

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

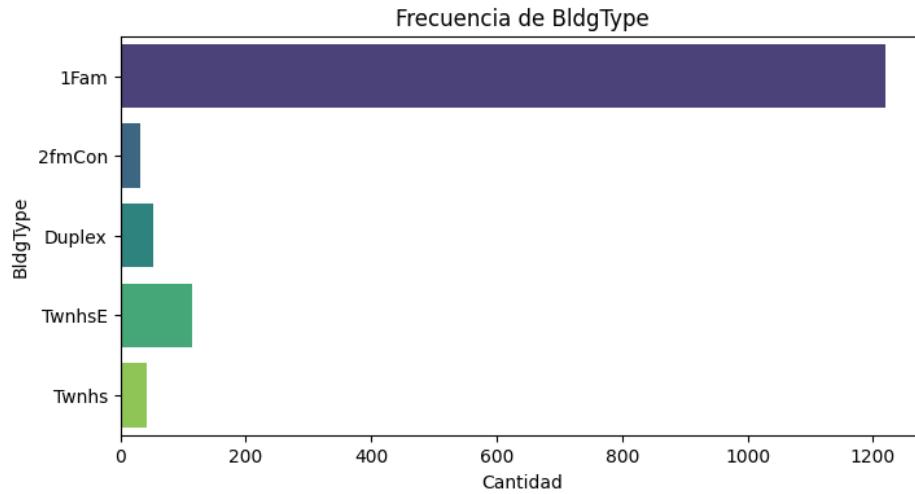
	count	mean	std	min	Q1.25%	Median.50%	Q3.75%	max	mediana
SalePrice	1460	180921.20	79442.50	34900	129975.00	163000.0	214000.00	755000	163000.0

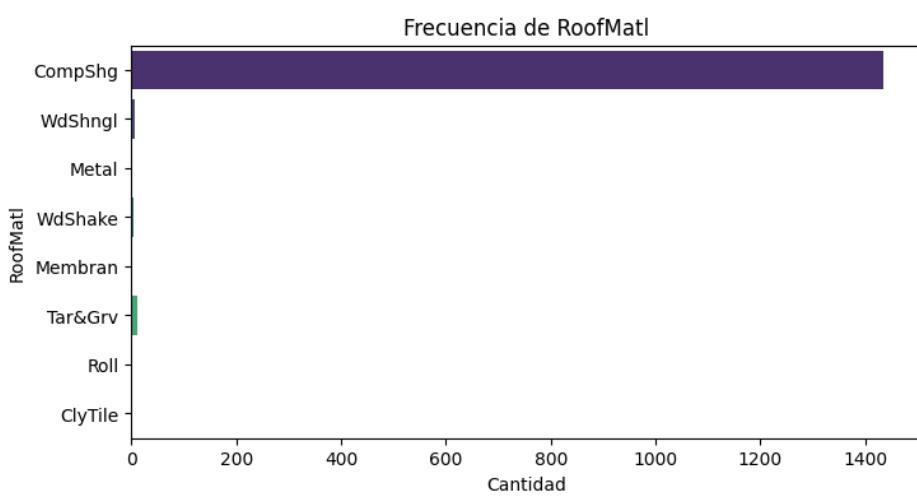
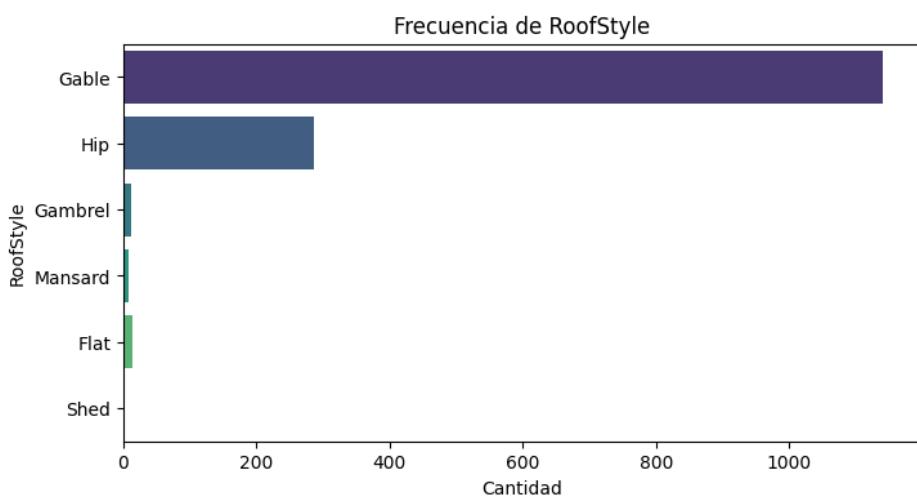
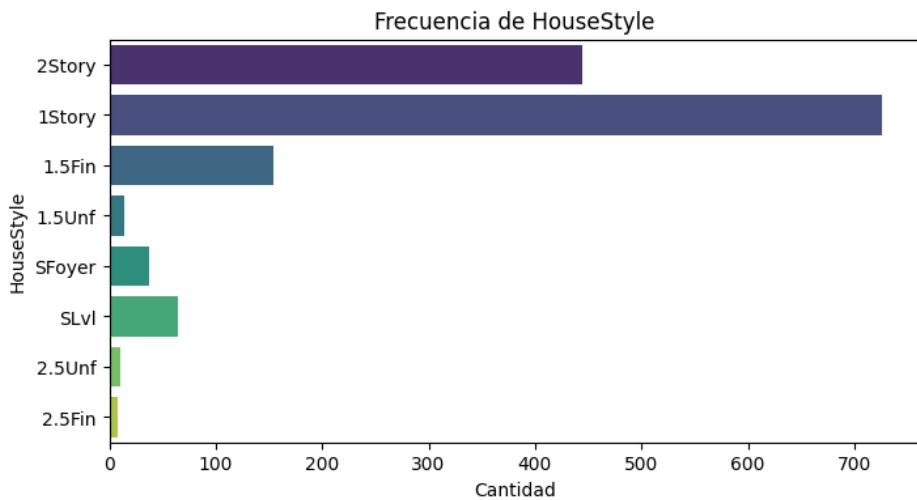
Estas estadísticas descriptivas nos permiten tener una idea general de la distribución de las variables numéricas en el dataset. A partir de estos datos podemos explorar variables con gran variabilidad y outliers como:

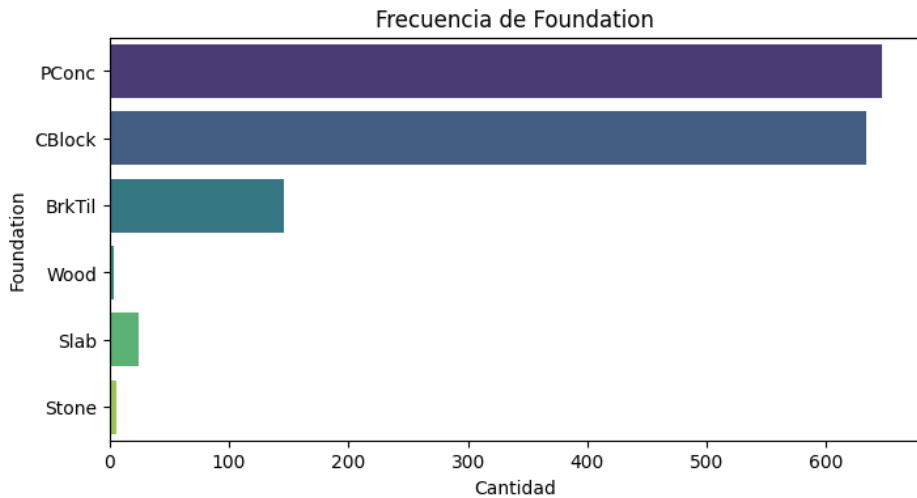
- **SalePrice:** Es la variable objetivo; analizar su distribución es esencial para detectar sesgos o valores atípicos que puedan afectar modelos predictivos.
- **GrLivArea, LotArea, X1stFlrSF y TotalBsmtSF:** Estas variables relacionadas con áreas muestran amplios rangos y desviaciones estándar elevadas, 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

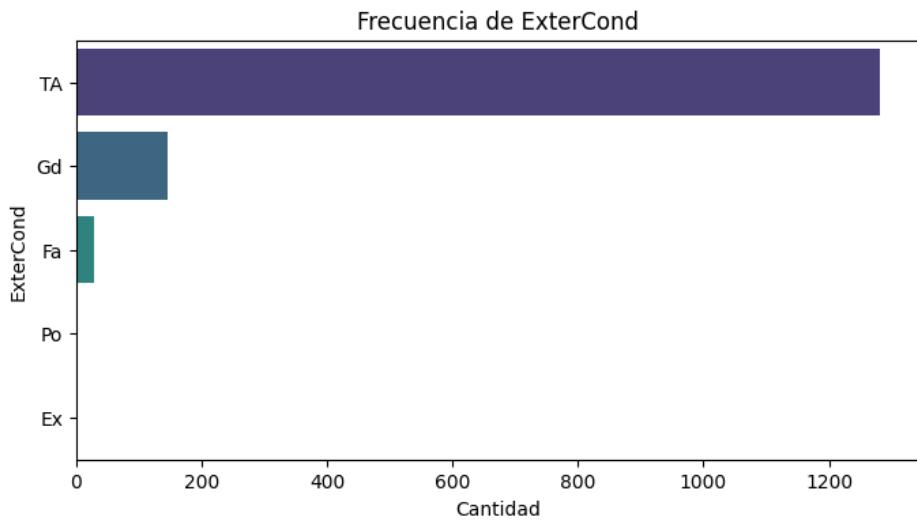


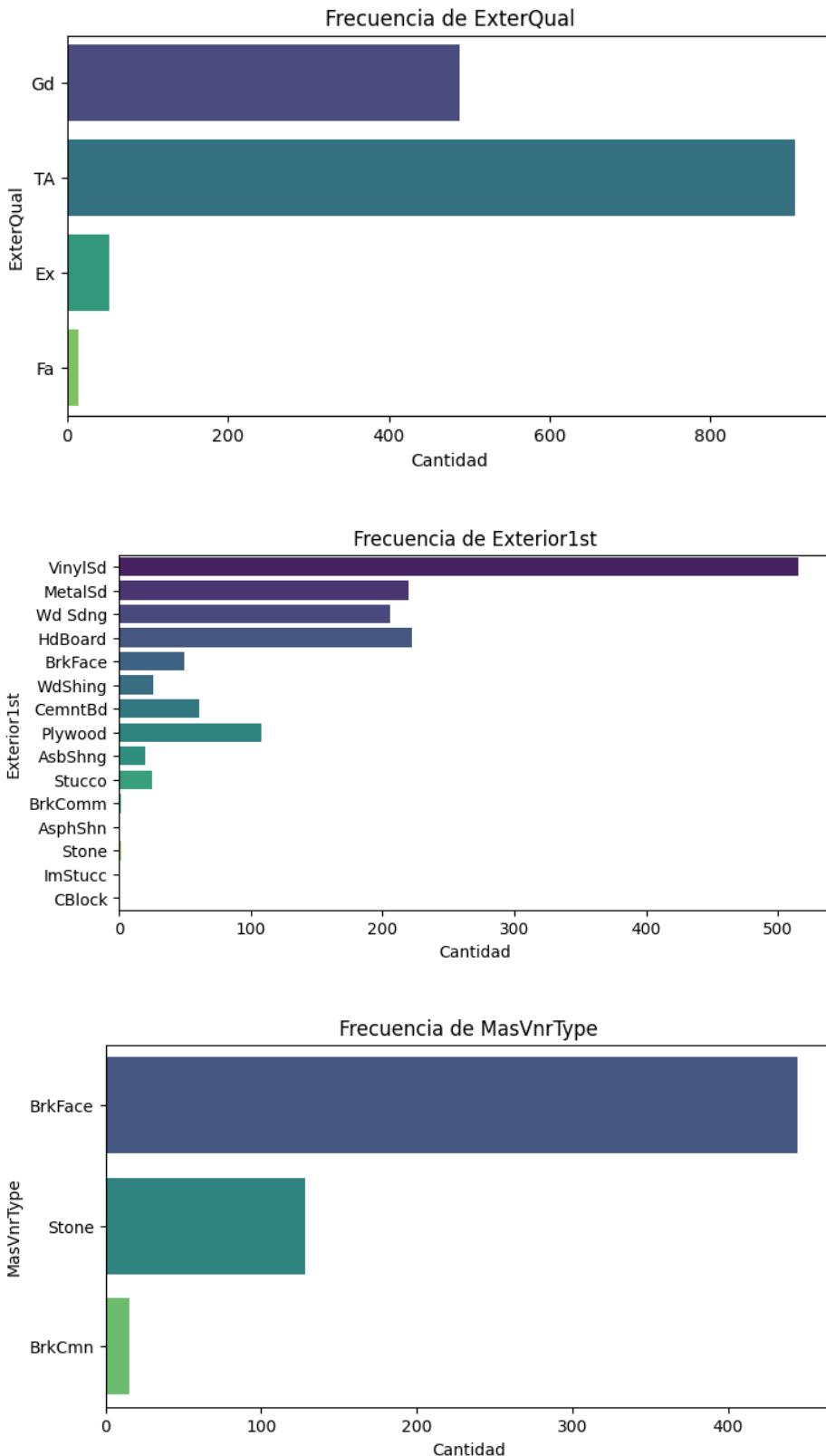




Este grupo de variables muestra una mayoría de casas unifamiliares, predominancia en casas de 2 y 1 piso, techos de tipo Gable y materiales de techos CompShg. La mayoría de las casas tienen cimientos de concreto y madera. Estos patrones pueden ser útiles para identificar características comunes en la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

Variables relacionadas con el exterior y materiales

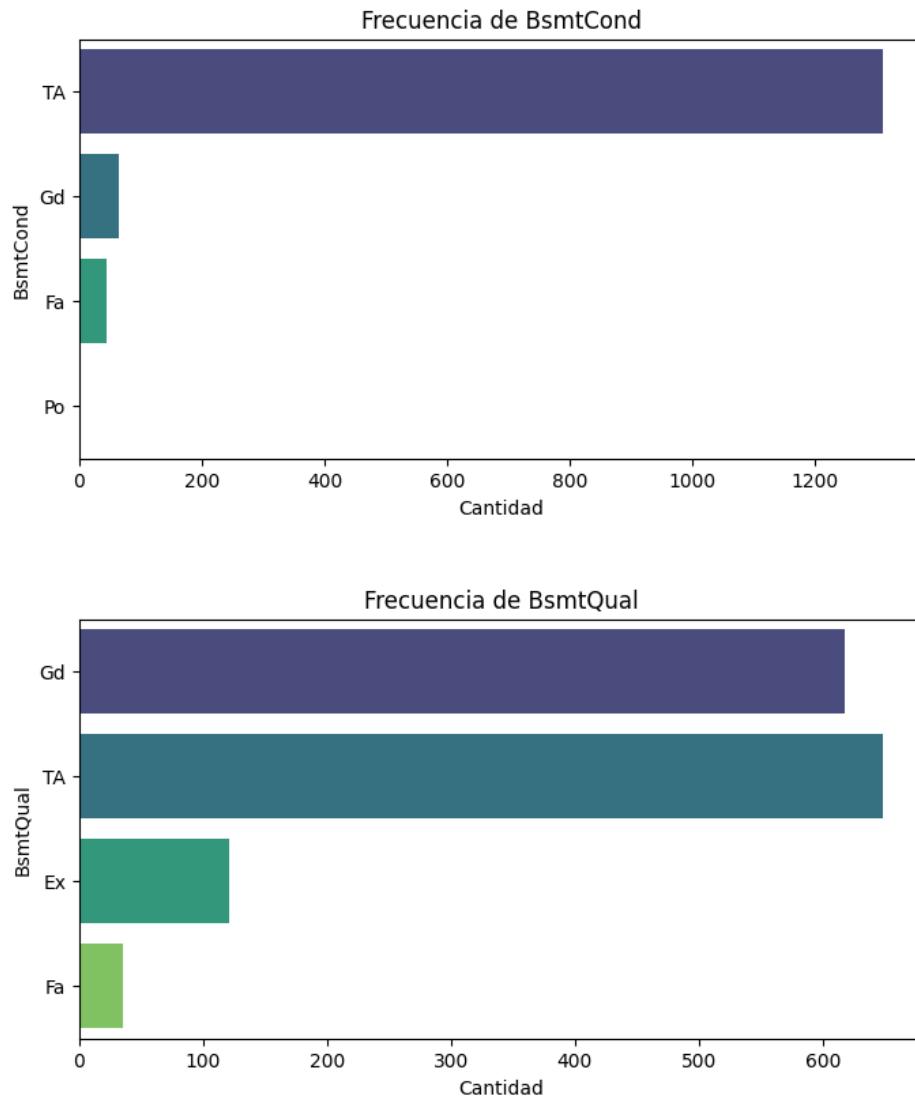


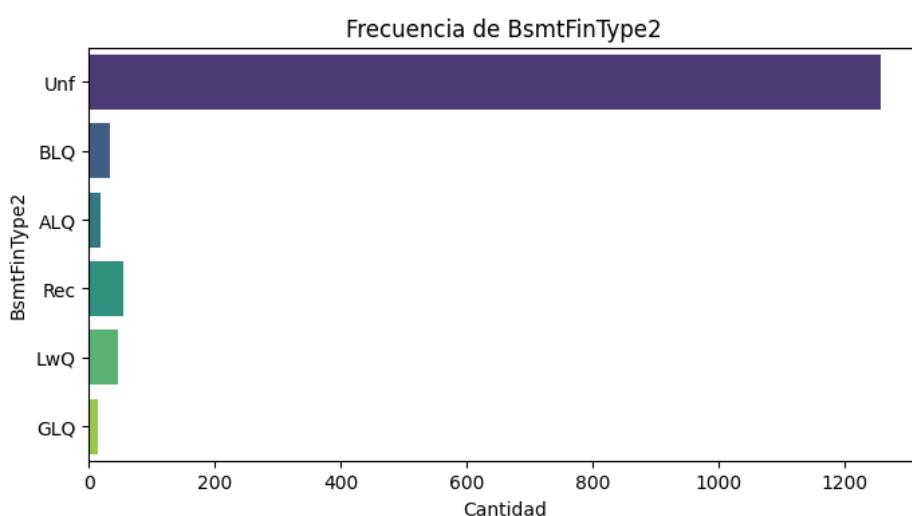
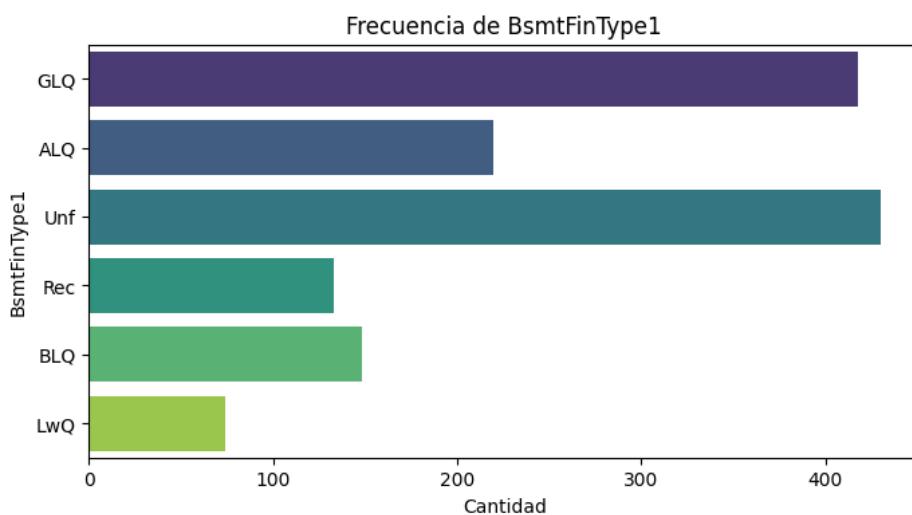
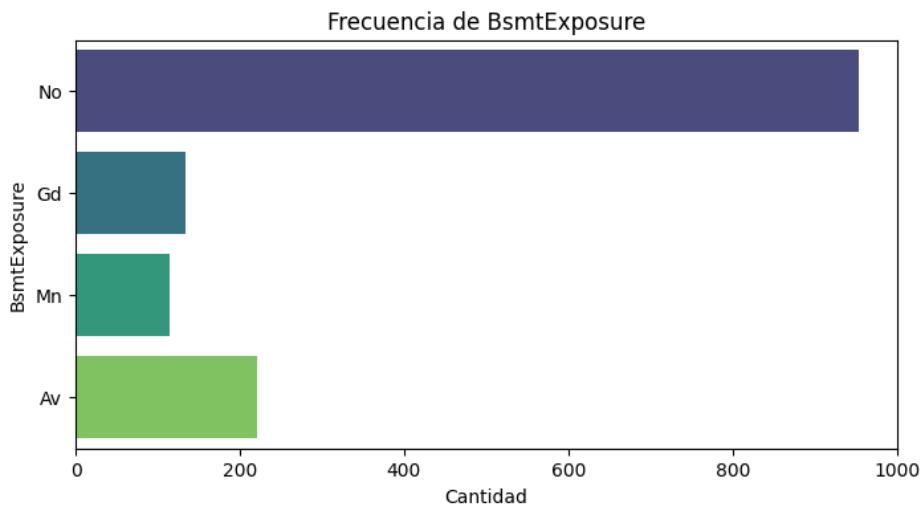


La mayoría de las casas presentan una condición y calidad exterior promedio, con pocas en estado excelente o deficiente. En las cubiertas exteriores, domina “VinylSd” tanto en la primera como en la segunda capa,

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

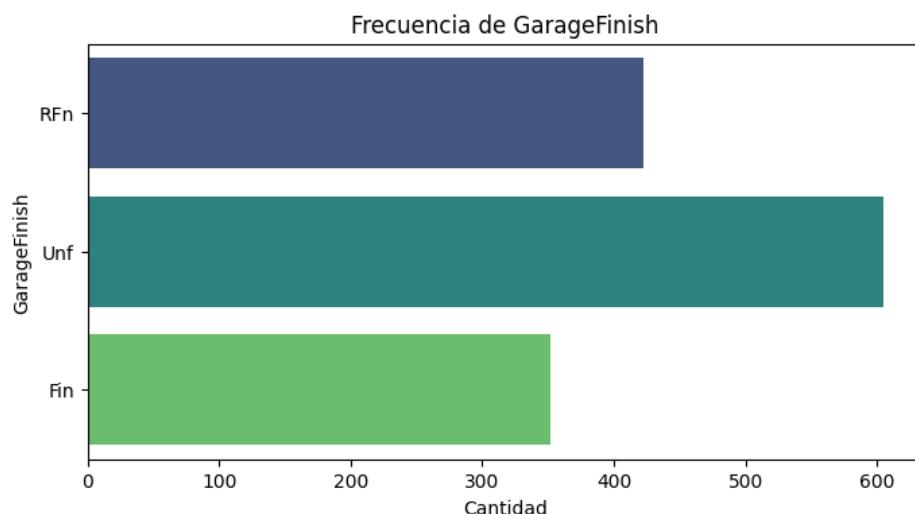
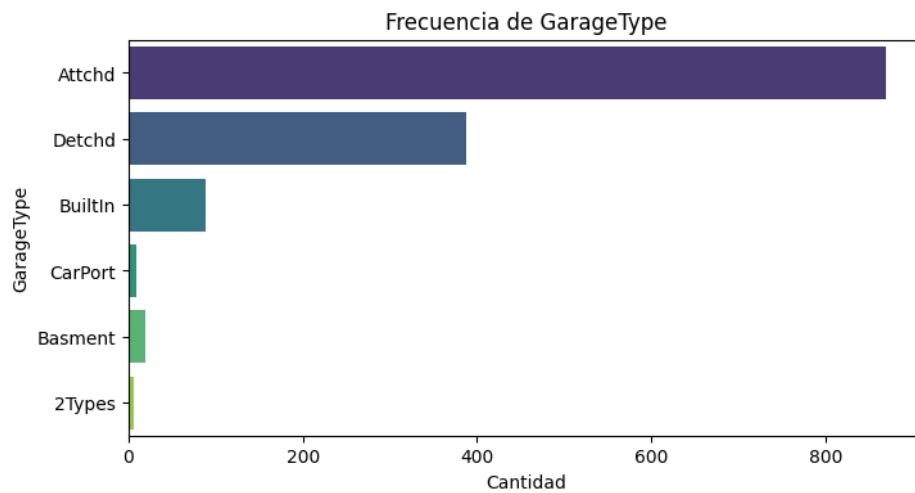


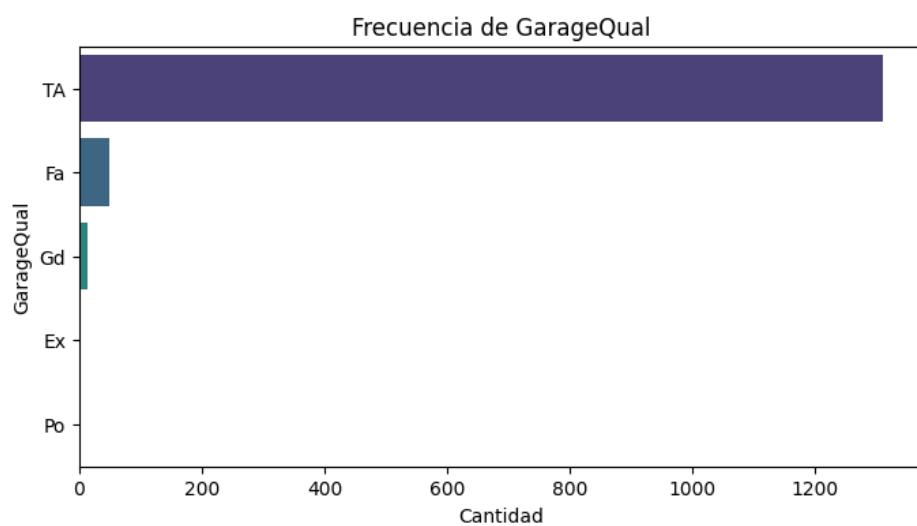
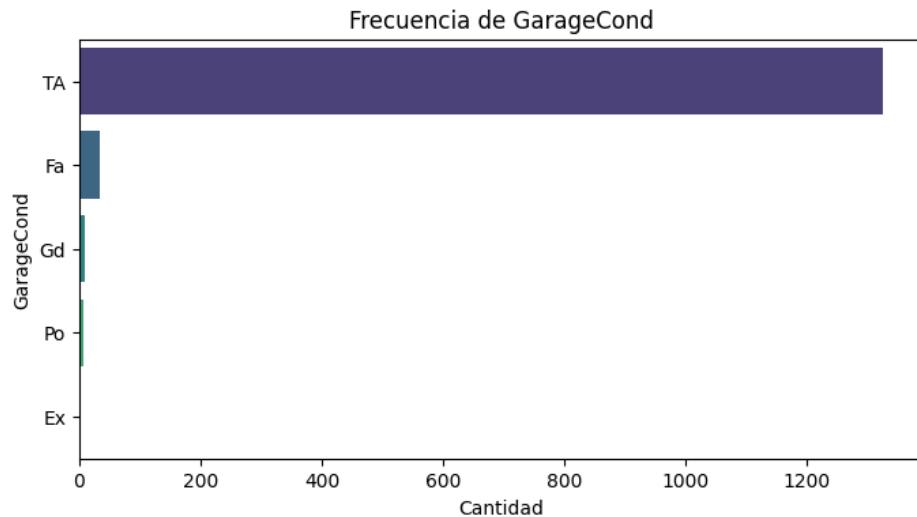


La mayoría de los sótanos están en condición “TA” y calidad “TA” o “Gd”, con pocos casos “Ex” o “Fa”. La exposición del sótano suele ser “No” (sin exposición), aunque también hay un grupo con “Gd”, “Mn” y

“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.

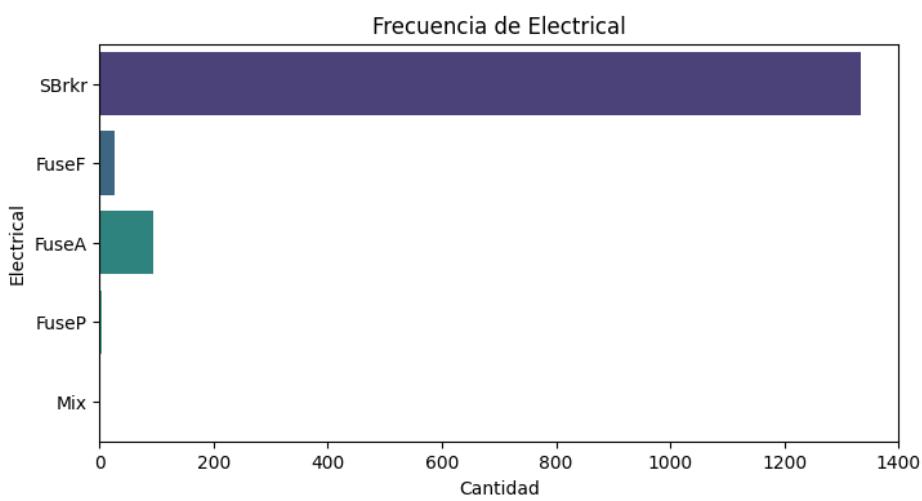
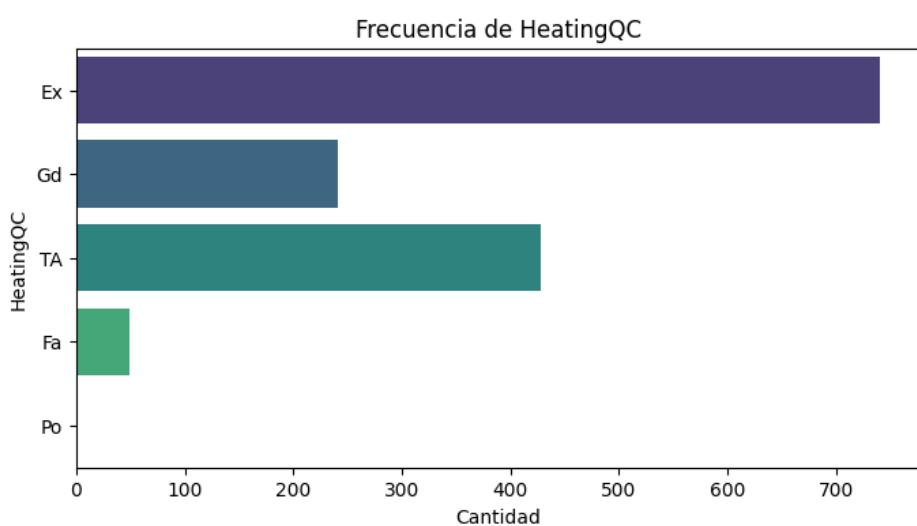
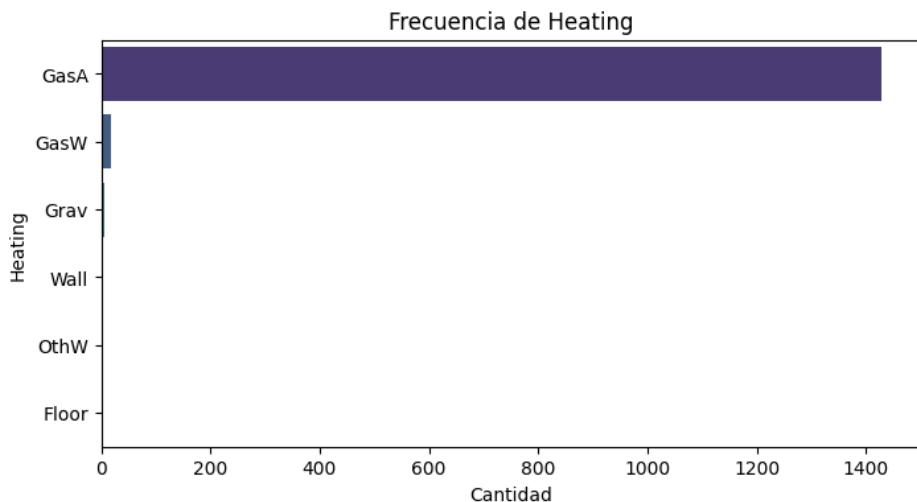
Variables relacionadas con el garaje

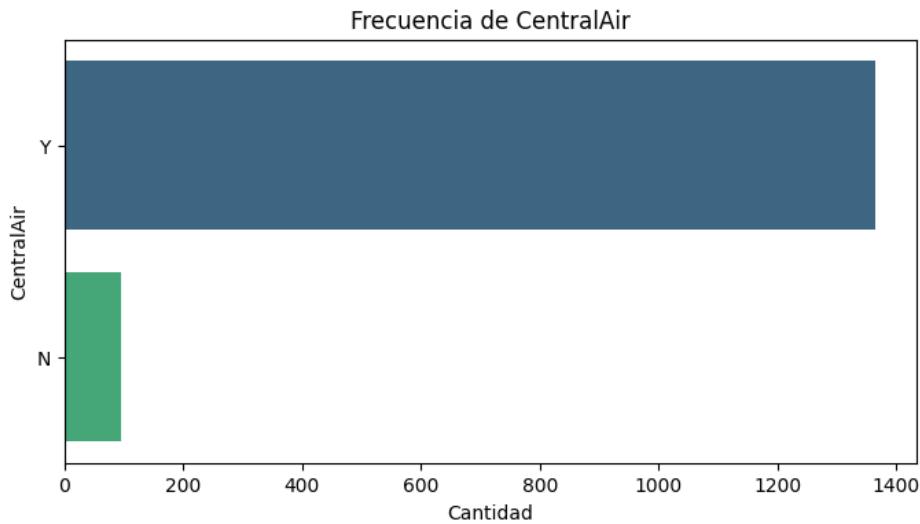




La mayoría de las casas tienen garajes adjuntos, seguidos por garajes separados y sin garaje. En cuanto al acabado del garaje, predominan los garajes sin acabado o con acabado de calidad estándar. La calidad y condición del garaje tienden a ser promedio, con pocos casos en los extremos. Estos patrones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

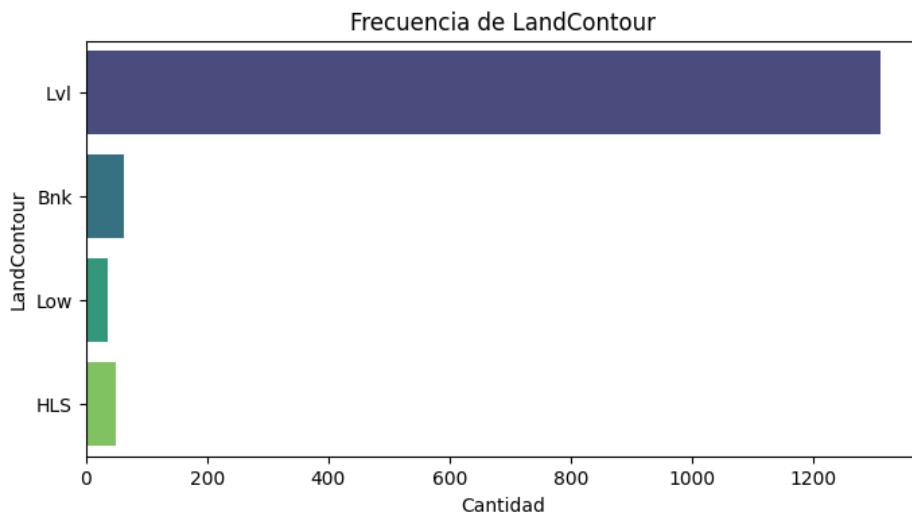
Variables relacionadas con calefacción y electricidad

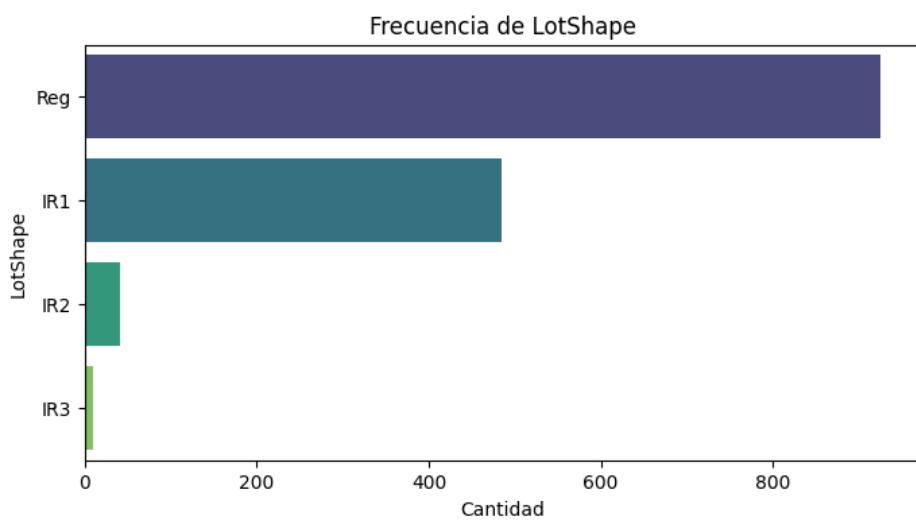
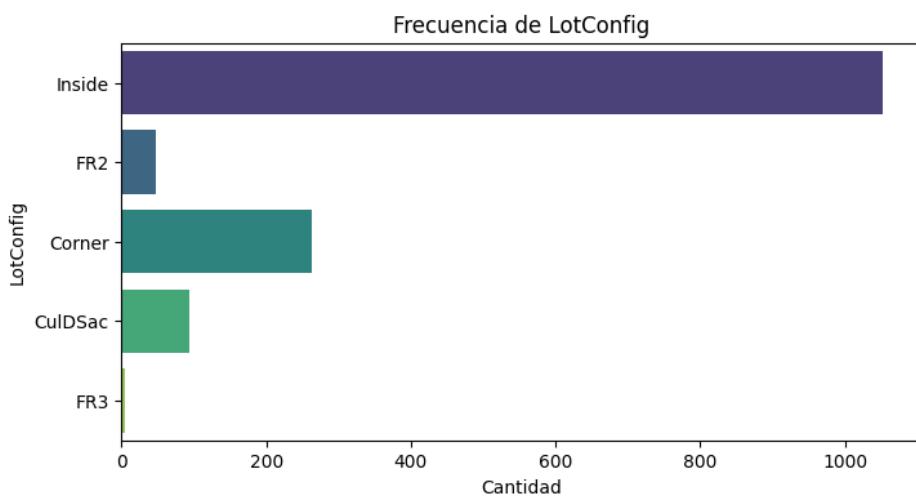
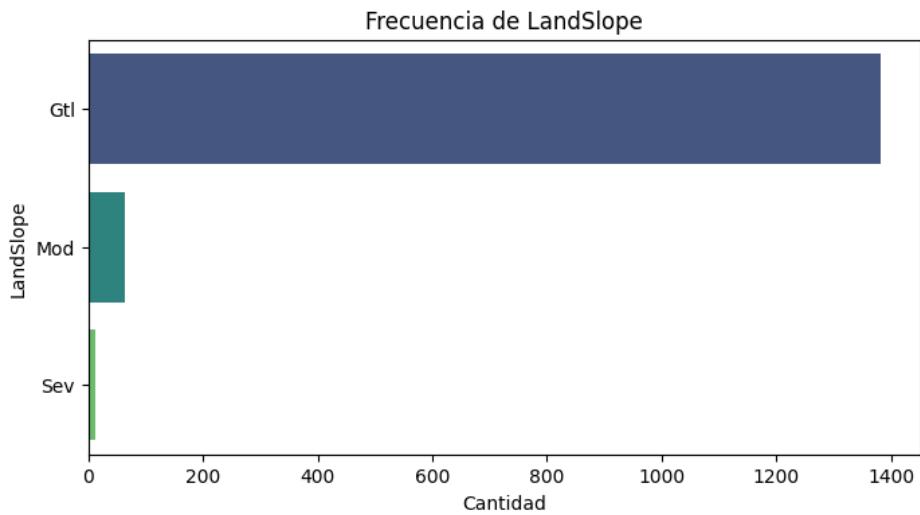




La mayoría de las casas tienen calefacción estándar (GasA) y calidad promedio (TA). La electricidad es principalmente SBrkr, con algunos casos de FuseA y FuseF. La mayoría de las casas tienen aire acondicionado central, lo que sugiere un nivel de comodidad y eficiencia energética estándar en la mayoría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

Variables relacionadas con la ubicación del terreno

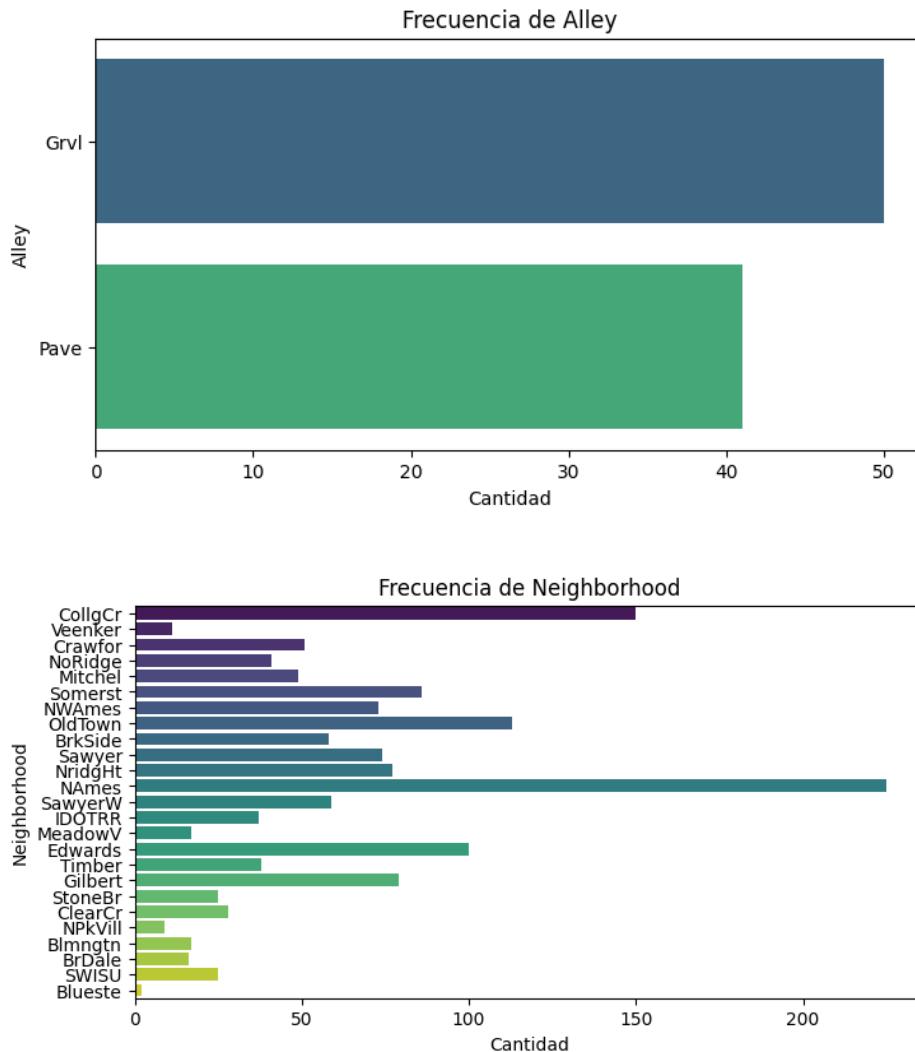


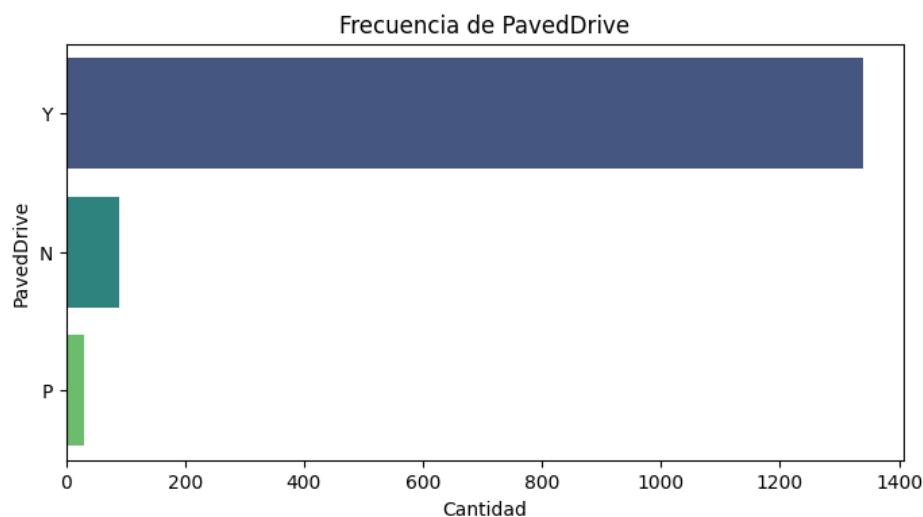
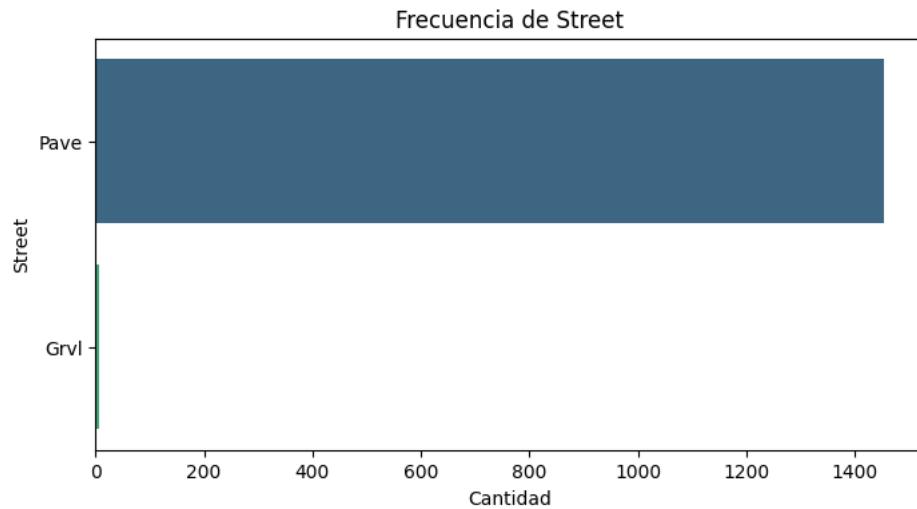


La mayoría de las propiedades tienen terrenos planos o ligeramente inclinados, con configuraciones de lote

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.

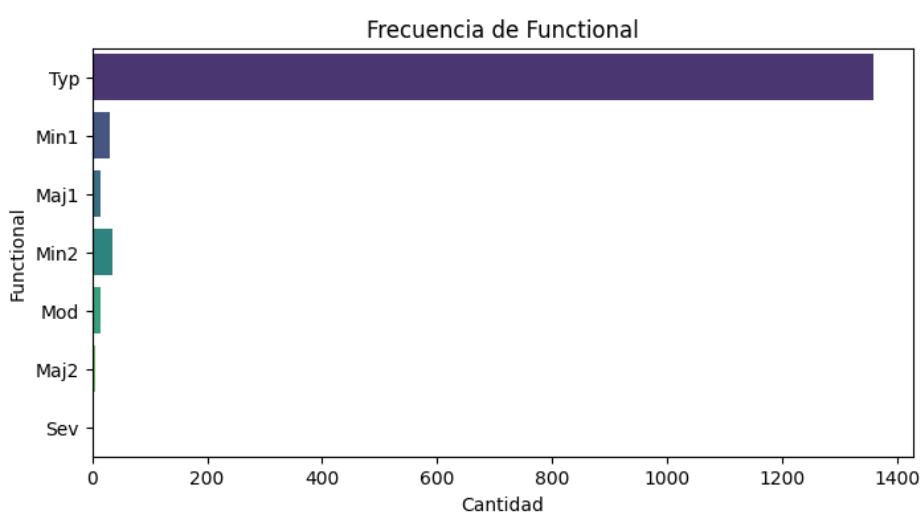
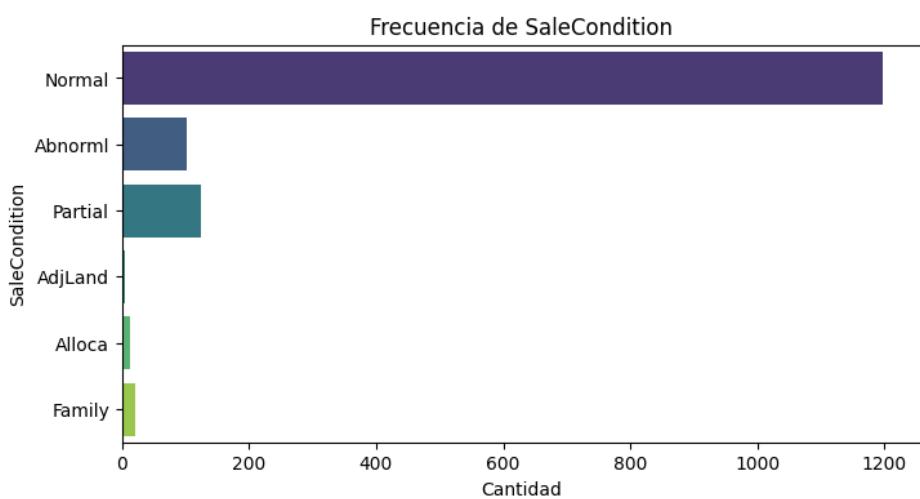
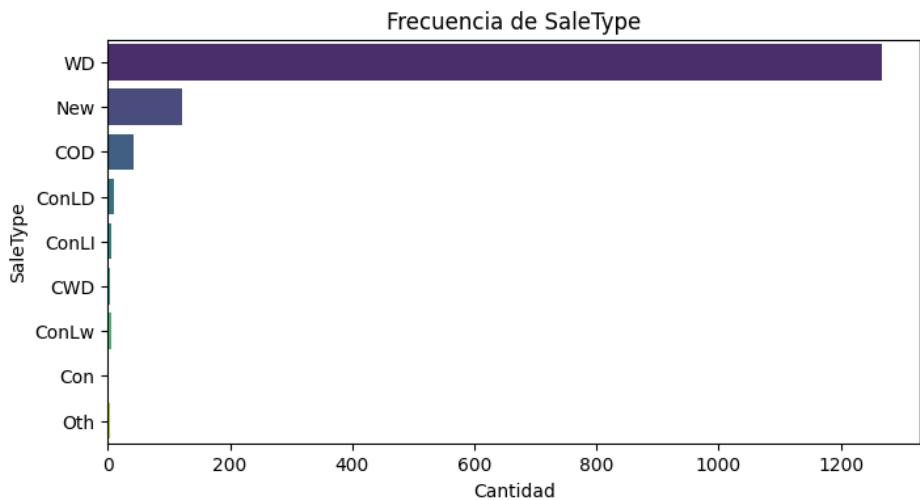
Variables relacionadas con vecindario y accesibilidad

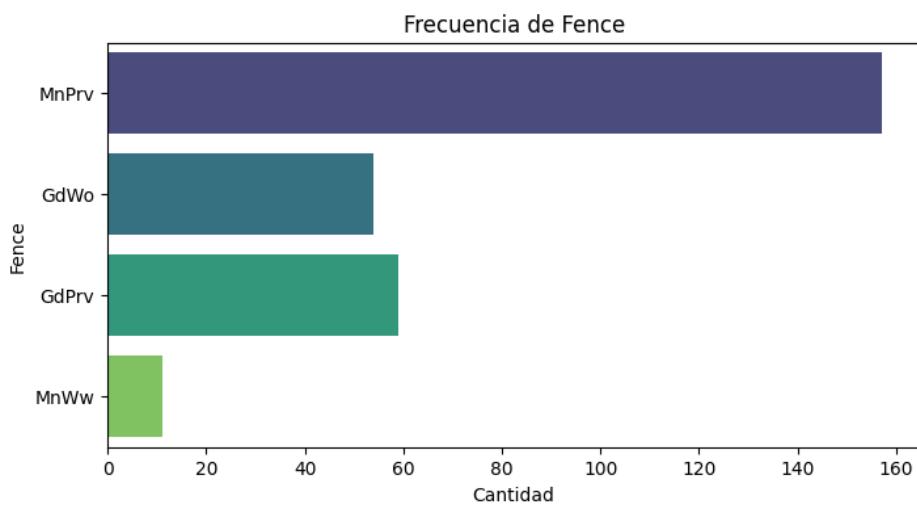
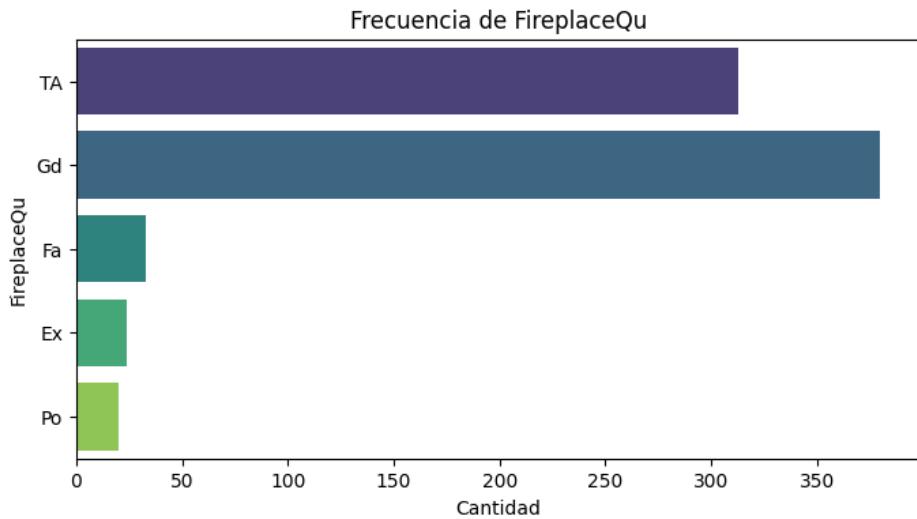




La mayoría de las propiedades tienen acceso por calle pavimentada y no tienen acceso a callejón. Los vecindarios más comunes son NAmes, CollgCr y OldTown, lo que sugiere una concentración en áreas urbanas o suburbanas. La mayoría de las propiedades tienen acceso pavimentado, lo que indica una buena accesibilidad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

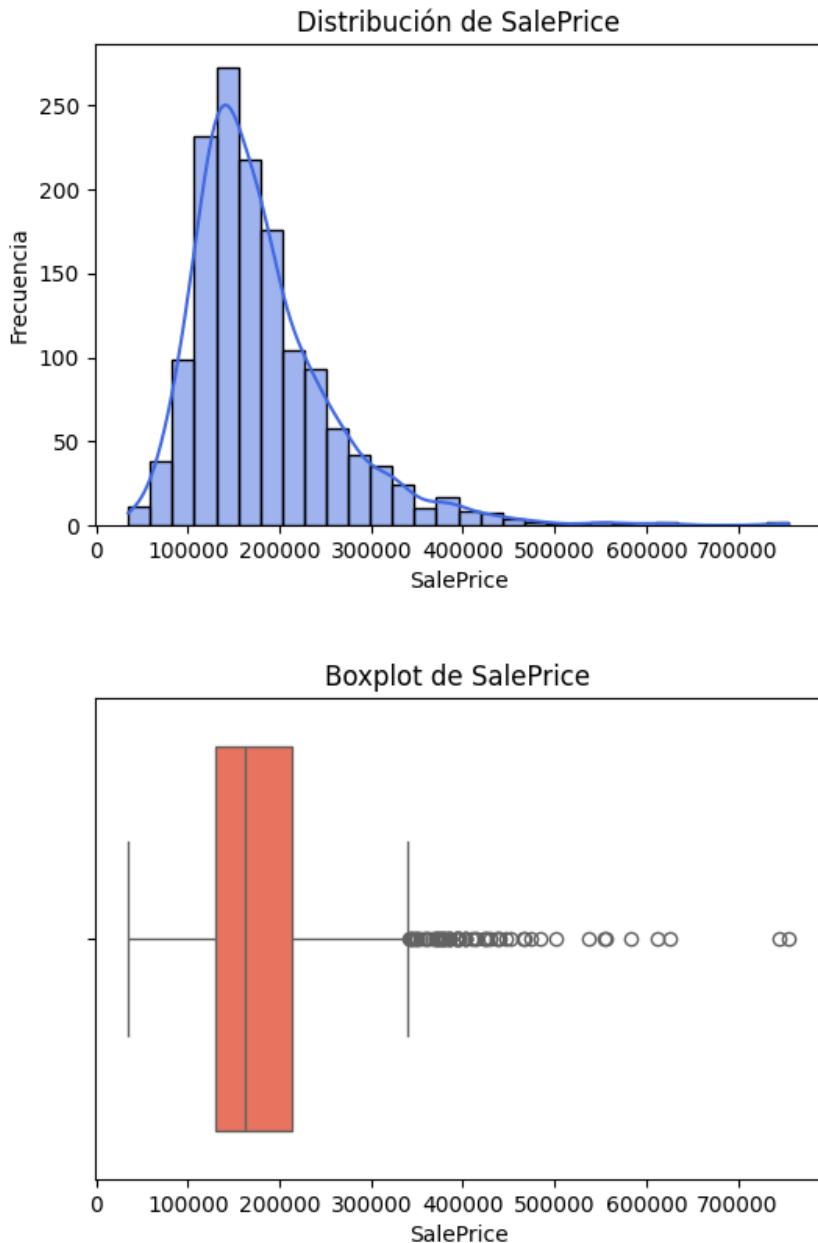
Variables relacionadas con seguridad y condiciones





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.

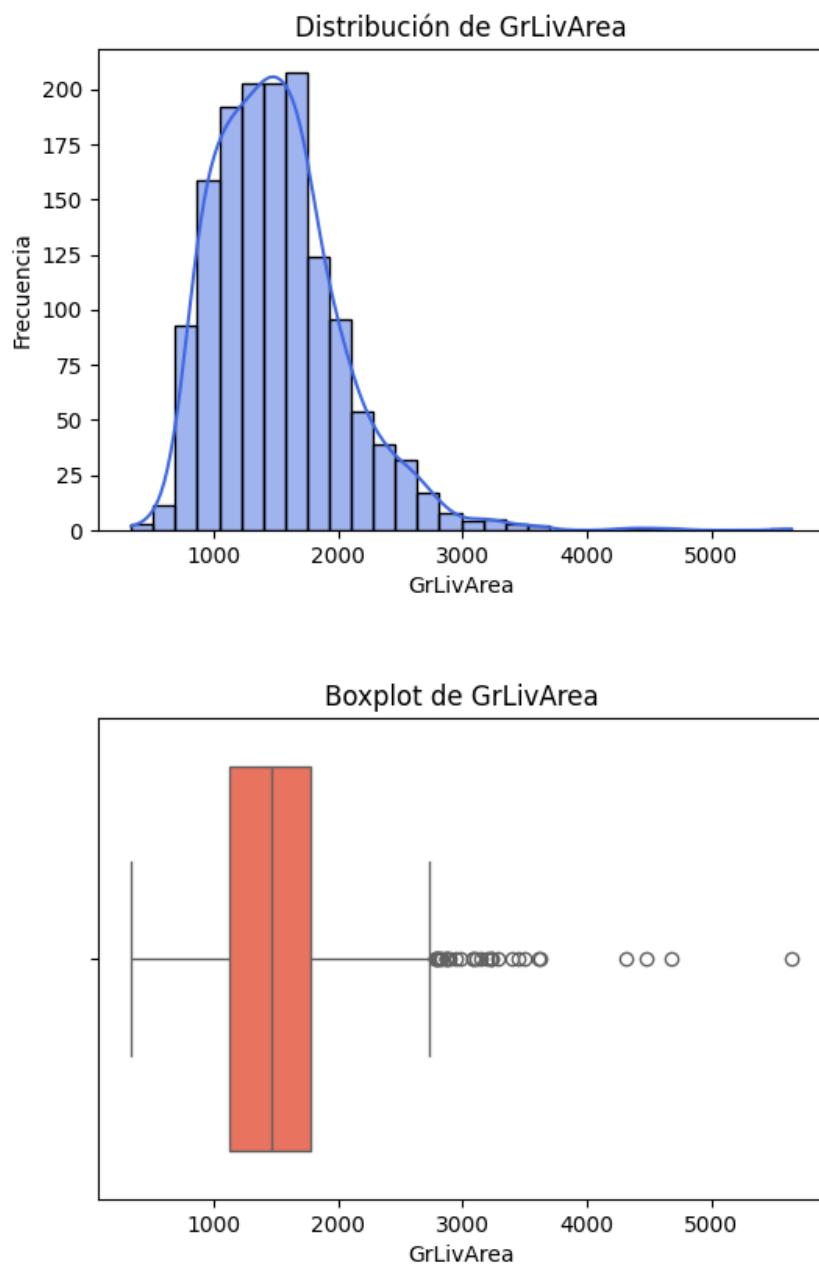
Visualización de Variables Numéricas

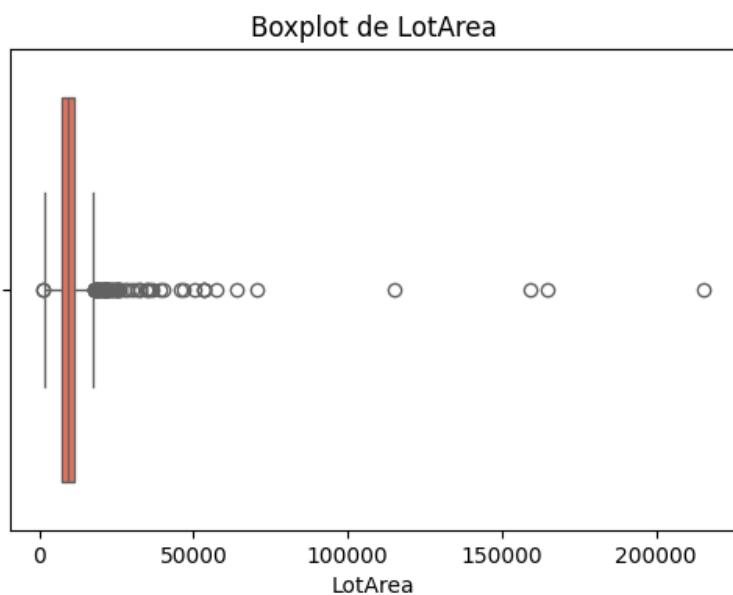
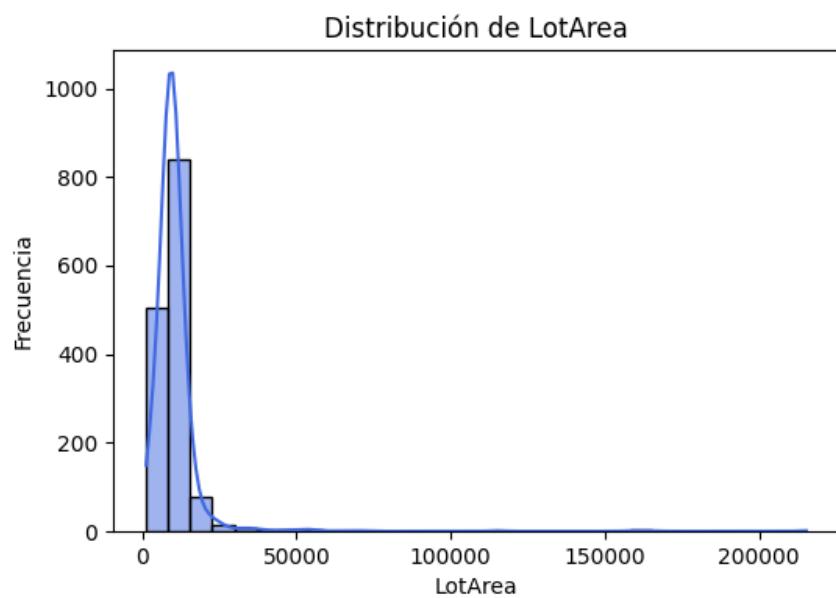


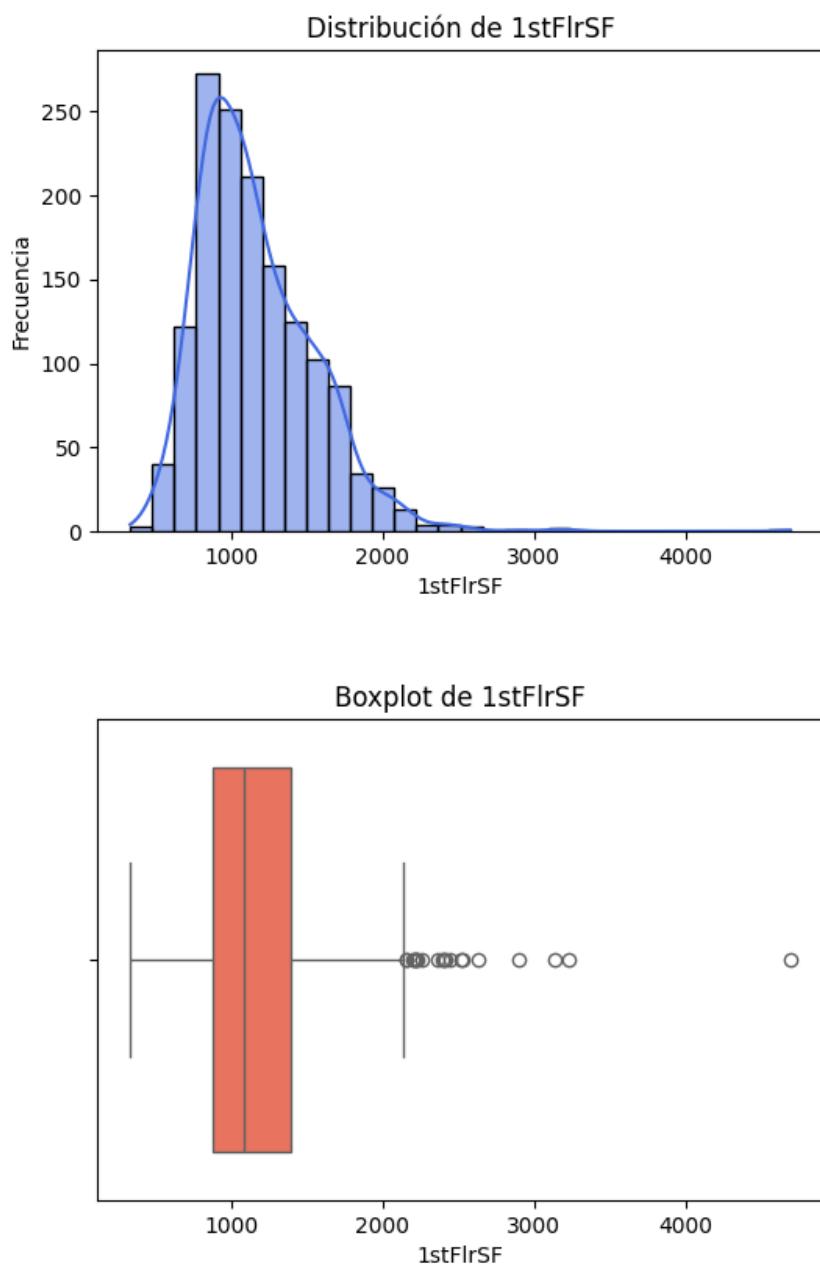
Siendo esta la variable objetivo, observamos una distribución sesgada a la derecha. Esta asimetría puede afectar métodos estadísticos que asumen distribuciones normales. Observamos outliers en la parte superior de la distribución, lo que sugiere la presencia de propiedades muy caras que pueden afectar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

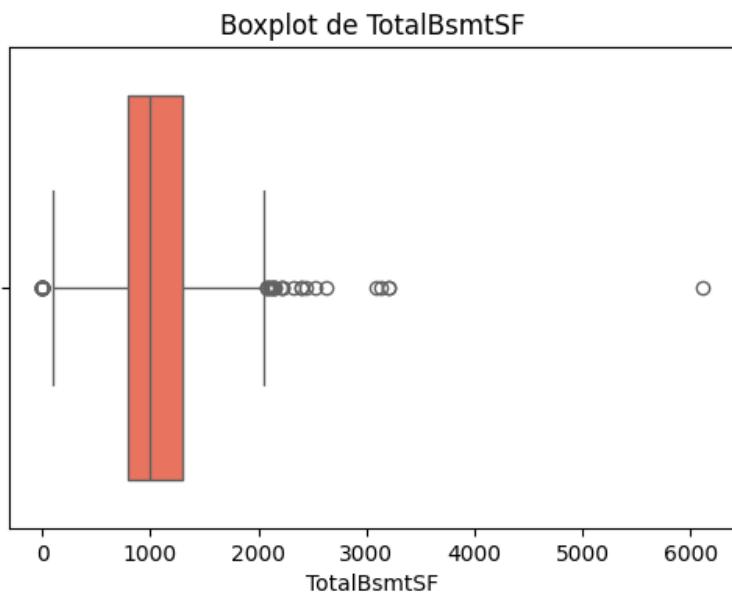
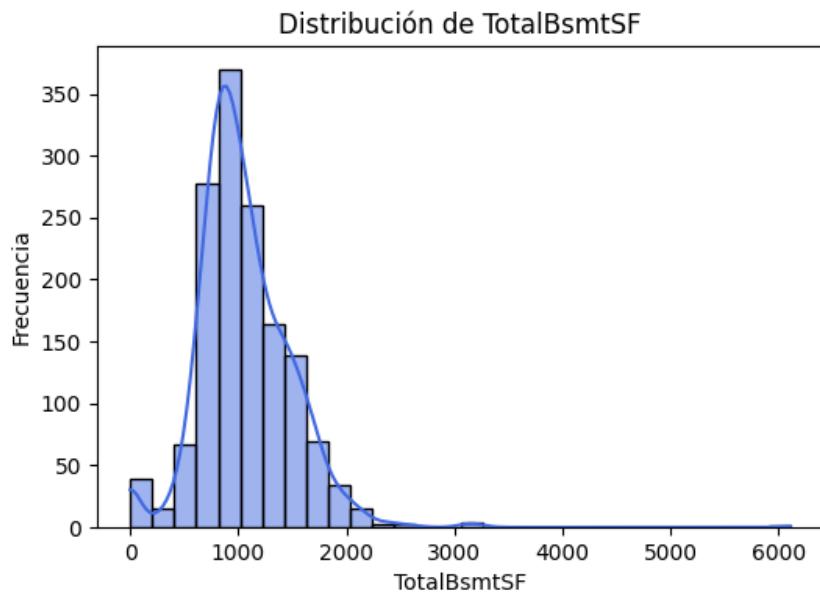
- **Boxplot:** Se aprecia que la mayoría de los precios se concentran en un rango intercuartílico entre 130,000 y 210,000 dólares, pero existen varios puntos extremos en la cola superior. Esto indica la presencia de propiedades con precios significativamente más altos.
- **Histograma con curva de densidad:** 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





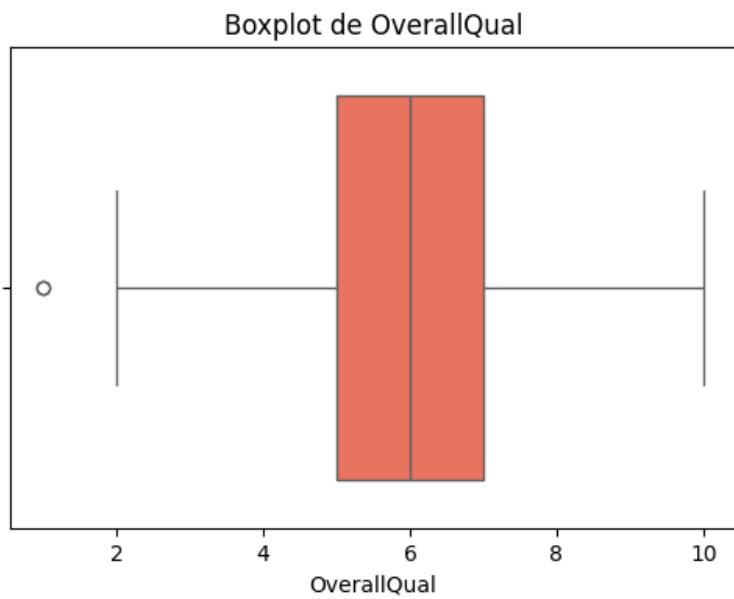
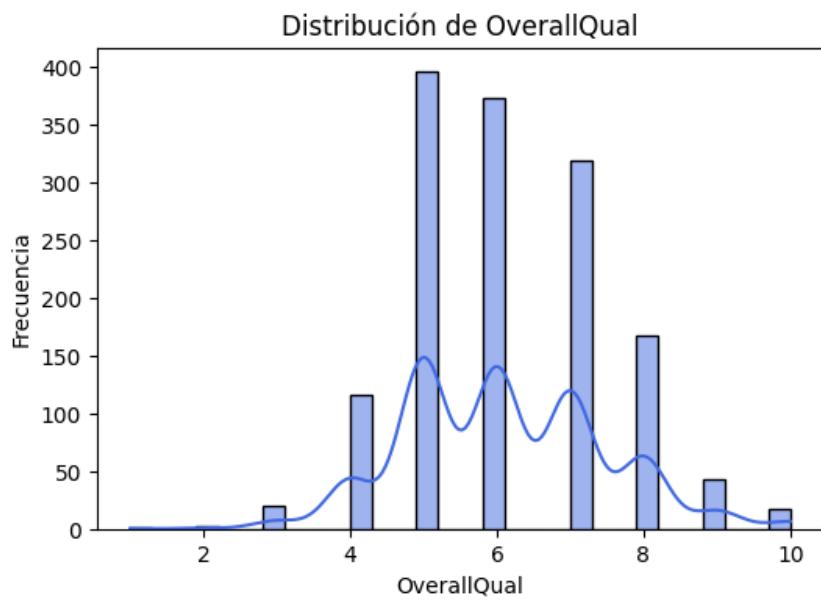




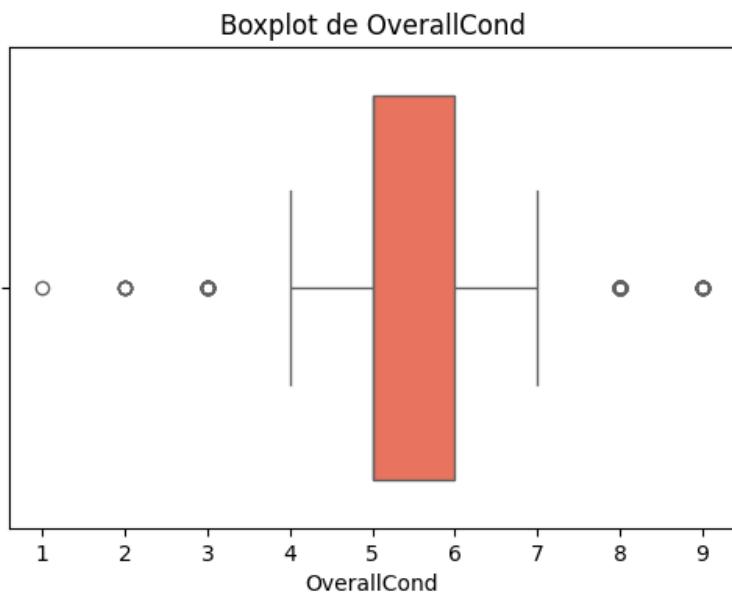
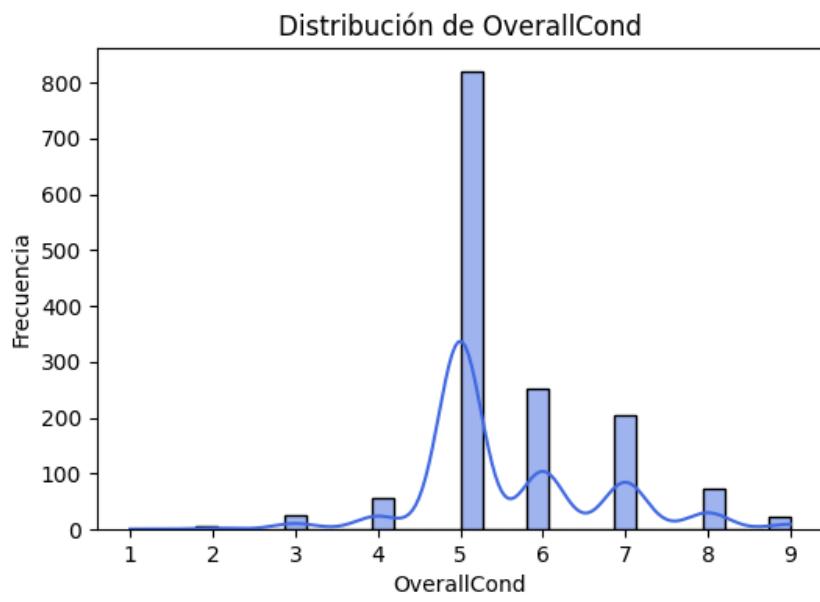
GrLivArea, LotArea, X1stFlrSF y TotalBsmtSF:

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.

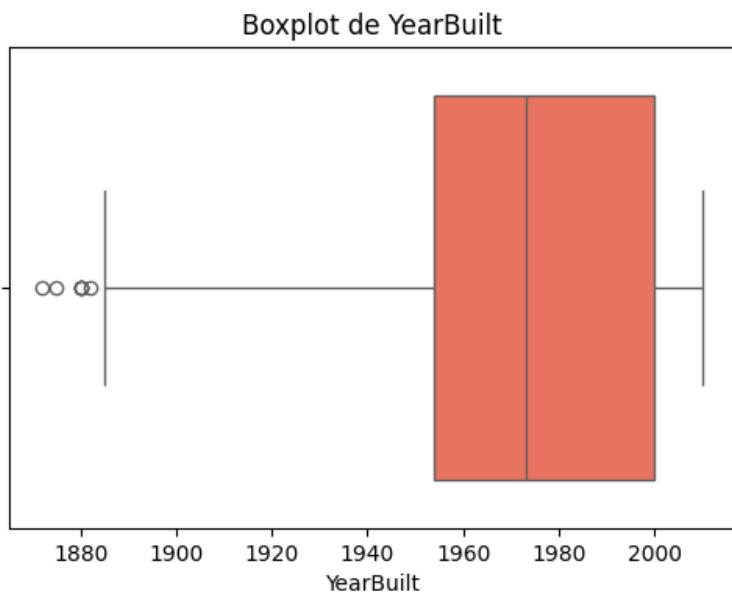
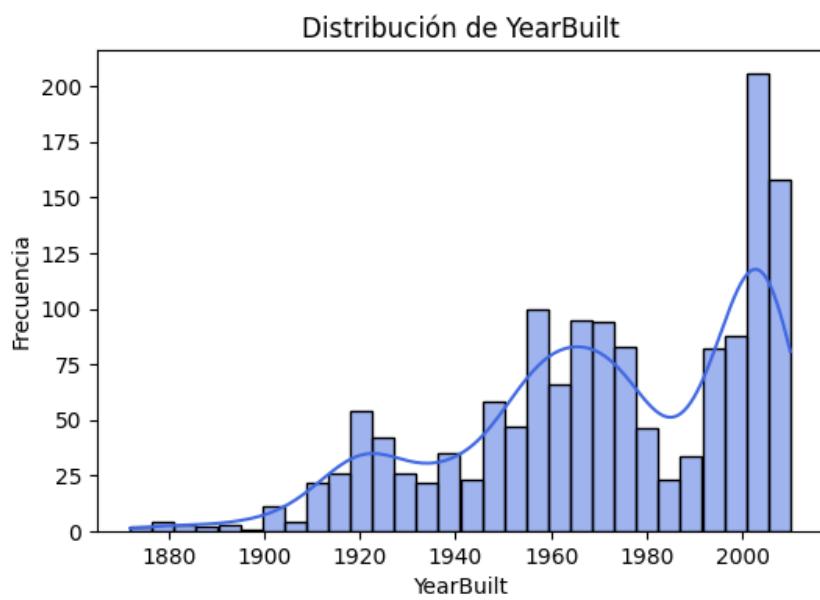
Variables Numéricas Relacionadas con Calidad y Años



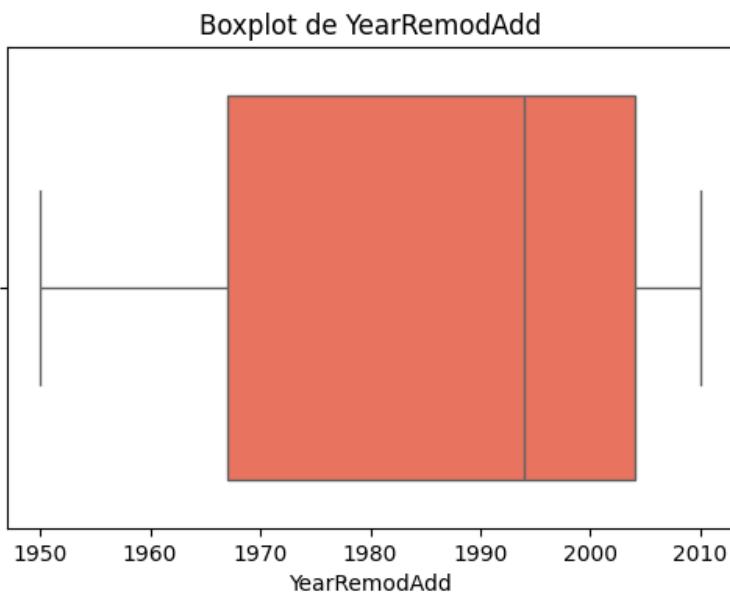
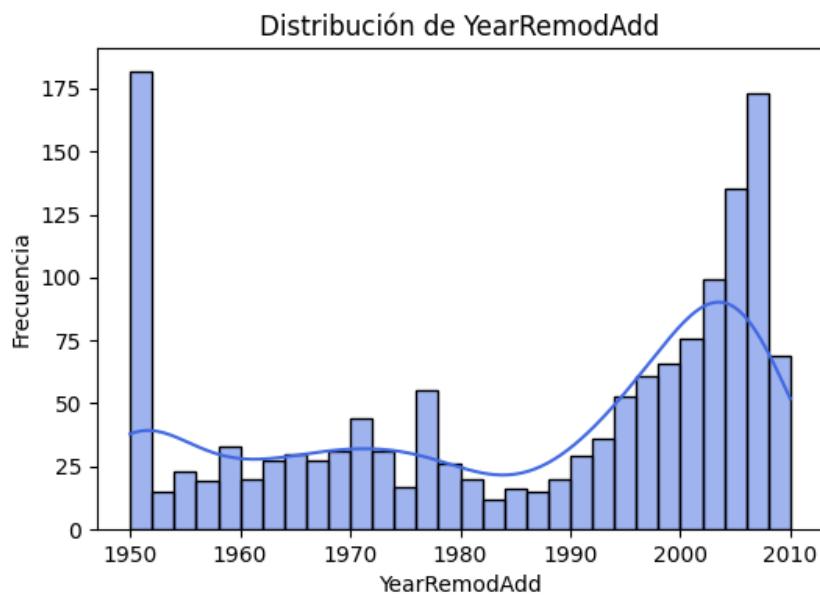
- Valores entre 1 y 10.
- Mayoría entre 5 y 7.
- Pico alrededor de 5-6.
- Pocos valores en los extremos.



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

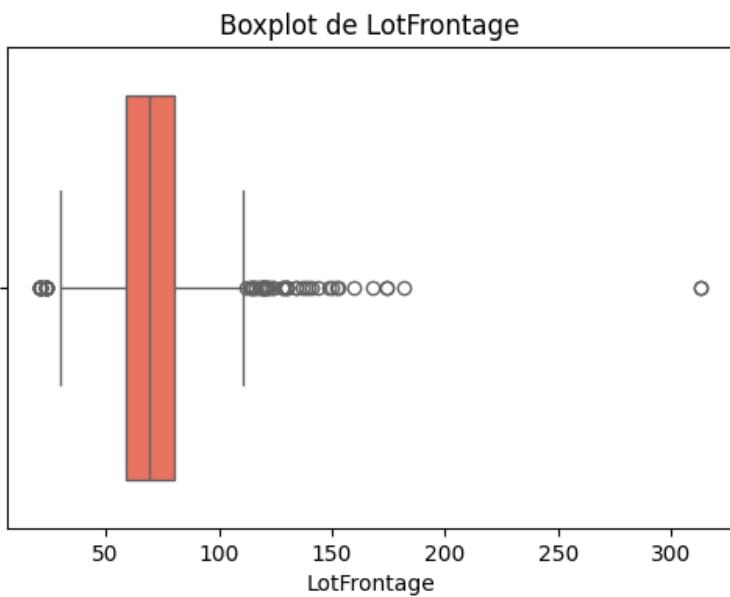
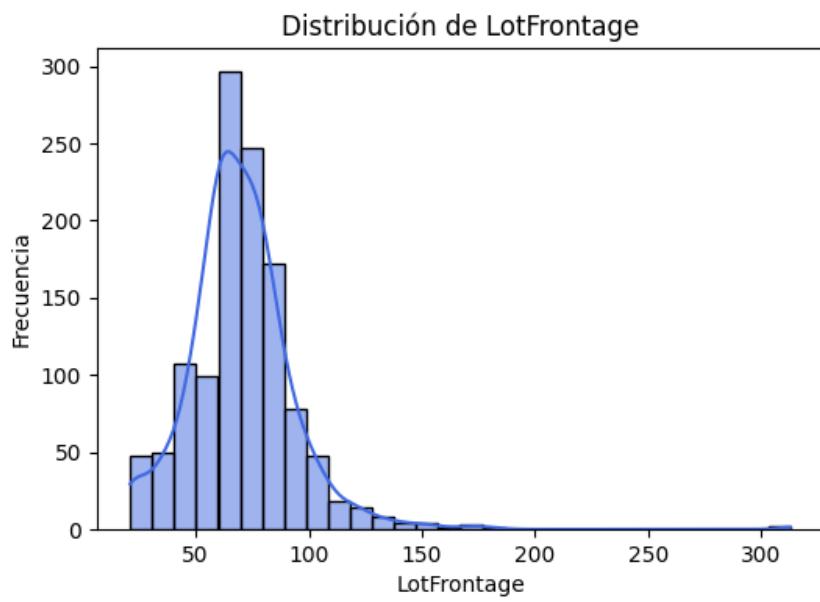


- Rango amplio (1870–2010).
- Incremento progresivo hasta 2000.
- Concentración alta en décadas recientes.
- Boxplot concentrado en 1950–2000.

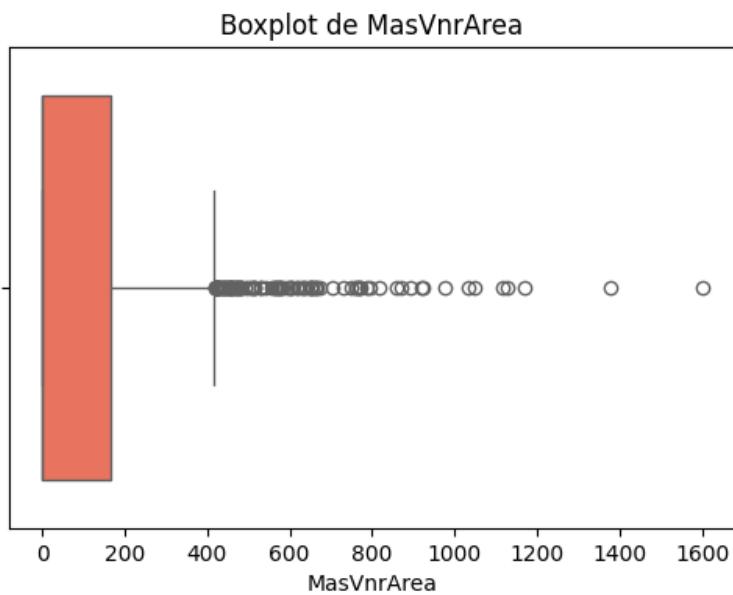
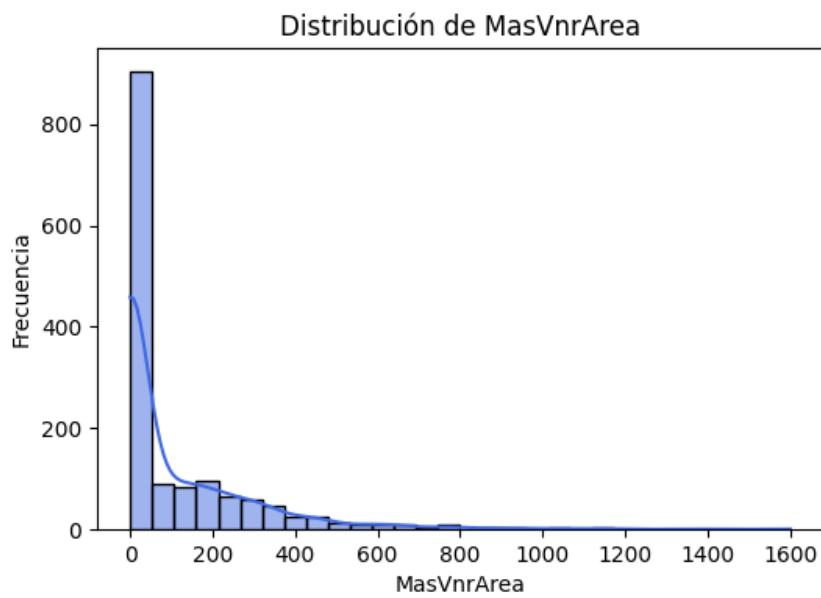


- Rango 1950–2010.
- Mayor actividad de remodelación cerca de 1990 y 2010, 1950 presenta remodelaciones altas.
- Boxplot abarca 1960–2000.
- Pocos valores anteriores a 1960.

Variables Numéricas Relacionadas con Áreas y Calidad

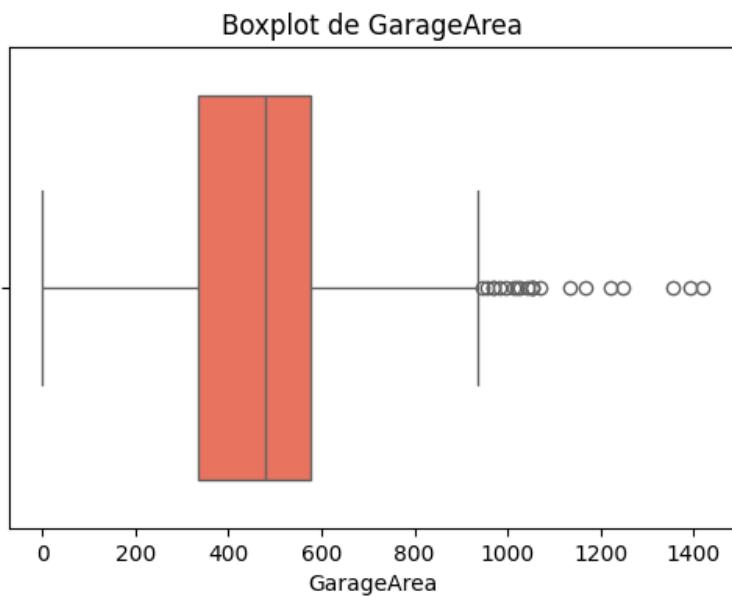
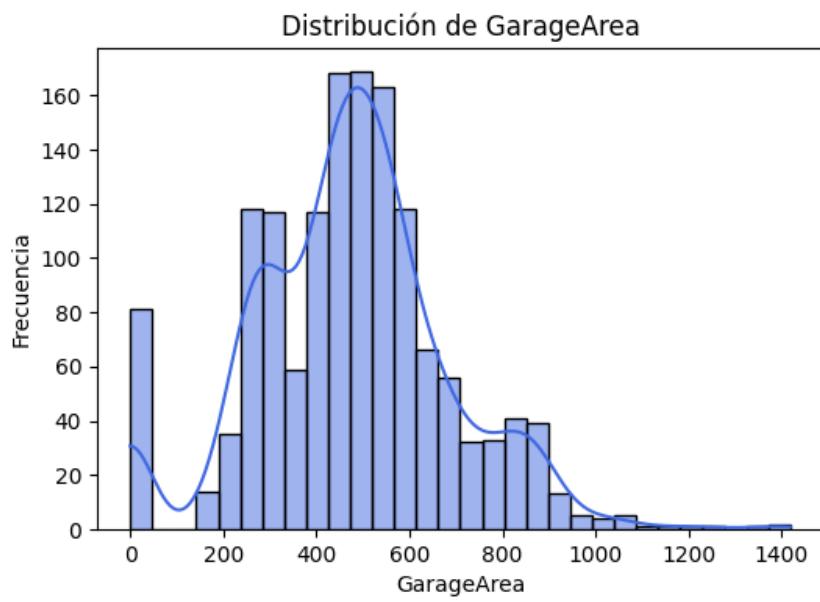


- Pico cercano a 60-70.
- Muchos valores faltantes.
- Cola derecha larga, outliers por encima de 150.

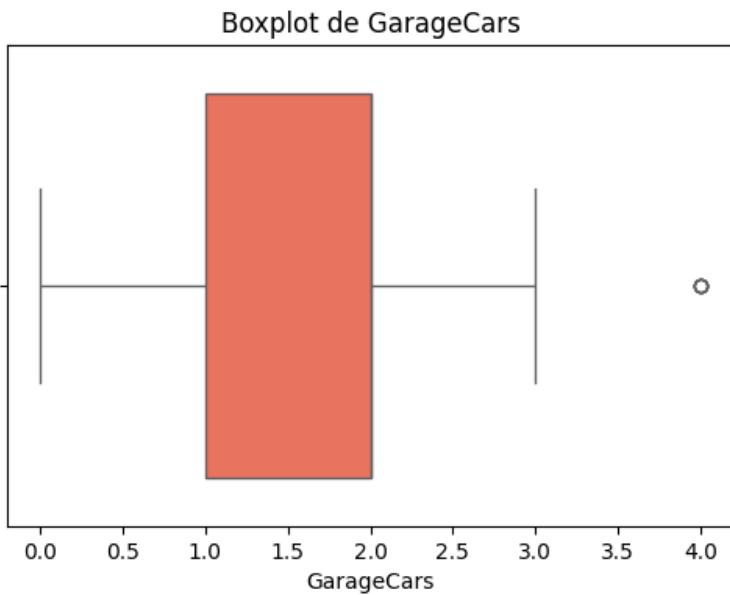
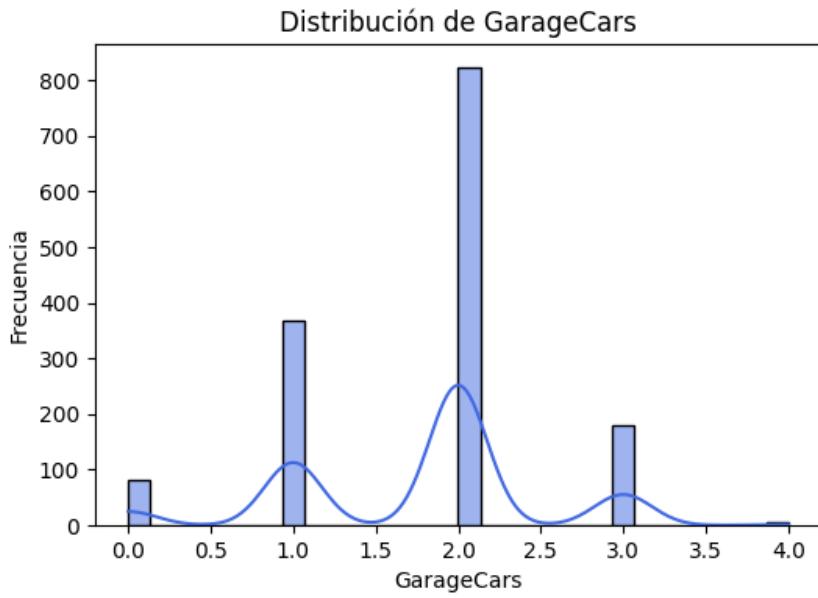


- Mayoría en 0 (sin acabado de mampostería).
- Fuerte sesgo a la derecha.
- Outliers hasta 1600.

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|
```

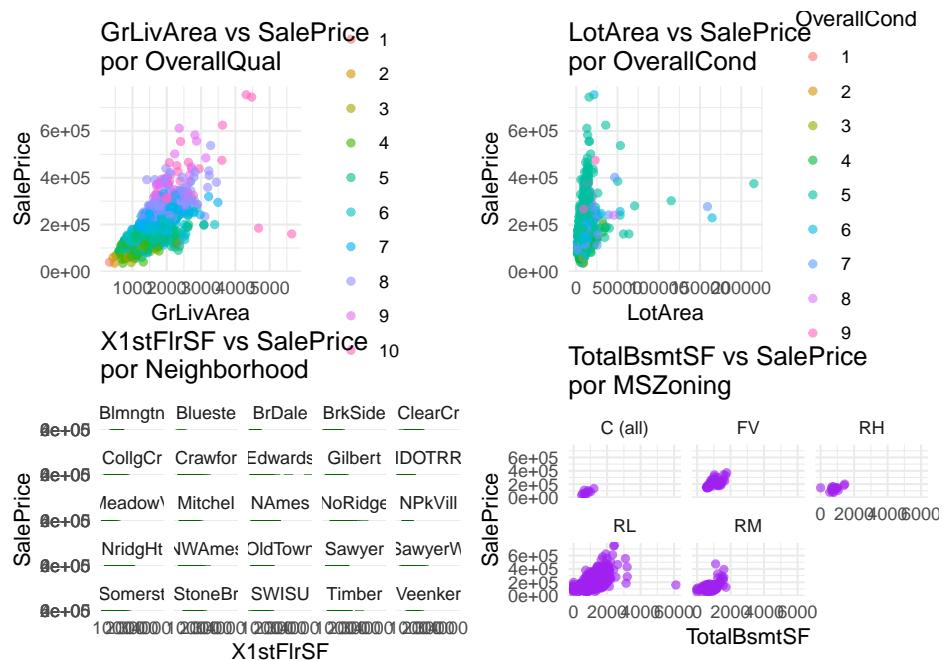
Las pruebas de normalidad en todos los grupos de variables arrojan p-valores extremadamente bajos ($p < 2.2e-16$ en la mayoría de los casos), lo que indica que ninguna de estas variables sigue una distribución normal según los test de Shapiro-Wilk, Anderson-Darling, Kolmogorov-Smirnov y Lilliefors. Esto 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.

Preguntas Exploratorias

A partir de esta exploración inicial, se identificaron patrones y características clave en las variables categóricas y numéricicas. 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)?



- **GrLivArea vs SalePrice**

En el diagrama de dispersión se aprecia una **tendencia claramente positiva**: a mayor superficie habitable (GrLivArea), mayor tiende a ser el precio de venta.

- Se observa que los puntos con OverallQual más alto se concentran en la parte superior de la nube de puntos, indicando que casas con más área y mejor calidad se venden a precios notablemente superiores.

- **LotArea vs SalePrice**

Existe también una relación positiva, pero es más dispersa que GrLivArea. Se ven valores muy altos de LotArea que no siempre llevan precios igual de altos, lo cual sugiere que el tamaño del lote por sí solo no determina el precio de forma tan directa como el área habitable.

- Al superponer la variable OverallCond , se aprecia que las viviendas con mejor condición se ubican en rangos de precio más elevados, aun con lotes de tamaño similar.

- **X1stFlrSF vs SalePrice (por Neighborhood)**

En la gráfica se percibe nuevamente una relación creciente entre la superficie del primer piso y el precio.

- Sin embargo, Neighborhood introduce diferencias: barrios de mayor nivel muestran precios más altos incluso para valores de X1stFlrSF relativamente moderados, mientras que en barrios de menor nivel se requieren superficies mucho mayores para alcanzar precios similares.

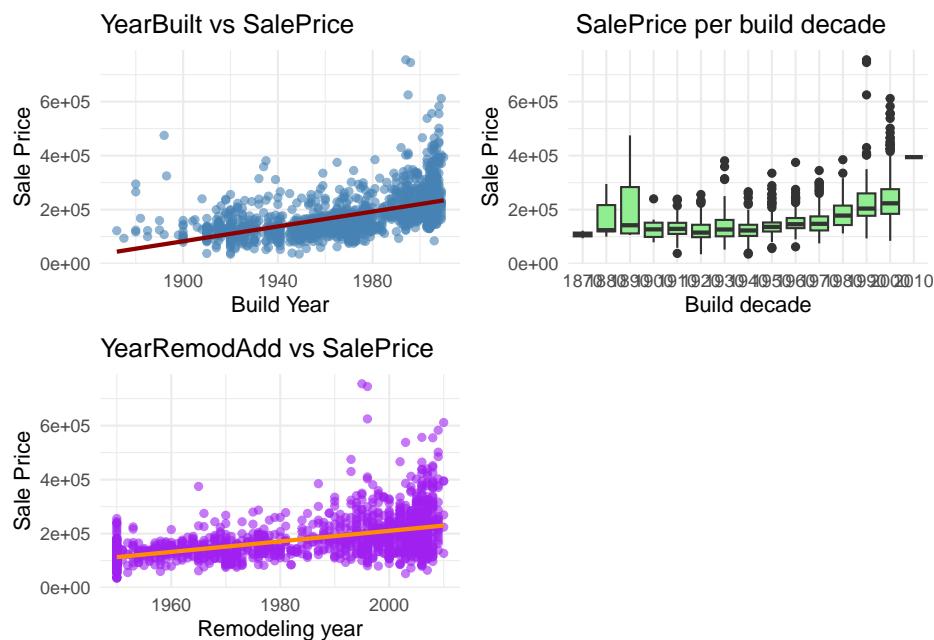
- **TotalBsmtSF vs SalePrice (por MSZoning)**

De igual modo, se ve correlación positiva entre el tamaño del sótano y el precio.

- La zonificación segmenta el mercado: en zonas residenciales de baja densidad los precios suelen ser más elevados que en zonas multifamiliares, a igualdad de TotalBsmtSF.

Las áreas de la vivienda guardan una relación positiva con SalePrice. Esa relación se modula por la calidad/condición de la vivienda y por la ubicación .

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?



- **YearBuilt vs SalePrice**

El gráfico de dispersión con una línea de tendencia sugiere que las casas más nuevas suelen tener precios promedio más altos.

- Sin embargo, hay puntos antiguos (antes de 1940) con precios elevados, lo cual indica que algunas casas históricas o muy bien conservadas también pueden alcanzar precios altos, probablemente por estar en barrios deseados o haber sido remodeladas.

- **SalePrice por década de construcción**

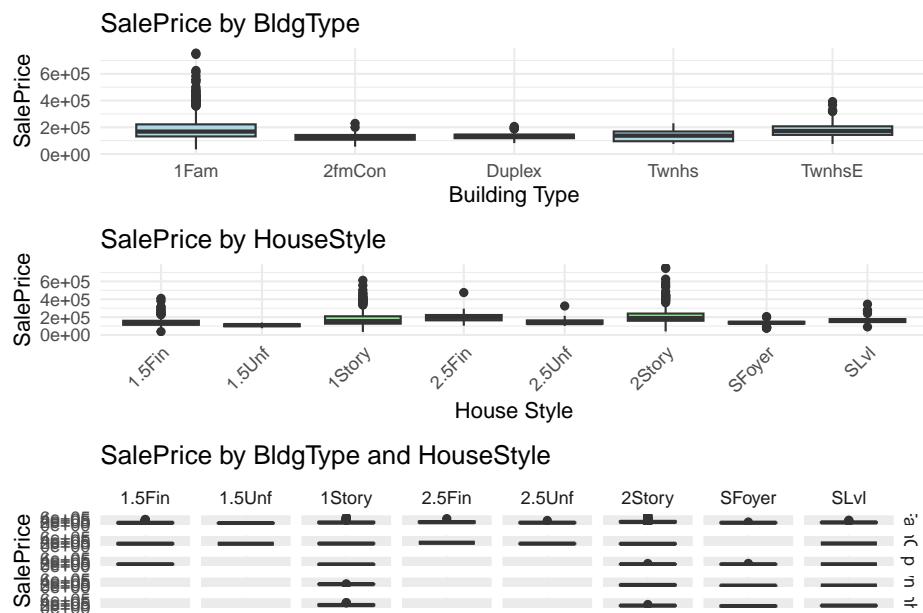
En la gráfica de cajas (boxplot) por década, se ve un incremento gradual en el precio mediano con cada década más reciente, aunque hay traslapos entre décadas y algunos outliers altos en décadas anteriores.

- **YearRemodAdd vs SalePrice**

El diagrama de dispersión muestra una tendencia similar: las casas con remodelaciones más recientes suelen presentar precios mayores. Se evidencia que la remodelación eleva el valor de propiedades antiguas.

Las viviendas construidas o renovadas más recientemente tienden a tener precios mayores, aunque propiedades muy antiguas y con alto mantenimiento pueden equipararse a precios de casas más nuevas.

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?



- **SalePrice por BldgType**

El boxplot muestra que 1Fam suele tener la mediana de precios más alta. Otras tipologías presentan mediana y dispersión de precios algo menores.

- **SalePrice por HouseStyle**

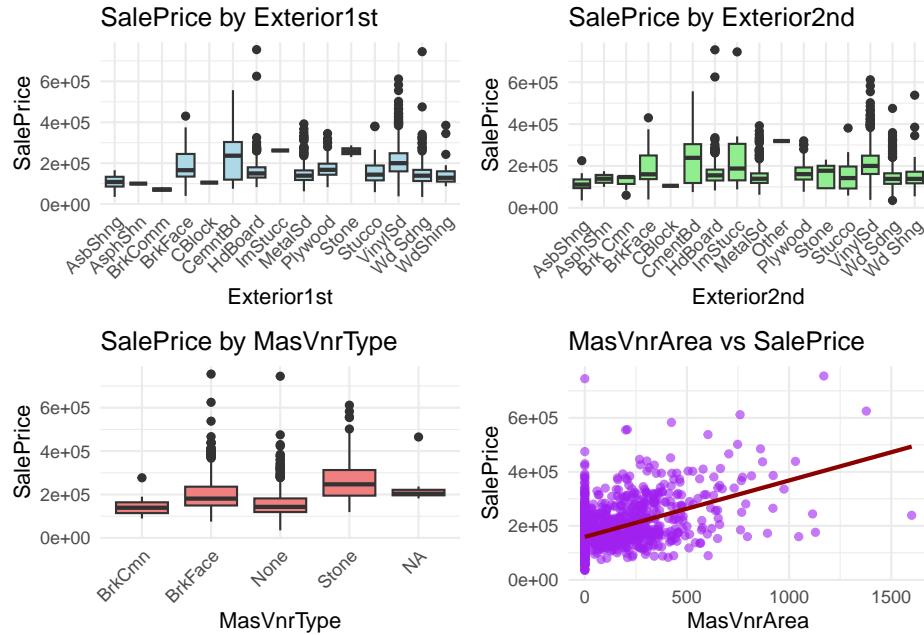
Se ven estilos como 1Story, 1.5Fin, 2Story, etc.

- Generalmente, 2Story presenta una mediana algo más elevada que 1Story. Estilos con 1.5 pisos tienen una mediana menor, aunque con bastante dispersión.

- **SalePrice por BldgType y HouseStyle (combinados)**

Se confirma que las unifamiliares de 2 pisos tienden a precios más altos. Los demás estilos y tipos presentan menor valor medio, aunque con outliers en todos los grupos.

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?



- **SalePrice por Exterior1st y Exterior2nd**

Los boxplots muestran diferencias entre materiales: algunos como “Stone” o “Brick” tienen medianas de precio más altas. Acabados más económicos tienden a mediana inferior.

- **SalePrice por MasVnrType**

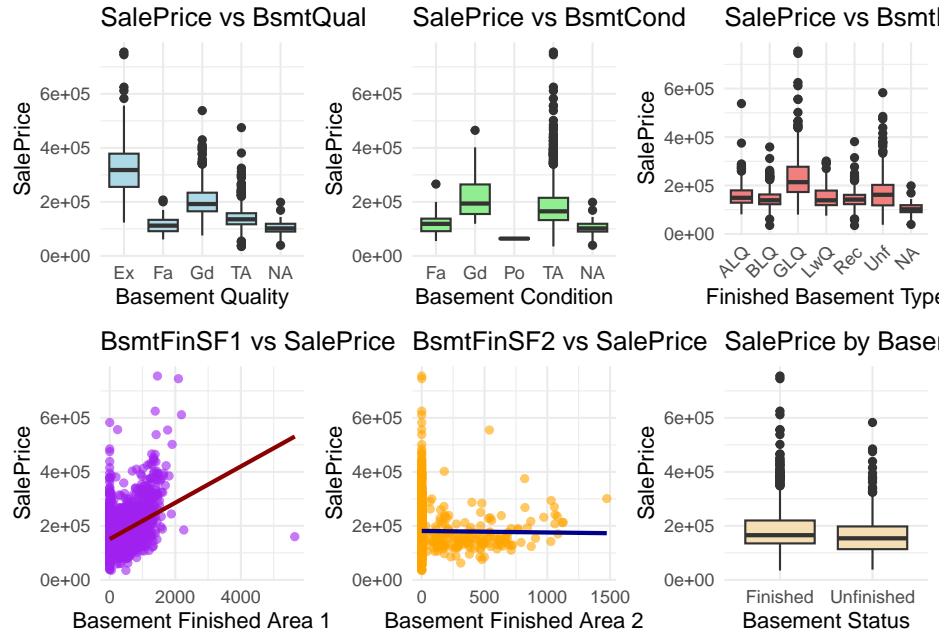
Se ven categorías como “BrkFace”, “Stone”, “None”. “Stone” y “BrkFace” suelen asociarse a valores más altos que “None”.

- **MasVnrArea vs SalePrice**

El diagrama de dispersión y la línea de tendencia reflejan una correlación positiva: cuanto mayor es el área de recubrimiento en mampostería (ladrillo, piedra, etc.), mayor suele ser el precio de venta.

Los acabados exteriores y la presencia de mampostería se asocian con precios más altos, indicando que la calidad y estética exterior añade valor.

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?



- **SalePrice vs BsmtQual, BsmtCond, BsmtFinType**

Los boxplots muestran que calidades altas (Ex, Gd) y condiciones buenas se asocian con precios medianos superiores. BsmtFinType (GLQ, ALQ) —acabados de mayor nivel— también suben el precio respecto a un sótano sin terminar (Unf).

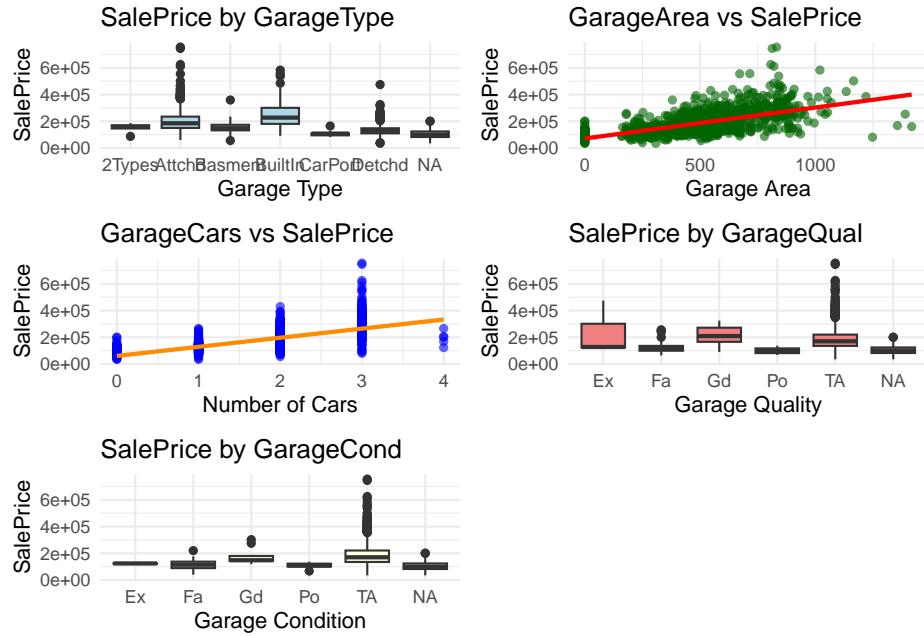
- **BsmtFinSF1 y BsmtFinSF2 vs SalePrice**

Los diagramas de dispersión evidencian una relación positiva: a más metros cuadrados terminados en el sótano, mayor precio.

- BsmtFinSF1 suele tener un impacto más claro que BsmtFinSF2, probablemente porque es el área de acabado principal.

Un sótano bien calificado y con superficies terminadas aumenta el espacio habitable y, por ende, el valor de la vivienda.

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?



- **SalePrice por GarageType**

Viviendas con garajes “Attached” o “BuiltIn” suelen tener precios medianos mayores que aquellas con “CarPort” o “NA” .

- **GarageArea vs SalePrice**

Se ve una **correlación positiva**: un garaje más grande tiende a asociarse con precios más altos.

- **GarageCars vs SalePrice**

La línea que conecta la media según el número de coches sube de forma notable: garajes de 2-3 plazas suelen estar en rangos de precio más elevados que los de 1 plaza.

- **SalePrice por GarageQual y GarageCond**

Garajes con calidades superiores (Ex, Gd) presentan precios medianos notablemente más altos. Condiciones regulares (TA) o pobres (Po) reducen la mediana.

Un garaje amplio, con capacidad suficiente y buena calidad incrementa el valor de la vivienda, confirmando su importancia en la percepción del comprador.

- **¿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?**

Table 7: Frequency of Alley

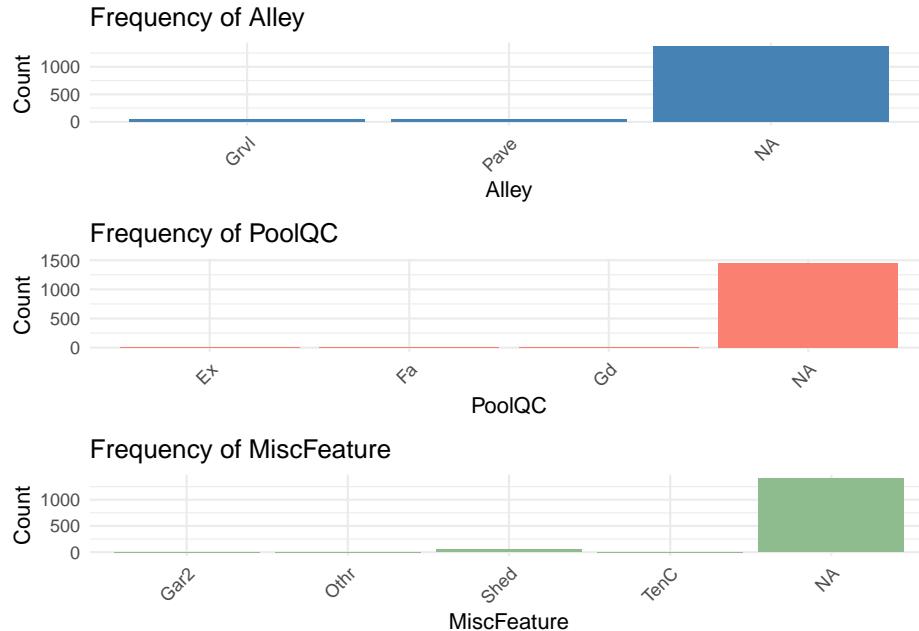
Alley	Count
Grvl	50
Pave	41
NA	1369

Table 8: Frequency of PoolQC

PoolQC	Count
Ex	2
Fa	2
Gd	3
NA	1453

Table 9: Frequency of MiscFeature

MiscFeature	Count
Gar2	2
Othr	2
Shed	49
TenC	1
NA	1406



- **Frecuencia de Alley**

El gráfico de barras muestra que la mayoría de los registros están en “NA”, y muy pocos tienen “Grvl” o “Pave”. Esto indica fuerte desequilibrio.

- **Frecuencia de PoolQC**

La gran mayoría también aparece como “NA”, y solo un puñado de viviendas tiene calificaciones de piscina (Ex, Gd, etc.). Claramente hay pocas casas con piscina.

- **Frecuencia de MiscFeature**

De nuevo, “NA” es dominante. Las categorías como “Shed”, “Tenc”, etc. son muy minoritarias.

Estas variables tienen muchos valores nulos o categorías con muy pocas observaciones, por lo que, para un análisis o modelado predictivo, probablemente se necesite agrupar, recodificar o descartar en ciertos casos.

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?

Table 10: Frequency of LotShape

LotShape	Count
IR1	484
IR2	41
IR3	10
Reg	925

Table 11: Frequency of LandContour

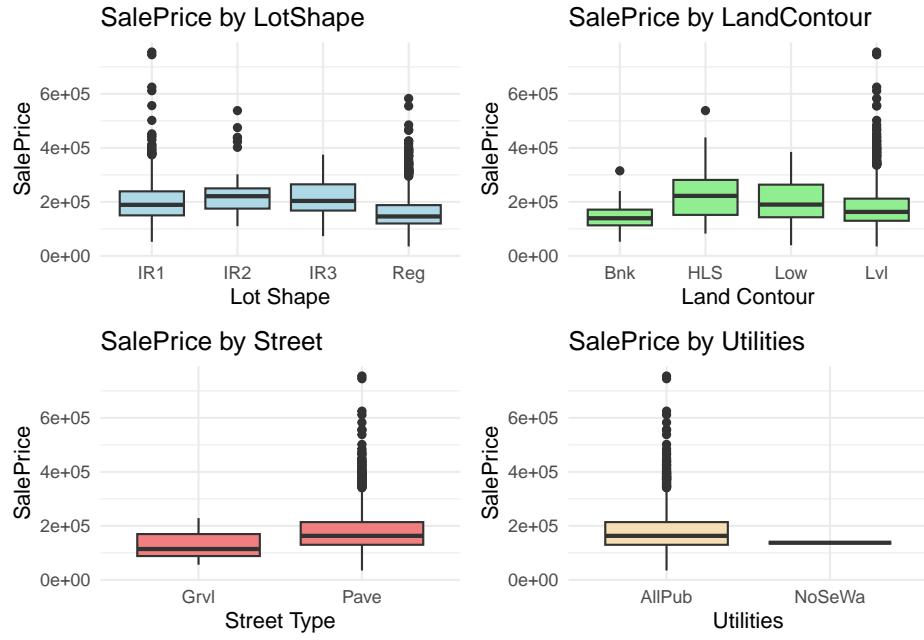
LandContour	Count
Bnk	63
HLS	50
Low	36
Lvl	1311

Table 12: Frequency of Street

Street	Count
Grvl	6
Pave	1454

Table 13: Frequency of Utilities

Utilities	Count
AllPub	1459
NoSeWa	1



- **SalePrice por LotShape**

Los boxplots muestran que lotes de forma regular (Reg) tienden a un precio mediano más alto, mientras que lotes muy irregulares (IR3) suelen tener precios más bajos.

- **SalePrice por LandContour**

Los terrenos “Lvl” (nivelados) muestran, en general, medianas más altas que “Bnk” o “HLS” (terrenos con pendientes). No obstante, se observan outliers en todos los grupos.

- **SalePrice por Street**

Calles pavimentadas (Pave) se asocian a precios más elevados que calles de grava (Grvl). La diferencia no es tan marcada como en otras variables, pero sí visible.

- **SalePrice por Utilities**

Tener todos los servicios públicos (AllPub) presenta una mediana superior frente a “NoSeWa”. La mayoría de propiedades se concentran en “AllPub”, con pocas en la otra categoría.

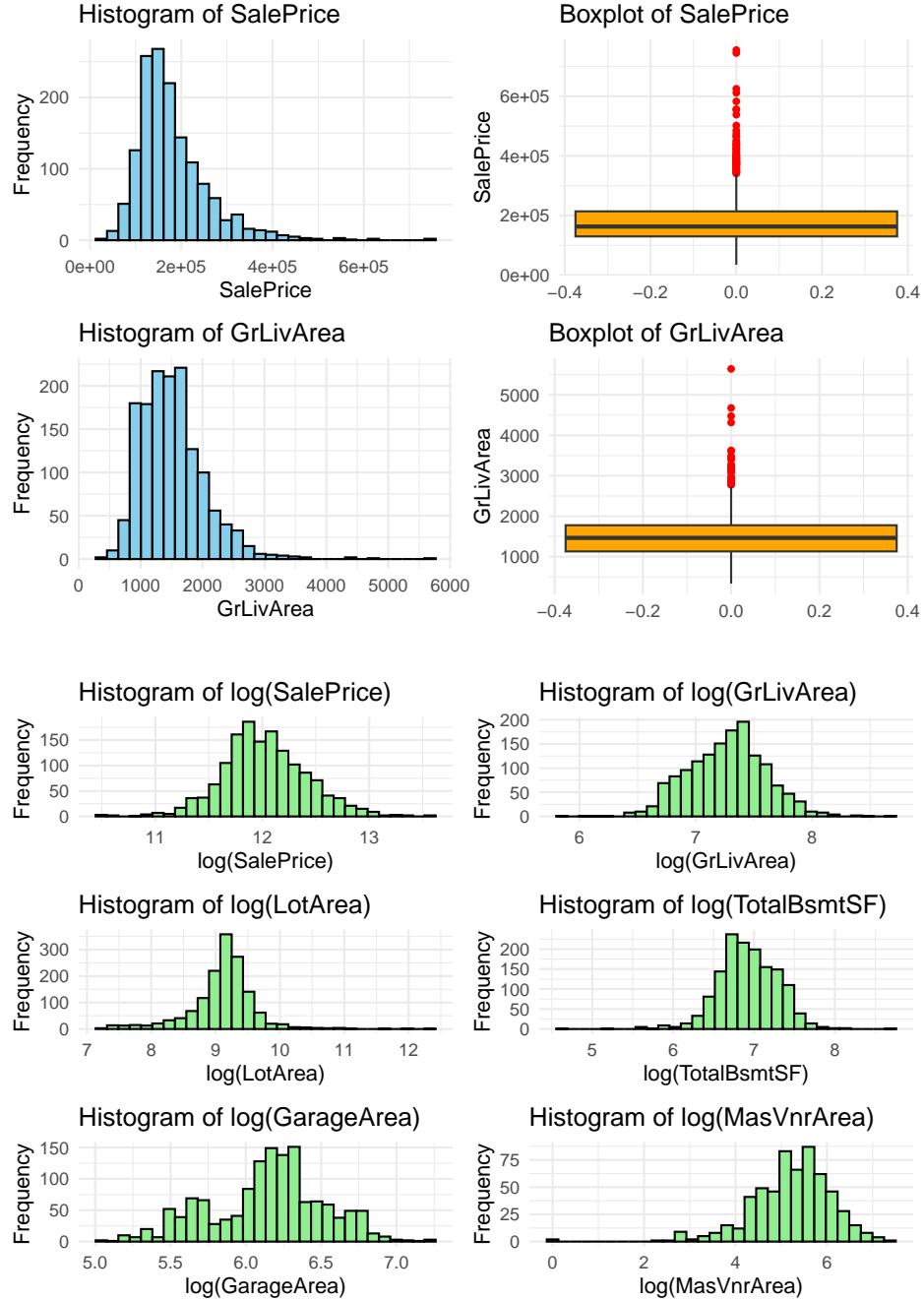
Aunque no tan determinantes como el área o la calidad de construcción, estas variables de configuración y servicios del terreno influyen en la valoración final, especialmente cuando se combinan con la ubicación .

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?**

Table 14: Skewness of Selected Variables

	Variable	Skewness
SalePrice	SalePrice	1.88
GrLivArea	GrLivArea	1.37
LotArea	LotArea	12.20
TotalBsmtSF	TotalBsmtSF	1.52

GarageArea	GarageArea	0.18
MasVnrArea	MasVnrArea	2.67

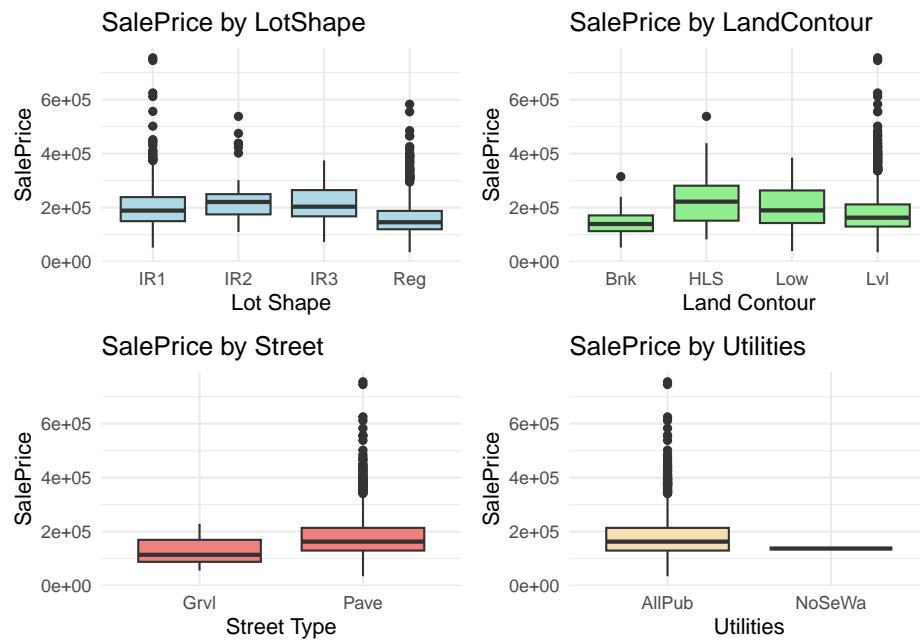


- Histogramas y boxplots de SalePrice y GrLivArea

- **SalePrice:** Presenta una distribución sesgada a la derecha (right-skewed) con algunos outliers muy altos.
- **GrLivArea:** También muestra outliers en la cola derecha y una distribución asimétrica. Esto sugiere que, para un **modelo de regresión**, podría ser beneficioso aplicar **transformaciones logarítmicas** o alguna técnica de robustez que maneje valores extremos.

SalePrice y GrLivArea tienen outliers y sesgo. Para un análisis predictivo, es habitual considerar log(SalePrice) y, a veces, log(GrLivArea), o bien detectar y tratar outliers que puedan distorsionar la estimación.

10. ¿Cómo se combinan las variables de calidad, área y ubicación para explicar de forma conjunta la variabilidad en el precio de las propiedades?



A partir de todos los gráficos:

- **Calidad (OverallQual, BsmtQual, GarageQual, etc.):**
Las viviendas de mejor calidad y en buen estado destacan con precios altos en todos los ejes (área, sótano, garaje).
- **Área (GrLivArea, TotalBsmtSF, LotArea):**
El tamaño habitable es uno de los principales impulsores del precio; sin embargo, si la calidad es baja o la ubicación desfavorable, el precio no sube tanto.
- **Ubicación y servicios (Neighborhood, MSZoning, Street, Utilities):**
Zonas residenciales codiciadas y servicios completos pueden hacer que, incluso con áreas menores, se alcancen precios similares a los de casas grandes en barrios menos deseados.

En conjunto, una casa grande, con acabados de calidad y en un vecindario atractivo, se sitúa en la parte alta del rango de precios. Por el contrario, deficiencias en cualquiera de estas dimensiones pueden limitar el valor final de la vivienda.

Transformación de Datos y Preprocesamiento

La exploración inicial de los datos permitió identificar que para una mejor comprensión y modelado de los datos es necesario transformar y preprocesar el conjunto de datos. Dentro de las transformaciones necesarias a realizar se detallan las siguientes:

- **Manejo de NAs:**
 - Reemplazar NAs con “None” en variables categóricas donde la ausencia sea semánticamente “no existe”.
 - Colocar 0 en variables de área donde no exista sótano/garaje.
 - Decidir si eliminar variables con demasiados NAs irrelevantes.
- **Agrupación de categorías poco frecuentes:**
 - Unir en “Other” o “Rare” para evitar demasiados dummies con muy pocos registros.
- **Codificación:**
 - One-Hot para nominales (Neighborhood, BldgType, etc.).
 - Ordinal para calidades y condiciones (Ex > Gd > TA > Fa > Po).
- **Outliers:**
 - Evaluar la eliminación o recorte (capping) de valores extremadamente altos en variables como SalePrice, GrLivArea, LotArea.
 - Transformar SalePrice y otras variables con log para reducir skew.
- **Feature engineering:**
 - Crear variables de área total, antigüedad, total de baños, puntuaciones de calidad, etc.
 - Comprobar su correlación con SalePrice para validarlas.
- **Escalado:**
 - Normalizar o estandarizar variables numéricas según el algoritmo y la magnitud de los valores.
- **Validación:**
 - Separar datos de entrenamiento y testantes de encodings y escalados, para no sobreajustar.

Análisis de Grupos

Estadístico de Hopkins

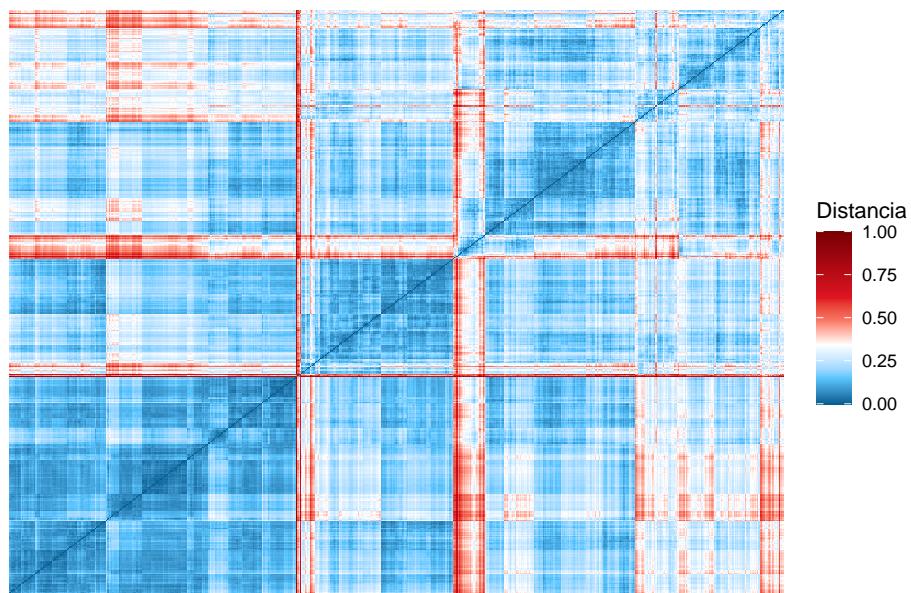
```
## [1] "Hopkins statistic: 0.8947"
```

El estadístico de Hopkins confirma una fuerte tendencia a clusterizar.

VAT

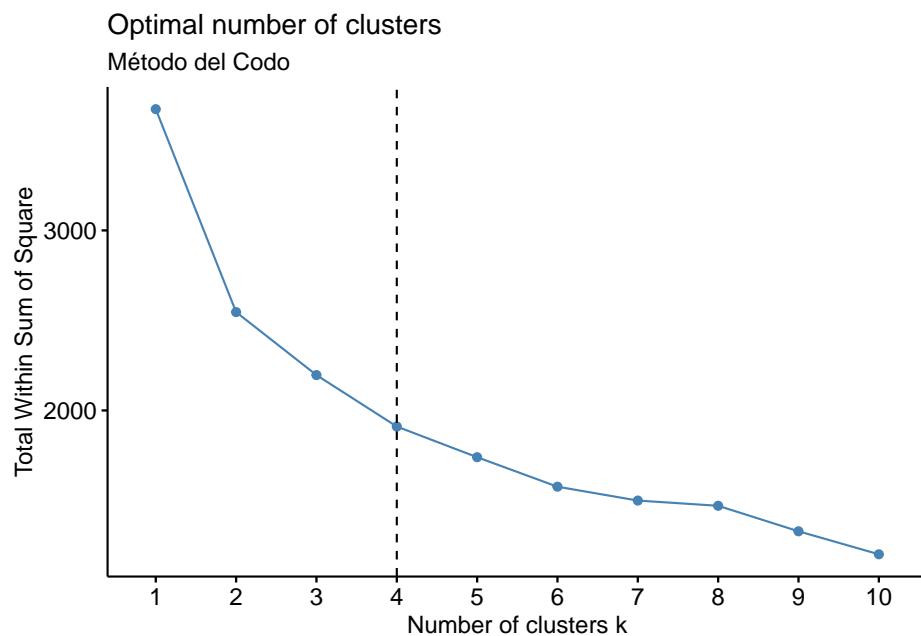
```
## [1] "dist"
## [1] 0 1
```

VAT sobre muestra de 1120 observaciones



Se observa que la matriz de distancias presenta una estructura clara, con bloques de observaciones similares en color blanco y líneas oscuras que separan grupos de observaciones. Esto sugiere que los datos tienen una estructura no aleatoria y son adecuados para el clustering.

Número óptimo de clusters (Método del Codo)



La gráfica muestra que el codo se encuentra entre $k=3$ y $k = 4$, lo que sugiere que este es el número óptimo de clusters para el conjunto de datos.

K-Means

Agrupamiento con K=3

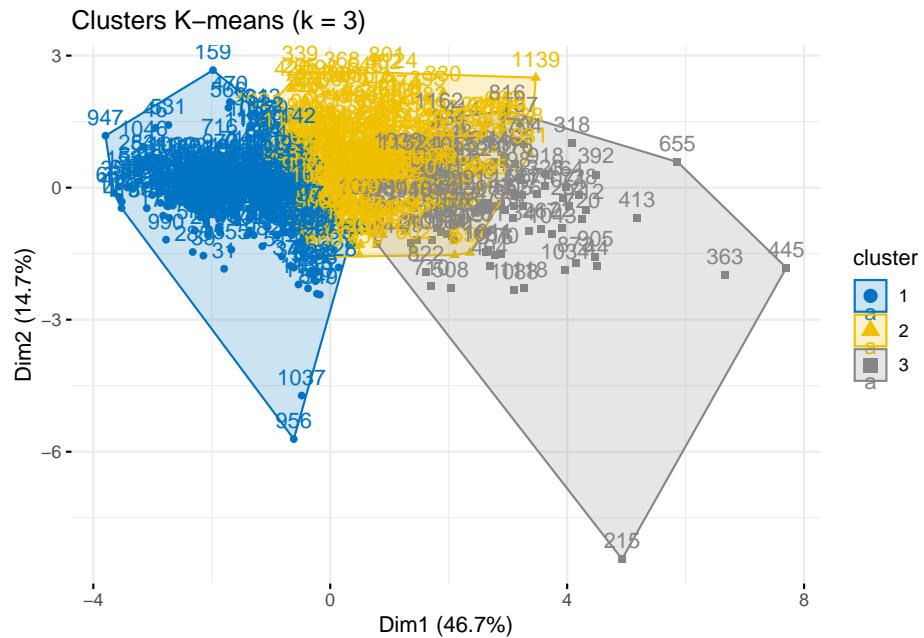
Table 15: Resumen General del Modelo K-means

Cluster	Size	Tot_WithinSS
1	313	2184.212
2	279	2184.212
3	114	2184.212

Table 16: Centros de los Clusters

SalePrice_log	GrLivArea_log	LotArea_log	OverallQual	YearBuilt	TotalBsmtSF_log	GarageArea_log	Cluster
0.906	0.565	0.329	1.136	0.916	0.747	0.646	1
0.125	0.325	0.316	-0.107	-0.154	-0.261	0.070	2
-0.660	-1.442	-0.205	-0.610	-0.325	0.091	-0.671	3

Visualización de Clusters



Agrupamiento con K=4

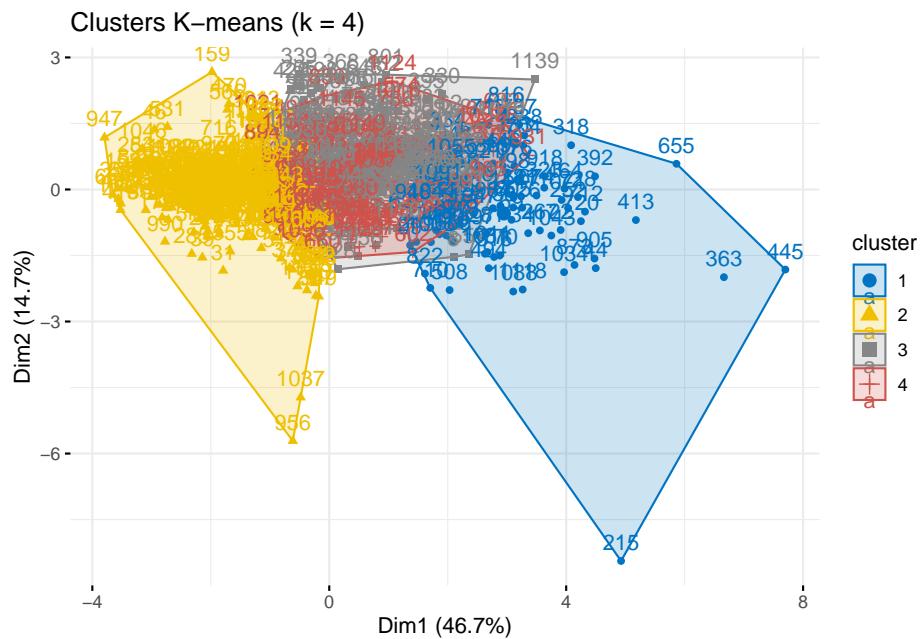
Table 17: Resumen General del Modelo K-means

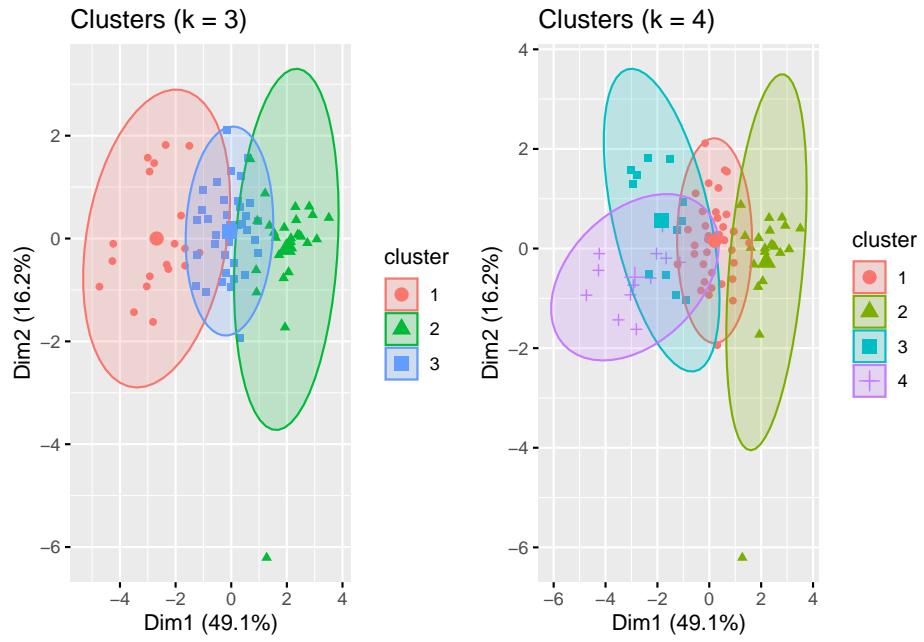
Cluster	Size	Tot_WithinSS
1	313	2184.212
2	279	2184.212

Table 18: Centros de los Clusters

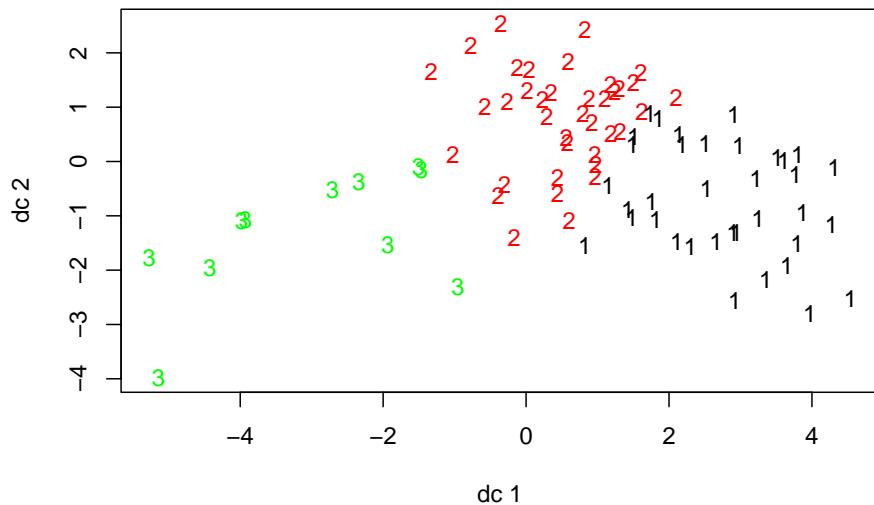
SalePrice_log	GrLivArea_log	LotArea_log	OverallQual	YearBuilt	TotalBsmtSF_log	GarageArea_log	Clust
-0.7433	-1.6085	-0.3407	-0.6400	-0.4149	-0.0384	-0.8610	1
0.9454	0.5778	0.3542	1.2152	0.9326	0.8199	0.6859	2
0.0054	0.1156	0.3924	-0.3180	-0.3387	0.5308	0.0899	3
0.2992	0.4275	0.1652	0.2295	0.3218	-1.0689	0.0865	4

Visualización de Clusters

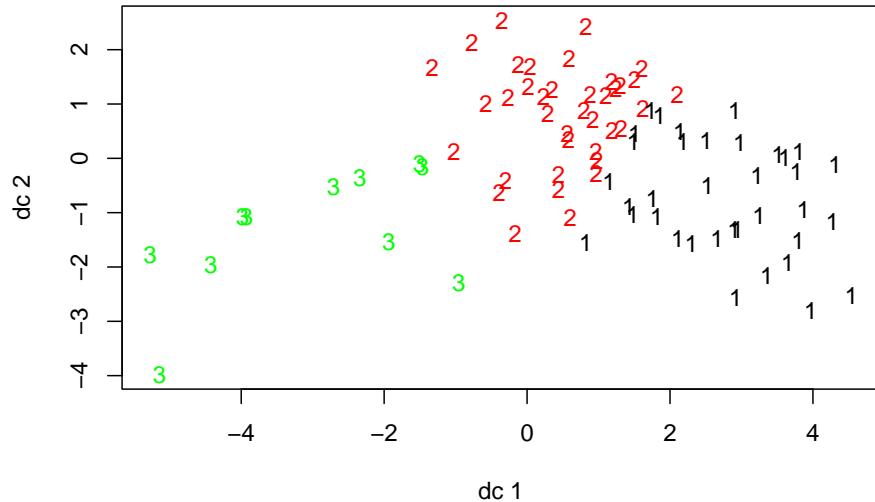




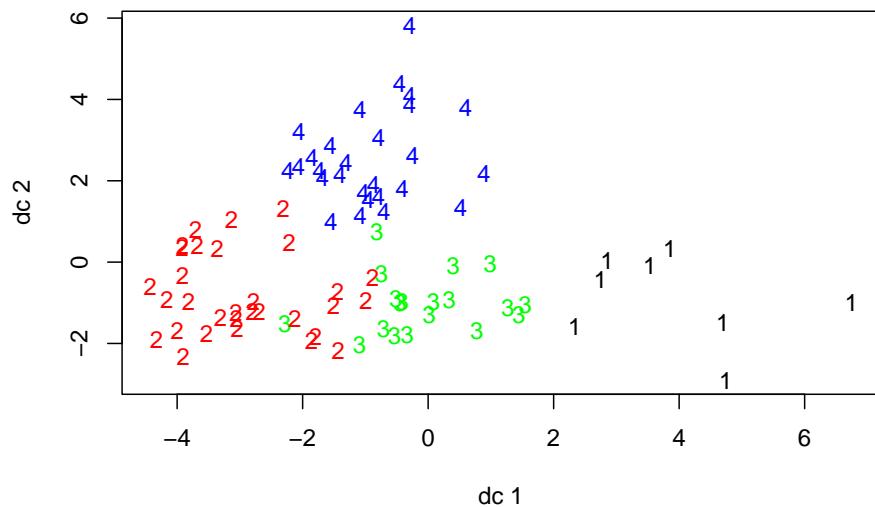
Ubicación de los Clusters (k = 3)



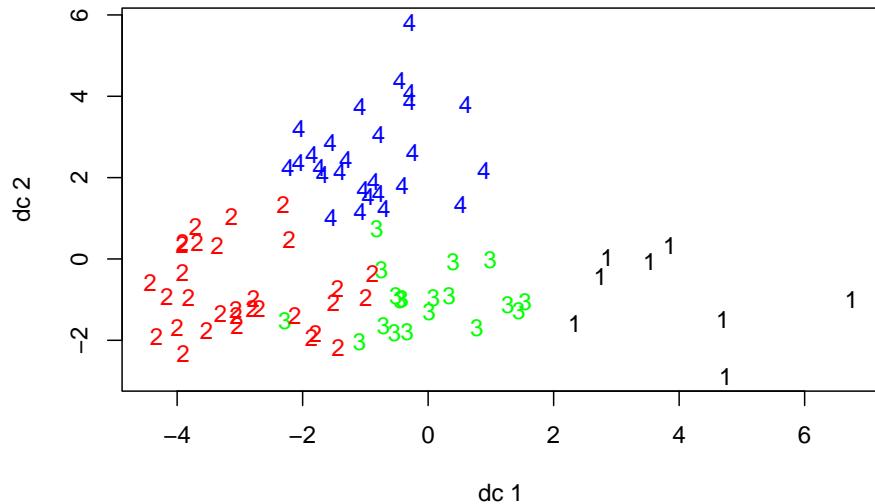
Ubicación de los Clusters ($k = 3$)



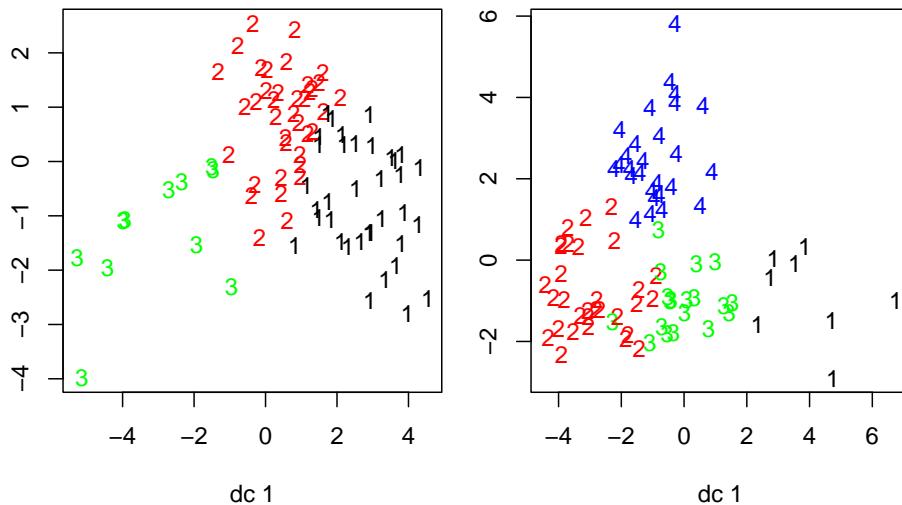
Ubicación de los Clusters ($k = 4$)



Ubicación de los Clusters ($k = 4$)



Ubicación de los Clusters ($k = 3$) Ubicación de los Clusters ($k = 4$)

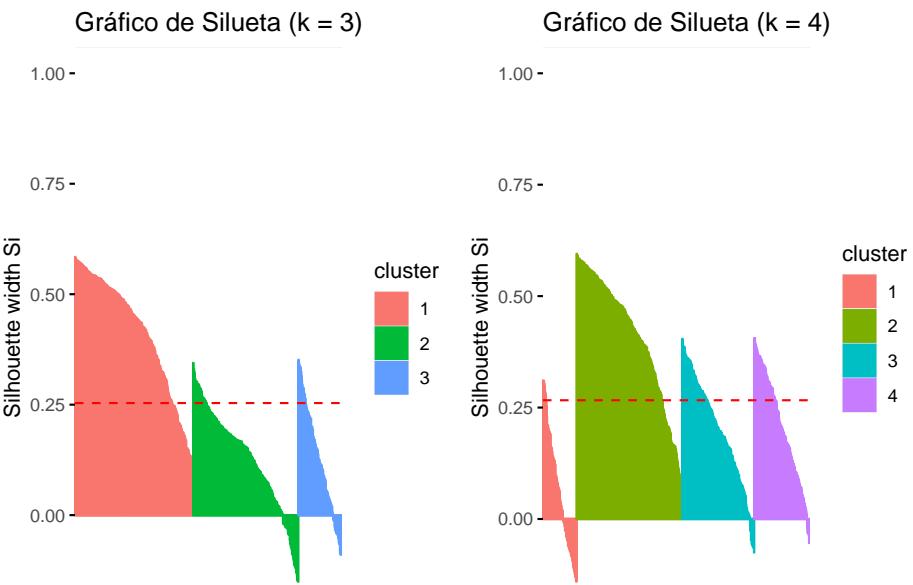


- El modelo con 3 clusters agrupa el mercado de viviendas en tres segmentos (alta, media, baja) de forma razonable y más sencilla.
- El modelo con 4 clusters incrementa la separación estadística y reduce la variabilidad interna, pero añade complejidad en la interpretación. La elección final depende tanto de la validez estadística como de la utilidad práctica para el análisis o toma de decisiones .
- Con 4 clusters se logra una mayor segmentación del mercado, dividiendo uno de los grupos grandes en dos.

Calidad del Agrupamiento

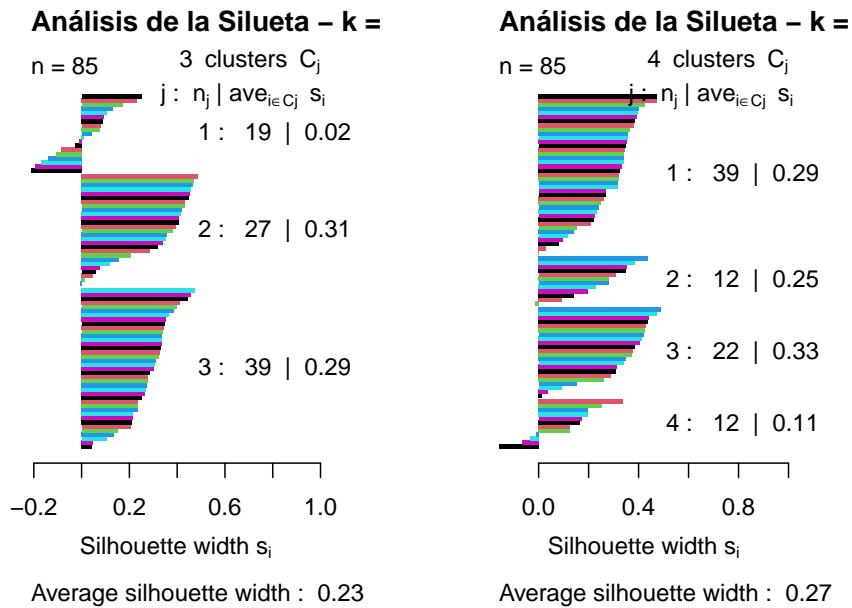
```
##   cluster size ave.sil.width
## 1       1   313      0.41
## 2       2   279      0.13
## 3       3   114      0.12

##   cluster size ave.sil.width
## 1       1    89      0.06
## 2       2   280      0.41
## 3       3   191      0.20
## 4       4   146      0.21
```



```
## Silhouette promedio (k = 3): 0.2343003
```

```
## Silhouette promedio (k = 4): 0.2681828
```



Ambas configuraciones ($k=3$ y $k=4$) muestran un promedio de silueta, bajo. Esto indica que:

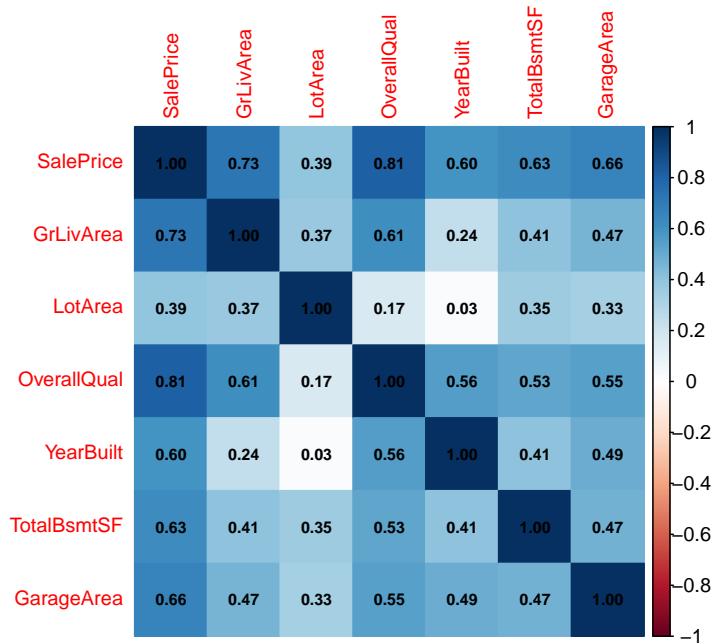
1. La separación entre clusters no es muy marcada (hay solapamiento).
2. Con $k=4$ aparece un cluster pequeño que mejora su silueta, pero otro obtiene valores muy bajos, compensando el beneficio.
3. En términos globales, no hay una gran diferencia en la calidad de la segmentación entre 3 y 4 clusters.

En resumen, los datos sugieren que la estructura de los clusters no está fuertemente definida o que tal vez se necesiten otras técnicas (por ejemplo, reducción de dimensionalidad o índices adicionales) para encontrar una segmentación más clara.

Análisis de Componentes Principales (PCA)

Mátriz de Correlación

```
## Determinante de la matriz de correlacion: 0.01917
```



Determinante de la Matriz de Correlación: 0.01917

Observervamos que el determinante de la matriz de correlación es cercano a 0, lo que indica que las variables están altamente correlacionadas, lo cual es un requisito para aplicar PCA.

La matriz de correlación confirma que la calidad y el tamaño (habitable, sótano, garaje) son los principales impulsores del precio de venta de las viviendas, mientras que la antigüedad y el área del lote tienen un papel secundario

Indice de Kaiser-Meyer-Olkin (KMO)

Table 19: Índice de Kaiser-Meyer-Olkin (KMO) Global

Índice.KMO.Global
0.8144

Test de esfericidad de Bartlett

Table 20: Test de Esfericidad de Bartlett

Chi.cuadrado	Grados.de.Libertad	p.valor
4606.007	21	0

El índice KMO es mayor a 0.5 y el test de Bartlett es significativo con un p-valor = 0 , indicando que las variables están efectivamente correlacionadas., lo que indica que los datos son adecuados para realizar un análisis de componentes principales.

Análisis de Componentes Principales

Table 21: Proporción de Varianza Explicada por Componente

	Componente	Proporción_Varianza	Varianza_Acumulada
PC1	PC1	56.12	56.12
PC2	PC2	15.22	71.35
PC3	PC3	9.67	81.01
PC4	PC4	7.65	88.66
PC5	PC5	5.88	94.54
PC6	PC6	3.82	98.36
PC7	PC7	1.64	100.00

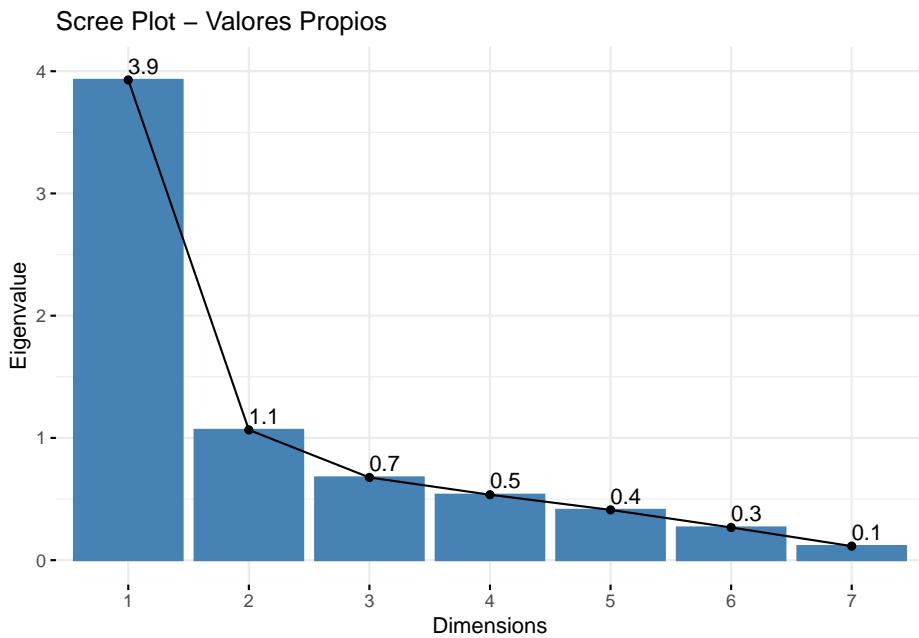
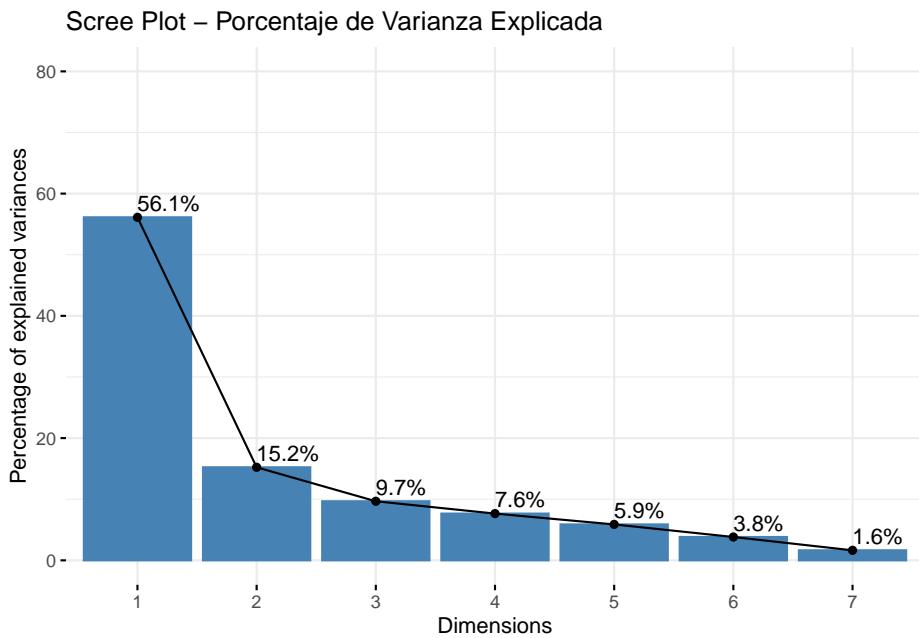
Regla de Kaiser

Table 22: Valores Propios de Cada Componente Principal

	Componente	Valor_Propio
PC1		3.9288
PC2		1.0656
PC3		0.6766
PC4		0.5354
PC5		0.4113
PC6		0.2674
PC7		0.1150

Se retienen los primeros 2 componentes principales cumplen con este criterio, por lo que se retendrán para el análisis.

Regla de Sedimentación



Gran parte de la variabilidad se concentra en la primera y segunda componente, y que la tercera todavía aporta una porción apreciable. Más allá de la tercera, las ganancias en varianza explicada son mucho menores.

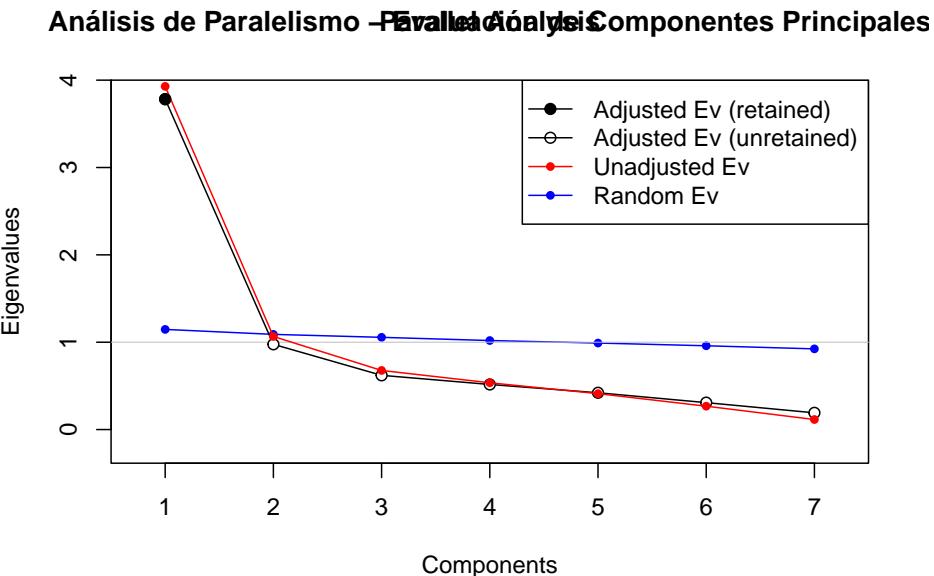
Paralelismo de los Componentes Principales

```
##  
## Using eigendecomposition of correlation matrix.  
## Computing: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
```

```

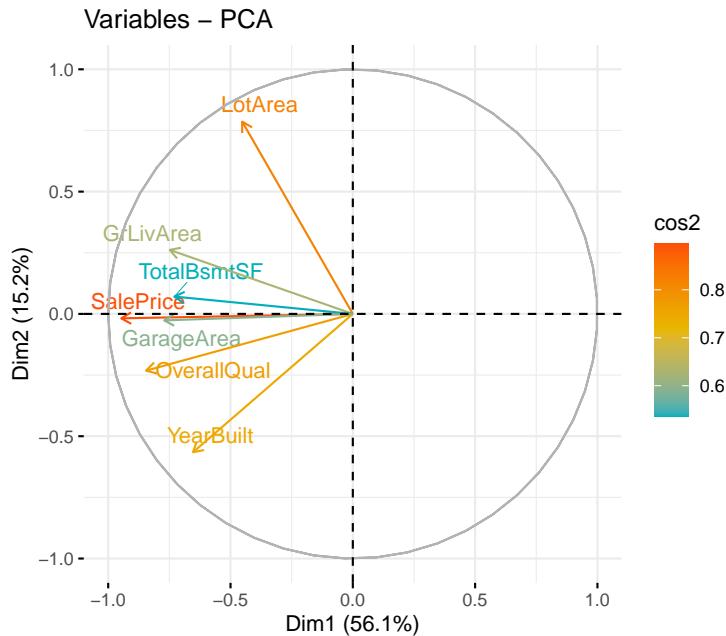
## 
## 
## Results of Horn's Parallel Analysis for component retention
## 210 iterations, using the 95 centile estimate
## 
## -----
## Component    Adjusted      Unadjusted     Estimated
##             Eigenvalue   Eigenvalue     Bias
## -----
## 1           3.782213    3.928780    0.146567
## -----
## 
## Adjusted eigenvalues > 1 indicate dimensions to retain.
## (1 components retained)

```



El análisis sugiere conservar 2 componentes principales, ya que la tercera (y siguientes) no se diferencian lo bastante de los valores propios generados al azar.

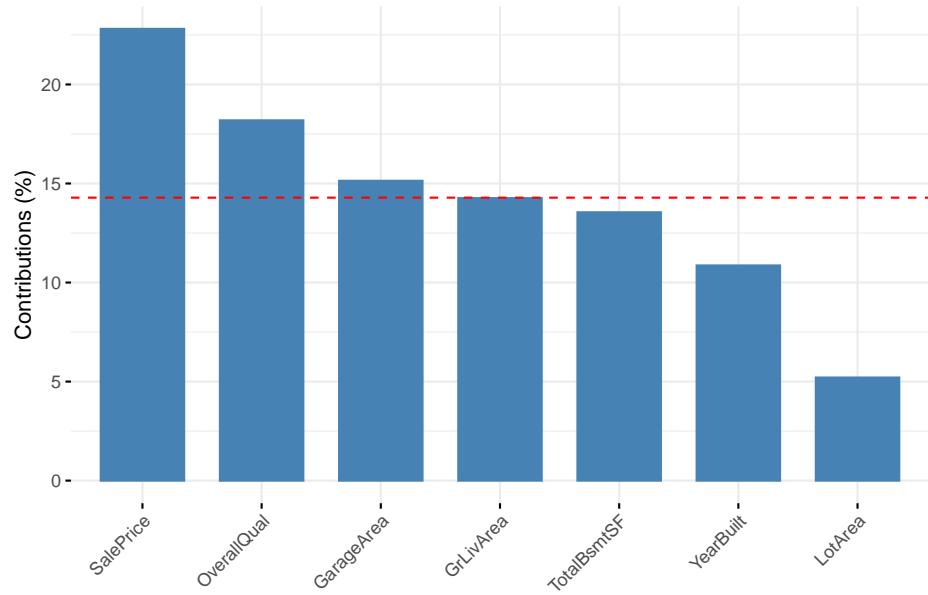
Carga Factorial



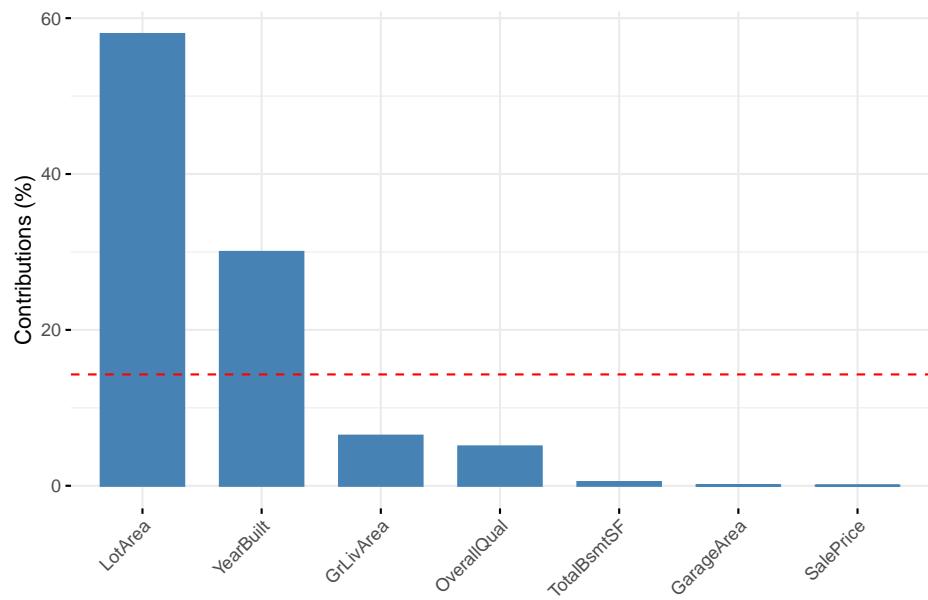
- La primera componente (Dim1) captura la mayor parte de la información relevante para el valor de la casa (tamaño, calidad, precio).
- La segunda componente aporta un matiz adicional (por ejemplo, tamaño de lote y antigüedad), pero su peso en la varianza total es menor.
- Variables como SalePrice, OverallQual, GrLivArea, TotalBsmtSF y GarageArea se mueven en conjunto, reforzando la idea de que la calidad y los espacios construidos son factores clave en la determinación del precio de venta.

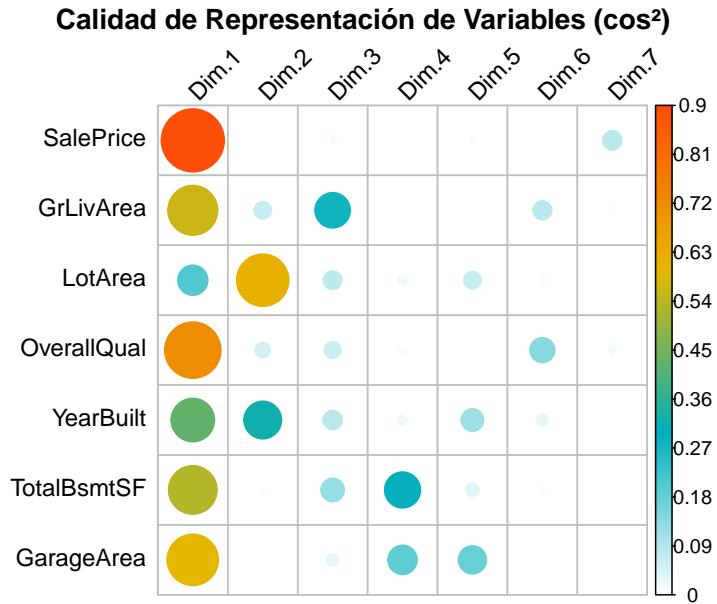
Contribución de Variables en las primeras dimensiones

Contribution of variables to Dim-1



Contribution of variables to Dim-2





- La dimensión 1 es la dimensión principal en la cual se concentran relaciones de precio, calidad y espacios de las casas.
- La segunda dimensión explica factores como el tamaño del lote y la antigüedad de la vivienda que no presentan una relación tan significativa con el precio como la primera.
- El análisis de **contribución** y \cos^2 refuerza que, en el espacio de las dos primeras componentes, se distinguen claramente dos ejes interpretables:
 - Eje 1: Valor/Calidad/Tamaño construidos
 - Eje 2: Tamaño de lote/Año de construcción

Modelado Predictivo (Regresión lineal)

Se separan los datos en conjuntos de entrenamiento y prueba, y se aplica un modelo de regresión lineal para predecir el precio de venta de las viviendas.

Table 23: Número de filas en cada dataset

Dataset	Filas
Entrenamiento	1169
Prueba	291

Modelo Univariado

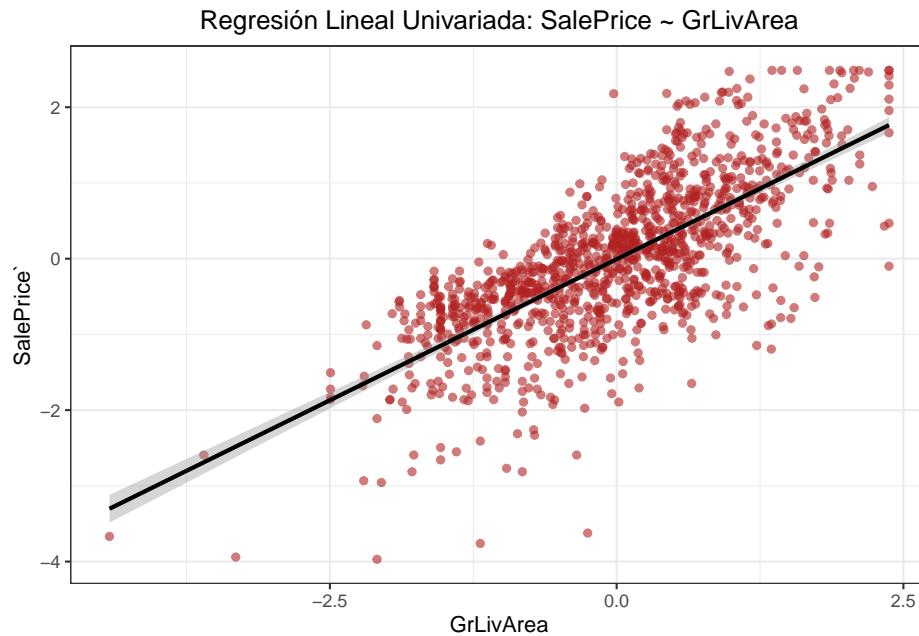
Se trabaja con la variable más correlacionada con el precio de venta: GrLivArea.

Table 24: Resumen del Modelo Lineal

term	estimate	std.error	statistic	p.value
(Intercept)	-0.0062902	0.0201771	-0.3117493	0.7552867
GrLivArea	0.7454009	0.0205889	36.2040340	0.0000000

GrLivArea es un predictor muy fuerte en el modelo, mientras que el intercepto no aporta información relevante.

Representación Gráfica



La gráfica muestra una relación lineal positiva entre el área habitable y el precio de venta, con una dispersión considerable en los datos.

Residuos

Table 25: Primeras Predicciones

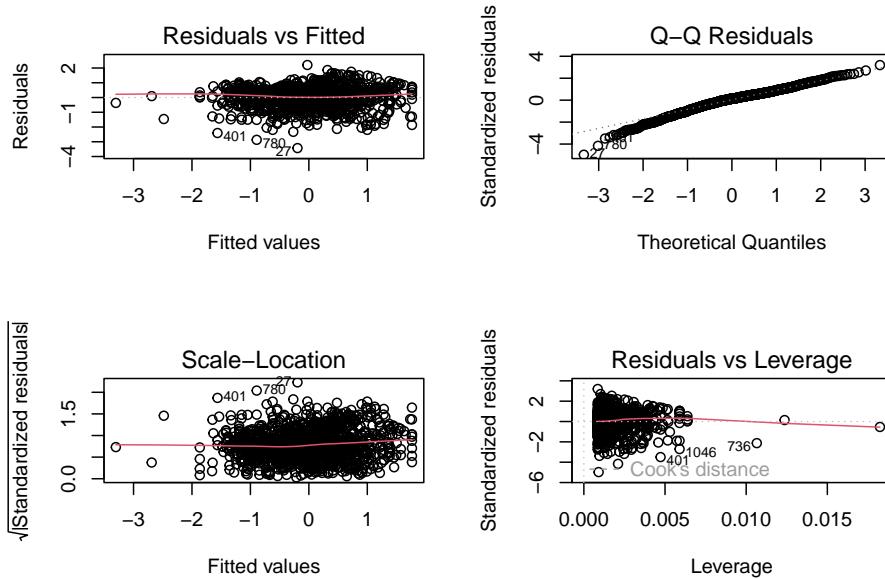
Predicciones
0.3975843
-0.2912703
0.4961974
0.4068484
0.9669432
-0.1183730

Table 26: Primeros Residuales

Residuales
0.1762891
0.5125934
0.2542833
-0.8454940
0.0683811
-0.2663739

Table 27: Número total de predicciones

Total
1169



- **Residuals vs Fitted:**

- Muestra cómo se distribuyen los residuos en función de las predicciones. No se aprecia un patrón en “U” o “ ” muy marcado, aunque se observan algunos puntos alejados.

- **Q-Q Residuals:**

- Varios puntos siguen la línea teórica, pero se observa **cierta desviación** en colas.
- Indica que la distribución de los residuos se aleja de la normalidad en los valores más altos y más bajos.

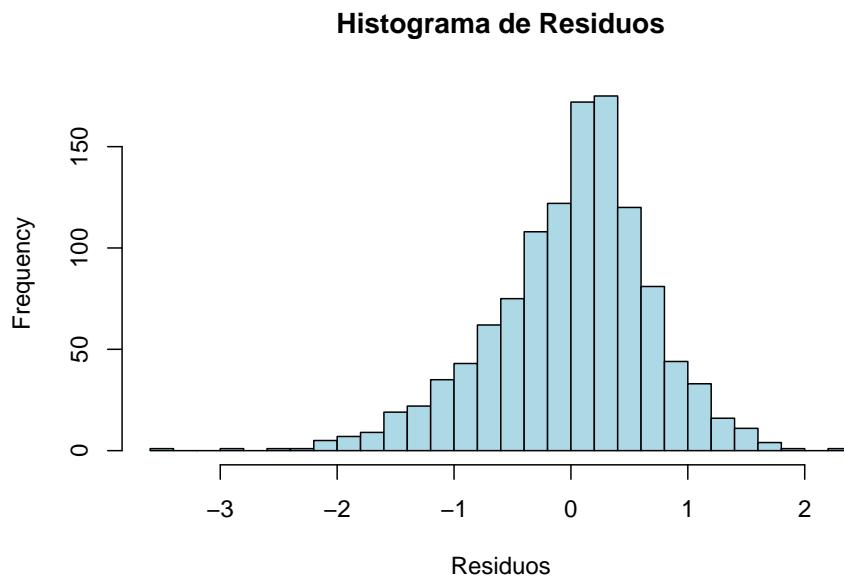
- **Scale-Location:**

- No se ve un patrón muy pronunciado, pero sí algo de dispersión variable.

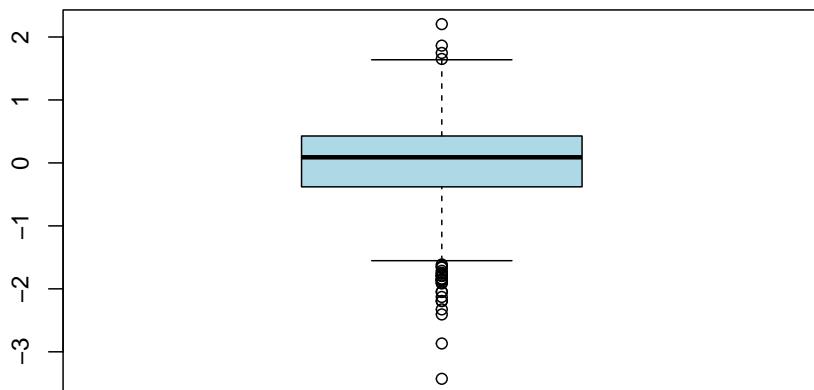
- **Residuals vs Leverage:**

- Algunas observaciones podrían tener un impacto desproporcionado en el ajuste.

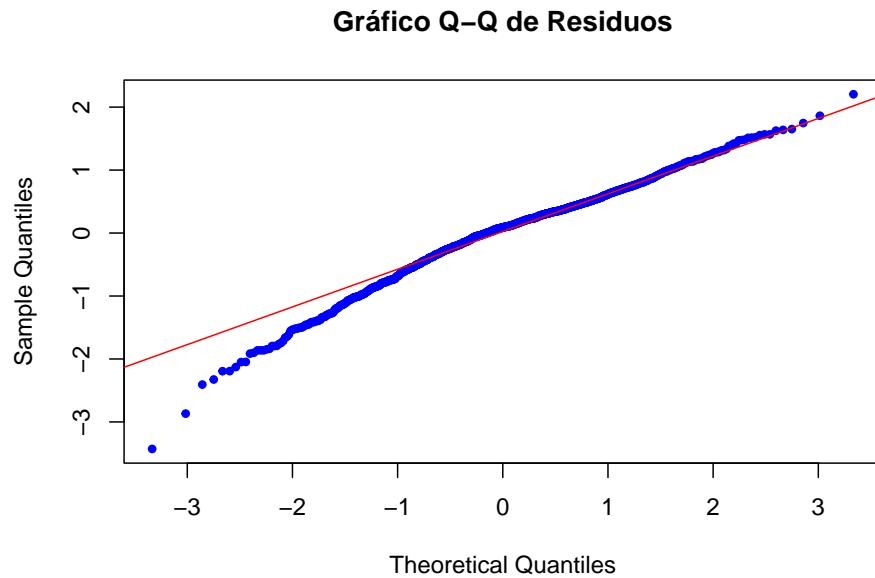
Distribución de Residuos



Boxplot de Residuos



- **Histograma:** Los residuos se concentran mayoritariamente alrededor de 0, con cola a la izquierda y derecha.
- **Boxplot:** Hay valores atípicos tanto por debajo como por encima. La mediana está cerca de 0.



Prueba de Normalidad

Table 28: Resultado del Test de Lilliefors

Estadístico	p_value
0.0716157	0

Predicción y Ecualización

```
##           1          2          3          4          5          6 
## 0.8526736 -0.7298945 -1.0275957 -2.3004966 -0.6089351 -0.6931393
## RMSE del modelo univariado: 0.6477968
```

Gráfico de Predicciones

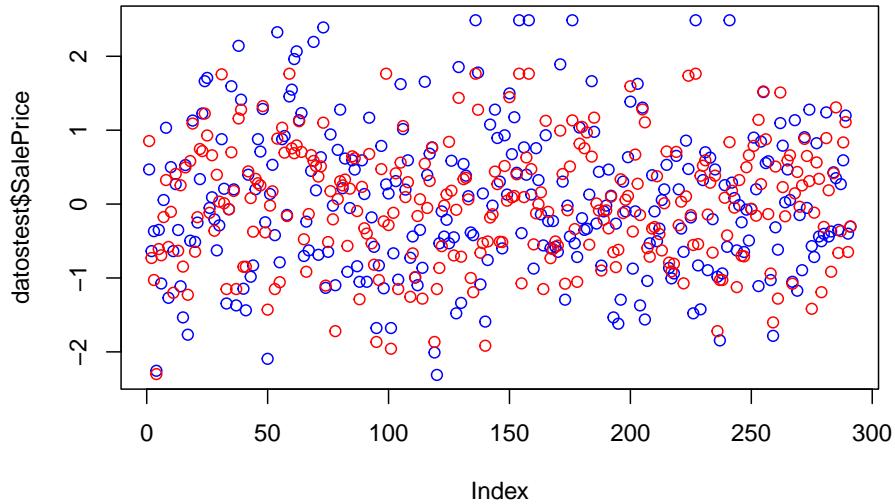


Table 29: Resumen de la diferencia ($\text{SalePrice} - \text{pred_uni}$)

Estadística	Valor
Min.	-2.6678587
1st Qu.	-0.2930908
Median	0.0664041
Mean	0.0315591
3rd Qu.	0.4555516
Max.	1.6497790

1. **Relación significativa:** GrLivArea tiene un efecto fuertemente positivo y significativo en SalePrice.
2. **Intercecpo no significativo:** No aporta gran información cuando GrLivArea = 0.
3. **Distribución de residuos:**
 - La media y mediana de ($\text{SalePrice} - \text{Predicción}$) están cercanas a 0, lo que sugiere que no hay un sesgo sistemático.
 - Existen puntos outlier y el test de Lilliefors indica **no normalidad** de los residuos.
 - Aun así, el modelo univariado capta la tendencia global: a mayor área habitable, mayor precio.

En suma, la **regresión univariada** $\text{SalePrice} \sim \text{GrLivArea}$ **confirma** que el área habitable es un predictor relevante y explica una parte importante de la variabilidad en el precio de las viviendas, pero **no** captura toda la complejidad del problema.

Modelo Multivariado (Todas las Variables numéricas)

Selección de Variables

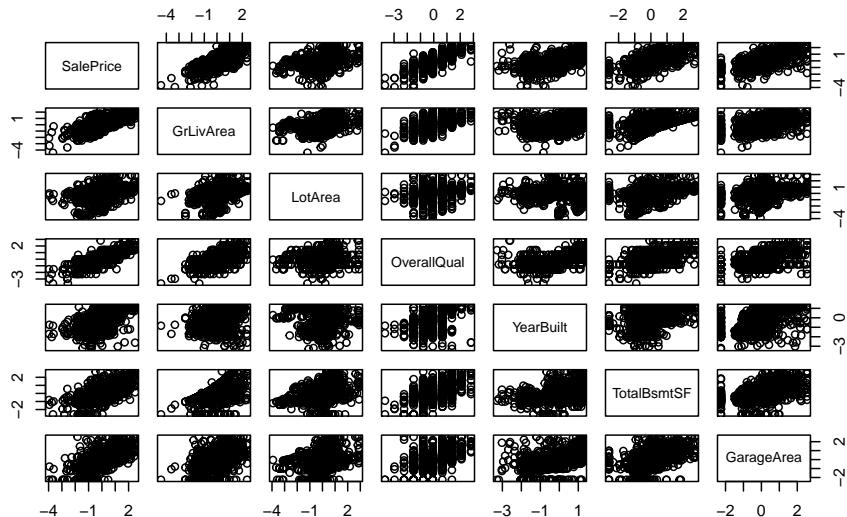
Se seleccionan todas las variables numéricas del dataset. Sin embargo, por motivos de visualización se escogen aquellas con mayor correlación para el cálculo de la matriz de correlación variables que presentan la mayor relación de acuerdo a la matriz de correlación.

Table 30: Variables numéricas comunes

Variables
Id
MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
MasVnrArea
ExterQual
ExterCond
BsmtQual
BsmtCond
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF
HeatingQC
X1stFlrSF
X2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
KitchenQual
TotRmsAbvGrd
Fireplaces
GarageYrBlt
GarageCars
GarageArea
GarageQual
GarageCond
WoodDeckSF
OpenPorchSF
EnclosedPorch
X3SsnPorch
ScreenPorch
PoolArea
MiscVal
MoSold
YrSold
SalePrice
Neighborhood.EDwards
Neighborhood.Gilbert
Neighborhood.NAmes
Neighborhood.NridgHt
Neighborhood.NWAmes

Variables
Neighborhood.OldTown
Neighborhood.Other
Neighborhood.Sawyer
Neighborhood.Somerst
BldgType.Other
BldgType.TwnhsE
HouseStyle.1Story
HouseStyle.2Story
HouseStyle.Other
Exterior1st.MetalSd
Exterior1st.Other
Exterior1st.Plywood
Exterior1st.VinylSd
Exterior1st.Wd.Sdng
Exterior2nd.MetalSd
Exterior2nd.Other
Exterior2nd.Plywood
Exterior2nd.VinylSd
Exterior2nd.Wd.Sdng
TotalArea
HouseAge
TotalBath

Matriz de Dispersion Reducida

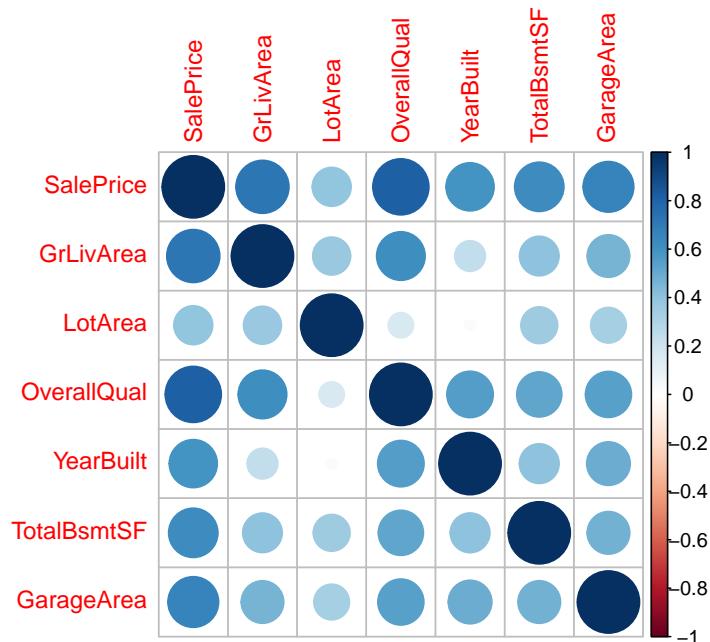


```
##          SalePrice GrLivArea   LotArea OverallQual YearBuilt TotalBsmtSF
## SalePrice    1.0000000 0.7273271 0.3929428  0.8131492 0.5991370  0.6296148
## GrLivArea    0.7273271 1.0000000 0.3740873  0.6109062 0.2447465  0.4062122
## LotArea      0.3929428 0.3740873 1.0000000  0.1678780 0.0292402  0.3518232
## OverallQual  0.8131492 0.6109062 0.1678780  1.0000000 0.5574899  0.5281824
## YearBuilt     0.5991370 0.2447465 0.0292402  0.5574899 1.0000000  0.4099544
## TotalBsmtSF  0.6296148 0.4062122 0.3518232  0.5281824 0.4099544  1.0000000
```

```

## GarageArea 0.6641182 0.4695250 0.3302363 0.5489846 0.4920082 0.4709556
## GarageArea
## SalePrice 0.6641182
## GrLivArea 0.4695250
## LotArea 0.3302363
## OverallQual 0.5489846
## YearBuilt 0.4920082
## TotalBsmtSF 0.4709556
## GarageArea 1.0000000

```



Definimos los predictores

Table 31: Predictores seleccionados

Predictor
Id
MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
MasVnrArea
ExterQual
ExterCond
BsmtQual
BsmtCond
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF

Predictor
TotalBsmtSF
HeatingQC
X1stFlrSF
X2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
KitchenQual
TotRmsAbvGrd
Fireplaces
GarageYrBlt
GarageCars
GarageArea
GarageQual
GarageCond
WoodDeckSF
OpenPorchSF
EnclosedPorch
X3SsnPorch
ScreenPorch
PoolArea
MiscVal
MoSold
YrSold
Neighborhood.Edwards
Neighborhood.Gilbert
Neighborhood.NAmes
Neighborhood.NridgHt
Neighborhood.NWAmes
Neighborhood.OldTown
Neighborhood.Other
Neighborhood.Sawyer
Neighborhood.Somerst
BldgType.Other
BldgType.TwnhsE
HouseStyle.1Story
HouseStyle.2Story
HouseStyle.Other
Exterior1st.MetalSd
Exterior1st.Other
Exterior1st.Plywood
Exterior1st.VinylSd
Exterior1st.Wd.Sdng
Exterior2nd.MetalSd
Exterior2nd.Other
Exterior2nd.Plywood
Exterior2nd.VinylSd

Predictor
Exterior2nd.Wd.Sdng
TotalArea
HouseAge
TotalBath

Table 32: Fórmula del modelo

Formula
SalePrice ~ Id + MSSubClass + LotFrontage + LotArea + OverallQual + OverallCond + YearBuilt + YearRemodAdd + MasVnrArea + ExterQual + ExterCond + BsmtQual + BsmtCond + BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF + TotalBsmtSF + HeatingQC + X1stFlrSF + X2ndFlrSF + LowQualFinSF + GrLivArea + BsmtFullBath + BsmtHalfBath + FullBath + HalfBath + BedroomAbvGr + KitchenAbvGr + KitchenQual + TotRmsAbvGrd + Fireplaces + GarageYrBlt + GarageCars + GarageArea + GarageQual + GarageCond + WoodDeckSF + OpenPorchSF + EnclosedPorch + X3SsnPorch + ScreenPorch + PoolArea + MiscVal + MoSold + YrSold + Neighborhood.Edwards + Neighborhood.Gilbert + Neighborhood.NAmes + Neighborhood.NridgHt + Neighborhood.NWAmes + Neighborhood.OldTown + Neighborhood.Other + Neighborhood.Sawyer + Neighborhood.Somerst + BldgType.Other + BldgType.TwnhsE + HouseStyle.1Story + HouseStyle.2Story + HouseStyle.Other + Exterior1st.MetalSd + Exterior1st.Other + Exterior1st.Plywood + Exterior1st.VinylSd + Exterior1st.Wd.Sdng + Exterior2nd.MetalSd + Exterior2nd.Other + Exterior2nd.Plywood + Exterior2nd.VinylSd + Exterior2nd.Wd.Sdng + TotalArea + HouseAge + TotalBath

Table 33: Dimensiones de los conjuntos filtrados

Conjunto	Filas	Columnas
train_filtered	1072	73
test_filtered	269	73

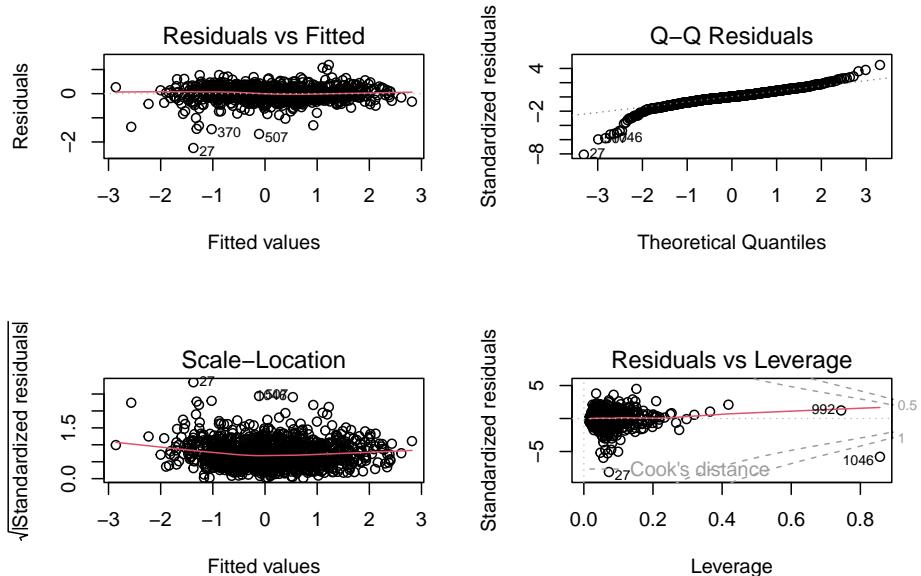
Table 34: Resumen del Modelo Lineal Múltiple

term	estimate	std.error	statistic	p.value
(Intercept)	-0.0057034	0.0095234	-0.5988861	0.5493842
Id	0.0013439	0.0091483	0.1469071	0.8832349
MSSubClass	-0.0020344	0.0340721	-0.0597096	0.9523988
LotFrontage	-0.0100581	0.0099056	-1.0153935	0.3101634
LotArea	0.1119300	0.0140911	7.9432989	0.0000000
OverallQual	0.2057266	0.0184352	11.1594743	0.0000000
OverallCond	0.1671981	0.0133021	12.5693247	0.0000000
YearBuilt	0.1568658	0.0265167	5.9157404	0.0000000
YearRemodAdd	0.0358082	0.0164892	2.1716070	0.0301193
MasVnrArea	0.0106720	0.0110142	0.9689266	0.3328156
ExterQual	0.0148884	0.0159504	0.9334191	0.3508285

term	estimate	std.error	statistic	p.value
ExterCond	-0.0335578	0.0106096	-3.1629703	0.0016087
BsmtQual	0.0322408	0.0167505	1.9247658	0.0545415
BsmtCond	0.0257093	0.0099840	2.5750500	0.0101650
BsmtFinSF1	-0.2919041	0.0478023	-6.1064907	0.0000000
BsmtFinSF2	-0.1176444	0.0190203	-6.1851898	0.0000000
BsmtUnfSF	-0.3620954	0.0478228	-7.5716124	0.0000000
TotalBsmtSF	0.4831023	0.0466859	10.3479276	0.0000000
HeatingQC	0.0192274	0.0126187	1.5237169	0.1278950
X1stFlrSF	-0.0074141	0.0468250	-0.1583355	0.8742244
X2ndFlrSF	0.0821875	0.0419518	1.9590913	0.0503791
LowQualFinSF	-0.0131931	0.0117202	-1.1256657	0.2605766
GrLivArea	0.2846448	0.0513077	5.5477952	0.0000000
BsmtFullBath	0.0342874	0.0136562	2.5107691	0.0122034
BsmtHalfBath	0.0021831	0.0102653	0.2126718	0.8316262
FullBath	0.0187390	0.0169284	1.1069559	0.2685787
HalfBath	0.0137377	0.0144034	0.9537794	0.3404253
BedroomAbvGr	-0.0311246	0.0150916	-2.0623776	0.0394291
KitchenAbvGr	-0.0652824	0.0170027	-3.8395275	0.0001310
KitchenQual	0.0296612	0.0151429	1.9587528	0.0504188
TotRmsAbvGrd	0.0507330	0.0207969	2.4394529	0.0148821
Fireplaces	0.0481744	0.0117469	4.1010496	0.0000445
GarageYrBlt	0.0120139	0.0202121	0.5943929	0.5523836
GarageCars	0.0747201	0.0222782	3.3539499	0.0008265
GarageArea	0.0063357	0.0227832	0.2780871	0.7810029
GarageQual	0.0182023	0.0123121	1.4784007	0.1396148
GarageCond	0.0109902	0.0116990	0.9394123	0.3477455
WoodDeckSF	0.0236146	0.0099488	2.3736175	0.0178025
OpenPorchSF	0.0093855	0.0106854	0.8783435	0.3799679
EnclosedPorch	0.0163502	0.0107565	1.5200319	0.1288184
X3SsnPorch	0.0117285	0.0095145	1.2326960	0.2179783
ScreenPorch	0.0202490	0.0095071	2.1298858	0.0334242
PoolArea	0.0086813	0.0097172	0.8934018	0.3718566
MiscVal	-0.0203265	0.0158274	-1.2842642	0.1993464
MoSold	-0.0028832	0.0092122	-0.3129782	0.7543624
YrSold	-0.0291952	0.0092121	-3.1692351	0.0015747
Neighborhood.Edwards	-0.0442198	0.0134102	-3.2974692	0.0010099
Neighborhood.Gilbert	-0.0124792	0.0115038	-1.0847885	0.2782763
Neighborhood.NAmes	-0.0146532	0.0157138	-0.9325063	0.3512996
Neighborhood.NridgHt	0.0321786	0.0116727	2.7567433	0.0059441
Neighborhood.NWAmes	-0.0412126	0.0122424	-3.3663753	0.0007906
Neighborhood.OldTown	-0.0308663	0.0152134	-2.0288860	0.0427339
Neighborhood.Other	0.0140400	0.0172156	0.8155406	0.4149567
Neighborhood.Sawyer	-0.0159491	0.0121214	-1.3157833	0.1885477
Neighborhood.Somerst	0.0444381	0.0115950	3.8325123	0.0001347
BldgType.Other	-0.0260472	0.0265997	-0.9792305	0.3277024
BldgType.TwnhsE	-0.0065371	0.0247421	-0.2642076	0.7916743
HouseStyle.1Story	0.0602117	0.0267847	2.2479903	0.0247936
HouseStyle.2Story	-0.0062351	0.0205707	-0.3031058	0.7618721
HouseStyle.Other	0.0463780	0.0160702	2.8859576	0.0039860
Exterior1st.MetalSd	0.0276516	0.0400042	0.6912169	0.4895894
Exterior1st.Other	0.0341625	0.0234588	1.4562725	0.1456306
Exterior1st.Plywood	0.0004584	0.0186103	0.0246306	0.9803545

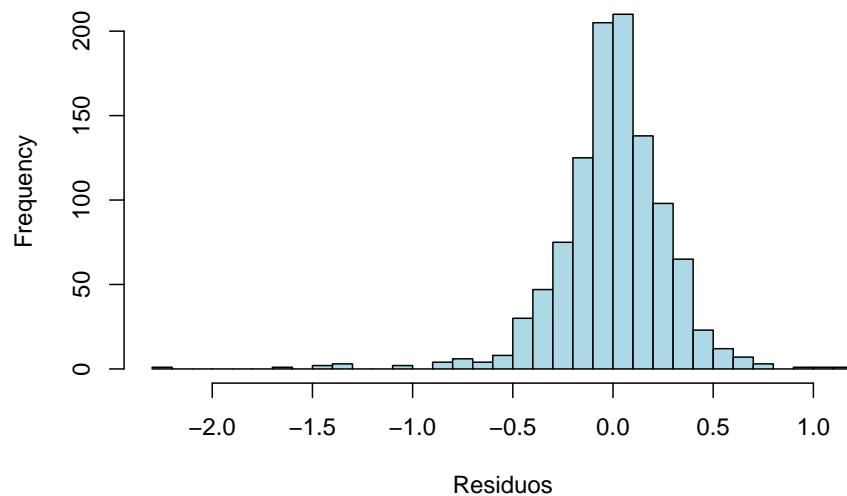
term	estimate	std.error	statistic	p.value
Exterior1st.VinylSd	0.0545221	0.0530089	1.0285466	0.3039410
Exterior1st.Wd.Sdng	-0.0279996	0.0282548	-0.9909668	0.3219410
Exterior2nd.MetalSd	0.0030025	0.0398330	0.0753778	0.9399292
Exterior2nd.Other	-0.0157355	0.0239848	-0.6560596	0.5119364
Exterior2nd.Plywood	-0.0096669	0.0196163	-0.4927989	0.6222627
Exterior2nd.VinylSd	-0.0315167	0.0531816	-0.5926230	0.5535672
Exterior2nd.Wd.Sdng	0.0292607	0.0275795	1.0609590	0.2889642
TotalArea	NA	NA	NA	NA
HouseAge	NA	NA	NA	NA
TotalBath	NA	NA	NA	NA

Analisis de Residuos

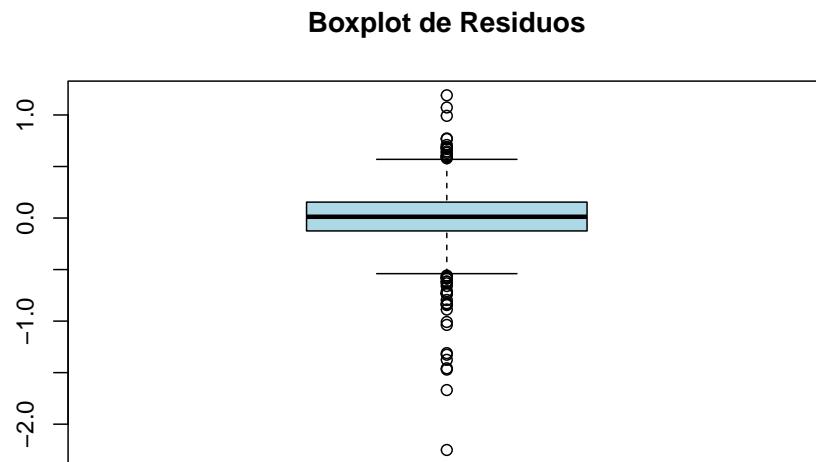


Histograma de Residuos

Histograma de Residuos



Boxplot de Residuos



Pruaba de Normalidad

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: model_multi$residuals  
## D = 0.082087, p-value < 2.2e-16
```

Predicción y Ecualización

```
## RMSE del modelo multivariado en test: 0.3109389
```

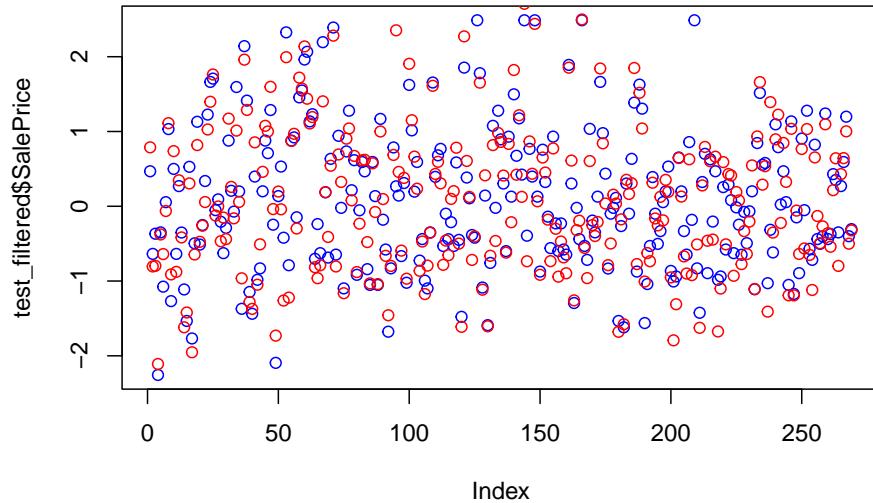


Table 35: Resumen de la diferencia ($\text{SalePrice} - \text{pred_test}$)

Estadística	Valor
Min	-2.0867039
1st Qu.	-0.1465720
Median	0.0396616
Mean	0.0188662
3rd Qu.	0.1961073
Max	0.8403423

Análisis de Errores

Table 36: MSE en Entrenamiento y Prueba

Conjunto	MSE
Entrenamiento	0.077664
Prueba	0.096683

Análisis del Modelo y Detección de Multicolinealidad

- **Multicolinealidad:**

Al ajustar el modelo completo se observa que algunas variables están fuertemente correlacionadas, lo que puede generar coeficientes inestables y altos errores estándar. La matriz de correlación y gráficos de dispersión confirman estas relaciones.

- **Variables Relevantes:**

Mediante análisis de correlación y la significancia en el resumen del modelo, se identificaron aquellas variables que aportan de manera significativa . Sin embargo, la inclusión de todas puede introducir redundancia.

2. Adaptación del Modelo y Sobreajuste

- **Ajuste al Dataset:**

Aunque el modelo completo muestra un R^2 alto , el análisis de residuos evidencia patrones y falta de normalidad en los residuos, lo que sugiere que el modelo no captura toda la complejidad y podría estar sobreajustado.

- **Sobreajuste (Overfitting):**

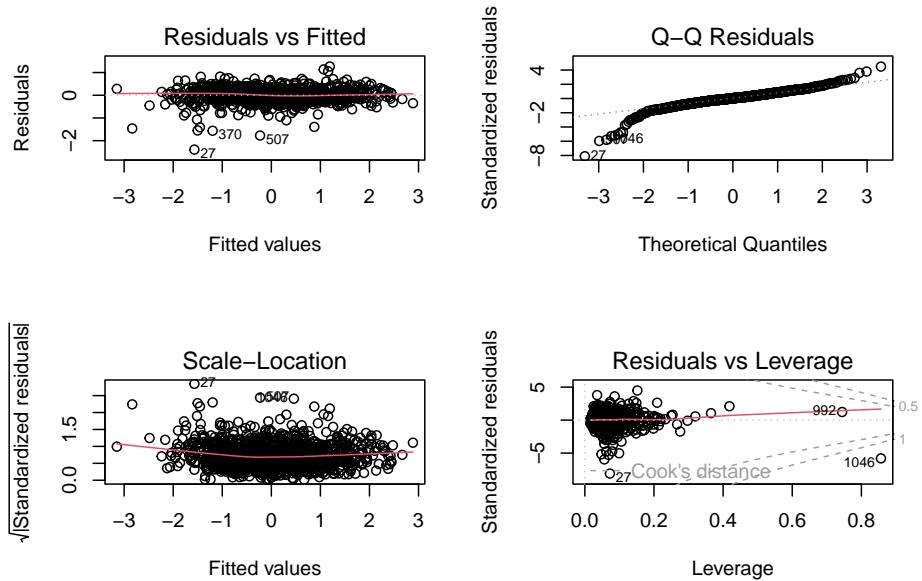
Se observa que el error de entrenamiento es significativamente menor que el error en el conjunto de prueba, lo que indica que el modelo completo podría estar ajustándose demasiado a los datos de entrenamiento.

Modelo Normalizado

Table 37: Resumen del Modelo Lineal Múltiple (Normalizado)

term	estimate	std.error	statistic	p.value
(Intercept)	0.0000000	0.0093587	0.0000000	1.0000000
Id	0.0014228	0.0096848	0.1469071	0.8832349
MSSubClass	-0.0020901	0.0350052	-0.0597096	0.9523988
LotFrontage	-0.0109943	0.0108276	-1.0153935	0.3101634
LotArea	0.1183905	0.0149045	7.9432989	0.0000000
OverallQual	0.2063767	0.0184934	11.1594743	0.0000000
OverallCond	0.1700178	0.0135264	12.5693247	0.0000000
YearBuilt	0.1634302	0.0276263	5.9157404	0.0000000
YearRemodAdd	0.0374163	0.0172298	2.1716070	0.0301193
MasVnrArea	0.0116006	0.0119726	0.9689266	0.3328156
ExterQual	0.0157899	0.0169162	0.9334191	0.3508285
ExterCond	-0.0342621	0.0108323	-3.1629703	0.0016087
BsmtQual	0.0341340	0.0177341	1.9247658	0.0545415
BsmtCond	0.0262421	0.0101909	2.5750500	0.0101650
BsmtFinSF1	-0.3130166	0.0512597	-6.1064907	0.0000000
BsmtFinSF2	-0.1352983	0.0218746	-6.1851898	0.0000000
BsmtUnfSF	-0.3821436	0.0504706	-7.5716124	0.0000000
TotalBsmtSF	0.4681708	0.0452430	10.3479276	0.0000000
HeatingQC	0.0199381	0.0130852	1.5237169	0.1278950
X1stFlrSF	-0.0077839	0.0491610	-0.1583355	0.8742244
X2ndFlrSF	0.0867587	0.0442852	1.9590913	0.0503791
LowQualFinSF	-0.0122019	0.0108397	-1.1256657	0.2605766
GrLivArea	0.2856188	0.0514833	5.5477952	0.0000000
BsmtFullBath	0.0363378	0.0144728	2.5107691	0.0122034
BsmtHalfBath	0.0022000	0.0103444	0.2126718	0.8316262
FullBath	0.0195522	0.0176630	1.1069559	0.2685787
HalfBath	0.0145833	0.0152900	0.9537794	0.3404253
BedroomAbvGr	-0.0318318	0.0154345	-2.0623776	0.0394291
KitchenAbvGr	-0.0537223	0.0139919	-3.8395275	0.0001310
KitchenQual	0.0307017	0.0156741	1.9587528	0.0504188

term	estimate	std.error	statistic	p.value
TotRmsAbvGrd	0.0517248	0.0212034	2.4394529	0.0148821
Fireplaces	0.0507590	0.0123771	4.1010496	0.0000445
GarageYrBlt	0.0127792	0.0214996	0.5943929	0.5523836
GarageCars	0.0675722	0.0201470	3.3539499	0.0008265
GarageArea	0.0058433	0.0210125	0.2780871	0.7810029
GarageQual	0.0200782	0.0135810	1.4784007	0.1396148
GarageCond	0.0123944	0.0131938	0.9394123	0.3477455
WoodDeckSF	0.0255635	0.0107699	2.3736175	0.0178025
OpenPorchSF	0.0096154	0.0109472	0.8783435	0.3799679
EnclosedPorch	0.0169770	0.0111688	1.5200319	0.1288184
X3SsnPorch	0.0120039	0.0097379	1.2326960	0.2179783
ScreenPorch	0.0218698	0.0102681	2.1298858	0.0334242
PoolArea	0.0093848	0.0105046	0.8934018	0.3718566
MiscVal	-0.0130391	0.0101529	-1.2842642	0.1993464
MoSold	-0.0030833	0.0098515	-0.3129782	0.7543624
YrSold	-0.0310120	0.0097853	-3.1692351	0.0015747
Neighborhood.Edwards	-0.0410623	0.0124527	-3.2974692	0.0010099
Neighborhood.Gilbert	-0.0135854	0.0125235	-1.0847885	0.2782763
Neighborhood.NAmes	-0.0157612	0.0169020	-0.9325063	0.3512996
Neighborhood.NridgHt	0.0357290	0.0129606	2.7567433	0.0059441
Neighborhood.NWAmes	-0.0431902	0.0128299	-3.3663753	0.0007906
Neighborhood.OldTown	-0.0324560	0.0159969	-2.0288860	0.0427339
Neighborhood.Other	0.0149091	0.0182813	0.8155406	0.4149567
Neighborhood.Sawyer	-0.0166071	0.0126215	-1.3157833	0.1885477
Neighborhood.Somerst	0.0475430	0.0124052	3.8325123	0.0001347
BldgType.Other	-0.0238711	0.0243774	-0.9792305	0.3277024
BldgType.TwnhsE	-0.0072555	0.0274614	-0.2642076	0.7916743
HouseStyle.1Story	0.0639991	0.0284695	2.2479903	0.0247936
HouseStyle.2Story	-0.0067121	0.0221444	-0.3031058	0.7618721
HouseStyle.Other	0.0497234	0.0172294	2.8859576	0.0039860
Exterior1st.MetalSd	0.0295071	0.0426887	0.6912169	0.4895894
Exterior1st.Other	0.0344814	0.0236778	1.4562725	0.1456306
Exterior1st.Plywood	0.0004949	0.0200924	0.0246306	0.9803545
Exterior1st.VinylSd	0.0585806	0.0569547	1.0285466	0.3039410
Exterior1st.Wd.Sdng	-0.0293295	0.0295969	-0.9909668	0.3219410
Exterior2nd.MetalSd	0.0032078	0.0425564	0.0753778	0.9399292
Exterior2nd.Other	-0.0159655	0.0243355	-0.6560596	0.5119364
Exterior2nd.Plywood	-0.0101323	0.0205608	-0.4927989	0.6222627
Exterior2nd.VinylSd	-0.0338871	0.0571816	-0.5926230	0.5535672
Exterior2nd.Wd.Sdng	0.0304942	0.0287421	1.0609590	0.2889642
TotalArea	NA	NA	NA	NA
HouseAge	NA	NA	NA	NA
TotalBath	NA	NA	NA	NA



Seleccionamos predictores

Table 38: Resumen del Modelo Multivariado (Stepwise Backward)

term	estimate	std.error	statistic	p.value
(Intercept)	-0.0073596	0.0092067	-0.7993791	0.4242553
LotArea	0.1119581	0.0116530	9.6076618	0.0000000
OverallQual	0.2092479	0.0174560	11.9871655	0.0000000
OverallCond	0.1679341	0.0129443	12.9735722	0.0000000
YearBuilt	0.1751872	0.0221931	7.8937764	0.0000000
YearRemodAdd	0.0384440	0.0157751	2.4370045	0.0149783
ExterCond	-0.0296403	0.0103398	-2.8666199	0.0042335
BsmtQual	0.0321587	0.0161957	1.9856323	0.0473395
BsmtCond	0.0252994	0.0098632	2.5650197	0.0104580
BsmtFinSF1	-0.2840501	0.0317148	-8.9563868	0.0000000
BsmtFinSF2	-0.1168373	0.0141334	-8.2667661	0.0000000
BsmtUnfSF	-0.3565092	0.0333038	-10.7047498	0.0000000
TotalBsmtSF	0.4860561	0.0390296	12.4535315	0.0000000
HeatingQC	0.0191216	0.0121764	1.5703806	0.1166343
X2ndFlrSF	0.0992295	0.0280280	3.5403657	0.0004175
GrLivArea	0.2873959	0.0277663	10.3505213	0.0000000
BsmtFullBath	0.0317319	0.0126267	2.5130766	0.0121204
BedroomAbvGr	-0.0308441	0.0140129	-2.2011270	0.0279487
KitchenAbvGr	-0.0718110	0.0154975	-4.6337198	0.0000041
KitchenQual	0.0335308	0.0142987	2.3450255	0.0192152
TotRmsAbvGrd	0.0484207	0.0197502	2.4516589	0.0143854
Fireplaces	0.0466517	0.0113190	4.1215316	0.0000407
GarageCars	0.0840175	0.0149541	5.6183748	0.0000000
GarageQual	0.0259645	0.0091957	2.8235579	0.0048411
WoodDeckSF	0.0235046	0.0095558	2.4597344	0.0140677
EnclosedPorch	0.0145032	0.0103781	1.3974830	0.1625699
ScreenPorch	0.0183550	0.0092910	1.9755711	0.0484705

term	estimate	std.error	statistic	p.value
YrSold	-0.0289262	0.0089190	-3.2431987	0.0012200
Neighborhood.Edwards	-0.0515849	0.0109395	-4.7154663	0.0000027
Neighborhood.Gilbert	-0.0145909	0.0098893	-1.4754129	0.1404078
Neighborhood.NAmes	-0.0228564	0.0109659	-2.0843181	0.0373777
Neighborhood.NridgHt	0.0305630	0.0099271	3.0787296	0.0021340
Neighborhood.NWAmes	-0.0424994	0.0099275	-4.2809856	0.0000203
Neighborhood.OldTown	-0.0389308	0.0116328	-3.3466495	0.0008476
Neighborhood.Sawyer	-0.0210389	0.0100553	-2.0923161	0.0366550
Neighborhood.Somerst	0.0430795	0.0100616	4.2815911	0.0000203
BldgType.Other	-0.0237120	0.0133993	-1.7696516	0.0770816
HouseStyle.1Story	0.0604469	0.0217138	2.7837969	0.0054712
HouseStyle.Other	0.0465417	0.0125238	3.7162639	0.0002131
Exterior1st.MetalSd	0.0306271	0.0117823	2.5994126	0.0094722
Exterior1st.Other	0.0216101	0.0119047	1.8152596	0.0697753
Exterior1st.VinylSd	0.0243981	0.0139548	1.7483663	0.0806991
Exterior1st.Wd.Sdng	-0.0372791	0.0210951	-1.7671919	0.0774928
Exterior2nd.Wd.Sdng	0.0403526	0.0192578	2.0953969	0.0363799

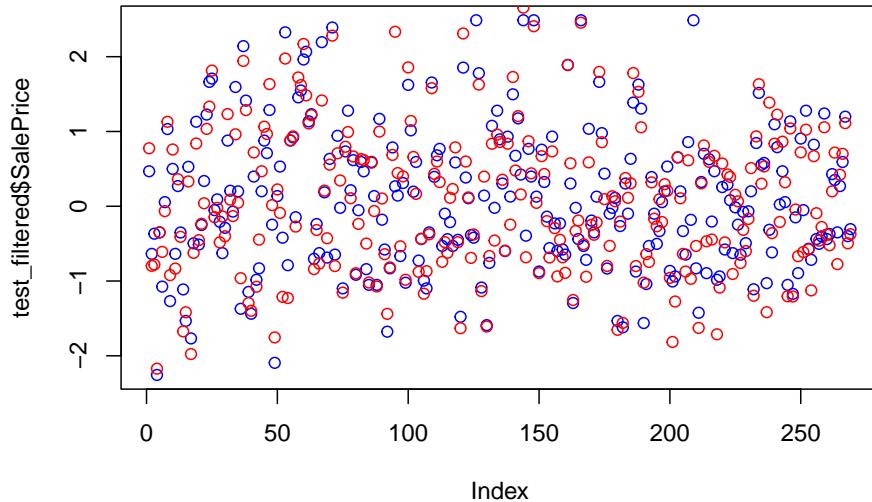
Test de Normalidad

```
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: modelo_mult2$residuals
## D = 0.078014, p-value < 2.2e-16
```

Predicción y Ecualización

Table 39: MSE del Modelo Stepwise Backward en Entrenamiento y Prueba

Conjunto	MSE
Entrenamiento	0.0787642
Prueba	0.0937194



- **Motivación y Efecto de la Normalización:**

- Todas las variables se escalan a media 0 y desviación estándar 1.
- Los coeficientes se vuelven directamente comparables: un coeficiente más grande indica mayor impacto relativo sobre el precio.

- **Principales Hallazgos en los Coeficientes:**

- Variables como *OverallQual*, *OverallCond*, *YearBuilt*, *LotArea* y *TotalBsmtSF* muestran coeficientes positivos y muy significativos.
- Muchas otras variables presentan p-valores altos, sugiriendo que su aporte es menor o está solapado por predictores más fuertes.

- **Residuos y Ajuste del Modelo:**

- Los gráficos de diagnóstico (Residuals vs Fitted, Q-Q) revelan patrones y desviaciones en las colas, indicando no normalidad de residuos.
- El test de Lilliefors confirma p-valor cercano a 0, por lo que se rechaza la hipótesis de normalidad.
- La alta dimensionalidad y varias variables con poca significancia sugieren posible sobreajuste.

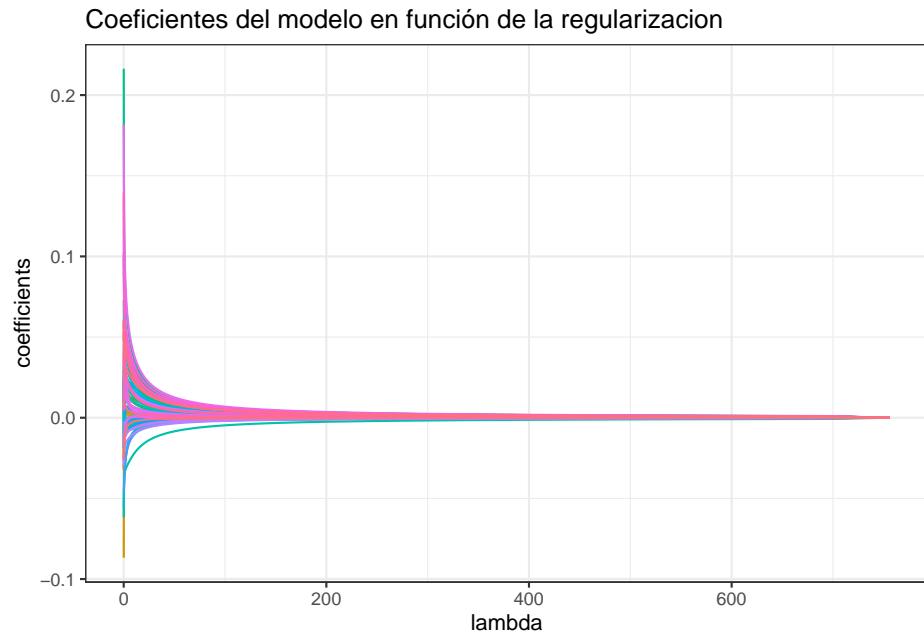
- **Conclusión:**

- El modelo normalizado facilita comparar la importancia relativa de cada predictor, pero no resuelve problemas de residuos ni sobreajuste.

Modelos Regularizados

Modelo Ridge

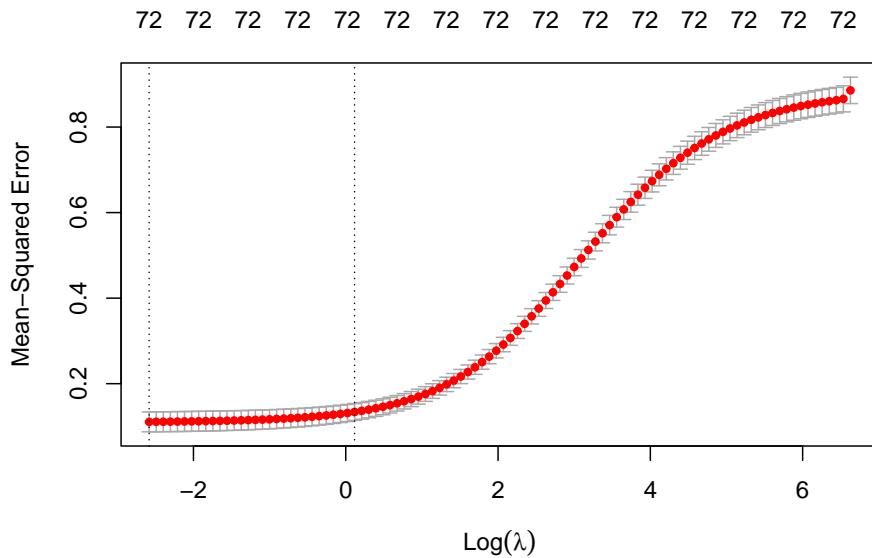
Se crean las matrices de entrenamiento y testeо, y se ajusta un modelo de regresión Ridge. La regularización de los coeficientes evita el sobreajuste y mejorar la generalización del modelo.



Se determina el valor óptimo de lambda mediante validación cruzada y se ajusta el modelo final.

Table 40: Valor óptimo de lambda y MSE en validación cruzada

Lambda_Optimo	MSE_CV
0.0756042	0.1107241



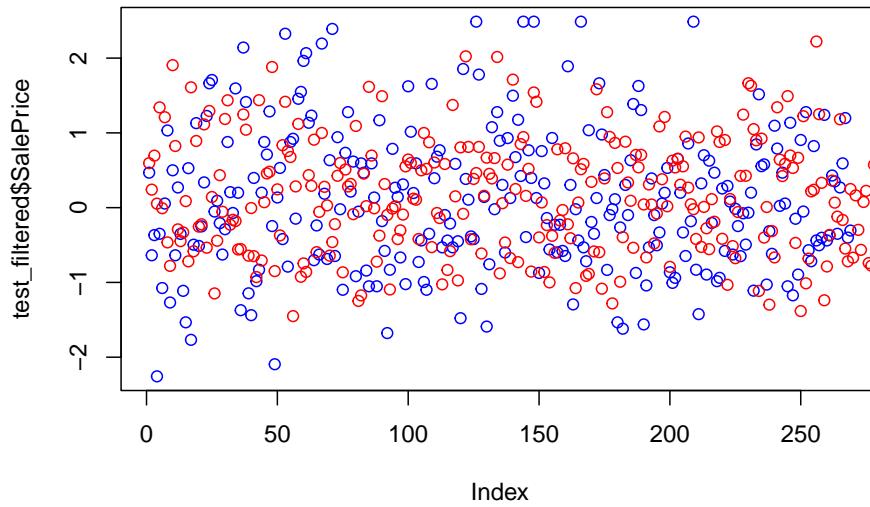
Se establece un modelo con el valor óptimo de lambda y se evalua su desempeño en el conjunto de testeo.

Table 41: Coeficientes del Modelo 3

	Coeficiente	Valor
(Intercept)	(Intercept)	0.0176678
Id	Id	-0.0034016
MSSubClass	MSSubClass	-0.0123911
LotFrontage	LotFrontage	0.0001376
LotArea	LotArea	0.0565166
OverallQual	OverallQual	0.0976979
OverallCond	OverallCond	0.0501548
YearBuilt	YearBuilt	0.0335311
YearRemodAdd	YearRemodAdd	0.0476448
MasVnrArea	MasVnrArea	0.0249511
ExterQual	ExterQual	0.0494185
ExterCond	ExterCond	0.0002527
BsmtQual	BsmtQual	0.0498759
BsmtCond	BsmtCond	0.0235713
BsmtFinSF1	BsmtFinSF1	0.0279078
BsmtFinSF2	BsmtFinSF2	0.0033711
BsmtUnfSF	BsmtUnfSF	-0.0027899
TotalBsmtSF	TotalBsmtSF	0.0638335
HeatingQC	HeatingQC	0.0345806
X1stFlrSF	X1stFlrSF	0.0499370
X2ndFlrSF	X2ndFlrSF	0.0341155
LowQualFinSF	LowQualFinSF	-0.0040830
GrLivArea	GrLivArea	0.0899822
BsmtFullBath	BsmtFullBath	0.0192836
BsmtHalfBath	BsmtHalfBath	-0.0013289
FullBath	FullBath	0.0358573
HalfBath	HalfBath	0.0240396
BedroomAbvGr	BedroomAbvGr	0.0124987
KitchenAbvGr	KitchenAbvGr	-0.0315744
KitchenQual	KitchenQual	0.0543552
TotRmsAbvGrd	TotRmsAbvGrd	0.0499982
Fireplaces	Fireplaces	0.0562081
GarageYrBlt	GarageYrBlt	0.0172847
GarageCars	GarageCars	0.0534889
GarageArea	GarageArea	0.0413511
GarageQual	GarageQual	0.0206244
GarageCond	GarageCond	0.0108605
WoodDeckSF	WoodDeckSF	0.0295109
OpenPorchSF	OpenPorchSF	0.0218718
EnclosedPorch	EnclosedPorch	-0.0015969
X3SsnPorch	X3SsnPorch	0.0082130
ScreenPorch	ScreenPorch	0.0186638
PoolArea	PoolArea	-0.0085506
MiscVal	MiscVal	-0.0070768
MoSold	MoSold	0.0075213
YrSold	YrSold	-0.0117609
Neighborhood.Edwards	Neighborhood.Edwards	-0.0314626
Neighborhood.Gilbert	Neighborhood.Gilbert	-0.0036731
Neighborhood.NAmes	Neighborhood.NAmes	-0.0052843
Neighborhood.NridgHt	Neighborhood.NridgHt	0.0263131
Neighborhood.NWAmes	Neighborhood.NWAmes	-0.0070832

	Coefficiente	Valor
Neighborhood.OldTown	Neighborhood.OldTown	-0.0179719
Neighborhood.Other	Neighborhood.Other	0.0165615
Neighborhood.Sawyer	Neighborhood.Sawyer	-0.0116634
Neighborhood.Somerst	Neighborhood.Somerst	0.0187277
BldgType.Other	BldgType.Other	-0.0296296
BldgType.TwnhsE	BldgType.TwnhsE	-0.0081762
HouseStyle.1Story	HouseStyle.1Story	-0.0095095
HouseStyle.2Story	HouseStyle.2Story	0.0127045
HouseStyle.Other	HouseStyle.Other	0.0034993
Exterior1st.MetalSd	Exterior1st.MetalSd	-0.0000707
Exterior1st.Other	Exterior1st.Other	0.0095248
Exterior1st.Plywood	Exterior1st.Plywood	-0.0004887
Exterior1st.VinylSd	Exterior1st.VinylSd	0.0093361
Exterior1st.Wd.Sdng	Exterior1st.Wd.Sdng	-0.0069127
Exterior2nd.MetalSd	Exterior2nd.MetalSd	-0.0002535
Exterior2nd.Other	Exterior2nd.Other	-0.0048346
Exterior2nd.Plywood	Exterior2nd.Plywood	-0.0017730
Exterior2nd.VinylSd	Exterior2nd.VinylSd	0.0096090
Exterior2nd.Wd.Sdng	Exterior2nd.Wd.Sdng	0.0009504
TotalArea	TotalArea	0.0638241
HouseAge	HouseAge	-0.0340596
TotalBath	TotalBath	0.0474601

Se observan las predicciones y se calcula el error cuadrático medio en el conjunto de testeo.



```
## MSE en prueba: 0.1202691
```

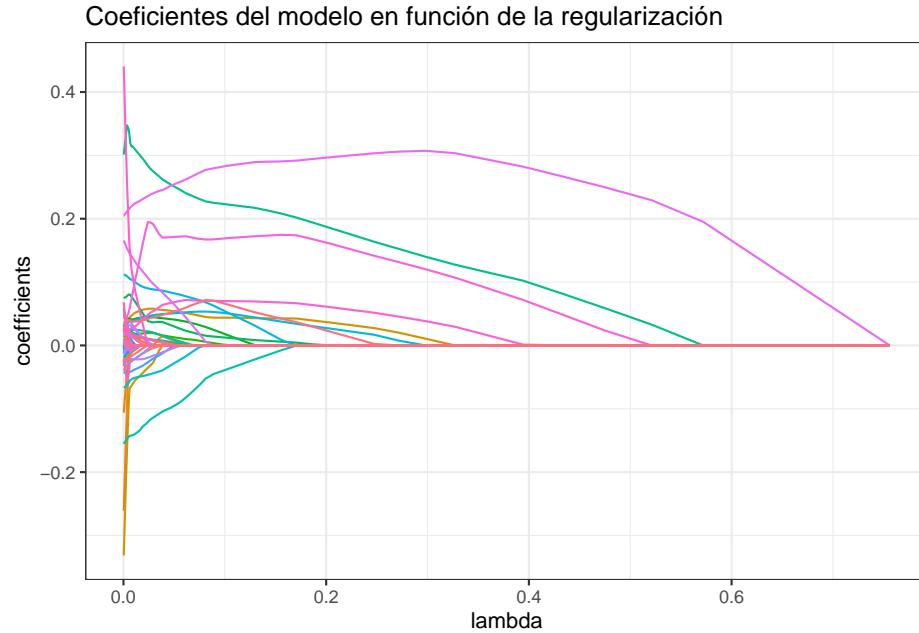
```
## MSE en entrenamiento: 0.1224638
```

- Elección de lambda Óptimo

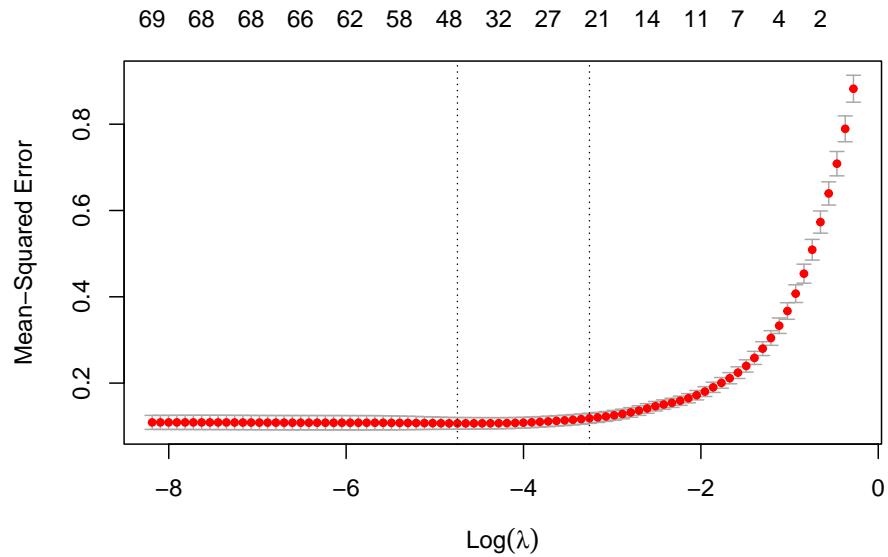
- Se aplicó validación cruzada para determinar el valor de 0.0756, que balancea la complejidad del modelo (penalizando los coeficientes) y su capacidad de predicción.
- Coeficientes Suavizados
 - A diferencia del modelo OLS, Ridge reduce la magnitud de todos los coeficientes sin llegar a anularlos.
 - Variables importantes (e.g., *OverallQual*, *GrLivArea*, *TotalArea*, *TotalBsmtSF*) conservan coeficientes positivos y relativamente altos, pero atenuados frente a un modelo sin penalización.
- Desempeño y Generalización
 - MSE de Entrenamiento: ~0.1225
 - MSE de Prueba: ~0.1203
 - La similitud entre ambos errores indica que el modelo **no está sobreajustado**, y la regularización contribuye a una mejor estabilidad frente a multicolinealidad.
- Conclusiones
 - Ridge reduce la varianza y mejora la robustez del modelo al encoger los coeficientes.
 - El error de prueba y entrenamiento son cercanos, lo que evidencia una buena generalización.

Modelo Lasso

Se construye un modelo Lasso y ajustamos los coeficientes mediante validación cruzada.



Se determina el valor óptimo de lambda mediante validación cruzada y ajustamos el modelo final.



El valor óptimo de lambda y el error cuadrático medio en validación cruzada son:

```
##
## Call: cv.glmnet(x = x_train, y = y_train, type.measure = "mse", nfolds = 10,      alpha = 1, standardize = TRUE)
##
## Measure: Mean-Squared Error
##
##       Lambda Index Measure      SE Nonzero
## min 0.00869     49  0.1067 0.01359      46
## 1se 0.03851     33  0.1184 0.01179      21
```

Ahora ajustamos el modelo con el valor óptimo de lambda y evaluamos su desempeño en el conjunto de prueba.

```
## 73 x 1 sparse Matrix of class "dgCMatrix"
##           s0
## (Intercept) 8.350982e-03
## Id          .
## MSSubClass   .
## LotFrontage  .
## LotArea      8.638218e-02
## OverallQual 2.449294e-01
## OverallCond  8.362104e-02
## YearBuilt    .
## YearRemodAdd 4.742973e-02
## MasVnrArea   .
## ExterQual    1.578677e-02
## ExterCond    .
## BsmtQual    5.620814e-02
## BsmtCond    3.895542e-03
## BsmtFinSF1  1.420147e-02
## BsmtFinSF2  .
```

```

## BsmtUnfSF          .
## TotalBsmtSF         .
## HeatingQC           1.533498e-02
## X1stFlrSF           .
## X2ndFlrSF           .
## LowQualFinSF        .
## GrLivArea            2.618888e-01
## BsmtFullBath         .
## BsmtHalfBath         .
## FullBath             .
## HalfBath             .
## BedroomAbvGr         .
## KitchenAbvGr         -3.941530e-02
## KitchenQual          4.982308e-02
## TotRmsAbvGrd         .
## Fireplaces            4.290521e-02
## GarageYrBlt          .
## GarageCars            3.671420e-02
## GarageArea            .
## GarageQual            7.561251e-03
## GarageCond            .
## WoodDeckSF            5.744336e-03
## OpenPorchSF           .
## EnclosedPorch         .
## X3SsnPorch            .
## ScreenPorch           .
## PoolArea              .
## MiscVal               .
## MoSold                .
## YrSold                .
## Neighborhood.Edwards -1.251079e-02
## Neighborhood.Gilbert .
## Neighborhood.NAmes   .
## Neighborhood.NridgHt  .
## Neighborhood.NWAmes  .
## Neighborhood.OldTown -1.080632e-02
## Neighborhood.Other   .
## Neighborhood.Sawyer  .
## Neighborhood.Somerst .
## BldgType.Other         -4.296169e-06
## BldgType.TwnhsE        .
## HouseStyle.1Story      .
## HouseStyle.2Story      .
## HouseStyle.Other        .
## Exterior1st.MetalSd   .
## Exterior1st.Other       .
## Exterior1st.Plywood    .
## Exterior1st.VinylSd    .
## Exterior1st.Wd.Sdng    .
## Exterior2nd.MetalSd   .
## Exterior2nd.Other       .
## Exterior2nd.Plywood    .
## Exterior2nd.VinylSd    .
## Exterior2nd.Wd.Sdng    .

```

```

## TotalArea          1.705055e-01
## HouseAge          -1.045245e-01
## TotalBath          6.439197e-02

```

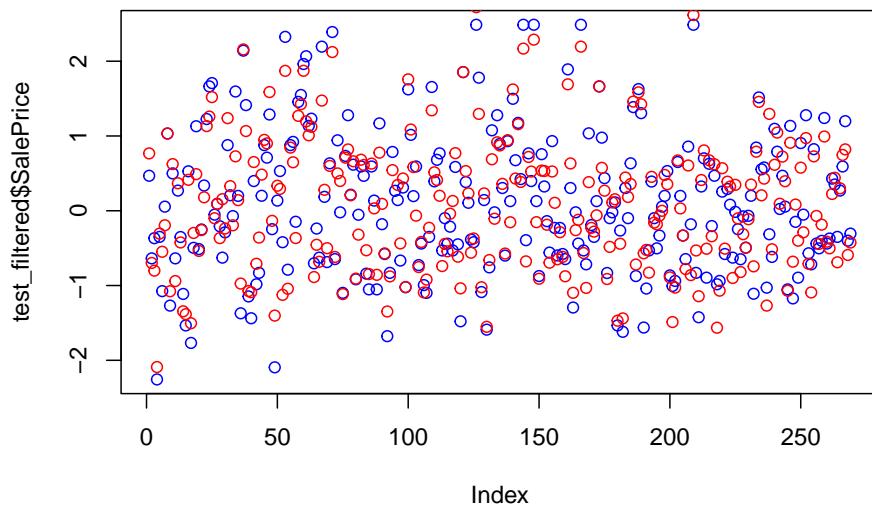
Se determinan los predictores con los coeficientes no nulos y evaluamos el modelo en el conjunto de testeo.

Se obtienen 18 predictores con coeficientes no nulos.

El error cuadrático medio en el conjunto de testeo es de:

```
## MSE en prueba: 0.1079552
```

```
## MSE en entrenamiento: 0.1081771
```



Análisis del Modelo Lasso

- Elección del lambda Óptimo
 - Se usó validación cruzada (10 folds) para estimar el error cuadrático medio (MSE) a diferentes valores de lambda
 - De los resultados, se destacan dos valores de lambda:
 - * lambda minimo: aquel que minimiza el MSE en validación cruzada (~0.00869).
 - * lambda1se: el más sencillo dentro de 1 desviación estándar del mínimo (~0.03851).
- Coeficientes No Nulos
 - Lasso tiende a poner en cero muchos coeficientes, quedándose con un subconjunto reducido de predictores (18 en este caso).
 - Entre las variables seleccionadas sobresalen:
 - * **OverallQual** (coef. ~0.24), **GrLivArea** (~0.26), **TotalArea** (~0.17) y **KitchenQual** (~0.05).

- * Algunas variables importantes en OLS (como *TotalBsmtSF* o *GarageArea*) quedan en cero si su aporte marginal no supera la penalización.

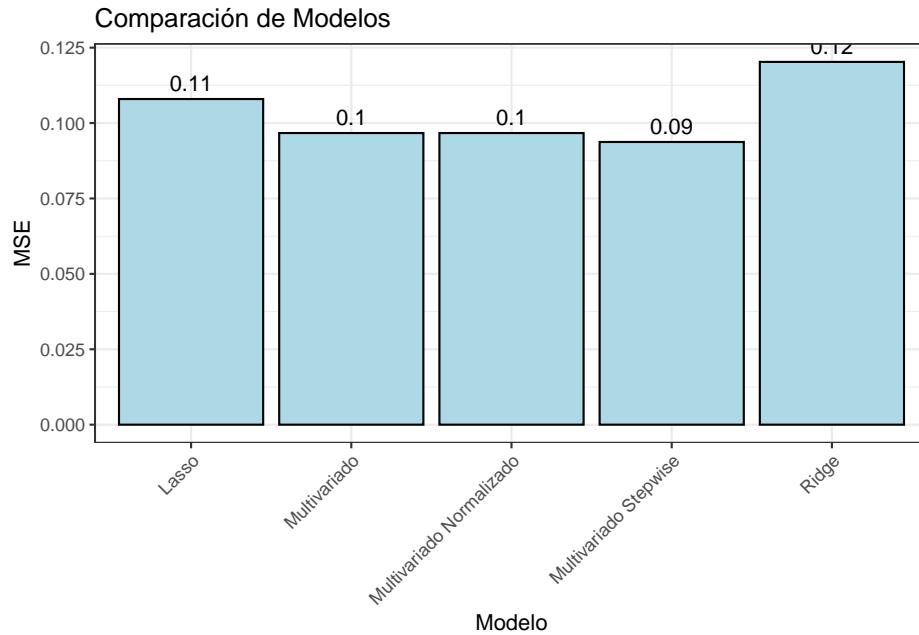
- **Desempeño Predictivo**

- **MSE en Entrenamiento:** ~0.1082
- **MSE en Prueba:** ~0.1080
- La cercanía de estos errores indica buena generalización y ausencia de sobreajuste significativo.
- El modelo final es más simple, pues sólo 18 predictores quedan con coeficientes distintos de cero.

- **Conclusiones**

- Reduce la complejidad del modelo, dejando un número manejable de variables relevantes.
- Error de prueba y entrenamiento similares, lo que evidencia buena capacidad de generalización.
- Algunas variables potencialmente útiles pueden ser forzadas a cero si su efecto es relativamente menor frente a la penalización.
- Puede omitirse información marginalmente significativa.

Comparación de Modelos



Calculo de AIC y BIC para los modelos

```
##           Modelo AIC_or_AICc      BIC
## 1 Lineal Múltiple    444.8552  798.2422
## 2       Stepwise    407.9337  631.9114
## 3        Ridge   -661.9737 -314.1320
## 4       Lasso   -786.8469 -678.3114
```

- **Modelo Stepwise**

- **Ventaja:** El MSE más bajo (0.09) en el conjunto de prueba.
- **Inconveniente:** Su AIC/BIC no es tan bueno como el de los modelos regularizados (Ridge, Lasso).
- **Uso recomendado:** Si la prioridad es la **precisión de predicción** y no se penaliza tanto la complejidad.

- **Modelo Lasso**

- **Ventaja:** Logra un MSE razonablemente bajo (0.11) y la mejor puntuación en AIC/BIC. Además, selecciona un número reducido de variables.
- **Uso recomendado:** Cuando se busca simplicidad y un buen balance entre ajuste y complejidad según criterios de información.

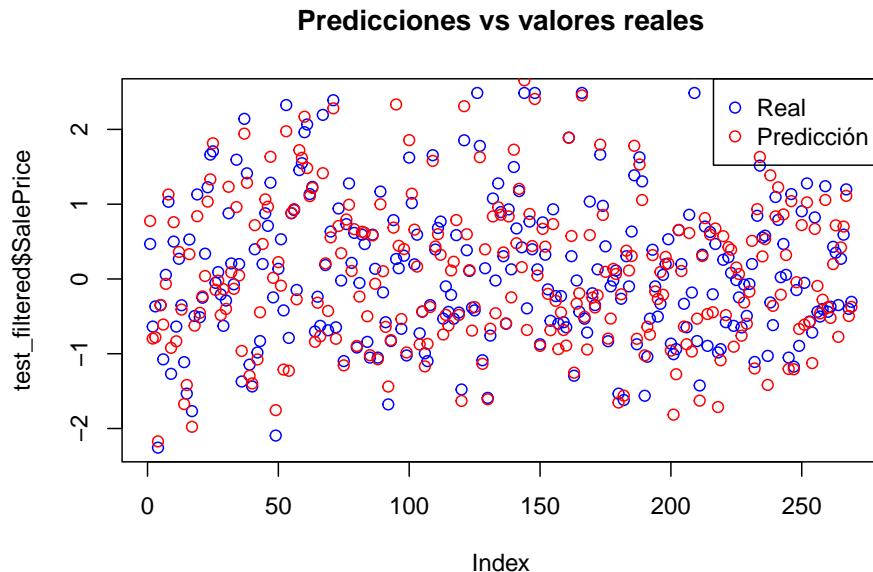
- **Modelo Ridge**

- **Ventaja:** Estabiliza coeficientes frente a multicolinealidad, con MSE de 0.12.
- **Inconveniente:** No alcanza ni el MSE más bajo ni el mejor AIC/BIC.
- **Uso recomendado:** Situaciones con alta correlación entre predictores donde se desee mantener **todos** los coeficientes.

- **Modelo OLS Completo y Normalizado**

- Ambos tienen MSE ~0.10, relativamente buenos, pero un AIC/BIC elevado , lo que sugiere mayor complejidad de la necesaria y potencial riesgo de sobreajuste.

Por ende en el caso de la predicción de precios de casas en el mercado inmobiliario tanto los modelos regularizados y la selección de variables pueden llegar a brindar el equilibrio necesario entre la calidad de interpretar con precisión. En este caso, se usa la predicción del modelo Stepwise dado que fue el que tuvo un mejor desempeño en comparación a los demás.



```
## standardGeneric for "summary" defined from package "base"
##
## function (object, ...)
## standardGeneric("summary")
## <environment: 0x0000021470d28f48>
## Methods may be defined for arguments: object
## Use showMethods(summary) for currently available ones.
```