UNIVERSIDAD DEL VALLE DE GUATEMALA

Minería de Datos

Avances
Proyecto No. 2

Isabella Miralles #22293

Guatemala, 2025

Haga un análisis exploratorio extenso de los datos. Explique bien todos los hallazgos. No ponga solo gráficas y código. Debe llegar a conclusiones interesantes para poder predecir. Explique el preprocesamiento que necesitó hacer.

Descripción de las variables

Los datos para explorar tienen 81 variables que describen diferentes características de las viviendas. Las mas importantes para este análisis son,

- SalePrice
- GrLivArea
- OverallQual
- YearBuilt
- TotalBsmtSF
- LotArea
- GarageCars
- GarageArea

Estas variables son consideradas para evaluar la influencia que tienen en el precio de las viviendas.

Análisis Exploratorio de los datos

Este análisis se realizó con el objetivo de entender la distribución de las variables, identificar los posibles valores atípicos y examinar las relaciones que puedan influir en el precio de las casas.

Revisión general y estadísticas descriptivas

- Dimensiones y tipos de datos
  El dataset tiene 1460 registros y 81 columnas. Se vieron los tipos de datos para diferenciar las variables numéricas de las categorías.
- Resumen estadístico
  Calcular medidas de tendencia central y dispersión para las variables numéricas. Esto ayuda a identificar rangos y detectar posibles inconsistencias en los datos.
- Datos faltantes
  Se identifican las columnas con valores ausentes. La cantidad de datos faltantes varían y se decide imputar o eliminar dichas variables si no aportan información significativa.

Visualización de distribuciones y detección de outliers

- Histogramas y diagramas de densidad
  Generar gráficos de distribución para variables clave
- Boxplots
  Se utilizarán para detectar valores atípicos en variables

Pruebas de Normalidad

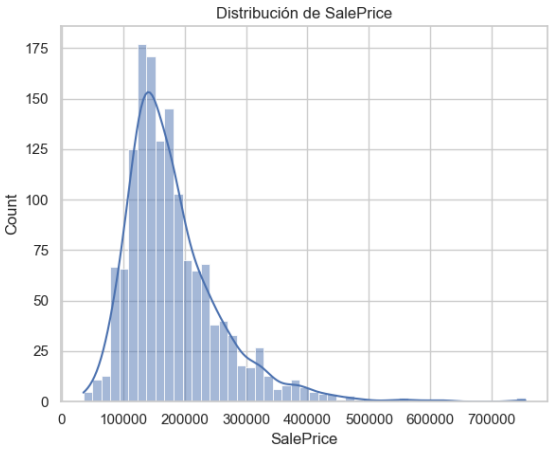
Se utilizarán para evaluar la normalidad de la variable respuesta y otras variables importantes.

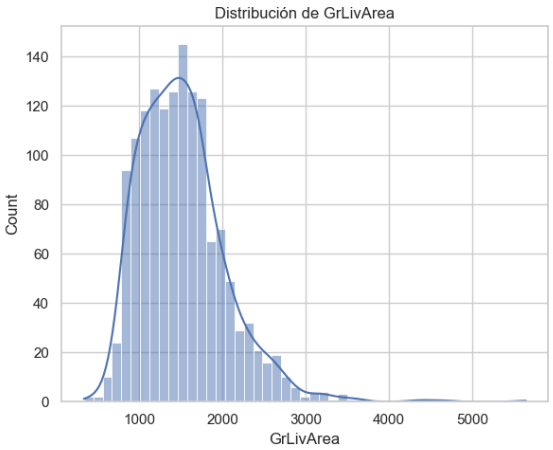
Análisis de correlación y relaciones con la variable respuesta

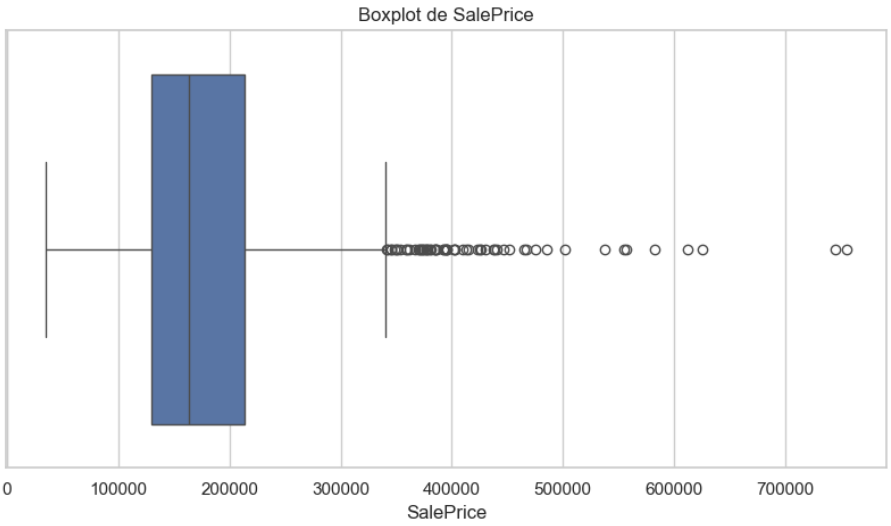
- Matriz de correlación
  Se calculará la matriz de correlación entre las variables numéricas para identificar aquellas que tienen una relación con SalePrice.
- Diagramas de dispersión
  Se realizarán para visualizar la relación entre SalePrice y las variables clave.

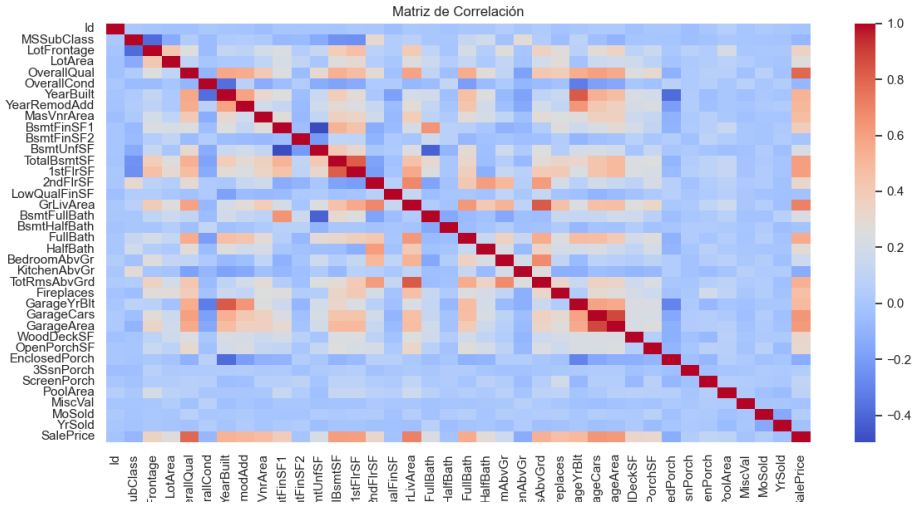
Análisis de variables a incluir en el modelo

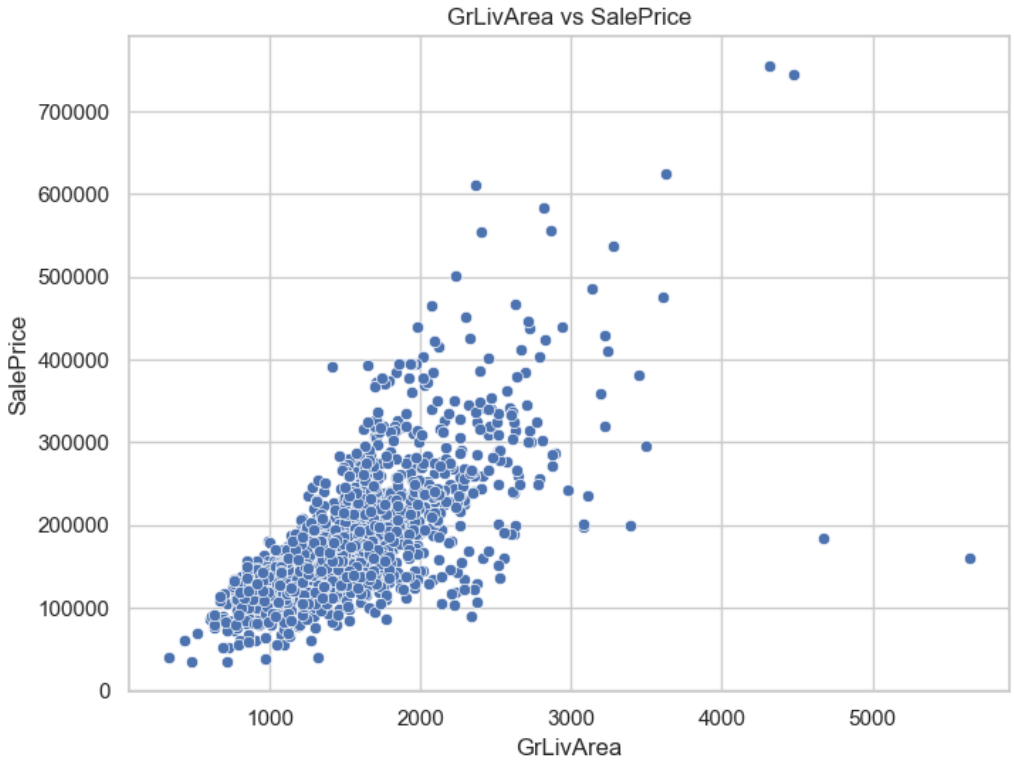
- **Correlación**
  Se priorizan las variables que muestran una correlación alta con el precio.
- **Distribución y normalidad**
  Se realizarán transformaciones a variables que presentan sesgos, así se garantizara que los supuestos de normalidad de los modelos de regresión se cumplan.
- **Relación entre variables**
  Se evaluará la relación con una matriz de correlación, esto para evitar incluir variables que aporten informacion redundante.
- **Análisis gráfico y estadístico**
  Se emplearán técnicas como el análisis de componentes principales para identificar patrones y agrupar variables.


Distribución de SalePrice
Count
SalePrice
0
25
50
75
100
125
150
175
0
100000
200000
300000
400000
500000
600000
700000


Distribución de GrLivArea
Count
GrLivArea
0
20
40
60
80
100
120
140
1000
2000
3000
4000
5000


Boxplot de SalePrice
SalePrice
0
100000
200000
300000
400000
500000
600000
700000


Matriz de Correlación
Id
MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
MasVnrArea
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF
1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageYrBlt
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF
EnclosedPorch
3SsnPorch
ScreenPorch
PoolArea
MiscVal
MoSold
YrSold
SalePrice
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


GrLivArea vs SalePrice
SalePrice
GrLivArea
0
100000
200000
300000
400000
500000
600000
700000
1000
2000
3000
4000
5000

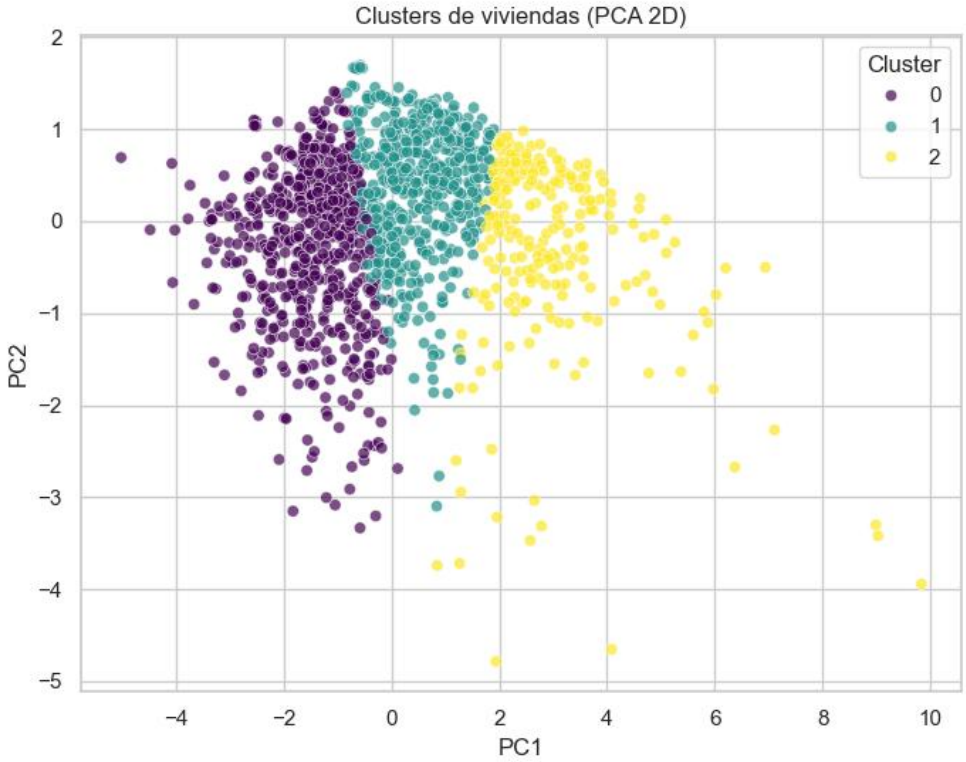
```
PS D:\Documentos\Septimo semestre\Mineria de Datos\Proyecto1-MD> python main.py
### Visión General de los Datos ###
Dimensiones del dataset: (1460, 81)

Tipos de variables:
 Id               int64
MSSubClass        int64
MSZoning         object
LotFrontage     float64
LotArea           int64
                 ...
MoSold            int64
YrSold            int64
SaleType         object
SaleCondition    object
SalePrice         int64
Length: 81, dtype: object

Resumen estadístico:
              Id   MSSubClass  LotFrontage        LotArea  ...       MiscVal       MoSold       YrSold      SalePrice
count  1460.000000  1460.000000  1201.000000    1460.000000  ...   1460.000000  1460.000000  1460.000000    1460.000000
mean    730.500000    56.897260    70.049958   10516.828082  ...     43.489041     6.321918  2007.815753  180921.195890
std     421.610009    42.300571    24.284752    9981.264932  ...    496.123024     2.703626     1.328095   79442.502883
min       1.000000    20.000000    21.000000    1300.000000  ...      0.000000     1.000000  2006.000000   34900.000000
25%     365.750000    20.000000    59.000000    7553.500000  ...      0.000000     5.000000  2007.000000  129975.000000
50%     730.500000    50.000000    69.000000    9478.500000  ...      0.000000     6.000000  2008.000000  163000.000000
75%    1095.250000    70.000000    80.000000   11601.500000  ...      0.000000     8.000000  2009.000000  214000.000000
max    1460.000000   190.000000   313.000000  215245.000000  ...  15500.000000    12.000000  2010.000000  755000.000000
```

```
[8 rows x 38 columns]

Datos faltantes por variable:
 PoolQC          1453
MiscFeature      1406
Alley            1369
Fence            1179
MasVnrType        872
FireplaceQu       690
LotFrontage       259
GarageQual         81
GarageFinish       81
GarageType         81
GarageYrBlt        81
GarageCond         81
BsmtFinType2       38
BsmtExposure       38
BsmtCond           37
BsmtQual           37
BsmtFinType1       37
MasVnrArea          8
Electrical          1
dtype: int64

### Visualizaciones ###

### Preprocesamiento de los Datos ###
Preprocesamiento completado. Nuevas dimensiones: (1460, 290)
PS D:\Documentos\Septimo semestre\Mineria de Datos\Proyecto1-MD>
```

- Variables como GrLivArea y OverallQual muestran una alta correlación con SalePrice.
- La calidad de la construcción y el tamaño de la vivienda son los dos factores que más influyen en el precio de la propiedad.
- El precio es significativamente mas elevado cuando las casas están en vecindarios exclusivos.

- Las propiedades que tienen un garage amplio y mayor cantidad de baños completos tienen un valor mas elevado.
- Se identificaron los valores atípicos que pueden alterar la predicción y se arreglaron.

Incluya un análisis de grupos en el análisis exploratorio. Explique las características de cada uno




```
PS D:\Documentos\Septimo semestre\Mineria de Datos\Proyecto1-MD> python clusterizacion.py
### Análisis Exploratorio Previo ###

### Análisis de Grupos (Clusterización) ###
Medias por Cluster:
             SalePrice    GrLivArea  OverallQual    YearBuilt  TotalBsmtSF
Cluster
0        125835.311550  1217.620061     5.045593  1947.541033   824.721884
1        188474.796099  1575.531915     6.491135  1989.051418  1094.556738
2        315317.336134  2196.567227     8.084034  1994.722689  1612.815126
PS D:\Documentos\Septimo semestre\Mineria de Datos\Proyecto1-MD>
```

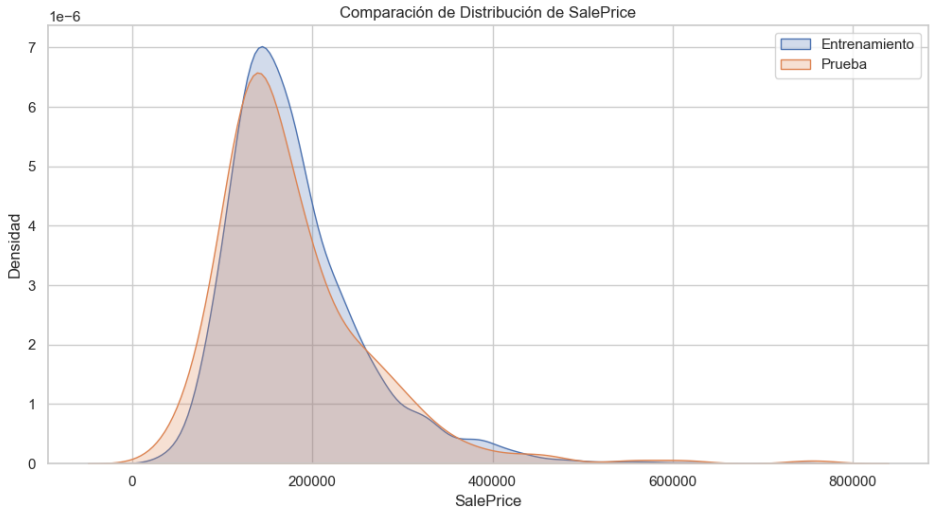
Se utilizo clusters para esta actividad, se eligió un subconjunto de variables numéricas y se imputaron valores faltantes, luego se estandarizaron los datos para que todas las variables tuvieran la misma escala.

Se aplicaron K-means para identificar grupos de viviendas. Aquí se utilizaron 3 clusters.

La dimensión de los datos estandarizados se redujo para poder visualizar los clusters en un gráfico 2D.

En el grafico cada punto representa una vivienda y el color indica el cluster que se le asigno.

Divida el set de datos preprocesados en dos conjuntos: Entrenamiento y prueba. Describa el criterio que usó para crear los conjuntos: número de filas de cada uno, estratificado o no, balanceado o no, etc. Use el conjunto de datos llamado “train.csv”. Extraiga de ahí su subconjunto de prueba




```
### División del Conjunto de Datos ###
Filas en entrenamiento: 1168
Filas en prueba: 292
PS D:\Documentos\Septimo semestre\Mineria de Datos\Proyecto1-MD>
```

El dataset se divide en un 80 % para entrenamiento y un 20 % para prueba, esto asegurando la reproducibilidad con una semilla fija que sería random_state=42. El grafico se incluye para verificar que la distribución es similar en ambos subconjuntos.

Haga ingeniería de características, ¿qué variables cree que puedan ser mejores predictores para el precio de las casas? Explique en que basó la selección o no de las variables

```
### Aplicando Ingeniería de Características ###
Antes de modificar el dataset: (1460, 81)
Después de modificar el dataset: (1460, 123)
```

Variables relevantes

Determinar que variables son mejores para predecir el precio de venta y se analiza su correlación.

Análisis de correlación

Se utilizó la matriz de correlación para identificar las variables numéricas que más influyen en el precio de la venta. Las variables con una mayor correlación positiva son,

- OverallQual
- GrLivArea
- GagrageCars
- GarageArea
- TotalBsmtSF
- 1stFlrSf
- FullBath
- TotRmAbvGrd

Las variables que tienen una correlación cercana a cero no se tomaron en cuenta para la predicción.

Creación de nuevas características

Transformaciones de Variables categóricas

Las variables categóricas que poseen una gran importancia se vuelven en numéricas,

- Codificación ordinal
  Para las variables categóricas con orden lógico como OverallQual, OverallCond, ExterQual, ExterCond, BsmtQual, BsmtCond, KitchenQual, entre otras.
- Para variables nominales como Neighborhood, HouseStyle, RoofStyle, Exterior1st, entre otras.

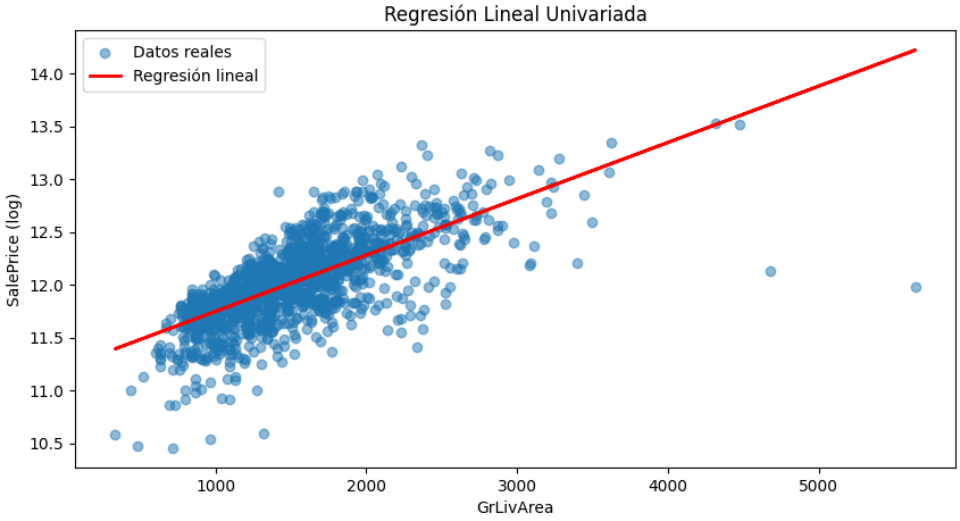
Eliminación de Variables Irrelevantes

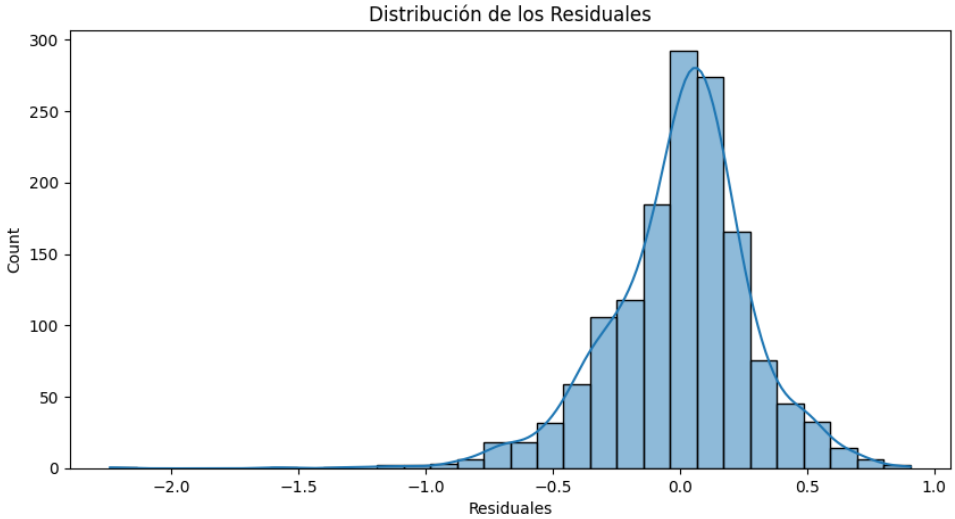
Se eliminan las variables que no aportan información valiosa o que son redundantes.

- Id no es predictivo
- YrSold y MoSold Son útiles si se analizan tendencias temporales, pero para este trabajo no son necesarias
- MiscFeature, Alley, PoolQC, Les hacen falta demasiados valores.
- LowQualFinSF, BsmtFinSF2, MiscVal, entre otras. Son variables con correlación muy baja.

En conclusión, se tiene un conjunto de variables más representativas para el valor de las casas. Las características derivadas obtienen mejor la información relevante y ayudan a mejorar el desempeño del modelo de predicción.

Seleccione una de las variables y haga un modelo univariado de regresión lineal para predecir el precio de las casas. Analice el modelo (resumen, residuos, resultados de la predicción). Muéstrelo gráficamente

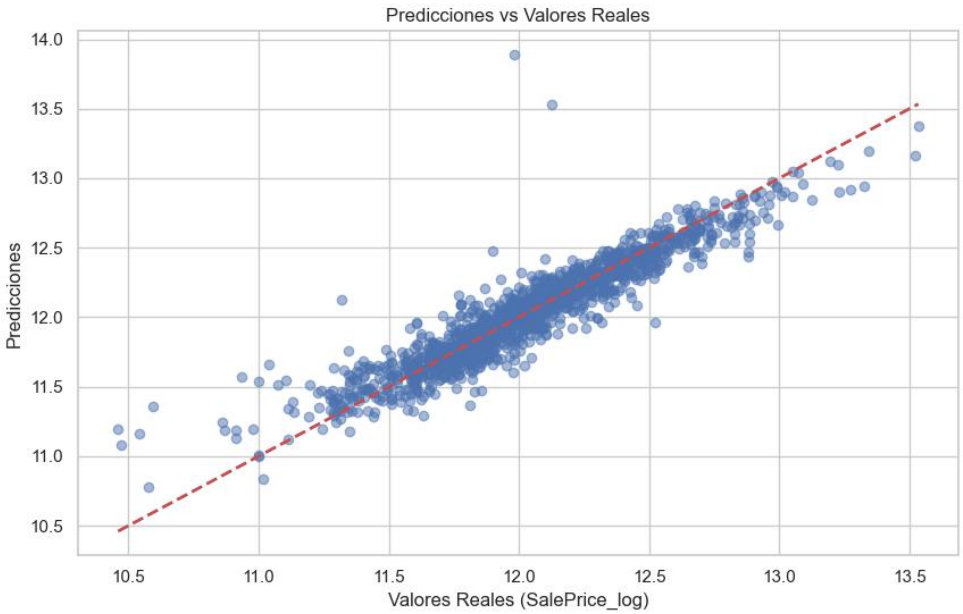


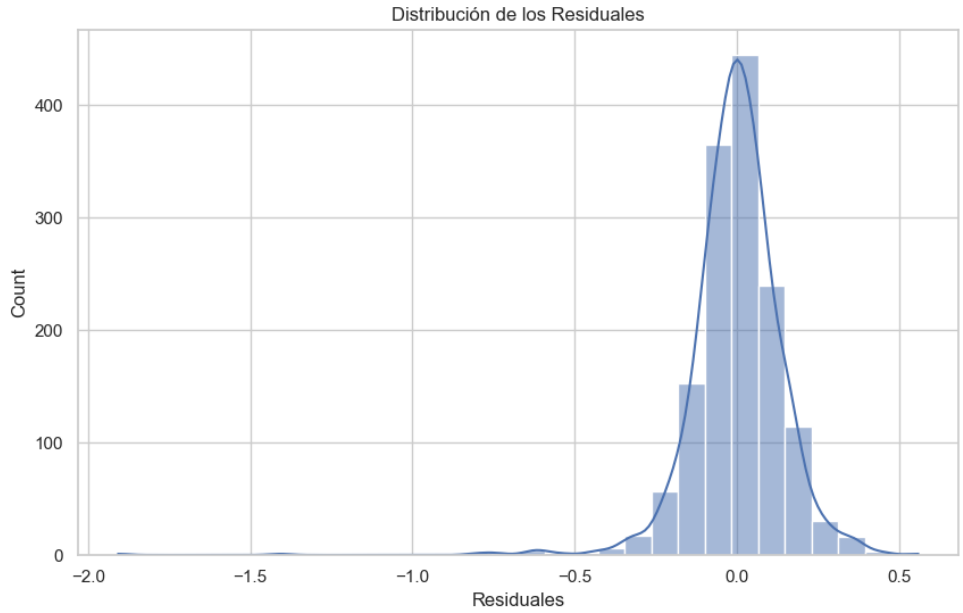


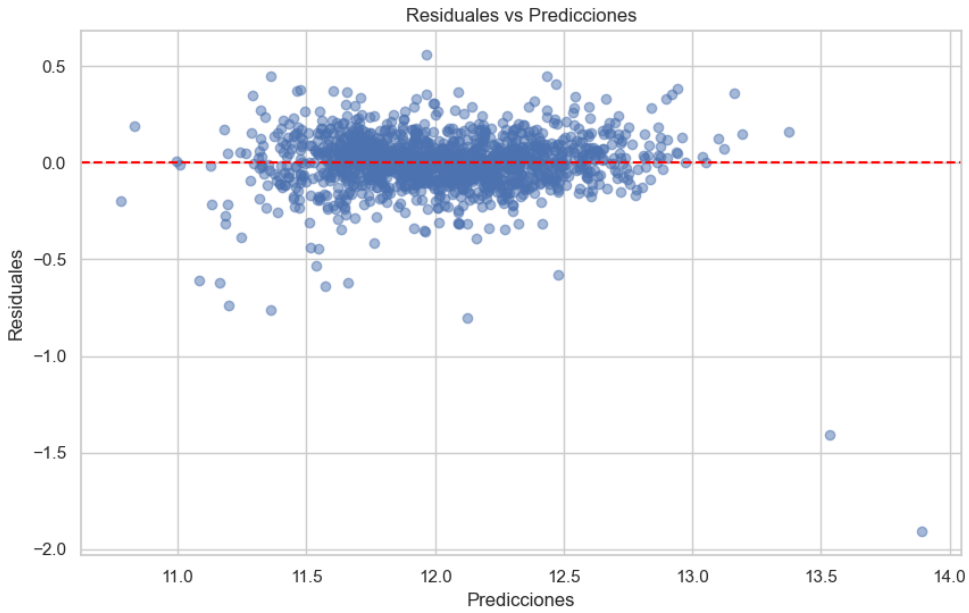
```
### Regression Model ###
                            OLS Regression Results
==============================================================================
Dep. Variable:          SalePrice_log   R-squared:                       0.491
Model:                            OLS   Adj. R-squared:                  0.491
Method:                 Least Squares   F-statistic:                     1408.
Date:                Mon, 03 Mar 2025   Prob (F-statistic):          3.06e-216
Time:                        14:44:35   Log-Likelihood:                -237.96
No. Observations:                1460   AIC:                             479.9
Df Residuals:                    1458   BIC:                             490.5
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         11.2166      0.023    492.511      0.000      11.172      11.261
GrLivArea      0.0005   1.42e-05     37.525      0.000       0.001       0.001
==============================================================================
Omnibus:                      284.575   Durbin-Watson:                   2.012
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1190.093
Skew:                          -0.876   Prob(JB):                    3.76e-259
Kurtosis:                       7.061   Cond. No.                     4.90e+03
==============================================================================
```

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.9e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
MSE: 0.0811
R²: 0.4913
```

Haga un modelo de regresión lineal con todas las variables numéricas para predecir el precio de las casas. Analice el modelo (resumen, residuos, resultados de la predicción). Muestre el modelo gráficamente.







Modelo de Regresión Lineal Multivariada con Todas las Variables Numéricas

OLS Regression Results

Dep. Variable:	SalePrice_log	R-squared:	0.868
Model:	OLS	Adj. R-squared:	0.865
Method:	Least Squares	F-statistic:	268.4
Date:	Fri, 07 Mar 2025	Prob (F-statistic):	0.00
Time:	19:43:50	Log-Likelihood:	748.81
No. Observations:	1460	AIC:	-1426.
Df Residuals:	1424	BIC:	-1235.
Df Model:	35		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	16.9944	5.968	2.847	0.004	5.287	28.702
Id	-4.726e-06	9.23e-06	-0.512	0.609	-2.28e-05	1.34e-05
MSSubClass	-0.0006	0.000	-5.506	0.000	-0.001	-0.000
LotFrontage	-0.0001	0.000	-0.687	0.492	-0.001	0.000
LotArea	1.867e-06	4.31e-07	4.329	0.000	1.02e-06	2.71e-06
OverallQual	0.0844	0.005	16.815	0.000	0.075	0.094
OverallCond	0.0486	0.004	11.148	0.000	0.040	0.057
YearBuilt	0.0030	0.000	10.526	0.000	0.002	0.004
YearRemodAdd	0.0011	0.000	3.953	0.000	0.001	0.002
MasVnrArea	5.072e-07	2.51e-05	0.020	0.984	-4.87e-05	4.97e-05
BsmtFinSF1	2.838e-05	1.07e-05	2.657	0.008	7.43e-06	4.93e-05
BsmtFinSF2	1.646e-05	1.9e-05	0.865	0.387	-2.09e-05	5.38e-05
BsmtUnfSF	4.735e-06	1.02e-05	0.465	0.642	-1.52e-05	2.47e-05
TotalBsmtSF	4.958e-05	1.42e-05	3.487	0.001	2.17e-05	7.75e-05

1stFlrSF	6.042e-05	2.62e-05	2.306	0.021	9.03e-06	0.000
2ndFlrSF	3.452e-05	2.42e-05	1.427	0.154	-1.29e-05	8.2e-05
LowQualFinSF	4.06e-05	6.33e-05	0.642	0.521	-8.35e-05	0.000
GrLivArea	0.0001	2.41e-05	5.622	0.000	8.82e-05	0.000
BsmtFullBath	0.0635	0.011	5.761	0.000	0.042	0.085
BsmtHalfBath	0.0191	0.017	1.105	0.269	-0.015	0.053
FullBath	0.0405	0.012	3.377	0.001	0.017	0.064
HalfBath	0.0222	0.011	1.971	0.049	0.000	0.044
BedroomAbvGr	-0.0021	0.007	-0.298	0.766	-0.016	0.012
KitchenAbvGr	-0.0507	0.022	-2.302	0.021	-0.094	-0.007
TotRmsAbvGrd	0.0157	0.005	3.006	0.003	0.005	0.026
Fireplaces	0.0450	0.008	6.000	0.000	0.030	0.060
GarageYrBlt	-0.0002	0.000	-0.847	0.397	-0.001	0.000
GarageCars	0.0663	0.012	5.461	0.000	0.043	0.090
GarageArea	3.238e-05	4.2e-05	0.771	0.441	-5e-05	0.000
WoodDeckSF	0.0001	3.39e-05	3.649	0.000	5.71e-05	0.000
OpenPorchSF	-3.387e-05	6.41e-05	-0.528	0.597	-0.000	9.19e-05
EnclosedPorch	0.0002	7.12e-05	2.364	0.018	2.86e-05	0.000
3SsnPorch	0.0002	0.000	1.647	0.100	-4.18e-05	0.000
ScreenPorch	0.0004	7.26e-05	4.991	0.000	0.000	0.001
PoolArea	-0.0004	0.000	-3.713	0.000	-0.001	-0.000
MiscVal	-3.565e-06	7.83e-06	-0.455	0.649	-1.89e-05	1.18e-05
MoSold	0.0004	0.001	0.260	0.795	-0.002	0.003
YrSold	-0.0070	0.003	-2.367	0.018	-0.013	-0.001

Omnibus:	1017.649	Durbin-Watson:	1.981
Prob(Omnibus):	0.000	Jarque-Bera (JB):	51369.392
Skew:	-2.646	Prob(JB):	0.00
Kurtosis:	31.573	Cond. No.	1.30e+16

Omnibus:	1017.649	Durbin-Watson:	1.981
Prob(Omnibus):	0.000	Jarque-Bera (JB):	51369.392
Skew:	-2.646	Prob(JB):	0.00
Kurtosis:	31.573	Cond. No.	1.30e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.93e-21. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

MSE: 0.0210
R^2: 0.8684

Analice el modelo. Determine si hay multicolinealidad entre las variables, y cuáles son las que aportan al modelo. Haga un análisis de correlación de las características del modelo y especifique si el modelo se adapta bien a los datos. Explique si hay sobreajuste (overfitting) o no. En caso de existir sobreajuste, haga otro modelo que lo corrija.

```
PS D:\Documentos\Septimo semestre\Mineria de Datos\Proyecto1-MD> python regression_model.py
### Modelo de Regresión Lineal Multivariada ###
                            OLS Regression Results
==============================================================================
Dep. Variable:          SalePrice_log   R-squared:                       0.868
Model:                            OLS   Adj. R-squared:                  0.865
Method:                 Least Squares   F-statistic:                     268.4
Date:                Fri, 07 Mar 2025   Prob (F-statistic):               0.00
Time:                        21:06:41   Log-Likelihood:                748.81
No. Observations:                1460   AIC:                            -1426.
Df Residuals:                    1424   BIC:                            -1235.
Df Model:                          35
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         16.9944      5.968      2.847      0.004       5.287      28.702
Id         -4.726e-06   9.23e-06     -0.512      0.609   -2.28e-05    1.34e-05
MSSubClass    -0.0006      0.000     -5.506      0.000      -0.001      -0.000
LotFrontage   -0.0001      0.000     -0.687      0.492      -0.001       0.000
LotArea     1.867e-06   4.31e-07      4.329      0.000    1.02e-06    2.71e-06
OverallQual    0.0844      0.005     16.815      0.000       0.075       0.094
OverallCond    0.0486      0.004     11.148      0.000       0.040       0.057
YearBuilt      0.0030      0.000     10.526      0.000       0.002       0.004
YearRemodAdd   0.0011      0.000      3.953      0.000       0.001       0.002
MasVnrArea  5.072e-07   2.51e-05      0.020      0.984   -4.87e-05    4.97e-05
BsmtFinSF1  2.838e-05   1.07e-05      2.657      0.008    7.43e-06    4.93e-05
BsmtFinSF2  1.646e-05    1.9e-05      0.865      0.387   -2.09e-05    5.38e-05
BsmtUnfSF   4.735e-06   1.02e-05      0.465      0.642   -1.52e-05    2.47e-05
```

TotalBsmtSF	4.958e-05	1.42e-05	3.487	0.001	2.17e-05	7.75e-05
1stFlrSF	6.042e-05	2.62e-05	2.306	0.021	9.03e-06	0.000
2ndFlrSF	3.452e-05	2.42e-05	1.427	0.154	-1.29e-05	8.2e-05
LowQualFinSF	4.06e-05	6.33e-05	0.642	0.521	-8.35e-05	0.000
GrLivArea	0.0001	2.41e-05	5.622	0.000	8.82e-05	0.000
BsmtFullBath	0.0635	0.011	5.761	0.000	0.042	0.085
BsmtHalfBath	0.0191	0.017	1.105	0.269	-0.015	0.053
FullBath	0.0405	0.012	3.377	0.001	0.017	0.064
HalfBath	0.0222	0.011	1.971	0.049	0.000	0.044
BedroomAbvGr	-0.0021	0.007	-0.298	0.766	-0.016	0.012
KitchenAbvGr	-0.0507	0.022	-2.302	0.021	-0.094	-0.007
TotRmsAbvGrd	0.0157	0.005	3.006	0.003	0.005	0.026
Fireplaces	0.0450	0.008	6.000	0.000	0.030	0.060
GarageYrBlt	-0.0002	0.000	-0.847	0.397	-0.001	0.000
GarageCars	0.0663	0.012	5.461	0.000	0.043	0.090
GarageArea	3.238e-05	4.2e-05	0.771	0.441	-5e-05	0.000
WoodDeckSF	0.0001	3.39e-05	3.649	0.000	5.71e-05	0.000
OpenPorchSF	-3.387e-05	6.41e-05	-0.528	0.597	-0.000	9.19e-05
EnclosedPorch	0.0002	7.12e-05	2.364	0.018	2.86e-05	0.000
3SsnPorch	0.0002	0.000	1.647	0.100	-4.18e-05	0.000
ScreenPorch	0.0004	7.26e-05	4.991	0.000	0.000	0.001
PoolArea	-0.0004	0.000	-3.713	0.000	-0.001	-0.000
MiscVal	-3.565e-06	7.83e-06	-0.455	0.649	-1.89e-05	1.18e-05
MoSold	0.0004	0.001	0.260	0.795	-0.002	0.003
YrSold	-0.0070	0.003	-2.367	0.018	-0.013	-0.001

==========

Omnibus:	1017.649	Durbin-Watson:	1.981
Prob(Omnibus):	0.000	Jarque-Bera (JB):	51369.392
Skew:	-2.646	Prob(JB):	0.00
Kurtosis:	31.573	Cond. No.	1.30e+16

==========

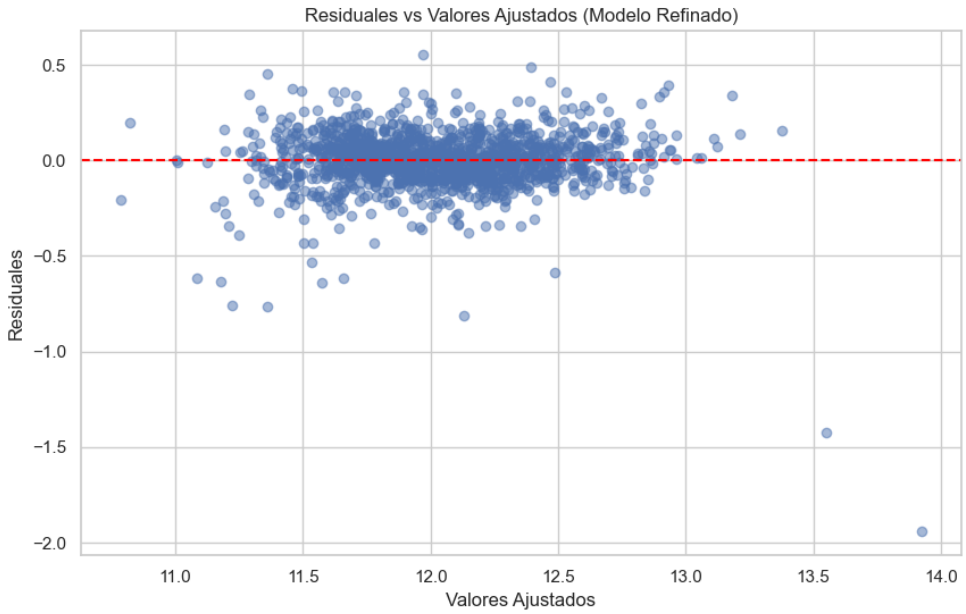
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.93e-21. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

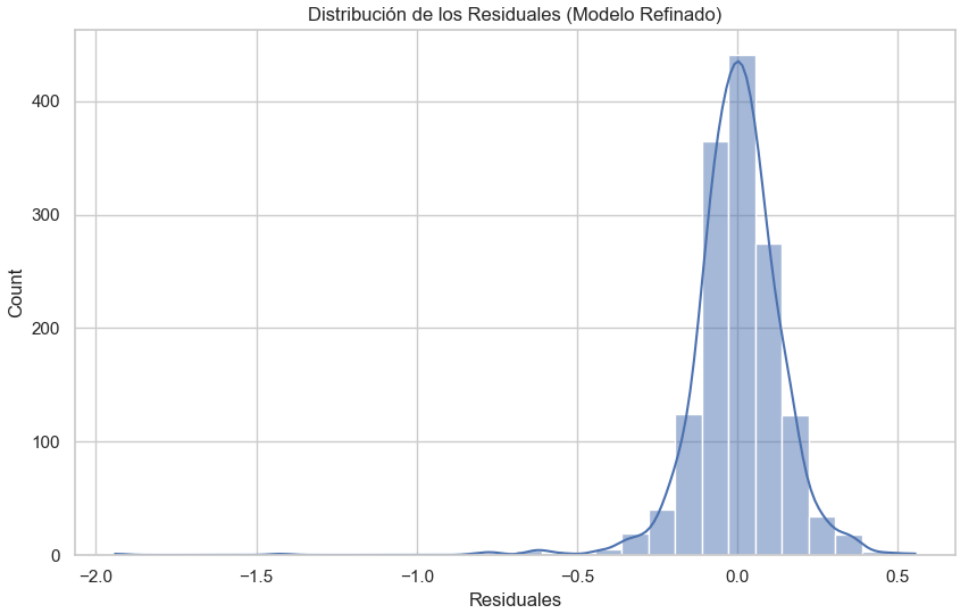
```
C:\Users\distelsa\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.9_qbz5n2kfra8p0\LocalCache\local-packages\Python39\site-packages\statsmodels\stats\outliers_influence.py:197: RuntimeWarning: divide by zero encountered in scalar divide
  vif = 1. / (1. - r_squared_i)
```

```
VIF para las variables:
         Variable           VIF
0           const  2.416497e+06
1              Id  1.026945e+00
2      MSSubClass  1.657272e+00
3     LotFrontage  1.568471e+00
4         LotArea  1.255983e+00
5     OverallQual  3.264369e+00
6     OverallCond  1.596944e+00
7       YearBuilt  5.008877e+00
8    YearRemodAdd  2.425908e+00
9      MasVnrArea  1.394015e+00
10     BsmtFinSF1           inf
11     BsmtFinSF2           inf
12      BsmtUnfSF           inf
13    TotalBsmtSF           inf
14       1stFlrSF           inf
15       2ndFlrSF           inf
16   LowQualFinSF           inf
17      GrLivArea           inf
18   BsmtFullBath  2.219862e+00
19   BsmtHalfBath  1.153241e+00
20       FullBath  2.951727e+00
21       HalfBath  2.168018e+00
22   BedroomAbvGr  2.329373e+00
23   KitchenAbvGr  1.597388e+00
24   TotRmsAbvGrd  4.889900e+00
25     Fireplaces  1.585984e+00
```

```
26    GarageYrBlt  3.314130e+00
27     GarageCars  5.586220e+00
28     GarageArea  5.460000e+00
29     WoodDeckSF  1.220581e+00
30    OpenPorchSF  1.222736e+00
31  EnclosedPorch  1.283853e+00
32      3SsnPorch  1.025865e+00
33    ScreenPorch  1.110277e+00
34       PoolArea  1.110201e+00
35        MiscVal  1.023410e+00
36         MoSold  1.050896e+00
37         YrSold  1.052044e+00

MSE: 0.0210
R²: 0.8684
Evaluación con Linear Regression:
Entrenamiento: MSE = 0.0207, R² = 0.8639
Prueba: MSE = 0.0231, R² = 0.8764
No se detecta sobreajuste significativo.
```

Si tiene multicolinealidad o sobreajuste, haga un modelo con las variables que sean mejores predictoras del precio de las casas. Determine la calidad del modelo realizando un análisis de los residuos. Muéstrelo gráficamente.





```
### Modelo Refinado con las Mejores Variables Predictoras ###
Eliminando MasVnrArea (p-value = 0.9839)
Eliminando MoSold (p-value = 0.7949)
Eliminando BedroomAbvGr (p-value = 0.7736)
Eliminando BsmtUnfSF (p-value = 0.6562)
Eliminando MiscVal (p-value = 0.6543)
Eliminando BsmtFinSF2 (p-value = 0.6547)
Eliminando OpenPorchSF (p-value = 0.6164)
Eliminando Id (p-value = 0.6080)
Eliminando LowQualFinSF (p-value = 0.4972)
Eliminando 2ndFlrSF (p-value = 0.9080)
Eliminando LotFrontage (p-value = 0.4869)
Eliminando GarageArea (p-value = 0.4873)
Eliminando GarageYrBlt (p-value = 0.4552)
Eliminando BsmtHalfBath (p-value = 0.2178)
Eliminando 1stFlrSF (p-value = 0.2061)
Eliminando HalfBath (p-value = 0.1223)
Eliminando 3SsnPorch (p-value = 0.0777)

Variables seleccionadas: ['MSSubClass', 'LotArea', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd', 'BsmtFin
SF1', 'TotalBsmtSF', 'GrLivArea', 'BsmtFullBath', 'FullBath', 'KitchenAbvGr', 'TotRmsAbvGrd', 'Fireplaces', 'GarageCar
s', 'WoodDeckSF', 'EnclosedPorch', 'ScreenPorch', 'PoolArea', 'YrSold']
                            OLS Regression Results
```

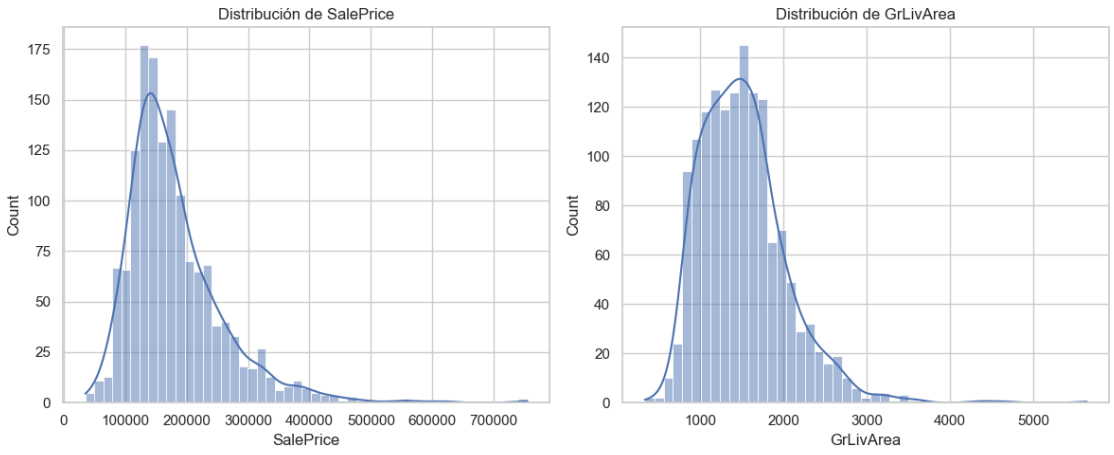
OLS Regression Results

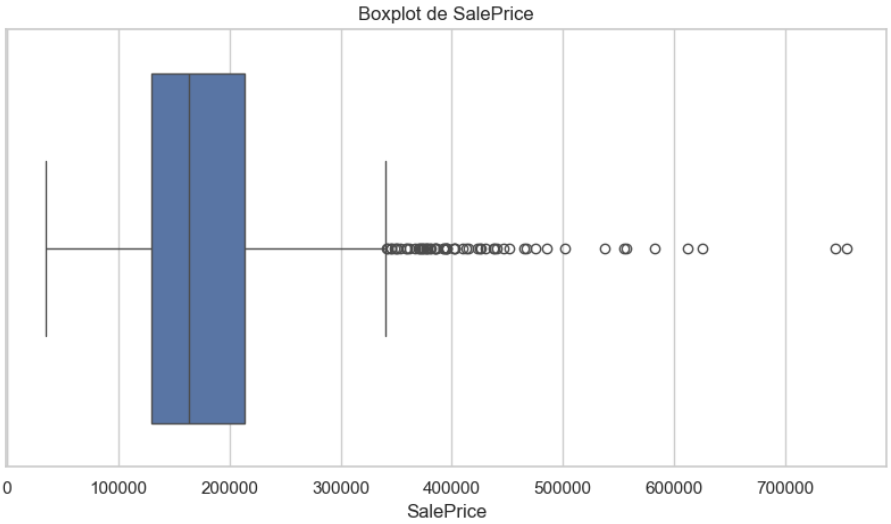
Dep. Variable:	SalePrice_log	R-squared:	0.867
Model:	OLS	Adj. R-squared:	0.865
Method:	Least Squares	F-statistic:	470.3
Date:	Fri, 07 Mar 2025	Prob (F-statistic):	0.00
Time:	21:22:27	Log-Likelihood:	743.10
No. Observations:	1460	AIC:	-1444.
Df Residuals:	1439	BIC:	-1333.
Df Model:	20		
Covariance Type:	nonrobust		

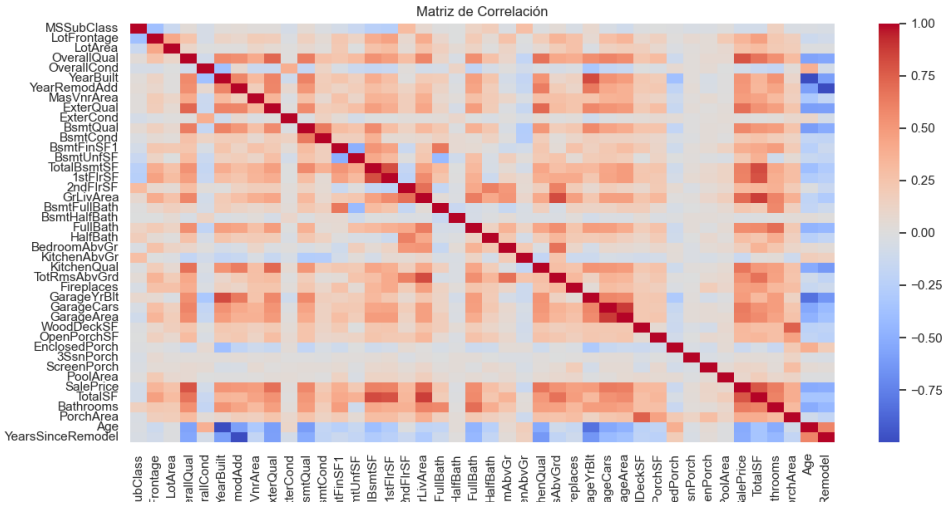
	coef	std err	t	P>|t|	[0.025	0.975]
const	16.4857	5.865	2.811	0.005	4.982	27.990
MSSubClass	-0.0006	0.000	-6.080	0.000	-0.001	-0.000
LotArea	1.872e-06	4.22e-07	4.438	0.000	1.04e-06	2.7e-06
OverallQual	0.0836	0.005	17.140	0.000	0.074	0.093
OverallCond	0.0498	0.004	11.753	0.000	0.041	0.058
YearBuilt	0.0030	0.000	12.724	0.000	0.003	0.004
YearRemodAdd	0.0011	0.000	4.021	0.000	0.001	0.002
BsmtFinSF1	2.68e-05	1.28e-05	2.099	0.036	1.75e-06	5.18e-05
TotalBsmtSF	6.033e-05	1.32e-05	4.574	0.000	3.45e-05	8.62e-05
GrLivArea	0.0002	1.73e-05	11.043	0.000	0.000	0.000
BsmtFullBath	0.0606	0.010	6.001	0.000	0.041	0.080
FullBath	0.0301	0.011	2.786	0.005	0.009	0.051
KitchenAbvGr	-0.0485	0.021	-2.303	0.021	-0.090	-0.007
TotRmsAbvGrd	0.0141	0.005	3.069	0.002	0.005	0.023
Fireplaces	0.0477	0.007	6.602	0.000	0.034	0.062

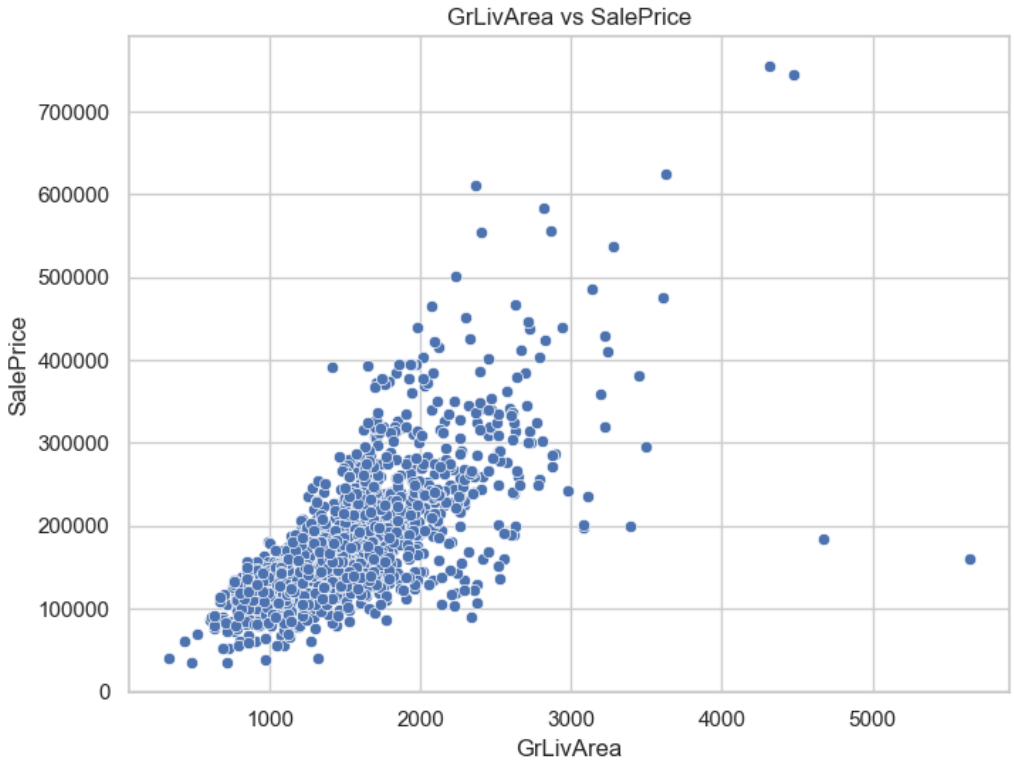
```
GarageCars       0.0744     0.007     10.521     0.000      0.060      0.088
WoodDeckSF       0.0001  3.33e-05      3.779     0.000   6.05e-05      0.000
EnclosedPorch    0.0002  7.06e-05      2.317     0.021   2.51e-05      0.000
ScreenPorch      0.0004  7.19e-05      5.146     0.000      0.000      0.001
PoolArea        -0.0004  9.87e-05     -4.010     0.000     -0.001     -0.000
YrSold          -0.0070     0.003     -2.397     0.017     -0.013     -0.001
==============================================================================
Omnibus:                     1043.613   Durbin-Watson:                   1.974
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            56474.226
Skew:                          -2.730   Prob(JB):                         0.00
Kurtosis:                      32.976   Cond. No.                     2.26e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.26e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

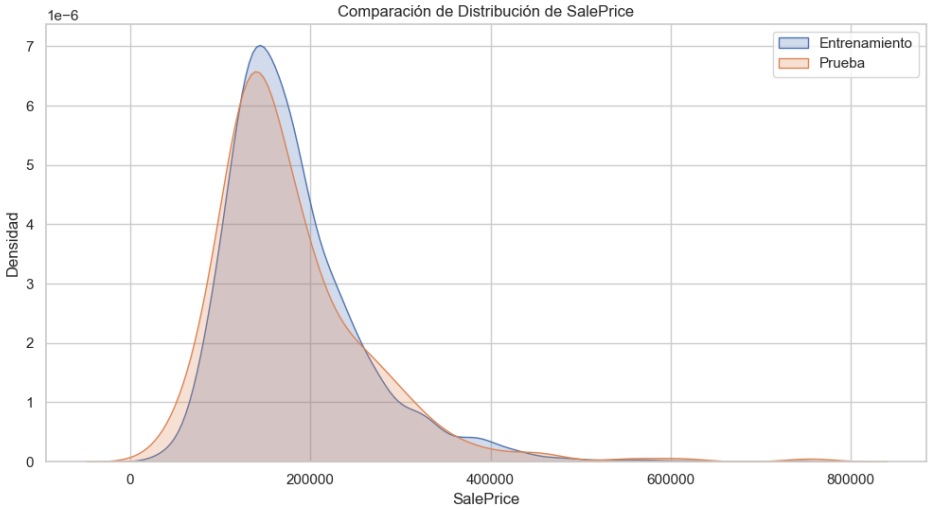
Utilice cada modelo con el conjunto de prueba y determine la eficiencia del algoritmo para predecir el precio de las casas. ¿Qué tan bien lo hizo? ¿Qué medidas usó para determinar la calidad de la predicción?






Matriz de Correlación
MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
MasVnrArea
ExterQual
ExterCond
BsmtQual
BsmtCond
BsmtFinSF1
BsmtUnfSF
TotalBsmtSF
1stFlrSF
2ndFlrSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
KitchenQual
TotRmsAbvGrd
Fireplaces
GarageYrBlt
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF
EnclosedPorch
3SsnPorch
ScreenPorch
PoolArea
SalePrice
TotalSF
Bathrooms
PorchArea
Age
YearsSinceRemodel
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


GrLivArea vs SalePrice
SalePrice
GrLivArea
0
100000
200000
300000
400000
500000
600000
700000
1000
2000
3000
4000
5000




```
### Visión General de los Datos ###
Dimensiones del dataset: (1460, 81)

Tipos de variables:
 Id               int64
MSSubClass        int64
MSZoning         object
LotFrontage     float64
LotArea           int64
                 ...
MoSold            int64
YrSold            int64
SaleType         object
SaleCondition    object
SalePrice         int64
Length: 81, dtype: object

Resumen estadístico:
                Id   MSSubClass  LotFrontage        LotArea  ...       MiscVal       MoSold       YrSold      SalePri
ce
count  1460.000000  1460.000000  1201.000000    1460.000000  ...   1460.000000  1460.000000  1460.000000    1460.00000
0
mean    730.500000    56.897260    70.049958   10516.828082  ...     43.489041     6.321918  2007.815753  180921.19589
0
std     421.610009    42.300571    24.284752    9981.264932  ...    496.123024     2.703626     1.328095   79442.50288
3
min       1.000000    20.000000    21.000000    1300.000000  ...      0.000000     1.000000  2006.000000   34900.00000
0
```

```
25%     365.750000    20.000000    59.000000    7553.500000  ...      0.000000     5.000000  2007.000000  129975.00000
0
50%     730.500000    50.000000    69.000000    9478.500000  ...      0.000000     6.000000  2008.000000  163000.00000
0
75%    1095.250000    70.000000    80.000000   11601.500000  ...      0.000000     8.000000  2009.000000  214000.00000
0
max    1460.000000   190.000000   313.000000  215245.000000  ...  15500.000000    12.000000  2010.000000  755000.00000
0

[8 rows x 38 columns]

Datos faltantes por variable:
 PoolQC          1453
MiscFeature     1406
Alley           1369
Fence           1179
MasVnrType       872
FireplaceQu      690
LotFrontage      259
GarageQual        81
GarageFinish      81
GarageType        81
GarageYrBlt       81
GarageCond        81
BsmtFinType2      38
BsmtExposure      38
BsmtCond          37
BsmtQual          37
BsmtFinType1      37
```

```
BsmtQual          37
BsmtFinType1      37
MasVnrArea         8
Electrical         1
dtype: int64

### Aplicando Ingeniería de Características ###
Antes de modificar el dataset: (1460, 81)
Después de modificar el dataset: (1460, 123)

### Visualizaciones ###

### Preprocesamiento de los Datos ###
Preprocesamiento completado. Nuevas dimensiones: (1460, 260)

### División del Conjunto de Datos ###
Filas en entrenamiento: 1168
Filas en prueba: 292

### Entrenamiento y Evaluación del Modelo ###
### Evaluación del Modelo en Conjunto de Prueba ###
MSE: 0.0231
RMSE: 0.1521
MAE: 0.1082
R²: 0.8761
```

Discuta sobre la efectividad de los modelos. ¿Cuál lo hizo mejor? ¿Cuál es el mejor modelo para predecir el precio de las casas? Haga los gráficos que crea que le pueden ayudar en la discusión.