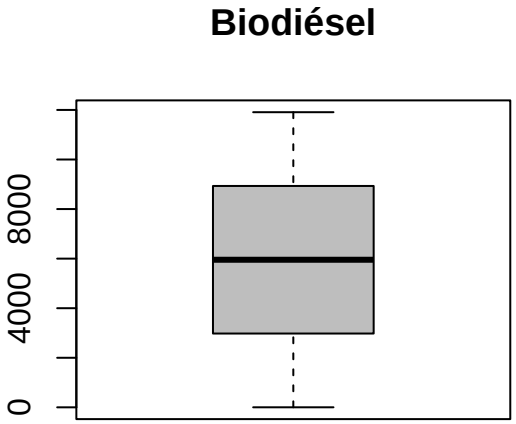
```
##      index        provincia          localidad          direccion
##  Min.   :    1   Length:11909       Length:11909       Length:11909
##  1st Qu.: 2978   Class :character   Class :character   Class :character
##  Median : 5955   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 5955
##  3rd Qu.: 8932
##  Max.   :11909
##
##     margen            horario             latitud          longitud
##  Length:11909       Length:11909       Min.   :-4.038   Min.   :-18.0119
##  Class :character   Class :character   1st Qu.:38.008   1st Qu.: -5.4449
##  Mode  :character   Mode  :character   Median :40.041   Median : -3.4731
##                                        Mean   :39.615   Mean   : -3.3195
##                                        3rd Qu.:41.681   3rd Qu.: -0.6822
##                                        Max.   :43.732   Max.   : 40.4712
##                                        NA's   :4        NA's   :4
##    empresa          fechaRevision      fechaUltima        ventaPublico
##  Length:11909       Length:11909       Length:11909       Length:11909
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    Biodiésel       Gasoleo_A     Gasoleo_A_Plus   Gas_N._Compr
##  Min.   :1.449   Min.   :1.315   Min.   :1.424   Min.   :1.299
##  1st Qu.:1.706   1st Qu.:1.829   1st Qu.:1.939   1st Qu.:1.379
##  Median :1.809   Median :1.889   Median :1.979   Median :1.468
##  Mean   :1.831   Mean   :1.865   Mean   :1.959   Mean   :1.522
##  3rd Qu.:1.948   3rd Qu.:1.919   3rd Qu.:2.009   3rd Qu.:1.580
##  Max.   :2.049   Max.   :2.309   Max.   :2.999   Max.   :3.700
##  NA's   :11871   NA's   :349     NA's   :4344    NA's   :11785
##  Gas_N._Licuado    Gasóleo_B       Gasóleo_C          GLP
##  Min.   :1.249   Min.   :1.009   Min.   :1.215   Min.   :0.639
##  1st Qu.:1.299   1st Qu.:1.438   1st Qu.:1.510   1st Qu.:1.019
##  Median :1.299   Median :1.500   Median :1.586   Median :1.057
##  Mean   :1.416   Mean   :1.505   Mean   :1.552   Mean   :1.047
##  3rd Qu.:1.390   3rd Qu.:1.569   3rd Qu.:1.615   3rd Qu.:1.089
##  Max.   :3.650   Max.   :1.949   Max.   :1.686   Max.   :1.260
##  NA's   :11829   NA's   :8904    NA's   :11886   NA's   :11081
##     Super_95        Super_98
##  Min.   :1.199   Min.   :1.330
##  1st Qu.:1.729   1st Qu.:1.891
##  Median :1.779   Median :1.929
##  Mean   :1.756   Mean   :1.908
##  3rd Qu.:1.809   3rd Qu.:1.959
##  Max.   :2.329   Max.   :2.999
##  NA's   :1165    NA's   :5927
```

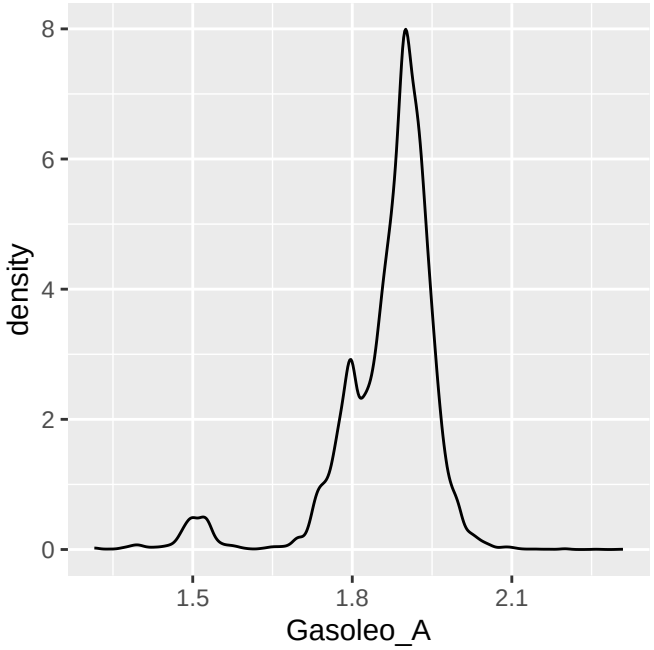
2.3 Variables relativas al precio

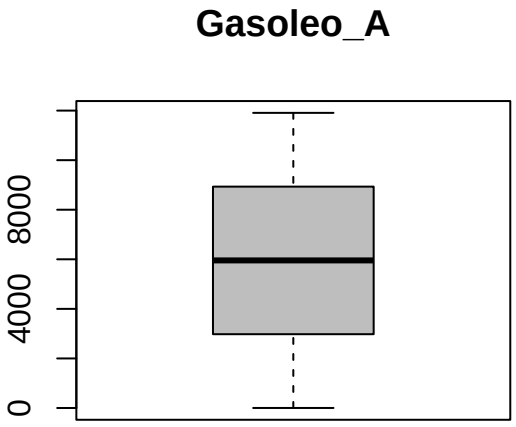
Veamos con mas detalle las variables relativas a los precios de los carburantes. Todas ellas contienen valores nulos. Esto se esperaba ya que no todos los productos estan disponibles en todas las estaciones de servicio.

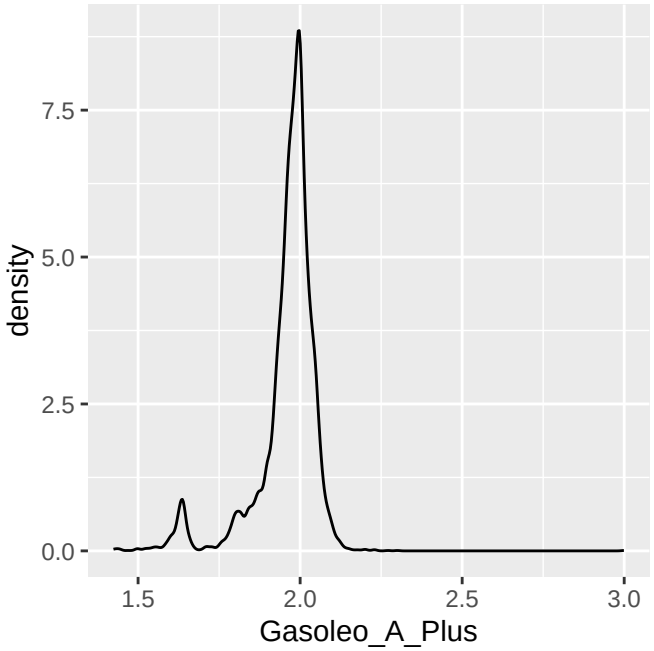
Para ver la distribucion de cada una, y encontrar posibles valores extremos, generamos un grafico de densidad y un diagrama de caja para cada una de ellas.

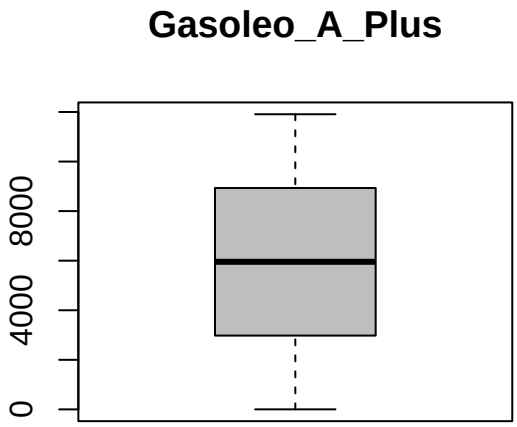


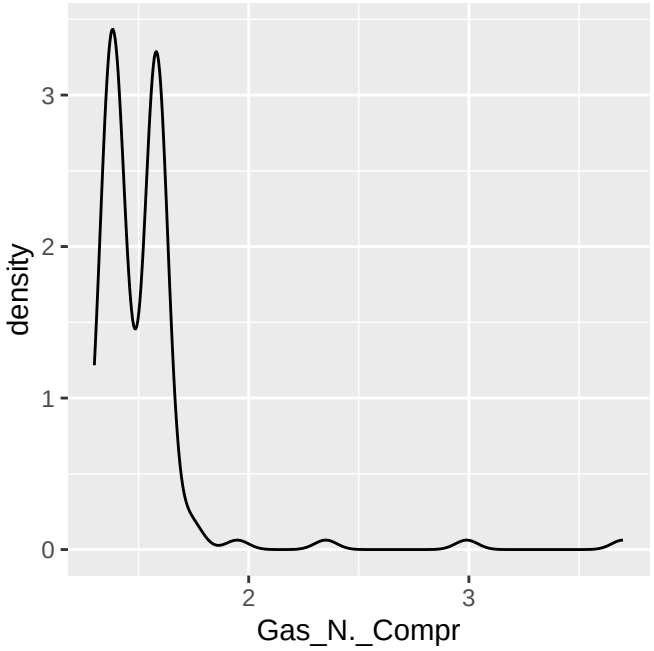


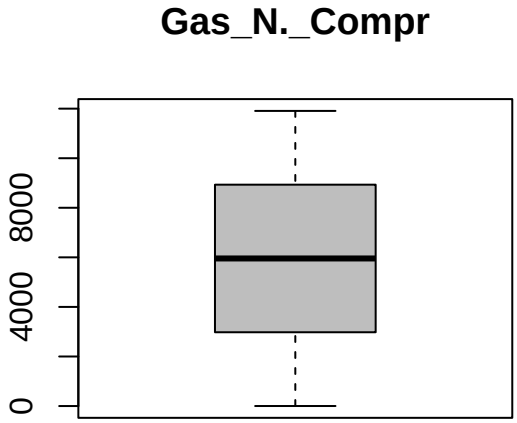


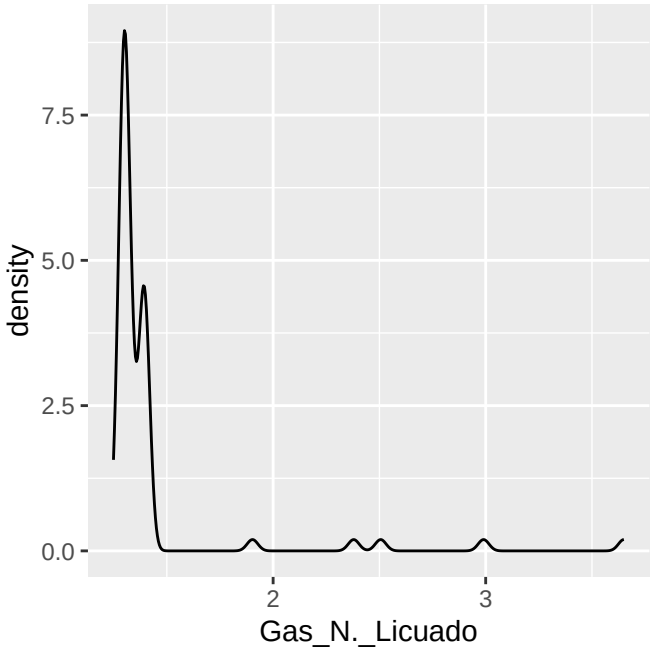


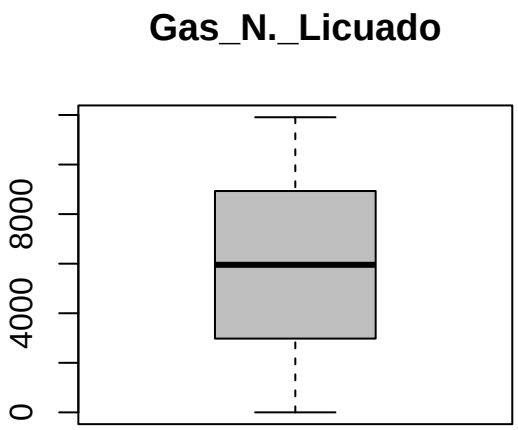

7.5
5.0
2.5
0.0
density
1.5
2.0
2.5
3.0
Gasoleo_A_Plus

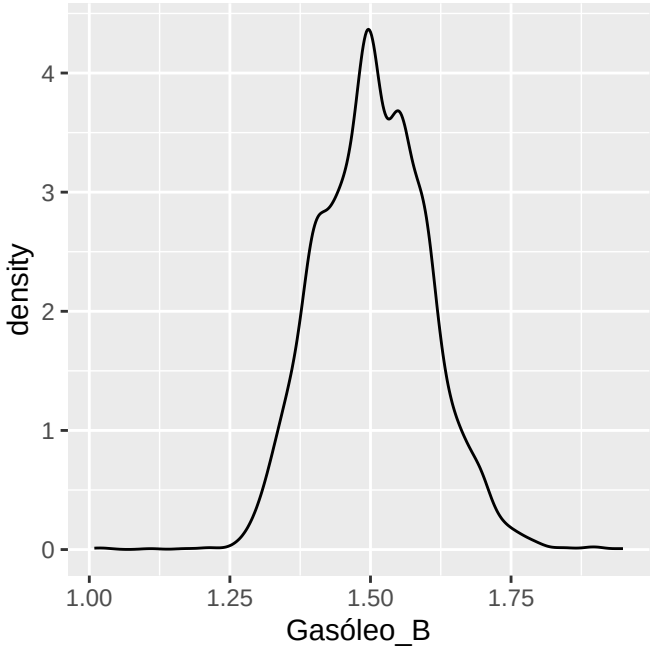

Gasoleo_A_Plus
0
4000
8000

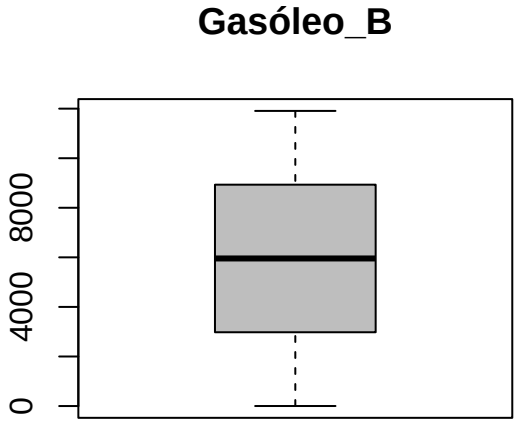

3
2
1
0
density
2
3
Gas_N._Compr

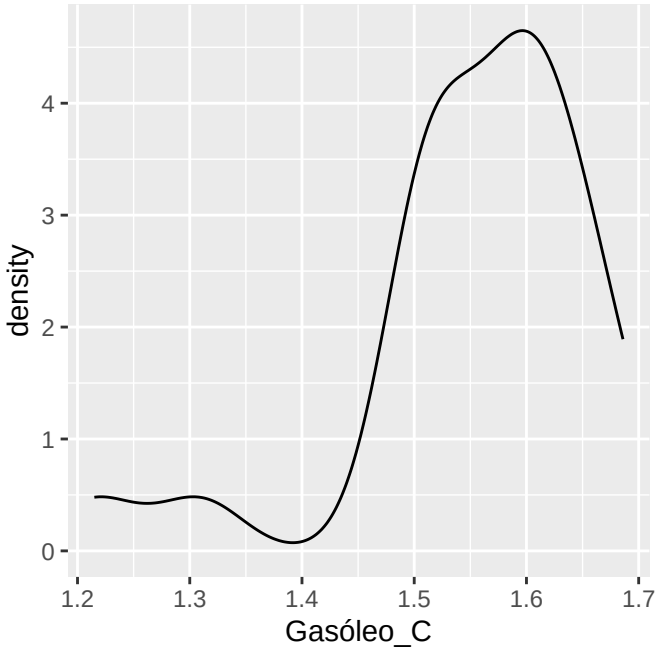

Gas_N._Compr
0
4000
8000

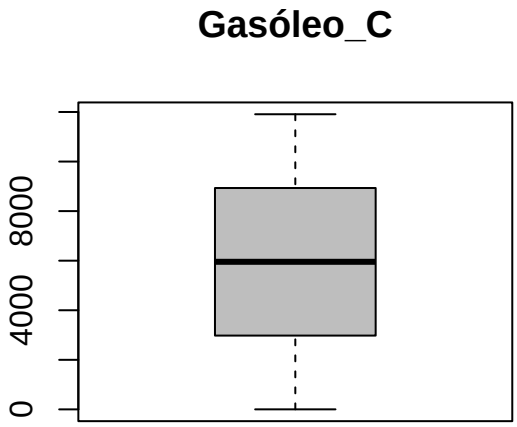

7.5
5.0
2.5
0.0
density
2
3
Gas_N._Licuado

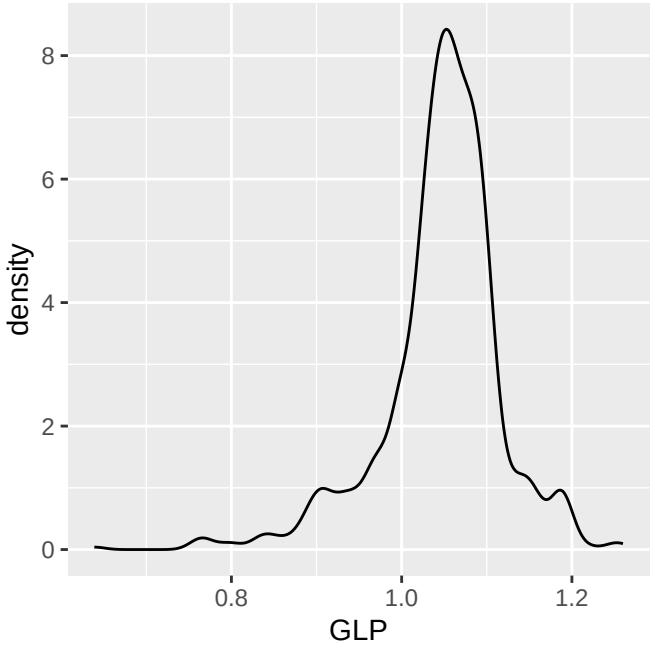

Gas_N._Licuado
0
4000
8000

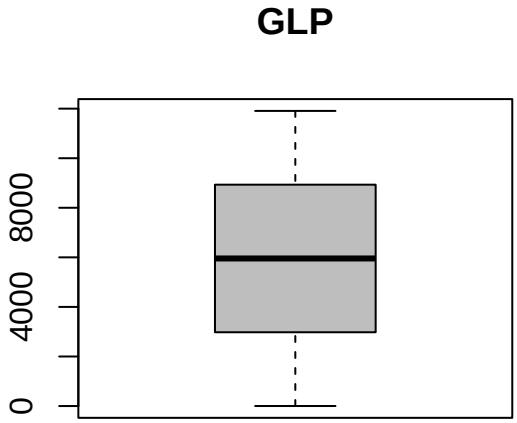

4
3
2
1
0
density
1.00
1.25
1.50
1.75
Gasóleo_B

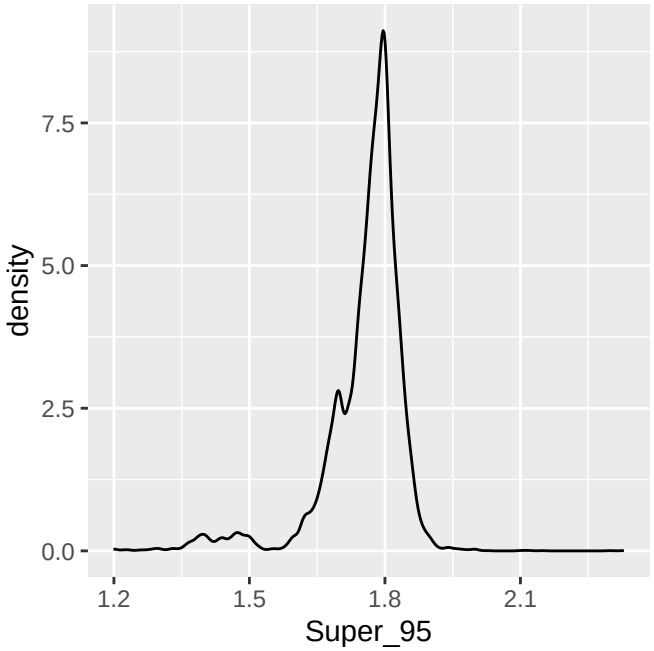

Gasóleo_B
0
4000
8000

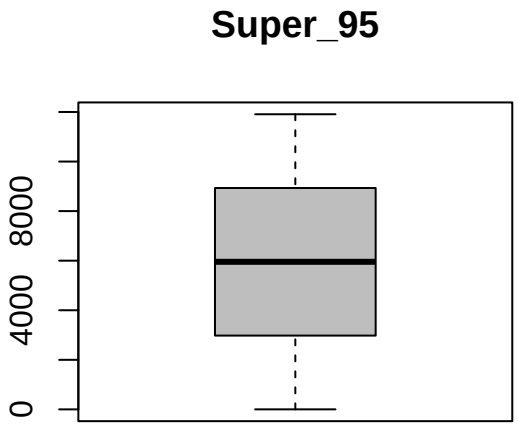

density
4
3
2
1
0
1.2
1.3
1.4
1.5
1.6
1.7
Gasóleo_C

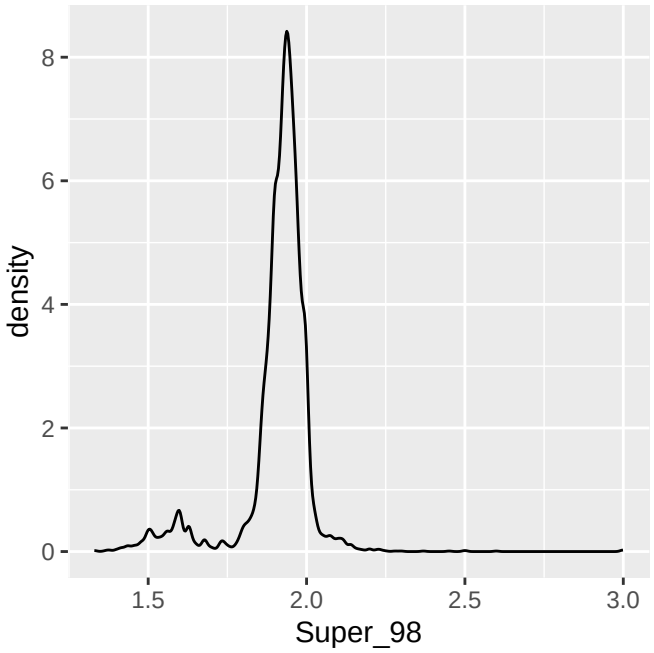

Gasóleo_C
0
4000
8000

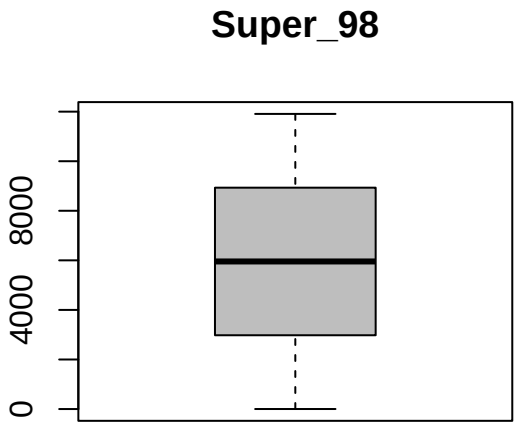

density
8
6
4
2
0
0.8
1.0
1.2
GLP


GLP
0
4000
8000









Todos los diagramas de cajas nos indican que no hay valores extremos en nuestros datos. Los graficos de densidad nos indican que en general casi todos los precios se distribuyen normalmente, con alguna distribuciones bimodales para los carburantes menos comunes (gas natural comprimido, gas natural licuado, biodiesel).

2.4 Variables latitud y longitud

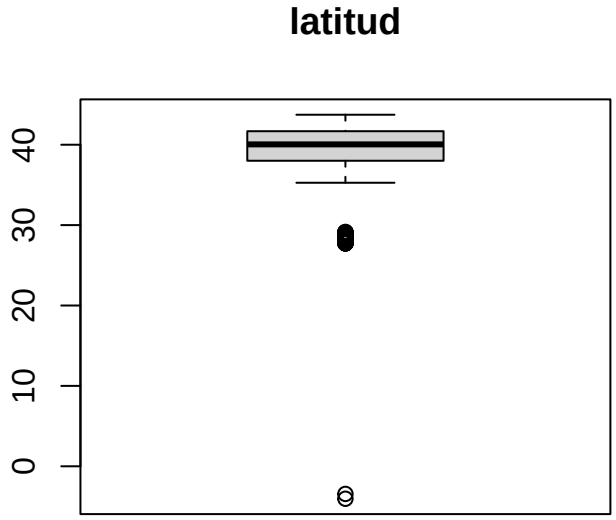
Ahora limpiaremos los valores de la variable latitud y longitud. Empezamos por solucionar los valores *NA*, como se trata solo de 4 instancias buscamos los valores de latitud y longitud de la gasolinera en internet para sustituir manualmente *Na* por el valor real.

	provincia	latitud	longitud
1916	BARCELONA	NA	NA
9882	SEVILLA	NA	NA
11376	VALLADOLID	NA	NA
11601	ZAMORA	NA	NA

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -4.038  38.008  40.042  39.615  41.681  43.732
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -18.0119  -5.4456  -3.4753  -3.3196  -0.6822  40.4712
```

Una vez solucionados los valores *Na* procedemos a comprobar la existencia de *outliers*.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -4.038  38.008  40.042  39.615  41.681  43.732
```

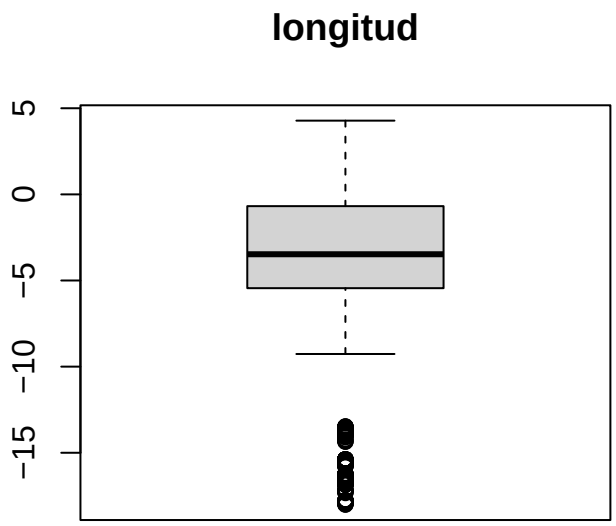
Observando los *outliers* de la latitud, los puntos en torno a 30 son correctos ya que pertenecen a las gasolineras de canarias, las cuales tienen latitudes cercanas a 30. El problema está en las latitudes inferiores a 0. Vamos a comprobar las instancias con estos valores.

	provincia	latitud	longitud
5762	JAÉN	-4.038306	38.03625
7275	MADRID	-3.436722	40.47117

Se puede ver que los datos de latitud y longitud están invertidos, así que procedemos a insertar los valores correctos.

	provincia	latitud	longitud
5762	JAÉN	38.03625	-4.038306
7275	MADRID	40.47117	-3.436722

Una vez solucionados los problemas con los *outliers* de la latitud vamos a ver los *outliers* de la longitud.




```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -18.0119  -5.4456  -3.4753  -3.3268  -0.6841   4.2795
```

Aunque tenemos *outliers* inferiores, los datos están dentro del rango aceptable. Estos outliers cercanos a -18 se deben a las islas canarias.

2.5 Variable Provincia

Ya hemos limpiado todos los datos numéricos, así que ahora pasaremos a limpiar los datos cualitativos.

Comenzamos con la variable provincia. Comprobamos si el número de provincias es el correcto y si los nombres son los correctos.

```
##  [1] "ALBACETE"              "ALICANTE"              "ALMERÍA"
##  [4] "ARABA/ÁLAVA"           "ASTURIAS"              "ÁVILA"
##  [7] "BADAJOZ"               "ILLES"                 "BARCELONA"
## [10] "BIZKAIA"               "BURGOS"                "CÁCERES"
## [13] "CÁDIZ"                 "CANTABRIA"             "CASTELLÓN / CASTELLÓ"
## [16] "CEUTA"                 "CIUDAD REAL"           "CÓRDOBA"
## [19] "A"                     "CUENCA"                "GIPUZKOA"
## [22] "GIRONA"                "GRANADA"               "GUADALAJARA"
## [25] "HUELVA"                "HUESCA"                "JAÉN"
## [28] "LEÓN"                  "LLEIDA"                "LUGO"
## [31] "MADRID"                "MÁLAGA"                "MELILLA"
## [34] "MURCIA"                "NAVARRA"               "OURENSE"
## [37] "PALENCIA"              "LAS"                   "PONTEVEDRA"
## [40] "LA"                    "SALAMANCA"             "SANTA CRUZ DE TENERIFE"
## [43] "SEGOVIA"               "SEVILLA"               "SORIA"
## [46] "TARRAGONA"             "TERUEL"                "TOLEDO"
```

```
## [49] "VALENCIA / VALÈNCIA"    "VALLADOLID"            "ZAMORA"
## [52] "ZARAGOZA"
```

El número de provincias es correcto, pero parece que hay tres provincias con el nombre incorrecto.

	index	provincia
4180	4180	A
4181	4181	A
4182	4182	A

	index	provincia
9013	9013	LA
9014	9014	LA
9015	9015	LA

	index	provincia
8571	8571	LAS
8572	8572	LAS
8573	8573	LAS

Parece que parte del nombre se ha quedado en la variable localidad, así que damos los valores pertinentes a las provincias y eliminamos de localidad la parte correspondiente a la provincia.

```
##  [1] "ALBACETE"              "ALICANTE"              "ALMERÍA"
##  [4] "ARABA/ÁLAVA"           "ASTURIAS"              "ÁVILA"
##  [7] "BADAJOZ"               "ILLES"                 "BARCELONA"
## [10] "BIZKAIA"               "BURGOS"                "CÁCERES"
## [13] "CÁDIZ"                 "CANTABRIA"             "CASTELLÓN / CASTELLÓ"
## [16] "CEUTA"                 "CIUDAD REAL"           "CÓRDOBA"
## [19] "A CORUÑA"              "CUENCA"                "GIPUZKOA"
## [22] "GIRONA"                "GRANADA"               "GUADALAJARA"
## [25] "HUELVA"                "HUESCA"                "JAÉN"
## [28] "LEÓN"                  "LLEIDA"                "LUGO"
## [31] "MADRID"                "MÁLAGA"                "MELILLA"
## [34] "MURCIA"                "NAVARRA"               "OURENSE"
## [37] "PALENCIA"              "LAS PALMAS"            "PONTEVEDRA"
## [40] "LA RIOJA"              "SALAMANCA"             "SANTA CRUZ DE TENERIFE"
## [43] "SEGOVIA"               "SEVILLA"               "SORIA"
## [46] "TARRAGONA"             "TERUEL"                "TOLEDO"
## [49] "VALENCIA / VALÈNCIA"   "VALLADOLID"            "ZAMORA"
## [52] "ZARAGOZA"
```

Corregidos los valores de las provincias pasaremos a comprobar si existe algún valor vacío en la variable localidad y en la variable dirección.

index	localidad

index	localidad

No existen valores vacíos así que damos ambas variables por buenas. Pasamos entonces a comprobar si las variables margen, horario, empresa, fechaRevision, fechaUltima y ventaPublico son correctas. Para ello, comprobaremos los valores únicos de margen, fechaRevision y fechaUltima. Mientras que veremos si existe algún valor vacío en horario, empresa y ventaPublico.

x
Derecho
Izquierdo
N

x
18/11/2022

```
##  [1] "17/11/2022" "01/11/2022" "18/11/2022" "16/11/2022" "14/11/2022"
##  [6] "15/11/2022" "04/11/2022" "09/11/2022" "03/11/2022" "10/11/2022"
## [11] "13/11/2022" "21/10/2022" "11/11/2022" "31/10/2022" "25/10/2022"
## [16] "07/11/2022" "08/11/2022" "02/11/2022" "22/10/2022" "12/11/2022"
## [21] "26/10/2022" "24/10/2022" "05/11/2022" "20/10/2022" "30/10/2022"
## [26] "28/10/2022" "27/10/2022" "23/10/2022" "29/10/2022" "06/11/2022"
```

index	horario

index	empresa

index	ventaPublico

Todos los datos parecen correctos, así que no es necesaria una limpieza de estos.

2.6 Variables empresa y ventaPublico

Comenzemos limpiando los datos vacios

index	empresa

index	ventaPublico

El siguiente paso es crear una tabla de frequencias para estas dos variables.

var	frequency	percentage	cumulative_perc
REPSOL	2770	23.26	23.26
CEPSA	1370	11.50	34.76
GALP	511	4.29	39.05
SHELL	366	3.07	42.12
BP	212	1.78	43.90
BALLENOIL	184	1.55	45.45
PETRONOR	177	1.49	46.94
AVIA	158	1.33	48.27

var	frequency	percentage	cumulative_perc
CARREFOUR	142	1.19	49.46
PLENOIL	139	1.17	50.63

En cuanto a la variable empresa, esta contiene 4027 diferentes valores. La petrolera Repsol gestiona el 23% de todas las gasolineras en el territorio nacional, seguida de Cepsa (11%).

La variable ventaPublico tiene solo dos posibles valores: "Venta al publico" (92%) y "Cerrada al publico general" (7%).

3. Analisis de los datos

En este apartado vamos a responder a las siguientes preguntas:

- ¿Cuales son las cadenas de gasolineras mas caras?
- ¿Se puede predecir el valor del combustible con los datos que tenemos?
- ¿Existe diferencia en el precio del Gasóleo A entre el norte y el sur de España? ¿y entre el este y el oeste?
- ¿Podríamos predecir si la gasolinera esta en el norte sabiendo el precio del Gasóleo A? Y si está en el este?

3.1 ¿Cuales son las cadenas de gasolineras mas caras?

3.1.1 Consideraciones iniciales y creacion del dataset Para responder a esta pregunta vamos a realizar las siguientes consideraciones:

- Analizaremos unicamente el precio del Gasoleo A, que es el que esta disponible en mas gasolineras (11.560 de las 11.909).
- Solo incluimos estaciones de servicio con venta al publico.
- Ya que existen mas 4.027 empresas asociadas con estaciones de servicio, compareremos unicamente las 10 cadenas principales, que gestionan algo mas del 50% de todas las gasolineras en el pais (REPSOL, CEPSA, GALP, SHELL, BP, BALLENOIL, PETRONOR, AVIA. CARREFOUR, PLENOIL).

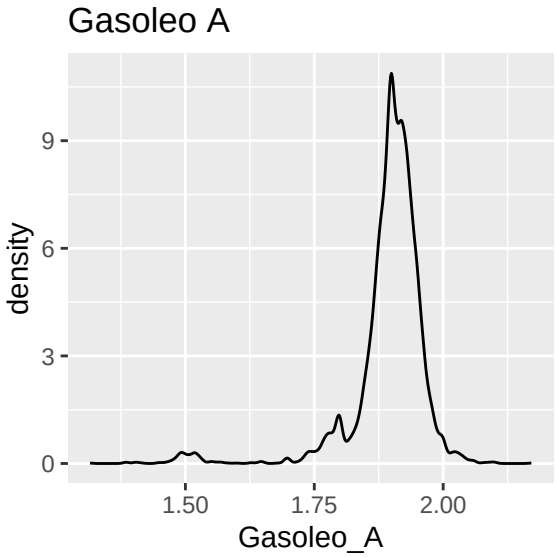
A continuacion creamos el dataset con las condiciones anteriores. Primero obtenemos las 10 empresas con mas gasolineras, y nos quedamos unicamente con las estaciones de servicio de dichas empresas

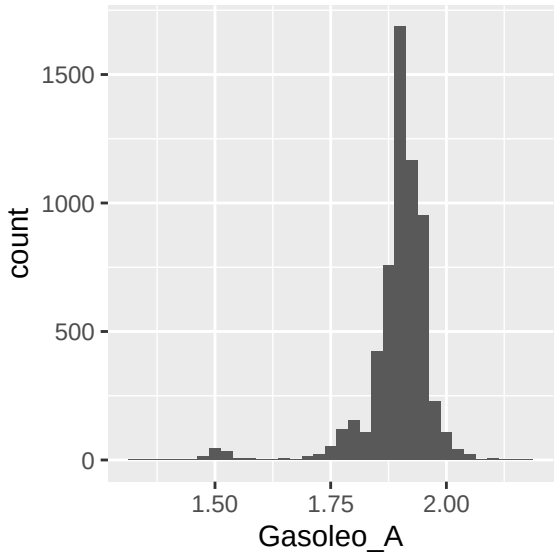
```
##  [1] "AVIA"      "BALLENOIL" "BP"        "CARREFOUR" "CEPSA"     "GALP"
##  [7] "PETRONOR"  "PLENOIL"   "REPSOL"    "SHELL"
```

Ahora eliminamos las gasolineras no abiertas al publico,y reducimos el volumen de datos utilizados seleccionando solo las columnas que pueden ser de utilidad. Tambien eliminamos todos los valores nulos en el precio del gasoleo

empresa	provincia	Gasoleo_A
AVIA	LA RIOJA	1.969
AVIA	LA RIOJA	1.929
AVIA	ARABA/ÁLAVA	1.949
AVIA	MADRID	1.889
AVIA	LA RIOJA	1.959
AVIA	A CORUÑA	1.959

3.1.2 Comprobacion de la normalidad y homogeneidad de varianzas Antes de continuar con el analisis, comprobemos la normalidad de la variable Gasoleo_A. En primer lugar creamos un diagrama de densidad y un histograma.

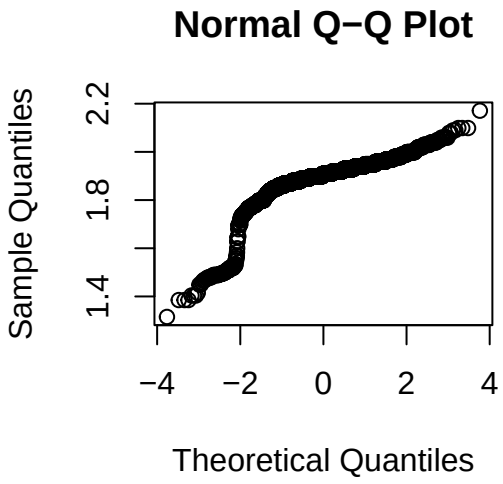





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.315   1.879   1.909   1.895   1.929   2.171
```

Visualmente se comprueba que la variable tiene cola hacia la izquierda. Por otra parte, la tabla resumen nos muestra que la media y la mediana son valores muy cercanos, con lo que la distribucion esta centrada.

Usemos ahora las curva Q-Q, que representa los quantiles de nuestras observaciones frente a una hipotetica distribucion normal. Si la representacion sigue una linea recta, significa que podemos asumir normalidad.



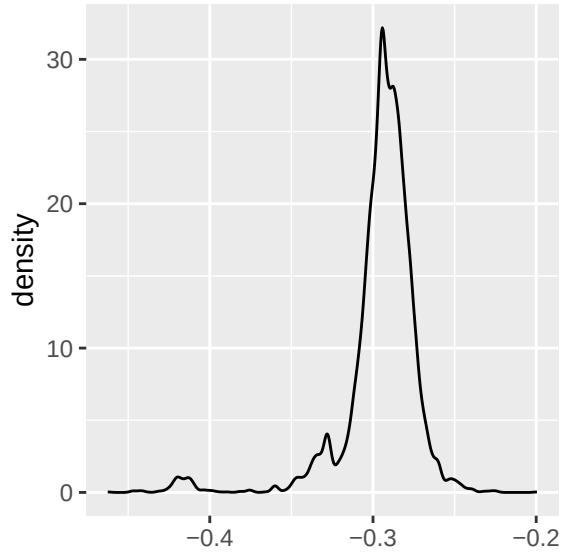
De las representaciones anteriores podemos concluir que la variable Gasoleo_A es normal por encima de 1.8. La cola de la izquierda afecta a la normalidad de la variable, y deberiamos estudiar los valores mas extremos para enterder si son outliers o considerar esas observaciones como una poblacion de estaciones de servicio diferente.

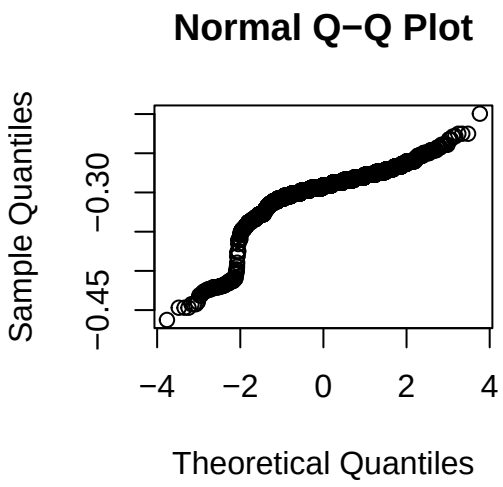
Finalmente realizamos el test de Kolmogorov-Smirnov.

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  top_10$Gasoleo_A
## D = 0.19127, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

p<0,05, con lo que corfirmamos que la distribucion de la variable Gasole_A no es normal.

Veamos si podemos normalizar la variable mediante la transformacion de Box-Cox.






```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  top_10_nr$Gasoleo_A
## D = 0.17985, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

La transformacion Box-Cox no parece que tenga ningun efecto en la normalidad. Por esa razon, utilizaremos la variable original.

El test que vamos a aplicar a continuacion (ANOVA) es muy robusto frente a la falta de normalidad, especialmente cuando el numero de observaciones es elevado. Por esa razon, pasaremos por alto la falta de normalidad de esta variables

Comprobemos tambien la homogeneidad de varianzas utilizando el test de Levene

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    9  8.2803 2.374e-12 ***
##       5967
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

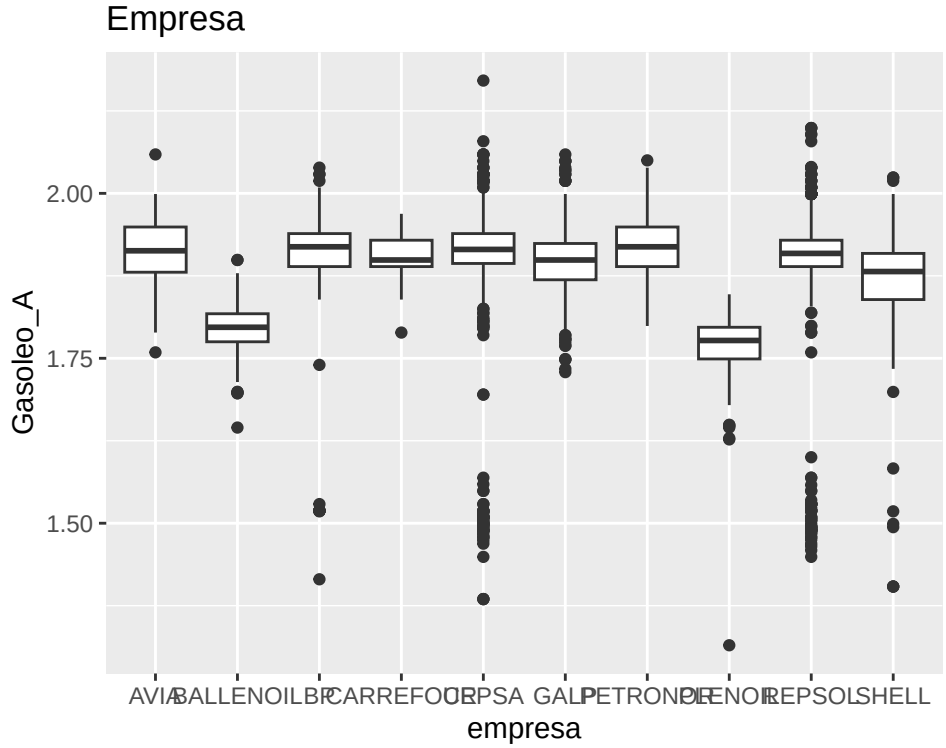
$p<0.05$, con lo que se rechaza la hipótesis nula de homocedasticidad y se concluye que la variable Gasoleo_A presenta varianzas estadísticamente diferentes para los diferentes grupos de empresa.

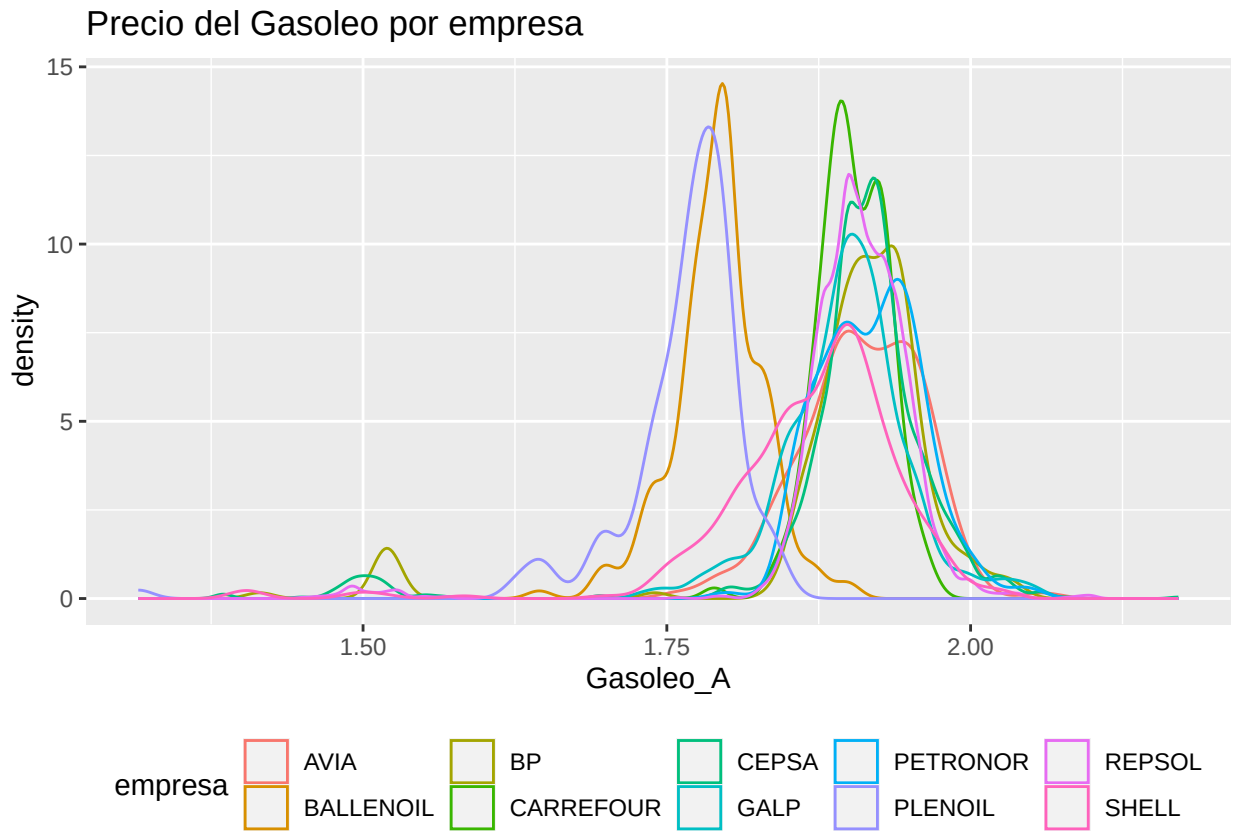
3.1.3 Analisis de los precios medios del gasoleo y modelo ANOVA Comparemos ahora los precios medios y medianos por empresa. Primero creamos una tabla para ver ambos valores.

empresa	Media	Mediana
AVIA	1.910721	1.9130
BALLENOIL	1.792331	1.7970
BP	1.900359	1.9190
CARREFOUR	1.902972	1.8990
CEPSA	1.902944	1.9150
GALP	1.898092	1.8990
PETRONOR	1.918294	1.9190

empresa	Media	Mediana
PLENOIL	1.764381	1.7770
REPSOL	1.903996	1.9090
SHELL	1.869238	1.8815

Los diagramas de caja y graficos de densidad nos ayudaran a enterder mejor los datos.





La primera conclusion que obtenemos es que Plenoil y Ballenoil parecen tener menores precios para el Gasoleo A.

Para entender si la diferencia de precios es estadisticamente significativa, aplicamos el test ANOVA. En el test ANOVA vamos a comparar las medias de precios de las 10 empresas, y estimar si estas son iguales o no.

- H0: $\mu 1 = \mu 2 = \mu 3 = \mu 4 = \mu 5 = \mu 6 = \mu 7 = \mu 8 = \mu 9 = \mu 10$
- H1: Las medias no son iguales.

```
## Call:
##    aov(formula = Gasoleo_A ~ empresa, data = top_10)
##
## Terms:
##                   empresa Residuals
## Sum of Squares   5.016313 28.246947
## Deg. of Freedom         9      5967
##
## Residual standard error: 0.06880306
## Estimated effects may be unbalanced
```

```
##
## Call:
## aov(formula = Gasoleo_A ~ empresa, data = top_10)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51794 -0.01738  0.00500  0.03206  0.26806
##
```

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.910722   0.005474 349.075  < 2e-16 ***
## empresaBALLENOIL -0.118390   0.007462 -15.865  < 2e-16 ***
## empresaBP        -0.010363   0.007231  -1.433   0.1519
## empresaCARREFOUR -0.007750   0.007971  -0.972   0.3310
## empresaCEPSA     -0.007777   0.005790  -1.343   0.1793
## empresaGALP      -0.012629   0.006266  -2.016   0.0439 *
## empresaPETRONOR   0.007572   0.007530   1.006   0.3147
## empresaPLENOIL   -0.146340   0.008001 -18.290  < 2e-16 ***
## empresaREPSOL    -0.006725   0.005628  -1.195   0.2322
## empresaSHELL     -0.041484   0.006549  -6.334 2.56e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0688 on 5967 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1495
## F-statistic: 117.7 on 9 and 5967 DF,  p-value: < 2.2e-16
```

El valor de P<0.05, por lo que rechazamos la hipotesis nula y concluimos que el precio medio por empresa no es igual.

Este resultado solo nos indica que el valor medio para cada empresa no es igual, pero no dice nada sobre que empresas son diferentes entre si. Para ello utilizamos el test the Tukey.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Gasoleo_A ~ empresa, data = top_10)
##
## $empresa
##                              diff          lwr          upr     p adj
## BALLENOIL-AVIA      -0.1183899972 -0.142007926 -0.094772068 0.0000000
## BP-AVIA             -0.0103630284 -0.033249066  0.012523009 0.9170003
## CARREFOUR-AVIA      -0.0077498878 -0.032976777  0.017477002 0.9937585
## CEPSA-AVIA          -0.0077772419 -0.026102413  0.010547929 0.9435545
## GALP-AVIA           -0.0126291811 -0.032460048  0.007201686 0.5879729
## PETRONOR-AVIA        0.0075722663 -0.016260444  0.031404977 0.9920069
## PLENOIL-AVIA        -0.1463402240 -0.171662836 -0.121017612 0.0000000
## REPSOL-AVIA         -0.0067251382 -0.024537151  0.011086875 0.9732881
## SHELL-AVIA          -0.0414838141 -0.062212070 -0.020755558 0.0000000
## BP-BALLENOIL         0.1080269688  0.086086952  0.129966986 0.0000000
## CARREFOUR-BALLENOIL  0.1106401095  0.086268207  0.135012012 0.0000000
## CEPSA-BALLENOIL      0.1106127554  0.093483671  0.127741840 0.0000000
## GALP-BALLENOIL       0.1057608162  0.087029641  0.124491991 0.0000000
## PETRONOR-BALLENOIL   0.1259622636  0.103036476  0.148888051 0.0000000
## PLENOIL-BALLENOIL   -0.0279502268 -0.052421196 -0.003479258 0.0112845
## REPSOL-BALLENOIL     0.1116648590  0.095085913  0.128243805 0.0000000
## SHELL-BALLENOIL      0.0769061832  0.057227406  0.096584960 0.0000000
## CARREFOUR-BP         0.0026131406 -0.021050201  0.026276483 0.9999987
## CEPSA-BP             0.0025857865 -0.013519172  0.018690745 0.9999671
## GALP-BP             -0.0022661526 -0.020065621  0.015533315 0.9999956
## PETRONOR-BP          0.0179352947 -0.004235765  0.040106354 0.2370108
## PLENOIL-BP          -0.1359771956 -0.159742558 -0.112211833 0.0000000
## REPSOL-BP            0.0036378902 -0.011880666  0.019156447 0.9992317
## SHELL-BP            -0.0311207856 -0.049914891 -0.012326680 0.0000073
## CEPSA-CARREFOUR     -0.0000273541 -0.019314525  0.019259816 1.0000000
## GALP-CARREFOUR      -0.0048792933 -0.025602380  0.015843793 0.9992041
## PETRONOR-CARREFOUR   0.0153221541 -0.009257942  0.039902251 0.6183338
## PLENOIL-CARREFOUR   -0.1385903362 -0.164617586 -0.112563087 0.0000000
## REPSOL-CARREFOUR     0.0010247495 -0.017775539  0.019825038 1.0000000
## SHELL-CARREFOUR     -0.0337339263 -0.055317338 -0.012150515 0.0000341
## GALP-CEPSA          -0.0048519392 -0.016203690  0.006499812 0.9410627
## PETRONOR-CEPSA       0.0153495082 -0.002074529  0.032773546 0.1405799
## PLENOIL-CEPSA       -0.1385629821 -0.157975186 -0.119150779 0.0000000
## REPSOL-CEPSA         0.0010521036 -0.006218852  0.008323059 0.9999865
## SHELL-CEPSA         -0.0337065722 -0.046561905 -0.020851239 0.0000000
## PETRONOR-GALP        0.0202014474  0.001200172  0.039202723 0.0266629
## PLENOIL-GALP        -0.1337110430 -0.154550549 -0.112871537 0.0000000
## REPSOL-GALP          0.0059040428 -0.004599203  0.016407289 0.7486707
## SHELL-GALP          -0.0288546330 -0.043778137 -0.013931129 0.0000000
## PLENOIL-PETRONOR    -0.1539124903 -0.178590818 -0.129234163 0.0000000
## REPSOL-PETRONOR     -0.0142974046 -0.031180917  0.002586108 0.1818757
## SHELL-PETRONOR      -0.0490560804 -0.068992123 -0.029120038 0.0000000
## REPSOL-PLENOIL       0.1396150858  0.120686548  0.158543624 0.0000000
## SHELL-PLENOIL        0.1048564100  0.083161195  0.126551625 0.0000000
## SHELL-REPSOL        -0.0347586758 -0.046871294 -0.022646058 0.0000000
```

Los resultado confirman lo que ya habiamos intuido anteriormente: Plenoil y Ballenoil son mas baratas que las otras 8 empresas. Tambien podemos observar que es Pleanoil es mas barata que Ballenoil (p=0.0112845), con lo que podemos considerarla la cadena mas barata.

Otros resultados significativos son SHELL-BP (p=0.0000073), SHELL-CARREFOUR (p=0.0000341) y PETRONOR-GALP (p=0.0266629)

Hemos visto que hay una diferencia significativa en las medias. Ahora veamos cual es la capacidad explicativa del modelo. Para ello calcularemos R^2 usando la funcion etaSquared del paquete lsr.

```
##            eta.sq eta.sq.part
## empresa 0.1508064   0.1508064
```

El resultado, que es la medida de la intensidad de la relacion, nos indica que el modelo explica el 15% de la variabilidad total.

Hemos visto que dos de las cadenas de carburantes son mas baratas que las otras 8. Pero veamos si la provincia tiene influencia en el precio. Lo que queremos averiguar es si los resultados anteriores pueden estar sesgados por la provincia es la que se localizan las estaciones de servicio de cada cadena.

Para simplicar los calculos, compararemos Plenoil, Ballenoil y Repsol.

Tambien seleccionamos solo las provincias que tengan estaciones de servicio de las tres empresas.

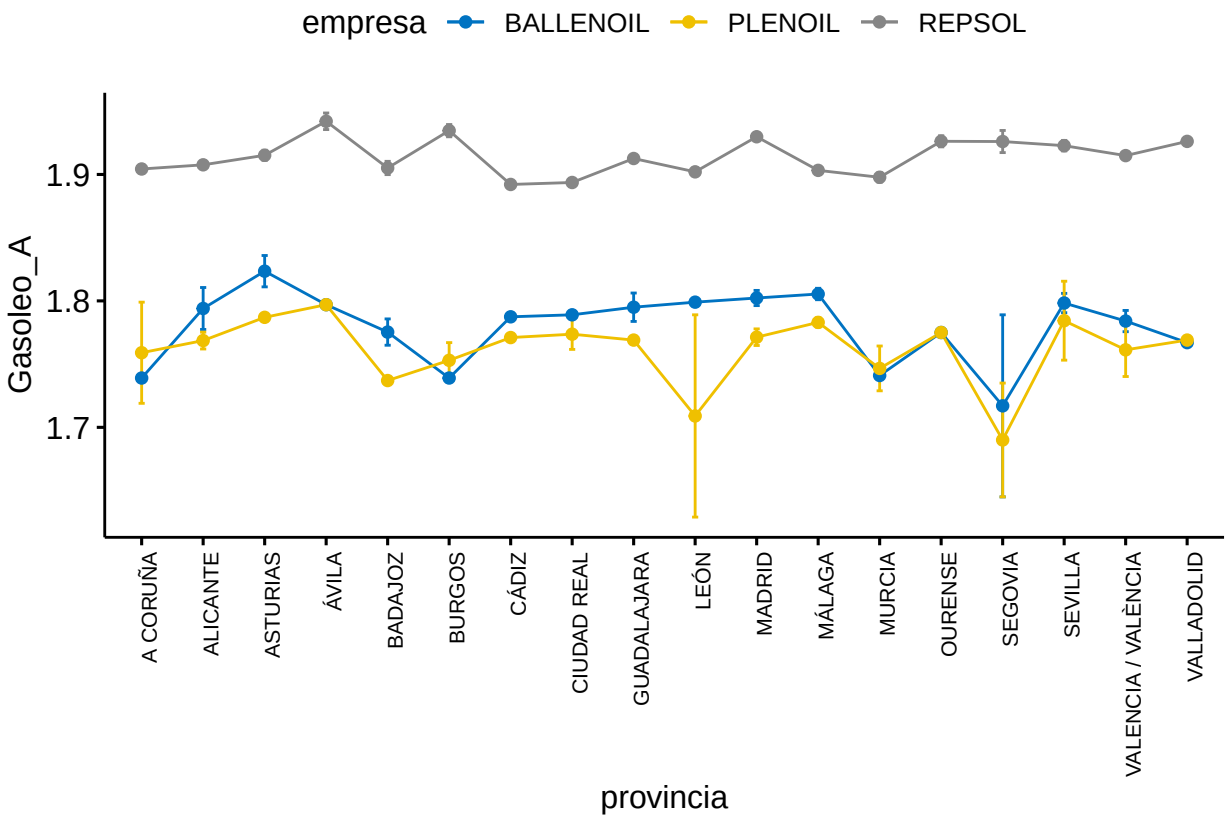
```
##  [1] "A CORUÑA"            "ALICANTE"            "ASTURIAS"
##  [4] "ÁVILA"               "BADAJOZ"             "BURGOS"
##  [7] "CÁDIZ"               "CIUDAD REAL"         "GUADALAJARA"
## [10] "LEÓN"                "MADRID"              "MÁLAGA"
## [13] "MURCIA"              "OURENSE"             "SEGOVIA"
## [16] "SEVILLA"             "VALENCIA / VALÈNCIA" "VALLADOLID"
```

Para ver el posible impacto de la provincia en el precio, creamos una tabla que agrupa los valores medios por provincia y empresa.

provincia	empresa	Media
A CORUÑA	BALLENOIL	1.739000
A CORUÑA	PLENOIL	1.759000
A CORUÑA	REPSOL	1.904319
ALICANTE	BALLENOIL	1.794000
ALICANTE	PLENOIL	1.768615
ALICANTE	REPSOL	1.907614

En esta tabla ya podemos comprobar que en general Ballenoil y Plenoil siguen siendo mas baratas que Repsol.

Representamos el grafico de perfil para visualizar las diferencias de precio entre empresas por provincia.



A continuacion aplicamos el modelo ANoVA de dos factores. Incluimos el termino empresa:provincia ya que queremos entender si existe interaccion entre estas dos variables.

```
## Call:
##    aov(formula = Gasoleo_A ~ empresa + provincia + empresa:provincia,
##     data = top_3)
##
## Terms:
##                   empresa provincia empresa:provincia Residuals
## Sum of Squares  3.723252  0.211078          0.077215  1.391420
## Deg. of Freedom        2        17                34      1488
##
## Residual standard error: 0.03057931
## Estimated effects may be unbalanced
```

```
##                     Df Sum Sq Mean Sq  F value   Pr(>F)
## empresa              2  3.723  1.8616 1990.844  < 2e-16 ***
## provincia           17  0.211  0.0124   13.278  < 2e-16 ***
## empresa:provincia   34  0.077  0.0023    2.429 9.72e-06 ***
## Residuals         1488  1.391  0.0009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los resultados nos indican que los tres factores (empresa, provincia y su interaccion) son estadisticamente significativos. Veamos el orden de importancia.

```
##                       eta.sq eta.sq.part
## empresa           0.68195488  0.72588250
```

```
## provincia          0.03906705  0.13171807
## empresa:provincia 0.01429121  0.05257598
```

Empresa es el mas significativo, seguido de provincia y de su interaccion.

```
##
## Call:
## aov(formula = Gasoleo_A ~ empresa + provincia + empresa:provincia,
##     data = top_3)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.114286 -0.018614 -0.001077  0.014681  0.193879
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             1.739e+00  3.058e-02  56.869
## empresaPLENOIL                          2.000e-02  3.745e-02   0.534
## empresaREPSOL                           1.653e-01  3.079e-02   5.369
## provinciaALICANTE                       5.500e-02  3.419e-02   1.609
## provinciaASTURIAS                       8.450e-02  3.419e-02   2.472
## provinciaÁVILA                          5.800e-02  4.325e-02   1.341
## provinciaBADAJOZ                        3.633e-02  3.303e-02   1.100
## provinciaBURGOS                         8.280e-15  4.325e-02   0.000
## provinciaCÁDIZ                          4.841e-02  3.147e-02   1.539
## provinciaCIUDAD REAL                    5.000e-02  4.325e-02   1.156
## provinciaGUADALAJARA                    5.600e-02  3.350e-02   1.672
## provinciaLEÓN                           6.000e-02  3.531e-02   1.699
## provinciaMADRID                         6.329e-02  3.085e-02   2.051
## provinciaMÁLAGA                         6.644e-02  3.223e-02   2.061
## provinciaMURCIA                         2.000e-03  3.745e-02   0.053
## provinciaOURENSE                        3.600e-02  4.325e-02   0.832
## provinciaSEGOVIA                       -2.200e-02  3.745e-02  -0.587
## provinciaSEVILLA                        5.931e-02  3.152e-02   1.882
## provinciaVALENCIA / VALÈNCIA            4.506e-02  3.142e-02   1.434
## provinciaVALLADOLID                     2.800e-02  3.531e-02   0.793
## empresaPLENOIL:provinciaALICANTE       -4.538e-02  4.089e-02  -1.110
## empresaREPSOL:provinciaALICANTE        -5.171e-02  3.451e-02  -1.498
## empresaPLENOIL:provinciaASTURIAS       -5.650e-02  5.071e-02  -1.114
## empresaREPSOL:provinciaASTURIAS        -7.366e-02  3.458e-02  -2.130
## empresaPLENOIL:provinciaÁVILA          -2.000e-02  5.721e-02  -0.350
## empresaREPSOL:provinciaÁVILA           -2.027e-02  4.396e-02  -0.461
## empresaPLENOIL:provinciaBADAJOZ        -5.833e-02  4.994e-02  -1.168
## empresaREPSOL:provinciaBADAJOZ         -3.553e-02  3.344e-02  -1.063
## empresaPLENOIL:provinciaBURGOS         -6.000e-03  5.296e-02  -0.113
## empresaREPSOL:provinciaBURGOS           3.033e-02  4.371e-02   0.694
## empresaPLENOIL:provinciaCÁDIZ          -3.641e-02  4.055e-02  -0.898
## empresaREPSOL:provinciaCÁDIZ           -6.065e-02  3.193e-02  -1.900
## empresaPLENOIL:provinciaCIUDAD REAL    -3.533e-02  5.147e-02  -0.686
## empresaREPSOL:provinciaCIUDAD REAL     -6.068e-02  4.359e-02  -1.392
## empresaPLENOIL:provinciaGUADALAJARA    -4.600e-02  4.536e-02  -1.014
## empresaREPSOL:provinciaGUADALAJARA     -4.771e-02  3.414e-02  -1.398
## empresaPLENOIL:provinciaLEÓN           -1.100e-01  4.671e-02  -2.355
## empresaREPSOL:provinciaLEÓN            -6.226e-02  3.575e-02  -1.742
## empresaPLENOIL:provinciaMADRID         -5.099e-02  3.813e-02  -1.337
## empresaREPSOL:provinciaMADRID          -3.779e-02  3.113e-02  -1.214
## empresaPLENOIL:provinciaMÁLAGA         -4.244e-02  4.050e-02  -1.048
## empresaREPSOL:provinciaMÁLAGA          -6.753e-02  3.262e-02  -2.070
## empresaPLENOIL:provinciaMURCIA         -1.440e-02  4.431e-02  -0.325
## empresaREPSOL:provinciaMURCIA          -8.567e-03  3.774e-02  -0.227
## empresaPLENOIL:provinciaOURENSE        -2.000e-02  5.721e-02  -0.350
## empresaREPSOL:provinciaOURENSE         -1.410e-02  4.373e-02  -0.322
## empresaPLENOIL:provinciaSEGOVIA        -4.700e-02  4.835e-02  -0.972
## empresaREPSOL:provinciaSEGOVIA          4.372e-02  3.812e-02   1.147
## empresaPLENOIL:provinciaSEVILLA        -3.398e-02  4.210e-02  -0.807
## empresaREPSOL:provinciaSEVILLA         -4.087e-02  3.190e-02  -1.281
## empresaPLENOIL:provinciaVALENCIA / VALÈNCIA -4.277e-02  3.985e-02  -1.073
## empresaREPSOL:provinciaVALENCIA / VALÈNCIA  -3.448e-02  3.174e-02  -1.086
## empresaPLENOIL:provinciaVALLADOLID      -1.800e-02  4.671e-02  -0.385
## empresaREPSOL:provinciaVALLADOLID       -6.166e-03  3.575e-02  -0.172
##                                        Pr(>|t|)
## (Intercept)                             < 2e-16 ***
## empresaPLENOIL                            0.5934
## empresaREPSOL                          9.17e-08 ***
## provinciaALICANTE                         0.1079
## provinciaASTURIAS                         0.0136 *
## provinciaÁVILA                            0.1801
## provinciaBADAJOZ                          0.2715
## provinciaBURGOS                           1.0000
## provinciaCÁDIZ                            0.1241
## provinciaCIUDAD REAL                      0.2478
## provinciaGUADALAJARA                      0.0948 .
## provinciaLEÓN                             0.0895 .
## provinciaMADRID                           0.0404 *
## provinciaMÁLAGA                           0.0394 *
## provinciaMURCIA                           0.9574
## provinciaOURENSE                          0.4053
## provinciaSEGOVIA                          0.5570
## provinciaSEVILLA                          0.0601 .
## provinciaVALENCIA / VALÈNCIA              0.1518
## provinciaVALLADOLID                       0.4279
## empresaPLENOIL:provinciaALICANTE          0.2673
## empresaREPSOL:provinciaALICANTE           0.1343
## empresaPLENOIL:provinciaASTURIAS          0.2654
## empresaREPSOL:provinciaASTURIAS           0.0333 *
## empresaPLENOIL:provinciaÁVILA             0.7267
## empresaREPSOL:provinciaÁVILA              0.6448
## empresaPLENOIL:provinciaBADAJOZ           0.2429
## empresaREPSOL:provinciaBADAJOZ            0.2881
## empresaPLENOIL:provinciaBURGOS            0.9098
## empresaREPSOL:provinciaBURGOS             0.4879
## empresaPLENOIL:provinciaCÁDIZ             0.3694
## empresaREPSOL:provinciaCÁDIZ              0.0577 .
## empresaPLENOIL:provinciaCIUDAD REAL       0.4925
## empresaREPSOL:provinciaCIUDAD REAL        0.1641
```

```
## empresaPLENOIL:provinciaGUADALAJARA           0.3107
## empresaREPSOL:provinciaGUADALAJARA            0.1625
## empresaPLENOIL:provinciaLEÓN                  0.0187 *
## empresaREPSOL:provinciaLEÓN                   0.0818 .
## empresaPLENOIL:provinciaMADRID                0.1814
## empresaREPSOL:provinciaMADRID                 0.2249
## empresaPLENOIL:provinciaMÁLAGA                0.2948
## empresaREPSOL:provinciaMÁLAGA                 0.0386 *
## empresaPLENOIL:provinciaMURCIA                0.7453
## empresaREPSOL:provinciaMURCIA                 0.8204
## empresaPLENOIL:provinciaOURENSE               0.7267
## empresaREPSOL:provinciaOURENSE                0.7472
## empresaPLENOIL:provinciaSEGOVIA               0.3312
## empresaREPSOL:provinciaSEGOVIA                0.2516
## empresaPLENOIL:provinciaSEVILLA               0.4198
## empresaREPSOL:provinciaSEVILLA                0.2003
## empresaPLENOIL:provinciaVALENCIA / VALÈNCIA   0.2833
## empresaREPSOL:provinciaVALENCIA / VALÈNCIA    0.2775
## empresaPLENOIL:provinciaVALLADOLID            0.7000
## empresaREPSOL:provinciaVALLADOLID             0.8631
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03058 on 1488 degrees of freedom
## Multiple R-squared:  0.7425, Adjusted R-squared:  0.7333
## F-statistic: 80.94 on 53 and 1488 DF,  p-value: < 2.2e-16
```

EL modelo explica el 74% de la variabilidad entre medias.

3.1.4 Conclusiones Los resultados obtenidos nos indican que el precio del combustible varia dependiendo de las empresas que gestionan las estaciones de servicio, y tambien depende de la provincia en la que se encuentren.

Hemos comprobado como Ballenoil y Plenoil son mas baratas, y los resultado son estadisticamente significativos.

Por motivos de simplicidad, y para faciliar la visualizacion de los resultados, utilizamos un numero reducido de empresas. Este analisis se prodria aplicar a todas las empresas en el juego de datos, pero para ello deberia considerarse agrupar las empresas con menor numero de gasolineras en grupos (por ejemplo, de 1 a 5 gasolineras, de 5 a 10, etc. . .)

3.2 ¿Se puede predecir el valor del combustible con los datos que tenemos?

En el apartado anterior hemos comprobado que el valor del gasoleo A se ve influenciado por la empresa que gestiona la estacion de servicio, y por la provincia en la que se localiza. En este apartado vamos a predecir el valor del gasoleo utilizando una regresion, usando variables en el juego de datos (latitud, longitud), y creando nuevas variables.

3.2.1 Consideraciones iniciales y creacion del dataset Para responder a esta pregunta vamos a realizar las siguientes consideraciones:

- Analizaremos unicamente el precio del Gasoleo A, que es el que esta disponible en mas gasolineras (11.560 de las 11.909).
- Solo incluimos estaciones de servicio con venta al publico.
- Ya que la variable a predecir (Gasoleo A) es continua, utilizaremos una regresion lineal.
- La regresion lineal solo admite variables independientes numericas. Para incluir el impacto de la empresa, de la provincia y de la localidad en el modelo, creamos tres nuevas variables (num_gas_empr, num_gas_prov, num_gas_locl) que indican el numero de gasolineras perteneciente a cada empresa, provincia y localidad.

Comenzamos creando tres nuevas tablas, que agrupan el numero de gasolineras por empresa, provincia y localidad

	empresa	num_gas_empr
3345	REPSOL	2770
1087	CEPSA	1370
2361	GALP	511

	empresa	num_gas_empr
3650	SHELL	366
427	BP	212
351	BALLENOIL	184

	provincia	num_gas_prov
33	MADRID	825
9	BARCELONA	806
49	VALENCIA / VALÈNCIA	642
44	SEVILLA	479
3	ALICANTE	468
36	MURCIA	458

	localidad	num_gas_locl
2124	MADRID	236
556	BARCELONA	90
3521	SEVILLA	73
3869	VALENCIA	64
4249	ZARAGOZA	57
2145	MALAGA	56

Unimos estas tres tablas a nuestros datos originales.

Eliminamos las gasolineras cerradas al publico, y seleccionamos solo los campos que vamos a utilizar en la regresion.

```
##     latitud         longitud         Gasoleo_A      num_gas_empr
##  Min.   :27.71   Min.   :-18.012   Min.   :1.315   Min.   :   1.0
##  1st Qu.:38.02   1st Qu.: -5.547   1st Qu.:1.839   1st Qu.:   1.0
##  Median :40.13   Median : -3.471   Median :1.890   Median : 177.0
##  Mean   :39.62   Mean   : -3.340   Mean   :1.868   Mean   : 929.2
##  3rd Qu.:41.73   3rd Qu.: -0.588   3rd Qu.:1.924   3rd Qu.:2770.0
##  Max.   :43.73   Max.   :  4.279   Max.   :2.309   Max.   :2770.0
##   num_gas_prov    num_gas_locl
##  Min.   :  9.0   Min.   :  1.00
##  1st Qu.:204.0   1st Qu.:  2.00
##  Median :273.0   Median :  5.00
##  Mean   :357.7   Mean   : 15.11
##  3rd Qu.:468.0   3rd Qu.: 14.00
##  Max.   :825.0   Max.   :236.00
```

3.2.2 Regresion Con los datos listos, aplicamos la regresion lineal

```
##
## Call:
## lm(formula = Gasoleo_A ~ num_gas_empr + num_gas_prov + num_gas_locl +
##     latitud + longitud, data = gas_rgrssn)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.42634 -0.04509  0.00385  0.04909  0.49217
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.231e+00  1.189e-02 103.522   <2e-16 ***
## num_gas_empr  1.897e-05  6.260e-07  30.308   <2e-16 ***
## num_gas_prov -7.755e-06  3.619e-06  -2.143   0.0322 *
## num_gas_locl  5.685e-05  2.215e-05   2.567   0.0103 *
## latitud       1.600e-02  2.801e-04  57.147   <2e-16 ***
## longitud      3.854e-03  2.347e-04  16.425   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0748 on 10867 degrees of freedom
## Multiple R-squared:  0.4328, Adjusted R-squared:  0.4325
## F-statistic:  1658 on 5 and 10867 DF,  p-value: < 2.2e-16
```

Los resultados del modelo nos indica que:

- Todas las variables independientes son significativas (p<0,05)
- El valor p del modelo es 2.2e-16 (<0,05), con lo que es estadisticamente significativo.
- EL modelo explica el 43% de la varianza de la variable dependiente

Intentemos mejorar el modelo. La latitud y la longitud determinan la localizacion de cada una de las estaciones. Ya que se necesitan ambas para poder saber la localizacion, y esta posiblemente tiene impacto en los precios, vamos a incluir la interaccion entre ambas en nuestro modelo.

```
##
## Call:
## lm(formula = Gasoleo_A ~ num_gas_empr + num_gas_prov + num_gas_locl +
##     latitud + longitud + latitud:longitud, data = gas_rgrssn)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49771 -0.03710  0.00341  0.03727  0.44677
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.038e+00  1.618e-02 125.934   <2e-16 ***
## num_gas_empr      1.734e-05  5.342e-07  32.454   <2e-16 ***
## num_gas_prov     -3.748e-06  3.086e-06  -1.215    0.225
## num_gas_locl      2.089e-05  1.889e-05   1.106    0.269
## latitud          -4.242e-03  3.965e-04 -10.700   <2e-16 ***
## longitud          8.584e-02  1.297e-03  66.162   <2e-16 ***
## latitud:longitud -2.149e-03  3.360e-05 -63.956   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06376 on 10866 degrees of freedom
## Multiple R-squared:  0.5879, Adjusted R-squared:  0.5877
## F-statistic:  2584 on 6 and 10866 DF,  p-value: < 2.2e-16
```

El modelo mejora su poder predictivo, y ahora explica el 58% del precio del gasoleo. Tambien las variables num_gas_prov y num_gas_locl dejan de ser significativas. Las excluimos del modelo y comprobamos si tenian algun impacto.

```
##
```

```
## Call:
## lm(formula = Gasoleo_A ~ num_gas_empr + latitud + longitud +
##     latitud:longitud, data = gas_rgrssn)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49623 -0.03703  0.00339  0.03720  0.44688
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.035e+00  1.586e-02  128.26   <2e-16 ***
## num_gas_empr      1.739e-05  5.327e-07   32.64   <2e-16 ***
## latitud          -4.196e-03  3.936e-04  -10.66   <2e-16 ***
## longitud          8.580e-02  1.296e-03   66.21   <2e-16 ***
## latitud:longitud -2.150e-03  3.358e-05  -64.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06376 on 10868 degrees of freedom
## Multiple R-squared:  0.5878, Adjusted R-squared:  0.5877
## F-statistic:  3875 on 4 and 10868 DF,  p-value: < 2.2e-16
```

3.2.3 Conclusiones La regresion lineal con las variables num_gas_empr (numero de gasolineras por empresa), latitud, longitud y la interaccion entre estas dos ultimas puede ser usada para calcular el precio del Gasoleo_A. El modelo explica el 58% de la variabilidad en la variable dependiente

3.3 ¿Existe diferencia en el precio del Gasoleo A entre el norte y el sur de España? ¿y entre el este y el oeste?

3.3.1 Consideraciones iniciales y creacion del dataset Para responder a esta pregunta vamos a realizar las siguientes consideraciones:

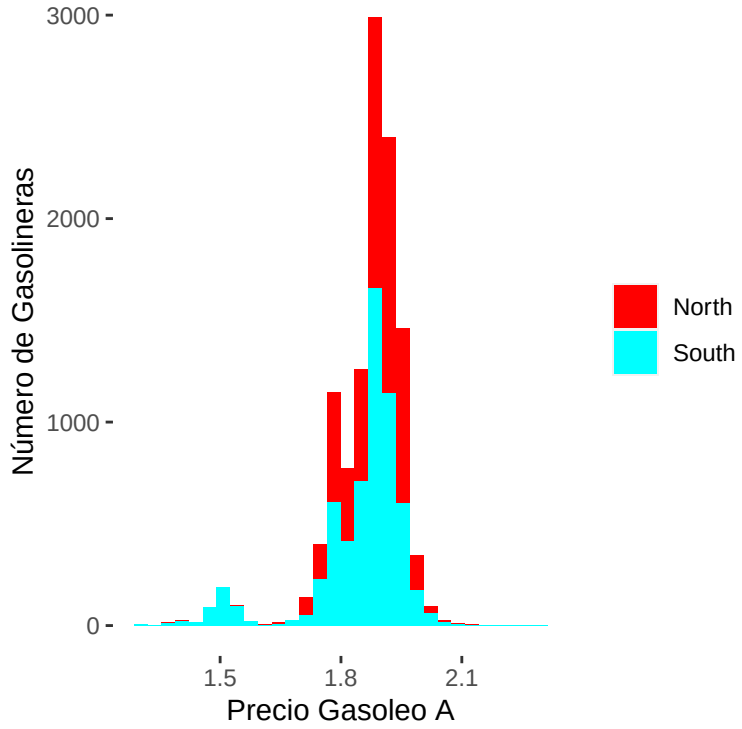
- Tomaremos todas las latitudes por encima de 40,3093 como el norte, debido a que se considera el centro de la latitud de la península.
- Tomaremos todas las longitudes por encima de -3,6842 como el este, debido a que se considera el centro de la longitud de la península.
- Solo tomaremos las gasolineras que tienen Gasoleo_A.

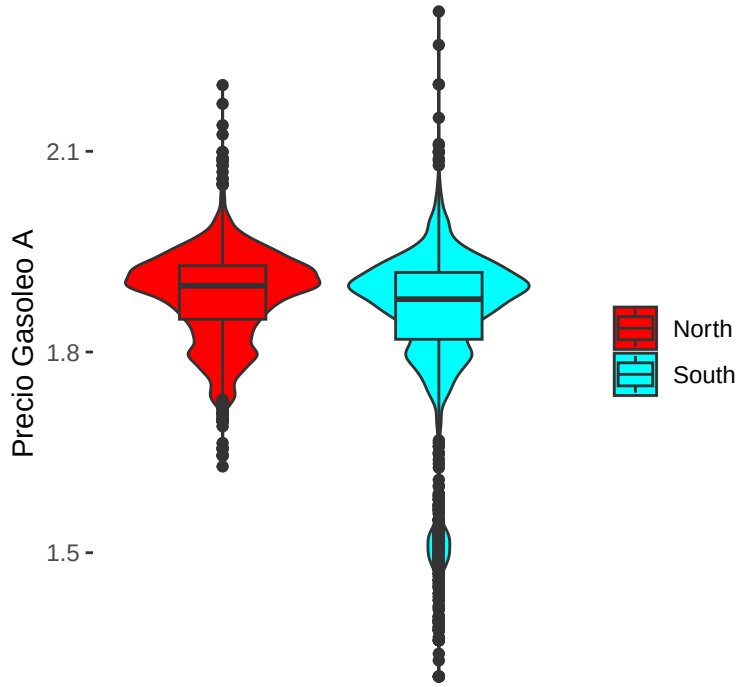
A continuación, creamos el dataset con las condiciones anteriores. Primero obtenemos las variables *latitud*, *longitud* y *Gasoleo_A*. Despues, seleccionamos las gasolineras con gasóleo A. Finalmente, en función su valor cambiamos la latitud por el valor North o South y lo mismo con la longitud, pero con los valores East y West.

latitud	longitud	Gasoleo_A
South	East	1.869
South	East	2.040
South	East	1.839

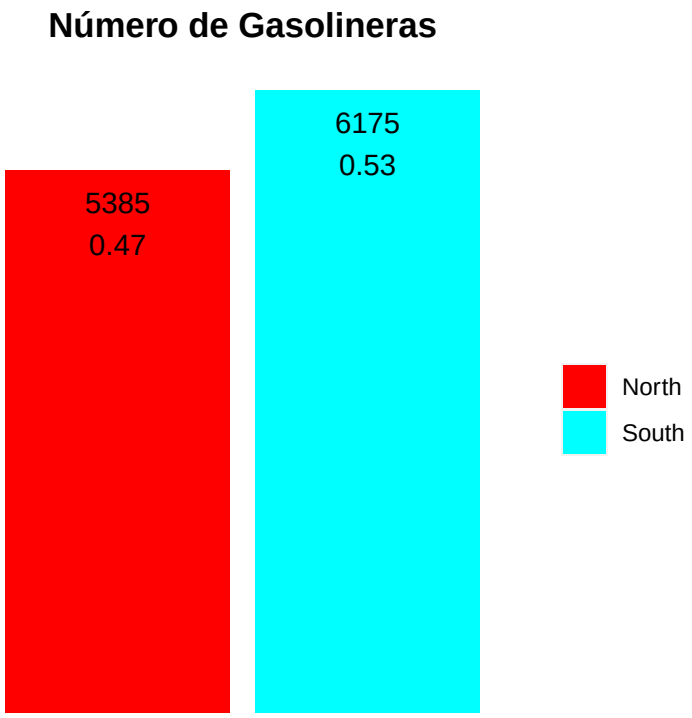
3.3.2 Análisis de la latitud

3.3.2.1 Análisis visual de la latitud Comprobamos visualmente la distribución del precio del Gasóleo A en función de si es el norte o el sur y la frecuencia de gasolinas en el norte y en el sur.






```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
```



Visualmente no vemos una gran diferencia en la distribución del precio en función de la latitud. Sin embargo, parece que la media de los precios es inferior en el sur. Así mismo, la proporción de gasolineras es muy similar, vemos que el 47% de las gasolineras están en el sur y el 53 en el norte.

3.3.2.2 Hipotesis nula y alternativa para la latitud La hipótesis nula es que la media del precio del Gasóleo A del norte es igual a la del sur y la alternativa es que son distintas.

$$H_0 : \mu_N = \mu_S$$

$$H_1 : \mu_N \neq \mu_S$$

3.3.2.3 Aplicación del test para la latitud Sabemos que son dos muestras independientes y por el teorema del límite central asumimos normalidad, ya que usamos la media de las muestras y estas son lo suficientemente grandes (n > 30). Además, no conocemos las varianzas de ambas muestras.

Ahora comprobamos si la varianza es la misma en el norte y en el sur.

```
##
##  F test to compare two variances
##
## data:  north and south
## F = 0.29815, num df = 5384, denom df = 6174, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2831403 0.3139783
## sample estimates:
## ratio of variances
##          0.2981471
```

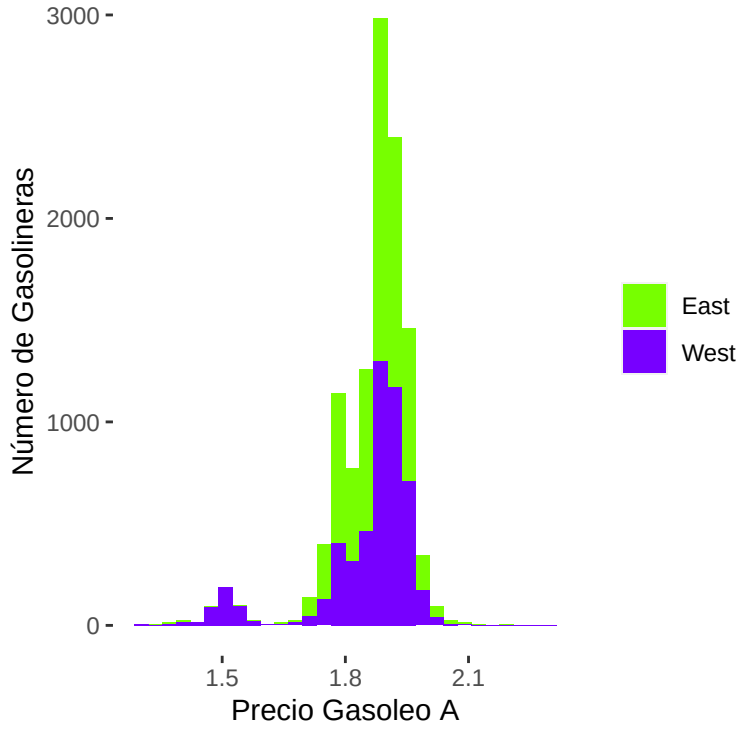
Al realizar el test de varianza vemos que las varianzas son distintas, ya que $p - value < 0.05$. Por lo que realizamos el test especificando que las varianzas son diferentes.

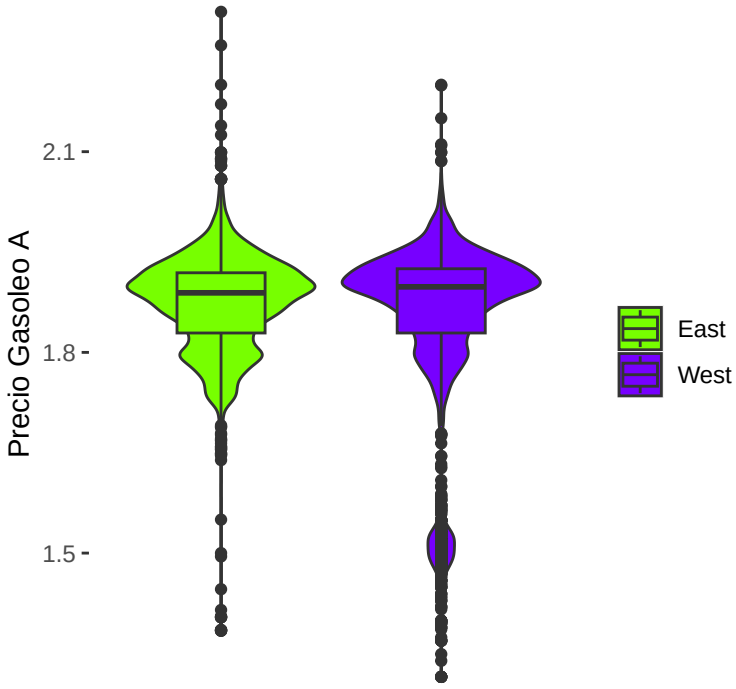
```
##
##  Welch Two Sample t-test
##
## data:  north and south
## t = 20.319, df = 9803.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03206140 0.03890783
## sample estimates:
## mean of x mean of y
##  1.884182  1.848697
```

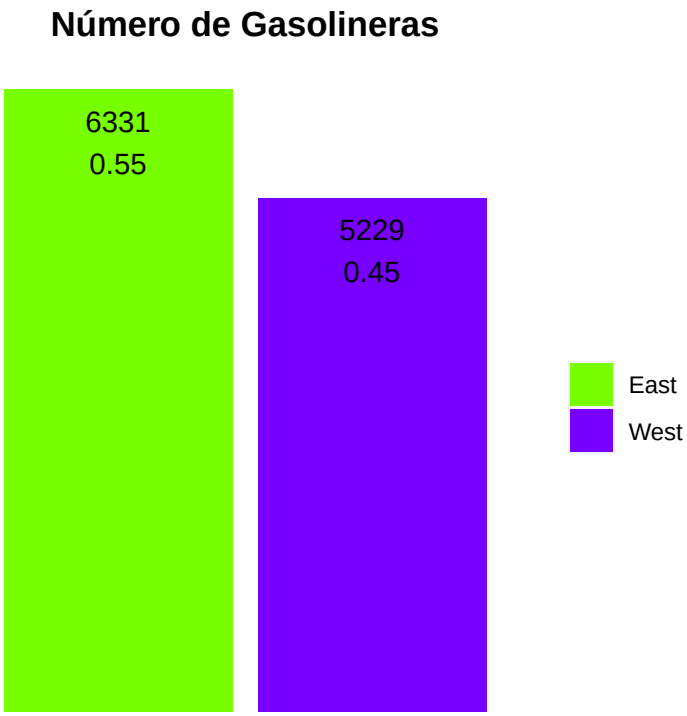
El valor p obtenido es menor que el nivel de significancia (0.05) y, por lo tanto, se puede rechazar la hipótesis nula de igualdad de medias de precio en las gasolineras entre el norte y el sur.

3.3.3 Análisis de la longitud

3.3.3.1 Análisis visual de la longitud Comprobamos visualmente la distribución del precio del Gasóleo A en función de si es el este o el oeste y la frecuencia de gasolinas en el este y en el oeste.







Visualmente, al igual que con la latitud, no vemos una gran diferencia en la distribución del precio en función de la longitud. Sin embargo, parece que la media de los precios es inferior en el este. Así mismo, la proporción de gasolineras es muy similar, vemos que el 55% de las gasolineras están en el este y el 45% en el oeste.

3.3.3.2 Hipotesis nula y alternativa para la longitud La hipótesis nula es que la media del precio del Gasóleo A del este es igual a la del oeste y la alternativa que es distinta.

$$H_0 : \mu_E = \mu_W$$

$$H_1 : \mu_E \neq \mu_W$$

3.3.3.3 Aplicación del test para la latitud Sabemos que son dos muestras independientes y por el teorema del límite central asumimos normalidad, ya que usamos la media de las muestras y estas son lo suficientemente grandes (n > 30). Además, no conocemos las varianzas de ambas muestras.

Ahora comprobamos si la varianza es la misma en el este y en el oeste.

```
##
##  F test to compare two variances
##
## data:  east and west
## F = 0.32031, num df = 6330, denom df = 5228, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3041170 0.3373173
## sample estimates:
## ratio of variances
##           0.320308
```

Al realizar el test de varianza vemos que las varianzas son distintas, ya que $p - value < 0.05$.

```
##
##  Welch Two Sample t-test
##
## data:  east and west
## t = 10.4, df = 7903.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01628152 0.02384439
## sample estimates:
## mean of x mean of y
##  1.874302  1.854239
```

El valor p obtenido es menor que el nivel de significancia (0.05) y, por lo tanto, se puede rechazar la hipótesis nula de igualdad de medias de precio en las gasolineras entre el este y el oeste.

3.3.4 Conclusiones Tras los resultados obtenidos podemos asegurar que hay diferencia de precios entre norte y sur y entre este y oeste. Para futuras líneas de estudio se podría comprobar que latitudes tienen menor precio y realizar una comparación por pares, por ejemplo, comparando el noroeste con el sudeste.

3.4 ¿Podríamos predecir si la gasolinera esta en el norte sabiendo el precio del Gasoleo A? Y si esta en el este?

3.4.1 Modelo de regresión logística para la latitud.

3.4.1.1 Generación de los conjuntos de entrenamiento y de test Primero creamos un juego de datos que contenga solo el precio y la clasificación de la latitud.

latitud	Gasoleo_A
0	1.869

latitud	Gasoleo_A
0	2.040
0	1.839

Dividimos los datos en entrenamiento y prueba.

3.4.1.2 Creación del modelo para la latitud Entrenamos el modelo con los datos de entrenamiento.

```
##
## Call:
## glm(formula = latitud ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8985  -1.1492  -0.6539   1.1887   1.6423
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.6620     0.4713  -16.26   <2e-16 ***
## Gasoleo_A     4.0207     0.2513   16.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12771  on 9247  degrees of freedom
## Residual deviance: 12472  on 9246  degrees of freedom
## AIC: 12476
##
## Number of Fisher Scoring iterations: 4
```

3.4.1.3 Comprobación de la bondad del modelo Comprobamos la bondad del modelo mediante una matriz de confusión. Para ello usamos los datos de test y predecimos la longitud con los precios de test.

```
## Confusion Matrix and Statistics
##
##
## pred   0   1
##    0 851 635
##    1 362 464
##
##                Accuracy : 0.5688
##                  95% CI : (0.5483, 0.5891)
##     No Information Rate : 0.5247
##     P-Value [Acc > NIR] : 1.141e-05
##
##                   Kappa : 0.1252
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4222
##             Specificity : 0.7016
```

```
##          Pos Pred Value : 0.5617
##          Neg Pred Value : 0.5727
##              Prevalence : 0.4753
##          Detection Rate : 0.2007
##    Detection Prevalence : 0.3573
##       Balanced Accuracy : 0.5619
##
##        'Positive' Class : 1
##
```

La precisión no es muy buena, siendo capaz el modelo de clasificar solo el 56,88% de los datos. Además, parece que clasifica mejor las gasolineras del sur, habiendo clasificado correctamente el 70,16% de las gasolineras del sur. Así mismo, solo ha clasificado correctamente el 42,22% de las gasolineras del norte.

3.4.2 Modelo de regresión logística para la longitud.

3.4.2.1 Generación de los conjuntos de entrenamiento y de test Primero creamos un juego de datos que contenga solo el precio y la clasificación de la longitud.

longitud	Gasoleo_A
1	1.869
1	2.040
1	1.839

Dividimos los datos en entrenamiento y prueba.

3.4.2.2 Creación del modelo para la latitud Entrenamos el modelo con los datos de entrenamiento.

```
##
## Call:
## glm(formula = latitud ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8985  -1.1492  -0.6539   1.1887   1.6423
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.6620     0.4713  -16.26   <2e-16 ***
## Gasoleo_A     4.0207     0.2513   16.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12771  on 9247  degrees of freedom
## Residual deviance: 12472  on 9246  degrees of freedom
## AIC: 12476
##
## Number of Fisher Scoring iterations: 4
```

3.4.2.3 Comprobación de la bondad del modelo Comprobamos la bondad del modelomediante una matriz de confusión. Para ello usamos los datos de test y predecimos la longitud con los precios de test.

```
## Confusion Matrix and Statistics
##
##
## pred    0    1
##    0  145  102
##    1  895 1170
##
##                Accuracy : 0.5688
##                  95% CI : (0.5483, 0.5891)
##     No Information Rate : 0.5502
##     P-Value [Acc > NIR] : 0.03767
##
##                   Kappa : 0.0637
##
##  Mcnemar's Test P-Value : < 2e-16
##
##             Sensitivity : 0.9198
##             Specificity : 0.1394
##          Pos Pred Value : 0.5666
##          Neg Pred Value : 0.5870
##              Prevalence : 0.5502
##          Detection Rate : 0.5061
##    Detection Prevalence : 0.8932
##       Balanced Accuracy : 0.5296
##
##        'Positive' Class : 1
##
```

La precisión no es muy buena, siendo capaz el modelo de clasificar solo el 56,88% de los datos. Además, parece que clasifica muy bien las gasolineras del este, habiendo clasificado correctamente el 91,98% de las gasolineras del este. Sin embargo, solo ha clasificado correctamente el 13,94% de las gasolineras del oeste.

3.4.3 Conclusiones Para ambos modelos, ya sea buscando clasificar la longitud o la latitud, hemos obtenido la misma precisión (0.5688), la cual no es buena. Así mismo, el modelo para predecir la longitud es capaz de clasificar mejor las gasolineras del sur, mientras que el modelo para la latitud clasificar mejor las gasolineras del este.

4. Tabla de contribuciones

Contribuciones	*Firma*
Investigación previa	Imanol Miguez Quintela, Ivan Cuevas Ortin
Redacción de las respuestas	Imanol Miguez Quintela, Ivan Cuevas Ortin
Desarrollo del código	Imanol Miguez Quintela, Ivan Cuevas Ortin
Participación en el vídeo	Imanol Miguez Quintela, Ivan Cuevas Ortin