Proyecto Estadistica Aplicada 2

Mauricio Diaz 200854, Luis Eduardo Suarez

Objetivo

Contexto: La base de datos ve los datos mensuales de aproximadamente 1200 estaciones de gas en todo México, distribuidos de manera aleatoria, basados en los datos para guardar información para crear dashboards sobre ellos

Objetivo: El objetivo del estudio es hacer un analisis sobre las ventas de estacion de gas. De esta manera desarrollar un modelo que nos permita tanto dar pronosticos como interpretaciones generales para lograr entender y ver que cambios hacer.

Las variables son las siguientes

Descripción de Variables % Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Mon, Dec 09, 2024 - 10:29:45

Table 1: Tabla Descriptiva para Variables

Variable	Descripción
cre_id	ID de la estación de gas
sales	Ventas mensuales de la estación de gas, medida en litros de gas
date	Mes de las ventas en particular en agosto
selling_price	Precio de venta de un litro de gas
ppc	Precio de compra de gas aproximado
quotient	Índice: >1 indica precios mayores que la competencia; <1, precios menores.
global	Nivel de tráfico alrededor de la gasolinera medida de 1 a 10
comps1km	# de estaciones de gas a 1km
comps1km_2km	# de estaciones de gas 1 a 2km
comps2km_5km	# de estaciones de gas 2 a 5km
comps5km_10km	# de estaciones de gas 5 a 10km
comps10km_plus	# de estaciones de gas a más de 10km
municipio	Municipio en el que se encuentra la estación de gas
entidad	Estado en el que se encuentra la estación de gas
population	Población del municipio en el que se encuentra la estación del gas
Cars	Número de carros en el municipio
Pib per capita	PIB per cápita del estado
Pib	PIB del estado

Analisis Exploratorio de Datos

El objetivo general de esta seccion es comprender la relación entre las variables independientes y la variable dependiente. Esto implica analizar la distribución de las variables, explorar posibles correlaciones y patrones

entre las variables predictoras y la respuesta, identificar relaciones lineales o no lineales. También se busca detectar anomalías o valores atípicos que puedan influir en el modelo. Esta parte nos permite tanto tomar decisiones informadas como respondernos preguntas sobre el comportamiento de las variables para ajustar un modelo de regresión que sea interpretable.

Estadisticas Descriptivas

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Mon, Dec 09, 2024 - 10:29:45

Table 2: Estadistica Descriptiva

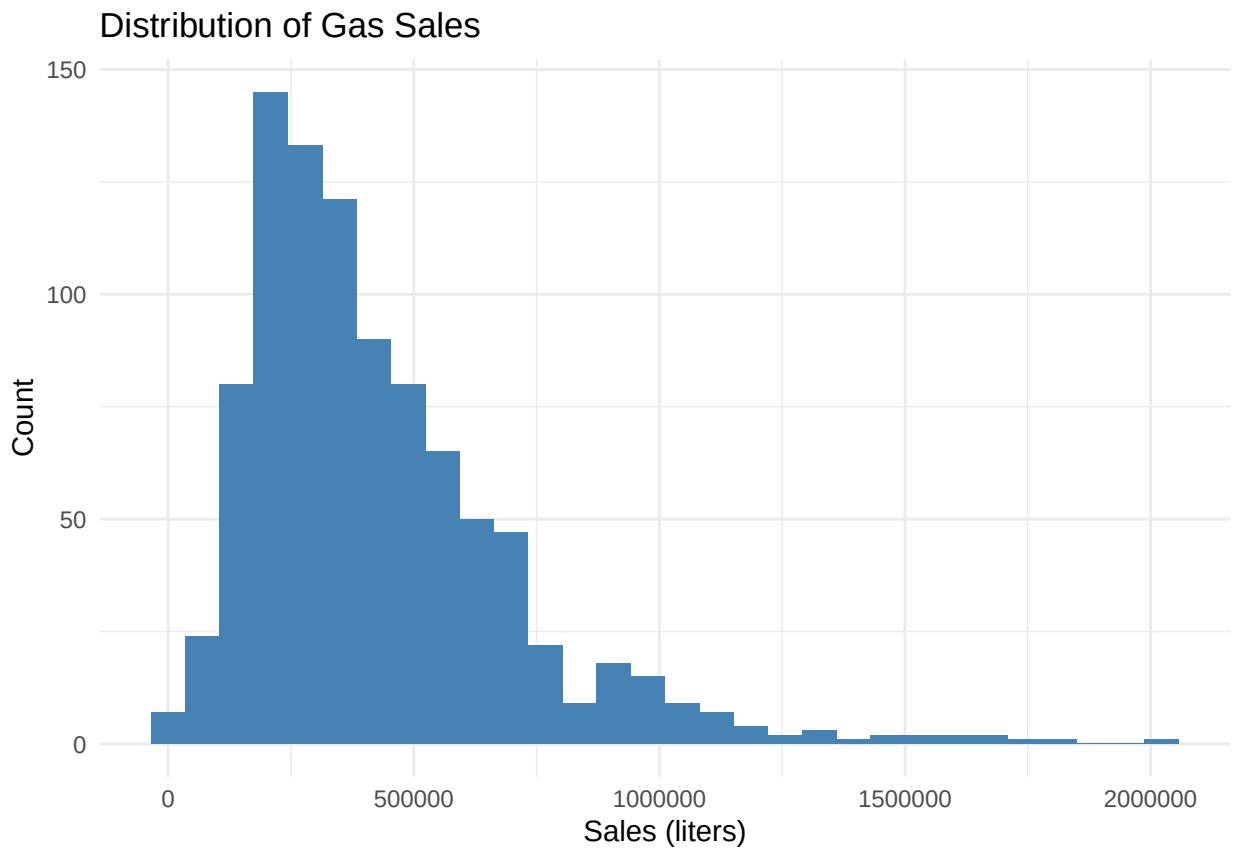
Statistic	N	Mean	St. Dev.	Min	Max
sales	943	427,810.800	273,913.500	3,602	2,026,495
selling_price	943	24.914	1.223	18.708	27.257
ppc	943	24.784	1.242	16.753	27.304
quotient	943	1.006	0.035	0.903	1.518
flag_min	943	12.811	22.296	0	93
global	939	6.581	1.481	0.000	9.923
comps1km	943	2.182	2.139	0	11
comps1km_2km	943	5.135	4.558	0	18
comps2km_5km	943	7.357	5.265	0	20
comps5km_10km	943	2.119	3.780	0	20
comps10km_plus	943	1.807	4.182	0	20
population	896	704,520.200	648,871.900	3,176	1,922,523
Cars	889	392,571.000	403,606.800	298	1,293,116
PIB.per.cápita..MXN.	936	257,422.900	79,627.260	83,827	499,492
PIB	936	1,313,488.000	726,210.700	186,670	4,652,611

Porcentaje de Nulos en los Datos

	column	null_count	null_percentage
Cars	Cars	54	5.73
population	population	47	4.98
PIB.per.cápita..MXN.	PIB.per.cápita..MXN.	7	0.74
PIB	PIB	7	0.74
global	global	4	0.42
sales	sales	0	0.00
selling_price	selling_price	0	0.00
ppc	ppc	0	0.00
quotient	quotient	0	0.00
flag_min	flag_min	0	0.00
comps1km	comps1km	0	0.00
comps1km_2km	comps1km_2km	0	0.00
comps2km_5km	comps2km_5km	0	0.00
comps5km_10km	comps5km_10km	0	0.00
comps10km_plus	comps10km_plus	0	0.00
municipio	municipio	0	0.00
entidad	entidad	0	0.00

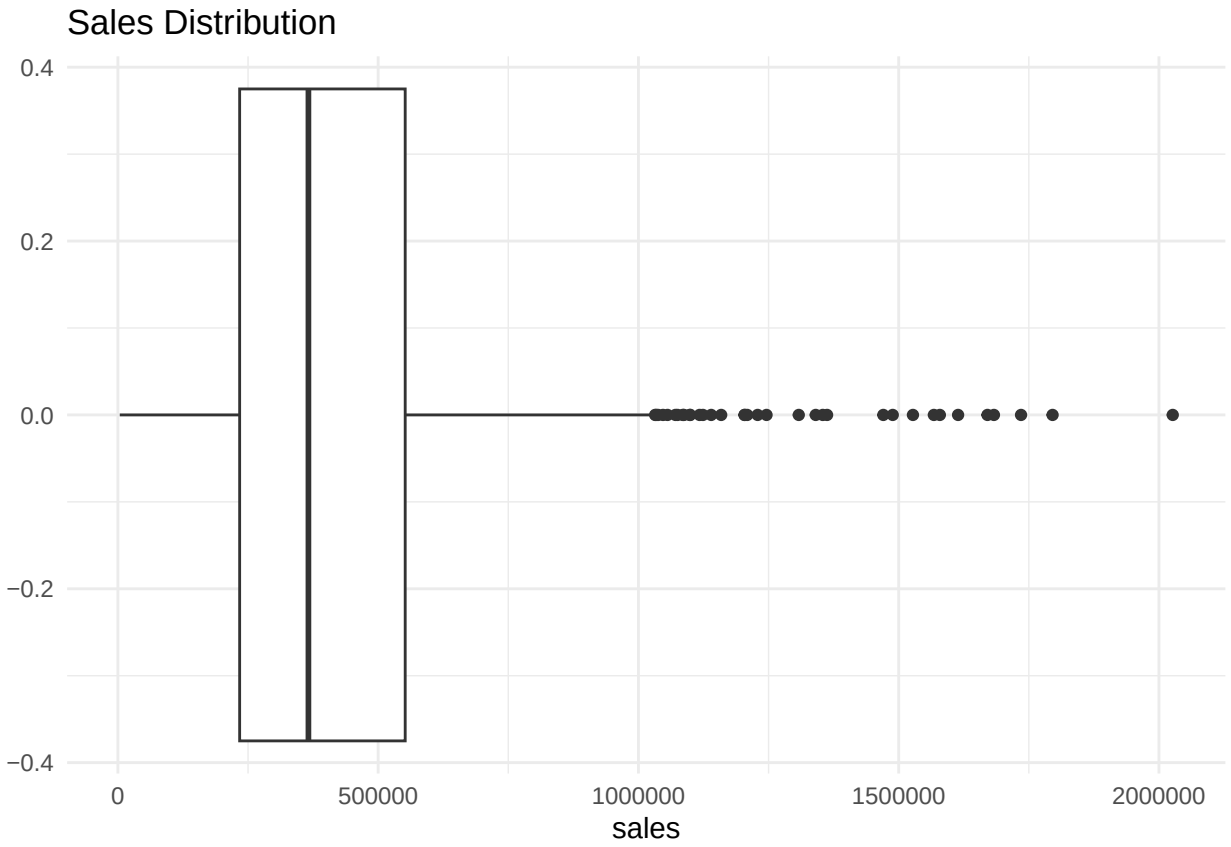
Distribucion General de la variable Objetivo

Aqui damos una visualización de la distribución general de la variable objetivo mediante un grafico de barras



En esta seccion elegimos usar una grafico de Caja para mostrar el comportamiento general de la variable objetivo y los valores que logra tener.

Vista simple de Valores Atipicos en la variable Objetivo



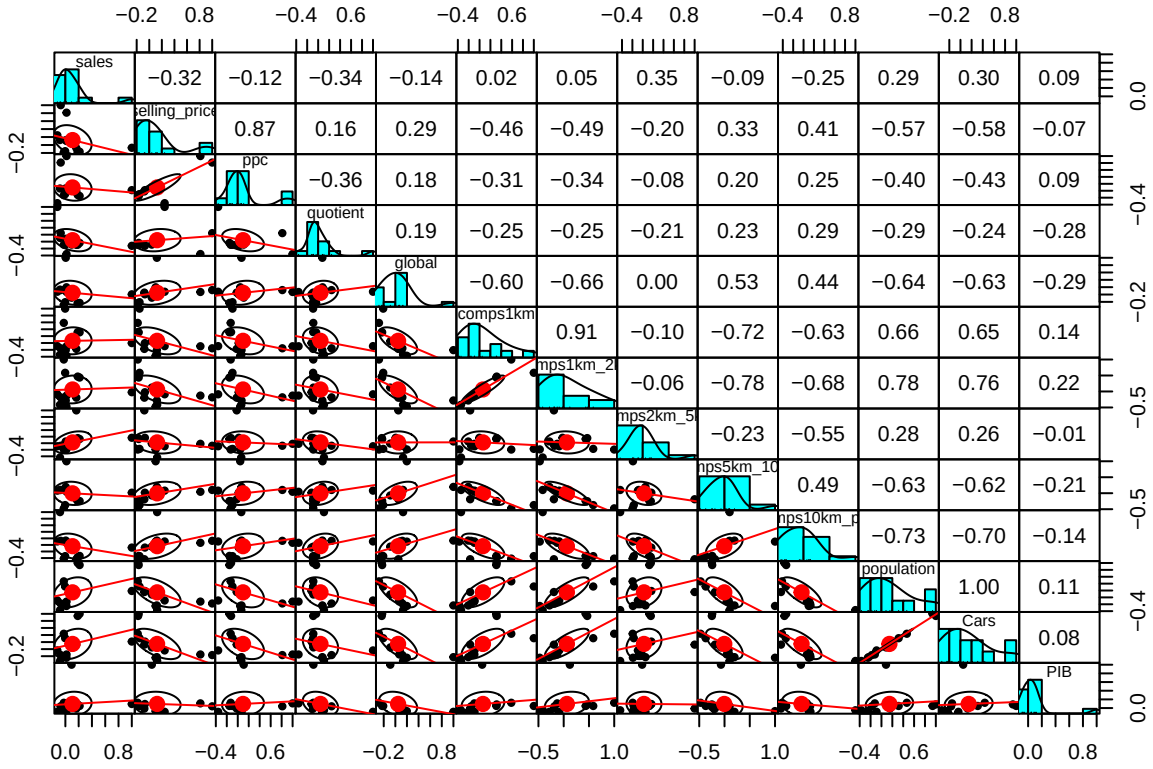
Busqueda de Valores Atipicos en variables Independientes

Con la tabla siguiente buscamos dar

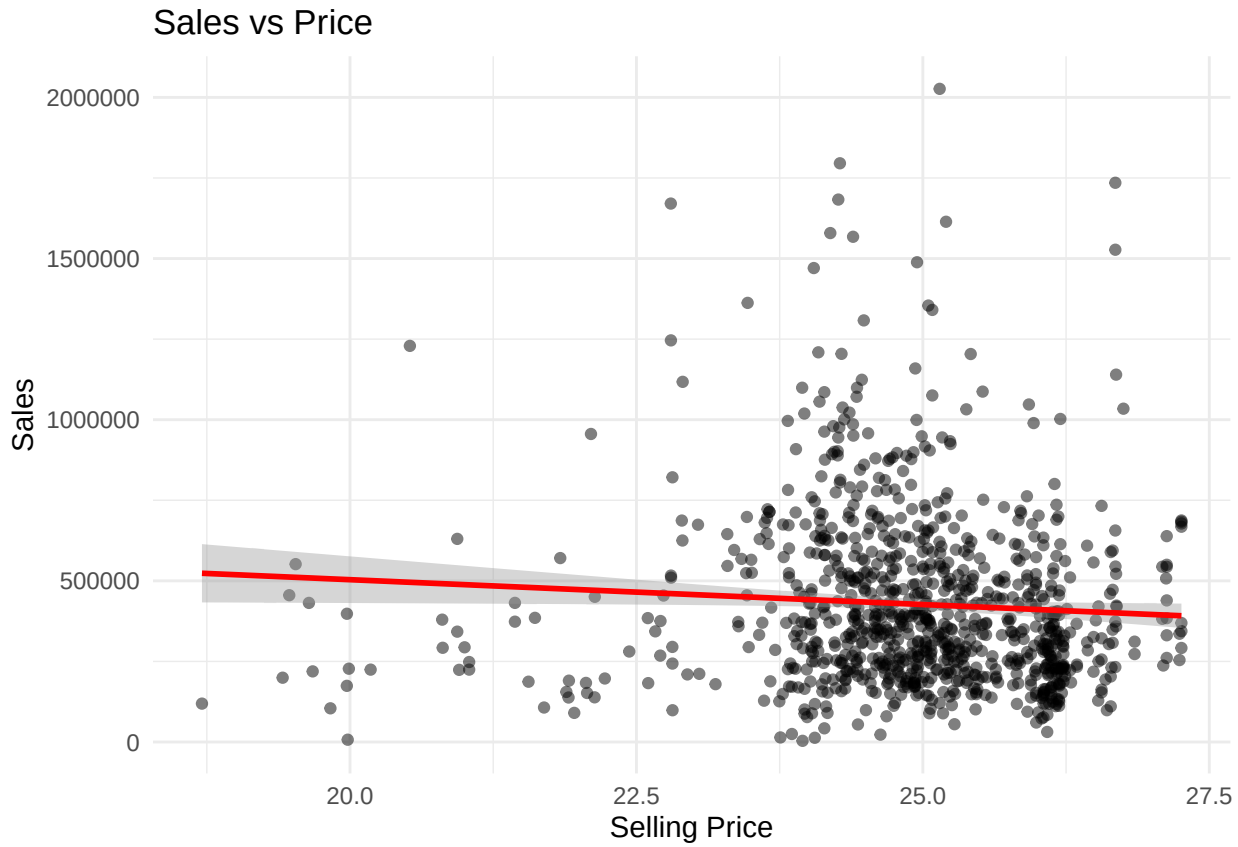
	column	null_count	null_percentage	outlier_count
Cars	Cars	54	5.73	0
population	population	47	4.98	0
PIB.per.cápita..MXN.	PIB.per.cápita..MXN.	7	0.74	0
PIB	PIB	7	0.74	4
global	global	4	0.42	0
sales	sales	0	0.00	0
selling_price	selling_price	0	0.00	0
ppc	ppc	0	0.00	0
quotient	quotient	0	0.00	0
flag_min	flag_min	0	0.00	0
comps1km	comps1km	0	0.00	0
comps1km_2km	comps1km_2km	0	0.00	0
comps2km_5km	comps2km_5km	0	0.00	0
comps5km_10km	comps5km_10km	0	0.00	0
comps10km_plus	comps10km_plus	0	0.00	0
municipio	municipio	0	0.00	0
entidad	entidad	0	0.00	0

Analisis de Correlacion entre las variables

Lo que estamos haciendo en esta seccion es generar una matriz de correlacion para ver la interaccion entre las variables y su comportamiento. Esto no lo estamos usando como metodo en este caso pero tambien nos permite ver multicolinealidad.



Analisis de variable de precio con Ventas



Ajuste del Modelo

En esta seccion ya llegamos al ajuste del modelo.

Tenemos el siguiente modelo donde,

- Variable Objetivo (Y): Ventas
- Variable Dependientes (X): precio, trafico, ppc, quotient...

$$Y = X\beta + \epsilon$$

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Mon, Dec 09, 2024 - 10:29:48

Table 5:

	Dependent variable:
	sales
selling_price	−231,114.300***
	(69,549.070)
ppc	236,569.400***
	(71,190.180)
quotient	4,368,491.000***
	(1,481,886.000)
flag_min	2,062.065***
	(410.517)
global	6,801.851
	(6,096.095)
comps1km	3,232.126
	(5,260.654)
comps1km_2km	−951.629
	(3,226.634)
comps2km_5km	9,024.147***
	(2,502.731)
comps5km_10km	8,954.540***
	(3,093.137)
comps10km_plus	3,576.871
	(3,251.908)
population	−0.133***
	(0.041)
Cars	0.351***
	(0.064)
PIB.per.cápita..MXN.	0.302**
	(0.123)
PIB	0.020
	(0.013)
Constant	−4,381,399.000***
	(1,533,317.000)
Observations	886
R^2	0.160
Adjusted R^2	0.146
Residual Std. Error	252,774.900 (df = 871)
F Statistic	11.841*** (df = 14; 871)
Note:	*p<0.1; **p<0.05; ***p<0.01

Analisis de los estimadores

En esta seccion vamos a ver de manera particular cada uno de los estimadores de Regresion, los coeficientes, tuvimos la oportunidad de ver cada uno al ajustar el modelo en esta seccion podremos ver:

- varianza
- Y gorro estimacion
- Y gorro pronostico

Varianza

sigma = 252774.9

Podemos ver que tenemos una varianza de 252774.9, no estamos asumiendo una varianza fija ya que puede haber heterestadicistidad, pero estamos usando la varianza que nos da el valor del modelo

Prediccion Dato Interno

Queremos ver la estimacion general para datos interno de una gasolinera con los siguientes datos

- selling_price = 24.96989,
- ppc = 25.01667,
- quotient = 0.9982151,
- flag_min = 0,
- global = 6.552333,
- comps1km = 0,
- comps1km_2km = 2,
- comps2km_5km = 11,
- comps5km_10km = 7,
- comps10km_plus = 0,
- population = 35223,
- Cars = 4029,
- PIB.per.cápita..MXN. = 165567,
- PIB = 1086962

Ybarra = 399911.5

CI = (356883.8, 442939.2)

Podemos ver en la seccion pasado que tenemos una estimacion de 399911.5.

Pronostico De Dato

YbarraPronostico = 399911.5

CI = (-98070.02, 897893.1)

Podemos ver el mismo pronostico que tiene el mismo valor pero cambia el intervalo de confianza y vemos que cruza el 0. El intervalo de predicción cruza el 0 porque esto puede ser porque incluye tanto la incertidumbre del modelo como la variabilidad de las predicciones individuales. Esto puede indicar que hay alta variabilidad en los residuos del modelo.

Anova

Queremos lograr entender de donde se origina la variacion de Y barra es por eso que hacemos un analisis de varianza (anova) que nos ayuda a entender como cada variable independiente explica la variacion total en los datos.

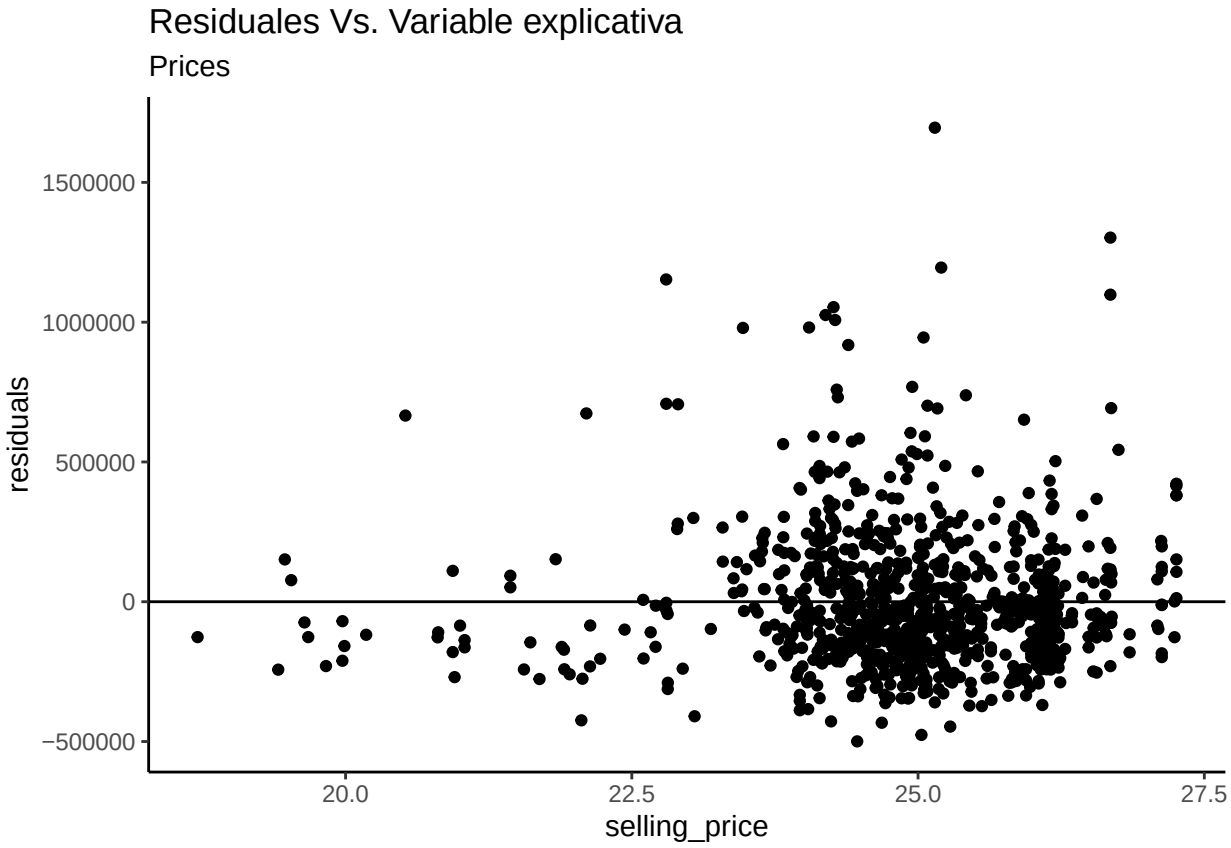
```
## Analysis of Variance Table
##
## Response: sales
##                        Df     Sum Sq    Mean Sq F value    Pr(>F)
## selling_price           1 3.2197e+11 3.2197e+11  5.0390 0.0250336 *
## ppc                     1 1.0778e+12 1.0778e+12 16.8684 4.384e-05 ***
## quotient                1 7.3654e+11 7.3654e+11 11.5273 0.0007168 ***
## flag_min                1 2.0335e+12 2.0335e+12 31.8259 2.280e-08 ***
## global                  1 4.3859e+10 4.3859e+10  0.6864 0.4076100
## comps1km                1 9.9230e+09 9.9230e+09  0.1553 0.6936159
## comps1km_2km            1 3.5220e+10 3.5220e+10  0.5512 0.4580198
## comps2km_5km            1 1.3985e+12 1.3985e+12 21.8876 3.350e-06 ***
## comps5km_10km           1 5.0125e+11 5.0125e+11  7.8449 0.0052094 **
## comps10km_plus          1 8.5351e+10 8.5351e+10  1.3358 0.2480945
## population              1 1.8102e+12 1.8102e+12 28.3310 1.301e-07 ***
## Cars                    1 1.8468e+12 1.8468e+12 28.9033 9.776e-08 ***
## PIB.per.cápita..MXN.    1 5.3311e+11 5.3311e+11  8.3435 0.0039666 **
## PIB                     1 1.5799e+11 1.5799e+11  2.4727 0.1162006
## Residuals             871 5.5653e+13 6.3895e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Podemos ver que en este caso la suma de cuadrados de los residuales es mayor a cualquier de los ajustes de los parametros lo que nos indica que la variacion de las ventas no es explicada por el modelo. Podemos ver de la misma manera como la tabla anova.

Diagnostico del modelo

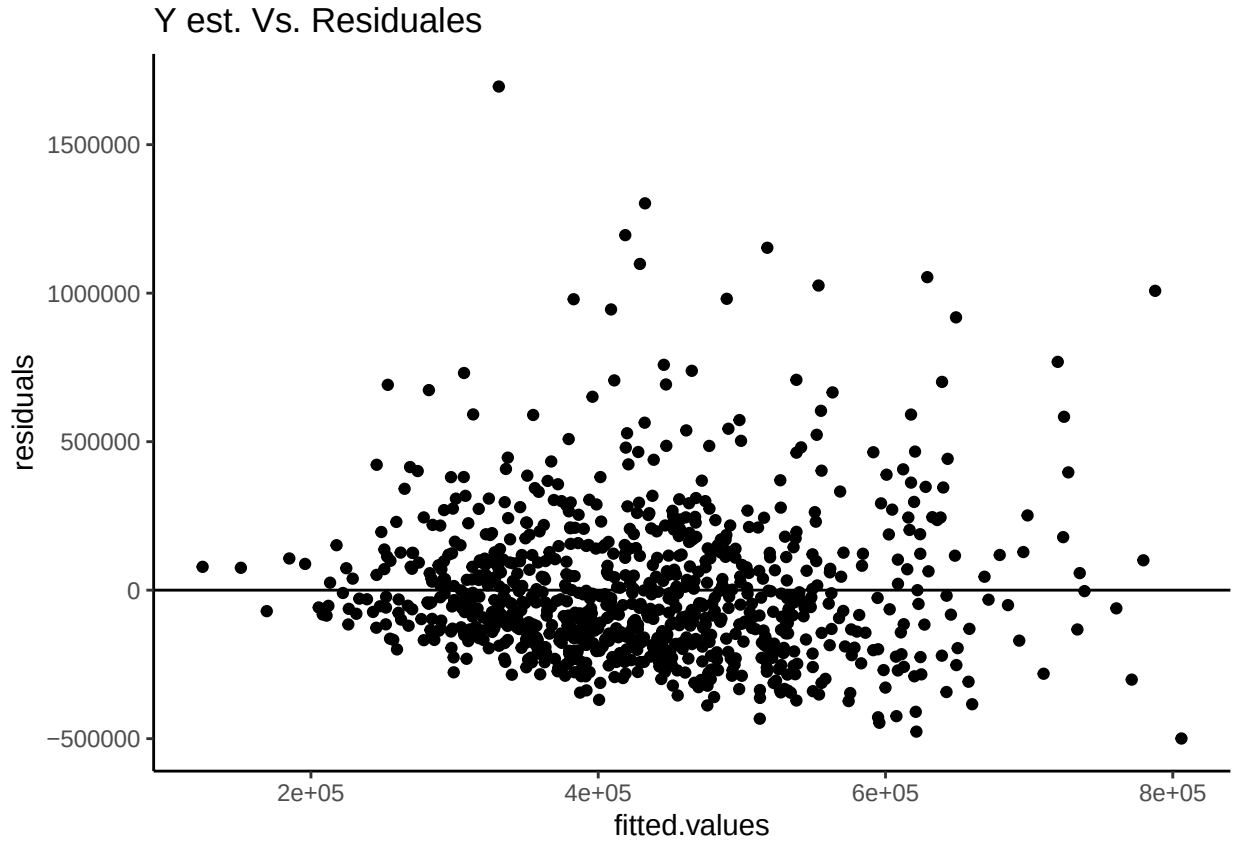
Analisis de Residuales

Residuales con una de las variables



Residuales y fitted values

heteroesquedasticidad




```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 20.877, df = 14, p-value = 0.1048
```

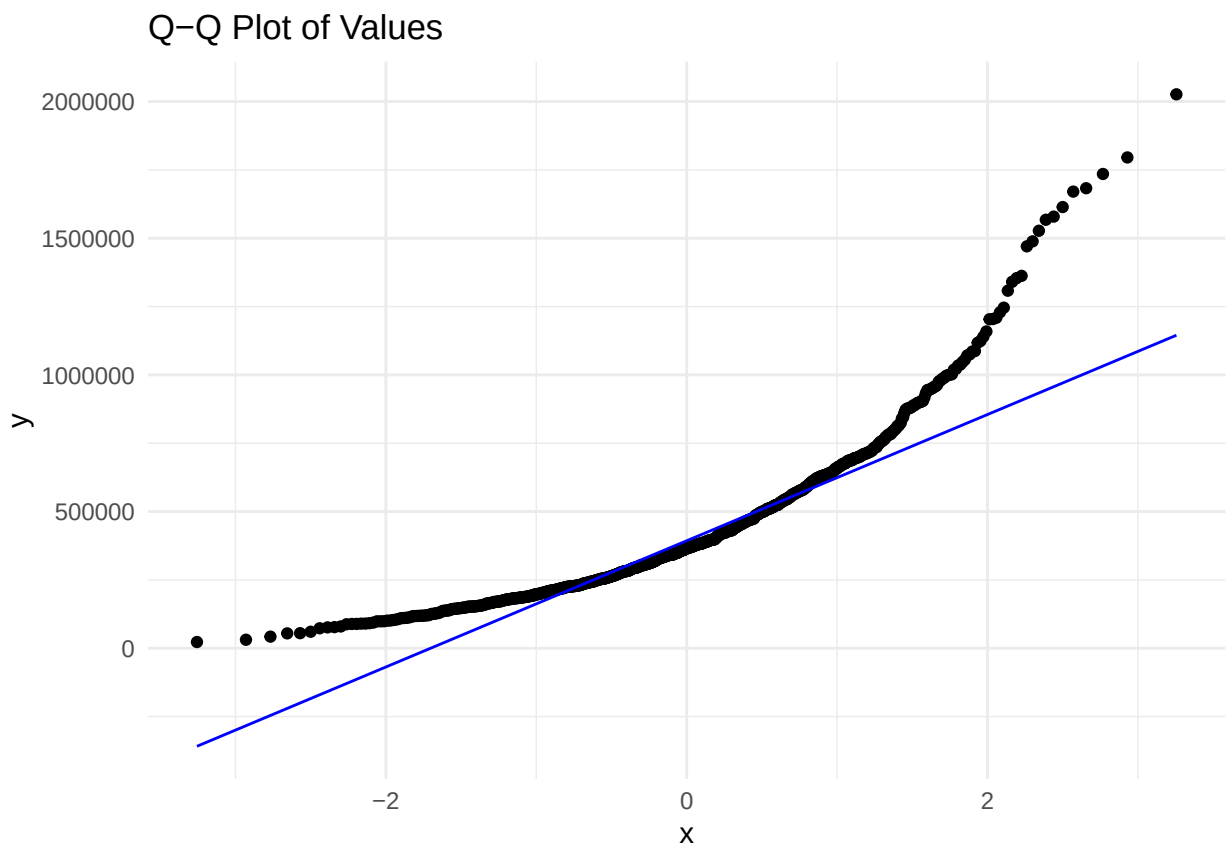
Linealidad

Rainbow test

```
##
##  Rainbow test
##
## data:  sales ~ .
## Rain = 0.81963, df1 = 443, df2 = 428, p-value = 0.981
```

Normalidad

qqplot



Multicolinealidad

VIF

```
##        selling_price                  ppc            quotient
##            101.152146           109.551955           38.166917
##              flag_min               global            comps1km
##              1.136728             1.127507             1.776348
##         comps1km_2km         comps2km_5km        comps5km_10km
##              3.056747             2.422142             1.916756
##       comps10km_plus           population                 Cars
##              2.423294             9.846311             9.281752
## PIB.per.cápita..MXN.                  PIB
##              1.256846             1.132448
```