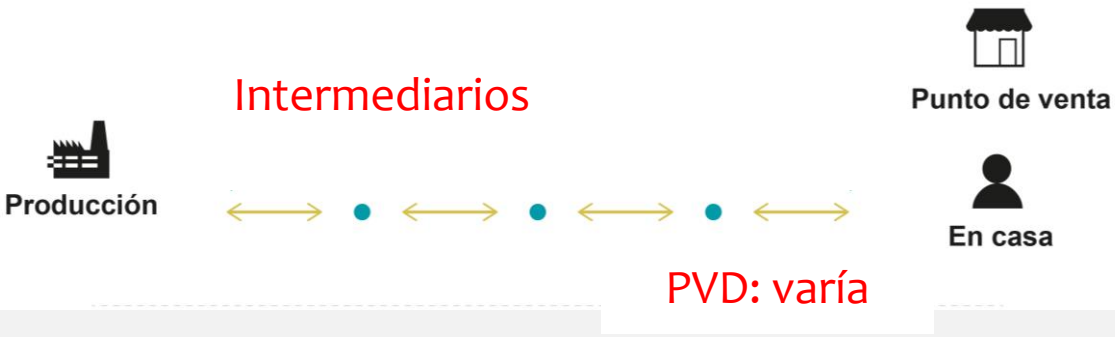
Determinación del precio de calzado de una empresa de consumo masivo a sus canales de distribución, a través de regresión lineal múltiple

Elaborado por: Jheremy Ron

Problemática: Nuestra empresa enfrenta el desafío de optimizar las predicciones de ventas de “zapatos”, debido a la variabilidad en la demanda influenciada por la estacionalidad, el tipo de tienda y las características del producto.

- **Objetivos:**

1. Desarrollar un modelo de regresión lineal múltiple para predecir ventas basadas en múltiples variables.
2. Identificar las variables más influyentes en las ventas, como el tipo de tienda, el material y la temporada.
3. Mejorar la planificación de inventarios y las estrategias de marketing con predicciones precisas.



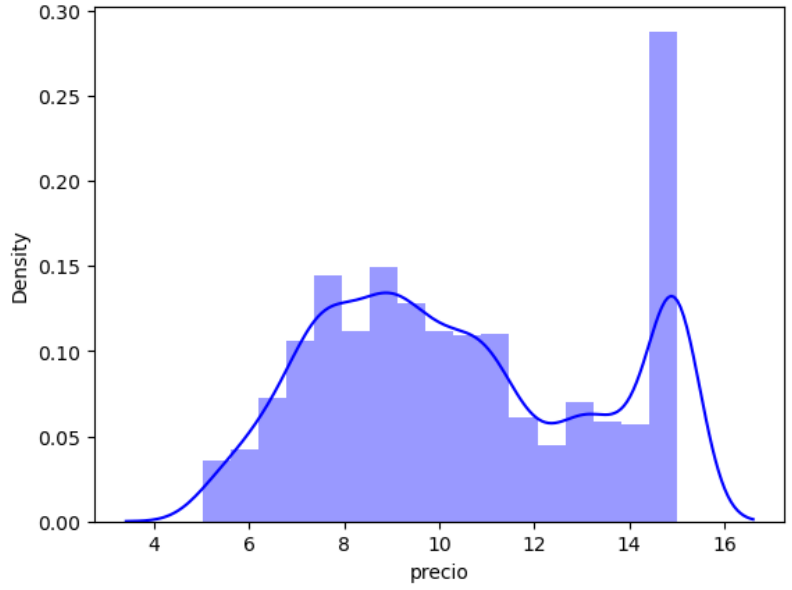
Introducción del problema y objetivos

Metodología y Resultados

1.- Análisis de valores extremos Identificación de variables dependientes e independientes “Precio”

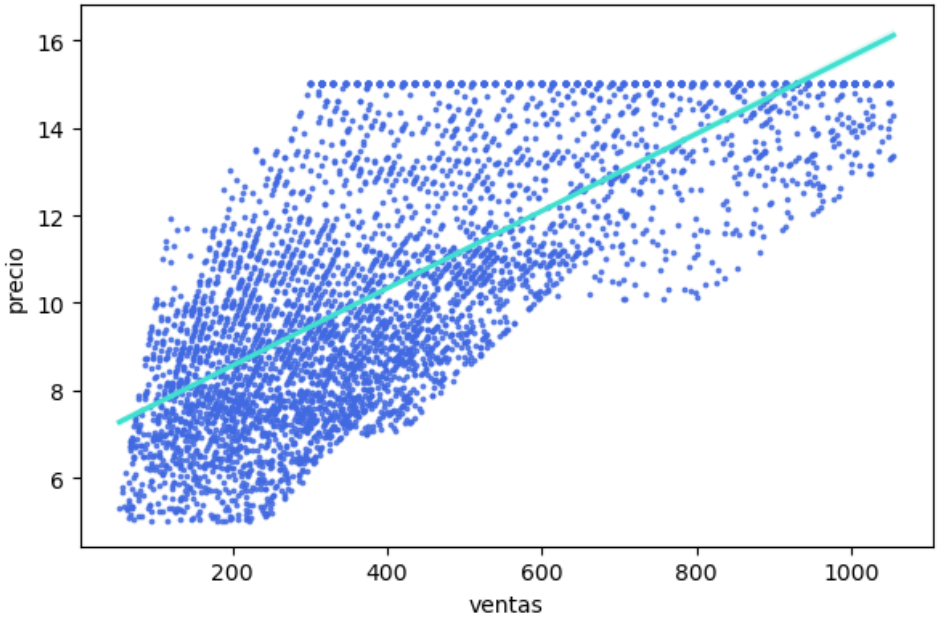
2.- Depuramos la información donde, se eliminó todos los datos atípicos a través de quartiles

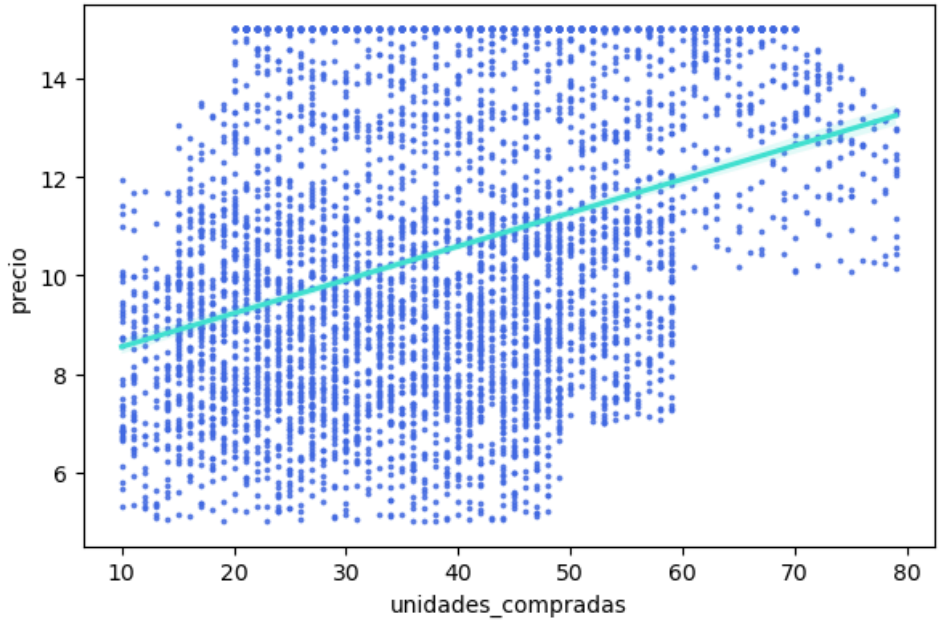
	mes	dia_semana	tamaño	material	ventas	unidades_compradas	compras_por_dia	precio
mes	1.000000	0.008341	-0.002684	-0.023619	0.002513	-0.026454	-0.021371	0.055382
dia_semana	0.008341	1.000000	0.033015	-0.009222	0.018342	-0.013014	-0.006888	0.069690
tamaño	-0.002684	0.033015	1.000000	0.005620	-0.009491	-0.003178	-0.016071	-0.014694
material	-0.023619	-0.009222	0.005620	1.000000	-0.009148	-0.002901	0.000935	-0.016965
ventas	0.002513	0.018342	-0.009491	-0.009148	1.000000	0.883470	0.008239	0.725010
unidades_compradas	-0.026454	-0.013014	-0.003178	-0.002901	0.883470	1.000000	-0.000330	0.362172
compras_por_dia	-0.021371	-0.006888	-0.016071	0.000935	0.008239	-0.000330	1.000000	0.023741
precio	0.055382	0.069690	-0.014694	-0.016965	0.725010	0.362172	0.023741	1.000000

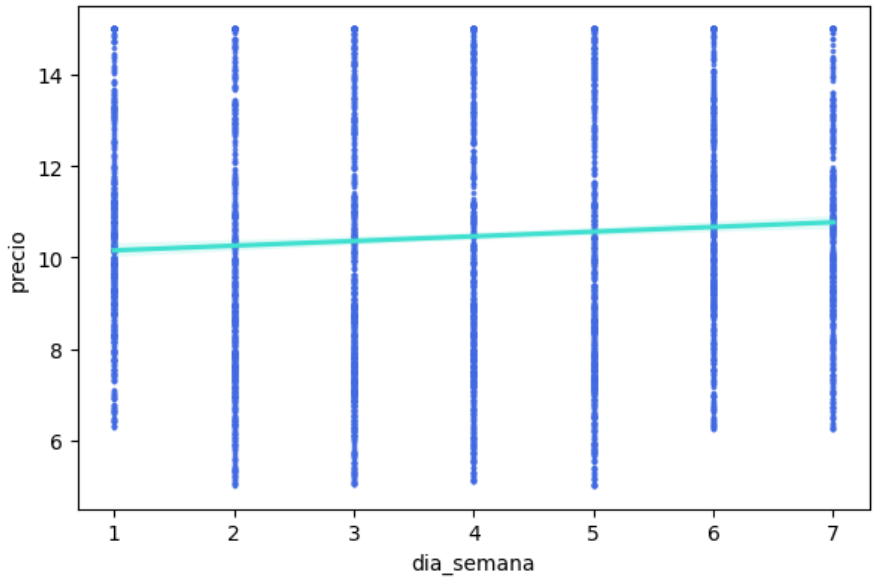


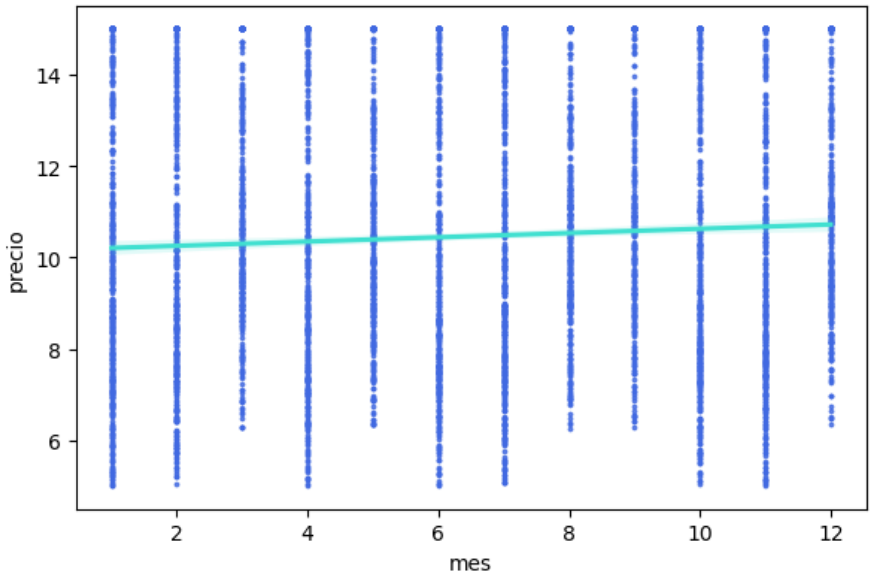
3 .- Correlación entre variables Y evaluación de los puntos más altos con las variables cuantitativas

Metodología y Resultados









Metodología y Resultados

Primero se aplicó label encoder, sin embargo, no había suficiente información para trabajar al momento de evaluar VIF, donde se procedió a dumificar las variables

	fecha	mes	dia_semana	sku	tamaño	material	ventas	unidades_compradas	compras_por_dia	precio	log_precio	charolero	complejos deportivos
0	2022-01-01	1	7	crocs	2	1	1037.420546	74	4	14.019197	2.640428	False	False
1	2022-01-01	1	7	crocs	3	1	720.000000	48	4	15.000000	2.708050	False	True
2	2022-01-01	1	7	dedo	2	1	286.459825	32	4	8.951870	2.191862	False	False
3	2022-01-01	1	7	sandalias	2	1	141.223098	21	4	6.724909	1.905818	False	False
4	2022-01-01	1	7	sandalias	2	1	296.679201	30	4	9.889307	2.291454	False	False

dia_semana	tamaño	material	ventas	unidades_compradas	compras_por_dia	precio	log_precio	charolero	complejos deportivos	stands en playa	tiendas de barrio	dedo	sandalia
7	2	1	1037.420546	74	4	14.019197	2.640428	0	0	1	0	0	
7	3	1	720.000000	48	4	15.000000	2.708050	0	1	0	0	0	
7	2	1	286.459825	32	4	8.951870	2.191862	0	0	1	0	1	
7	2	1	141.223098	21	4	6.724909	1.905818	0	0	0	0	0	
7	2	1	296.679201	30	4	9.889307	2.291454	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	

Metodología y Resultados

Se procedió a aplicar regresión lineal y descartar las variables que no eran representativas, hasta tener información dentro de los parámetros establecidos

OLS Regression Results

Dep. Variable:	log_precio	R-squared:	0.925
Model:	OLS	Adj. R-squared:	0.911
Method:	Least Squares	F-statistic:	66.59
Date:	Thu, 10 Oct 2024	Prob (F-statistic):	0.00
Time:	04:52:34	Log-Likelihood:	5320.2
No. Observations:	4707	AIC:	-9168.
Df Residuals:	3971	BIC:	-4416.
Df Model:	735		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.3642	0.011	219.780	0.000	2.343	2.385
fecha[T.Timestamp('2022-01-02 00:00:00')]	0.1111	0.038	2.926	0.003	0.037	0.185
fecha[T.Timestamp('2022-01-03 00:00:00')]	-0.0210	0.029	-0.730	0.466	-0.077	0.035
fecha[T.Timestamp('2022-01-04 00:00:00')]	-0.0430	0.033	-1.313	0.189	-0.107	0.021
fecha[T.Timestamp('2022-01-05 00:00:00')]	-0.0670	0.035	-1.924	0.054	-0.135	0.001

	feature	VIF
0	charolero	1.000171
1	complejos_deportivos	0.999675
2	compras_por_dia	0.992578
3	dedo	1.000113
4	dia_semana	0.999773
5	fecha	48.130615
6	material	0.998569
7	mes	1.019729
8	sandalias	0.999932
9	stands_en_playa	1.000052
10	tamaño	0.999760
11	tiendas_de_barrio	0.999958
12	unidades_compradas	0.997859
13	ventas	0.999317

	feature	VIF
0	charolero	1.953391
1	complejos_deportivos	1.939984
2	compras_por_dia	9.029213
3	dedo	3.987815
4	dia_semana	4.835794
5	material	1.928871
6	mes	4.191789
7	sandalias	5.106870
8	stands_en_playa	1.989545
9	tamaño	6.124492
10	tiendas_de_barrio	1.931711
11	unidades_compradas	51.518782
12	ventas	46.431827

Metodología y Resultados

Se procedió a aplicar regresión lineal y descartar las variables que no eran representativas, hasta tener información dentro de los parámetros establecidos

OLS Regression Results

Dep. Variable:	log_precio	R-squared:	0.025
Model:	OLS	Adj. R-squared:	0.023
Method:	Least Squares	F-statistic:	15.10
Date:	Thu, 10 Oct 2024	Prob (F-statistic):	4.56e-22
Time:	04:52:34	Log-Likelihood:	-714.70
No. Observations:	4707	AIC:	1447.
Df Residuals:	4698	BIC:	1506.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.2562	0.015	146.355	0.000	2.226	2.286
charolero	0.0129	0.013	0.989	0.323	-0.013	0.039
complejos_deportivos	-0.0007	0.013	-0.053	0.958	-0.026	0.025
dedo	-0.0727	0.009	-8.377	0.000	-0.090	-0.056
dia_semana	0.0107	0.002	5.144	0.000	0.007	0.015
material	-0.0098	0.008	-1.189	0.235	-0.026	0.006
mes	0.0051	0.001	4.261	0.000	0.003	0.007
stands_en_playa	0.0178	0.013	1.371	0.170	-0.008	0.043
tiendas_de_barrio	-0.0004	0.013	-0.031	0.975	-0.026	0.025

Omnibus:	422.547	Durbin-Watson:	1.706
Prob(Omnibus):	0.000	Jarque-Bera (JB):	209.396
Skew:	-0.349	Prob(JB):	3.39e-46
Kurtosis:	2.237	Cond. No.	46.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

GLSAR Regression Results

Dep. Variable:	log_precio	R-squared:	0.800
Model:	GLSAR	Adj. R-squared:	0.800
Method:	Least Squares	F-statistic:	3756.
Date:	Thu, 10 Oct 2024	Prob (F-statistic):	0.00
Time:	04:52:34	Log-Likelihood:	2762.4
No. Observations:	4706	AIC:	-5513.
Df Residuals:	4700	BIC:	-5474.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.5597	0.012	221.681	0.000	2.537	2.582
dedo	-0.3722	0.004	-87.262	0.000	-0.381	-0.364
dia_semana	0.0090	0.002	5.439	0.000	0.006	0.012
material	-0.0047	0.006	-0.742	0.458	-0.017	0.008
mes	0.0055	0.001	4.691	0.000	0.003	0.008
sandalias	-0.5837	0.004	-135.190	0.000	-0.592	-0.575

Omnibus:	7.389	Durbin-Watson:	2.304
Prob(Omnibus):	0.025	Jarque-Bera (JB):	7.370
Skew:	-0.088	Prob(JB):	0.0251
Kurtosis:	3.083	Cond. No.	24.2

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifie

Metodología y Resultados

Se procedió a aplicar regresión lineal y descartar las variables que no eran representativas, hasta tener información dentro de los parámetros establecidos

WLS Regression Results

Dep. Variable:	log_precio	R-squared:	0.979
Model:	WLS	Adj. R-squared:	0.979
Method:	Least Squares	F-statistic:	4.451e+04
Date:	Thu, 10 Oct 2024	Prob (F-statistic):	0.00
Time:	04:52:34	Log-Likelihood:	1036.8
No. Observations:	4707	AIC:	-2062.
Df Residuals:	4701	BIC:	-2023.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.2582	0.001	3321.030	0.000	2.257	2.260
dedo	-0.0734	0.000	-225.648	0.000	-0.074	-0.073
dia_semana	0.0107	0.000	96.299	0.000	0.011	0.011
material	-0.0136	0.000	-42.293	0.000	-0.014	-0.013
mes	0.0064	5.38e-05	119.213	0.000	0.006	0.007
sandalias	-0.0042	0.001	-7.750	0.000	-0.005	-0.003

Omnibus:	1122.286	Durbin-Watson:	1.817
Prob(Omnibus):	0.000	Jarque-Bera (JB):	52540.007
Skew:	-0.278	Prob(JB):	0.00
Kurtosis:	19.358	Cond. No.	99.2

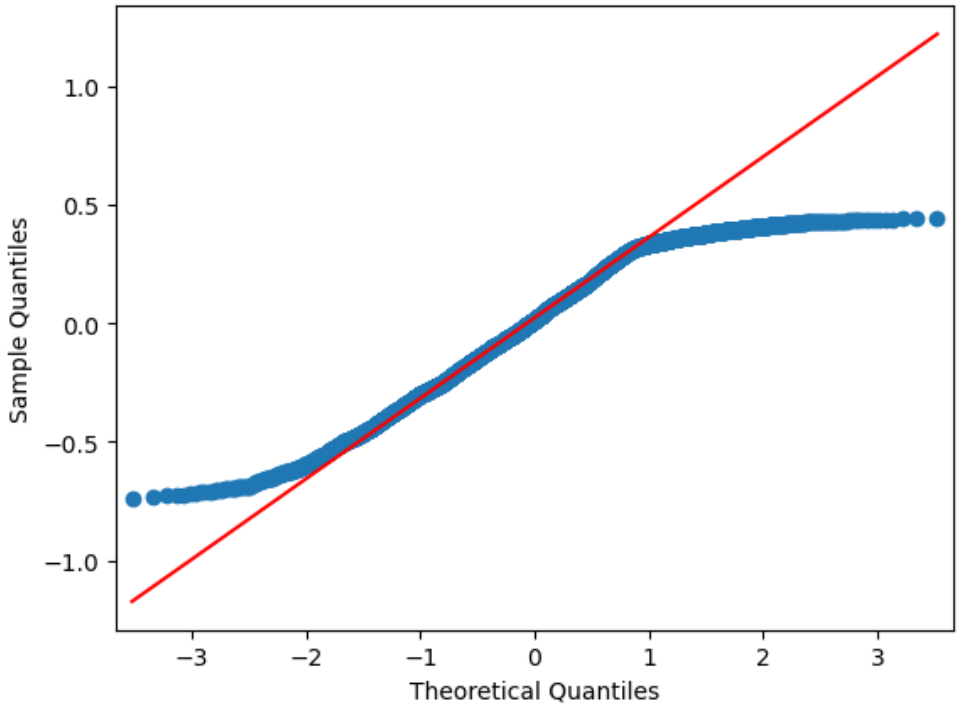
Robust linear Model Regression Results

Dep. Variable:	log_precio	No. Observations:	4707
Model:	RLM	Df Residuals:	4701
Method:	IRLS	Df Model:	5
Norm:	HuberT		
Scale Est.:	mad		
Cov Type:	H1		
Date:	Thu, 10 Oct 2024		
Time:	04:52:34		
No. Iterations:	19		

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.5489	0.008	317.601	0.000	2.533	2.565
dedo	-0.3673	0.006	-63.524	0.000	-0.379	-0.356
dia_semana	0.0122	0.001	10.215	0.000	0.010	0.015
material	-0.0119	0.005	-2.508	0.012	-0.021	-0.003
mes	0.0056	0.001	8.232	0.000	0.004	0.007
sandalias	-0.5787	0.006	-99.460	0.000	-0.590	-0.567

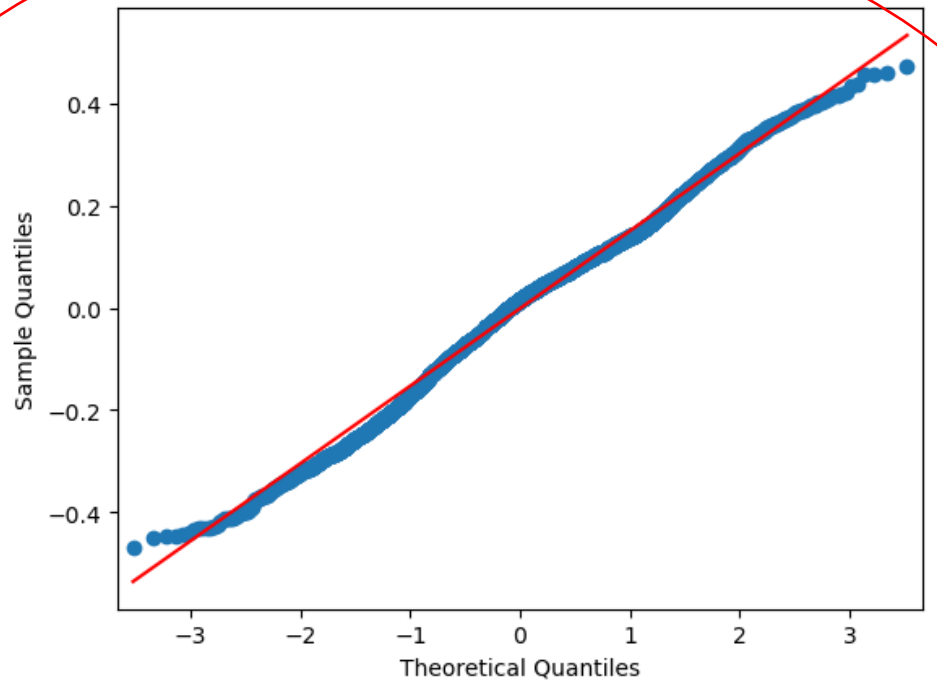
Metodología y Resultados

Se evaluó la normalidad de los residuos por los diferentes tipos de regresión.




```
[('Jarque-Bera', 209.39589936237704),
 ('Chi^2 two-tail prob.', 3.3904565794865562e-46),
 ('Skew', -0.34853436369678126),
 ('Kurtosis', 2.2372676055929266)]
```

```
In [102]: results_3.resid.mean()

Out[102]: -7.13410870873642e-15
```

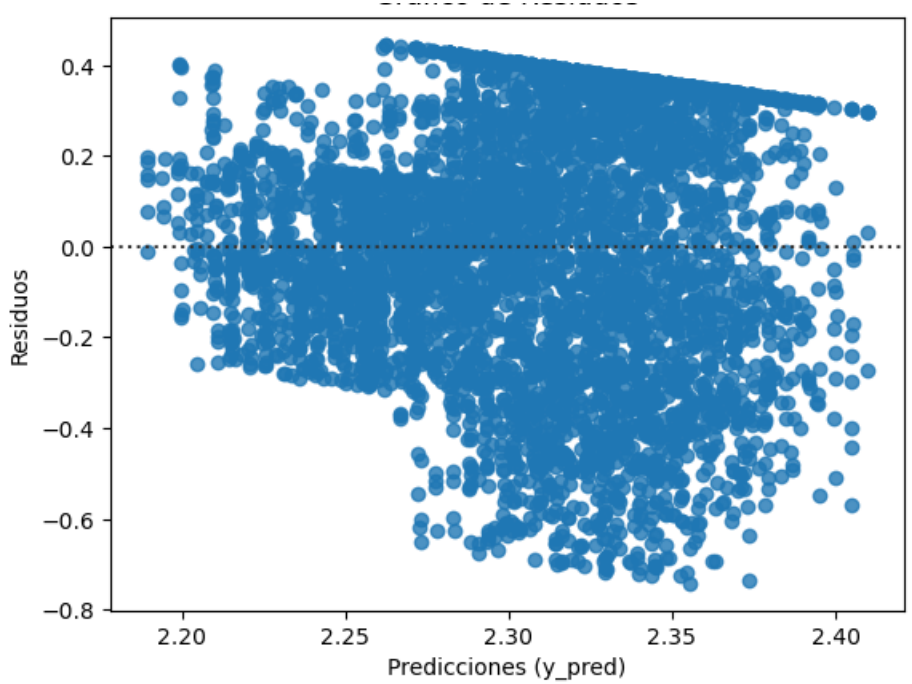



```
[('Jarque-Bera', 21.978446432280354),
 ('Chi^2 two-tail prob.', 1.6882664763152557e-05),
 ('Skew', -0.1577248634018685),
 ('Kurtosis', 2.8879512088779355)]
```

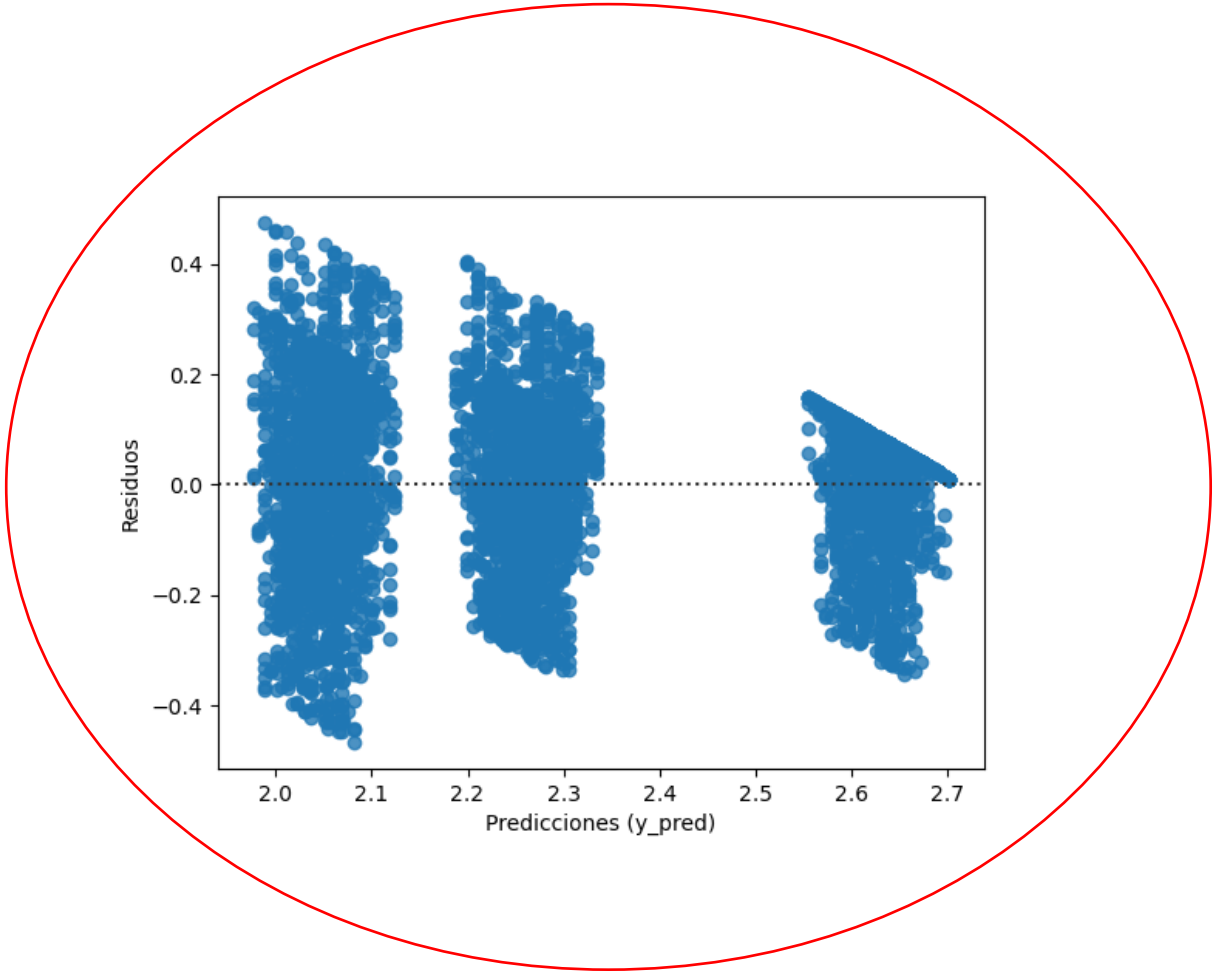
```
In [44]: results_rlm.resid.mean()

Out[44]: -0.0024929895729863833
```

Metodología y Resultados

Se evaluó los residuos por método.




```
[('Lagrange multiplier statistic', 1001.135756901658),
 ('p-value', 8.491328699924583e-211),
 ('f-value', 158.64503788432415),
 ('f p-value', 2.323988555175008e-237)]
```




```
[('Lagrange multiplier statistic', 478.3264754677434),
 ('p-value', 3.800598501358145e-101),
 ('f-value', 106.35073850599936),
 ('f p-value', 1.132399779997633e-106)]
```

Predicción

In [51]:
```
df3=df2[df2.columns.difference(['fecha','compras_por_dia', 'sandalias', 'tamaño', 'unidades_compradas', 'ventas'])]
# df3['intercepto']=1
df3=df3[[ 'charolero', 'complejos_deportivos' , 'dedo', 'dia_semana', 'material', 'mes', 'stands_en_playa', 'tiendas_de_barrio']]
df3
```

Out[51]:

	charolero	complejos_deportivos	dedo	dia_semana	material	mes	stands_en_playa	tiendas_de_barrio
0	0	0	0	7	1	1	1	0
1	0	1	0	7	1	1	0	0
2	0	0	1	7	1	1	1	0
3	0	0	0	7	1	1	0	0
4	0	0	0	7	1	1	0	0
...	...	...	...	...	...	...	...	...
4825	1	0	1	1	1	12	0	0
4826	0	0	1	1	1	12	0	0
4827	0	0	0	1	1	12	0	0
4828	0	0	0	1	1	12	0	1
4829	1	0	1	1	1	12	0	0

4707 rows × 8 columns

In [52]:
```
results_3.predict()
```

Out[52]:
```
array([2.34427687, 2.32576045, 2.2716235 , ..., 2.31817766, 2.31777157,
       2.2584626 ])
```

Estrategias empresariales, recomendaciones, innovación y conclusiones

- No ha existido mayor innovación en este SKU, dado que son sandalias; el modelo es funcional y sistemático, por cuanto se podría utilizar en SKUS que permitan mayor innovación.

- Posiblemente se pueda realizar un análisis directo de tendero a consumidor, donde exista más información, dado que los precios tienden a medir la capacidad de pago de las personas, y no al revés a buscar descuentos

- El modelo arroja que existe una "crecimiento" de 2.25 USD, esto es claro debido a que el descuento máximo es de 2 USD. En crocs, el precio mínimo es de 10 usd mientras que el máximo es de 15 USD; en sandalias de dedo, el máximo es de 13,49 USD y el mínimo es de 7 USD y finalmente en sandalias el máximo es de 12 USD y el mínimo es de 5 USD.

- Es posible que al momento de enmascaramiento de los datos y el haber alterado la información con factores expansivos, hayan cambiado la tendencia y es por eso su disminución en indicadores de fiabilidad.

Gracias!!