Proyecto 2

Rodrigo Mansilla, Javier Chen

2025-02-25

Introducción

Metodología

Exploración inicial de Datos

Dimensiones del Dataset

Table 1: Dimensiones del Dataset

Métrica	Valor
Número de filas	1460
Número de columnas	81

Primeras filas

Table 2: Primeras 6 filas (5 columnas)

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities
1	60	RL	65	8450	Pave	NA	Reg	Lvl	AllPub
2	20	RL	80	9600	Pave	NA	Reg	Lvl	AllPub
3	60	RL	68	11250	Pave	NA	IR1	Lvl	AllPub
4	70	RL	60	9550	Pave	NA	IR1	Lvl	AllPub
5	60	RL	84	14260	Pave	NA	IR1	Lvl	AllPub
6	50	RL	85	14115	Pave	NA	IR1	Lvl	AllPub

Ultimas filas

Table 3: Últimas 6 filas (5 columnas)

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities
1455	1455	20	FV	62	7500	Pave	Pave	Reg	Lvl	AllPub
1456	1456	60	RL	62	7917	Pave	NA	Reg	Lvl	AllPub
1457	1457	20	RL	85	13175	Pave	NA	Reg	Lvl	AllPub

1458	1458	70	RL	66	9042	Pave	NA	Reg	Lvl	AllPub
1459	1459	20	RL	68	9717	Pave	NA	Reg	Lvl	AllPub
1460	1460	20	RL	75	9937	Pave	NA	Reg	Lvl	AllPub

Observamos que el dataset contiene una complejidad adecuada y es necesaria la limpieza y transformación de datos para poder detectar relaciones, outliers y patrones en los datos.

Análisis descriptivo y Exploración de Variables

Estadísticas Descriptivas de variables numéricas

Table 4: Resumen Estadístico de Variables Numéricas

	count	mean	std	min	Q1.25%	Median.50%	Q3.75%	max	mediana
Id	1460	730.50	421.61	1	365.75	730.5	1095.25	1460	730.5
MSSubClass	1460	56.90	42.30	20	20.00	50.0	70.00	190	50.0
LotFrontage	1201	70.05	24.28	21	59.00	69.0	80.00	313	69.0
LotArea	1460	10516.83	9981.26	1300	7553.50	9478.5	11601.50	215245	9478.5
OverallQual	1460	6.10	1.38	1	5.00	6.0	7.00	10	6.0
OverallCond	1460	5.58	1.11	1	5.00	5.0	6.00	9	5.0
YearBuilt	1460	1971.27	30.20	1872	1954.00	1973.0	2000.00	2010	1973.0
YearRemodAdd	1460	1984.87	20.65	1950	1967.00	1994.0	2004.00	2010	1994.0
MasVnrArea	1452	103.69	181.07	0	0.00	0.0	166.00	1600	0.0
BsmtFinSF1	1460	443.64	456.10	0	0.00	383.5	712.25	5644	383.5
BsmtFinSF2	1460	46.55	161.32	0	0.00	0.0	0.00	1474	0.0
BsmtUnfSF	1460	567.24	441.87	0	223.00	477.5	808.00	2336	477.5
TotalBsmtSF	1460	1057.43	438.71	0	795.75	991.5	1298.25	6110	991.5
X1stFlrSF	1460	1162.63	386.59	334	882.00	1087.0	1391.25	4692	1087.0
X2ndFlrSF	1460	346.99	436.53	0	0.00	0.0	728.00	2065	0.0
LowQualFinSF	1460	5.84	48.62	0	0.00	0.0	0.00	572	0.0
GrLivArea	1460	1515.46	525.48	334	1129.50	1464.0	1776.75	5642	1464.0
BsmtFullBath	1460	0.43	0.52	0	0.00	0.0	1.00	3	0.0
BsmtHalfBath	1460	0.06	0.24	0	0.00	0.0	0.00	2	0.0
FullBath	1460	1.57	0.55	0	1.00	2.0	2.00	3	2.0
HalfBath	1460	0.38	0.50	0	0.00	0.0	1.00	2	0.0
BedroomAbvGr	1460	2.87	0.82	0	2.00	3.0	3.00	8	3.0
KitchenAbvGr	1460	1.05	0.22	0	1.00	1.0	1.00	3	1.0
TotRmsAbvGrd	1460	6.52	1.63	2	5.00	6.0	7.00	14	6.0
Fireplaces	1460	0.61	0.64	0	0.00	1.0	1.00	3	1.0
GarageYrBlt	1379	1978.51	24.69	1900	1961.00	1980.0	2002.00	2010	1980.0
GarageCars	1460	1.77	0.75	0	1.00	2.0	2.00	4	2.0
GarageArea	1460	472.98	213.80	0	334.50	480.0	576.00	1418	480.0
WoodDeckSF	1460	94.24	125.34	0	0.00	0.0	168.00	857	0.0
OpenPorchSF	1460	46.66	66.26	0	0.00	25.0	68.00	547	25.0
EnclosedPorch	1460	21.95	61.12	0	0.00	0.0	0.00	552	0.0
X3SsnPorch	1460	3.41	29.32	0	0.00	0.0	0.00	508	0.0
ScreenPorch	1460	15.06	55.76	0	0.00	0.0	0.00	480	0.0
PoolArea	1460	2.76	40.18	0	0.00	0.0	0.00	738	0.0
MiscVal	1460	43.49	496.12	0	0.00	0.0	0.00	15500	0.0
MoSold	1460	6.32	2.70	1	5.00	6.0	8.00	12	6.0
YrSold	1460	2007.82	1.33	2006	2007.00	2008.0	2009.00	2010	2008.0

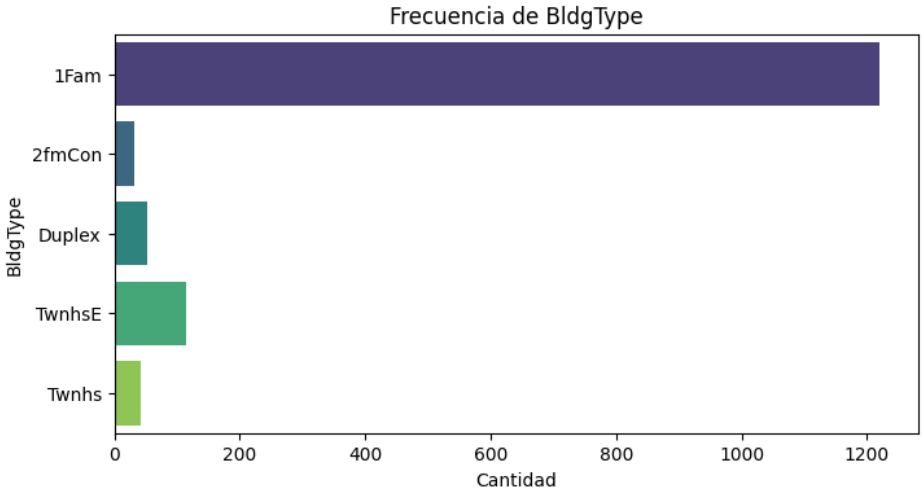
	count	mean	std	min	Q1.25%	Median.50%	Q3.75%	max	mediana
SalePrice	1460	180921.20	79442.50	34900	129975.00	163000.0	214000.00	755000	163000.0

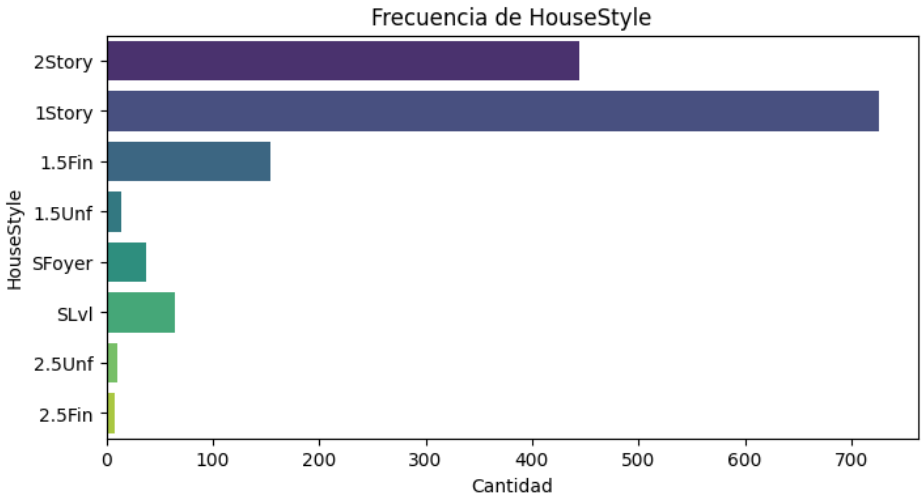
Estas estadísticas descriptivas nos permiten tener una idea general de la distribución de las variables numéricas en el dataset. A oartir de estos datos podemos explorar variables con gran variabilidad y outliers como:

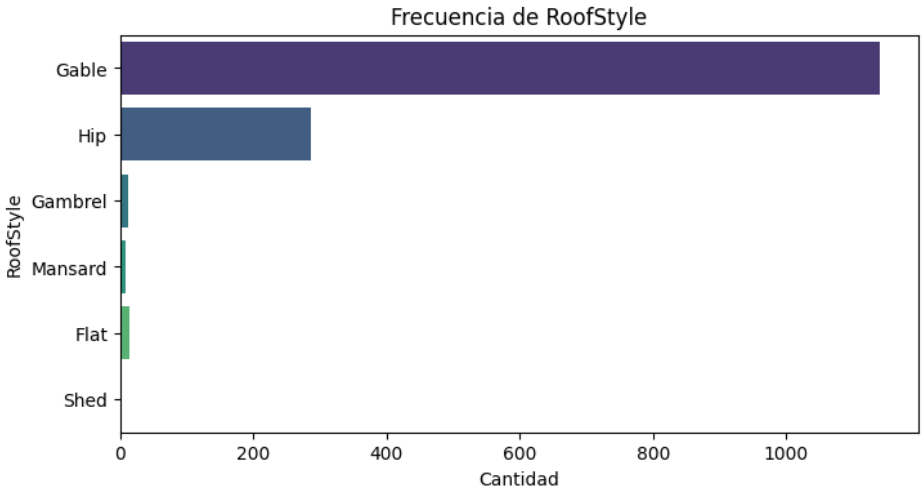
- **SalePrice:** Es la variable objetivo; analizar su distribución es esencial para detectar sesgos o valores atípicos que puedan afectar modelos predictivos.
- **GrLivArea, LotArea, X1stFlrSF y TotalBsmtSF**: Estas variables relacionadas con áreas muestran amplios rangos y desviaciones estándar elevadas, lo que indica una variabilidad considerable. Evaluar su distribución ayudará a entender cómo influyen en el precio.
- **OverallQual y OverallCond**: Son escalas de calidad y condición que, a pesar de ser discretas, pueden tener un impacto directo en el precio.
- **YearBuilt y YearRemodAdd:** La antigüedad y el año de remodelación pueden explicar cambios en la valoración de las viviendas. Su distribución puede revelar tendencias históricas y patrones de renovación.
- **LotFrontage y MasVnrArea:** Aunque LotFrontage presenta datos faltantes, es relevante para entender la exposición del lote. MasVnrArea muestra muchos ceros y algunos valores altos, lo que sugiere la presencia de outliers que vale la pena investigar.
- **GarageArea y GarageCars:** Estas variables relacionadas con el garaje también presentan variabilidad notable y pueden influir en el precio, es útil evaluar si existen distribuciones sesgadas o valores extremos.

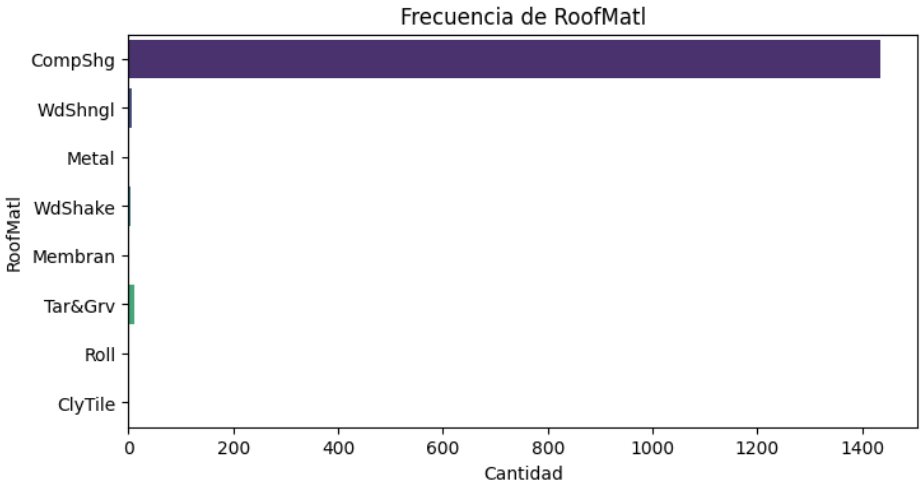
Exploración de variables categóricas

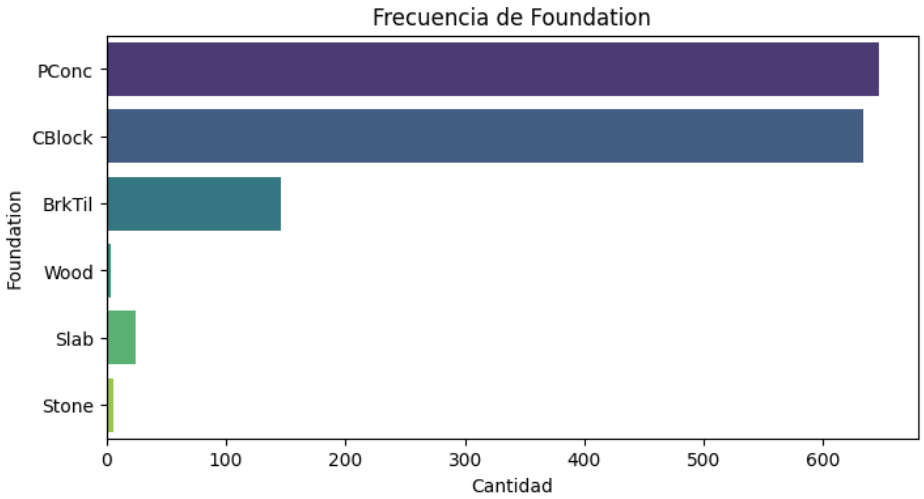
Variables relacionadas con la construcción y estructura




Frecuencia de HouseStyle
2Story
1Story
1.5Fin
1.5Unf
SFoyer
SLvl
2.5Unf
2.5Fin
HouseStyle
0
100
200
300
400
500
600
700
Cantidad

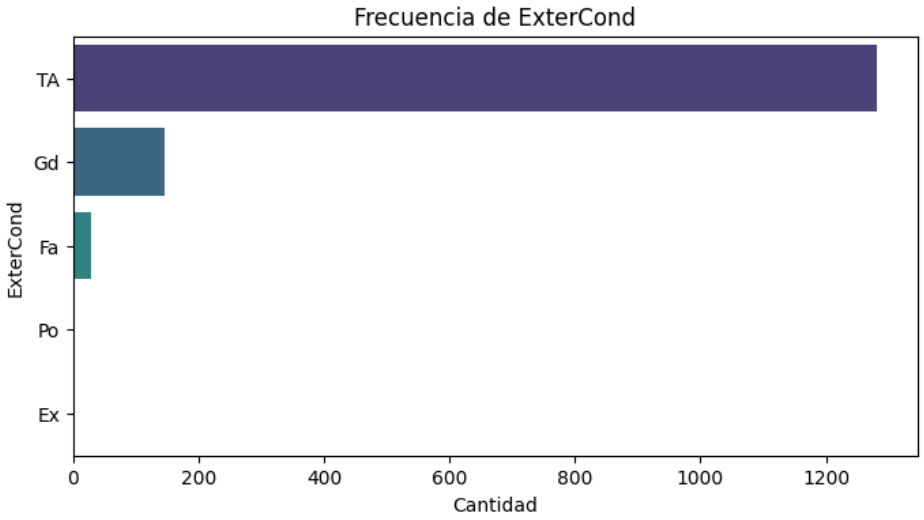

Frecuencia de RoofStyle
Gable
Hip
Gambrel
Mansard
Flat
Shed
RoofStyle
0
200
400
600
800
1000
Cantidad

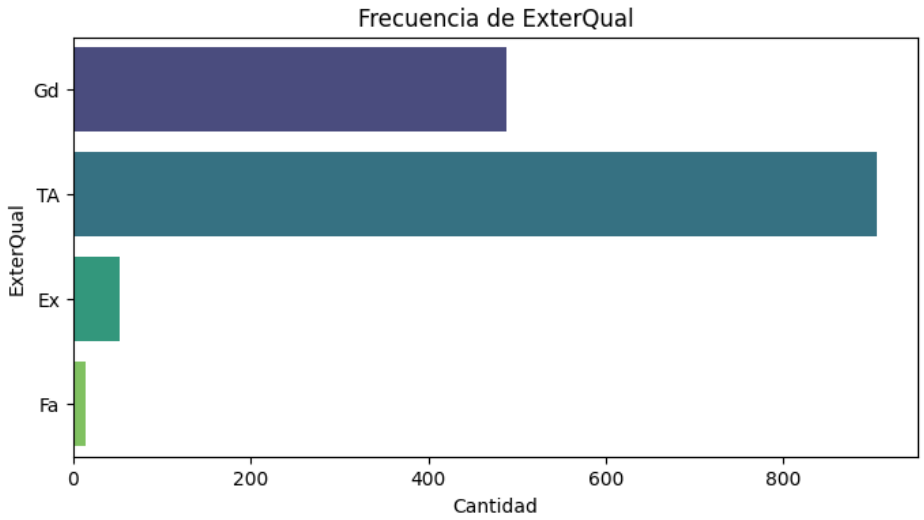

Frecuencia de RoofMatl
CompShg
WdShngl
Metal
WdShake
Membran
Tar&Grv
Roll
ClyTile
RoofMatl
0
200
400
600
800
1000
1200
1400
Cantidad

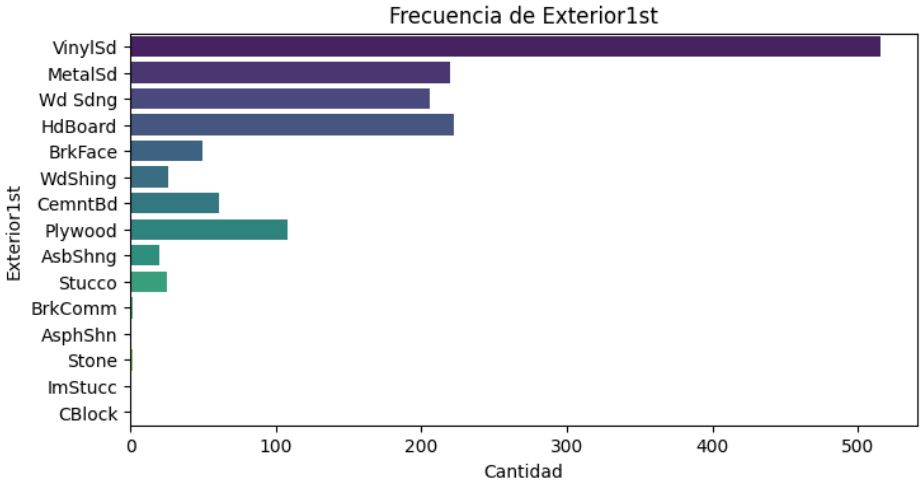


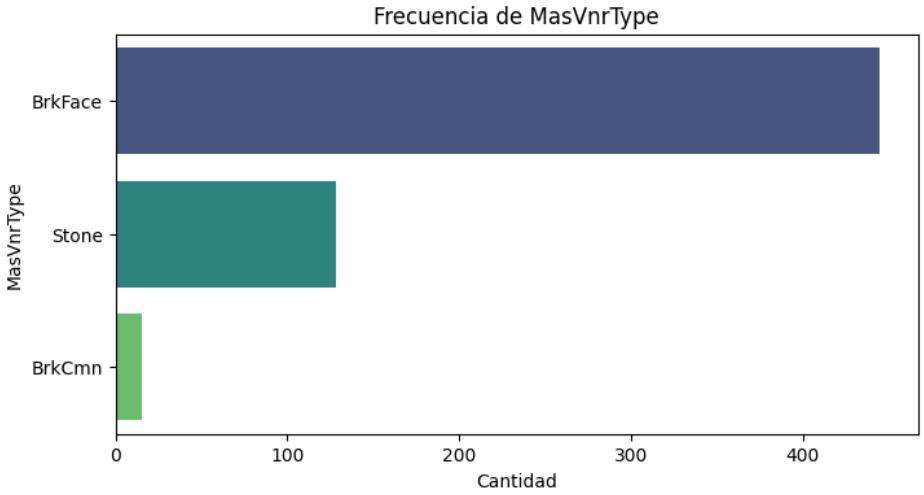
Este grupo de variables muestra una mayoría de casas unifamiliares, predominancia en casasde 2 y 1 piso, techos de tipo Gable y materiales de techos CompShg. La mayoría de las casas tienen cimientos de concreto y madera. Estos patrones pueden ser útiles para identificar características comunes en la construcción de las propiedades.

Variables relacionadas con el exterior y materiales





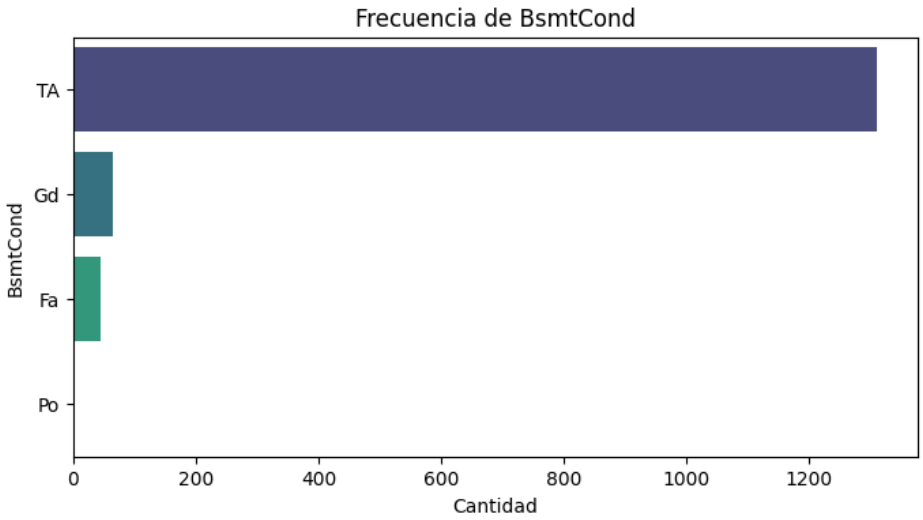


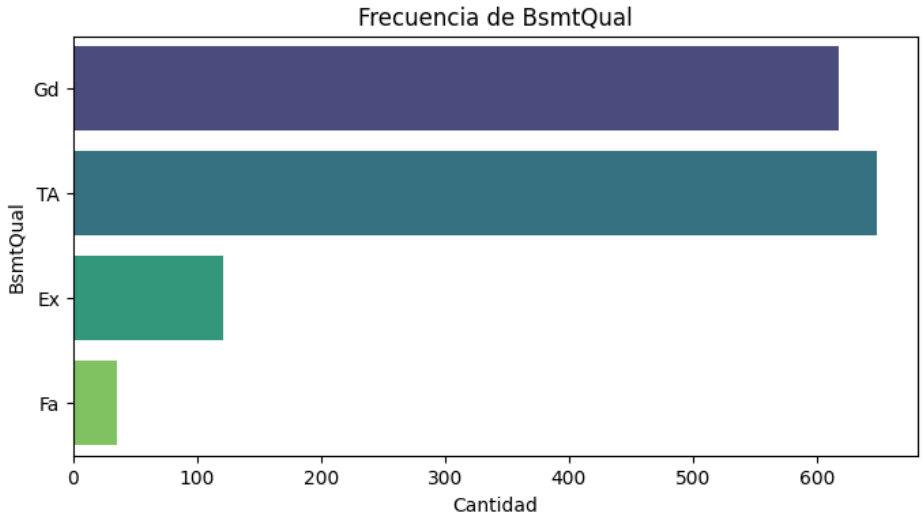


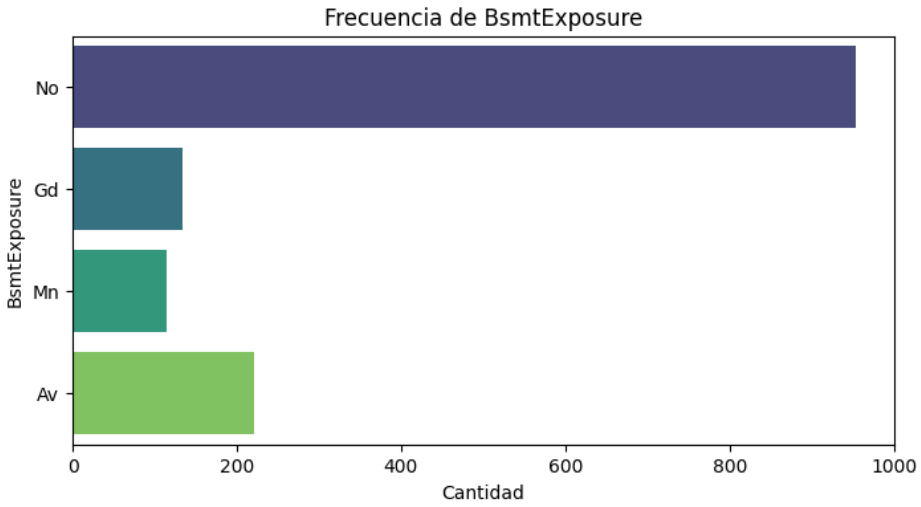
La mayoría de las casas presentan una condición y calidad exterior promedio, con pocas en estado excelente o deficiente. En las cubiertas exteriores, domina “VinylSd” tanto en la primera como en la segunda capa,

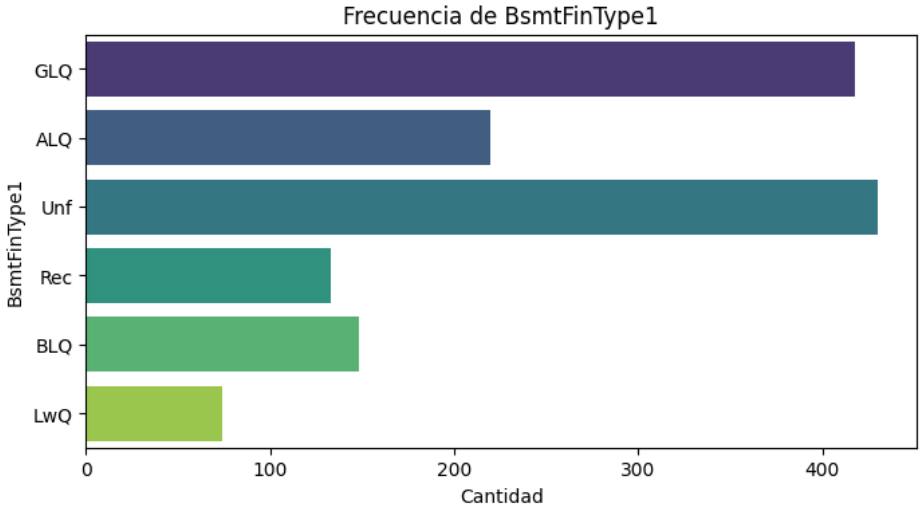
seguido a cierta distancia por “MetalSd”, “Wd Sdng” y “HdBoard”. La mampostería vista (MasVnrType) más frecuente es “BrkFace”, con “Stone” como segunda opción. Esto sugiere un mercado residencial donde predomina un nivel de acabado estándar y revestimientos vinílicos o de metal, con menos variedad en acabados de alta o baja calidad.

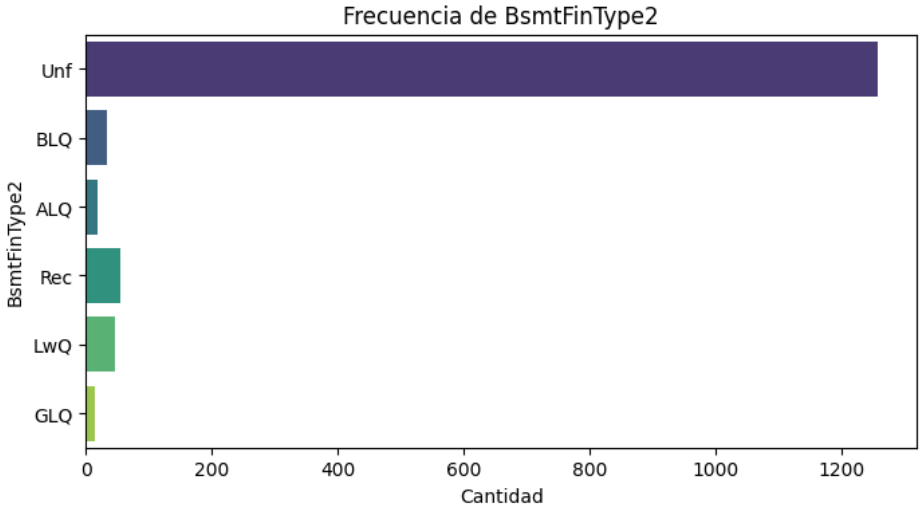
Variables relacionadas con el sótano







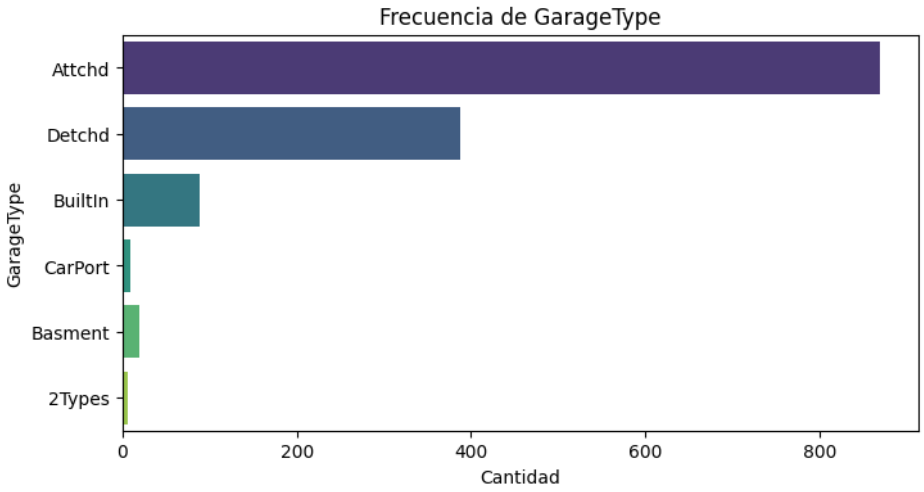


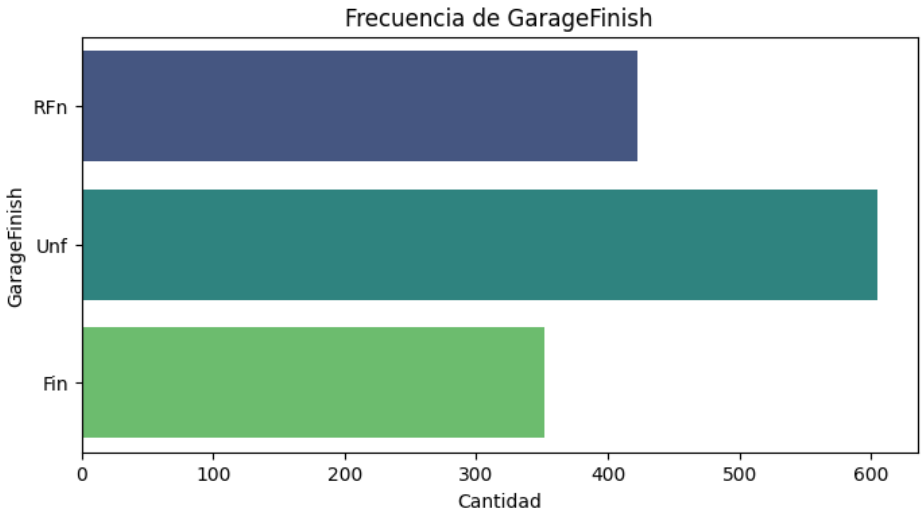


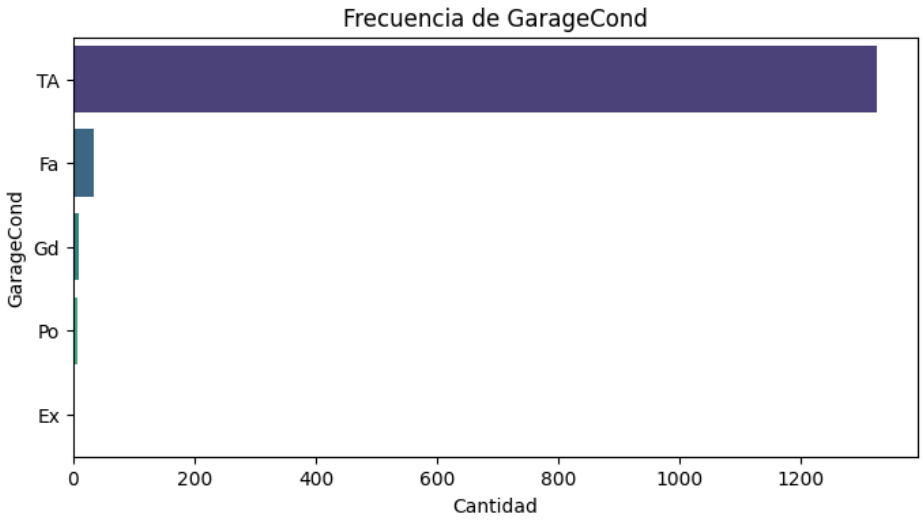
La mayoría de los sótanos están en condición “TA” y calidad “TA” o “Gd”, con pocos casos “Ex” o “Fa”. La exposición del sótano suele ser “No” (sin exposición), aunque también hay un grupo con “Gd”, “Mn” y

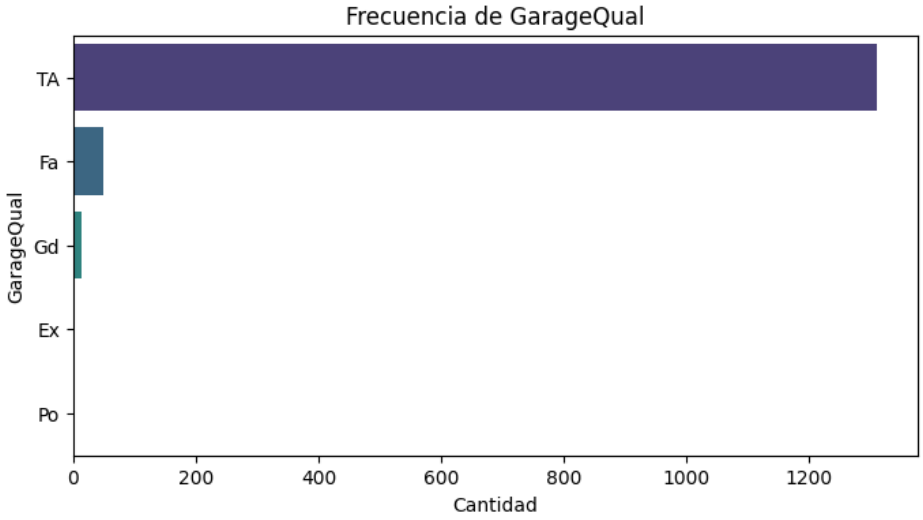
“Av”. Para la terminación del sótano, “GLQ” y “Unf” predominan en BsmtFinType1, mientras que “Unf” es casi absoluto en BsmtFinType2, indicando que muchos sótanos adicionales están sin terminar o tienen acabados básicos.

Variables relacionadas con el garaje



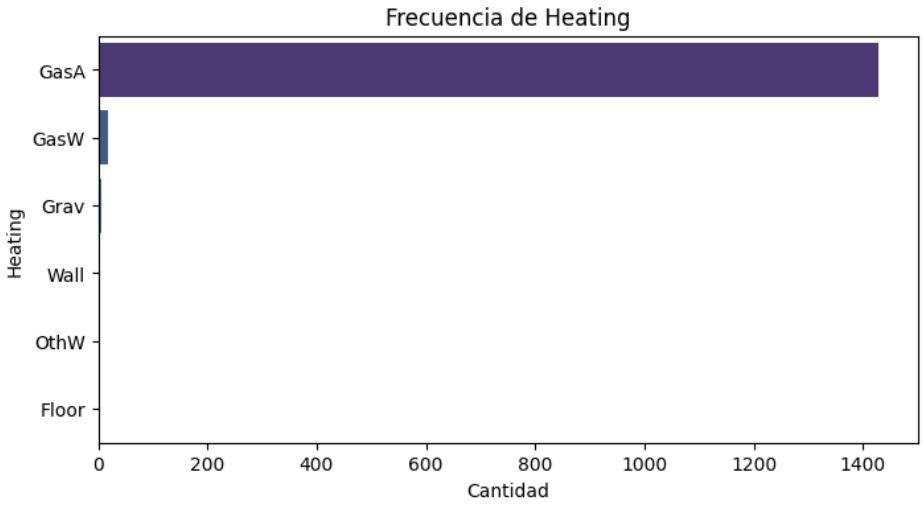


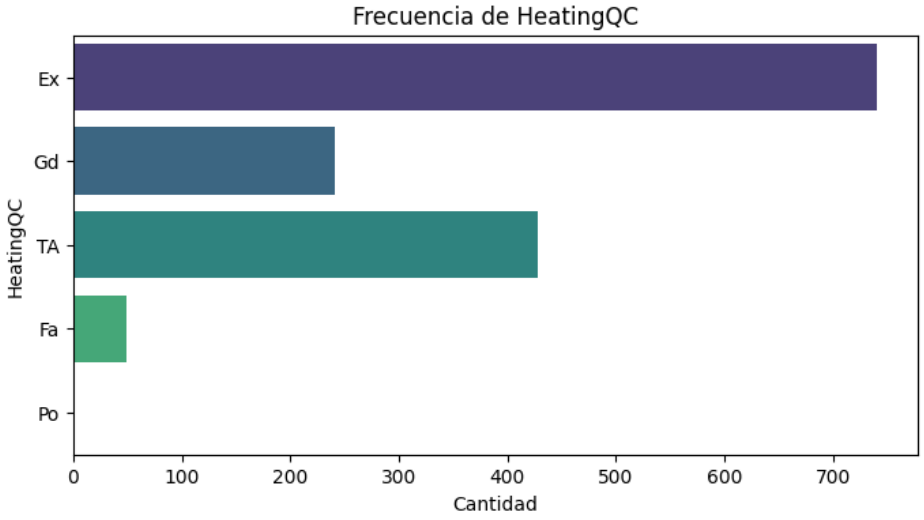


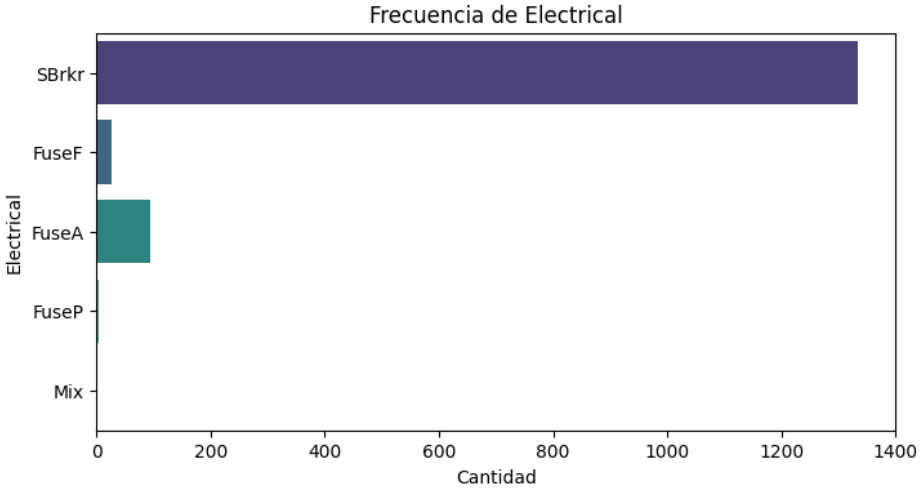


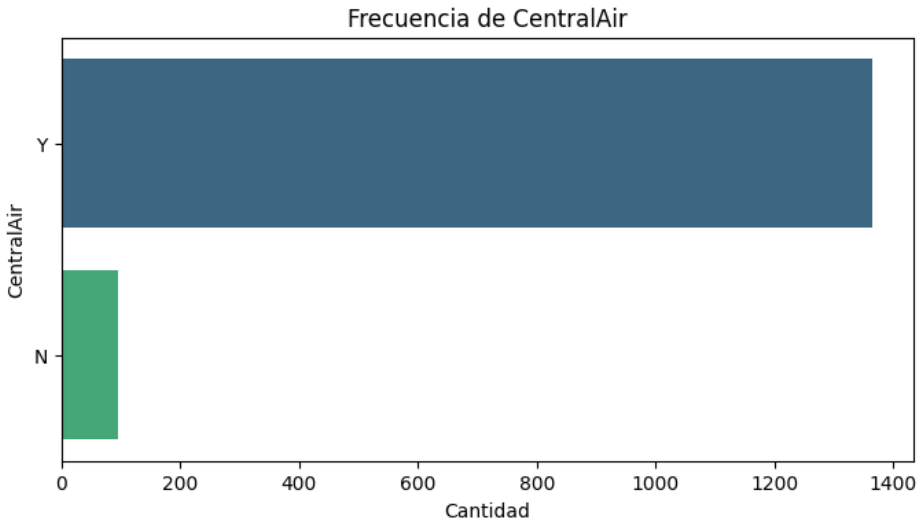
La mayoría de las casas tienen garajes adjuntos, seguidos por garajes separados y sin garaje. En cuanto al acabado del garaje, predominan los garajes sin acabado o con acabado de calidad estándar. La calidad y condición del garaje tienden a ser promedio, con pocos casos en los extremos. Estos patrones sugieren que la mayoría de las propiedades tienen garajes estándar o básicos, lo que puede influir en el precio de venta.

Variables relacionadas con calefacción y electricidad


Frecuencia de Heating
GasA
GasW
Grav
Wall
OthW
Floor
Heating
0
200
400
600
800
1000
1200
1400
Cantidad

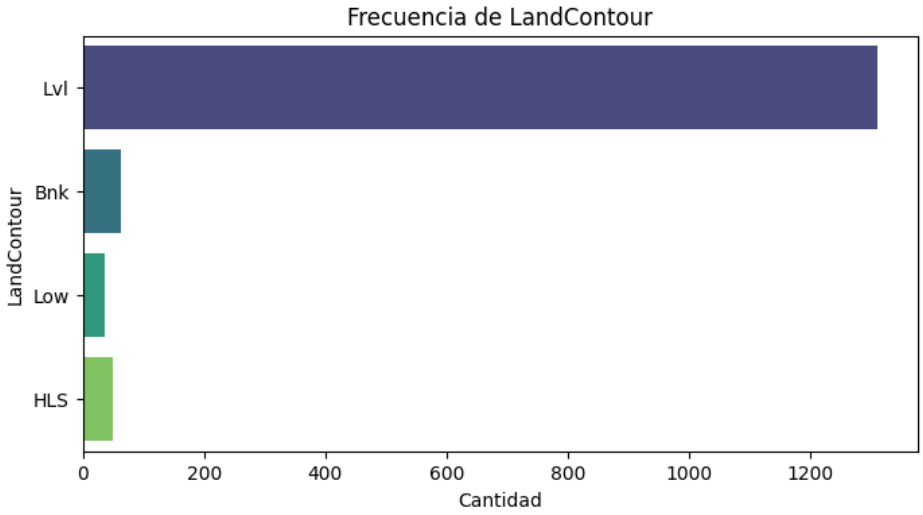

Frecuencia de HeatingQC
Ex
Gd
TA
Fa
Po
HeatingQC
0
100
200
300
400
500
600
700
Cantidad

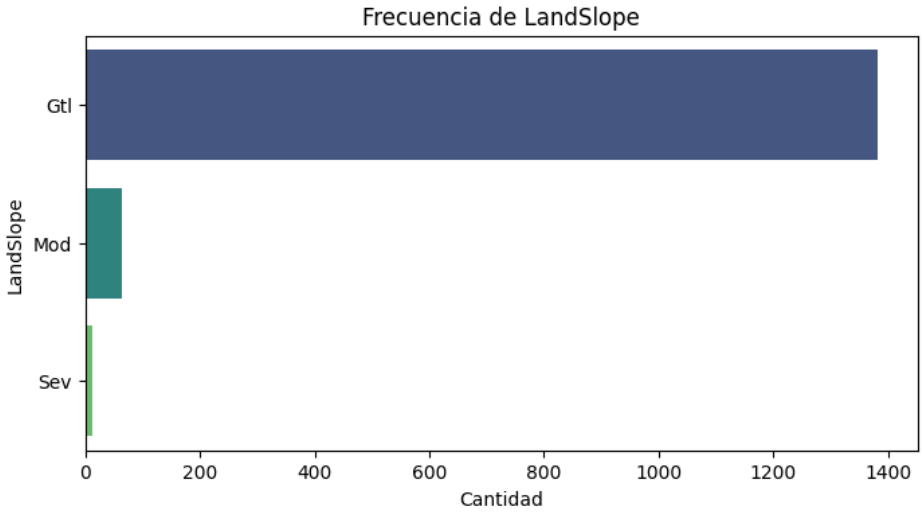

Frecuencia de Electrical
SBrkr
FuseF
FuseA
FuseP
Mix
Electrical
0
200
400
600
800
1000
1200
1400
Cantidad

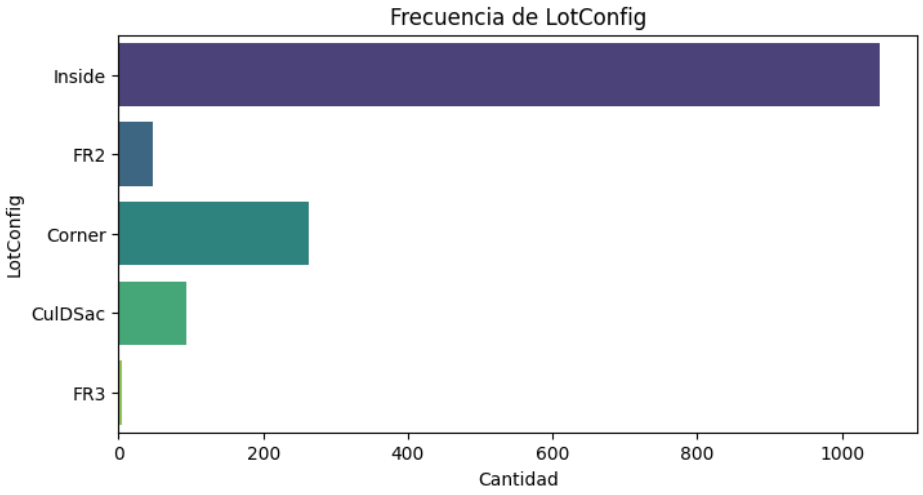


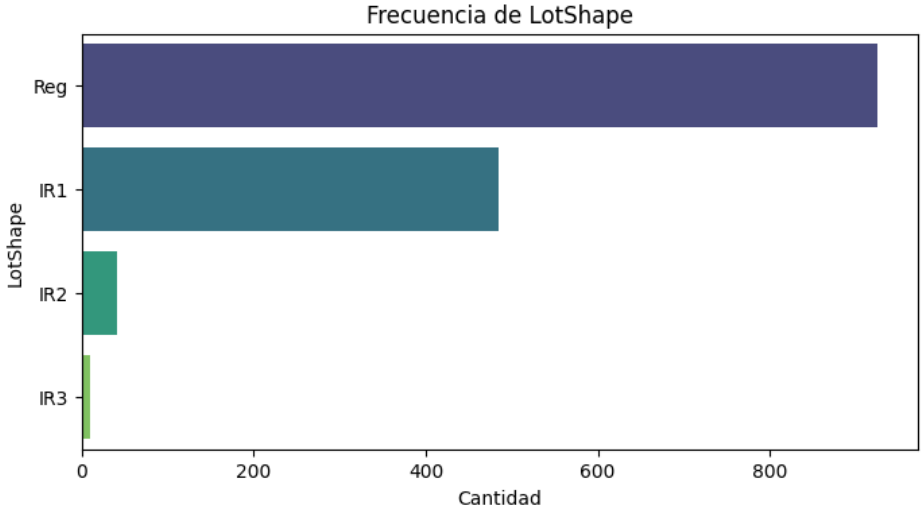
La mayoría de las casas tienen calefacción estándar (GasA) y calidad promedio (TA). La electricidad es principalmente SBrkr, con algunos casos de FuseA y FuseF. La mayoría de las casas tienen aire acondicionado central, lo que sugiere un nivel de comodidad y eficiencia energética estándar en la mayoría de las propiedades.

Variables relacionadas con la ubicación del terreno





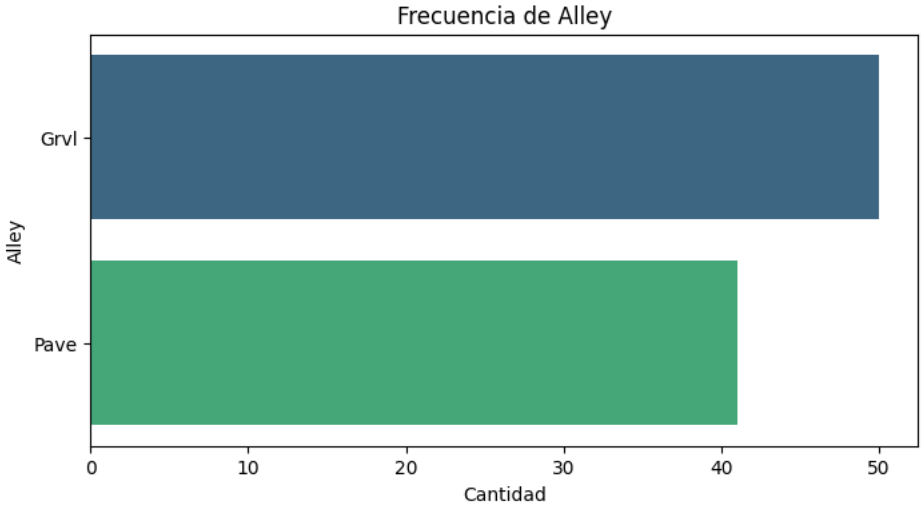


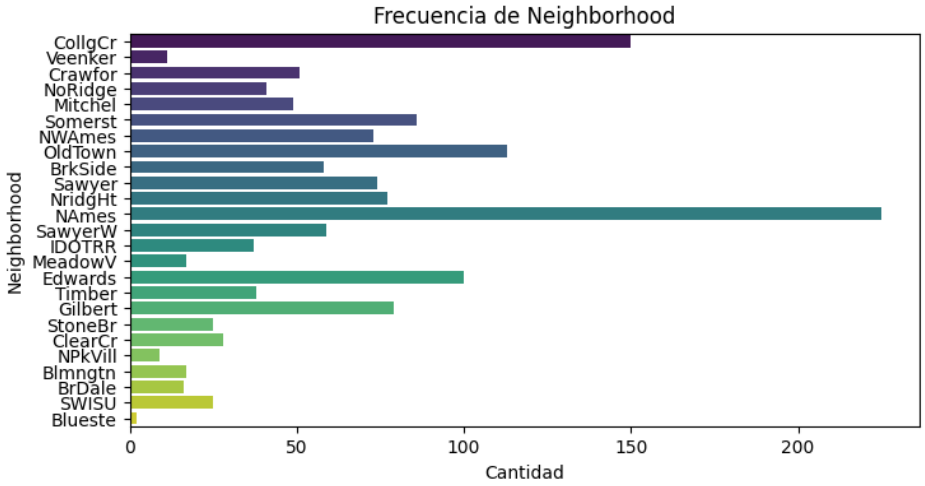


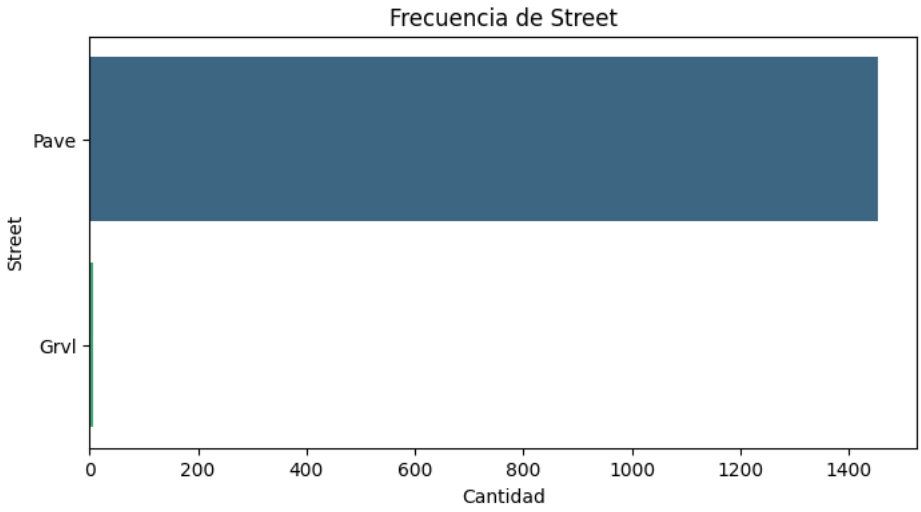
La mayoría de las propiedades tienen terrenos planos o ligeramente inclinados, con configuraciones de lote

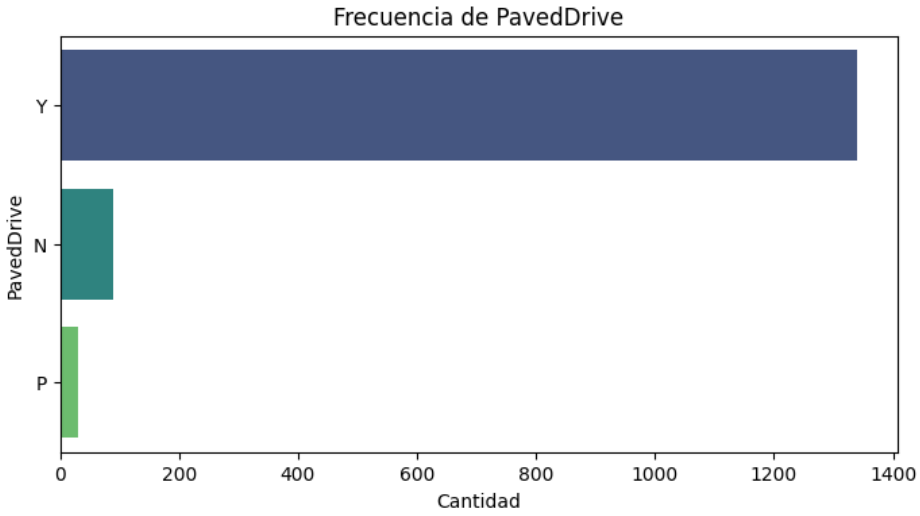
internas y formas regulares. Estos patrones sugieren que la mayoría de las propiedades están en áreas urbanas o suburbanas, con lotes estándar y fácil acceso a servicios y vías de comunicación.

Variables relacionadas con vecindario y accesibilidad



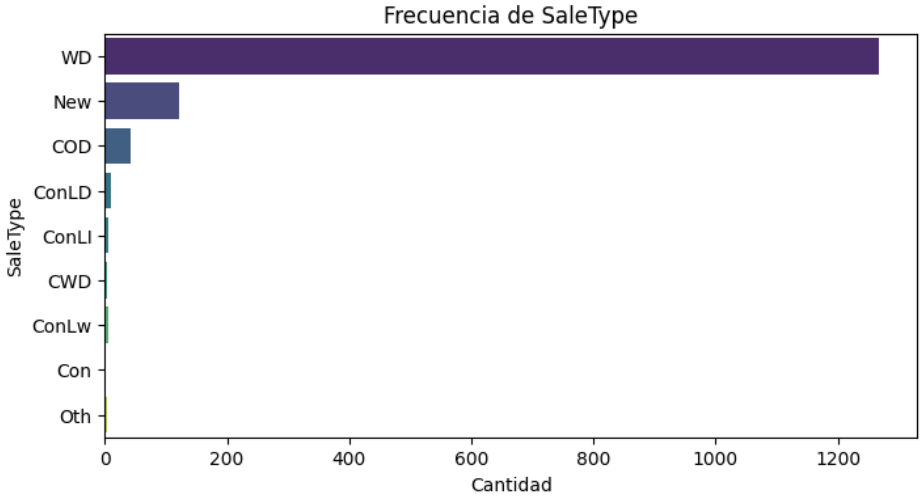


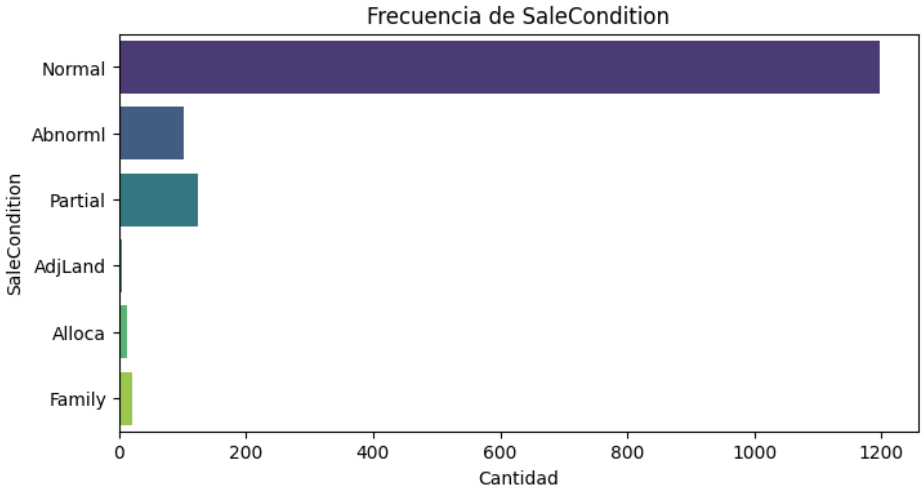


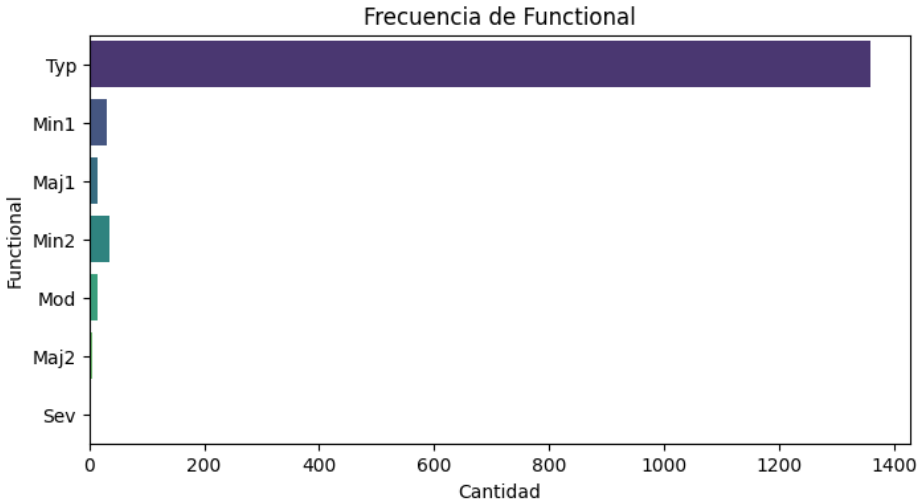


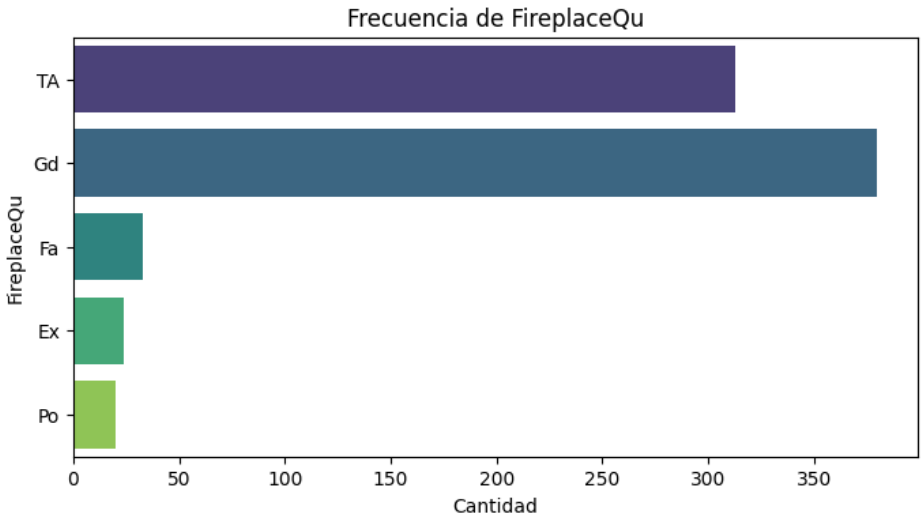
La mayoría de las propiedades tienen acceso por calle pavimentada y no tienen acceso a callejón. Los vecindarios más comunes son NAmes, CollgCr y OldTown, lo que sugiere una concentración en áreas urbanas o suburbanas. La mayoría de las propiedades tienen acceso pavimentado, lo que indica una buena accesibilidad a las vías principales.

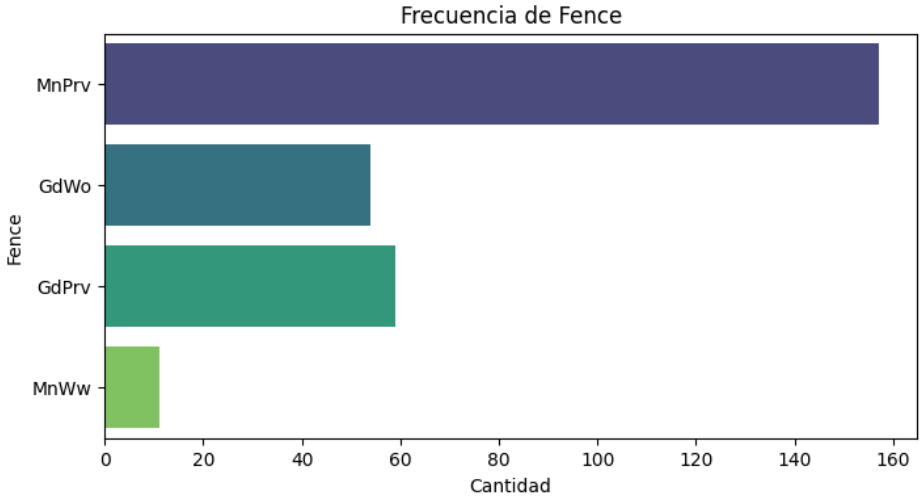
Variables relacionadas con seguridad y condiciones


Frecuencia de SaleType
WD
New
COD
ConLD
ConLI
CWD
ConLw
Con
Oth
SaleType
0
200
400
600
800
1000
1200
Cantidad


Frecuencia de SaleCondition
Normal
Abnorml
Partial
AdjLand
Alloca
Family
SaleCondition
0
200
400
600
800
1000
1200
Cantidad

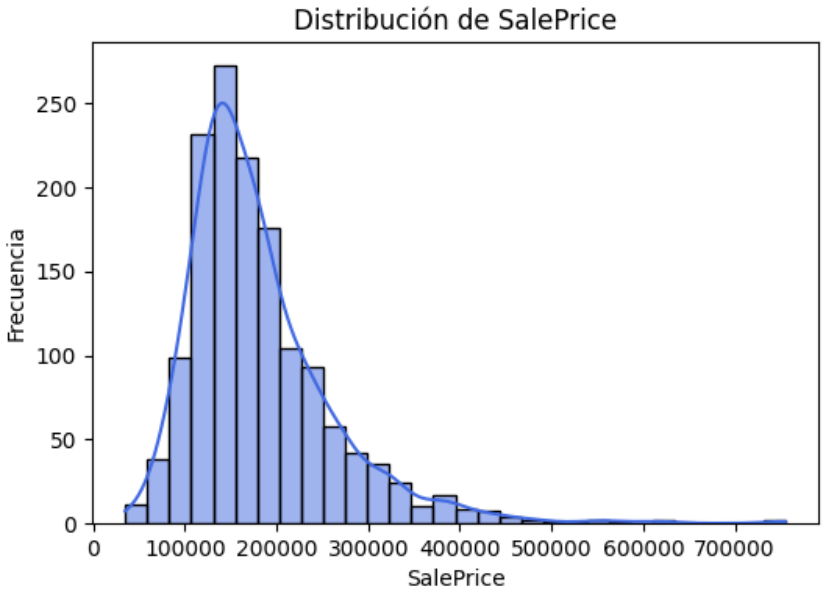

Frecuencia de Functional
Typ
Min1
Maj1
Min2
Mod
Maj2
Sev
Functional
0
200
400
600
800
1000
1200
1400
Cantidad

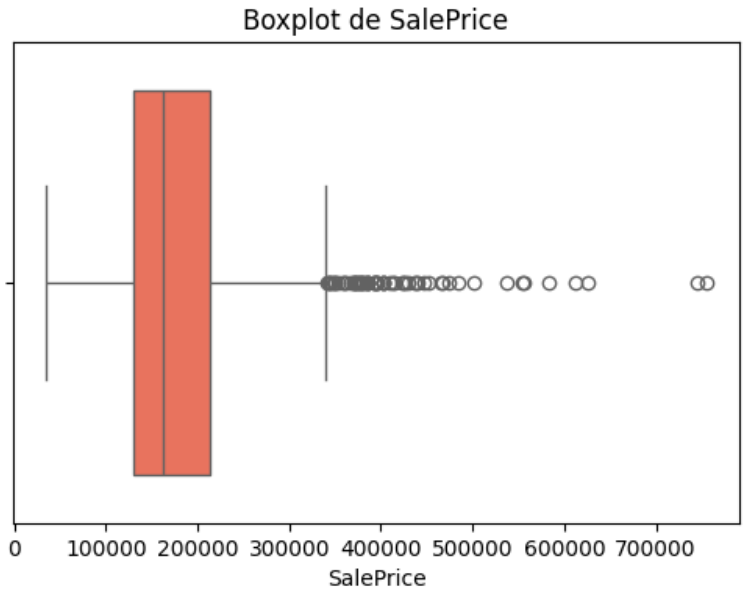




La mayoría de las propiedades se venden bajo condiciones normales y tienen funcionalidad típica. La calidad de la chimenea es promedio, con pocos casos en los extremos. La mayoría de las propiedades no tienen cercas, lo que sugiere una baja preocupación por la seguridad o privacidad en el vecindario.

Visualización de Variables Numéricas

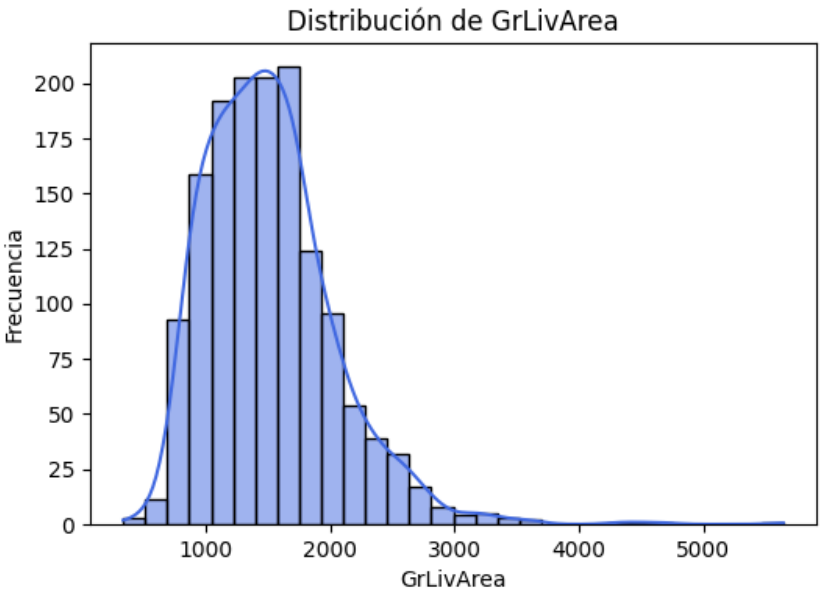


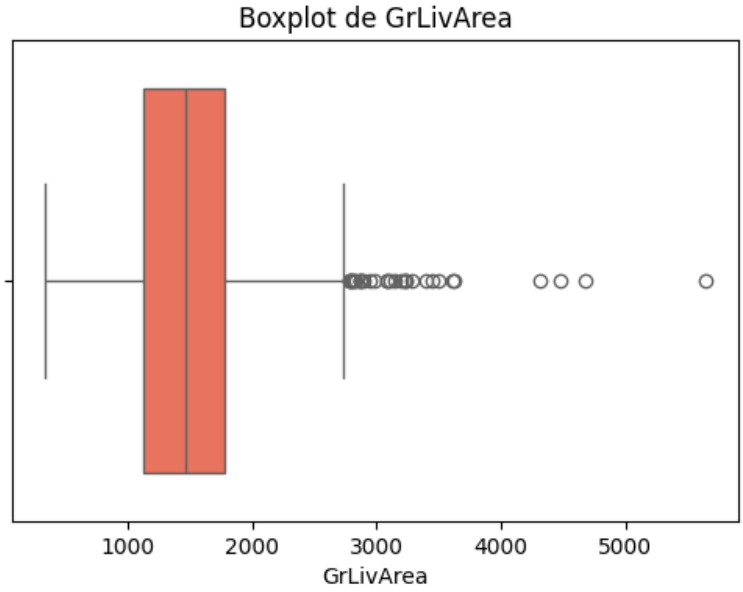


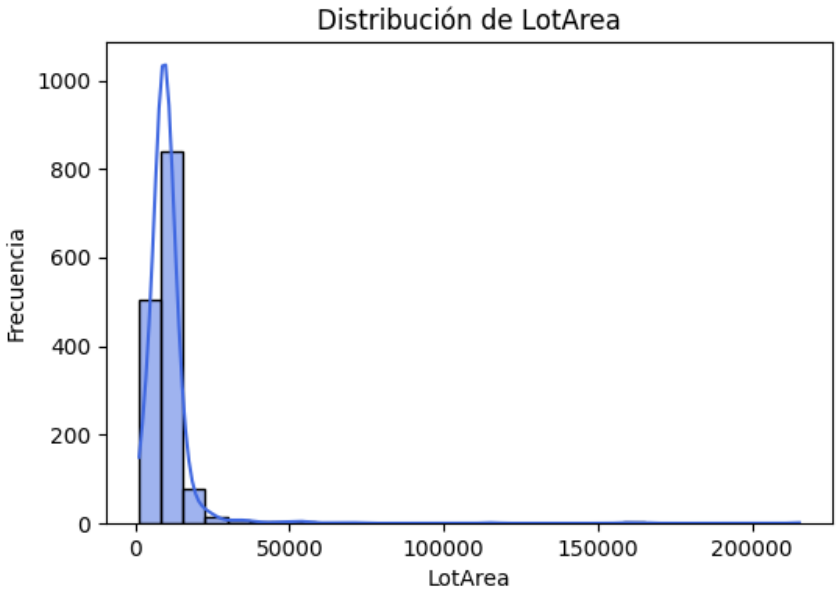
Siendo esta la variable objetivo,observamos una distribución sesgada a la derecha, Esta asimetría puede afectar métodos estadísticos que asumen distribuciones normales. Observamos outliers en la parte superior de la distribución, lo que sugiere la presencia de propiedades muy caras que pueden afectar la predicción de precios.

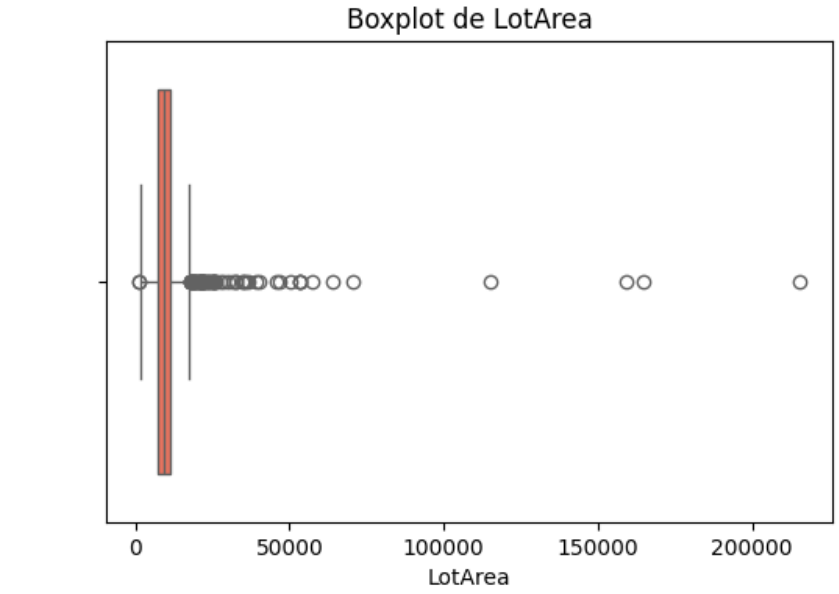
- **Boxplot:** Se aprecia que la mayoría de los precios se concentran en un rango intercuartílico entre 130,000 y 210,000 dólares, pero existen varios puntos extremos en la cola superior. Esto indica la presencia de propiedades con precios significativamente más altos.
- **Histograma con curva de densidad**: La distribución se observa sesgada a la derecha, lo que se confirma por la diferencia entre la mediana y la media. Esto sugiere que, para algunos análisis o modelado, podría ser útil aplicar una transformación para aproximar una distribución normal.

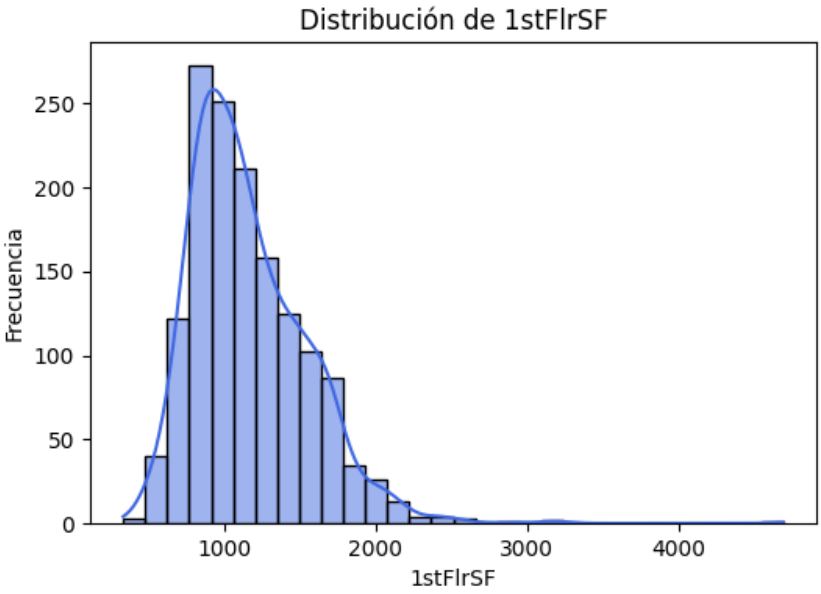
Variables Numéricas Relacionadas con Áreas y Calidad

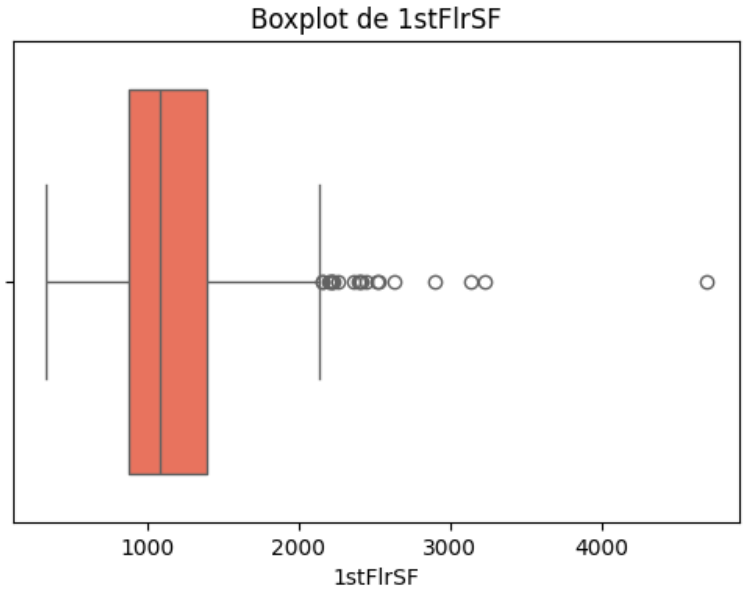


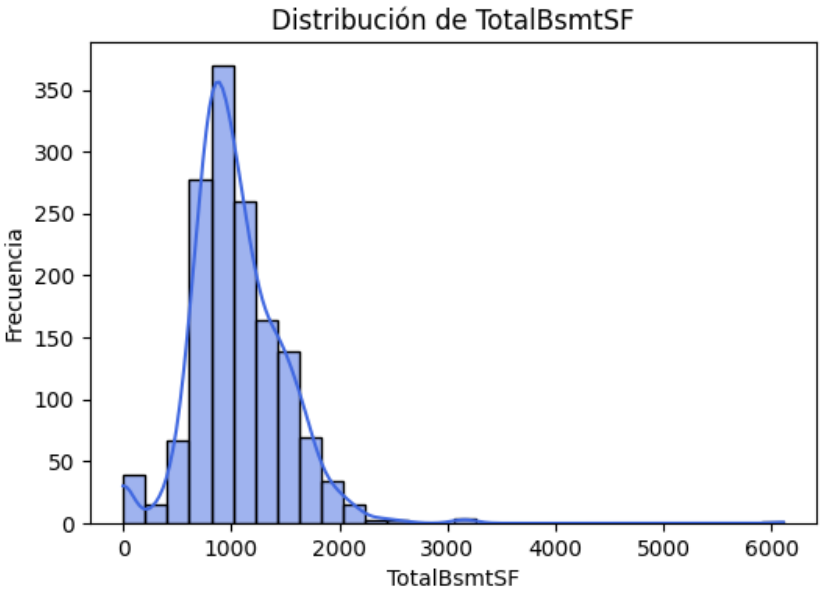


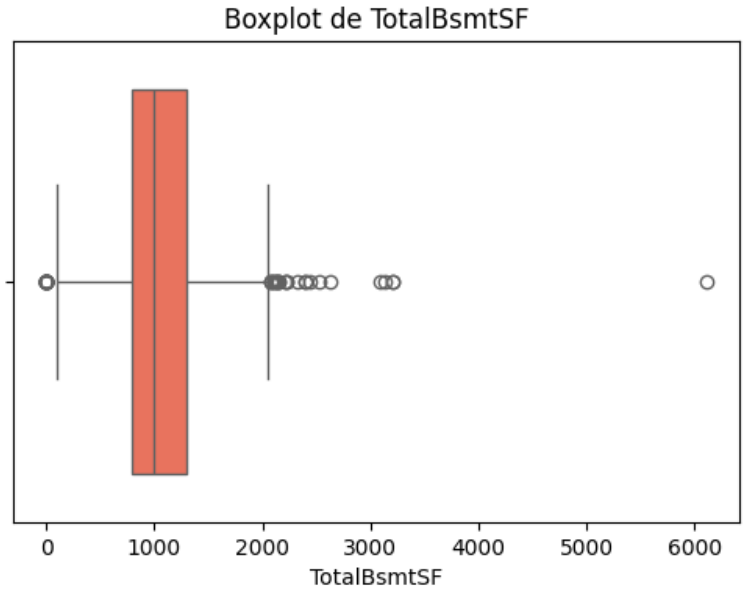

Distribución de LotArea
Frecuencia
1000
800
600
400
200
0
0
50000
100000
150000
200000
LotArea


Boxplot de LotArea
0
50000
100000
150000
200000
LotArea


Distribución de 1stFlrSF
Frecuencia
0
50
100
150
200
250
1000
2000
3000
4000
1stFlrSF


Boxplot de 1stFlrSF
1000
2000
3000
4000
1stFlrSF

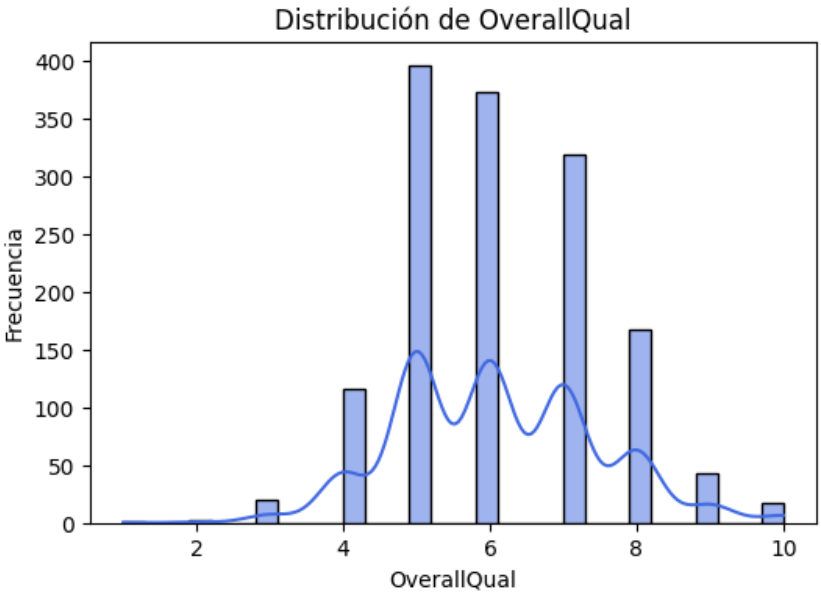


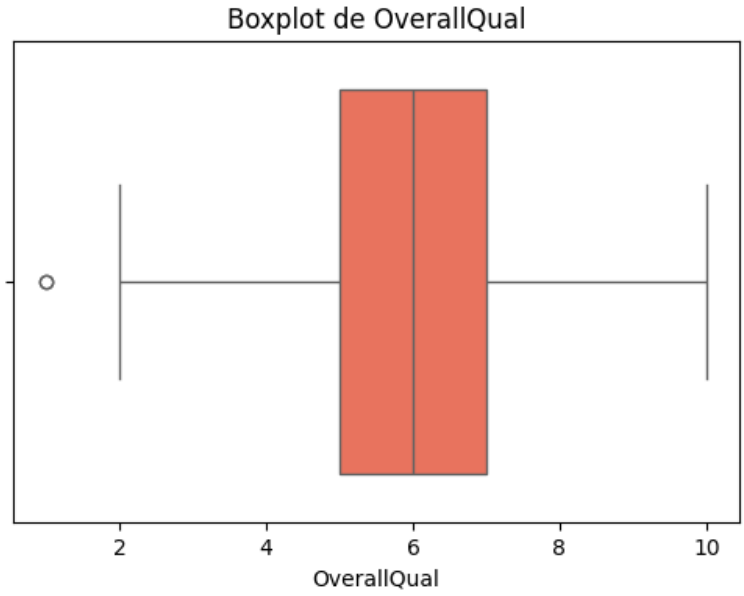


GrLivArea, LotArea, X1stFlrSF y TotalBsmtSF:

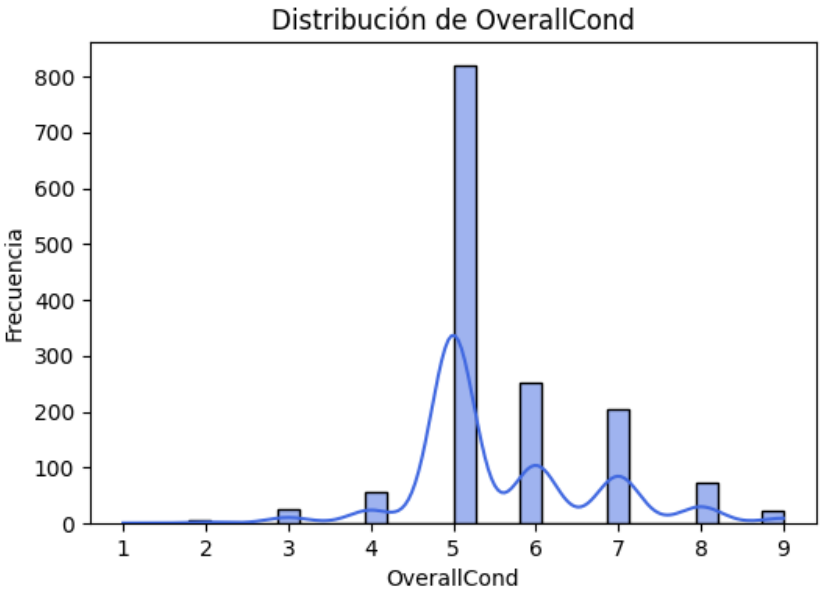
Los gráficos confirman que las variables de área tienden a ser **altamente asimétricas** y presentan **outliers**. Esto será fundamental al momento de construir modelos predictivos y al realizar inferencias estadísticas, ya que puede ser necesario **transformar** o **estratificar** estas variables para obtener resultados más confiables.

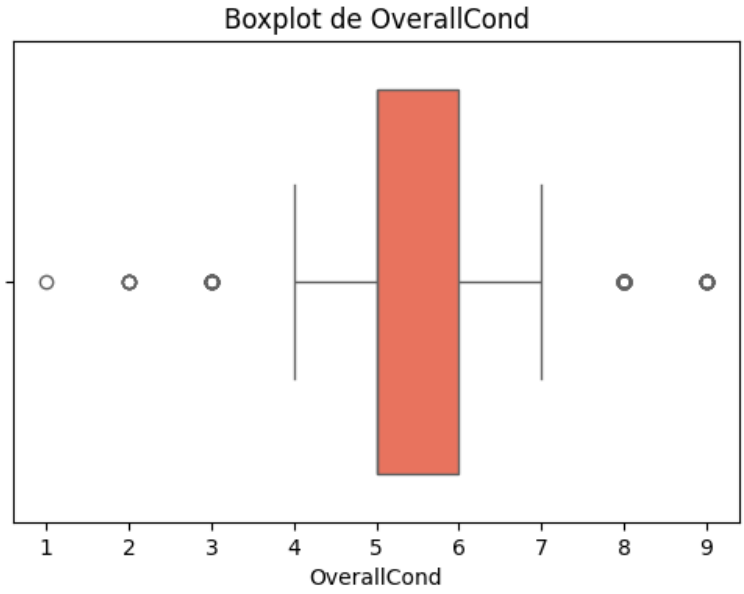
Variables Numéricas Relacionadas con Calidad y Años



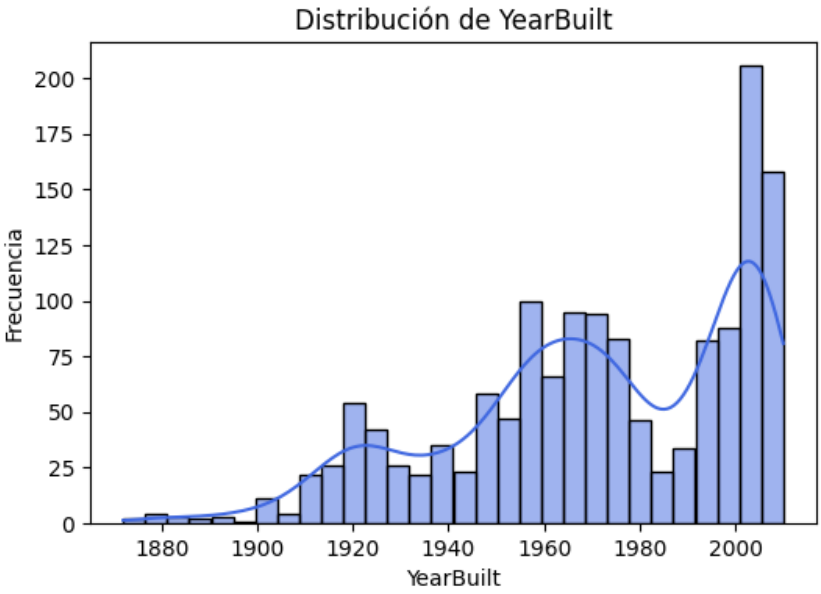


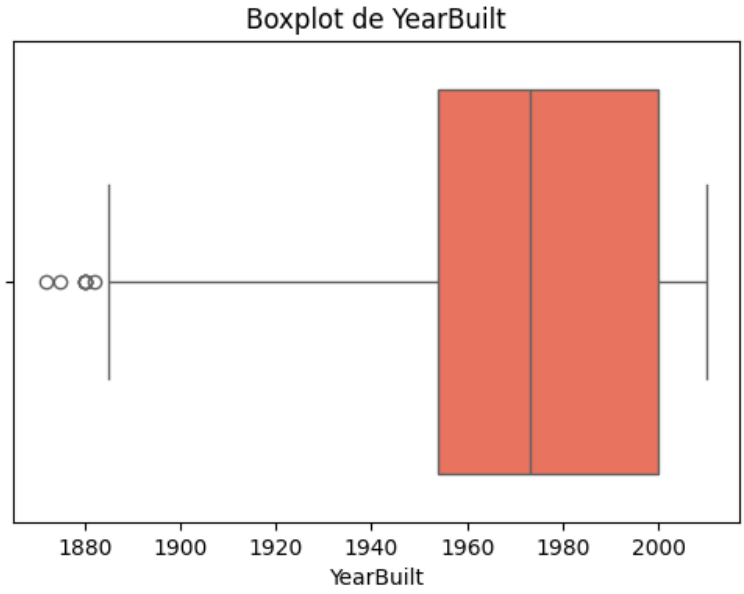
- Valores entre 1 y 10.
- Mayoría entre 5 y 7.
- Pico alrededor de 5-6.
- Pocos valores en los extremos.



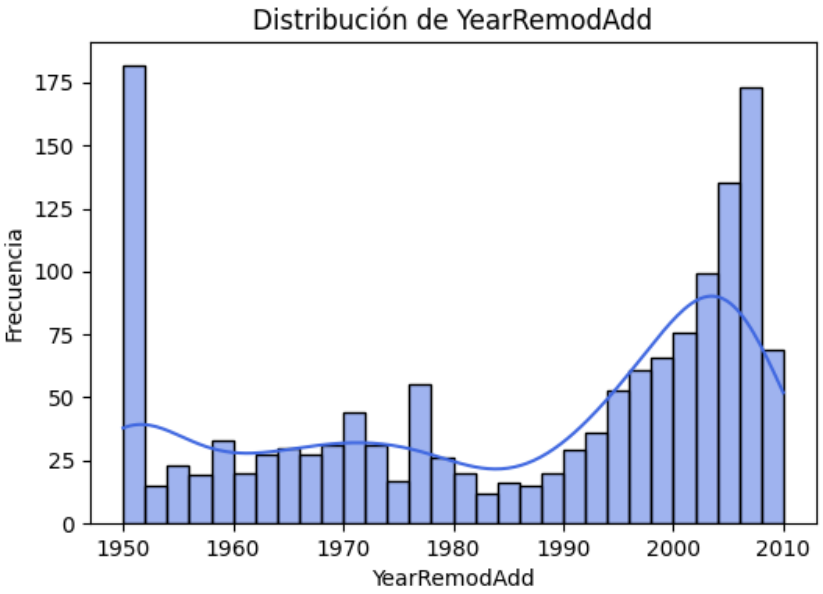


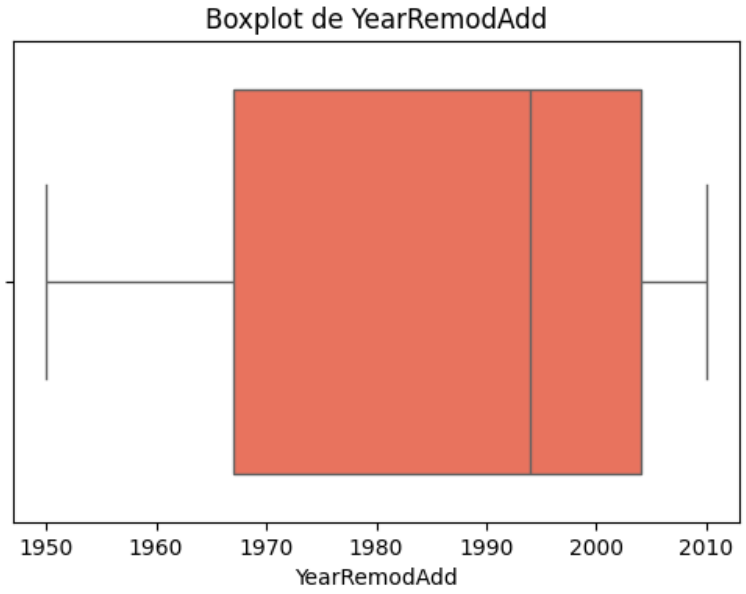
- Valores entre 1 y 9.
- Pico muy marcado en 5.
- Caja centrada en 5-6.
- Pocos casos en extremos (1, 9).





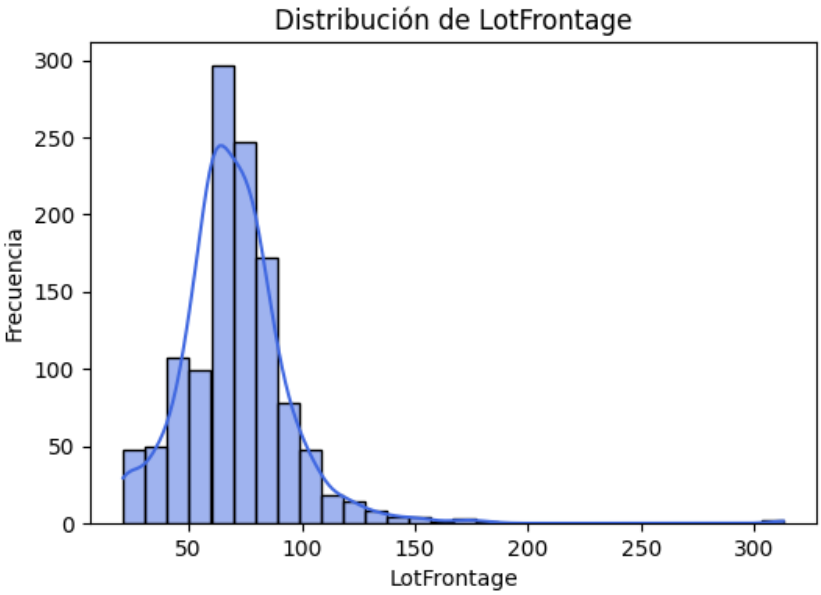
- Rango amplio (1870–2010).
- Incremento progresivo hasta 2000.
- Concentración alta en décadas recientes.
- Boxplot concentrado en 1950–2000.

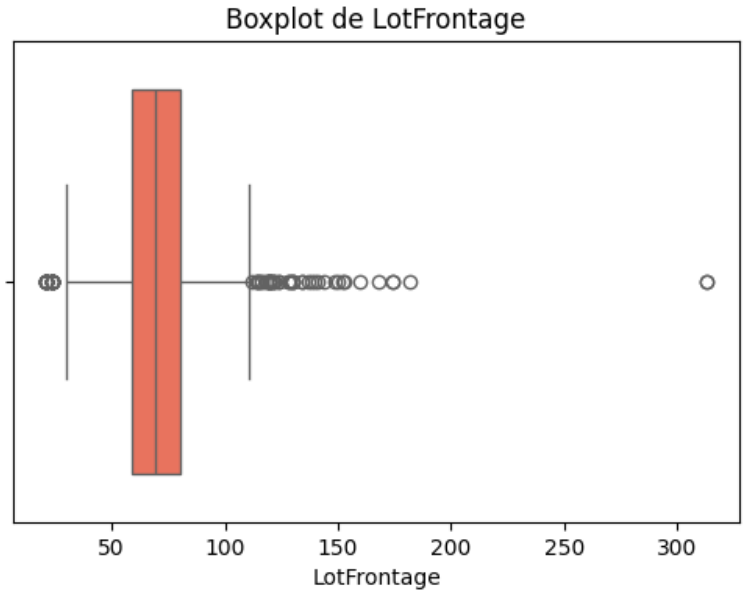




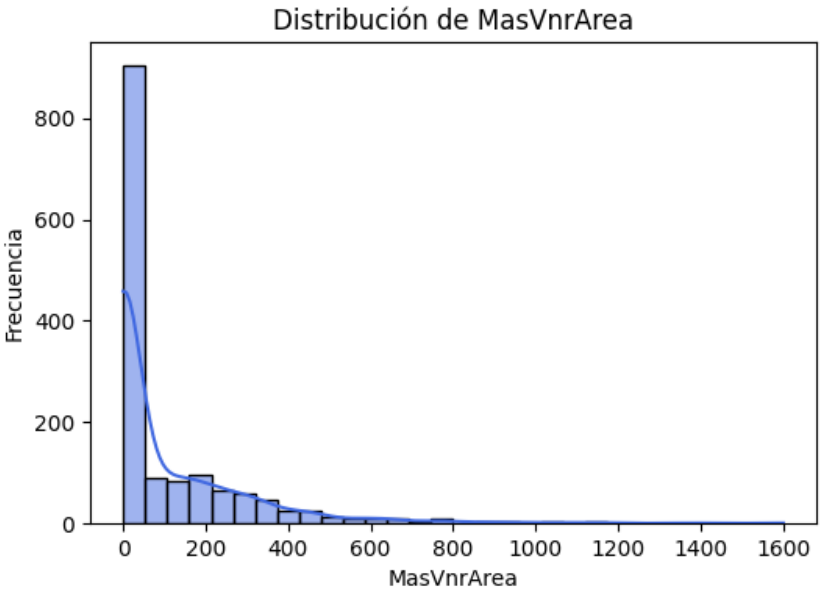
- Rango 1950–2010.
- Mayor actividad de remodelación cerca de 1990 y 2010, 1950 presenta remodelaciones altas.
- Boxplot abarca 1960–2000.
- Pocos valores anteriores a 1960.

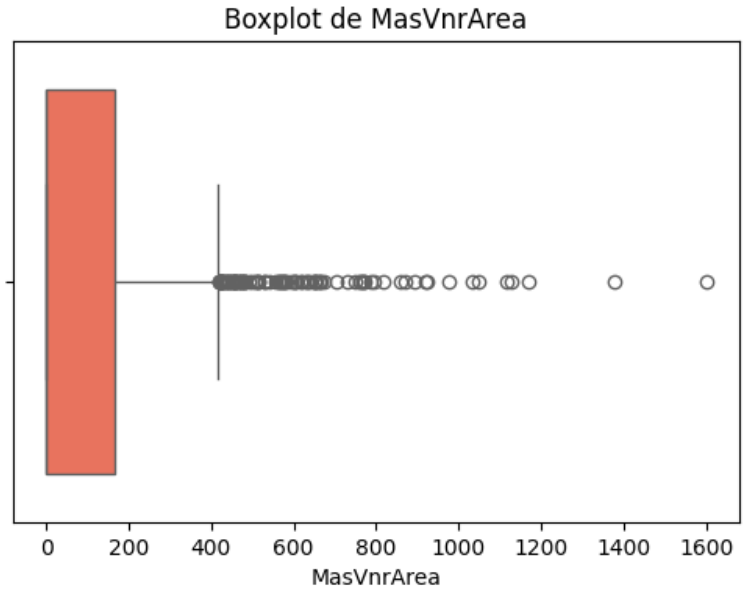
Variables Numéricas Relacionadas con Áreas y Calidad





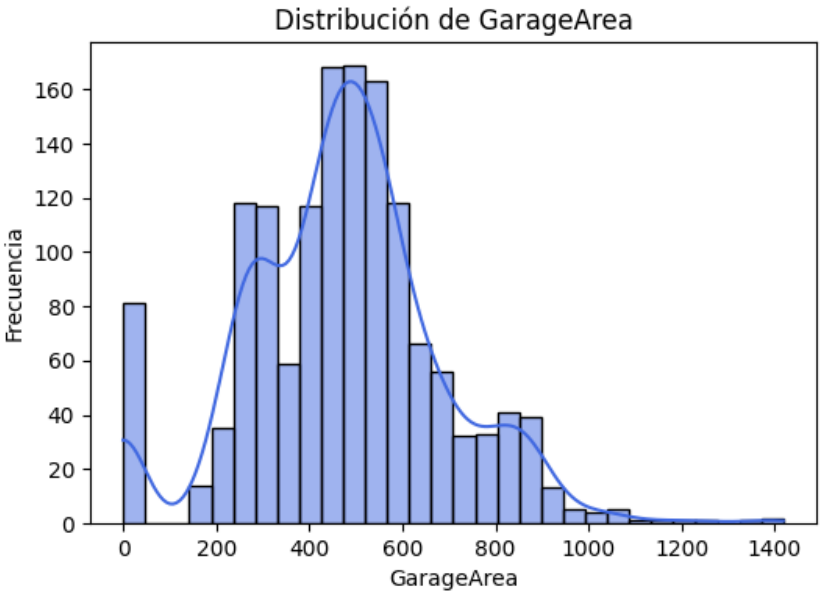
- Pico cercano a 60-70.
- Muchos valores faltantes.
- Cola derecha larga, outliers por encima de 150.

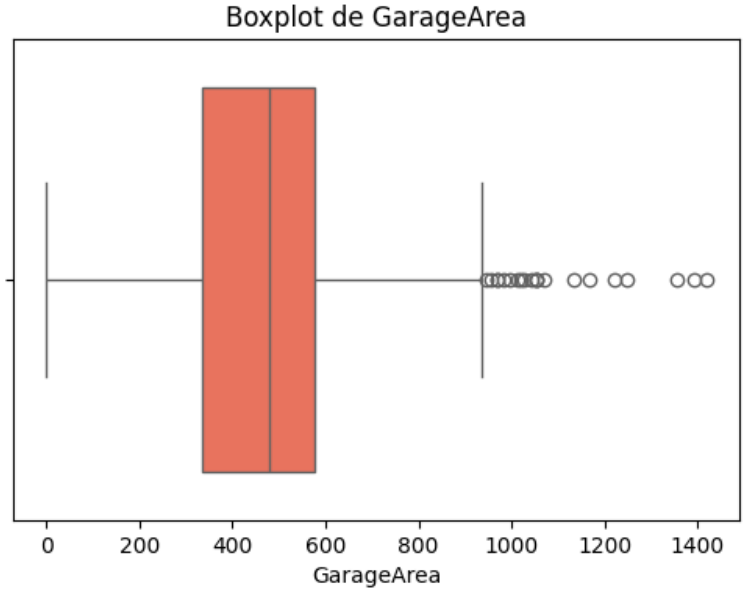




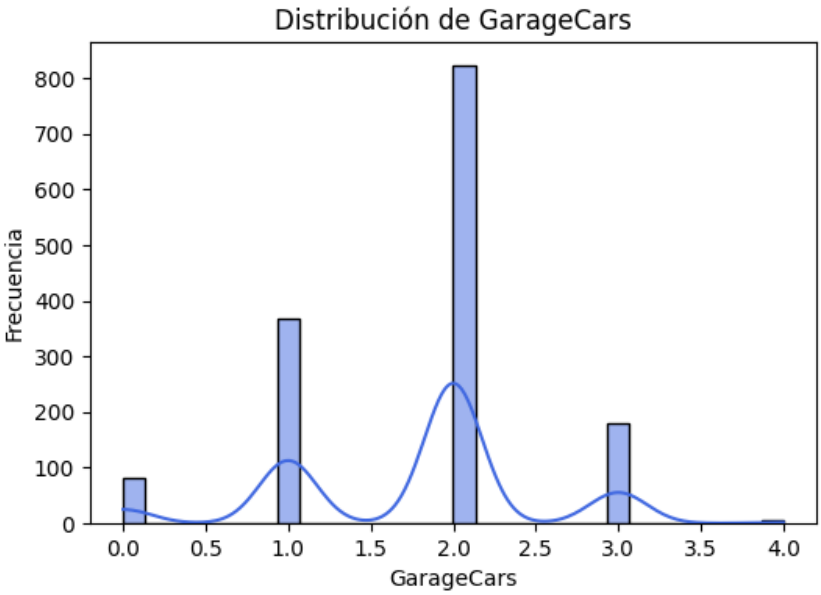
- Mayoría en 0 (sin acabado de mampostería).
- Fuerte sesgo a la derecha.
- Outliers hasta 1600.

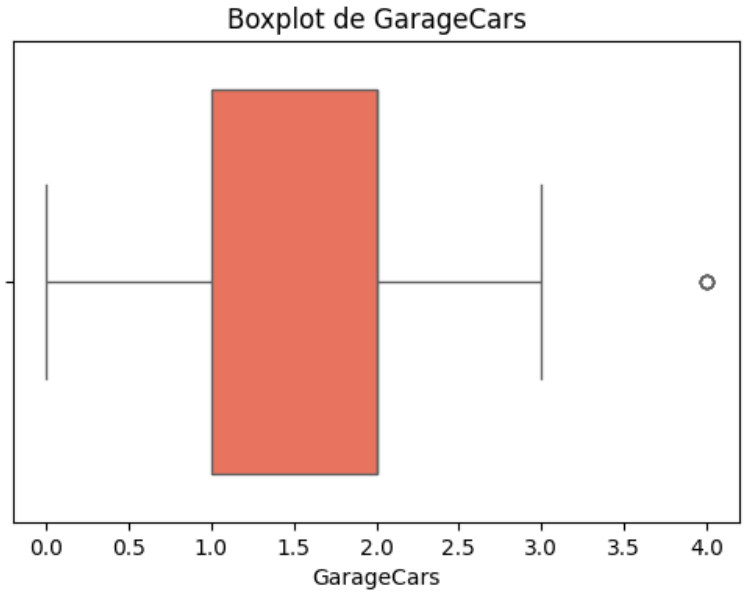
Variables Numéricas Relacionadas con el Garaje





- Mayoría entre 400–600.
- Distribución sesgada a la derecha.
- Outliers por encima de 1000.





- Pico en 2 autos.
- Rango 0–4.
- Pocos outliers en 4.

Las variables numéricas, como áreas y precios, se distribuyen con asimetría a la derecha y tienen outliers significativos. Las variables de calidad se concentran en rangos medios y se detectan datos faltantes en algunas. Esto indica que será necesario aplicar transformaciones , tratar outliers y profundizar en el análisis de las variables categóricas para extraer patrones relevantes en la valoración de propiedades

Identificación de faltantes y outliers

Variable	MissingCount	MissingPercent	UniqueValues
LotFrontage	259	17.74	65, 80, 68, 60, 84, 85, 75, NA, 51, 50, 70, 91, 72, 66, 101, 57, 44, 110, 98,
Alley	1369	93.77	NA, Grvl, Pave
PoolQC	1453	99.52	NA, Ex, Fa, Gd
Fence	1179	80.75	NA, MnPrv, GdWo, GdPrv, MnWw
MiscFeature	1406	96.30	NA, Shed, Gar2, Othr, TenC

Análisis de Outliers

Table 6: Resumen Estadístico y Cuantiles para Análisis de Outliers

Variable	Min	X1.	X5.	X25.	Median	X75.	X95.	X99.	Max
SalePrice	34900	61815.97	88000.00	129975.00	163000.0	214000.00	326100.00	442567.01	755000
GrLivArea	334	692.18	848.00	1129.50	1464.0	1776.75	2466.10	3123.48	5642
LotArea	1300	1680.00	3311.70	7553.50	9478.5	11601.50	17401.15	37567.64	215245
X1stFlrSF	334	520.00	672.95	882.00	1087.0	1391.25	1831.25	2219.46	4692
TotalBsmtSF	0	0.00	519.30	795.75	991.5	1298.25	1753.00	2155.05	6110
MasVnrArea	0	0.00	0.00	0.00	0.0	166.00	456.00	791.92	1600
GarageArea	0	0.00	0.00	334.50	480.0	576.00	850.10	1002.79	1418

Pruebas de Normalidad

Se definen grupos de variables como la variable objetivo y las variables numéricas de área, calidad y años, para evaluar su normalidad mediante pruebas estadísticas. Los resultados de las pruebas de normalidad se presentan a continuación:

Grupo 1: Variable objetivo y áreas

```
##
##
## Table: Pruebas de Normalidad para SalePrice
##
## |Variable  |Test               | Statistic| P.value|
## |:---------|:------------------|---------:|-------:|
## |SalePrice |Shapiro-Wilk       |    0.8697|       0|
## |SalePrice |Anderson-Darling   |   41.6920|       0|
## |SalePrice |Kolmogorov-Smirnov |    0.1237|       0|
## |SalePrice |Lilliefors         |    0.1237|       0|

##
##
## Table: Pruebas de Normalidad para GrLivArea
##
## |Variable  |Test               | Statistic| P.value|
## |:---------|:------------------|---------:|-------:|
```

```
## |GrLivArea |Shapiro-Wilk       |    0.9280|       0|
## |GrLivArea |Anderson-Darling   |   14.5322|       0|
## |GrLivArea |Kolmogorov-Smirnov |    0.0675|       0|
## |GrLivArea |Lilliefors         |    0.0675|       0|
```

```
##
##
## Table: Pruebas de Normalidad para LotArea
##
## |Variable |Test               | Statistic| P.value|
## |:--------|:------------------|---------:|-------:|
## |LotArea  |Shapiro-Wilk       |    0.3511|       0|
## |LotArea  |Anderson-Darling   |  198.4183|       0|
## |LotArea  |Kolmogorov-Smirnov |    0.2515|       0|
## |LotArea  |Lilliefors         |    0.2515|       0|
```

```
##
##
## Table: Pruebas de Normalidad para X1stFlrSF
##
## |Variable  |Test               | Statistic| P.value|
## |:---------|:------------------|---------:|-------:|
## |X1stFlrSF |Shapiro-Wilk       |    0.9269|       0|
## |X1stFlrSF |Anderson-Darling   |   19.1651|       0|
## |X1stFlrSF |Kolmogorov-Smirnov |    0.0869|       0|
## |X1stFlrSF |Lilliefors         |    0.0869|       0|
```

```
##
##
## Table: Pruebas de Normalidad para TotalBsmtSF
##
## |Variable    |Test               | Statistic| P.value|
## |:-----------|:------------------|---------:|-------:|
## |TotalBsmtSF |Shapiro-Wilk       |    0.9174|       0|
## |TotalBsmtSF |Anderson-Darling   |   17.2764|       0|
## |TotalBsmtSF |Kolmogorov-Smirnov |    0.0760|       0|
## |TotalBsmtSF |Lilliefors         |    0.0760|       0|
```

Grupo 2: Variables de calidad y construcción

```
##
##
## Table: Pruebas de Normalidad para OverallQual
##
## |Variable    |Test               | Statistic| P.value|
## |:-----------|:------------------|---------:|-------:|
## |OverallQual |Shapiro-Wilk       |    0.9480|       0|
## |OverallQual |Anderson-Darling   |   35.2300|       0|
## |OverallQual |Kolmogorov-Smirnov |    0.1552|       0|
## |OverallQual |Lilliefors         |    0.1552|       0|
```

```
##
```

```
##
## Table: Pruebas de Normalidad para OverallCond
##
## |Variable    |Test               | Statistic| P.value|
## |:-----------|:------------------|---------:|-------:|
## |OverallCond |Shapiro-Wilk       |    0.8289|       0|
## |OverallCond |Anderson-Darling   |  125.2851|       0|
## |OverallCond |Kolmogorov-Smirnov |    0.3200|       0|
## |OverallCond |Lilliefors         |    0.3200|       0|

##
##
## Table: Pruebas de Normalidad para YearBuilt
##
## |Variable  |Test               | Statistic| P.value|
## |:---------|:------------------|---------:|-------:|
## |YearBuilt |Shapiro-Wilk       |    0.9256|       0|
## |YearBuilt |Anderson-Darling   |   30.9635|       0|
## |YearBuilt |Kolmogorov-Smirnov |    0.1209|       0|
## |YearBuilt |Lilliefors         |    0.1209|       0|

##
##
## Table: Pruebas de Normalidad para YearRemodAdd
##
## |Variable     |Test               | Statistic| P.value|
## |:------------|:------------------|---------:|-------:|
## |YearRemodAdd |Shapiro-Wilk       |    0.8628|       0|
## |YearRemodAdd |Anderson-Darling   |   71.4944|       0|
## |YearRemodAdd |Kolmogorov-Smirnov |    0.1745|       0|
## |YearRemodAdd |Lilliefors         |    0.1745|       0|
```

Grupo 3: Variables relacionadas con acabados y garaje

```
##
##
## Table: Pruebas de Normalidad para MasVnrArea
##
## |Variable   |Test               | Statistic| P.value|
## |:----------|:------------------|---------:|-------:|
## |MasVnrArea |Shapiro-Wilk       |    0.6393|       0|
## |MasVnrArea |Anderson-Darling   |  182.6180|       0|
## |MasVnrArea |Kolmogorov-Smirnov |    0.3095|       0|
## |MasVnrArea |Lilliefors         |    0.3095|       0|

##
##
## Table: Pruebas de Normalidad para GarageArea
##
## |Variable   |Test               | Statistic| P.value|
## |:----------|:------------------|---------:|-------:|
## |GarageArea |Shapiro-Wilk       |    0.9753|       0|
## |GarageArea |Anderson-Darling   |    9.2333|       0|
```

```
## |GarageArea |Kolmogorov-Smirnov |    0.0753|       0|
## |GarageArea |Lilliefors         |    0.0753|       0|
```

Las pruebas de normalidad en todos los grupos de variables arrojan p-valores extremadamente bajos ($p < 2.2e\text{-}16$ en la mayoría de los casos), lo que indica que ninguna de estas variables sigue una distribución normal según los test de Shapiro-Wilk, Anderson-Darling, Kolmogorov-Smirnov y Lilliefors. Esto es especialmente notable en variables como SalePrice, LotArea y MasVnrArea, que presentan un marcado sesgo a la derecha y outliers. Aunque algunas variables tienen valores de W relativamente altos, el tamaño de la muestra (n=1460) hace que incluso desviaciones leves se vuelvan estadísticamente significativas. En conclusión, la evidencia sugiere que es necesario aplicar transformaciones y/o estrategias de manejo de outliers para aproximar la normalidad y estabilizar la varianza antes de proceder con el modelado predictivo.

A partir de esta exploración inicial, se identificaron patrones y características clave en las variables categóricas y numéricas. Estos insights serán fundamentales para la limpieza, transformación y modelado de los datos, permitiendo construir modelos predictivos precisos y robustos.

Adicionalmente,surgen interrogantes sobre la relación entre las variables y su impacto en el precio de venta, por lo que previo a las transformaciones las cuales se responden de manera iterativa en el análisis exploratorio de datos. A continuación, se presentan las preguntas de investigación que guiarán el análisis y modelado de los datos:

1. **¿Cómo se relacionan las variables de área (GrLivArea, LotArea, X1stFlrSF, TotalBsmtSF) con el precio de venta y cómo varían estas relaciones según categorías de calidad (OverallQual, OverallCond) y ubicación (Neighborhood, MSZoning)?**
2. **¿Qué impacto tienen los años de construcción y remodelación (YearBuilt, YearRemodAdd) en el precio? ¿Existen tendencias o agrupaciones de propiedades antiguas versus modernas que influyan en SalePrice?**
3. **¿Cuáles son las diferencias en la distribución de precios entre los distintos tipos de construcción y estilos de vivienda (BldgType, HouseStyle), y qué patrones se observan en función de la estructura física de la propiedad?**
4. **¿De qué manera afectan los acabados exteriores y materiales (Exterior1st, Exterior2nd, MasVnrType, MasVnrArea) la valoración de las viviendas? ¿Se observa que ciertos materiales o condiciones exteriores se asocian a precios más altos o más bajos?**
5. **¿Cómo influyen las condiciones y características del sótano (BsmtQual, BsmtCond, BsmtFinType1, BsmtFinSF1, BsmtFinSF2) en el precio?¿Existe un efecto diferencial entre casas con sótanos terminados y sin terminar?**
6. **¿Qué rol juegan las variables relacionadas con el garaje (GarageType, GarageArea, GarageCars, GarageQual, GarageCond) en la determinación del precio de venta? ¿Están las propiedades con garajes de mejor calidad o mayor capacidad asociadas a precios superiores?**
7. **¿Existen patrones de desequilibrio o baja representatividad en ciertas variables categóricas (por ejemplo, Alley, PoolQC, MiscFeature) que requieran agrupar categorías o realizar recodificaciones para un análisis más fiable?**
8. **¿Cómo se comportan las variables relacionadas con la ubicación y configuración del terreno (LotShape, LandContour, Street, Utilities) y qué relación tienen con el precio de venta?**
9. **¿Qué variables muestran mayor presencia de outliers o sesgo en su distribución, y cuál es el impacto de estos extremos en los modelos predictivos? ¿Es necesario aplicar transformaciones (como logaritmos) o segmentaciones específicas?**
10. **¿Cómo se combinan las variables de calidad, área y ubicación para explicar de forma conjunta la variabilidad en el precio de las propiedades?**

Análisis de Grupos

Modelado

Discusión y Conclusiones