



Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Estado de México

Reto House Price: Reporte

Exploración y Análisis de Datos (EDA)

TC3006C. Inteligencia artificial avanzada para la ciencia de datos

Grupo: 101

Integrantes:

Alan Contreras Prieto - A01749667

Alan Rodrigo Vega Reza - A01750658

Ameyalli Contreras Sánchez - A01749075

Iván Alexander Ramos Ramírez - A01750817

Yóse Miguel Sotomayor Carneado - A01750908

Carlos Alberto Zamudio Velázquez - A01799283

Profa. Andrea Torres Calderón

Fecha de entrega: 25 de agosto de 2025

Semestre agosto - diciembre 2025

Índice

Introducción.....	3
1. Descripción de los Datos.....	3
2. Análisis de la variable objetivo.....	6
2.1. Distribución y detección de outliers.....	6
2.2. Pruebas de normalidad.....	7
3. Análisis univariado.....	7
3.1. Variables numéricas.....	7
3.1.1. Detección de Outliers.....	7
3.1.2. Pruebas de normalidad.....	10
3.2. Variables categóricas.....	11
4. Análisis bivariado.....	20
4.1. Correlaciones.....	20
4.1.1. Correlación entre variables predictoras.....	20
4.1.2. Correlación con la variable objetivo.....	22
4.2. Relación SalePrice - Variables.....	24
4.2.1. Relaciones lineales.....	24
4.2.2. Relaciones curvas.....	27
5. Valores faltantes y outliers.....	28
5.1. Valores faltantes.....	28
5.2. Outliers en las variables principales.....	30
6. Recomendaciones basadas en los hallazgos.....	33
Conclusiones.....	35
Anexo.....	35

Introducción

El objetivo principal del presente reporte es generar un análisis exploratorio de datos sobre el dataset obtenido de la plataforma Kaggle *train.csv*, correspondiente al reto de regresión *House Price - Advanced Regression*. Este análisis servirá como base para el desarrollo de un modelo de regresión que permita estimar el precio de diferentes viviendas según sus características.

A través de este EDA, se busca entender la estructura del dataset, evaluar la calidad de los datos y las distribuciones de las variables, así como descubrir las relaciones clave entre las características de las viviendas y su precio. Se explorarán preguntas como: ¿qué variables tienen mayor influencia en el precio?, y ¿existen datos atípicos que puedan afectar al modelo?. Este análisis inicial es fundamental para asegurar la calidad de los datos, guiar la selección de características para el modelo predictivo y proporcionar recomendaciones relevantes sobre las transformaciones que deben hacerse a los datos para que el modelo pueda alcanzar un mejor desempeño.

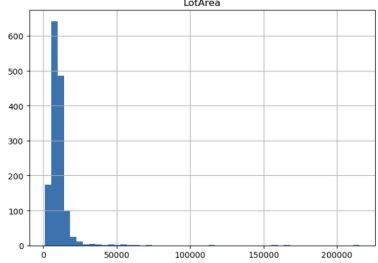
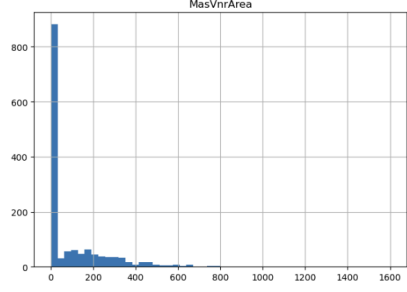
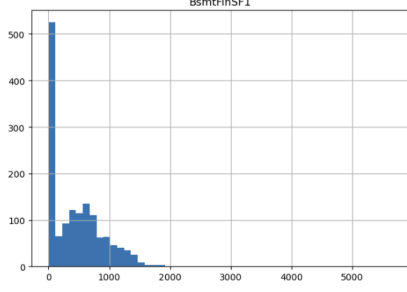
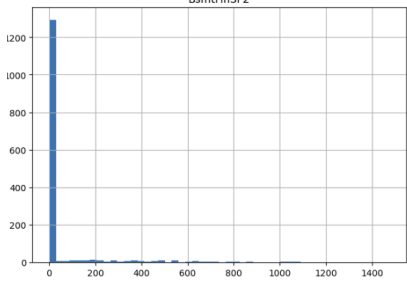
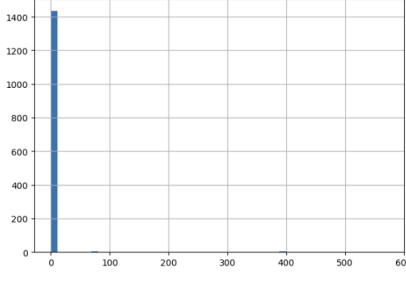
1. Descripción de los Datos

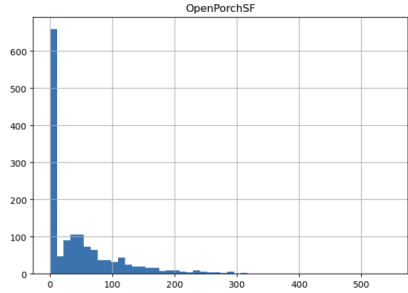
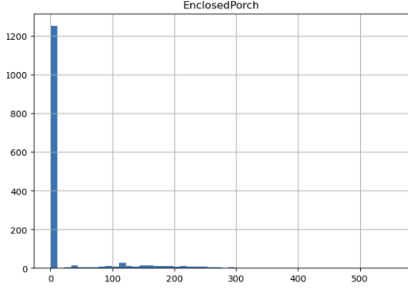
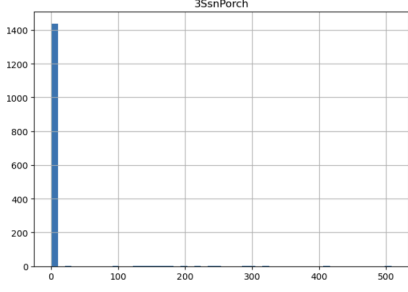
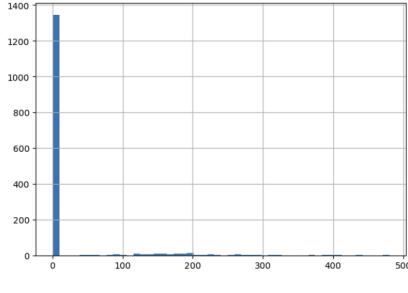
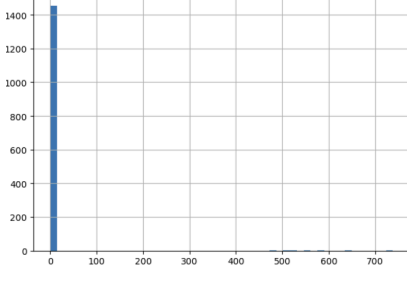
En primera instancia se hizo un análisis descriptivo general de la base de datos, donde se encontró que este cuenta con 1460 datos y 80 variables, entre las cuales se encuentra nuestra variable objetivo *SalePrice*. El dataset cuenta con 37 variables numéricas (incluyendo la variable objetivo) y 43 variables categóricas. Sin embargo, al leer la descripción de lo que representa cada una de las variables, se encontró que hay tres variables numéricas que realmente cumplen una función de variables categóricas, ya que los números de estas columnas representan el estatus o código de alguna característica de las casas, estas son:

- *MSSubClass*: Que con códigos numéricos describe el tipo de vivienda involucrada en la venta.
- *OverallQual*: Que califica del 1 al 10 el material general y el acabado de la casa.
- *OverallCond*: Que califica del 1 al 10 el estado general en el que se encuentra la casa.

Posteriormente se analizó la estadística descriptiva de los datos obtenida con *.describe()* dentro de la cual se detectó una alta variación en los rangos de valores de algunas variables, para las cuales se obtuvieron sus histogramas y tras visualizarlas, varias de ellas presentaron una distribución altamente desbalanceada, es decir, que la mayoría de los datos están concentrados en un rango muy pequeño de valores del dominio total de la variable, las cuales

se presentan a continuación con su rango de valores (mínimo y máximo), su media y su histograma:

Variable	Min	Max	Media	Histograma
LotArea	1300	215245	10516.828082	
MasVnrArea	0	1600	103.685262	
BsmtFinSF1	0	5644	443.639726	
BsmtFinSF2	0	1474	46.549315	
LowQualFinSF	0	572	5.844521	

OpenPorchSF	0	547	46.660274	 <p>Histogram for OpenPorchSF. The x-axis ranges from 0 to 500, and the y-axis ranges from 0 to 600. The distribution is highly right-skewed, with a very high frequency (over 600) at the lowest values (near 0) and a long tail extending towards 500.</p>
EnclosedPorch	0	552	21.95411	 <p>Histogram for EnclosedPorch. The x-axis ranges from 0 to 500, and the y-axis ranges from 0 to 1200. The distribution is highly right-skewed, with a peak frequency of approximately 1200 at the lowest values (near 0) and a long tail extending towards 500.</p>
3SsnPorch	0	508	3.409589	 <p>Histogram for 3SsnPorch. The x-axis ranges from 0 to 500, and the y-axis ranges from 0 to 1400. The distribution is highly right-skewed, with a peak frequency of approximately 1400 at the lowest values (near 0) and a long tail extending towards 500.</p>
ScreenPorch	0	480	15.060959	 <p>Histogram for ScreenPorch. The x-axis ranges from 0 to 500, and the y-axis ranges from 0 to 1400. The distribution is highly right-skewed, with a peak frequency of approximately 1400 at the lowest values (near 0) and a long tail extending towards 500.</p>
PoolArea	0	738	2.758904	 <p>Histogram for PoolArea. The x-axis ranges from 0 to 700, and the y-axis ranges from 0 to 1400. The distribution is highly right-skewed, with a peak frequency of approximately 1400 at the lowest values (near 0) and a long tail extending towards 700.</p>

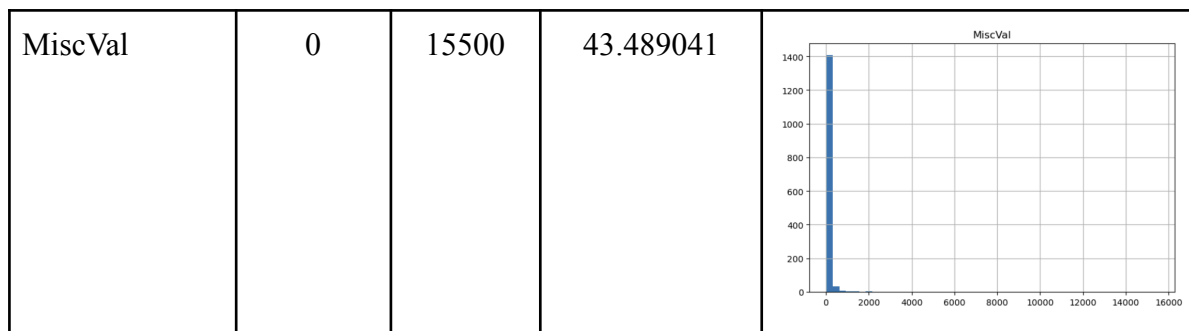


Tabla 1. Variación de rango de datos

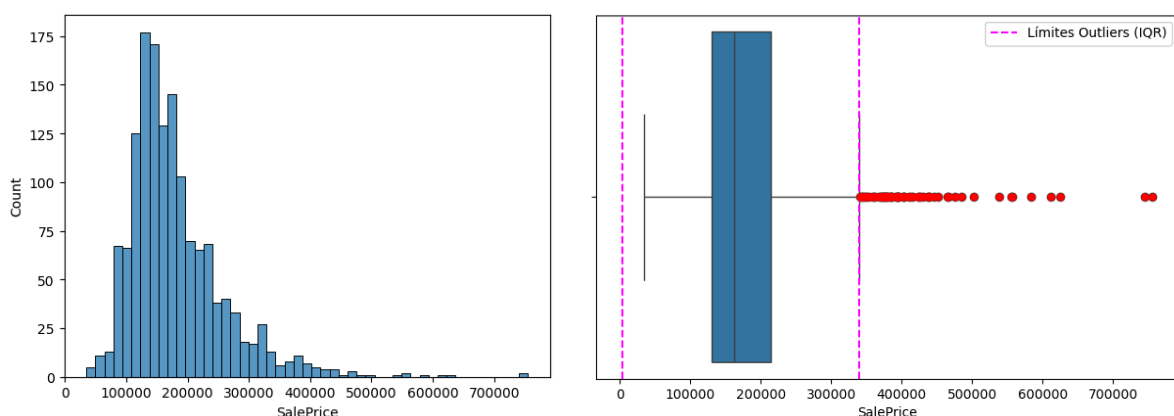
2. Análisis de la variable objetivo

2.1. Distribución y detección de outliers

Enseguida, se realizó el análisis de la variable objetivo *SalePrice*. Dentro del cual se identificó que la variable tiene en general una distribución relativamente simétrica según se observa en el histograma y el boxplot, debido a que su mediana se mantiene cerca del centro del rango intercuartil, aunque con un ligero sesgo hacia la derecha. Esto es normal debido a que, en general, la mayoría de las casas tienen precios moderados, mientras que un pequeño número de casas de lujo tienen precios mucho más altos.

Sin embargo, esta distribución se encuentra cargada hacia la izquierda y se detecta la presencia de una gran cantidad de datos atípicos, es decir, valores notablemente alejados de la mayoría de las observaciones, ya que se encuentran muy por encima del rango intercuartílico (IQR), lo que las convierte en observaciones inusuales sugiriendo que se trata de una desviación extrema del comportamiento habitual de los datos.

Para identificar los outliers se utilizó el método del *rango intercuartílico (IQR) 1.5*, ya que ayuda a encontrar valores que se desvían de la norma sin asumir una distribución normal de los datos, consiguiendo así visualizarlos de mejor manera, como resultado de esto se encontró un total de 61 outliers o valores extremos en la distribución de *SalePrice*.



2.2. Pruebas de normalidad

Además se le aplicó un análisis de normalidad Jarque Bera y Shapiro Wilks, donde se observó que pese a su distribución aparentemente balanceada, la variable objetivo no sigue una distribución normal:

Estadístico Jarque-Bera: 3438.8721241220583	Estadístico Shapiro-Wilk: 0.8696714665902145
Valor p: 0.0	Valor p: 3.2061412312021656e-33
La distribución de SalePrice NO es normal	La distribución de SalePrice NO es normal

Lo que se podría hacer para poder hacer válido el modelo final para inferencia de parámetros y no solo para predicciones, sería probar aplicar alguna transformación del tipo logarítmica, raíz cuadrada, Yeo-Johnson, entre otros a *SalePrice* para intentar normalizarla, en caso de que esto no ocurra, se tendrá que verificar la normalidad de residuos al concluir el modelo. De no conseguir ninguno de estos supuestos, sólo se podrá decir que el modelo es meramente predictivo y no puede ser usado para inferencia de parámetros.

3. Análisis univariado

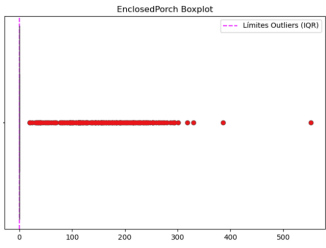
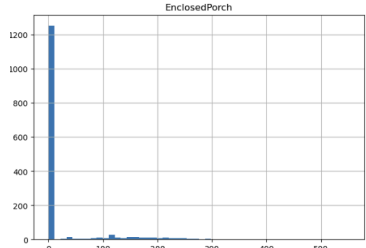
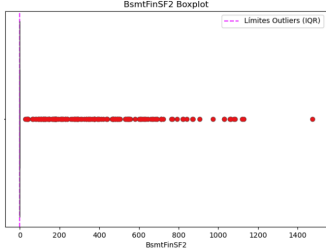
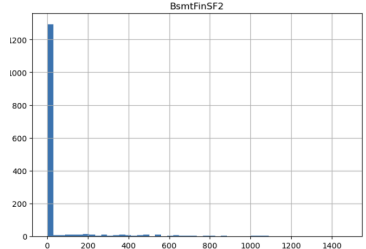
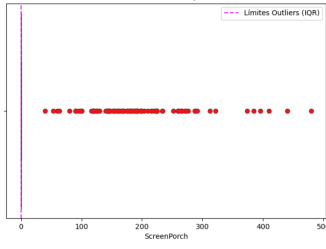
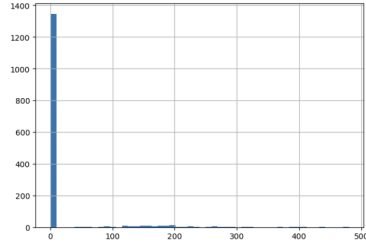
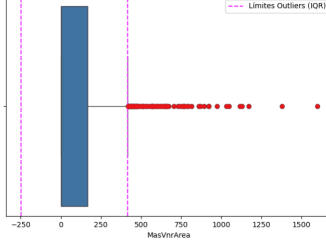
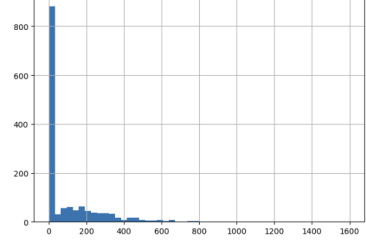
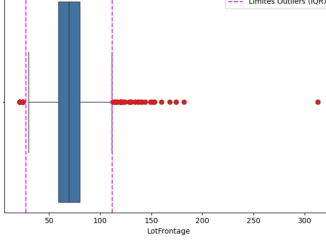
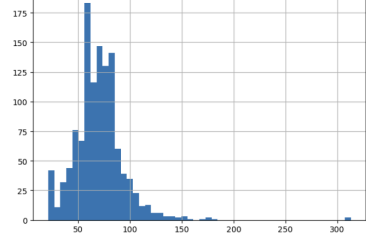
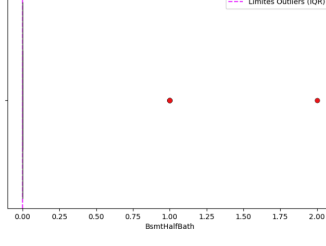
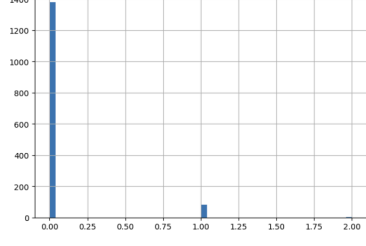
El siguiente paso del análisis exploratorio consistió en analizar las variables individualmente, se sesionará esta parte del análisis en variables numéricas y categóricas, debido a que las herramientas de análisis son distintas y con el objetivo de proporcionar una mejor comprensión de las mismas.

3.1. Variables numéricas

3.1.1. Detección de Outliers

Se obtuvieron el histograma y boxplot de cada una de las variables numéricas, haciendo también la identificación de outliers con el método de *IQR 1.5*. De forma general se puede decir que a excepción de *YearRemodAdd*, *FullBath*, *HalfBath*, *GarageYrBlt*, *MoSold* y *YrSold*, las otras 30 variables numéricas tienen outliers.

Utilizando un criterio del 4%, aquellas variables que tienen mayor presencia de outliers son las presentadas en la siguiente tabla (se omitieron aquellas variables que son categóricas mencionadas en la sección *Descripción de los datos* que son de tipo numérico y la variable objetivo pues ya se analizó en la sección anterior):

Variable	# Outliers	% Outliers	Boxplot	Histograma
EnclosedPorch	208	14.246575		
BsmtFinSF2	167	11.438356		
ScreenPorch	116	7.945205		
MasVnrArea	96	6.575342		
LotFrontage	88	6.027397		
BsmtHalfBath	82	5.616438		

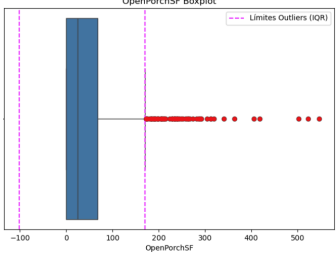
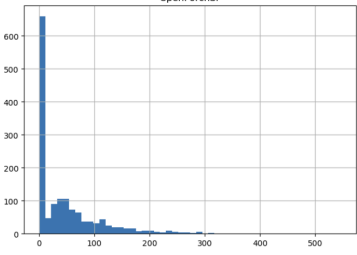
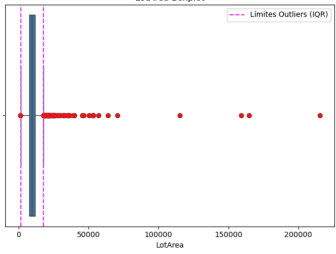
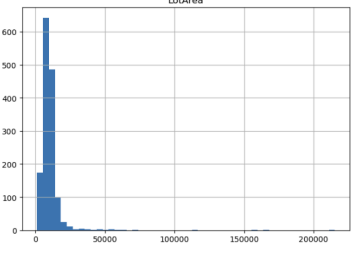
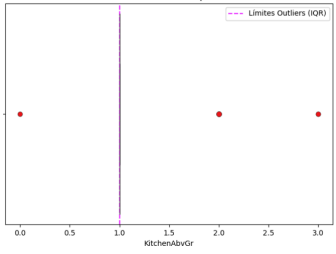
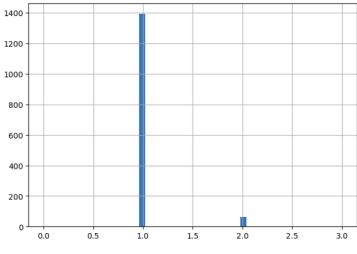
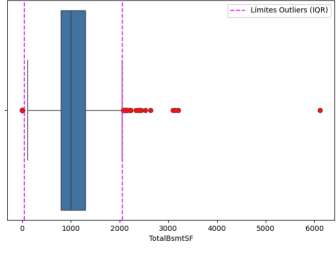
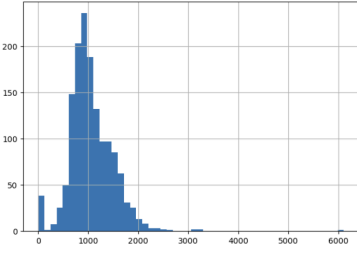
OpenPorchSF	77	5.273973		
LotArea	69	4.726027		
KitchenAbvGr	68	4.657534		
TotalBsmtSF	61	4.178082		

Tabla 2. Variables numéricas con más del 4% de outliers

Como se puede apreciar en los gráficos de *BsmtHalfBath* y *KitchenAbvGr*, sus valores numéricos son enteros pequeños (de 0 a 2 y de 0 a 3 respectivamente), por lo que los valores menos frecuentes son tomados como outliers por el *IQR 1.5*, sin embargo, al tratarse de un rango total tan pequeño de valores ni siquiera podemos apreciar claramente su boxplot y por ello, no se deben considerar como outliers peligrosos para el modelo, ya que son valores perfectamente válidos y lógicamente posibles. En lugar de ser errores o valores extremos que desvían el análisis, son simplemente valores poco frecuentes dentro de un rango de posibilidades válidas, cosa que se puede apreciar en sus histogramas.

Además se observa que algunas de las variables con más outliers, son también parte de la *Tabla 1. Variación de rango de datos*, presente en la sección *Descripción de los datos*, lo que indica que el que una variable tenga un rango de datos demasiado grande y además no tenga

una distribución balanceada puede presentar mayor cantidad de outliers. Esto podría tener un impacto perjudicial en modelos de regresión, principalmente en los que utilizan el método de *Mínimos Cuadrados Ordinarios (OLS)*, ya que pueden generar sesgos debido a que su error residual es mayor, provocando una especie de atracción que puede alterar la precisión del modelo en los demás datos. En consecuencia esto puede ocasionar que el modelo haga predicciones menos precisas para nuevas observaciones. Además, en modelos como la regresión lineal podría ocasionar la violación de supuestos como la normalidad de los residuos y la homocedasticidad, lo que puede llevar a intervalos de confianza y *p-values* inválidos, lo que a su vez genera conclusiones erróneas. Finalmente, los outliers pueden incrementar el error estándar de los coeficientes, haciendo que estos parezcan no significativos, cuando sí lo son.

Una manera de evitar este tipo de problemas en el modelo, es eliminando los outliers más lejanos, sin embargo, es importante considerar su relación con nuestra variable objetivo (análisis que se hará en la sección siguiente *Análisis Bivariado*), ya que puede que se trate de valores extremos que son perfectamente posibles dada la naturaleza del problema, por ejemplo, que se trate de una mansión y por consiguiente el precio sea bastante elevado. En este tipo de casos se justifica la presencia de estos outliers y su eliminación podría ocasionar que el modelo no pueda predecir con precisión el precio de este tipo de propiedades.

3.1.2. Pruebas de normalidad

Adicionalmente se analizó la normalidad de cada una de las variables predictoras con la finalidad de poder realizar una mejor selección de modelo de regresión y saber si es necesario realizar algún tipo de transformación en los datos. Para ello se aplicaron las pruebas de Jarque Bera y Shapiro Wilks, las cuales arrojaron como resultado que ninguna de las variables predictoras cuenta con una distribución normal. Esto indica que el conjunto de datos es heterogéneo y el conjunto de datos en general no cumple con el supuesto de normalidad que requieren algunos modelos estadísticos.

Algunas de las razones por las que estas variables resultaron no ser normales, son el hecho de que prácticamente todas tienen valores atípicos extremos, distribuciones sesgadas, o un rango muy pequeño de valores. Esta falta de normalidad afecta principalmente a modelos de regresión lineal tradicionales, ya que estos asumen que los residuos están distribuidos normalmente y la falta de normalidad en las variables viola este supuesto, lo que puede afectar la precisión de los coeficientes del modelo y la validez de las inferencias estadísticas.

Algunas soluciones que podrían ayudar a mitigar este problema son las siguientes:

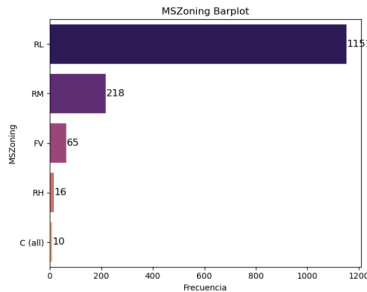
- Utilizar transformaciones como la logarítmica, la de raíz cuadrada o la de box-cox sobre los datos, para que estos asemejen más a una distribución normal, lo que mejoraría el rendimiento de los modelos lineales.
- Emplear modelos de regresión que no asuman la normalidad, como los modelos basados en árboles (Random Forest, Gradient Boosting) o de regresión robusta (Huber, Quantile regression) ya que estos no son sensibles a la distribución de las variables y manejan los valores extremos de manera más eficiente.

3.2. Variables categóricas

Para el análisis de variables categóricas se emplearon gráficos de barras para visualizar la frecuencia de cada clase de cada variable.

Tras realizar un análisis de estos gráficos, se observó que ninguna de las variables categóricas está distribuida de manera uniforme, ya que en todas hay una categoría predominante, y aquellas variables que poseen más de dos categorías tienden presentar distribuciones mucho más desiguales, lo que puede interpretarse como un sesgo en los datos, sin embargo, aquellas categorías con barras muy cortas, a pesar de ser poco comunes, pueden ser relevantes para predecir el precio de una vivienda.

A continuación se presenta una tabla con la categoría más frecuente para cada una de las variables, junto con su gráfico para tener una referencia visual de su distribución:

Variable	Categoría más frecuente	Categoría menos frecuente	Barplot												
MSZoning	RL	C (all)	 <p>MSZoning Barplot</p> <table><thead><tr><th>MSZoning</th><th>Frecuencia</th></tr></thead><tbody><tr><td>RL</td><td>1151</td></tr><tr><td>RM</td><td>218</td></tr><tr><td>FV</td><td>65</td></tr><tr><td>RH</td><td>16</td></tr><tr><td>C (all)</td><td>10</td></tr></tbody></table>	MSZoning	Frecuencia	RL	1151	RM	218	FV	65	RH	16	C (all)	10
MSZoning	Frecuencia														
RL	1151														
RM	218														
FV	65														
RH	16														
C (all)	10														

Street	Pave	Grvl	<div>Street Barplot</div> <table><tr><th>Category</th><th>Frecuencia</th></tr><tr><td>Pave</td><td>1454</td></tr><tr><td>Grvl</td><td>6</td></tr></table>	Category	Frecuencia	Pave	1454	Grvl	6				
Category	Frecuencia												
Pave	1454												
Grvl	6												
Alley	Grvl	Pave	<div>Alley Barplot</div> <table><tr><th>Category</th><th>Frecuencia</th></tr><tr><td>Grvl</td><td>50</td></tr><tr><td>Pave</td><td>41</td></tr></table>	Category	Frecuencia	Grvl	50	Pave	41				
Category	Frecuencia												
Grvl	50												
Pave	41												
LotShape	Reg	IR3	<div>LotShape Barplot</div> <table><tr><th>Category</th><th>Frecuencia</th></tr><tr><td>Reg</td><td>925</td></tr><tr><td>IR1</td><td>484</td></tr><tr><td>IR2</td><td>41</td></tr><tr><td>IR3</td><td>10</td></tr></table>	Category	Frecuencia	Reg	925	IR1	484	IR2	41	IR3	10
Category	Frecuencia												
Reg	925												
IR1	484												
IR2	41												
IR3	10												
LandContour	Lvl	Low	<div>LandContour Barplot</div> <table><tr><th>Category</th><th>Frecuencia</th></tr><tr><td>Lvl</td><td>1311</td></tr><tr><td>Brk</td><td>63</td></tr><tr><td>HLS</td><td>50</td></tr><tr><td>Low</td><td>36</td></tr></table>	Category	Frecuencia	Lvl	1311	Brk	63	HLS	50	Low	36
Category	Frecuencia												
Lvl	1311												
Brk	63												
HLS	50												
Low	36												
Utilities	AllPub	NoSeWa	<div>Utilities Barplot</div> <table><tr><th>Category</th><th>Frecuencia</th></tr><tr><td>AllPub</td><td>1459</td></tr><tr><td>NoSeWa</td><td>1</td></tr></table>	Category	Frecuencia	AllPub	1459	NoSeWa	1				
Category	Frecuencia												
AllPub	1459												
NoSeWa	1												

LotConfig	Inside	FR3	<div><div>LotConfig Barplot</div></div>
LandSlope	Gtl	Sev	<div><div>LandSlope Barplot</div></div>
Neighborhood	NAMES	Blueste	<div><div>Neighborhood Barplot</div></div>
Condition1	Norm	RRNe	<div><div>Condition1 Barplot</div></div>
Condition2	Norm	PosA	<div><div>Condition2 Barplot</div></div>

BldgType	1Fam	2fmCon	<div>BldgType Barplot</div> <table><thead><tr><th>BldgType</th><th>Frecuencia</th></tr></thead><tbody><tr><td>1Fam</td><td>1220</td></tr><tr><td>TwHseE</td><td>114</td></tr><tr><td>Duplex</td><td>52</td></tr><tr><td>TwHse</td><td>43</td></tr><tr><td>2fmCon</td><td>31</td></tr></tbody></table>	BldgType	Frecuencia	1Fam	1220	TwHseE	114	Duplex	52	TwHse	43	2fmCon	31																				
BldgType	Frecuencia																																		
1Fam	1220																																		
TwHseE	114																																		
Duplex	52																																		
TwHse	43																																		
2fmCon	31																																		
HouseStyle	1Story	2.5Fin	<div>HouseStyle Barplot</div> <table><thead><tr><th>HouseStyle</th><th>Frecuencia</th></tr></thead><tbody><tr><td>1Story</td><td>726</td></tr><tr><td>2Story</td><td>445</td></tr><tr><td>1.5Fin</td><td>154</td></tr><tr><td>SLvl</td><td>65</td></tr><tr><td>SFoyer</td><td>37</td></tr><tr><td>1.5Unf</td><td>14</td></tr><tr><td>2.5Unf</td><td>11</td></tr><tr><td>2.5Fin</td><td>8</td></tr></tbody></table>	HouseStyle	Frecuencia	1Story	726	2Story	445	1.5Fin	154	SLvl	65	SFoyer	37	1.5Unf	14	2.5Unf	11	2.5Fin	8														
HouseStyle	Frecuencia																																		
1Story	726																																		
2Story	445																																		
1.5Fin	154																																		
SLvl	65																																		
SFoyer	37																																		
1.5Unf	14																																		
2.5Unf	11																																		
2.5Fin	8																																		
RoofStyle	Gable	Shed	<div>RoofStyle Barplot</div> <table><thead><tr><th>RoofStyle</th><th>Frecuencia</th></tr></thead><tbody><tr><td>Gable</td><td>1141</td></tr><tr><td>Hip</td><td>286</td></tr><tr><td>Flat</td><td>13</td></tr><tr><td>Gambrel</td><td>11</td></tr><tr><td>Mansard</td><td>7</td></tr><tr><td>Shed</td><td>2</td></tr></tbody></table>	RoofStyle	Frecuencia	Gable	1141	Hip	286	Flat	13	Gambrel	11	Mansard	7	Shed	2																		
RoofStyle	Frecuencia																																		
Gable	1141																																		
Hip	286																																		
Flat	13																																		
Gambrel	11																																		
Mansard	7																																		
Shed	2																																		
RoofMatl	CompShg	Metal	<div>RoofMatl Barplot</div> <table><thead><tr><th>RoofMatl</th><th>Frecuencia</th></tr></thead><tbody><tr><td>CompShg</td><td>1434</td></tr><tr><td>Tar&Grv</td><td>11</td></tr><tr><td>WdShngl</td><td>6</td></tr><tr><td>WdShake</td><td>5</td></tr><tr><td>Metal</td><td>1</td></tr><tr><td>Membran</td><td>1</td></tr><tr><td>Roll</td><td>1</td></tr><tr><td>ClyTile</td><td>1</td></tr></tbody></table>	RoofMatl	Frecuencia	CompShg	1434	Tar&Grv	11	WdShngl	6	WdShake	5	Metal	1	Membran	1	Roll	1	ClyTile	1														
RoofMatl	Frecuencia																																		
CompShg	1434																																		
Tar&Grv	11																																		
WdShngl	6																																		
WdShake	5																																		
Metal	1																																		
Membran	1																																		
Roll	1																																		
ClyTile	1																																		
Exterior1st	VinylSd	AsphShn	<div>Exterior1st Barplot</div> <table><thead><tr><th>Exterior1st</th><th>Frecuencia</th></tr></thead><tbody><tr><td>VinylSd</td><td>515</td></tr><tr><td>HdBoard</td><td>222</td></tr><tr><td>MetalSd</td><td>220</td></tr><tr><td>Wd Sdng</td><td>206</td></tr><tr><td>Plywood</td><td>108</td></tr><tr><td>CemntBd</td><td>61</td></tr><tr><td>BrkFace</td><td>50</td></tr><tr><td>WdShing</td><td>26</td></tr><tr><td>Stucco</td><td>25</td></tr><tr><td>AsbShng</td><td>20</td></tr><tr><td>BrkComm</td><td>2</td></tr><tr><td>Stone</td><td>2</td></tr><tr><td>AsphShn</td><td>1</td></tr><tr><td>ImStucc</td><td>1</td></tr><tr><td>CBlock</td><td>1</td></tr></tbody></table>	Exterior1st	Frecuencia	VinylSd	515	HdBoard	222	MetalSd	220	Wd Sdng	206	Plywood	108	CemntBd	61	BrkFace	50	WdShing	26	Stucco	25	AsbShng	20	BrkComm	2	Stone	2	AsphShn	1	ImStucc	1	CBlock	1
Exterior1st	Frecuencia																																		
VinylSd	515																																		
HdBoard	222																																		
MetalSd	220																																		
Wd Sdng	206																																		
Plywood	108																																		
CemntBd	61																																		
BrkFace	50																																		
WdShing	26																																		
Stucco	25																																		
AsbShng	20																																		
BrkComm	2																																		
Stone	2																																		
AsphShn	1																																		
ImStucc	1																																		
CBlock	1																																		

Exterior2nd	VinylSd	Other	<div>Exterior1st Barplot</div> <table><thead><tr><th>Exterior1st</th><th>Frecuencia</th></tr></thead><tbody><tr><td>VinylSd</td><td>513</td></tr><tr><td>HdBoard</td><td>222</td></tr><tr><td>MetaSd</td><td>220</td></tr><tr><td>Wd Sdng</td><td>206</td></tr><tr><td>Plywood</td><td>108</td></tr><tr><td>CemntBd</td><td>61</td></tr><tr><td>BrkFace</td><td>50</td></tr><tr><td>WdShing</td><td>26</td></tr><tr><td>Stucco</td><td>25</td></tr><tr><td>AsbShng</td><td>20</td></tr><tr><td>BrkComm</td><td>2</td></tr><tr><td>Stone</td><td>2</td></tr><tr><td>AsphShn</td><td>1</td></tr><tr><td>ImStucc</td><td>1</td></tr><tr><td>CBlock</td><td>1</td></tr></tbody></table>	Exterior1st	Frecuencia	VinylSd	513	HdBoard	222	MetaSd	220	Wd Sdng	206	Plywood	108	CemntBd	61	BrkFace	50	WdShing	26	Stucco	25	AsbShng	20	BrkComm	2	Stone	2	AsphShn	1	ImStucc	1	CBlock	1
Exterior1st	Frecuencia																																		
VinylSd	513																																		
HdBoard	222																																		
MetaSd	220																																		
Wd Sdng	206																																		
Plywood	108																																		
CemntBd	61																																		
BrkFace	50																																		
WdShing	26																																		
Stucco	25																																		
AsbShng	20																																		
BrkComm	2																																		
Stone	2																																		
AsphShn	1																																		
ImStucc	1																																		
CBlock	1																																		
MasVnrType	BrkFace	BrkCmn	<div>MasVnrType Barplot</div> <table><thead><tr><th>MasVnrType</th><th>Frecuencia</th></tr></thead><tbody><tr><td>BrkFace</td><td>445</td></tr><tr><td>Stone</td><td>128</td></tr><tr><td>BrkCmn</td><td>15</td></tr></tbody></table>	MasVnrType	Frecuencia	BrkFace	445	Stone	128	BrkCmn	15																								
MasVnrType	Frecuencia																																		
BrkFace	445																																		
Stone	128																																		
BrkCmn	15																																		
ExterQual	TA	Fa	<div>ExterQual Barplot</div> <table><thead><tr><th>ExterQual</th><th>Frecuencia</th></tr></thead><tbody><tr><td>TA</td><td>906</td></tr><tr><td>Gd</td><td>488</td></tr><tr><td>Ex</td><td>52</td></tr><tr><td>Fa</td><td>14</td></tr></tbody></table>	ExterQual	Frecuencia	TA	906	Gd	488	Ex	52	Fa	14																						
ExterQual	Frecuencia																																		
TA	906																																		
Gd	488																																		
Ex	52																																		
Fa	14																																		
ExterCond	TA	Po	<div>ExterCond Barplot</div> <table><thead><tr><th>ExterCond</th><th>Frecuencia</th></tr></thead><tbody><tr><td>TA</td><td>1282</td></tr><tr><td>Gd</td><td>146</td></tr><tr><td>Fa</td><td>28</td></tr><tr><td>Ex</td><td>3</td></tr><tr><td>Po</td><td>1</td></tr></tbody></table>	ExterCond	Frecuencia	TA	1282	Gd	146	Fa	28	Ex	3	Po	1																				
ExterCond	Frecuencia																																		
TA	1282																																		
Gd	146																																		
Fa	28																																		
Ex	3																																		
Po	1																																		
Foundation	PConc	Wood	<div>Foundation Barplot</div> <table><thead><tr><th>Foundation</th><th>Frecuencia</th></tr></thead><tbody><tr><td>PConc</td><td>647</td></tr><tr><td>CBlock</td><td>634</td></tr><tr><td>BrkTil</td><td>146</td></tr><tr><td>Slab</td><td>24</td></tr><tr><td>Stone</td><td>6</td></tr><tr><td>Wood</td><td>3</td></tr></tbody></table>	Foundation	Frecuencia	PConc	647	CBlock	634	BrkTil	146	Slab	24	Stone	6	Wood	3																		
Foundation	Frecuencia																																		
PConc	647																																		
CBlock	634																																		
BrkTil	146																																		
Slab	24																																		
Stone	6																																		
Wood	3																																		

BsmtQual	TA	Fa	<div><div>BsmtQual Barplot</div><table><tr><th>BsmtQual</th><th>Frecuencia</th></tr><tr><td>TA</td><td>649</td></tr><tr><td>Gd</td><td>618</td></tr><tr><td>Ex</td><td>121</td></tr><tr><td>Fa</td><td>35</td></tr></table></div>	BsmtQual	Frecuencia	TA	649	Gd	618	Ex	121	Fa	35				
BsmtQual	Frecuencia																
TA	649																
Gd	618																
Ex	121																
Fa	35																
BsmtCond	TA	Po	<div><div>BsmtCond Barplot</div><table><tr><th>BsmtCond</th><th>Frecuencia</th></tr><tr><td>TA</td><td>1311</td></tr><tr><td>Gd</td><td>65</td></tr><tr><td>Fa</td><td>45</td></tr><tr><td>Po</td><td>2</td></tr></table></div>	BsmtCond	Frecuencia	TA	1311	Gd	65	Fa	45	Po	2				
BsmtCond	Frecuencia																
TA	1311																
Gd	65																
Fa	45																
Po	2																
BsmtExposure	No	Mn	<div><div>BsmtExposure Barplot</div><table><tr><th>BsmtExposure</th><th>Frecuencia</th></tr><tr><td>No</td><td>953</td></tr><tr><td>Av</td><td>221</td></tr><tr><td>Gd</td><td>134</td></tr><tr><td>Mn</td><td>114</td></tr></table></div>	BsmtExposure	Frecuencia	No	953	Av	221	Gd	134	Mn	114				
BsmtExposure	Frecuencia																
No	953																
Av	221																
Gd	134																
Mn	114																
BsmtFinType1	Unf	LwQ	<div><div>BsmtFinType1 Barplot</div><table><tr><th>BsmtFinType1</th><th>Frecuencia</th></tr><tr><td>Unf</td><td>430</td></tr><tr><td>GLQ</td><td>418</td></tr><tr><td>ALQ</td><td>220</td></tr><tr><td>BLO</td><td>148</td></tr><tr><td>Rec</td><td>133</td></tr><tr><td>LwQ</td><td>74</td></tr></table></div>	BsmtFinType1	Frecuencia	Unf	430	GLQ	418	ALQ	220	BLO	148	Rec	133	LwQ	74
BsmtFinType1	Frecuencia																
Unf	430																
GLQ	418																
ALQ	220																
BLO	148																
Rec	133																
LwQ	74																
BsmtFinType2	Unf	GLQ	<div><div>BsmtFinType2 Barplot</div><table><tr><th>BsmtFinType2</th><th>Frecuencia</th></tr><tr><td>Unf</td><td>1256</td></tr><tr><td>Rec</td><td>54</td></tr><tr><td>LwQ</td><td>46</td></tr><tr><td>BLO</td><td>33</td></tr><tr><td>ALQ</td><td>19</td></tr><tr><td>GLQ</td><td>14</td></tr></table></div>	BsmtFinType2	Frecuencia	Unf	1256	Rec	54	LwQ	46	BLO	33	ALQ	19	GLQ	14
BsmtFinType2	Frecuencia																
Unf	1256																
Rec	54																
LwQ	46																
BLO	33																
ALQ	19																
GLQ	14																

Heating	GasA	Floor	<div><div>Heating Barplot</div><table><tr><th>Category</th><th>Frecuencia</th></tr><tr><td>GasA</td><td>1428</td></tr><tr><td>GasW</td><td>18</td></tr><tr><td>Grav</td><td>7</td></tr><tr><td>Wall</td><td>4</td></tr><tr><td>OthW</td><td>2</td></tr><tr><td>Floor</td><td>1</td></tr></table></div>	Category	Frecuencia	GasA	1428	GasW	18	Grav	7	Wall	4	OthW	2	Floor	1
Category	Frecuencia																
GasA	1428																
GasW	18																
Grav	7																
Wall	4																
OthW	2																
Floor	1																
HeatingQC	Ex	Po	<div><div>HeatingQC Barplot</div><table><tr><th>Category</th><th>Frecuencia</th></tr><tr><td>Ex</td><td>741</td></tr><tr><td>TA</td><td>428</td></tr><tr><td>Gd</td><td>241</td></tr><tr><td>Fa</td><td>49</td></tr><tr><td>Po</td><td>1</td></tr></table></div>	Category	Frecuencia	Ex	741	TA	428	Gd	241	Fa	49	Po	1		
Category	Frecuencia																
Ex	741																
TA	428																
Gd	241																
Fa	49																
Po	1																
CentralAir	Y	N	<div><div>CentralAir Barplot</div><table><tr><th>Category</th><th>Frecuencia</th></tr><tr><td>Y</td><td>1365</td></tr><tr><td>N</td><td>95</td></tr></table></div>	Category	Frecuencia	Y	1365	N	95								
Category	Frecuencia																
Y	1365																
N	95																
Electrical	SBrkr	Mix	<div><div>Electrical Barplot</div><table><tr><th>Category</th><th>Frecuencia</th></tr><tr><td>SBrkr</td><td>1334</td></tr><tr><td>FuseA</td><td>94</td></tr><tr><td>FuseF</td><td>27</td></tr><tr><td>FuseP</td><td>3</td></tr><tr><td>Mix</td><td>1</td></tr></table></div>	Category	Frecuencia	SBrkr	1334	FuseA	94	FuseF	27	FuseP	3	Mix	1		
Category	Frecuencia																
SBrkr	1334																
FuseA	94																
FuseF	27																
FuseP	3																
Mix	1																
KitchenQual	TA	Fa	<div><div>KitchenQual Barplot</div><table><tr><th>Category</th><th>Frecuencia</th></tr><tr><td>TA</td><td>735</td></tr><tr><td>Gd</td><td>586</td></tr><tr><td>Ex</td><td>100</td></tr><tr><td>Fa</td><td>39</td></tr></table></div>	Category	Frecuencia	TA	735	Gd	586	Ex	100	Fa	39				
Category	Frecuencia																
TA	735																
Gd	586																
Ex	100																
Fa	39																

Functional	Typ	Sev	<div><p>Functional Barplot</p><table><thead><tr><th>functional</th><th>Frecuencia</th></tr></thead><tbody><tr><td>Typ</td><td>1360</td></tr><tr><td>Min2</td><td>34</td></tr><tr><td>Min1</td><td>31</td></tr><tr><td>Mod</td><td>15</td></tr><tr><td>Maj1</td><td>14</td></tr><tr><td>Maj2</td><td>5</td></tr><tr><td>Sev</td><td>1</td></tr></tbody></table></div>	functional	Frecuencia	Typ	1360	Min2	34	Min1	31	Mod	15	Maj1	14	Maj2	5	Sev	1
functional	Frecuencia																		
Typ	1360																		
Min2	34																		
Min1	31																		
Mod	15																		
Maj1	14																		
Maj2	5																		
Sev	1																		
FireplaceQu	Gd	Po	<div><p>FireplaceQu Barplot</p><table><thead><tr><th>FireplaceQu</th><th>Frecuencia</th></tr></thead><tbody><tr><td>Gd</td><td>380</td></tr><tr><td>TA</td><td>313</td></tr><tr><td>Fa</td><td>33</td></tr><tr><td>Ex</td><td>24</td></tr><tr><td>Po</td><td>20</td></tr></tbody></table></div>	FireplaceQu	Frecuencia	Gd	380	TA	313	Fa	33	Ex	24	Po	20				
FireplaceQu	Frecuencia																		
Gd	380																		
TA	313																		
Fa	33																		
Ex	24																		
Po	20																		
GarageType	Attchd	2Types	<div><p>GarageType Barplot</p><table><thead><tr><th>GarageType</th><th>Frecuencia</th></tr></thead><tbody><tr><td>Attchd</td><td>870</td></tr><tr><td>Detchd</td><td>387</td></tr><tr><td>BuiltIn</td><td>88</td></tr><tr><td>Basement</td><td>19</td></tr><tr><td>CarPort</td><td>9</td></tr><tr><td>2Types</td><td>6</td></tr></tbody></table></div>	GarageType	Frecuencia	Attchd	870	Detchd	387	BuiltIn	88	Basement	19	CarPort	9	2Types	6		
GarageType	Frecuencia																		
Attchd	870																		
Detchd	387																		
BuiltIn	88																		
Basement	19																		
CarPort	9																		
2Types	6																		
GarageFinish	Unf	Fin	<div><p>GarageFinish Barplot</p><table><thead><tr><th>GarageFinish</th><th>Frecuencia</th></tr></thead><tbody><tr><td>Unf</td><td>605</td></tr><tr><td>RFin</td><td>422</td></tr><tr><td>Fin</td><td>352</td></tr></tbody></table></div>	GarageFinish	Frecuencia	Unf	605	RFin	422	Fin	352								
GarageFinish	Frecuencia																		
Unf	605																		
RFin	422																		
Fin	352																		
GarageQual	TA	Ex	<div><p>GarageQual Barplot</p><table><thead><tr><th>GarageQual</th><th>Frecuencia</th></tr></thead><tbody><tr><td>TA</td><td>1311</td></tr><tr><td>Fa</td><td>48</td></tr><tr><td>Gd</td><td>14</td></tr><tr><td>Ex</td><td>3</td></tr><tr><td>Po</td><td>3</td></tr></tbody></table></div>	GarageQual	Frecuencia	TA	1311	Fa	48	Gd	14	Ex	3	Po	3				
GarageQual	Frecuencia																		
TA	1311																		
Fa	48																		
Gd	14																		
Ex	3																		
Po	3																		

GarageCond	TA	Ex	<div><div>GarageCond Barplot</div><table><tr><th>GarageCond</th><th>Frecuencia</th></tr><tr><td>TA</td><td>1326</td></tr><tr><td>Fa</td><td>35</td></tr><tr><td>Gd</td><td>9</td></tr><tr><td>Po</td><td>7</td></tr><tr><td>Ex</td><td>2</td></tr></table></div>	GarageCond	Frecuencia	TA	1326	Fa	35	Gd	9	Po	7	Ex	2
GarageCond	Frecuencia														
TA	1326														
Fa	35														
Gd	9														
Po	7														
Ex	2														
PavedDrive	Y	P	<div><div>PavedDrive Barplot</div><table><tr><th>PavedDrive</th><th>Frecuencia</th></tr><tr><td>Y</td><td>1340</td></tr><tr><td>N</td><td>90</td></tr><tr><td>P</td><td>30</td></tr></table></div>	PavedDrive	Frecuencia	Y	1340	N	90	P	30				
PavedDrive	Frecuencia														
Y	1340														
N	90														
P	30														
PoolQC	Gd	Ex	<div><div>PoolQC Barplot</div><table><tr><th>PoolQC</th><th>Frecuencia</th></tr><tr><td>Gd</td><td>3</td></tr><tr><td>Ex</td><td>2</td></tr><tr><td>Fa</td><td>2</td></tr></table></div>	PoolQC	Frecuencia	Gd	3	Ex	2	Fa	2				
PoolQC	Frecuencia														
Gd	3														
Ex	2														
Fa	2														
Fence	MnPrv	MnWw	<div><div>Fence Barplot</div><table><tr><th>Fence</th><th>Frecuencia</th></tr><tr><td>MnPrv</td><td>157</td></tr><tr><td>GdPrv</td><td>59</td></tr><tr><td>GdWo</td><td>54</td></tr><tr><td>MnWw</td><td>11</td></tr></table></div>	Fence	Frecuencia	MnPrv	157	GdPrv	59	GdWo	54	MnWw	11		
Fence	Frecuencia														
MnPrv	157														
GdPrv	59														
GdWo	54														
MnWw	11														
MiscFeature	Shed	TenC	<div><div>MiscFeature Barplot</div><table><tr><th>MiscFeature</th><th>Frecuencia</th></tr><tr><td>Shed</td><td>49</td></tr><tr><td>Gar2</td><td>2</td></tr><tr><td>Othr</td><td>2</td></tr><tr><td>TenC</td><td>1</td></tr></table></div>	MiscFeature	Frecuencia	Shed	49	Gar2	2	Othr	2	TenC	1		
MiscFeature	Frecuencia														
Shed	49														
Gar2	2														
Othr	2														
TenC	1														

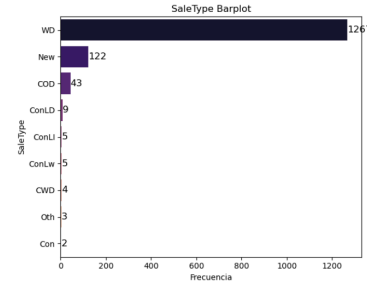
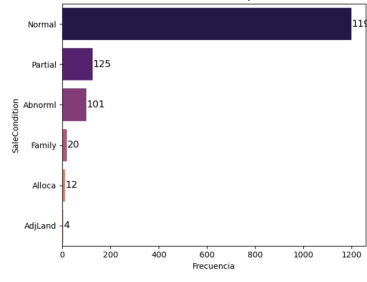
SaleType	WD	Con	 <table><caption>SaleType Barplot</caption><thead><tr><th>SaleType</th><th>Frecuencia</th></tr></thead><tbody><tr><td>WD</td><td>1267</td></tr><tr><td>New</td><td>122</td></tr><tr><td>COD</td><td>43</td></tr><tr><td>ConLD</td><td>9</td></tr><tr><td>ConLI</td><td>5</td></tr><tr><td>ConLW</td><td>5</td></tr><tr><td>CWD</td><td>4</td></tr><tr><td>Oth</td><td>3</td></tr><tr><td>Con</td><td>2</td></tr></tbody></table>	SaleType	Frecuencia	WD	1267	New	122	COD	43	ConLD	9	ConLI	5	ConLW	5	CWD	4	Oth	3	Con	2
SaleType	Frecuencia																						
WD	1267																						
New	122																						
COD	43																						
ConLD	9																						
ConLI	5																						
ConLW	5																						
CWD	4																						
Oth	3																						
Con	2																						
SaleCondition	Normal	AdjLand	 <table><caption>SaleCondition Barplot</caption><thead><tr><th>SaleCondition</th><th>Frecuencia</th></tr></thead><tbody><tr><td>Normal</td><td>1198</td></tr><tr><td>Partial</td><td>125</td></tr><tr><td>Abnorml</td><td>101</td></tr><tr><td>Family</td><td>20</td></tr><tr><td>Alloca</td><td>12</td></tr><tr><td>AdjLand</td><td>4</td></tr></tbody></table>	SaleCondition	Frecuencia	Normal	1198	Partial	125	Abnorml	101	Family	20	Alloca	12	AdjLand	4						
SaleCondition	Frecuencia																						
Normal	1198																						
Partial	125																						
Abnorml	101																						
Family	20																						
Alloca	12																						
AdjLand	4																						

Tabla 3. Categoría más y menos frecuente por variable categórica

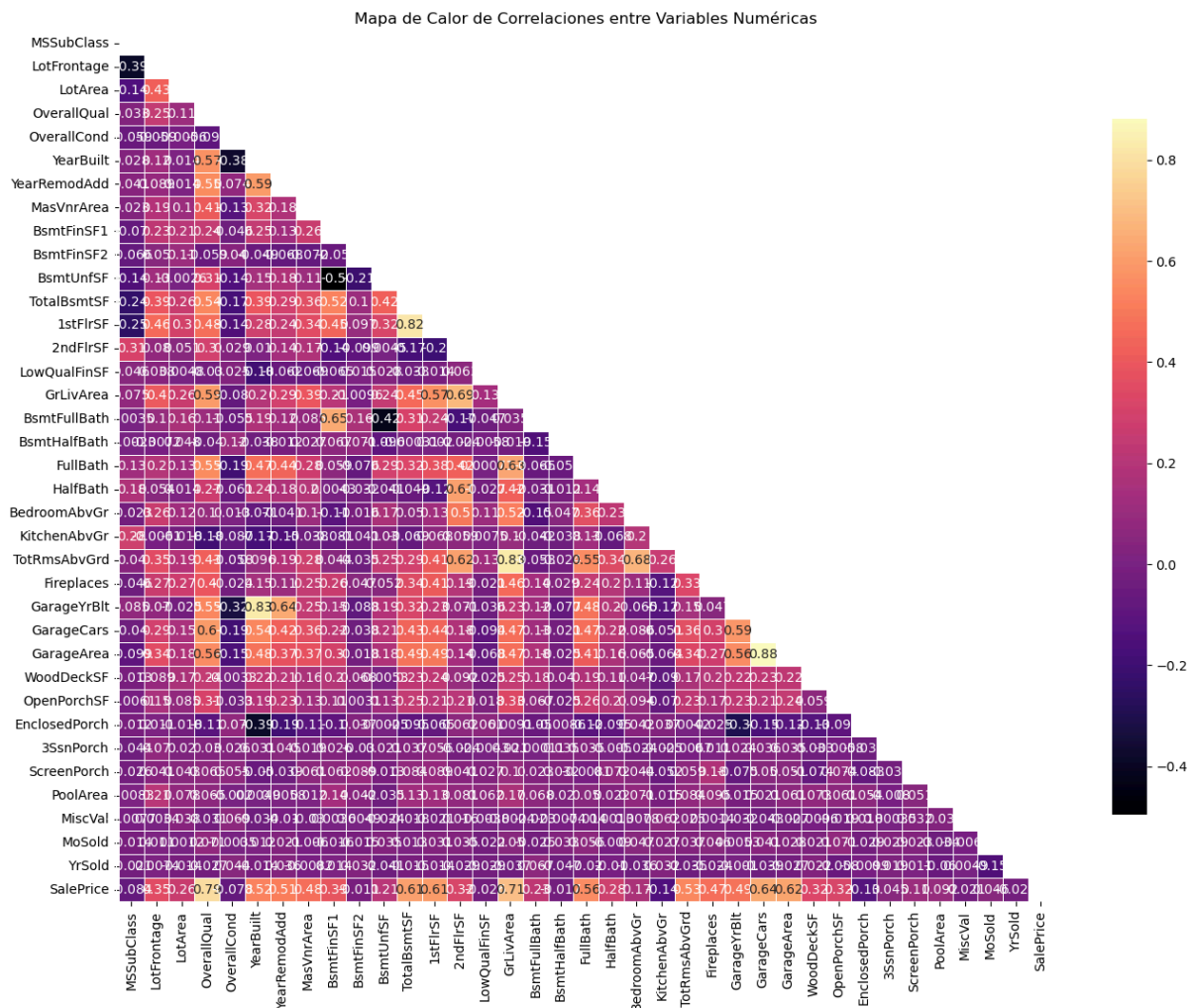
4. Análisis bivariado

Como siguiente paso se realizó el análisis multivariado y bivariado, donde se puede apreciar la relación entre variables y en especial la relación que guardan las variables predictoras con la variable objetivo *SalePrice*.

4.1. Correlaciones

4.1.1. Correlación entre variables predictoras

En primera instancia se calculó la correlación entre las variables, con el propósito de generar un heatmap que haga visualmente más fácil identificar aquellas variables que guardan mayor correlación entre sí, tanto positivas como negativas.

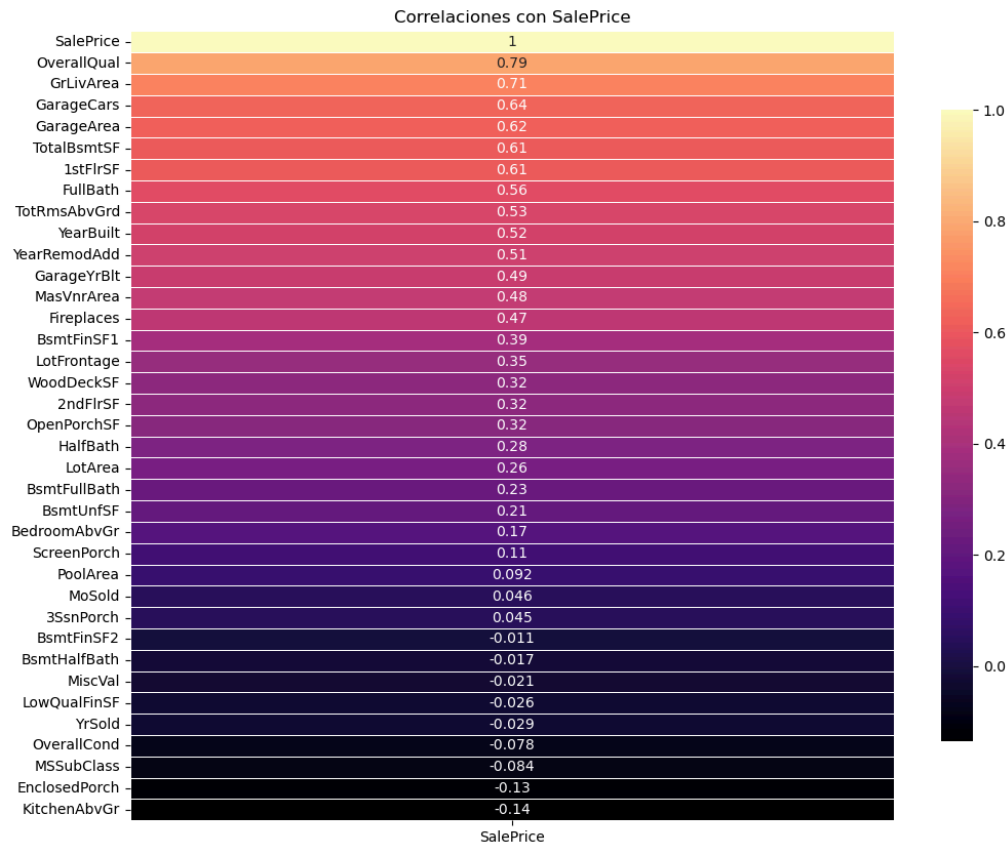


Según lo observado en el mapa de correlaciones, se puede notar que la gran mayoría de las variables no guardan una relación estrecha entre sí, sin embargo es posible notar algunas que sí la tienen (correlación mayor a 0.45 o menor que -0.45). Esto puede ser de ayuda al momento de elegir las variables predictoras para el modelo, debido a que si dos o más variables predictoras tienen una alta correlación entre sí pueden causar un problema conocido como multicolinealidad. El cual es bastante perjudicial en modelos de regresión porque provoca que los coeficientes del modelo se vuelvan inestables y difíciles de interpretar ya que el modelo no puede determinar cuál de las variables correlacionadas está influyendo en la variable objetivo. Además las variables altamente correlacionadas aportan información redundante. Pero la correlación también puede ser de ayuda cuando se quiere hacer ingeniería de características, ya que ésta se trata de combinar variables si éstas hacen sentido.

4.1.2. Correlación con la variable objetivo

Posteriormente se procedió a hacer el análisis de correlación de cada variable predictora con la variable objetivo *SalePrice*, con el objetivo de averiguar cuáles son aquellas variables que guardan una relación lineal fuerte entre ellas y también ayudan a descartar aquellas que no sean relevantes para el modelo.

A continuación se muestra un heatmap de correlación de cada variable con *SalePrice*:



Como se puede observar hay muchas variables que no tienen gran relación con la variable objetivo, por lo que, si no se encuentra una manera de hacer que éstas aporten algo al modelo, podría ser una buena idea quitarlas del dataset.

Enseguida, se presenta una tabla con las variables que guardan una relación estrecha con *SalePrice*, se consideró un umbral de correlación de ± 0.45 :

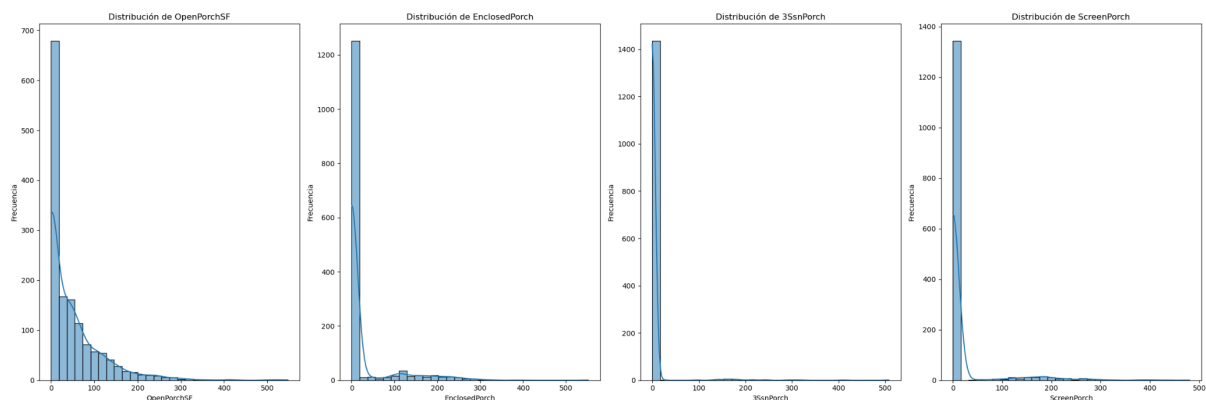
Variable	Correlación con <i>SalePrice</i>
OverallQual	0.790982
GrLivArea	0.708624
GarageCars	0.640409

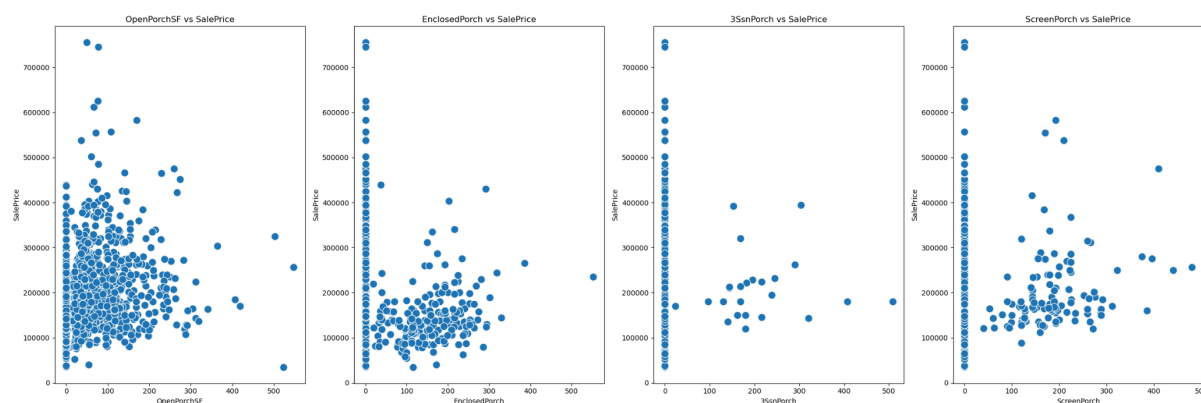
GarageArea	0.623431
TotalBsmtSF	0.613581
1stFlrSF	0.605852
FullBath	0.560664
TotRmsAbvGrd	0.533723
YearBuilt	0.522897
YearRemodAdd	0.507101
GarageYrBlt	0.486362
MasVnrArea	0.477493
Fireplaces	0.466929

Tabla 4. Correlaciones más fuertes con SalePrice

Con la finalidad de no eliminar todas las variables restantes, se decidió estudiar un caso particular de las variables *porch*, debido a que una de ellas guarda una correlación mucho más fuerte con la variable objetivo que el resto.

Las variables que representan el área en pies cuadrados de diferentes tipos de pórticos (*OpenPorchSF*, *EnclosedPorch*, *3SsnPorch* y *ScreenPorch*) presentan una distribución altamente desbalanceada, ya que en la mayoría de los registros el valor corresponde a “0”, indicando ausencia de pórtico como se muestra a continuación:





Salvo por *OpenPorchSF*, cuya correlación con el precio es más significativa (0.31586), las demás variables mencionadas muestran una correlación muy baja con la variable objetivo (*EnclosedPorch*: -0.128578, *3SsnPorch*: 0.044584, *ScreenPorch*: 0.111447), por lo que, se podría considerar viable eliminarlas del análisis, o en su defecto hacer un PCA para combinarlas y obtener una variable con su información, una última opción podría ser hacer uso de la ingeniería de variables y simplemente convertirla en una variable binaria para indicar si la casa tiene un pórtico o no. Si alguna de estas técnicas proporciona una mejoría en los resultados del modelo, podría ser de utilidad probarlo con otras variables que se encuentren en una situación similar.

4.2. Relación SalePrice - Variables

El análisis de los gráficos de dispersión contribuyó a identificar variables que guardan una relación lineal con la variable objetivo *SalePrice*, pero también ayudó a identificar aquellas que guardan más bien una relación no lineal o curva con la variable dependiente.

4.2.1. Relaciones lineales

Varias variables predictoras mostraron tener una relación lineal fuerte y positiva con el precio de venta (*SalePrice*). Esto es esencial para la fase de modelado, ya que estas variables son excelentes candidatas para un modelo de regresión lineal sin necesidad de aplicar demasiadas transformaciones. Dichas variables se muestran en la tabla a continuación:

Variable	Scatterplot
OverallQual	
FullBath	
BsmtFinSF1	
TotalBasmntSF	
1stFlrSF	

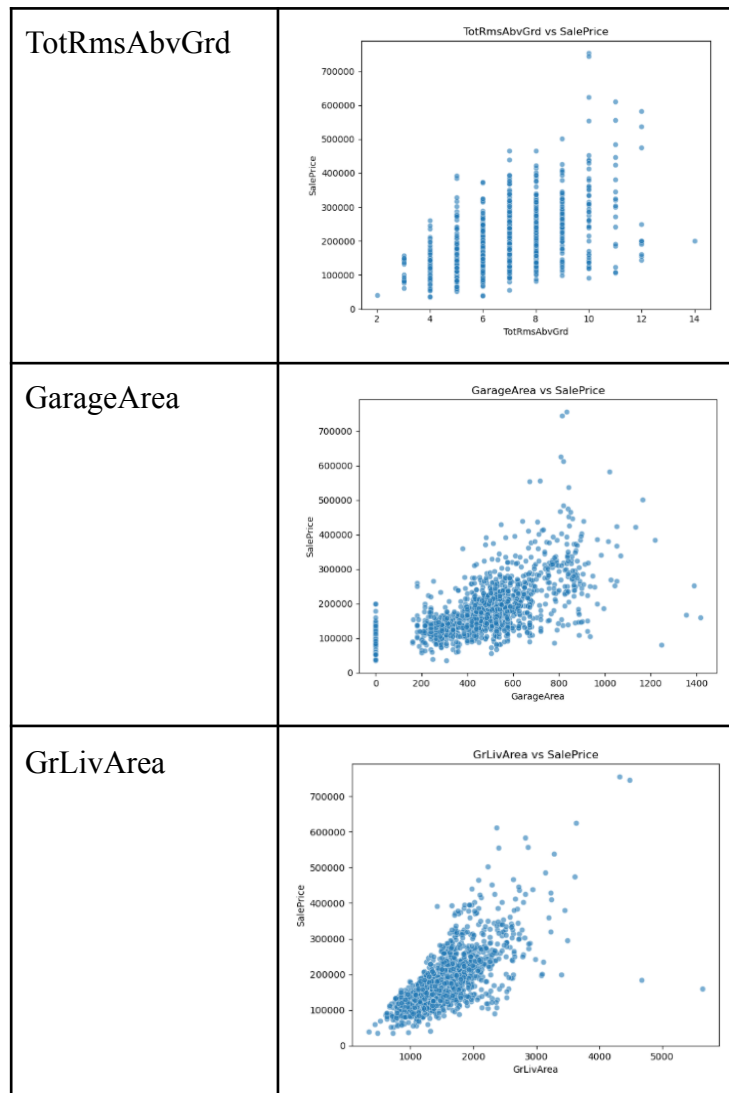


Tabla 5. Variables que presentan una relación lineal con SalePrice

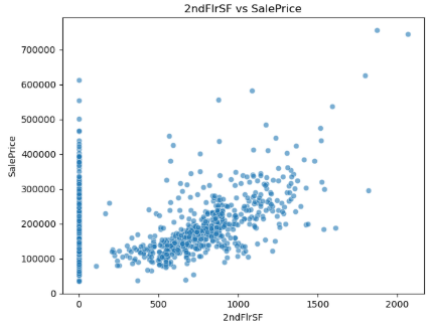
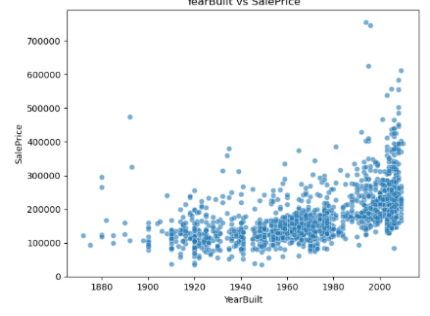
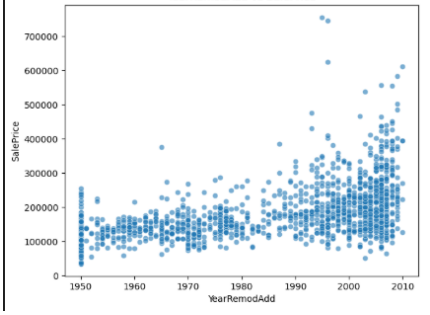
El análisis de las variables de calidad y área (*OverallQual*, *GrLivArea*, *TotalBsmntSF*, *1stFlrSF* y *GarageArea*) ha revelado que estas son los predictores más importantes del precio de venta de una vivienda, mostrando una relación fuertemente lineal y positiva. En conjunto, estos hallazgos confirman que las características relacionadas con el tamaño y la calidad de la vivienda son los principales impulsores de su valor, lo que convierte a estas variables en una buena base para cualquier modelo de regresión.

También se pueden apreciar dos variables que muestran un patrón lineal a pesar de que los puntos se agrupan directamente en columnas discretas (*FullBath* y *TotRmsAbvGrd*). Al observar el precio promedio para cada número de baños o habitaciones, se nota una tendencia clara y lineal positiva, lo que tiene sentido pues cada baño o habitación adicional contribuye a un aumento en el valor de la vivienda, además éstas guardan una correlación fuerte con la variable objetivo.

A diferencia de las variables anteriores, la relación de *BsmtFinSF1* con el precio de venta es una relación positiva, pero con mucha dispersión. Aunque se puede identificar una tendencia general, los puntos están muy dispersos y no se ajustan a una línea recta tan claramente como las otras variables. Esto puede resultar en que la variable tenga un bajo poder predictivo dentro de un modelo, debido a que a la vez su correlación con la variable dependiente tampoco es especialmente fuerte.

4.2.2. Relaciones curvas

Las variables predictoras que no guardan una relación lineal con el precio de venta (*SalePrice*) sino, presentan relaciones no lineales o curvas, representan un hallazgo importante para la selección y el tratamiento de las variables en el modelo predictivo. Estas variables se muestran en la siguiente tabla:

Variable	Scatterplot
2ndFlrSF	
YearBuilt	
YearRemodAdd	

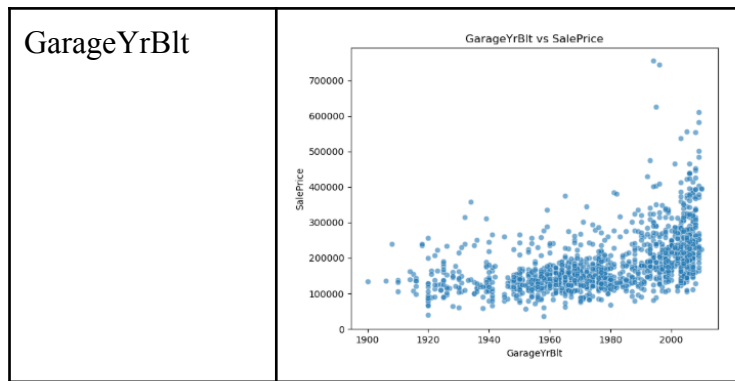


Tabla 6. Variables que presentan una relación curva con SalePrice

Las variables *YearBuilt*, *YearRemodAdd* y *GarageYrBlt* son variables de tiempo que muestran un patrón similar de aceleración en el precio. La relación no es lineal, ya que el aumento en el precio de venta es mucho más pronunciado para las propiedades construidas, remodeladas o con garaje añadido en años recientes, mientras que para las casas más antiguas sus precios son más estables.

La variable *2ndFlrSF* también presenta una relación no lineal, pero con una particularidad. Un gran número de registros tienen un valor de 0, lo que representa las viviendas que no cuentan con un segundo piso. Estos valores son información valiosa y deben ser considerados en el modelo. Para los valores mayores a cero, el patrón es similar a los de las demás variables de la *Tabla 6*, mostrando una relación positiva y curva con el precio, con una dispersión que aumenta a medida que el área del segundo piso se incrementa.

El hecho de que estas relaciones no sean lineales implica que un modelo de regresión lineal simple podría no capturar la complejidad de los datos de manera óptima. Por lo tanto, se recomienda aplicar transformaciones como el logaritmo para estas variables. Estas transformaciones ayudarán a linealizar las relaciones y a reducir la heterocedasticidad (varianza no constante), lo que permitirá que el modelo de regresión se ajuste de manera más precisa y confiable.

5. Valores faltantes y outliers

5.1. Valores faltantes

Durante el análisis de datos se identificó que muchas variables presentan datos faltantes, aquellas que tienen un porcentaje muy alto de valores faltantes se presentan en la tabla a continuación:

Variable	# Valores faltantes	% Valores faltantes
PoolQC	1453	99.520548
MiscFeature	1406	96.301370
Alley	1369	93.767123

Fence	1179	80.753425
MasVnrType	872	59.726027
FireplaceQu	690	47.260274

Tabla 7. Variables con más del 40% de datos faltantes

La escasa presencia de datos limita su utilidad y representa un inconveniente para el tratamiento de la información, así como para el modelo, ya que podría generar complicaciones como sesgos, resultados poco confiables proporcionando predicciones sin sentido, también puede generar un aumento en la varianza del modelo ya que la falta de información disponible puede hacer que el modelo sea más susceptible al ruido y las fluctuaciones, además el hecho de que falten cerca o más de la mitad de los datos hace que sea más probable que las variables se vuelvan irrelevantes, con lo que el modelo perdería su capacidad predictiva. En este caso sería bueno considerar la eliminación definitiva de aquellas variables que guarden poca correlación con la variable objetivo.

Sin embargo, se encontraron otras variables que tienen datos faltantes, pero no sobrepasa el 20% de datos faltantes, por lo que podría considerarse realizar una imputación de valores según convenga (media, mediana, moda, constante, KNN, regresión) y en los casos en los que faltan muy pocos datos se podría considerar la alternativa de eliminarlos, pero dado que la cantidad de datos de entrenamiento es muy limitado, se va a procurar evitar utilizarla a menos de que se considere completamente necesario.

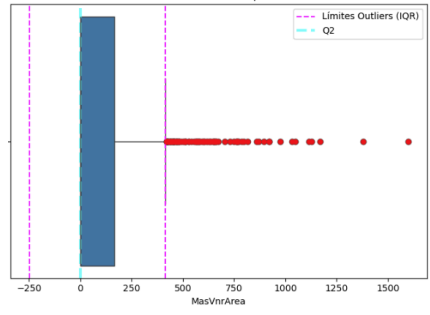
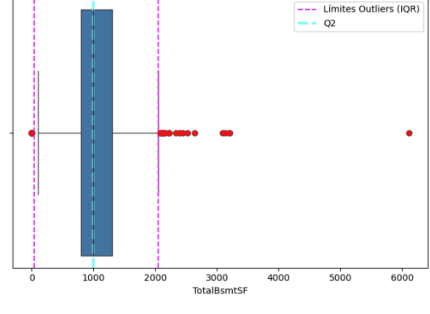
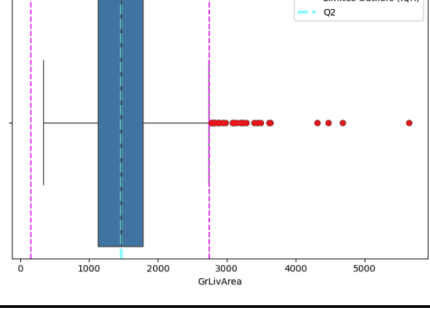
Variable	# Valores faltantes	% Valores faltantes
LotFrontage	259	17.739726
GarageType	81	5.547945
GarageYrBlt	81	5.547945
GarageFinish	81	5.547945
GarageQual	81	5.547945
GarageCond	81	5.547945
BsmtExposure	38	2.602740
BsmtFinType2	38	2.602740
BsmtQual	37	2.534247
BsmtCond	37	2.534247

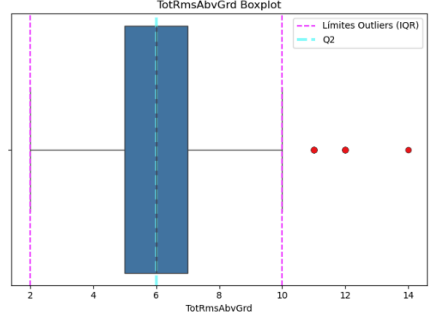
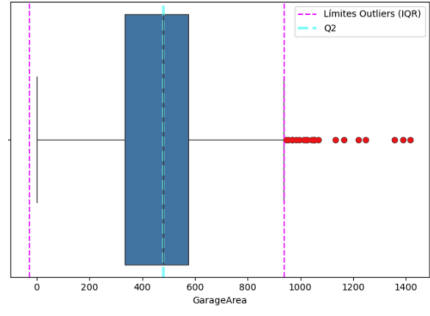
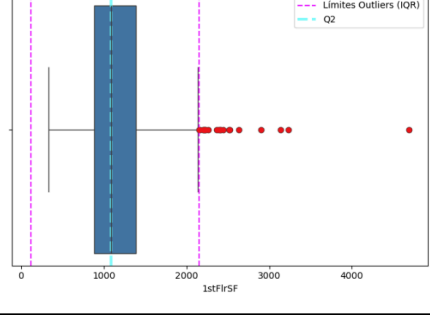
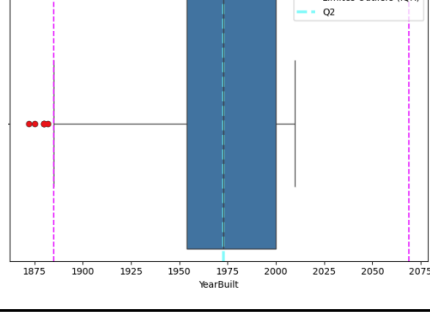
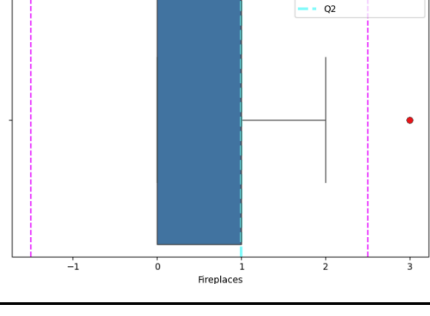
BsmtFinType1	37	2.534247
MasVnrArea	8	0.547945
Electrical	1	0.068493

Tabla 8. Variables con menos del 20% de datos faltantes

5.2. Outliers en las variables principales

Dentro de las variables que poseen una mayor correlación con la variable objetivo *SalePrice*, se identificó la presencia de outliers, a continuación se muestra su conteo, porcentaje y boxplot para cada una de ellas:

Variable	Correlación	# Outliers	% Outliers	Boxplot
MasVnrArea	0.477493	96	6.575342	
TotalBsmtSF	0.613581	61	4.178082	
GrLivArea	0.708624	31	2.123288	

TotRmsAbvGrd	0.533723	30	2.054795	 <p>TotRmsAbvGrd Boxplot</p>
GarageArea	0.623431	21	1.438356	 <p>GarageArea Boxplot</p>
1stFlrSF	0.605852	20	1.369863	 <p>1stFlrSF Boxplot</p>
YearBuilt	0.522897	7	0.479452	 <p>YearBuilt Boxplot</p>
Fireplaces	0.466929	5	0.342466	 <p>Fireplaces Boxplot</p>

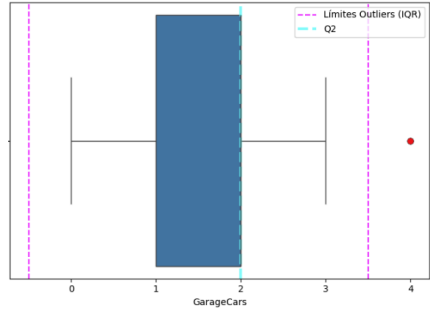
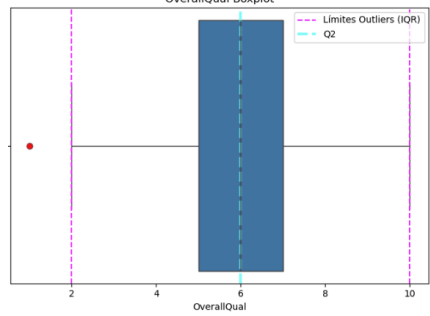
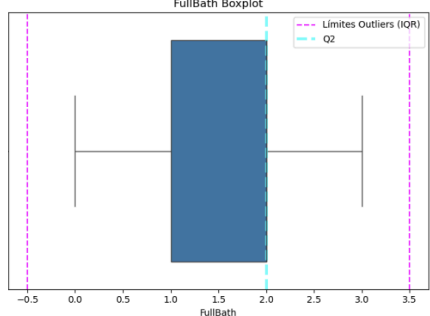
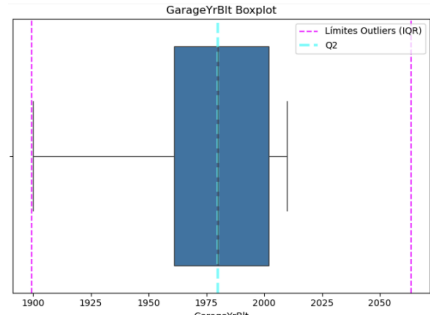
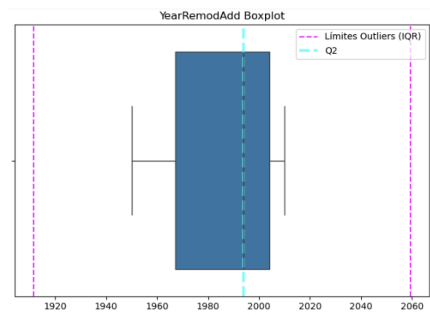
GarageCars	0.640409	5	0.342466	 <p>GarageCars Boxplot</p> <p>This boxplot shows the distribution of GarageCars. The median is approximately 1.5. The interquartile range (IQR) is from 1.0 to 2.0. Whiskers extend from 0.5 to 3.5. A single outlier is present at 4.0.</p>
OverallQual	0.790982	2	0.136986	 <p>OverallQual Boxplot</p> <p>This boxplot shows the distribution of OverallQual. The median is approximately 5.5. The IQR is from 4.5 to 6.5. Whiskers extend from 2.5 to 10.0. A single outlier is present at 2.0.</p>
FullBath	0.560664	0	0.000000	 <p>FullBath Boxplot</p> <p>This boxplot shows the distribution of FullBath. The median is approximately 1.5. The IQR is from 1.0 to 2.0. Whiskers extend from 0.0 to 3.5. No outliers are present.</p>
GarageYrBlt	0.486362	0	0.000000	 <p>GarageYrBlt Boxplot</p> <p>This boxplot shows the distribution of GarageYrBlt. The median is approximately 1975. The IQR is from 1970 to 1980. Whiskers extend from 1960 to 2000. No outliers are present.</p>
YearRemodAdd	0.507101	0	0.000000	 <p>YearRemodAdd Boxplot</p> <p>This boxplot shows the distribution of YearRemodAdd. The median is approximately 1980. The IQR is from 1975 to 1985. Whiskers extend from 1960 to 2000. No outliers are present.</p>

Tabla 9. Outliers de variables más correlacionadas con SalePrice

Como se observa en la *Tabla 9*, las variables mayormente relacionadas con la variable objetivo, presentan una cantidad muy baja de outliers (entre 0% y 7%) lo que se puede solucionar aplicando alguna técnica de imputación de datos para no perder la información que pueden aportar estas variables al modelo.

Al analizar los boxplots se puede observar que la mayoría de las variables tienen su mediana cerca del centro de la distribución según lo indica el IQR, es decir tienen una distribución simétrica. Sin embargo hay algunas que poseen su segundo cuartil o mediana en los límites del IQR, como es el caso de *MasVnrArea*, *Fireplaces*, *GarageCars*, *FullBath* y *YearRemodAdd*. Todas estas variables, excepto *MasVnrArea* tienen su mediana muy cerca o incluso sobre el tercer cuartil de la distribución, lo que indica que es una distribución sesgada hacia la izquierda, es decir, que la mayoría de los datos se concentran en la mitad superior del rango intercuartílico.

Por otra parte *MasVnrArea* tiene su mediana sobre el primer cuartil de la distribución, lo que indica que se trata de una distribución sesgada a la derecha, es decir, que la mayor parte de los datos se concentran en la mitad inferior del rango, y los valores más altos están más dispersos. Hay otras variables que también cuentan con un sesgo menor, pero visible hacia la derecha, como se puede observar en los boxplot de *TotalBsmntSF*, *1stFlrSF* y *YearBuilt*.

6. Recomendaciones basadas en los hallazgos

El análisis exploratorio del dataset, ha proporcionado una comprensión profunda de las características del dataset, revelando la necesidad de implementar estrategias de preprocesamiento de datos para optimizar el modelo de regresión de precios de viviendas. Los hallazgos principales y las recomendaciones clave para la siguiente fase de desarrollo del modelo son los siguientes:

1. Normalidad de variables predictoras y objetivo

Para evitar tener problemas con las variables predictoras, la variable objetivo (*SalePrice*) y el supuesto de normalidad considerado en varios de los modelos lineales tradicionales y con el objetivo de tener un mayor abanico de alternativas de modelos, se considera importante explorar y aplicar transformaciones matemáticas a las variables como el logaritmo para reducir el sesgo de las variables, intentando que se asemejen a una distribución normal y así poder cumplir con el supuesto de normalidad de variable objetivo y normalidad de residuos. De forma que la inferencia estadística sea válida y las predicciones más precisas.

2. Tratamiento de Outliers

En el caso de los outliers encontrados en muchas de las variables predictoras y de la variable objetivo, es importante verificar si estos se tratan de un error de captura de datos o si estos hacen sentido, para posteriormente poder tomar una decisión sobre lo que hay que hacer con cada caso. En este caso se presentan dos posibles soluciones:

- Eliminación de outliers extremos, especialmente en casos donde se trate de registros poco representativos o errores evidentes de captura.
- Utilizar transformaciones matemáticas, como logaritmos o raíces cuadradas, para mitigar su influencia en el modelo, garantizando que el modelo sea robusto y preciso incluso con datos extremos.

3. Simplificación y redundancia (Ingeniería de características)

Se observó una alta correlación y redundancia entre ciertas variables, en esta situación se puede considerar la opción de utilizar técnicas como PCA, conservación de la variable mayormente correlacionada con la variable objetivo o incluso la aplicación de ingeniería de características para evitar que el modelo reciba información que perjudique la interpretabilidad de resultados y/o su desempeño predictivo.

En el caso de las variables relacionadas con los pórticos (*OpenPorchSF*, *EnclosedPorch*, *3SsnPorch* y *ScreenPorch*), se proponen las alternativas con los distintos enfoques mencionados anteriormente:

- Conservar únicamente *OpenPorchSF*, dado que presenta la correlación más significativa con el precio de venta (0.31586), y eliminar las demás variables para reducir ruido.
- Aplicar un Análisis de Componentes Principales (PCA) que permita combinar estas variables en un único componente representativo, preservando la varianza explicada de manera más eficiente y reduciendo la dimensionalidad sin perder información relevante.
- Crear una variable categórica binaria que indique simplemente si la vivienda cuenta o no con algún tipo de pórtico, lo que ayudaría a simplificar el análisis al convertir múltiples columnas numéricas en una característica interpretativa más clara.

Respecto a las variables de baños, se sugiere unificar *BsmtFullBath* con *FullBath* y, de forma análoga, *BsmtHalfBath* con *HalfBath*, reduciendo así las cuatro variables a solamente dos *Fullbath* y *Halfbath*, obteniendo así únicamente un conteo total de

baños completos y de medios baños. Esto permitirá simplificar el modelo y evitar redundancias, conservando la información de manera más práctica y manejable.

4. Manejo de valores faltantes

El análisis de valores faltantes reveló que ciertas variables tienen una cantidad de datos ausentes tan alta que su imputación no es viable y podrían introducir ruido al modelo, puesto que no aportarían información valiosa al mismo. Por ello, se recomienda la eliminación de las variables *PoolQC*, *MiscFeature*, *Alley*, *Fence* y *FireplaceQu* antes de la fase de modelado, pues tampoco cuentan con una correlación significativa con la variable objetivo *SalePrice*.

Conclusiones

Este análisis exploratorio ha sentado una base sólida para comenzar el desarrollo de un modelo de regresión robusto, pues se ha demostrado que el éxito de un modelo predictivo no depende únicamente del algoritmo elegido, sino de la calidad de los datos con los que se entrena. Las recomendaciones propuestas se centran en la limpieza, simplificación y transformación de los datos para asegurar que el modelo final no solo sea preciso, sino también interpretable y fiable. El siguiente paso en este proyecto será la implementación de estas estrategias de preprocesamiento y la experimentación con diversos modelos de regresión para encontrar la solución más óptima posible.

Anexo

A continuación se presenta el enlace para acceder al notebook de Python utilizado para la realización del presente reporte de análisis exploratorio de datos:
[*AnalisisExploratorioRetoHousePrice.ipynb*](#)