PRÁCTICA – DATA MINING

0.- OBJETIVO

El objetivo de la práctica es abordar un problema de data mining realista siguiendo la metodología y buenas prácticas explicadas durante las clases teóricas.

Datos: La fuente recoge la información actualizada diariamente de los precios de los combustibles en España. Se muestran las gasolineras y se puede consultar el precio, horario, marcha y fecha de actualización del dato. Los datos sobre gasolineras proceden del Ministerio de Energía, Turismo y Agenda Digital.

El objetivo general es **predecir el precio de la gasolina 95 en España a nivel de provincia, desde el punto de vista del consumidor.**

Modelización: GLM y su comparativa con Regresión y Redes Neuronales

1.- ANÁLISIS PREVIO DE LOS DATOS

Analizaremos la variable dependiente (*y*), es decir, el objetivo de estudio, para saber cómo debemos tratarla o prepararla.

El primer paso es la identificación y formulación del problema de negocio, esto es:

- Identificación del tipo de problema: análisis predictivo
- Objetivo: predecir el precio de la gasolina 95
- Nombre de la variable objetivo: precio_gasolina_95
- Tipo de variable objetivo: Continua

El objetivo general es **predecir el precio de la gasolina 95 en España a nivel de provincia, desde el punto de vista del consumidor.**

El dataset del que partimos tiene las siguientes variables y características:

target	tipo	descripción
precio_gasolina_95	numérico	precio de la gasolina 95 octanos en España

variable	tipo	descripción
Fecha	numérica	fecha de extracción de la información
Provincia	alfanumérica	provincia donde está ubicada la gasolinera
Municipio	alfanumérica	municipio donde está ubicada la gasolinera
Localidad	alfanumérica	localidad donde está ubicada la gasolinera

Codigo_postal	alfanumérica	código postal
Direccion	alfanumérica	dirección de la gasolinera
Margen	alfanumérica	margen (ver siglas*)
Longitud	alfanumérica	longitud donde está ubicada la gasolinera
Latitud	alfanumérica	latitud donde está ubicada la gasolinera
Precio_gasolina_95	numérica	precio gasolina 95 octanos
Precio_gasoleo_A	numérica	precio gasóleo A
Precio_gasoleo_B	numérica	precio gasóleo B
Precio_bioetanol	numérica	precio bioetanol
Precio_nuevo_gasoleo_A	numérica	precio nuevo gasóleo A
Precio_biodiesel	numérica	precio biodiesel
porcentaje_ester_metilico	numérica	porcentaje éster metílico
porcentaje_bioalcohol	numérica	porcentaje bioalcohol
Precio_gas_natural_comprimido	numérica	precio gas natural comprimido
Precio_gas_natural_licuado	numérica	precio gas natural licuado
Precio_gases_licuados_del_petroleo	numérica	precio gases licuados del petróleo
Rotulo	alfanumérica	rótulo de la gasolinera
Tipo_venta	alfanumérica	tipo venta (ver siglas*)
Rem	alfanumérica	procedencia (ver siglas*)
Horario	alfanumérica	horario de atención

Fuente:
https://datos.gob.es/es/catalogo/e04990201-precio-de-carburantes-en-las-gasolineras-espanolas

Siglas (*):

- Margen: D: Derecho, I: Izquierdo, N: No aplica
- Tipo venta: P: Venta al público en general, R: Venta restringida a socios o cooperativistas
- Rem: OM: Datos procedentes del operador mayorista, dm: Datos procedentes del distribuidor minorista.

El siguiente paso es la exploración de datos, que consiste en un análisis previo donde se identificarán inconsistencias, duplicidades, missings, outliers y se realizará un análisis estadístico básico.

```
data gasolina95;
  set lib_in.gasolina_in_spain;
run;
```

Al abrir el dataset observamos que partimos de **72.113 observaciones y 25 variables** (incluida la variable objetivo).

1.1.- ANALISIS DE LA VARIABLE TARGET

Analisis de missing en la variable target

Lo siguiente que haré es comprobar si mi variable target tiene valores missings y de ser así eliminaré esas observaciones. Para ello utilizo el **proc means** con la opción **nmiss**

```
proc means data=gasolina95 nmiss;
  var Precio_gasolina_95;
run;
```

Procedimiento MEANS

Analysis Variable : Precio_gasolina_95
N Miss
2818

Se observa que tenemos **2818 valores missing**. Elimino estas observaciones:

```
data gasolina95;
  set gasolina95;
    if Precio_gasolina_95=. then delete;
run;
```

```
NOTE: There were 72113 observations read from the data set
WORK.GASOLINA95.
NOTE: The data set WORK.GASOLINA95 has 69295 observations and 25
variables.
```

Nos quedan por tanto: **69.295 observaciones y 25 variables**

Análisis de duplicados en base a la clave primaria del fichero

Establecemos que la clave primaria del fichero (primary key) será la ubicación de las gasolineras con sus coordenadas (longitud y latitud) y la fecha de extracción.

	Fecha	Longitud	Latitud	Codigo_postal	Provincia
1	28/10/2019	-2,519194	42,842917	01240	ÁLAVA
2	28/10/2019	-2,509361	42,846028	01240	ÁLAVA
3	28/10/2019	-2,989111	43,044333	01468	ÁLAVA
4	28/10/2019	-2,967611	43,031889	01450	ÁLAVA
5	28/10/2019	-2,477917	42,753194	01120	ÁLAVA
6	28/10/2019	-2,639528	42,941417	01510	ÁLAVA

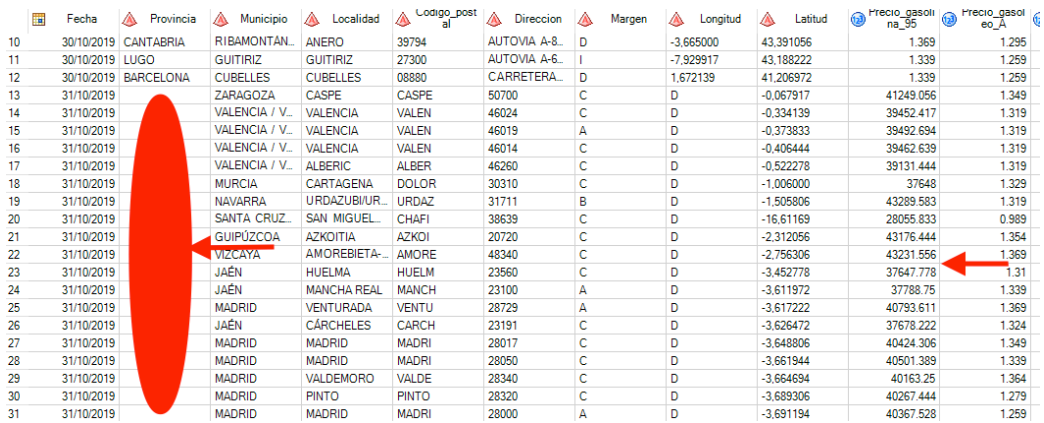
Deberíamos tener en una única ubicación la gasolinera por día en el que se recogió el dato, luego de estar bien la base de la información no debería contener duplicidad de registros en la primary key de la base de la información. Usaremos el **proc sort** para asegurar que no haya registros duplicados.

```
proc sort data=gasolina95 nodupkey dupout=duplicados;
  by fecha Longitud Latitud;
run;
```

Obtenemos que **existen duplicidades** en las observaciones, como se puede ver en la log de SAS Studio:

```
NOTE: There were 69295 observations read from the data set
WORK.GASOLINA95.
NOTE: 91 observations with duplicate key values were deleted.
NOTE: The data set WORK.DUPLICADOS has 91 observations and 25
variables.
NOTE: The data set WORK.GASOLINA95 has 69204 observations and 25
variables.
```

Además, analizando la salida “DUPLICADOS” se comprueba que hay una parte de la tabla de datos que parece desplazada una columna a la derecha, quedando la columna Provincia vacía. Se podría tratar de recuperar esos datos desplazando columnas pero hay datos que se han perdido, por ejemplo en la columna Margen parece que está la información de la dirección pero solo aparece un carácter. No tendríamos además la información de la última columna que es la del horario.



	Fecha	Provincia	Municipio	Localidad	Codigo_postal	Direccion	Margen	Longitud	Latitud	Precio_gasolina_95	Precio_gasoleo_A
10	30/10/2019	CANTABRIA	RIBAMONTÁN...	ANERO	39794	AUTOVIA A-8...	D	-3,665000	43,391056	1.369	1.295
11	30/10/2019	LUGO	GUITIRIZ	GUITIRIZ	27300	AUTOVIA A-6...	I	-7,929917	43,188222	1.339	1.259
12	30/10/2019	BARCELONA	CUBELLES	CUBELLES	08880	CARRETERA...	D	1,672139	41,206972	1.339	1.259
13	31/10/2019		ZARAGOZA	CASPE	CASPE	50700	C	D	-0,067917	41249.056	1.349
14	31/10/2019		VALENCIA / V...	VALENCIA	VALEN	46024	C	D	-0,334139	39452.417	1.319
15	31/10/2019		VALENCIA / V...	VALENCIA	VALEN	46019	A	D	-0,373833	39492.694	1.319
16	31/10/2019		VALENCIA / V...	VALENCIA	VALEN	46014	C	D	-0,406444	39462.639	1.319
17	31/10/2019		VALENCIA / V...	ALBERIC	ALBER	46260	C	D	-0,522278	39131.444	1.319
18	31/10/2019		MURCIA	CARTAGENA	DOLOR	30310	C	D	-1,006000	37648	1.329
19	31/10/2019		NAVARRA	URDAZUBI/UR...	URDAZ	31711	B	D	-1,505806	43289.583	1.319
20	31/10/2019		SANTA CRUZ...	SAN MIGUEL...	CHAFI	38639	C	D	-16,61169	28055.833	0.989
21	31/10/2019		GUIPÚZCOA	AZKOITIA	AZKOI	20720	C	D	-2,312056	43176.444	1.354
22	31/10/2019		VIZCAYA	AMOREBIETA-...	AMORE	48340	C	D	-2,756306	43231.556	1.369
23	31/10/2019		JAÉN	HUELMA	HUELM	23560	C	D	-3,452778	37647.778	1.31
24	31/10/2019		JAÉN	MANCHA REAL	MANCH	23100	A	D	-3,611972	37788.75	1.339
25	31/10/2019		MADRID	VENTURADA	VENTU	28729	A	D	-3,617222	40793.611	1.369
26	31/10/2019		JAÉN	CÁRCHELES	CARCH	23191	C	D	-3,626472	37678.222	1.324
27	31/10/2019		MADRID	MADRID	MADRI	28017	C	D	-3,648806	40424.306	1.349
28	31/10/2019		MADRID	MADRID	MADRI	28050	C	D	-3,661944	40501.389	1.339
29	31/10/2019		MADRID	VALDEMORO	VALDE	28340	C	D	-3,664694	40163.25	1.364
30	31/10/2019		MADRID	PINTO	PINTO	28320	C	D	-3,689306	40267.444	1.279
31	31/10/2019		MADRID	MADRID	MADRI	28000	A	D	-3,691194	40367.528	1.259

Por tanto eliminamos duplicados y nos quedamos con **69.204 observaciones y 25 variables**

Análisis de estadísticos básicos

Pasamos ahora a analizar los estadísticos básicos de nuestra variable objetivo, como es de tipo continuo utilizo **proc means**:

```
proc means data=gasolina95;
  var Precio_gasolina_95;
run;
```

Procedimiento MEANS

Analysis Variable : Precio_gasolina_95				
N	Media	Desv. est.	Mínimo	Máximo
69204	5860.07	14118.17	0.8690000	43731.81

Este resultado es bastante raro. El máximo no tiene ningún sentido y tiene que ver con lo mismo que comentábamos en el punto anterior. Parte de la tabla tiene las columnas como desplazadas hacia la derecha de forma que en el Precio de la gasolina de 95 aparecen datos correspondientes a la coordenada geográfica de latitud.

Vuelvo a visualizar el dataset actual y me ayudo de las opciones de Filtro y Orden y Clasificación para poder ver la variable Precio_gasolina_95 ordenada en orden ascendente. Se puede ver que el valor más pequeño es 0,869 que coincide claro está con lo que nos mostró el proc means y si me desplazo hacia abajo se observa el salto debido al error en la tabla, pasa de 1.489 a 277708.333

	Precio_gasolina_95
58963	1.434
58964	1.434
58965	1.434
58966	1.439
58967	1.439
58968	1.489
58969	1.489
58970	1.489
58971	1.489
58972	1.489
58973	1.489
58974	27705.333
58975	27751.944
58976	27753.306
58977	27760.861
58978	27763.889

Procedo a eliminar todas esas observaciones erróneas

```
data gasolina95;
  set gasolina95;
  if Precio_gasolina_95 > 20000 then delete;
run;
```

```
NOTE: There were 69204 observations read from the data set
WORK.GASOLINA95.
NOTE: The data set WORK.GASOLINA95 has 58973 observations and 25
variables.
```

Y me quedan finalmente **58.973 observaciones**

Ejecuto de nuevo el **proc means** para comprobar que ahora si obtengo valores más normales:

Procedimiento MEANS

Analysis Variable : Precio_gasolina_95				
N	Media	Desv. est.	Mínimo	Máximo
58973	1.2911433	0.0776148	0.8690000	1.4890000

Análisis de outliers, missings e incongruencias

Realizo ahora un análisis más profundo utilizando el proc univariate:

```
proc univariate data=gasolina95;
  var Precio_gasolina_95;
run;
```

Las tablas complementarias al anterior análisis y más importantes son las siguientes:

Momentos			
N	58973	Sumar pesos	58973
Media	1.2911433	Observ suma	76142.594
Desviación std	0.07761481	Varianza	0.00602406
Asimetría	-2.3456239	Curtosis	6.52443503
SC no corregida	98666.2511	SC corregida	355.250796
Coef. variación	6.01132425	Media error std	0.00031961

Cuantiles (Definición 5)	
Nivel	Cuantil
100% Máx	1.489
99%	1.389
95%	1.365
90%	1.355
75% Q3	1.339
50% Mediana	1.309
25% Q1	1.269
10%	1.217
5%	1.153
1%	0.977
0% Mín	0.869

Observaciones extremas			
Inferior		Superior	
Valor	Obs	Valor	Obs
0.869	22239	1.489	13053
0.869	22086	1.489	22902
0.869	12390	1.489	32719
0.869	12235	1.489	42533
0.869	2546	1.489	52344

Se observa que no hay valores missing. No hay valores 2.5 por encima de la media, por tanto no tenemos valores outliers o anómalos. Tampoco se detectan incongruencias entre los datos.

1.2.- SEPARACION DEL RESTO DE VARIABLES

Una vez analizada la variable objetivo pasamos a ver el resto de variables y lo primero es separarlas en variables de tipo categórico y de tipo analítico

Variables Categóricas	Variables Analíticas
Fecha	Precio_gasolina_95
Provincia	Precio_gasoleo_A
Municipio	Precio_gasoleo_B
Localidad	Precio_bioetanol
Codigo_postal	Precio_nuevo_gasoleo_A
Direccion	Precio_biodiesel
Margen	Porcentaje_ester_metilico
Longitud	Porcentaje_bioalcohol

Latitud	Precio_gas_natural_comprimido
Rotulo	Precio_gas_natural_licuado
Tipo_venta	Precio_gases_licuados_del_petrol
Rem	
Horario	

1.3.- ANALISIS DE VARIABLES CATEGÓRICAS

Las variables longitud y latitud determinan la posición exacta de la gasolinera. Tenemos además la información del código postal que permitiría agrupar las gasolineras por zonas por lo que el código postal va a dar más información a nuestro modelo GLM que las coordenadas. **Eliminamos la variables longitud y latitud del estudio**
Por otro lado, el código postal en España es una secuencia de 5 números, los dos primeros hacen referencia a la provincia y los otros 3 al municipio por lo que como ya tenemos las variables Provincia y Municipio vamos a **eliminar el código postal del estudio**

```
data gasolina95;
  set gasolina95;
  drop longitude latitude codigo_postal;
run;
```

Análisis de la variable fecha

Analizamos esta variable con un procedimiento de frecuencias

```
proc freq data=gasolina95;
  tables Fecha;
run;
```

Procedimiento FREQ

Fecha de extracción

Fecha	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
28/10/2019	9847	16.70	9847	16.70
29/10/2019	9859	16.72	19706	33.42
30/10/2019	9829	16.67	29535	50.08
01/11/2019	9813	16.64	39348	66.72
02/11/2019	9807	16.63	49155	83.35
03/11/2019	9818	16.65	58973	100.00

No se observan missing ni outliers. Está muy balanceada, teniendo todos los días una distribución similar. **No hay información del jueves 31/10/2019**

Análisis de las variables Provincia, Municipio y Localidad

Se analizan con el mismo procedimiento de frecuencias y no se observan missings ni outliers. No se incluyen tablas al no aportar información relevante.
Como el objetivo es conocer el precio a nivel de provincia voy a eliminar del estudio Municipio y Localidad. De esta forma quito los más granulares

```
data gasolina95;
  set gasolina95;
  drop Municipio Localidad;
run;
```

Análisis de la variable Dirección

La dirección exacta de la gasolinera no nos da grandes ventajas, pasa igual que con las coordenadas geográficas que hemos eliminado. Sin embargo, es posible que podamos extraer información interesante sobre la situación: si es una calle, una autovía….. Si los clientes tienen más propensión a utilizar una gasolinera de una autovía seguramente el precio sea mayor en estas gasolineras. Por tanto, es posible que el precio pueda depender en cierta medida de la ubicación de la gasolinera. Se crea por tanto una variable Zona que contendrá la ubicación de la gasolinera y que extraeremos de la dirección antes de eliminarla.
El código que he utilizado para la creación de la variable Zona que se extrae desde Dirección es el siguiente:

```
data gasolina95;
  set gasolina95;
  format zona $50.;
  if findw(upcase(direccion), "AEROPUERTO")OR
      findw(upcase(direccion), "AÉROPUERTO") then
      zona="AEROPUERTO";
  else if findw(upcase(direccion), "PLAZA") OR
      findw(upcase(direccion), "PZA.") then zona="PLAZA";
  else if findw(upcase(direccion), "AUTOVIA")OR
      findw(upcase(direccion), "CR A")OR findw(upcase(direccion),
      "AUTOVÍA")OR findw(upcase(direccion), "AU") then
      zona="AUTOVIA";
  else if findw(upcase(direccion), "CARRETERA") OR
      findw(upcase(direccion), "KM") OR findw(upcase(direccion),
      "CARRER")OR findw(upcase(direccion), "CR")OR
      findw(upcase(direccion), "CTRA")OR findw(upcase(direccion),
      "CTRA.")OR findw(upcase(direccion), "CRTA.")OR
      findw(upcase(direccion), "CRA.")OR findw(upcase(direccion),
      "P.K.") OR findw(upcase(direccion), "PK")OR
      findw(upcase(direccion), "CARRERA")then zona="CARRETERA";
  else if findw(upcase(direccion), "CALLE") OR
      findw(upcase(direccion), "RUA") OR findw(upcase(direccion),
      "BARRIO")OR findw(upcase(direccion), "CL") OR
      findw(upcase(direccion), "C/")OR findw(upcase(direccion),
      "VIA")OR findw(upcase(direccion), "VÍA")OR
      findw(upcase(direccion), "LUGAR") OR
      findw(upcase(direccion), "URBANIZACIÓN")OR
      findw(upcase(direccion), "BARRIADA")OR
      findw(upcase(direccion), "URB.")OR findw(upcase(direccion),
      "CALZADA")OR findw(upcase(direccion), "RAMBLA")OR
      findw(upcase(direccion), "CUESTA")OR
      findw(upcase(direccion), "PASAJE")OR
      findw(upcase(direccion), "PASSATGE")OR
      findw(upcase(direccion), "UR")OR findw(upcase(direccion),
      "BDA.")OR findw(upcase(direccion), "PARTIDA")OR
      findw(upcase(direccion), "BULEVAR")OR
      findw(upcase(direccion), "BDA.")OR findw(upcase(direccion),
      "CALLEJA")then zona="CALLE";
  else if findw(upcase(direccion), "AVENIDA") OR
      findw(upcase(direccion), "AVD") OR findw(upcase(direccion),
      "AVINGUDA") OR findw(upcase(direccion), "AV.") OR
      findw(upcase(direccion), "AVD.") OR
      findw(upcase(direccion), "AVDA.")OR
```

```
      findw(upcase(direccion), "AV")OR findw(upcase(direccion),
      "AVDA")then zona="AVENIDA";
  else if findw(upcase(direccion), "POLIGONO") OR
      findw(upcase(direccion), "P.I.")OR findw(upcase(direccion),
      "POLG")OR findw(upcase(direccion), "POLÍGONO")OR
      findw(upcase(direccion), "INDUSTRIAL") OR
      findw(upcase(direccion), "POL.IND.")OR
      findw(upcase(direccion), "IND.")OR findw(upcase(direccion),
      "POLIG.")OR findw(upcase(direccion), "POL.")OR
      findw(upcase(direccion), "PGNO.")OR
      findw(upcase(direccion), "PG")then zona="POLIGONO";
  else if findw(upcase(direccion), "PASEO")OR
      findw(upcase(direccion), "PASSEIG") then zona="PASEO";
  else if findw(upcase(direccion), "CAMINO")OR
      findw(upcase(direccion), "CAMI")OR findw(upcase(direccion),
      "VEREDA")OR findw(upcase(direccion), "PARAJE")OR
      findw(upcase(direccion), "CAÑADA")OR
      findw(upcase(direccion), "PARATGE") then zona="CAMINO";
  else if findw(upcase(direccion), "ROTONDA") OR
      findw(upcase(direccion), "GLORIETA") then zona="ROTONDA";
  else if findw(upcase(direccion), "MUELLE") then zona="MUELLE";
  else if findw(upcase(direccion), "ESTACION") then
      zona="ESTACION";
  else if findw(upcase(direccion), "TRAVESIA") OR
      findw(upcase(direccion), "TRAVESÍA")OR
      findw(upcase(direccion), "TRAVESSERA")then zona="TRAVESIA";
  else if findw(upcase(direccion), "RONDA") OR
      findw(upcase(direccion), "RDA.")OR findw(upcase(direccion),
      "RD") then zona="RONDA";
  else if findw(upcase(direccion), "CRUCE") then zona="CRUCE";
  else if findw(upcase(direccion), "COMERCIAL") then
      zona="CENTRO COMERCIAL";
  else if findw(upcase(direccion), "PARQUE") OR
      findw(upcase(direccion), "PARC")then zona="PARQUE";
  else if findw(upcase(direccion), "AP")OR
      findw(upcase(direccion), "AT") then zona="AUTOPISTA";
  else if findw(upcase(direccion), "PARQUE EMPRESARIAL") then
      zona="PARQUE EMPRESARIAL";
  else zona='missings';


run;
```

Tras crear la variable Zona realizo un estudio de frecuencias:

```
proc freq data=gasolina95;
  tables zona;
run;
```

Procedimiento FREQ

zona	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
AEROPUERTO	126	0.21	126	0.21
AUTOPISTA	42	0.07	168	0.28
AUTOVIA	3849	6.53	4017	6.81
AVENIDA	9582	16.25	13599	23.06
CALLE	12582	21.34	26181	44.39
CAMINO	621	1.05	26802	45.45
CARRETERA	27888	47.29	54690	92.74
CENTRO COMERCIAL	54	0.09	54744	92.83
CRUCE	18	0.03	54762	92.86
MUELLE	41	0.07	54803	92.93
PARQUE	48	0.08	54851	93.01
PASEO	482	0.82	55333	93.83
PLAZA	503	0.85	55836	94.68
POLIGONO	1901	3.22	57737	97.90
RONDA	248	0.42	57985	98.32
ROTONDA	96	0.16	58081	98.49
TRAVESIA	96	0.16	58177	98.65
missings	796	1.35	58973	100.00

Se observa que hay 796 observaciones que no se han podido clasificar correctamente. El porcentaje es de **1.35** por lo que se podría eliminar, pero observando esos registros:

	Direccion	zona
58181	SILVANO, 88-90	missings
58182	HNOS. GARCIA NOBLEJ...	missings
58183	ACCESO PLAN PARCIAL...	missings
58184	BENITO PEREZ GALDOS...	missings
58185	C/Alcalde de Lavadores	missings
58186	ESTRADA PORRIÑO-GO...	missings
58187	POBLADURA DEL VALLE...	missings
58188	ESTRADA ESTRADA PO...	missings
58189	XOAN CARLOS I S/N	missings
58190	SECTOR ARROYO ENME...	missings
58191	SEVILLA, 133	missings
58192	PEDRO FERNANDEZ VA...	missings

se ve que la mayoría se corresponden con CALLES y zonas urbanas, así que vamos **a asignarles a los missings la clasificación de CALLE y eliminamos la variable dirección.** Para ello solo tenemos que modificar la última instrucción del código anterior:

```
/*else zona='missings';*/
      else zona="CALLE";
      drop direccion;
run;
```

Así queda ahora la tabla de frecuencias donde vemos que ya no tenemos missings:

Procedimiento FREQ

zona	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
AEROPUERTO	126	0.21	126	0.21
AUTOPISTA	42	0.07	168	0.28
AUTOVIA	3849	6.53	4017	6.81
AVENIDA	9582	16.25	13599	23.06
CALLE	13378	22.68	26977	45.74
CAMINO	621	1.05	27598	46.80
CARRETERA	27888	47.29	55486	94.09
CENTRO COMERCIAL	54	0.09	55540	94.18
CRUCE	18	0.03	55558	94.21
MUELLE	41	0.07	55599	94.28
PARQUE	48	0.08	55647	94.36
PASEO	482	0.82	56129	95.18
PLAZA	503	0.85	56632	96.03
POLIGONO	1901	3.22	58533	99.25
RONDA	248	0.42	58781	99.67
ROTONDA	96	0.16	58877	99.84
TRAVESIA	96	0.16	58973	100.00

Se ve que la mayoría de las gasolineras las tenemos en carreteras.

Análisis de la variable Margen

Analizamos esta variable con un procedimiento de frecuencias:

```
proc freq data=gasolina95;
  tables Margen;
run;
```

Procedimiento FREQ

Margen	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
D	29401	49.86	29401	49.86
I	15592	26.44	44993	76.29
N	13980	23.71	58973	100.00

La variable Margen no presenta missings ni outliers. El descriptivo N es de no aplica y se refiere a zonas urbanas, ya que la variable se utiliza para recoger si la gasolinera está en margen derecho o izquierdo de carreteras, autovías… La distribución no está balanceada ya que se ve que hay una mayor probabilidad de que la gasolinera esté en el margen derecho que en el izquierdo.

Análisis de la variable Rotulo

En la variable Rotulo tenemos la marca de la gasolinera. Queremos diferenciar los precios por marca. Voy a crear una variable Cartel donde iré clasificando las gasolineras por marca obtenida desde la variable Rotulo. Esto es necesario porque por ejemplo las gasolineras REPSOL pueden tener rótulos diferentes pero siempre incluyendo la palabra REPSOL así que tengo que agruparlas. Esto mismo nos ocurre con varias marcas del dataset.
Caso de no poderla clasificar dejo la información contenida en Rotulo.

En primer lugar voy a eliminar el carácter “ que hay en algunos rótulos. Lo sustituyo por un espacio en blanco.

```
data gasolina95;
  set gasolina95;
  format rotulo_limpio $60.;
  rotulo_limpio = translate (rotulo, ' ', '"');

run;
```

A continuación indico el código utilizado para la agrupación. En el último paso **eliminamos la variable rotulo y rotulo_limpio quedándonos en su lugar con la variable cartel**

```
data gasolina95;
  set gasolina95;
  format cartel $50.;
  if findw(upcase(rotulo_limpio), "CEPSA") then cartel="CEPSA";
  else if findw(upcase(rotulo_limpio), "BP") OR
      findw(upcase(rotulo_limpio), "B.P.") then cartel="BP";
```

```
else if findw(upcase(rotulo_limpio), "SHELL") then
    cartel="SHELL";
else if findw(upcase(rotulo_limpio), "DISA") then
    cartel="DISA";
else if findw(upcase(rotulo_limpio), "REPSOL") then
    cartel="REPSOL";
else if findw(upcase(rotulo_limpio), "CEPSA") then
    cartel="CEPSA";
else if findw(upcase(rotulo_limpio), "GALP") then
    cartel="GALP";
else if findw(upcase(rotulo_limpio), "CARREFOUR") OR
    findw(upcase(rotulo_limpio), "CARRREFOUR")then
    cartel="CARREFOUR";
else if findw(upcase(rotulo_limpio), "ALCAMPO") then
    cartel="ALCAMPO";
 else if findw(upcase(rotulo_limpio), "AGLA") then
    cartel="AGLA";
else if findw(upcase(rotulo_limpio), "ALAMEDA") then
    cartel="ALAMEDA";
else if findw(upcase(rotulo_limpio), "AN ENERGETICOS") then
     cartel="AN ENERGETICOS";
else if findw(upcase(rotulo_limpio), "ANDAMUR") then
    cartel="ANDAMUR";
else if findw(upcase(rotulo_limpio), "ARENTO") then
    cartel="ARENTO";
else if findw(upcase(rotulo_limpio), "ASC") then cartel="ASC";
else if findw(upcase(rotulo_limpio), "AVANZA") then
    cartel="AVANZA";
else if findw(upcase(rotulo_limpio), "AVIA") then
    cartel="AVIA";
else if findw(upcase(rotulo_limpio), "BDMED") then
    cartel="BDMED";
else if findw(upcase(rotulo_limpio), "BENZINA") then
    cartel="BENZINA";
else if findw(upcase(rotulo_limpio), "BENZINERA") then
    cartel="BENZINERA";
else if findw(upcase(rotulo_limpio), "BEROIL") then
    cartel="BEROIL";
else if findw(upcase(rotulo_limpio), "BIOMAR") then
    cartel="BIOMAR";
else if findw(upcase(rotulo_limpio), "CAMPSA") then
    cartel="CAMPSA";
else if findw(upcase(rotulo_limpio), "CANARY") then
    cartel="CANAY";
else if findw(upcase(rotulo_limpio), "CLC") then cartel="CLC";
else if findw(upcase(rotulo_limpio), "DST") then cartel="DST";
else if findw(upcase(rotulo_limpio), "E.LECLERC") OR
    findw(upcase(rotulo_limpio), "E-LECLERC")OR
    findw(upcase(rotulo_limpio), "LECLERC") then
    cartel="LECLERC";
else if findw(upcase(rotulo_limpio), "ECOSUMINISTROS") then
    cartel="ECOSUMINISTROS";
else if findw(upcase(rotulo_limpio), "EROSKI") then
    cartel="EROSKI";
else if findw(upcase(rotulo_limpio), "EXOIL") then
    cartel="EXOIL";
else if findw(upcase(rotulo_limpio), "FAMILY ENERGY") then
    cartel="FAMILY ENERGY";
else if findw(upcase(rotulo_limpio), "FAST FUEL") then
    cartel="FAST FUEL";
```

```
else if findw(upcase(rotulo_limpio), "GACOSUR") then
    cartel="GACOSUR";
else if findw(upcase(rotulo_limpio), "RUNNER") then
    cartel="GAS RUNNER";
else if findw(upcase(rotulo_limpio), "GASEXPRESS") then
    cartel="GASEXPRESS";
else if findw(upcase(rotulo_limpio), "GHC") then cartel="GHC";
else if findw(upcase(rotulo_limpio), "GLOBALTANK") then
    cartel="GLOBALTANK";
else if findw(upcase(rotulo_limpio), "GP") then cartel="GP";
else if findw(upcase(rotulo_limpio), "HAFESA") then
    cartel="HAFESA";
else if findw(upcase(rotulo_limpio), "IBERDOEX") then
    cartel="IBERDOEX";
else if findw(upcase(rotulo_limpio), "INLOCOR") then
    cartel="INLOCOR";
else if findw(upcase(rotulo_limpio), "JAENCOOP") then
    cartel="JAENCOOP";
else if findw(upcase(rotulo_limpio), "JULIA-OIL") then
    cartel="JULIA-OIL";
else if findw(upcase(rotulo_limpio), "LABOIL") then
    cartel="LABOIL";
else if findw(upcase(rotulo_limpio), "LLORSOIL") then
    cartel="LLORSOIL";
else if findw(upcase(rotulo_limpio), "MEROIL") then
    cartel="MEROIL";
else if findw(upcase(rotulo_limpio), "MINIOIL") then
    cartel="MINIOIL";
else if findw(upcase(rotulo_limpio), "MKT") then
    cartel="MKTOIL";
else if findw(upcase(rotulo_limpio), "BAROL") then cartel="OIL
    BAROL";
else if findw(upcase(rotulo_limpio), "PRIX") then
    cartel="OILPRIX";
else if findw(upcase(rotulo_limpio), "OPTYME") then
    cartel="OPTYME";
else if findw(upcase(rotulo_limpio), "PCAN") then
    cartel="PCAN";
else if findw(upcase(rotulo_limpio), "PETROCASH") then
    cartel="PETROCASH";
else if findw(upcase(rotulo_limpio), "PETROCAT") then
    cartel="PETROCAT";
else if findw(upcase(rotulo_limpio), "PETROMAX") then
    cartel="PETROMAX";
else if findw(upcase(rotulo_limpio), "PETROMIRALLES") then
    cartel="PETROMIRALLES";
else if findw(upcase(rotulo_limpio), "PETRONIEVES") then
    cartel="PETRONIEVES";
else if findw(upcase(rotulo_limpio), "Q8") then cartel="Q8";
else if findw(upcase(rotulo_limpio), "REPOSTAR") OR
    findw(upcase(rotulo_limpio), "REPOSTAR.")then
    cartel="REPOSTAR";
else if findw(upcase(rotulo_limpio), "ROYMAGA") then
    cartel="ROYMAGA";
else if findw(upcase(rotulo_limpio), "SARAS") then
    cartel="SARAS";
else if findw(upcase(rotulo_limpio), "STAROIL") then
    cartel="STAROIL";
else if findw(upcase(rotulo_limpio), "TAMOIL") then
    cartel="TAMOIL";
```

```
  else if findw(upcase(rotulo_limpio), "TECNOIL") then
      cartel="TECNOIL";
  else if findw(upcase(rotulo_limpio), "TGAS") then
      cartel="TGAS";
  else if findw(upcase(rotulo_limpio), "VALCARCE") then
      cartel="VALCARCE";
  else if findw(upcase(rotulo_limpio), "ZOIL") then
      cartel="ZOIL";
  else cartel=rotulo_limpio;

    drop rotulo rotulo_limpio;


run;
```

Ahora realizo el análisis de frecuencias de la variable cartel con un **proc freq**:

```
proc freq data=gasolina95;
  tables cartel;
run;
```

Procedimiento FREQ

cartel	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
PETROL STATION	4	0.01	4	0.01
AGRICAR	6	0.01	10	0.02
E.S. LA TORRETA	6	0.01	16	0.03
E.S. SUR	6	0.01	22	0.04
EA SANTA POLA	6	0.01	28	0.05
LAS GABIAS	6	0.01	34	0.06
OCTAPLUS	6	0.01	40	0.07
T. ALBERO	6	0.01	46	0.08
(SIN RÓTULO)	69	0.12	115	0.20
(sin rótulo)	6	0.01	121	0.21
+B ENERGIAS	3	0.01	124	0.21
+OIL	6	0.01	130	0.22
-	6	0.01	136	0.23
.ES CABURANTES	6	0.01	142	0.24
.ES CARBURANTES	60	0.10	202	0.34
06/32718	6	0.01	208	0.35
15909	6	0.01	214	0.36
1PRIMER	6	0.01	220	0.37
23ESO68F	6	0.01	226	0.38
24H	6	0.01	232	0.39
3 PUNT TRES	6	0.01	238	0.40
40 PIES	3	0.01	241	0.41
63182867	6	0.01	247	0.42
7267	6	0.01	253	0.43
7345	6	0.01	259	0.44
96053	6	0.01	265	0.45
A CHAN DE AMOEDO	6	0.01	271	0.46
A PALMEIRA	6	0.01	277	0.47
A&B	6	0.01	283	0.48
A. G.	6	0.01	289	0.49
A.AGUAZA	6	0.01	295	0.50

Se puede comprobar que hay valores missings (gasolineras sin rótulo, sin determinar, con un guion o que no se pueden identificar bien). Como estos valores missings no representan un alto porcentaje del conjunto de datos procedo a eliminarlos:

```
data gasolina95;
  set gasolina95;
  if cartel in ("(SIN RÓTULO)", "(sin rótulo)", "-",
  "06/32718", "15909", "23ESO68F", "63182867", "7267", "7345",
  "96053", "A800-02", "NINGUNO", "NO", "NO ROTULO", "NO TIENE",
  "N° 10.935", "N° 15.526", "N° 7374", "SIN DETERMINAR", "SIN
  ROTULO")
   then delete;

run;
```

Vuelvo a hacer el análisis de frecuencias y ordeno:

```
proc freq data=gasolina95 ORDER=freq;
  tables cartel;
run;
```

Procedimiento FREQ

cartel	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
REPSOL	16297	27.73	16297	27.73
CEPSA	8445	14.37	24742	42.10
BP	4236	7.21	28978	49.30
GALP	3189	5.43	32167	54.73
SHELL	2581	4.39	34748	59.12
PETRONOR	1144	1.95	35892	61.07
AVIA	1038	1.77	36930	62.83
CAMPSA	883	1.50	37813	64.34
CARREFOUR	838	1.43	38651	65.76
DISA	807	1.37	39458	67.13
BALLENOIL	666	1.13	40124	68.27
SARAS	420	0.71	40544	68.98
PETROPRIX	414	0.70	40958	69.69
AGLA	406	0.69	41364	70.38
MEROIL	300	0.51	41664	70.89
BONAREA	294	0.50	41958	71.39
ALCAMPO	289	0.49	42247	71.88
ESCLATOIL	276	0.47	42523	72.35
EROSKI	255	0.43	42778	72.78
VALCARCE	255	0.43	43033	73.22
GASEXPRESS	234	0.40	43267	73.61
PETROCAT	234	0.40	43501	74.01
IBERDOEX	176	0.30	43677	74.31
PLENOIL	174	0.30	43851	74.61

Con esto reducimos el **número de observaciones a 58.775**

Si vemos el número de niveles de cartel que nos quedan son 1.844 que es muy elevado y complicaría el modelo en el caso de que la variable cartel sea una de las que intervienen (que es algo bastante probable).
Para reducir los niveles de cartel opto por escoger las 15 más frecuentes y el resto las agrupo como RESTO.

Mirando el resultado del proc freq anterior veo que las 15 gasolineras más frecuentes son:
REPSOL, CEPSA, BP, GALP, SHELL, PETRONOR, AVIA, CAMPSA, CARREFOUR, DISA, BALLENOIL, SARAS, PETROPRIX, AGLA, MEROIL

Con el código siguiente me quedo con estas 15 y agrupo el resto en RESTO.

```
data gasolina95;
  set gasolina95;
  if cartel not in ("REPSOL", "CEPSA", "BP", "GALP", "SHELL",
   "PETRONOR", "AVIA", "CAMPSA", "CARREFOUR", "DISA",
   "BALLENOIL", "SARAS", "PETROPRIX", "AGLA", "MEROIL")
    then cartel = "RESTO";

run;
```

Procedimiento FREQ

cartel	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
RESTO	17111	29.11	17111	29.11
REPSOL	16297	27.73	33408	56.84
CEPSA	8445	14.37	41853	71.21
BP	4236	7.21	46089	78.42
GALP	3189	5.43	49278	83.84
SHELL	2581	4.39	51859	88.23
PETRONOR	1144	1.95	53003	90.18
AVIA	1038	1.77	54041	91.95
CAMPSA	883	1.50	54924	93.45
CARREFOUR	838	1.43	55762	94.87
DISA	807	1.37	56569	96.25
BALLENOIL	666	1.13	57235	97.38
SARAS	420	0.71	57655	98.09
PETROPRIX	414	0.70	58069	98.80
AGLA	406	0.69	58475	99.49
MEROIL	300	0.51	58775	100.00

Análisis de la variable Tipo_Venta

Realizamos un análisis de frecuencias:

```
proc freq data=gasolina95;
  tables Tipo_venta;
run;
```

Procedimiento FREQ

Tipo_venta	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
P	58727	99.92	58727	99.92
p	48	0.08	58775	100.00

En este caso podemos ver que todos los valores son P ó p (venta al público) y no se observa ningún valor R (venta restringida a socios cooperativistas), por tanto, al ser todas las observaciones de venta al público, **procedo a eliminar esta variable Tipo_venta del estudio**, ya que no nos aporta información.

```
data gasolina95;
  set gasolina95;
  drop Tipo_venta;
run;
```

Análisis de la variable Rem

Realizamos un análisis de frecuencia de esta variable que es de tipo dicotómico.

```
proc freq data=gasolina95;
  tables Rem;
run;
```

Procedimiento FREQ

Rem	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
OM	22475	38.24	22475	38.24
dm	36300	61.76	58775	100.00

Se observa que no hay valores missings ni outliers. La distribución no está del todo balanceada, se puede ver que hay más datos procedentes de minoristas (dm) que de mayoristas (OM).

Análisis de la variable Horario

La variable Horario contiene información del día de la semana y del horario de atención, así como si es de 24H. Vamos a separar esta información porque a priori una gasolinera que esté abierta más días puede que sea más cara. Lo mismo ocurriría con las gasolineras que estén abiertas las 24H, a priori resultarán más caras.
Así pues vamos a crear dos nuevas variables: dia y atención

Este es el código para crear la nueva variable categórica dia

```
data gasolina95;
 set gasolina95 ;
 if find( horario, "D") then dia="L-D";
 else if find( horario, "S") then dia="L-S";
 else if find( horario, "V") then dia="L-V";
 else if find( horario, "J") then dia="L-J";
 else dia= tranwrd ( horario, ":", "");
run;
```

Realizo un estudio de frecuencias de la variable dia

```
proc freq data=gasolina95;
  tables dia;
run;
```

Procedimiento FREQ

dia	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
L	3129	5.32	3129	5.32
L-D	52701	89.67	55830	94.99
L-J	6	0.01	55836	95.00
L-M	6	0.01	55842	95.01
L-S	2541	4.32	58383	99.33
L-V	392	0.67	58775	100.00

La variable dia no tiene missings ni outliers, presenta una distribución en la que la mayoría de las gasolineras presta servicio de lunes a domingo (89,67%) y el resto no incluyen el domingo. Hay un 5,32% de los datos que no tenemos información de la semana completa ya que no sabemos cuando finaliza el servicio pero si que empieza en lunes.

Este es el código para crear la nueva variable atención:

```
data gasolina95;
   set gasolina95;
      format atencion $10.;
      if find( horario, "24H") then atencion="24H";
      else atencion = "NO ES 24H";
run;
```

Realizo un estudio de frecuencias de la variable atencion:

```
proc freq data=gasolina95;
  tables atencion;
run;
```

Procedimiento FREQ

atencion	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
24H	24011	40.85	24011	40.85
NO ES 24H	34764	59.15	58775	100.00

La variable atencion no tiene ni missings ni outliers, presenta una distribución en el que la mayoría de las gasolineras no tienen atención de 24H (59.15%) frente al 40.85% que si presta ese servicio.

Como último paso elimino la variable horario que queda reemplazada por dia y atencion.

```
data gasolina95;
  set gasolina95;
  drop horario;
run;
```

1.4.- ANALISIS DE VARIABLES ANALÍTICAS

Primero voy a realizar un **proc means** para ver los valores estadísticos básicos de las variables analíticas y su distribución:

```
proc means data=gasolina95;
  var _numeric_;
run;
```

Procedimiento MEANS

Variable	Etiqueta	N	Media	Desv. est.	Mínimo	Máximo
Fecha	Fecha de extracción	58775	21853.00	2.1605606	21850.00	21856.00
Precio_gasolina_95		58775	1.2911487	0.0776128	0.8690000	1.4890000
Precio_gasoleo_A		58620	1.2117630	0.0750855	0.8640000	1.3990000
Precio_gasoleo_B		13063	0.8988305	0.0671285	0.6970000	1.2490000
Precio_bioetanol		36	1.4790000	0.0565433	1.3660000	1.5280000
Precio_nuevo_gasoleo_A		43245	1.2951772	0.0683600	0.8980000	1.4790000
Precio_biodiesel		282	1.2154362	0.1011859	1.0590000	1.4990000
porcentaje_ester_metilico		0	.	.	.	.
porcentaje_bioalcohol		0	.	.	.	.
Precio_gasolina_98		37225	1.4259168	0.0892145	0.9690000	1.6790000
Precio_gas_natural_comprimido		62	0.9042258	0.0389095	0.8600000	0.9950000
Precio_gas_natural_licuado		48	0.8607500	0.0476416	0.8190000	0.9800000
Precio_gases_licuados_del_petrol	Precio_gases_licuados_del_petroleo	3705	0.7167968	0.0469703	0.5150000	0.7800000

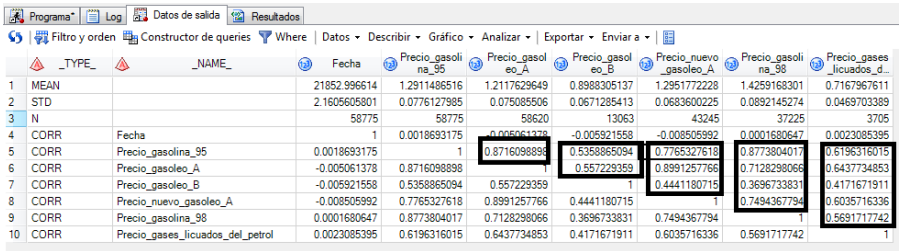
Aquí se puede ver que las variables **porcentaje_ester_metalico y porcentaje_bioalcohol** están totalmente vacías, por lo que podemos **eliminarlas del estudio.**
Igualmente se aprecia que hay muy pocas observaciones del **Precio_gas_natural_licuado, Precio_gas_natural_comprimido, Precio_bioetanol, Precio_biodiesel**, por lo que las vamos también a **eliminar del estudio**

```
data gasolina95;
  set gasolina95;
  drop porcentaje_ester_metilico porcentaje_bioalcohol
       Precio_gas_natural_licuado
       Precio_gas_natural_comprimido Precio_bioetanol
       Precio_biodiesel;
run;
```

En el resto no se observan missings, ni outliers.

Procedemos ahora a realizar un estadio de correlación de las variables analíticas que nos quedan:

```
proc corr data=gasolina95 outs=correlaciones;
  var _numeric_;
run;
```



Programa | Log | Datos de salida | Resultados

Filtro y orden | Constructor de queries | Where | Datos | Describir | Gráfico | Analizar | Exportar | Enviar a

	TYPE	_NAME_	Fecha	Precio_gasolina_95	Precio_gasoleo_A	Precio_gasoleo_B	Precio_nuevo_gasoleo_A	Precio_gasolina_98	Precio_gases_licuados_d...
1	MEAN		21852.996614	1.2911486516	1.2117629649	0.8988305137	1.2951772228	1.4259168301	0.7167967611
2	STD		2.1605605801	0.0776127985	0.075085506	0.0671285413	0.0683600225	0.0892145274	0.0469703389
3	N		58775	58775	58620	13063	43245	37225	3705
4	CORR	Fecha	1	0.0018693175	-0.005061378	-0.005921558	-0.008505992	0.0001680647	0.0023085395
5	CORR	Precio_gasolina_95	0.0018693175	1	0.8716098898	0.5358865094	0.7765327618	0.8773804017	0.6196316015
6	CORR	Precio_gasoleo_A	-0.005061378	0.8716098898	1	0.557229359	0.8991257766	0.7128298066	0.6437734853
7	CORR	Precio_gasoleo_B	-0.005921558	0.5358865094	0.557229359	1	0.4441180715	0.3696733831	0.4171671911
8	CORR	Precio_nuevo_gasoleo_A	-0.008505992	0.7765327618	0.8991257766	0.4441180715	1	0.7494367794	0.6035716336
9	CORR	Precio_gasolina_98	0.0001680647	0.8773804017	0.7128298066	0.3696733831	0.7494367794	1	0.5691717742
10	CORR	Precio_gases_licuados_del_petrol	0.0023085395	0.6196316015	0.6437734853	0.4171671911	0.6035716336	0.5691717742	1

Podemos ver que las variables están correladas positivamente y que existe una fuerte correlación entre algunas de ellas.

- El precio de la gasolina de 98 con el precio de la gasolina de 95 (87.77%)
- El precio del nuevo gasóleo A con el precio del gasóleo A (89.91%)
- El precio del gasóleo A con el precio de la gasolina de 95 (87.16%)

Independientemente, en nuestro caso, como el objetivo es la predicción del precio de la gasolina de 95 desde el punto de vista del consumidor podemos eliminar el resto de precios del estudio.

```
data gasolina95;
  set gasolina95;
  drop Precio_gasoleo_A Precio_gasoleo_B Precio_nuevo_gasoleo_A
       Precio_gasolina_98 Precio_gases_licuados_del_petrol;
run;
```

Finalmente me quedo con un dataset, cuyo objetivo es predecir el precio de la gasolina de 95 desde el punto de vista del consumidor que tiene **9 variables explicativas y 58.775 observaciones.**

Tras realizar todo el proceso de data cooking, almaceno en una librería permanente lib_in el dataset limpio trabajado hasta este punto:

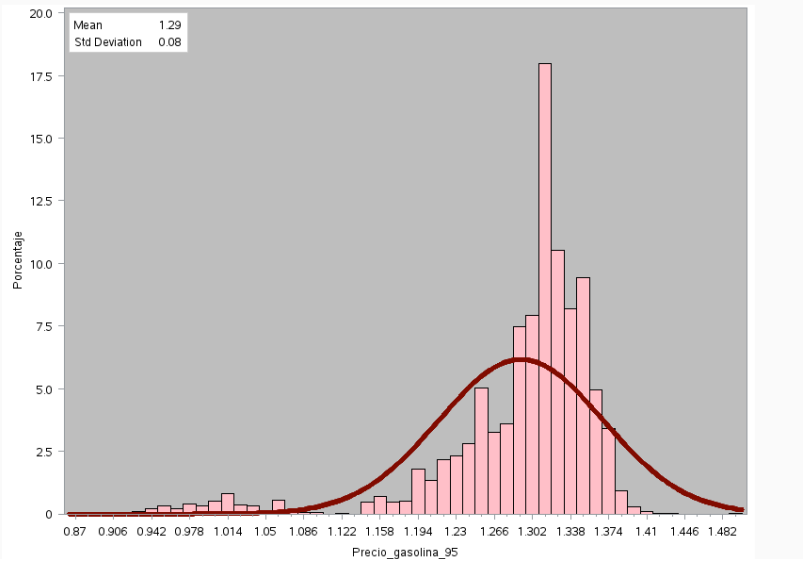
```
data lib_in.gasolina95_limpio;
  set gasolina95;
run;
```

Para determinar cuál es el algoritmo adecuado realizo un análisis de normalidad:

```
proc univariate data=lib_in.gasolina95_limpio normal plot;
   var precio_gasolina_95;
   qqplot precio_gasolina_95 / NORMAL (MU=EST SIGMA=EST
    COLOR=RED L=1);
   HISTOGRAM / NORMAL(COLOR=MAROON W=4) CFILL = pink CFRAME =
    LIGR;
   INSET MEAN STD /CFILL=BLANK FORMAT=5.2;
run;
```

Obtengo el test de normalidad y su gráfica:

Test para normalidad				
Test	Estadístico		P valor	
Kolmogorov-Smirnov	D	0.192675	Pr > D	<0.0100
Cramer-von Mises	W-Sq	622.6815	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3675.679	Pr > A-Sq	<0.0050



El p-valor vemos que es menor a 0.05, por tanto rechazamos la hipótesis nula, es decir los datos **no vienen de una distribución normal**.

2.- MODELADOS CON SAS

2.1.- MODELO LINEAL GENERAL (GLM)

Realizo ahora la partición de los datos para comenzar a entrenar el modelo. De las **58.775 observaciones** realizo estas particiones:

- Entrenamiento (70%): 41.143
- Validación (15%): 8.816
- Test (15%): 8.816

Con este código obtengo los tres datasets para el entrenamiento, la validación y el test según el criterio 70%-15%-15%.

```
data gasolina95_train gasolina95_valida gasolina95_test;
     set lib_in.gasolina95_limpio;
     if _N_ <= 41143 then output gasolina95_train;
     else if _N_ <= 49959 then output gasolina95_valida;
     else output gasolina95_test;
run;
```

Queremos predecir y explicar las causas del precio de la gasolina 95 y, como estamos con una variable de tipo continua, se determina utilizar un Modelo Lineal General como algoritmo para el análisis de modelización.

Utilizo un **GLM select** para obtener los efectos que intervienen en el modelo. Voy a realizar dos procesos **con y sin interacciones en el modelo.** En ambos comenzaré con la

semila 12345 y realizaré 200 modelos para cada uno escogiendo el método de validación cruzada.

Macro sin interacciones utilizada para generar un fichero txt que contenga los efectos, el ASE y la semilla:

```
%let lib=
"/home/u44690176/my_courses/Mariceli/mariceli0/output/macrogas95_1.txt";

%macro macro_GLM_select;
%do semilla=12345 %to 12545;
  ods graphics on;
  ods output SelectionSummary=modelos;
  ods output SelectedEffects=efectos;
  ods output Glmselect.SelectedModel.FitStatistics=ajuste;

  proc glmselect data=gasolina95_train plots=all seed=&semilla;
  partition fraction(validate=0.4);
  class Fecha Provincia Margen Rem zona cartel dia atencion;
  model precio_gasolina_95 = Fecha Provincia Margen Rem zona cartel dia
   atencion
   / selection=stepwise(select=aic choose=cv) details=all stats=all;
  run;
  ods graphics off;
  ods html close;
  data union; i=12; set efectos; set ajuste point=i; run;
  data;semilla=&semilla;file &lib mod;set union;put effects @80 nvalue1
   @95 semilla;run;
%end;
proc sql; drop table modelos,efectos,ajuste,union; quit;
%mend;

%macro_GLM_select;
```

Creo una tabla SAS con la información que he obtenido en el txt. La ordeno por orden de frecuencia de los efectos de manera descendente

```
data modelo_sin_interacciones;
length modelo $100;
input modelo $1-76 ase semilla;
cards;
Intercept Provincia Margen Rem zona cartel dia atencion 0.001189   12345
Intercept Provincia Margen Rem zona cartel dia atencion 0.001170   12346
Intercept Provincia Margen Rem zona cartel dia atencion 0.001165   12347
Intercept Provincia Margen Rem zona cartel dia atencion 0.001177   12348
Intercept Provincia Margen Rem zona cartel dia atencion 0.001214   12349
Intercept Provincia Margen Rem zona cartel dia atencion 0.001173   12350
……

run;

proc sql;
  create table modelos_sin_interacciones as
  select modelo as Modelo_GLM_sin_interacciones,
         count(modelo) as Count_model, min(ase) as Min_Ase,
          max(ase) as Max_Ase,
         avg(ase) as AVG_Ase, STD(ase) as STD_Ase
  from modelo_sin_interacciones
  group by modelo
  order by Count_model desc;
quit;
```

Se analizan los efectos de los modelos GLM, observando la frecuencia de ocurrencia, el mínimo y máximo alcanzado del ASE junto con su media y desviación

	Modelo_GLM_sin_interacciones	Count_model	Min_Ase	Max_Ase	AVG_Ase	STD_Ase
1	Intercept Provincia Margen Rem zona cartel dia atencion	200	0.001144	0.001222	0.0011753	0.0000146449
2	Intercept Fecha Provincia Margen Rem zona cartel dia atencion	1	0.001158	0.001158	0.001158	.

Resultado de la macro sin interacciones

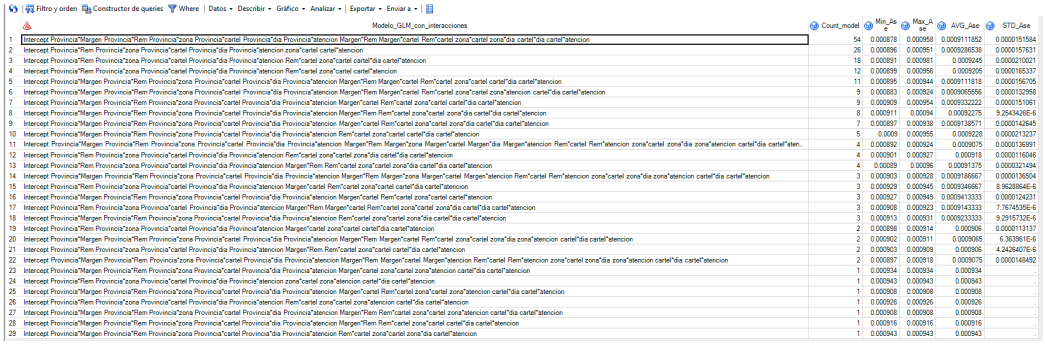
Repito ahora la misma macro pero **incluyendo 28 interacciones dos a dos**

Fecha*Provincia	Fecha*Margen	Fecha*Rem	Fecha*zona	Fecha*cartel
Fecha*dia	Fecha*atencion	Provincia*Margen	Provincia*Rem	Provincia*zona
Provincia*cartel	Provincia*dia	Provincia*atencion	Margen*Rem	Margen*zona
Margen*cartel	Margen*dia	Margen*atencion	Rem*zona	Rem*cartel
Rem*dia	Rem*atencion	zona*cartel	zona*dia	zona*atencion
cartel*dia	cartel*atencion	dia*atencion		

```
%let lib=
"/home/u44690176/my_courses/Mariceli/mariceli0/output/macrogas95_2.txt";

%macro macro_GLM_select_int;
%do semilla=12345 %to 12545;
  ods graphics on;
  ods output SelectionSummary=modelos;
  ods output SelectedEffects=efectos;
  ods output Glmselect.SelectedModel.FitStatistics=ajuste;

  proc glmselect data=gasolina95_train plots=all seed=&semilla;
  partition fraction(validate=0.4);
  class Fecha Provincia Margen Rem zona cartel dia atencion;
  model precio_gasolina_95 = Fecha*Provincia Fecha*Margen Fecha*Rem
                             Fecha*zona Fecha*cartel
                             Fecha*dia Fecha*atencion Provincia*Margen
                             Provincia*Rem Provincia*zona
                             Provincia*cartel Provincia*dia
                             Provincia*atencion
                             Margen*Rem Margen*zona Margen*cartel
                             Margen*dia Margen*atencion
                             Rem*zona Rem*cartel Rem*dia Rem*atencion
                             zona*cartel zona*dia
                             zona*atencion cartel*dia cartel*atencion
                             dia*atencion

    / selection=stepwise(select=aic choose=cv) details=all stats=all;
  run;
  ods graphics off;
  ods html close;
  data union; i=12; set efectos; set ajuste point=i; run;
  data;semilla=&semilla;file &lib mod;set union;put effects @80 nvalue1
   @95 semilla;run;
%end;
proc sql; drop table modelos,efectos,ajuste,union; quit;
%mend;

%macro_GLM_select_int;
```

Creo al igual que antes una tabla SAS con la información que he obtenido en el segundo txt. La ordeno por orden de frecuencia de los efectos de manera descendente.

Los resultados obtenidos de la macro con interacciones, los efectos del modelo GLM, la frecuencia de ocurrencia, el mínimo y el máximo alcanzado del ASE junto con su media y desviación son:



	Modelo_GLM_con_interacciones	Count_model	Min_Ase	Max_Ase	AVG_Ase	STD_Ase
1	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"Rem Margen"cartel Rem"cartel zona"cartel zona"dia cartel"dia cartel"atencion	54	0.000878	0.000958	0.0009111852	0.0000151584
2	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion zona"cartel cartel"atencion	26	0.000896	0.000951	0.0009286538	0.0000157631
3	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Rem"cartel zona"cartel cartel"dia cartel"atencion	18	0.000891	0.000981	0.0009245	0.0000210021
4	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Rem"cartel zona"cartel cartel"atencion	12	0.000899	0.000956	0.0009205	0.0000165337
5	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"Rem Margen"cartel Rem"cartel zona"cartel cartel"dia cartel"atencion	11	0.000895	0.000944	0.0009111818	0.0000156705
6	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"Rem Margen"cartel Rem"cartel zona"cartel zona"atencion cartel"dia cartel"atencion	9	0.000883	0.000924	0.0009065556	0.0000132958
7	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"cartel Rem"cartel zona"cartel cartel"dia cartel"atencion	9	0.000909	0.000954	0.0009332222	0.0000151061
8	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"Rem Rem"cartel zona"cartel zona"dia cartel"dia cartel"atencion	8	0.000911	0.00094	0.00092275	9.2543426E-6
9	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"cartel Rem"cartel zona"cartel zona"dia cartel"dia cartel"atencion	7	0.000897	0.000938	0.0009138571	0.0000142645
10	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Rem"cartel zona"cartel cartel"dia cartel"atencion	5	0.0009	0.000955	0.0009228	0.0000213237
11	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"Rem Margen"zona Margen"cartel Margen"dia Margen"atencion Rem"cartel Rem"atencion zona"cartel zona"dia zona"atencion cartel"dia cartel"aten..	4	0.000892	0.000924	0.0009075	0.0000136991
12	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Rem"cartel zona"cartel zona"dia cartel"dia cartel"atencion	4	0.000901	0.000927	0.000918	0.0000116046
13	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"Rem Rem"cartel zona"cartel zona"dia cartel"dia cartel"atencion	4	0.00089	0.00096	0.00091375	0.0000321494
14	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"Rem Margen"zona Margen"cartel Margen"atencion Rem"cartel Rem"atencion zona"cartel zona"dia zona"atencion cartel"dia cartel"atencion	3	0.000903	0.000928	0.0009186667	0.0000136504
15	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"cartel Rem"cartel zona"cartel cartel"dia cartel"atencion	3	0.000929	0.000945	0.0009346667	8.9628864E-6
16	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"cartel Rem"cartel zona"cartel zona"dia cartel"atencion	3	0.000927	0.000949	0.0009413333	0.0000124231
17	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"Rem Margen"cartel Rem"cartel zona"cartel zona"dia cartel"dia cartel"atencion	3	0.000908	0.000923	0.0009143333	7.7674535E-6
18	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Rem"cartel zona"cartel zona"dia cartel"dia cartel"atencion	3	0.000913	0.000931	0.0009233333	9.2915732E-6
19	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"cartel zona"cartel cartel"dia cartel"atencion	2	0.000898	0.000914	0.000906	0.0000113137
20	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"Rem Margen"cartel Rem"cartel zona"cartel zona"dia zona"atencion cartel"dia cartel"atencion	2	0.000902	0.000911	0.0009065	6.363961E-6
21	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"Rem Rem"cartel zona"cartel cartel"dia cartel"atencion	2	0.000903	0.000909	0.000906	4.2426407E-6
22	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"Rem Margen"cartel Margen"atencion Rem"cartel Rem"atencion zona"cartel zona"dia zona"atencion cartel"dia cartel"atencion	2	0.000897	0.000918	0.0009075	0.0000148492
23	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"cartel zona"cartel zona"atencion cartel"dia cartel"atencion	1	0.000934	0.000934	0.000934	.
24	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion zona"cartel zona"atencion cartel"dia cartel"atencion	1	0.000943	0.000943	0.000943	.
25	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"cartel Rem"cartel zona"cartel zona"atencion cartel"dia cartel"atencion	1	0.000908	0.000908	0.000908	.
26	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Rem"cartel zona"cartel zona"atencion cartel"dia cartel"atencion	1	0.000926	0.000926	0.000926	.
27	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"Rem Rem"cartel zona"cartel zona"atencion cartel"dia cartel"atencion	1	0.000908	0.000908	0.000908	.
28	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Margen"Rem Rem"cartel zona"cartel cartel"dia cartel"atencion	1	0.000916	0.000916	0.000916	.
29	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Rem"cartel zona"cartel zona"dia cartel"atencion	1	0.000943	0.000943	0.000943	.

Tras esto integro todos los modelos en una única tabla. Ordeno por frecuencia de los modelos y por el ASE de manera descendente con este código:

```
data all_model;
    set modelos_sin_interacciones
        (rename=Modelo_GLM_sin_interacciones = Modelo_GLM)
       modelos_con_interacciones
        (rename=Modelo_GLM_con_interacciones = Modelo_GLM);
run;

proc sort data=all_model;
  by descending Count_model AVG_Ase Min_Ase Max_Ase;
run;
```

La tabla final resultante con todos los modelos GLM (con y sin interacciones) es:

	Modelo_GLM	Count_model	Min_Ase	Max_Ase	AVG_Ase	STD_Ase
1	Intercept Provincia Margen Rem zona cartel dia atencion	200	0.001144	0.001222	0.0011753	0.0000146449
2	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	54	0.000878	0.000958	0.0009111852	0.0000151584
3	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion zona"c..	26	0.000896	0.000951	0.0009286538	0.0000157631
4	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Rem"c..	18	0.000891	0.000981	0.0009245	0.0000210021
5	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Rem"c..	12	0.000899	0.000956	0.0009205	0.0000165337
6	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	11	0.000895	0.000944	0.0009111818	0.0000156705
7	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	9	0.000883	0.000924	0.0009065556	0.0000132958
8	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	9	0.000909	0.000954	0.0009332222	0.0000151061
9	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	8	0.000911	0.00094	0.00092275	9.2543426E-6
10	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	7	0.000897	0.000938	0.0009138571	0.0000142645
11	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	5	0.0009	0.000955	0.0009228	0.0000213237
12	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	4	0.000892	0.000924	0.0009075	0.0000136991
13	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Marge..	4	0.00089	0.00096	0.00091375	0.0000321494
14	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Rem"c..	4	0.000901	0.000927	0.000918	0.0000116046
15	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Marge..	3	0.000908	0.000923	0.0009143333	7.7674535E-6
16	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	3	0.000903	0.000928	0.0009186667	0.0000136504
17	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	3	0.000913	0.000931	0.0009233333	9.2915732E-6
18	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Marge..	3	0.000929	0.000945	0.0009346667	8.9628864E-6
19	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	3	0.000927	0.000949	0.0009413333	0.0000124231
20	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Marge..	2	0.000898	0.000914	0.000906	0.0000113137
21	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Marge..	2	0.000903	0.000909	0.000906	4.2426407E-6
22	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	2	0.000902	0.000911	0.0009065	6.363961E-6
23	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	2	0.000897	0.000918	0.0009075	0.0000148492
24	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	1	0.000908	0.000908	0.000908	.
25	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	1	0.000908	0.000908	0.000908	.
26	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	1	0.000916	0.000916	0.000916	.
27	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion Rem"c..	1	0.000926	0.000926	0.000926	.
28	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	1	0.000934	0.000934	0.000934	.
29	Intercept Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provincia"atencion zona"c..	1	0.000943	0.000943	0.000943	.
30	Intercept Provincia"Margen Provincia"Rem Provincia"zona Provincia"cartel Provincia"dia Provinci..	1	0.000943	0.000943	0.000943	.
31	Intercept Fecha Provincia Margen Rem zona cartel dia atencion	1	0.001158	0.001158	0.001158	.

Podemos ver que el primer modelo:

Provincia Margen Rem zona cartel dia atencion

es el que no tiene interacciones, pero presenta mayor error que el resto que son con interacciones. Aunque al ser sin interacciones es más estable en el tiempo, consume menos tiempo de memoria y sería más fácil de implementar.

El segundo modelo:

*Provincia*Margen Provincia*Rem Provincia*zona Provincia*cartel Provincia*dia Provincia*atencion Margen*Rem Margen*cartel Rem*cartel zona*cartel zona*dia cartel*dia cartel*atencion*

Presenta menos error pero tiene un número elevado de interacciones con lo que debe ser más difícil de implementar.

En principio me quedaré con el segundo modelo, el que contiene interacciones:

*Precio gasolina 95 = Intercept + Provincia*Margen + Provincia*Rem + Provincia*zona + Provincia*cartel + Provincia*dia + Provincia*atencion + Margen*Rem + Margen*cartel + Rem*cartel + zona*cartel + zona*dia + cartel*dia + cartel*atencion + Error*

Introduzco estos efectos seleccionados en un proc glm:

```
proc glm data=gasolina95_train;
class Provincia Margen Rem zona cartel dia atencion;
model precio_gasolina_95 = Provincia*Margen Provincia*Rem
Provincia*zona Provincia*cartel Provincia*dia Provincia*atencion
Margen*Rem Margen*cartel Rem*cartel zona*cartel zona*dia
cartel*dia cartel*atencion
/ solution e;
run;
```

Este es el resultado:

R-cuadrado	Var Coef.	Raíz MSE	Media de Precio_gasolina_95
0.856141	2.285283	0.029504	1.291058

Para ver el resultado completo de todos los parámetros (es una tabla muy grande) dejo este enlace:

https://drive.google.com/file/d/1ObEH6AUxQPWSCZKlbjQeVx4Zv8011Qtn/view?usp=sharing

Parte inicial de la tabla de parámetros:

Parámetro		Estimación		Error estándar	t valor	Pr > |t|
T. independiente		1.291870224	B	0.03347189	38.60	<.0001
Provincia*Margen	ALBACETE D	0.005698773	B	0.04287597	0.13	0.8943
Provincia*Margen	ALBACETE I	0.011502847	B	0.04335525	0.27	0.7908
Provincia*Margen	ALBACETE N	0.016048844	B	0.04287405	0.37	0.7082
Provincia*Margen	ALICANTE D	-0.039517743	B	0.04984853	-0.79	0.4279
Provincia*Margen	ALICANTE I	-0.032114307	B	0.04977505	-0.65	0.5188
Provincia*Margen	ALICANTE N	-0.024822685	B	0.04970920	-0.50	0.6175
Provincia*Margen	ALMERÍA D	0.045490021	B	0.03841753	1.18	0.2364
Provincia*Margen	ALMERÍA I	0.051095600	B	0.03835188	1.33	0.1828
Provincia*Margen	ALMERÍA N	0.074525689	B	0.03849275	1.94	0.0529
Provincia*Margen	ASTURIAS D	-0.012741164	B	0.05291214	-0.24	0.8097
Provincia*Margen	ASTURIAS I	-0.011044535	B	0.05299604	-0.21	0.8349
Provincia*Margen	ASTURIAS N	-0.013304326	B	0.05272432	-0.25	0.8008
Provincia*Margen	BADAJOZ D	-0.143625794	B	0.05469616	-2.63	0.0086
Provincia*Margen	BADAJOZ I	-0.144560253	B	0.05477509	-2.64	0.0083
Provincia*Margen	BADAJOZ N	-0.140287433	B	0.05469252	-2.57	0.0103
Provincia*Margen	BALEARS (ILLES) D	-0.053526068	B	0.04921078	-1.09	0.2767
Provincia*Margen	BALEARS (ILLES) I	-0.050639189	B	0.04924020	-1.03	0.3038
Provincia*Margen	BALEARS (ILLES) N	-0.049231242	B	0.04915199	-1.00	0.3165
Provincia*Margen	BARCELONA D	-0.031931056	B	0.05113979	-0.62	0.5324
Provincia*Margen	BARCELONA I	-0.032790697	B	0.05115447	-0.64	0.5215
Provincia*Margen	BARCELONA N	-0.021916412	B	0.05106274	-0.43	0.6678
Provincia*Margen	BURGOS D	0.047814727	B	0.04032467	1.19	0.2357
Provincia*Margen	BURGOS I	0.056641476	B	0.04039648	1.40	0.1609
Provincia*Margen	BURGOS N	0.067883948	B	0.04049663	1.68	0.0937
Provincia*Margen	CANTABRIA D	0.037994303	B	0.04121812	0.92	0.3566
Provincia*Margen	CANTABRIA I	0.045061385	B	0.04143419	1.09	0.2768
Provincia*Margen	CANTABRIA N	0.045948173	B	0.04082208	1.13	0.2604
Provincia*Margen	CASTELLÓN / CASTELLÓ D	-0.029515000	B	0.05517696	-0.53	0.5927
Provincia*Margen	CASTELLÓN / CASTELLÓ I	-0.023367747	B	0.05517819	-0.42	0.6719
Provincia*Margen	CASTELLÓN / CASTELLÓ N	-0.020943643	B	0.05526723	-0.38	0.7047
Provincia*Margen	CEUTA D	-0.311810416	B	0.05551851	-5.62	<.0001
Provincia*Margen	CEUTA N	-0.297321733	B	0.05552895	-5.35	<.0001
Provincia*Margen	CIUDAD REAL D	-0.083264029	B	0.05369501	-1.55	0.1210
Provincia*Margen	CIUDAD REAL I	-0.073855348	B	0.05360325	-1.38	0.1683
Provincia*Margen	CIUDAD REAL N	-0.071655826	B	0.05368252	-1.33	0.1819
Provincia*Margen	CORUÑA (A) D	0.011003716	B	0.03360369	0.33	0.7433
Provincia*Margen	CORUÑA (A) I	0.013533824	B	0.03358893	0.40	0.6870
Provincia*Margen	CORUÑA (A) N	0.013365921	B	0.03353070	0.40	0.6902
Provincia*Margen	CUENCA D	-0.050412058	B	0.05316230	-0.95	0.3430
Provincia*Margen	CUENCA I	-0.057805558	B	0.05328803	-1.08	0.2780
Provincia*Margen	CUENCA N	-0.039472441	B	0.05292591	-0.75	0.4558

……..

Tengo un **R-Cuadrado de 0.8561** por lo que parece un buen resultado para la banda de ajuste

<u>Conclusión:</u> tras comparar 400 modelos de GLM, he obtenido uno con un **R=0.8561**, con un ASE de media de 0.000911 y una desviación de 0.000015 por lo que este sería el modelo "champion GLM" para predecir el precio de la gasolina 95 a través de las interacciones de variables: Provincia*Margen Provincia*Rem Provincia*zona Provincia*cartel Provincia*dia Provincia*atencion Margen*Rem Margen*cartel Rem*cartel zona*cartel zona*dia cartel*dia cartel*atencion

Precio_gasolina_95 = 1,2918 + Provincia*Margen + Provincia*Rem + Provincia*zona + Provincia*cartel + Provincia*dia + Provincia*atencion + Margen*Rem + Margen*cartel + Rem*cartel + zona*cartel + zona*dia + cartel*dia + cartel*atencion + épsilon

Donde el valor de cada parámetro viene dado por la tabla de resultado anterior y épsilon es un error aleatorio

A continuación voy a comprobar la robustez del modelo pasándole los datos de **validación**

```
proc glm data=gasolina95_valida;
class Provincia Margen Rem zona cartel dia atencion;
model precio_gasolina_95 = Provincia*Margen Provincia*Rem
Provincia*zona Provincia*cartel Provincia*dia Provincia*atencion
Margen*Rem Margen*cartel Rem*cartel zona*cartel zona*dia
cartel*dia cartel*atencion
/ solution e;
run;
```

R-cuadrado	Var Coef.	Raíz MSE	Media de Precio_gasolina_95
0.871867	2.432580	0.031412	1.291305

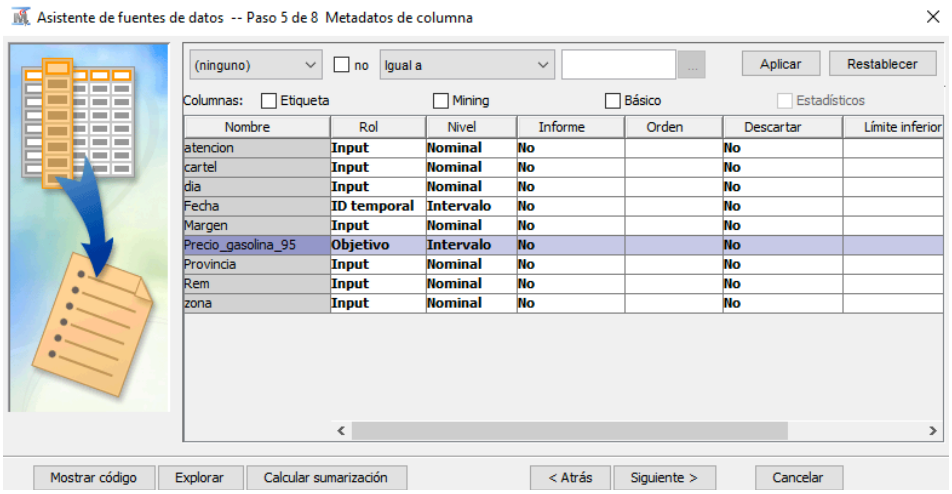
Observo que el R-cuadrado es de 0.8718 en las muestras de validación, bastante parecido a lo que he obtenido antes con las muestras de entrenamiento, por tanto podemos considerar que el modulo es robusto.

3.- COMPARACIÓN DE MODELOS

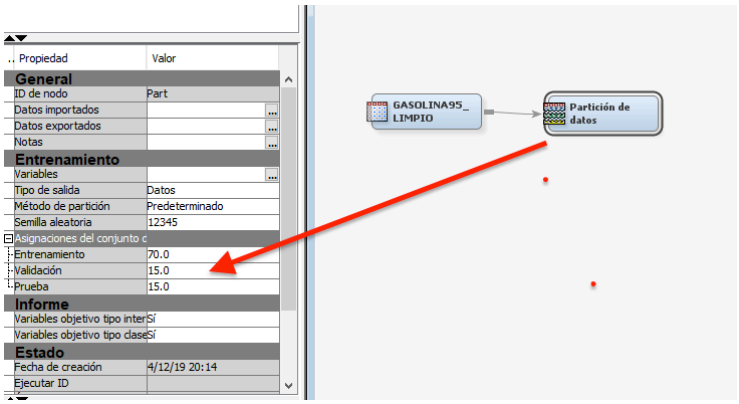
La parte de validación y test la voy a realizar con el Miner pero primero voy a realizar una comparativa de modelos para escoger el modelo "champion" y sobre ese realizar dicha validación y test.

Con la utilidad SAS Miner realizaré una comparación rápida de distintos modelos. Además del GLM compararé con la regresión lineal y con redes neuronales. Así podré ver si realmente el GLM es el mejor o es recomendable escoger otro algoritmo matemático. Utilizaré la versión HP (hight-performance) que equivale a realizar la comparativa con varios modelos.

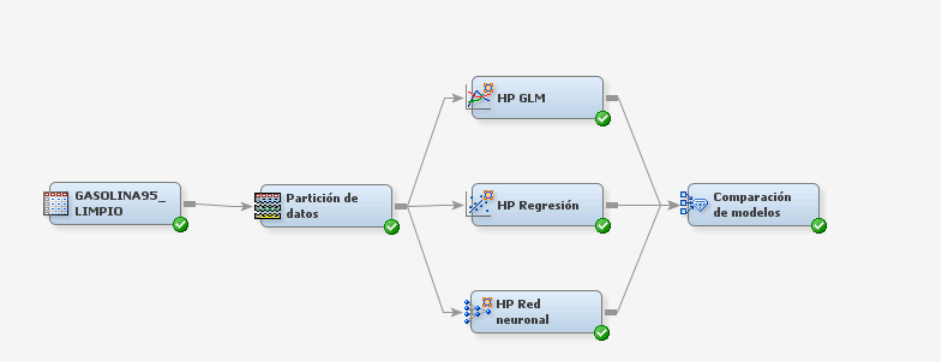
El primer paso con Miner es asignar la librería donde tengo la información y después agregar la tabla con la fuente de datos creada de gasolina (gasolina95_limpio). Escogeré dentro de la tabla nuestra variable objetivo que es **Precio_gasolina_95**



En la partición de datos aplico la misma proporción que en la primera parte del estudio 70-15-15



Añado los nodos para cada tipo de modelización de alto rendimiento (HP): GLM, regresión y red neuronal.



De todos los resultados que muestra el Miner, aquí están los más relevantes como la tabla de Train con la suma de los cuadrados de los errores y el ASE.

Data Role=Train

Statistics	HPGLM	HPNNA	HPReg
Train: Average Squared Error	0.00	0.00	0.00
Selection Criterion: Valid: Average Squared Error	0.00	0.00	0.00
Train: Divisor for ASE	23510.00	23510.00	23510.00
Train: Maximum Absolute Error	0.18	0.21	0.21
Train: Sum of Frequencies	23510.00	23510.00	23510.00
Train: Root Average Squared Error	0.03	0.03	0.03
Train: Sum of Squared Errors	18.88	22.59	27.23

Se analiza también los resultados obtenidos de validación de la comparativa.

Data Role=Valid

Statistics	HPGLM	HPNNA	HPReg
Valid: Average Squared Error	0.00	0.00	0.00
Valid: Divisor for ASE	17633.00	17633.00	17633.00
Valid: Maximum Absolute Error	0.32	0.21	0.21
Valid: Sum of Frequencies	17633.00	17633.00	17633.00
Valid: Root Average Squared Error	0.03	0.03	0.03
Valid: Sum of Squared Errors	16.74	18.40	21.12

Y por último, se analizará la parte resultante del test de los modelos comparados

Data Role=Test

Statistics	HPGLM	HPNNA	HPReg
Test: Average Squared Error	0.00	0.00	0.00
Test: Divisor for ASE	17632.00	17632.00	17632.00
Test: Maximum Absolute Error	0.32	0.21	0.21
Test: Sum of Frequencies	17632.00	17632.00	17632.00
Test: Root Average Squared Error	0.03	0.03	0.03
Test: Sum of Squared Errors	16.03	17.91	20.92

Según estos datos el modelo GLM parece ser el mejor ya que presenta una ventaja predictiva al ser la suma de los cuadrados de los errores inferior a la de los otros dos. Así pues nos quedamos con el modelo GLM

Conclusión final:

Precio_gasolina_95 = 1,2918 + Provincia*Margen + Provincia*Rem + Provincia*zona + Provincia*cartel + Provincia*dia + Provincia*atencion + Margen*Rem + Margen*cartel + Rem*cartel + zona*cartel + zona*dia + cartel*dia + cartel*atencion + épsilon

Donde el valor de cada parámetro viene dado por la tabla que aparece en fichero "Resultados_Practica95.sas.pdf" de este mismo repositorio y épsilon es un error aleatorio