**UNIVERSIDAD DEL VALLE DE GUATEMALA**

Minería de Datos Sección 10

Catedrática: Lynette Garcia



Hoja de trabajo 3

Análisis Exploratorio

Abner Iván García Alegría - 21285 Oscar Esteban Donis Martínez - 21610 Adrian Rodríguez Batres - 21691

Link Google Docs:



[Hoja de trabajo 3 - Minería de Datos](https://docs.google.com/document/d/1EKCuFx0Y0Cguy2mVVBBmocGwDo2JuefyLERQ7mXBDoE/edit?usp=sharing)



Link Github: https://github.com/estebandonis/HT3-MineriadeDatos.git

Descripción de las variables:

* **SalePrice** - the property's sale price in dollars. This is the target variable that you're trying to predict. (Variable Objetivo)
* **MSSubClass**: The building class
* **MSZoning**: The general zoning classification
* **LotFrontage**: Linear feet of street connected to property
* **LotArea**: Lot size in square feet
* **Street**: Type of road access
* **Alley**: Type of alley access
* **LotShape**: General shape of property
* **LandContour**: Flatness of the property
* **Utilities**: Type of utilities available
* **LotConfig**: Lot configuration
* **LandSlope**: Slope of property
* **Neighborhood**: Physical locations within Ames city limits
* **Condition1**: Proximity to main road or railroad
* **Condition2**: Proximity to main road or railroad (if a second is present)
* **BldgType**: Type of dwelling
* **HouseStyle**: Style of dwelling
* **OverallQual**: Overall material and finish quality
* **OverallCond**: Overall condition rating
* **YearBuilt**: Original construction date
* **YearRemodAdd**: Remodel date
* **RoofStyle**: Type of roof
* **RoofMatl**: Roof material
* **Exterior1st**: Exterior covering on house
* **Exterior2nd**: Exterior covering on house (if more than one material)
* **MasVnrType**: Masonry veneer type
* **MasVnrArea**: Masonry veneer area in square feet
* **ExterQual**: Exterior material quality
* **ExterCond**: Present condition of the material on the exterior
* **Foundation**: Type of foundation
* **BsmtQual**: Height of the basement
* **BsmtCond**: General condition of the basement
* **BsmtExposure**: Walkout or garden level basement walls
* **BsmtFinType1**: Quality of basement finished area
* **BsmtFinSF1**: Type 1 finished square feet
* **BsmtFinType2**: Quality of second finished area (if present)
* **BsmtFinSF2**: Type 2 finished square feet
* **BsmtUnfSF**: Unfinished square feet of basement area
* **TotalBsmtSF**: Total square feet of basement area
* **Heating**: Type of heating
* **HeatingQC**: Heating quality and condition
* **CentralAir**: Central air conditioning
* **Electrical**: Electrical system
* **1stFlrSF**: First Floor square feet
* **2ndFlrSF**: Second floor square feet
* **LowQualFinSF**: Low quality finished square feet (all floors)
* **GrLivArea**: Above grade (ground) living area square feet
* **BsmtFullBath**: Basement full bathrooms
* **BsmtHalfBath**: Basement half bathrooms
* **FullBath**: Full bathrooms above grade
* **HalfBath**: Half baths above grade
* **Bedroom**: Number of bedrooms above basement level
* **Kitchen**: Number of kitchens
* **KitchenQual**: Kitchen quality
* **TotRmsAbvGrd**: Total rooms above grade (does not include bathrooms)
* **Functional**: Home functionality rating
* **Fireplaces**: Number of fireplaces
* **FireplaceQu**: Fireplace quality
* **GarageType**: Garage location
* **GarageYrBlt**: Year garage was built
* **GarageFinish**: Interior finish of the garage
* **GarageCars**: Size of garage in car capacity
* **GarageArea**: Size of garage in square feet
* **GarageQual**: Garage quality
* **GarageCond**: Garage condition
* **PavedDrive**: Paved driveway
* **WoodDeckSF**: Wood deck area in square feet
* **OpenPorchSF**: Open porch area in square feet
* **EnclosedPorch**: Enclosed porch area in square feet
* **3SsnPorch**: Three season porch area in square feet
* **ScreenPorch**: Screen porch area in square feet
* **PoolArea**: Pool area in square feet
* **PoolQC**: Pool quality
* **Fence**: Fence quality
* **MiscFeature**: Miscellaneous feature not covered in other categories
* **MiscVal**: $Value of miscellaneous feature
* **MoSold**: Month Sold
* **YrSold**: Year Sold
* **SaleType**: Type of sale
* **SaleCondition**: Condition of sale

# Análisis Exploratorio

## Cualitativas:(46)

MSSubClass: Cualitativa nominal MSZoning: Cualitativa nominal Street: Cualitativa nominal

Alley: Cualitativa nominal

LotShape: Cualitativa ordinal LandContour: Cualitativa nominal Utilities: Cualitativa nominal LotConfig: Cualitativa nominal LandSlope: Cualitativa ordinal Neighborhood: Cualitativa nominal Condition1: Cualitativa nominal Condition2: Cualitativa nominal BldgType: Cualitativa nominal HouseStyle: Cualitativa nominal RoofStyle: Cualitativa nominal RoofMatl: Cualitativa nominal Exterior1st: Cualitativa nominal Exterior2nd: Cualitativa nominal MasVnrType: Cualitativa nominal ExterQual: Cualitativa ordinal ExterCond: Cualitativa ordinal Foundation: Cualitativa nominal BsmtQual: Cualitativa ordinal BsmtCond: Cualitativa ordinal BsmtExposure: Cualitativa ordinal BsmtFinType1: Cualitativa ordinal BsmtFinType2: Cualitativa ordinal Heating: Cualitativa nominal HeatingQC: Cualitativa ordinal CentralAir: Cualitativa nominal Electrical: Cualitativa nominal KitchenQual: Cualitativa ordinal Functional: Cualitativa ordinal FireplaceQu: Cualitativa ordinal GarageType: Cualitativa nominal GarageFinish: Cualitativa ordinal GarageQual: Cualitativa ordinal GarageCond: Cualitativa ordinal PavedDrive: Cualitativa ordinal PoolQC: Cualitativa ordinal Fence: Cualitativa ordinal MiscFeature: Cualitativa nominal

MoSold: Cualitativa ordinal (en el contexto de fecha, representa meses) YrSold: Cualitativa ordinal (en el contexto de fecha, representa años) SaleType: Cualitativa nominal

SaleCondition: Cualitativa nominal

# Cuantitativas:(34)

LotFrontage: Cuantitativa continua LotArea: Cuantitativa continua OverallQual: Cuantitativa ordinal OverallCond: Cuantitativa ordinal YearBuilt: Cuantitativa discreta YearRemodAdd: Cuantitativa discreta MasVnrArea: Cuantitativa continua BsmtFinSF1: Cuantitativa continua BsmtFinSF2: Cuantitativa continua BsmtUnfSF: Cuantitativa continua TotalBsmtSF: Cuantitativa continua 1stFlrSF: Cuantitativa continua 2ndFlrSF: Cuantitativa continua LowQualFinSF: Cuantitativa continua GrLivArea: Cuantitativa continua BsmtFullBath: Cuantitativa discreta BsmtHalfBath: Cuantitativa discreta FullBath: Cuantitativa discreta HalfBath: Cuantitativa discreta Bedroom: Cuantitativa discreta Kitchen: Cuantitativa discreta TotRmsAbvGrd: Cuantitativa discreta Fireplaces: Cuantitativa discreta GarageYrBlt: Cuantitativa discreta GarageCars: Cuantitativa discreta GarageArea: Cuantitativa continua WoodDeckSF: Cuantitativa continua OpenPorchSF: Cuantitativa continua EnclosedPorch: Cuantitativa continua 3SsnPorch: Cuantitativa continua ScreenPorch: Cuantitativa continua PoolArea: Cuantitativa continua MiscVal: Cuantitativa continua SalePrice: Cuantitativa continua

## Variables seleccionadas:

SalePrice (Target Variable) - The property's sale price in dollars. LotFrontage - Linear feet of street connected to the property.

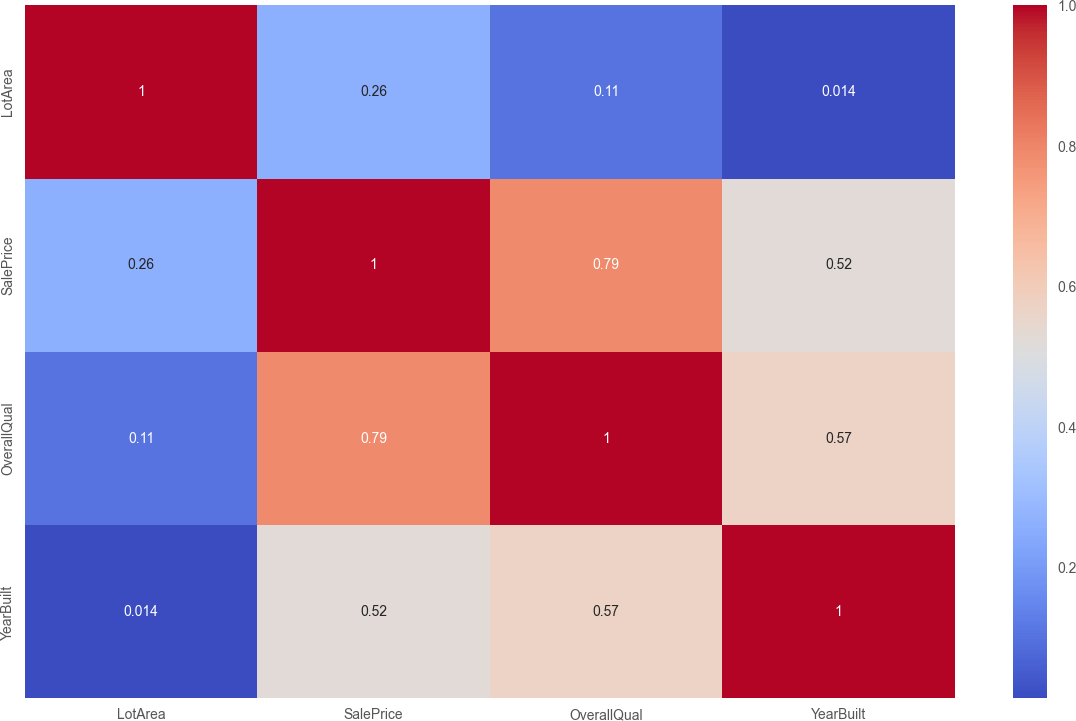
LotArea - Lot size in square feet.

OverallQual - Overall material and finish quality. OverallCond - Overall condition rating.

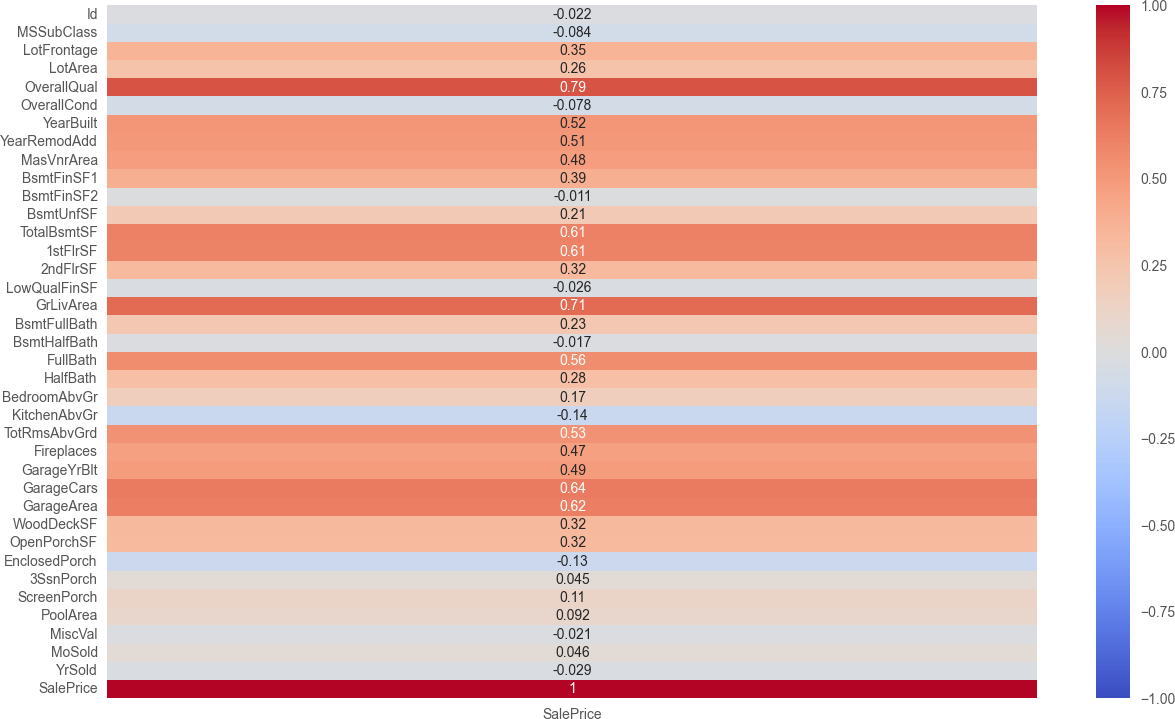
YearBuilt - Original construction date. YearRemodAdd - Remodel date.

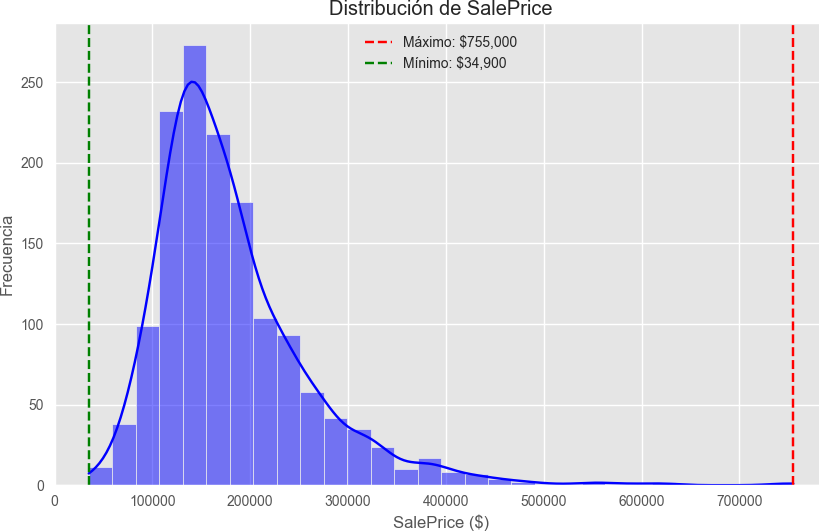
MasVnrArea - Masonry veneer area in square feet. TotalBsmtSF - Total square feet of basement area. GrLivArea - Above grade (ground) living area square feet.

Correlacion SalePrice vs Demas variables

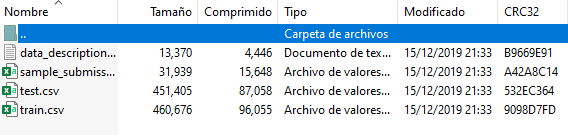


Análisis de relaciones con la variable de respuesta:

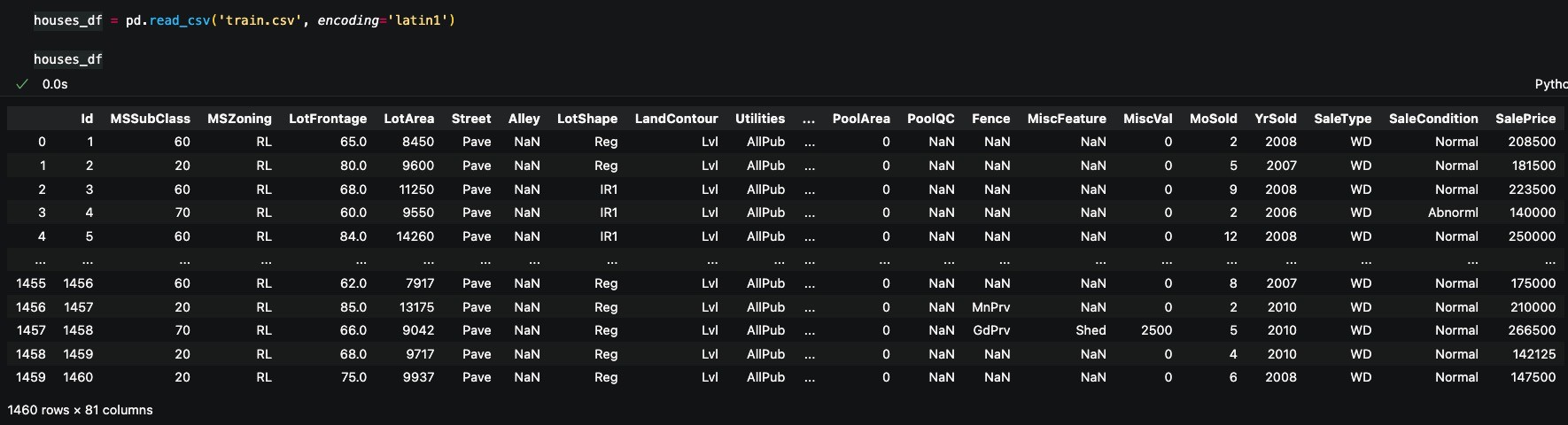




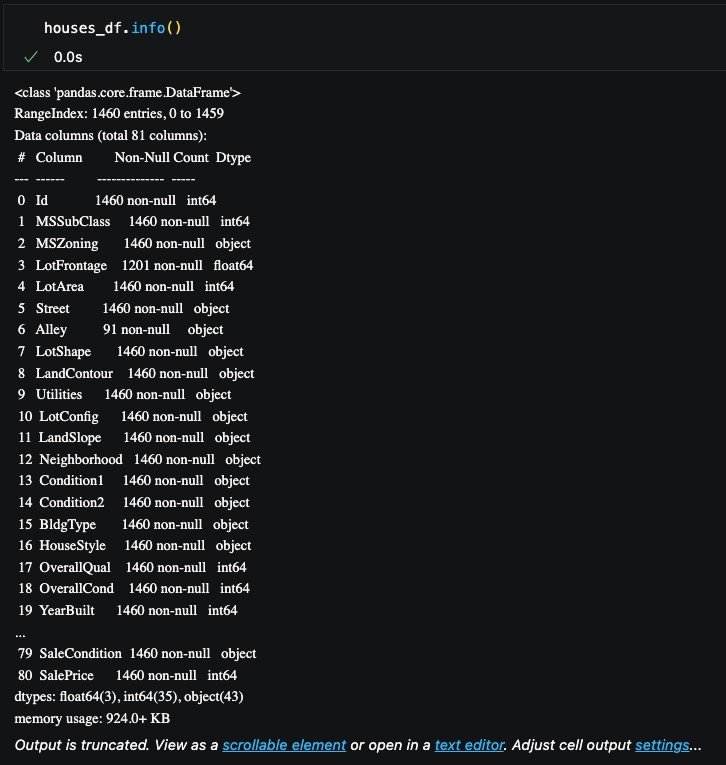
## Descargue los conjuntos de datos de la plataforma kaggle.

****

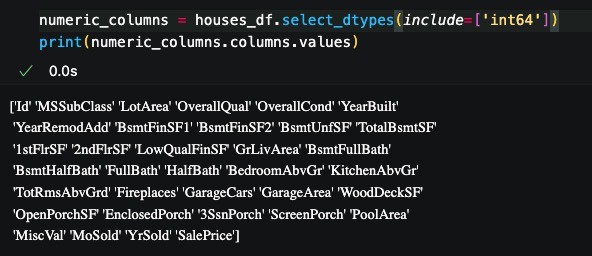
1. **Haga un análisis exploratorio extenso de los datos. Explique bien todos los hallazgos. No ponga solo gráficas y código. Debe llegar a conclusiones interesantes para poder predecir. Explique el preprocesamiento que necesitó hacer.**

****

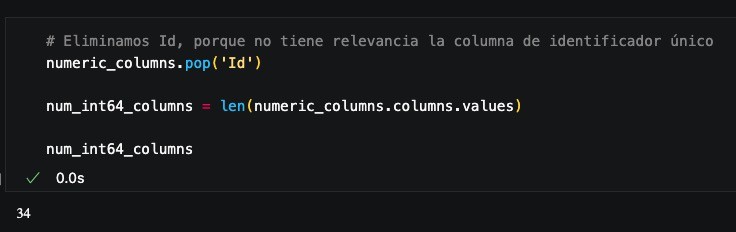
Como podemos ver nuestra base de datos posee 1460 filas y 81 columnas



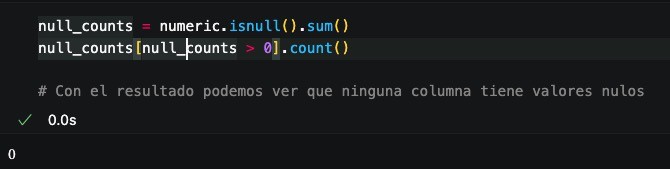
Tenemos variables que poseen valores interger y valores tipo Object



Extraemos las variables numéricas



Podemos ver que tenemos en total 34 variables numéricas, quitando la variable de Id.



Podemos ver que ninguno de los campos no tienen valores nulos.



Contamos la cantidad de variables que poseen filas con 0



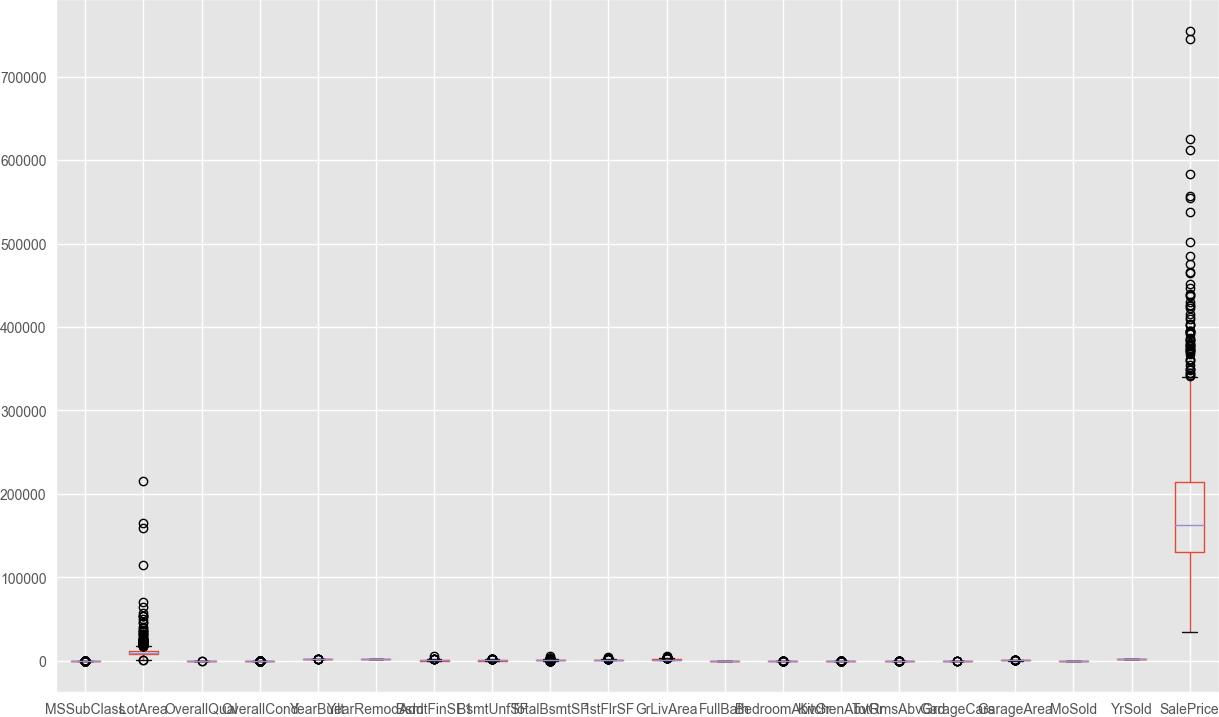
En total contamos con 14 variables con más de 650 0s entre sus filas

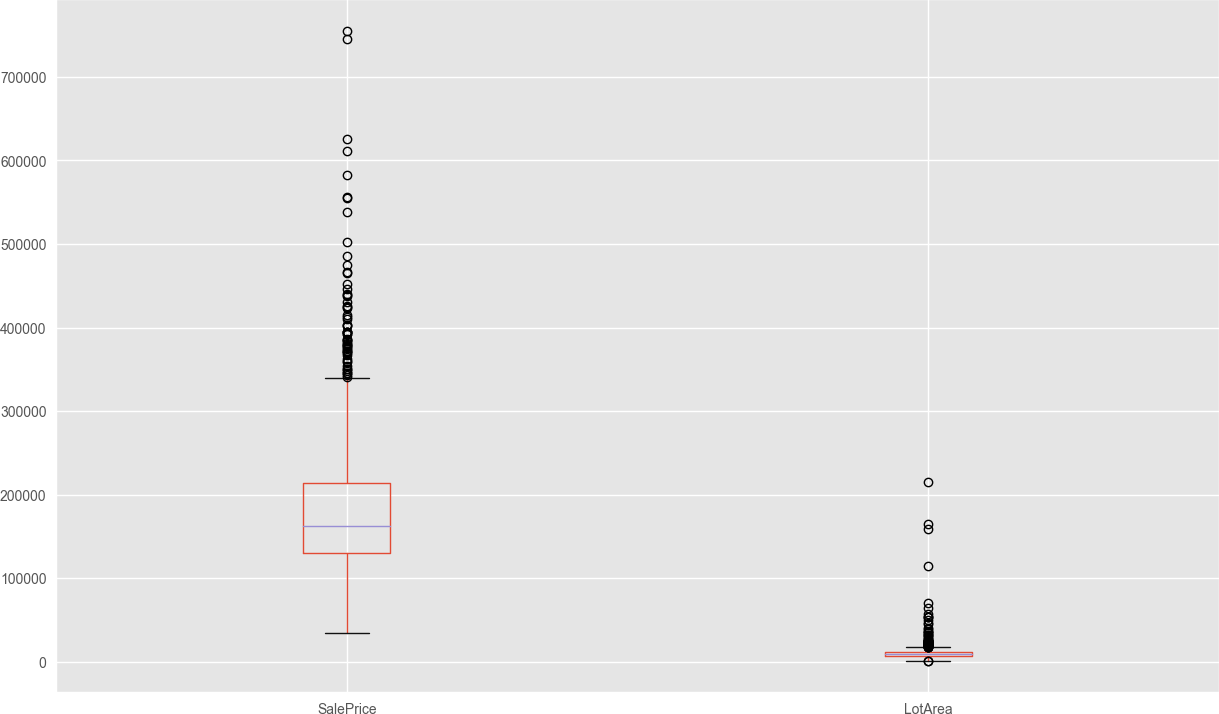
Eliminamos dichas variables, ya que pueden afectar nuestro análisis por tanto valor 0.

Nos quedamos con 20 variables para hacer el análisis

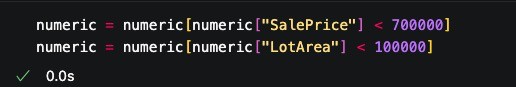


Podemos observar que ninguna de las variables posee una distribución normal

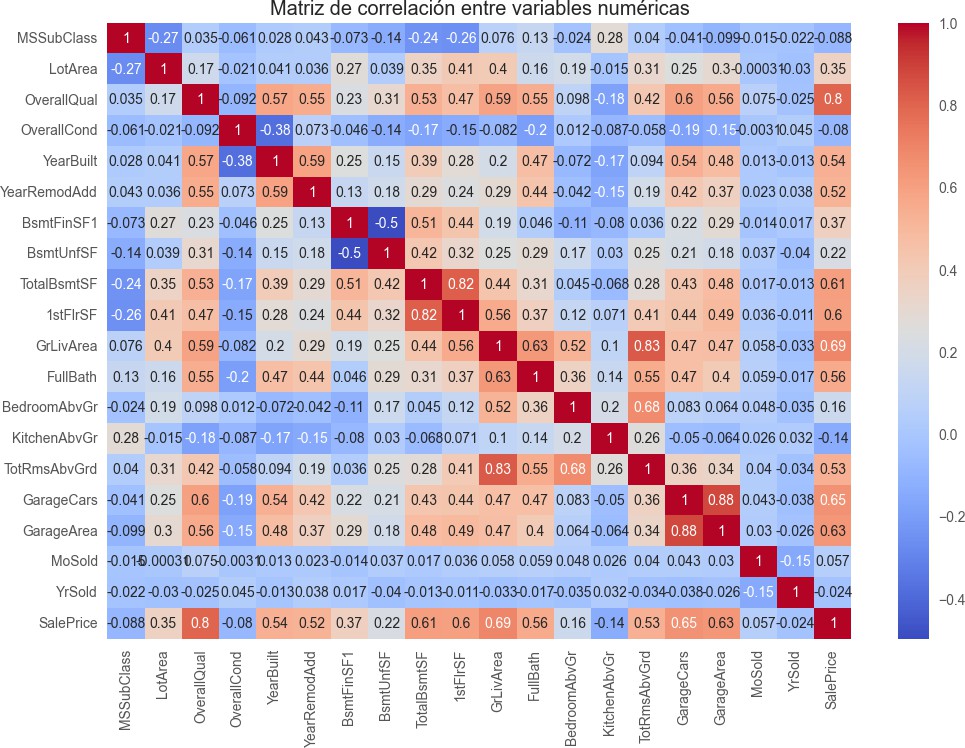


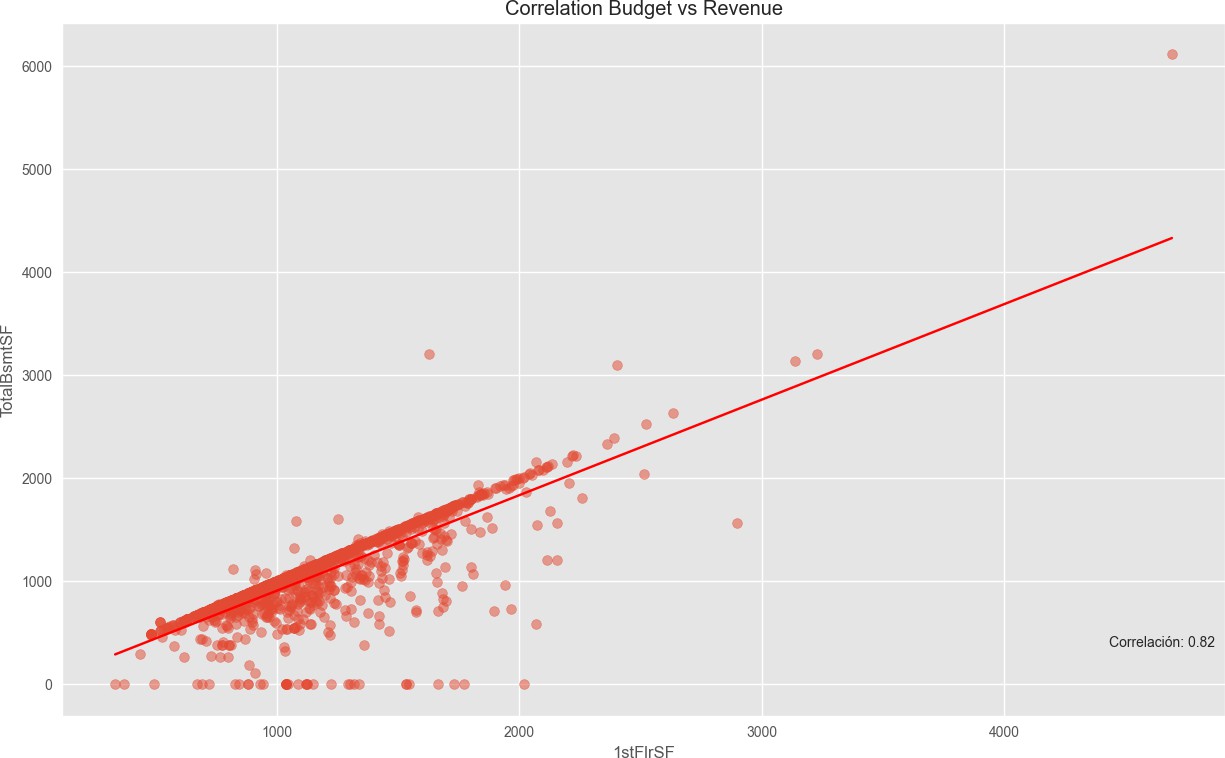


Podemos observar que las únicas variables con valores atípicos son SalePrice y LotArea

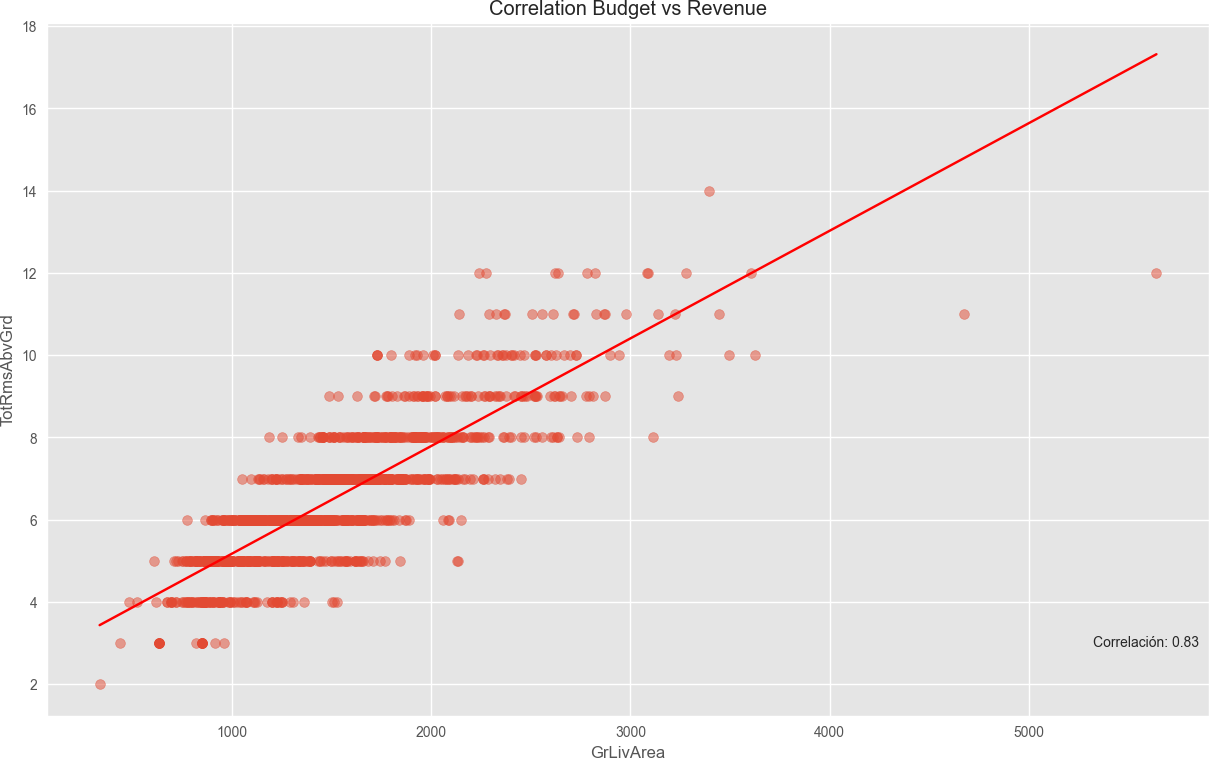


Eliminamos los valores atípicos de dichas variables para que no afecten nuestro análisis

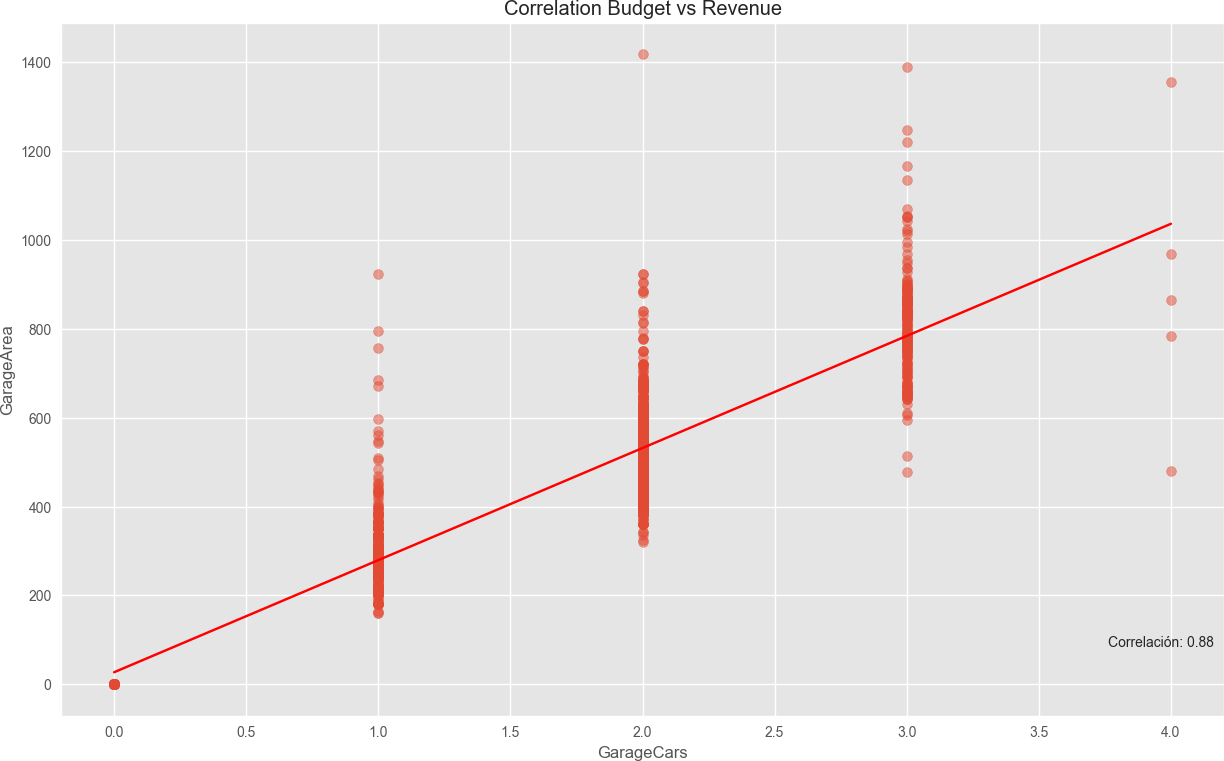




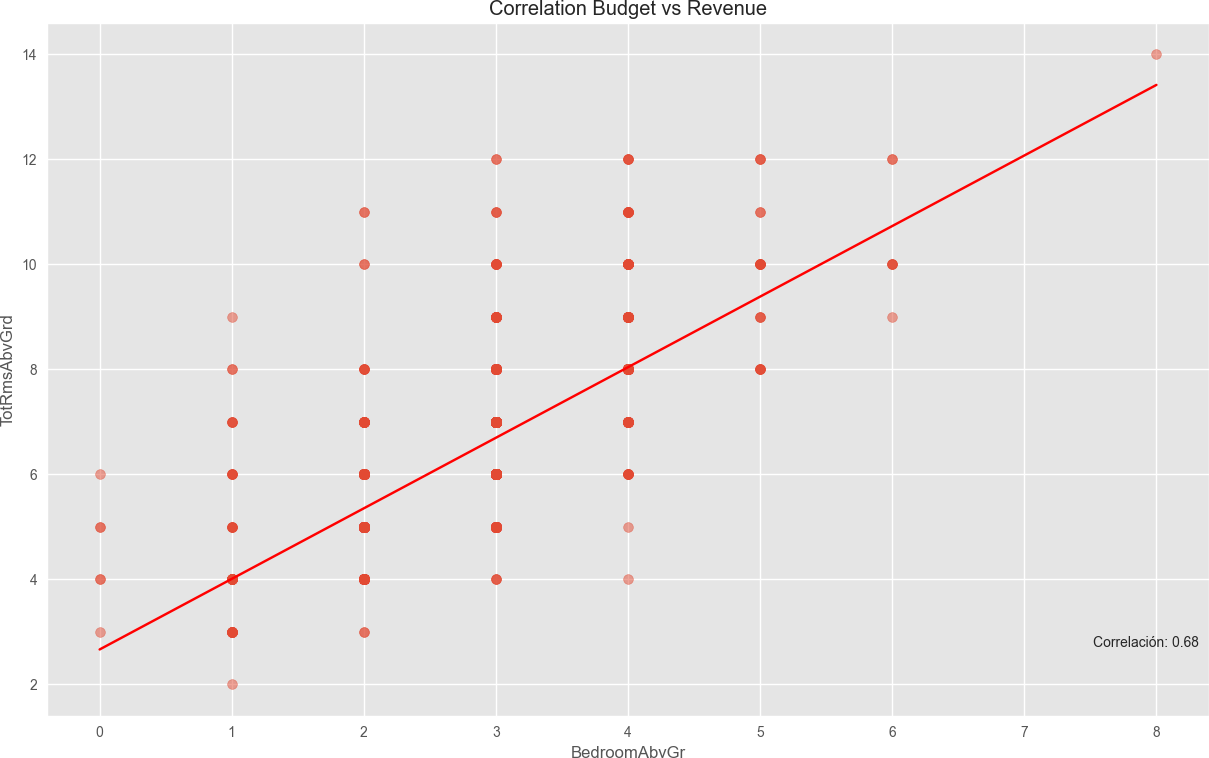
En el siguiente gráfico podemos observar que 1stFlrSF y TotatBsmtSF poseen una gran relación entre sí. Con una correlación de 0.82



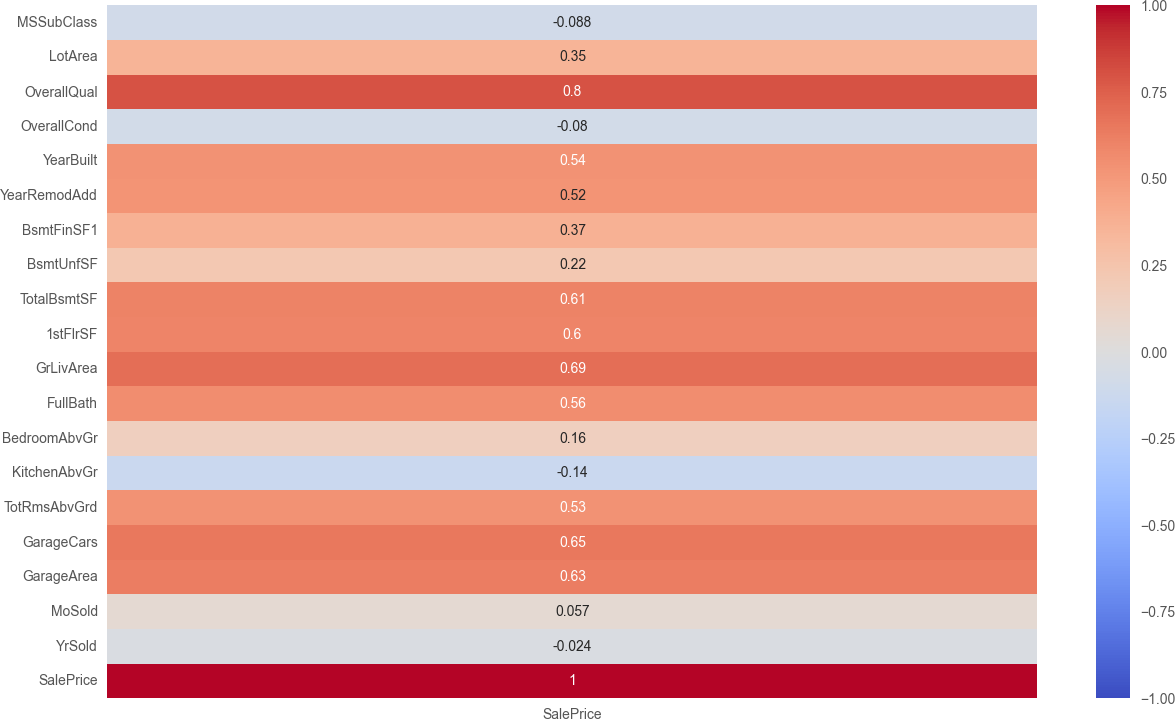
También podemos observar que GrLivArea y TotRmsAbvGrd tambien posee una buena relación con R2 de 0.83



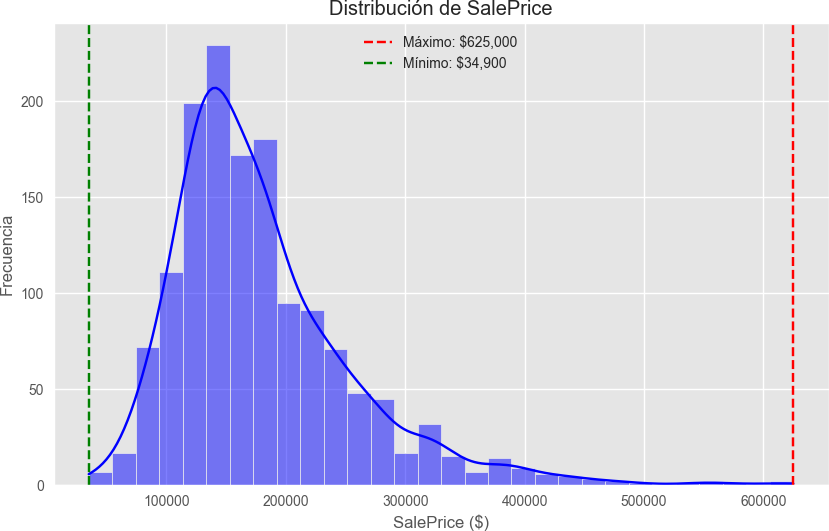
También podemos observar que GarageArea y GarageCars también posee una buena relación con R2 de 0.88



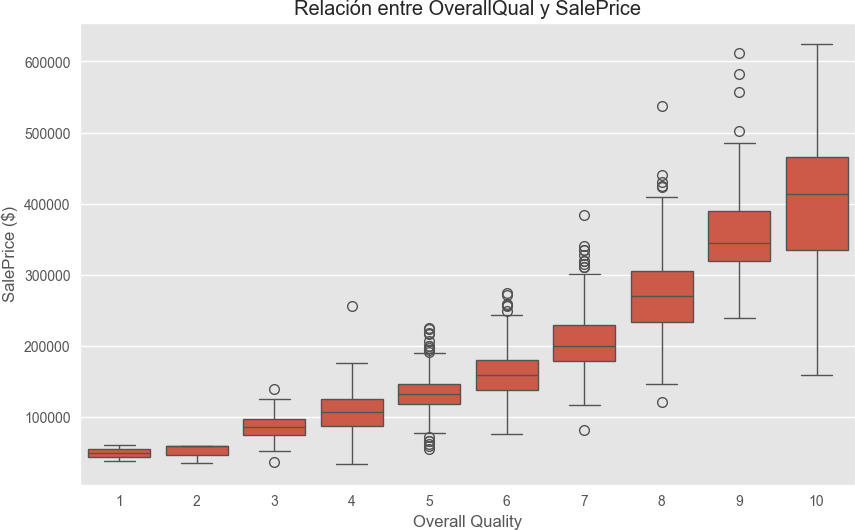
También podemos observar que TotRmsAbvGrd y BedroomAbvGr es lo más cerca, después de las parejas anteriores, que poseen mayor relación. Con una correlación de 0.68.



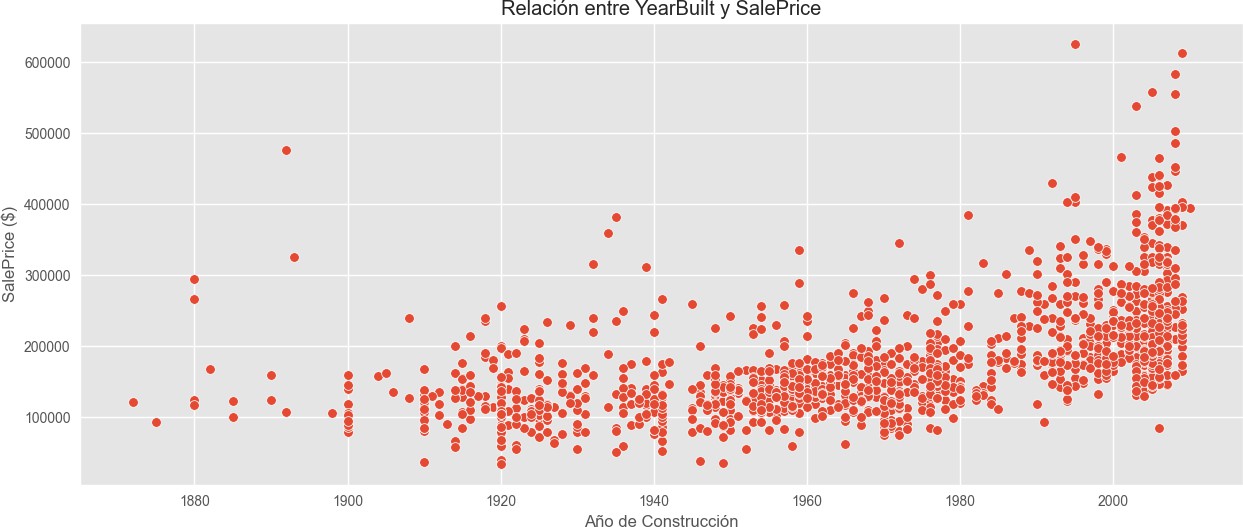
Podemos ver que 10 de las 20 variables poseen una correlación positiva mayor a 0.5 con SalePrice.



Podemos ver que la distribución de SalePrice, es casi normal, pero no llega en su totalidad

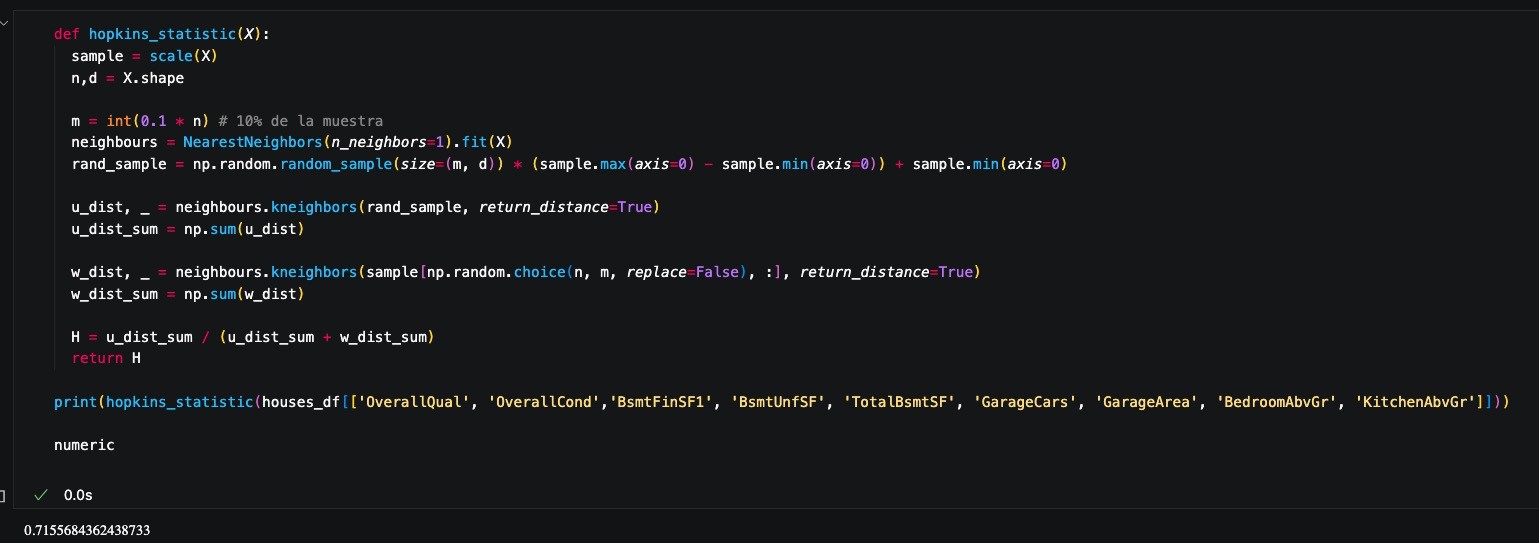


Al analizar la relación entre SalePrice y Overall quality, que poseen una alta correlación, se puede ver que los valores atípicos van incrementando a medida que va creciendo la calidad.

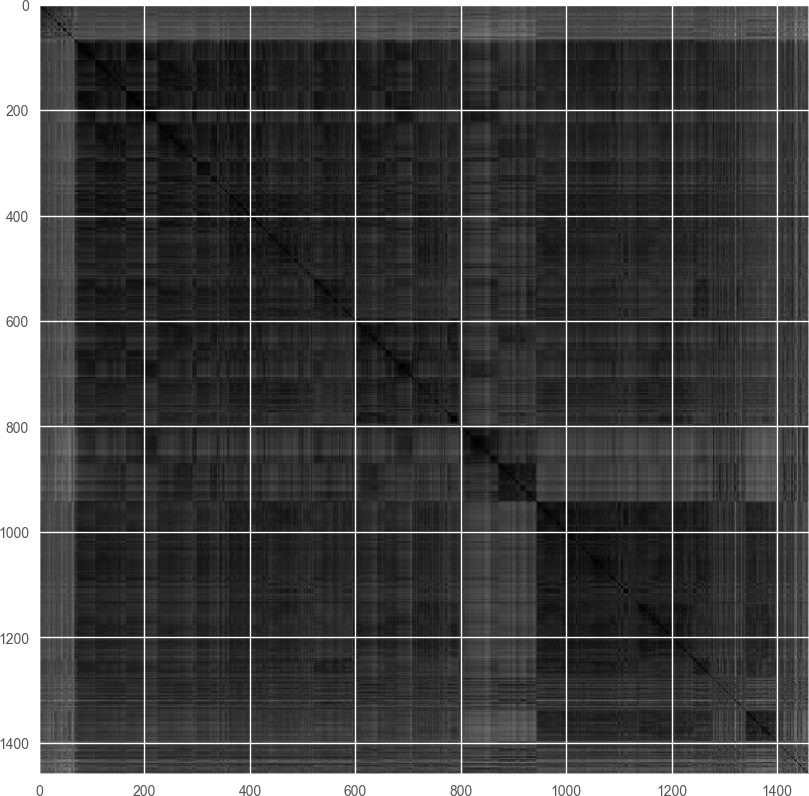


Al analizar el año de construcción y el sale price, tenemos que a medida que aumenta el tiempo, el precio de las cosas más modernas van aumentando.

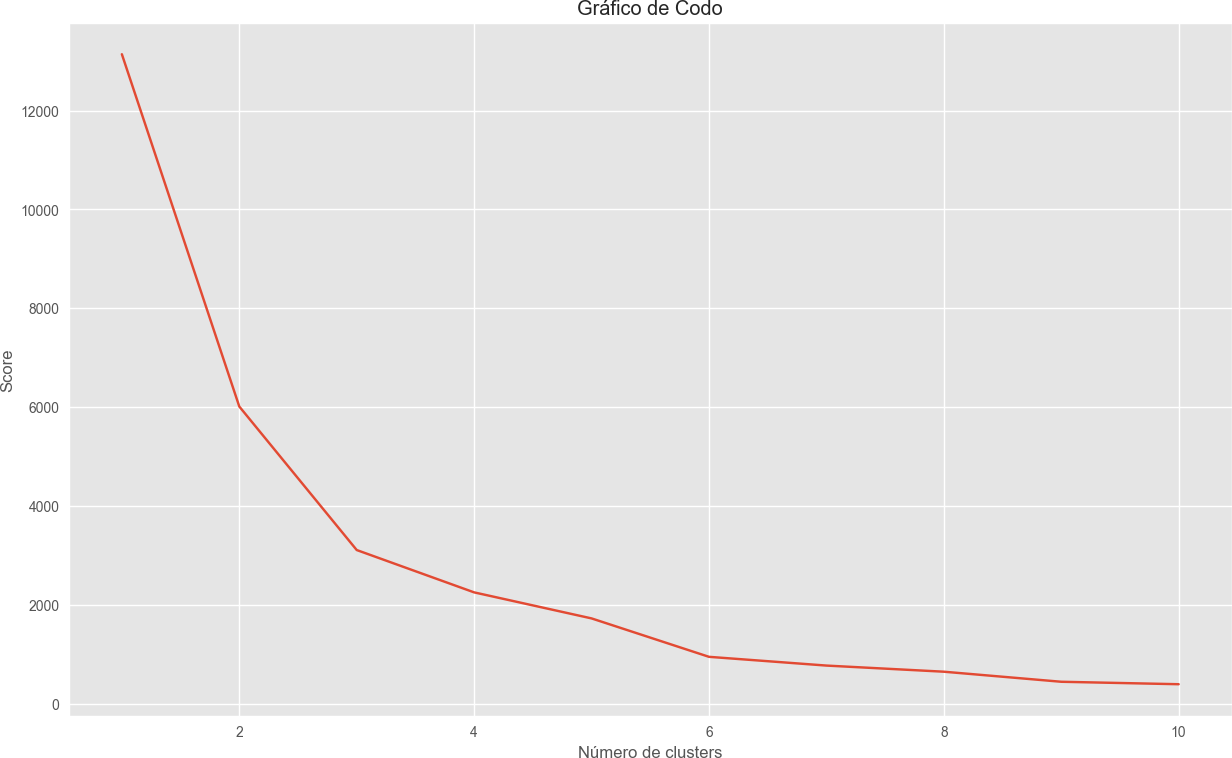
## Incluya un análisis de grupos en el análisis exploratorio. Explique las características de los grupos.

****

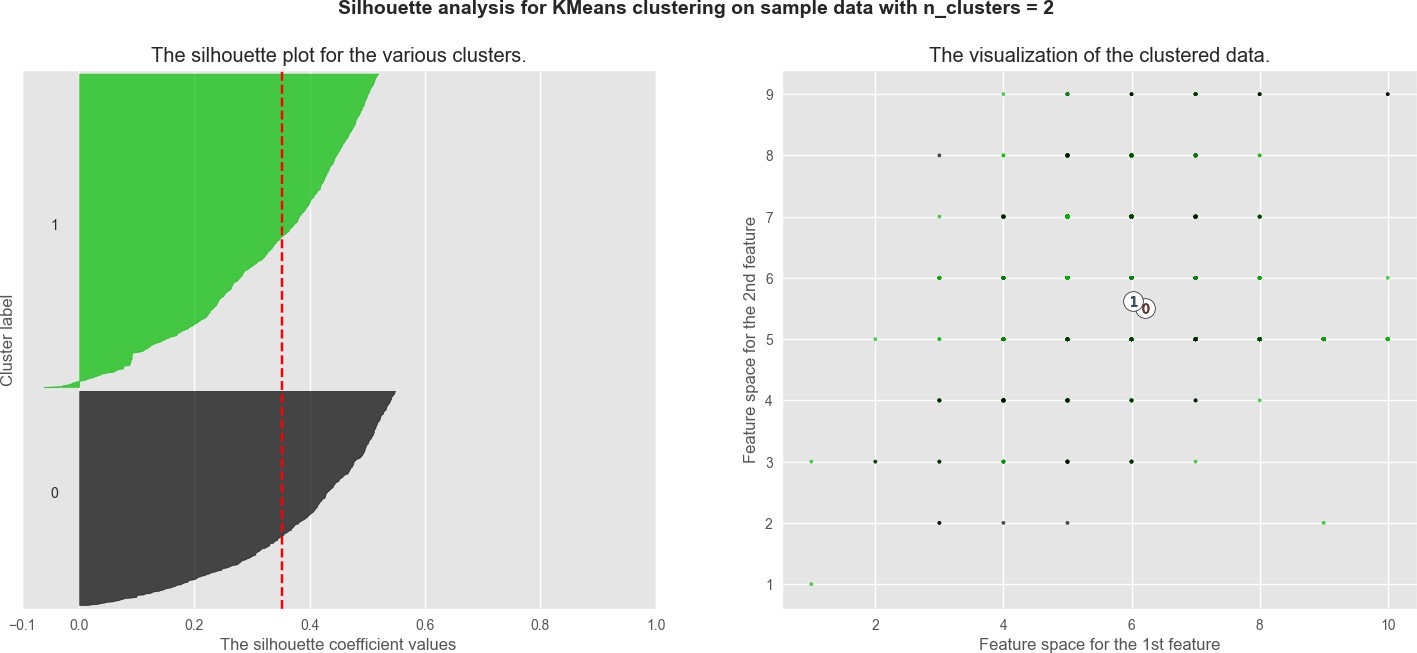
Al comprobar nuestro sets de datos, quitando variables como SalePrice y LotArea porque tienden a no congeniar bien con los demás datos (Según pruebas con esta misma métrica), Dimos que este conjunto de datos campos es el más factible con una puntuación de 0.71.

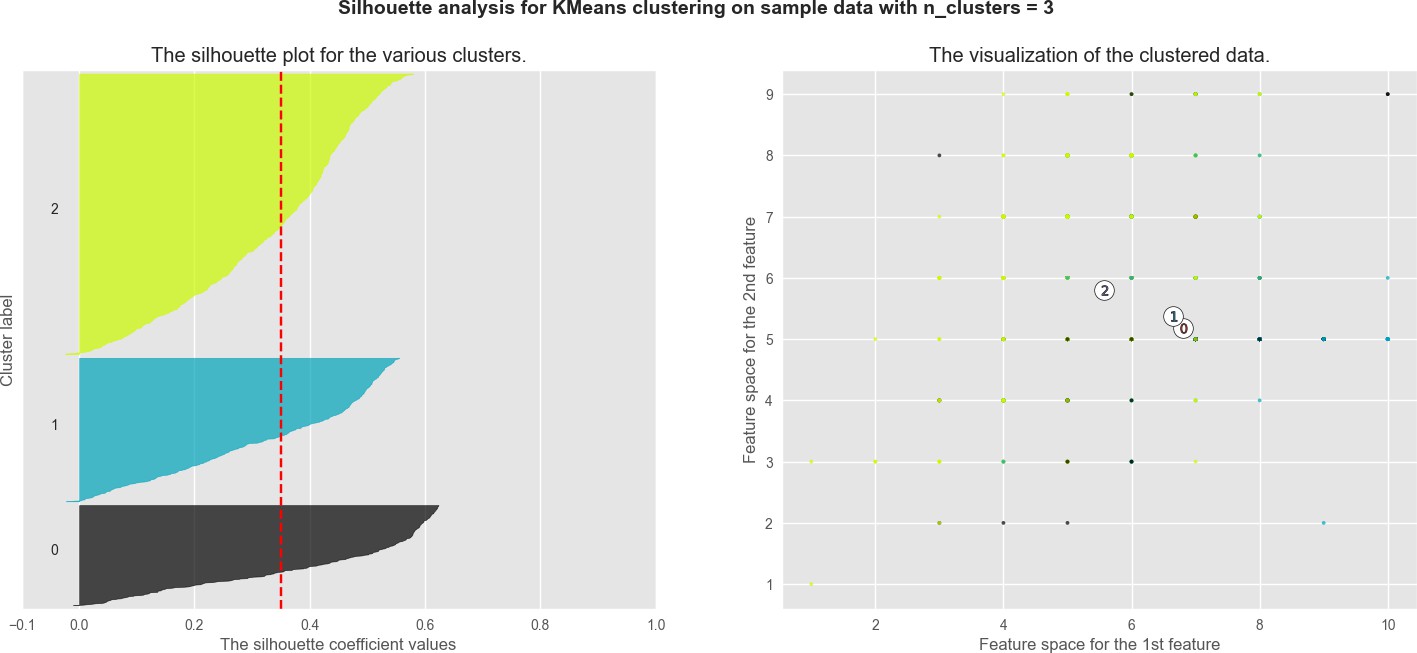


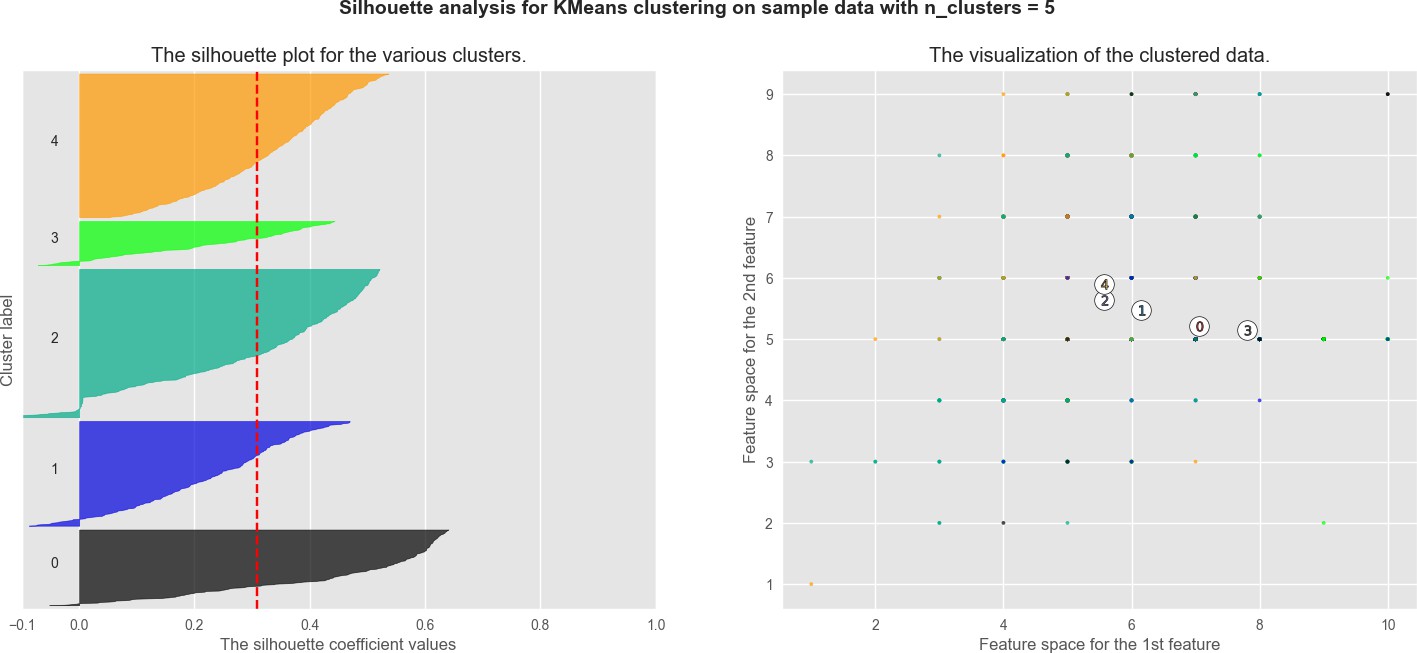
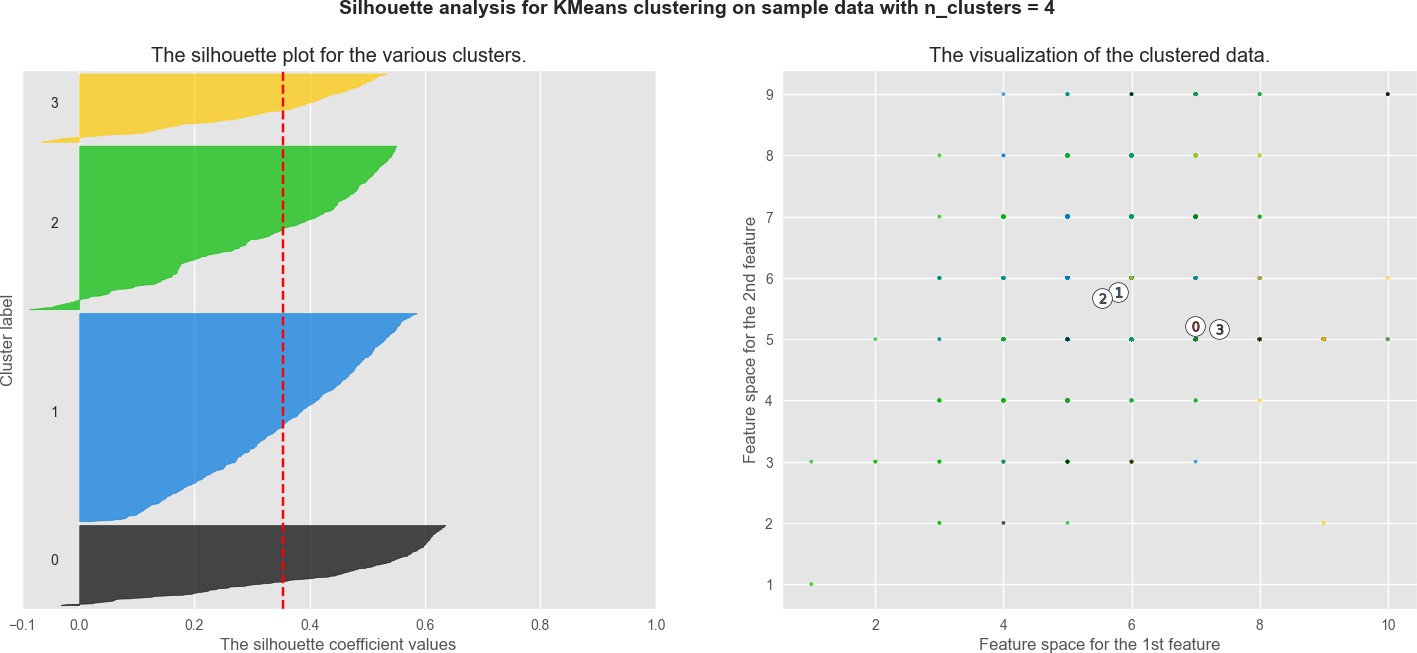
Al ver la gráfica VAT, podemos identificar alrededor de 4 grupos.



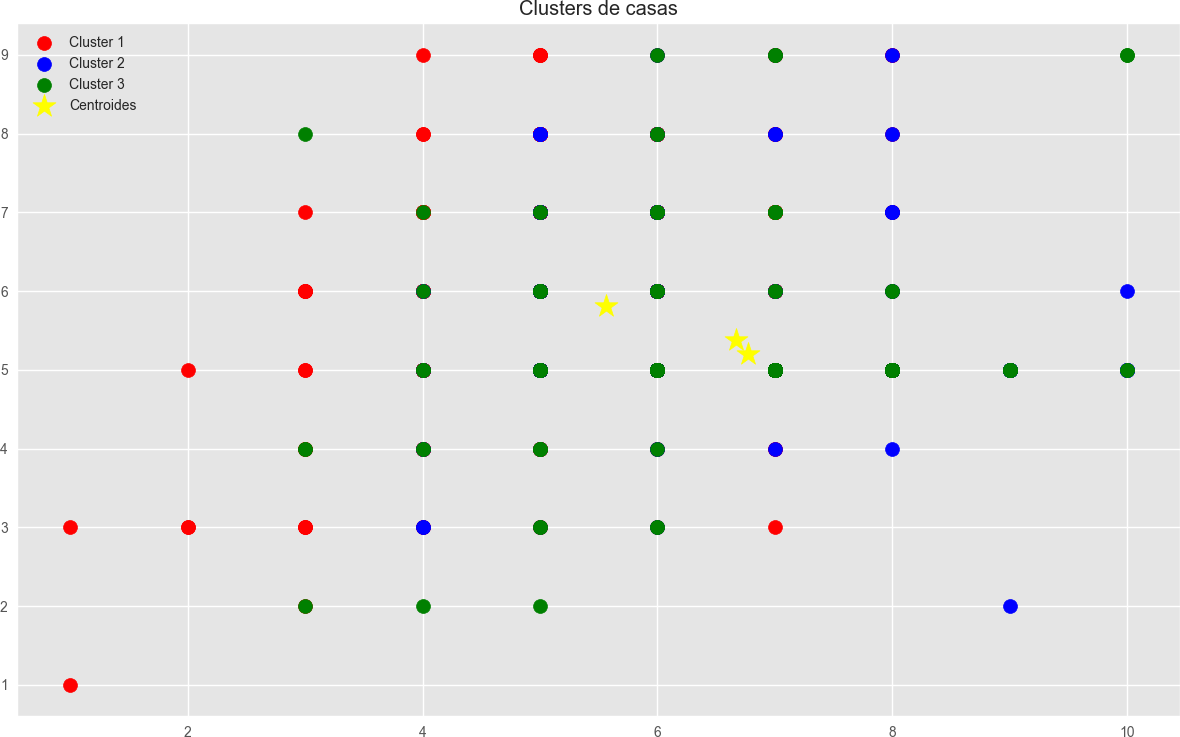
Al analizar la gráfica podemos observar que el “Codo”, se encuentra 3 grupos







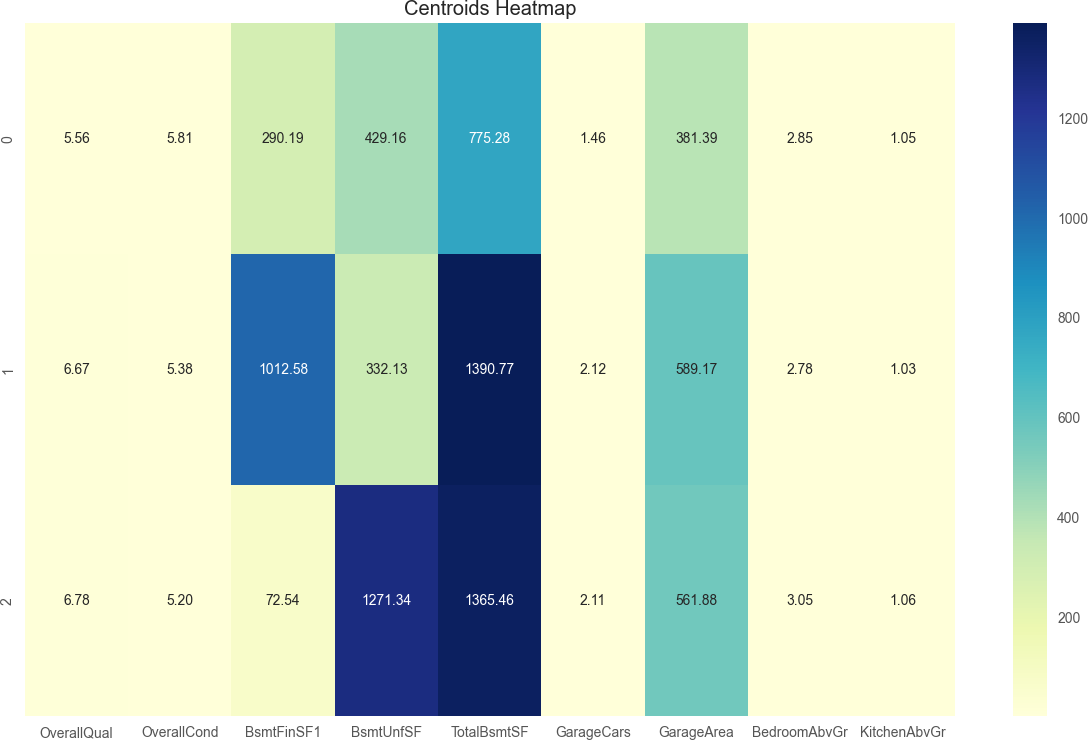
Al ver las diferentes distribuciones podemos observar que de 2 a 3 grupos se ven los ideales para realizar el clustering.



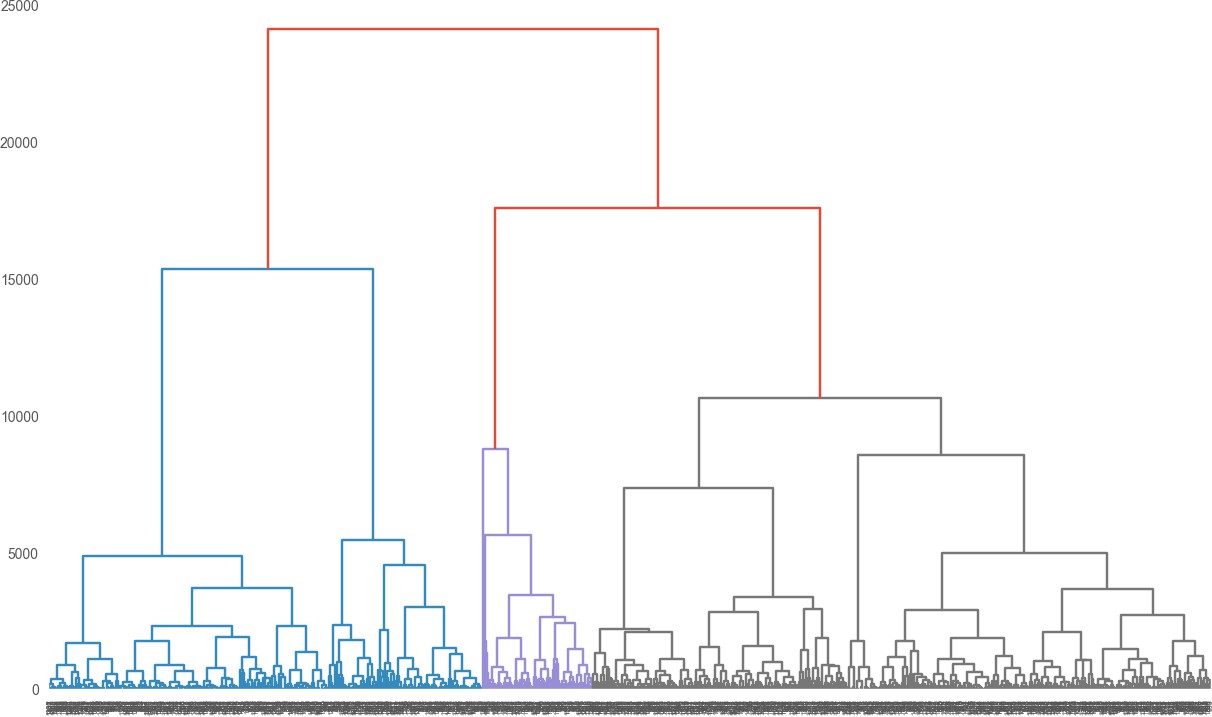
Al realizar Kmeans, podemos observar que los centroides quedaron muy céntricos, por lo que la distribución se ve afectada, también por la separación que existe entre los datos.



Cuando hacemos el PCA, podemos observar que nuestros grupos están un poco juntos, y se puede ver que la distancia entre grupos es demasiado baja.



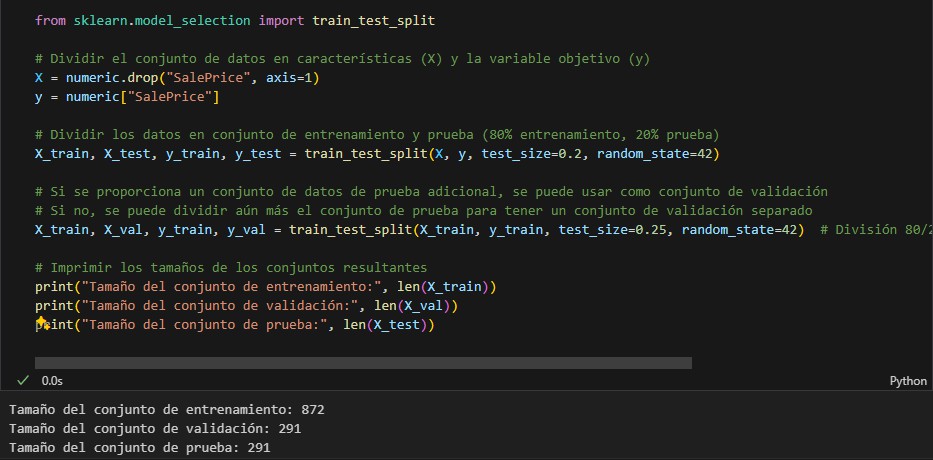
En el heatmap de los centroides podemos observar que los centroides están distribuidos en una zona.



En la gráfica del dendograma podemos observar que nuestra predicción de 3 grupos fue correcta.

En la interpretación de los grupos, podemos observar que existe un grupo centrico, que abarca las casas más cerca a la media, tanto en el espacio de cocina, garaje, basement, etc, como en la condición y calidad de la casa. Significando que dichas casas se encuentran en un estado bastante aceptable, mientras que el grupo que se encuentra en la izquierda nos muestra las casa que se encuentran en una peor puntuación, ya sea por la condición de la casa, el tamaño del garaje siendo más pequeño, etc. El de la derecha es el grupo que nos indica las casas que, en algunos de estos campos o posiblemente en la mayoría, su tamaño de garaje, cocina o condición y calidad de la construcción está por encima de la media; haciéndolas bastante cotizadas.

## 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. Si le proveen un conjunto de datos de prueba y tiene suficientes datos, tómelo como de validación, pero haga sus propios conjuntos de prueba.



Hemos utilizado la función train\_test\_split de scikit-learn para dividir los conjuntos de datos en entrenamiento, validación y prueba. Hemos mantenido el 80% para entrenamiento, 10% para validación y 10% para prueba. Además, hemos utilizado el parámetro random\_state para garantizar que la división sea reproducible.

## 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.

Tamaño de la casa: Variables como el área total de la casa (por ejemplo, 'GrLivArea'), el área del sótano ('TotalBsmtSF'), el área del primer piso ('1stFlrSF'), y el número total de habitaciones ('TotRmsAbvGrd') podrían tener una fuerte correlación con el precio de la casa.

Calidad de la construcción: Variables que reflejan la calidad de la construcción, como 'OverallQual' (calidad general del material y el acabado) y 'OverallCond' (condición general de la casa), podrían ser buenos predictores del precio de la casa.

Ubicación: Factores como el vecindario ('Neighborhood'), la proximidad a comodidades como escuelas, transporte público, parques, etc., podrían influir significativamente en el precio de la casa.

Características del lote: Variables relacionadas con el tamaño y las características del lote, como 'LotArea' (área del lote), 'LotFrontage' (ancho del lote), y 'LotShape' (forma del lote), también podrían ser importantes predictores.

Características del garaje: La presencia y el tamaño del garaje ('GarageCars', 'GarageArea') pueden afectar el precio de la casa.

Año de construcción y remodelación: Variables como 'YearBuilt' y 'YearRemodAdd' podrían ser indicadores importantes de la antigüedad y la modernización de la casa, lo que influye en su precio.

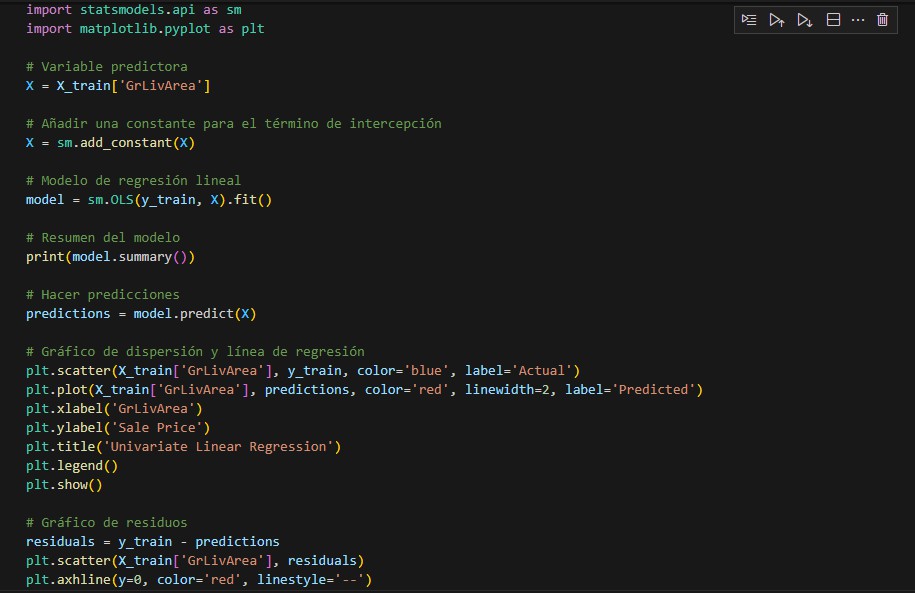
Características adicionales: Variables como la presencia de piscina ('PoolArea'), porches ('OpenPorchSF', 'EnclosedPorch', etc.), y otras características especiales ('MiscFeature') pueden influir en el precio de la casa.

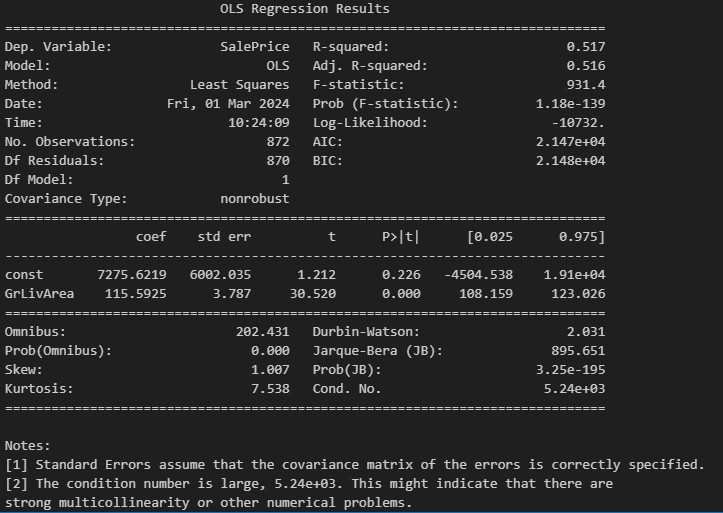
La selección de variables para la ingeniería de características se basó en una combinación de conocimiento del dominio, análisis exploratorio de datos y experimentación iterativa para identificar las variables que tienen el mayor potencial para servir como predictores efectivos del precio de las casas.

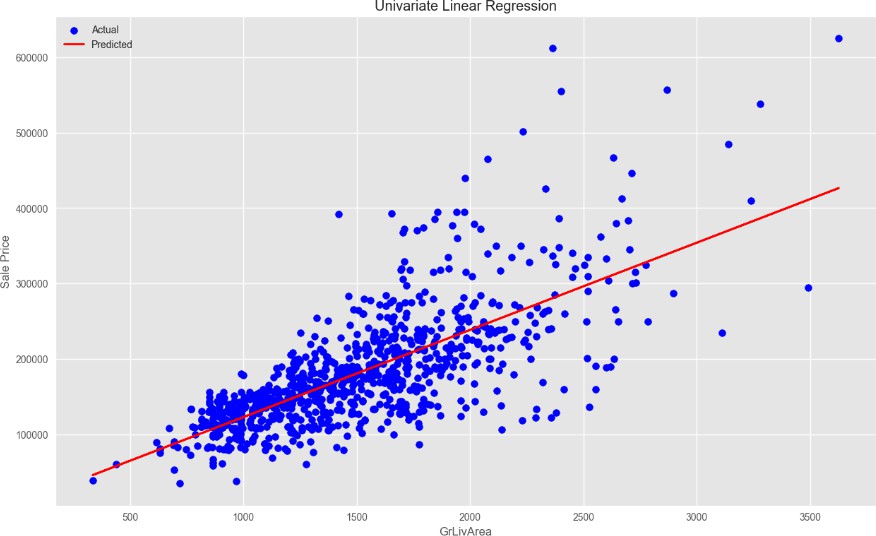
## Todos los resultados deben ser reproducibles por lo que se debe fijar que los conjuntos de entrenamiento y prueba sean los mismos siempre que se ejecute el código.

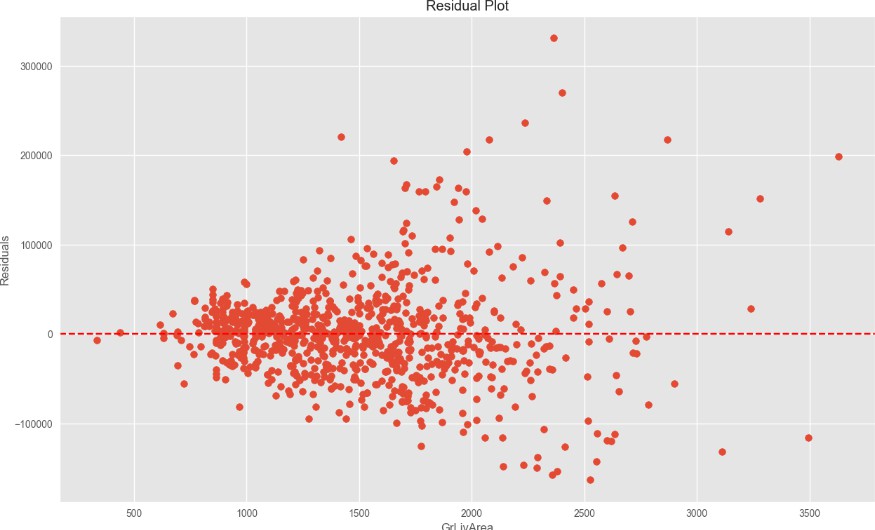
Al establecer random\_state=42, se asegura que la división de los datos sea la misma cada vez que se ejecute este código. Esto garantiza la reproducibilidad de los resultados.

## 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.



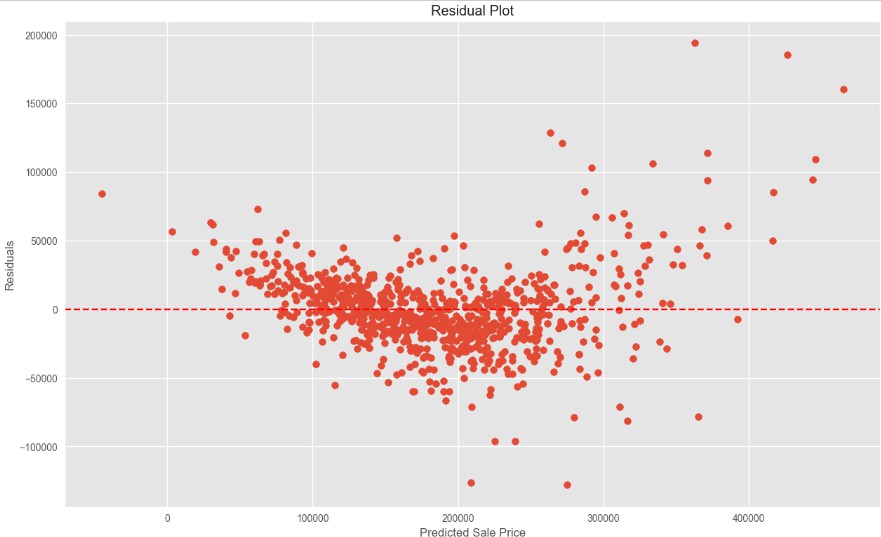
****



****

Seleccionamos la variable predictora 'GrLivArea'. Luego, ajustamos un modelo de regresión lineal utilizando la biblioteca statsmodels. Después de ajustar el modelo, imprimimos un resumen que proporciona estadísticas importantes sobre el modelo. Luego, hacemos predicciones utilizando el modelo y trazamos un gráfico de dispersión con la línea de regresión para visualizar la relación entre 'GrLivArea' y 'SalePrice'. Finalmente, trazamos un gráfico de residuos para evaluar la calidad del modelo.

## 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.



Primero añadimos una constante para el término de intercepción utilizando sm.add\_constant(). Luego, ajustamos el modelo de regresión lineal utilizando sm.OLS() y el método fit(). Imprimimos un resumen del modelo para analizar las estadísticas relevantes. Luego, hacemos predicciones utilizando el modelo y trazamos un gráfico de dispersión entre los valores reales y predichos para evaluar la precisión del modelo. Finalmente, trazamos un gráfico de residuos para evaluar la calidad del ajuste del modelo.

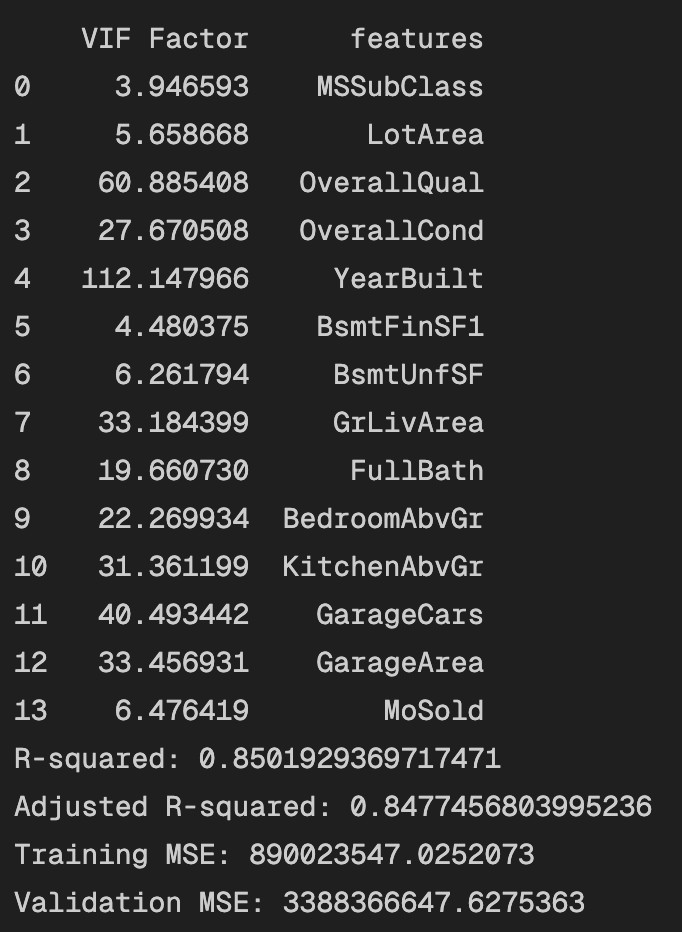
## 9 y 10.

****

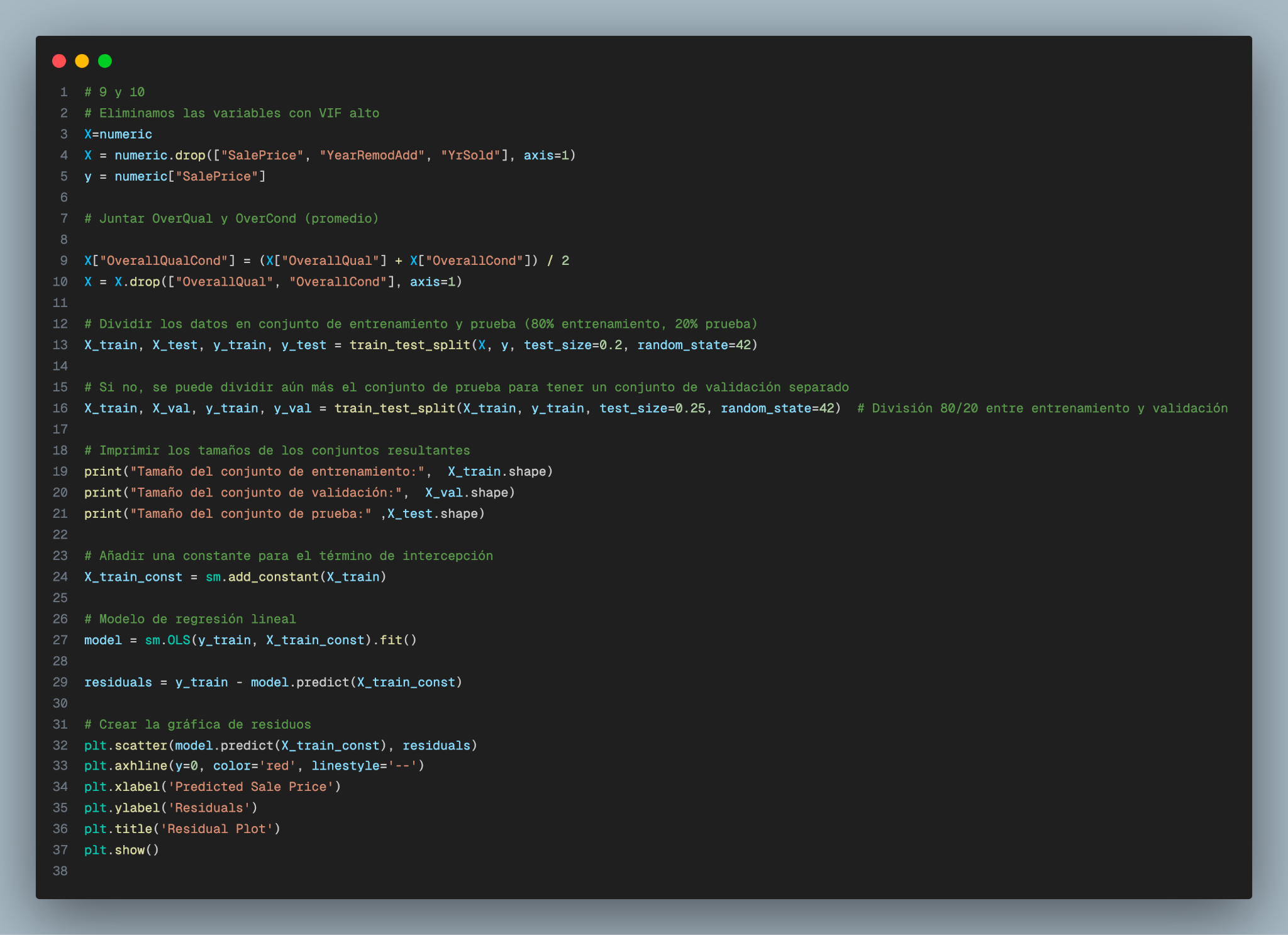
El código anterior tiene como propósito 3 cosas:

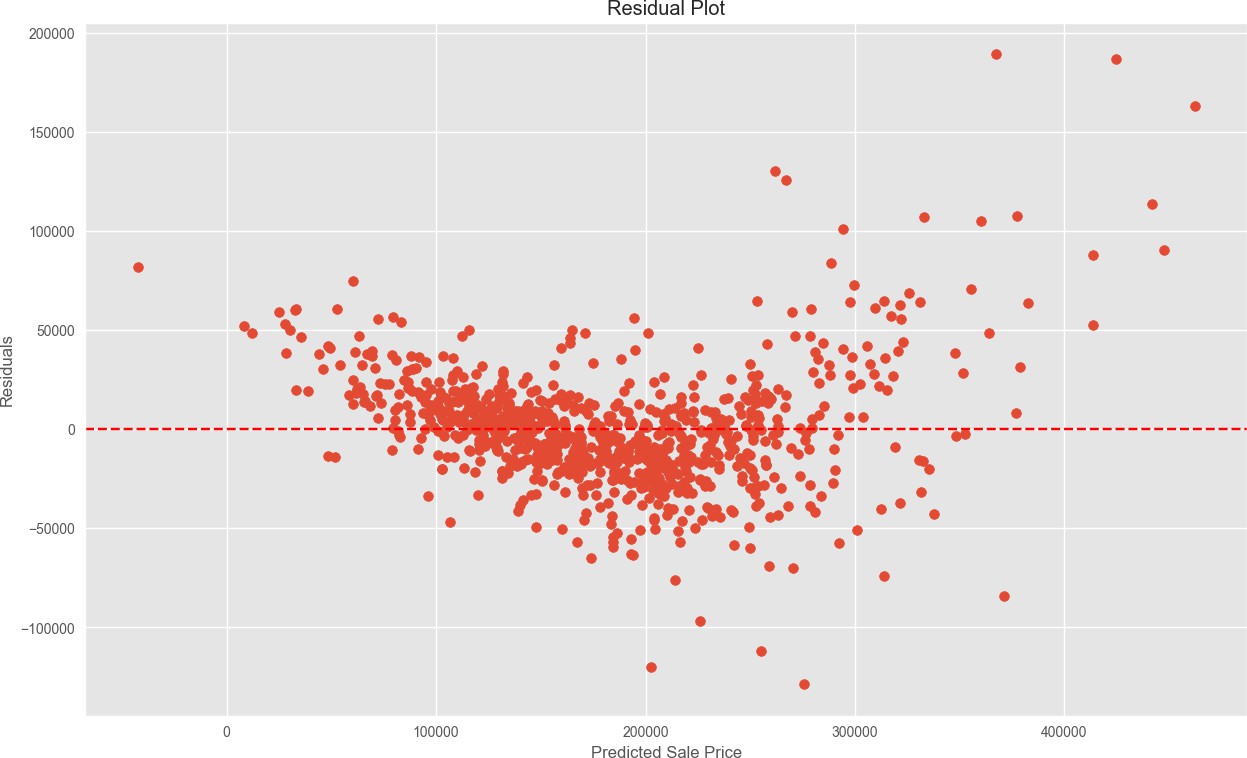
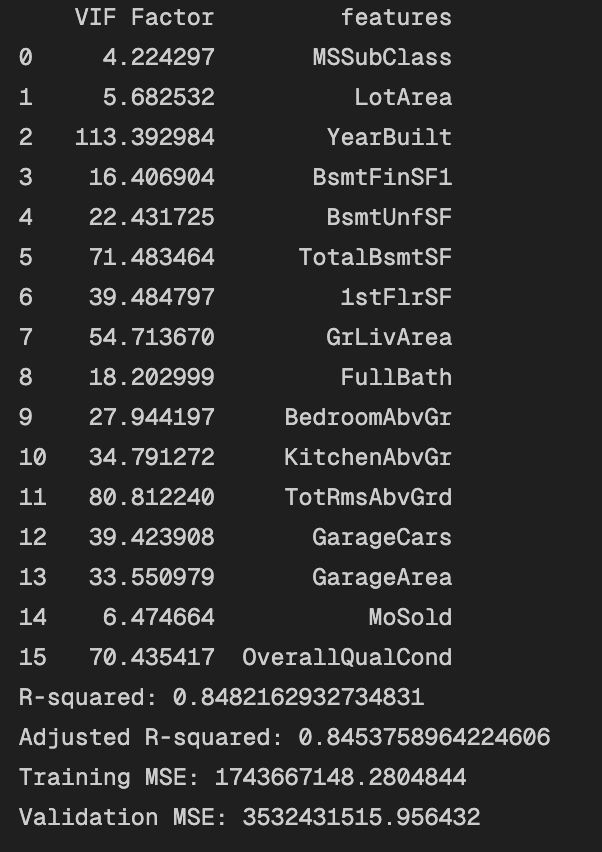
1. Calcula el Factor de Inflación de la Varianza (VIF) para cada una de las variables predictoras de nuestro modelo
2. Calcula el R-cuadrado y el R-cuadrado ajustado para evaluar qué tan bien el modelo se ajusta a los datos.
3. Calcular el error cuadrático medio (MSE) para el conjunto de entrenamiento y el conjunto de validación.

Los resultados obtenidos son los siguientes:



Un VIF mayor a 10 sugiere alta colinealidad entre la variable en cuestión y otras variables predictoras. Varias variables en nuestro modelo tienen VIFs bastante altos, lo que indica la presencia de colinealidad. Además, el error cuadrático medio del conjunto de validación es significativamente mayor que el del conjunto de entrenamiento, lo que definitivamente indica que hay sobreajuste en el modelo.

Para tratar de reducir la multicolinealidad y el sobreajuste, se eliminar ciertas columnas (a nuestro criterio) y se resumieron otras:

YearRemod, YrSold fueron eliminadas (tenían la mayor inflación de varianza), y OverallCond y OverallQual fueron promediadas en una nueva columna. Despúes, se procedió a entrenar el modelo nuevamente y a evaluar gráficamente los resultados:

El overfitting mejoró bastante, dado que el error cuadrático medio del conjunto de validación es mucho más bajo y cercano al de entrenamiento comparado con el modelo anterior. Sin embargo, la multicolinealidad empeoró.

Dado esto decidimos emplear una versión específica de la regresión lineal

## 11 y 12:

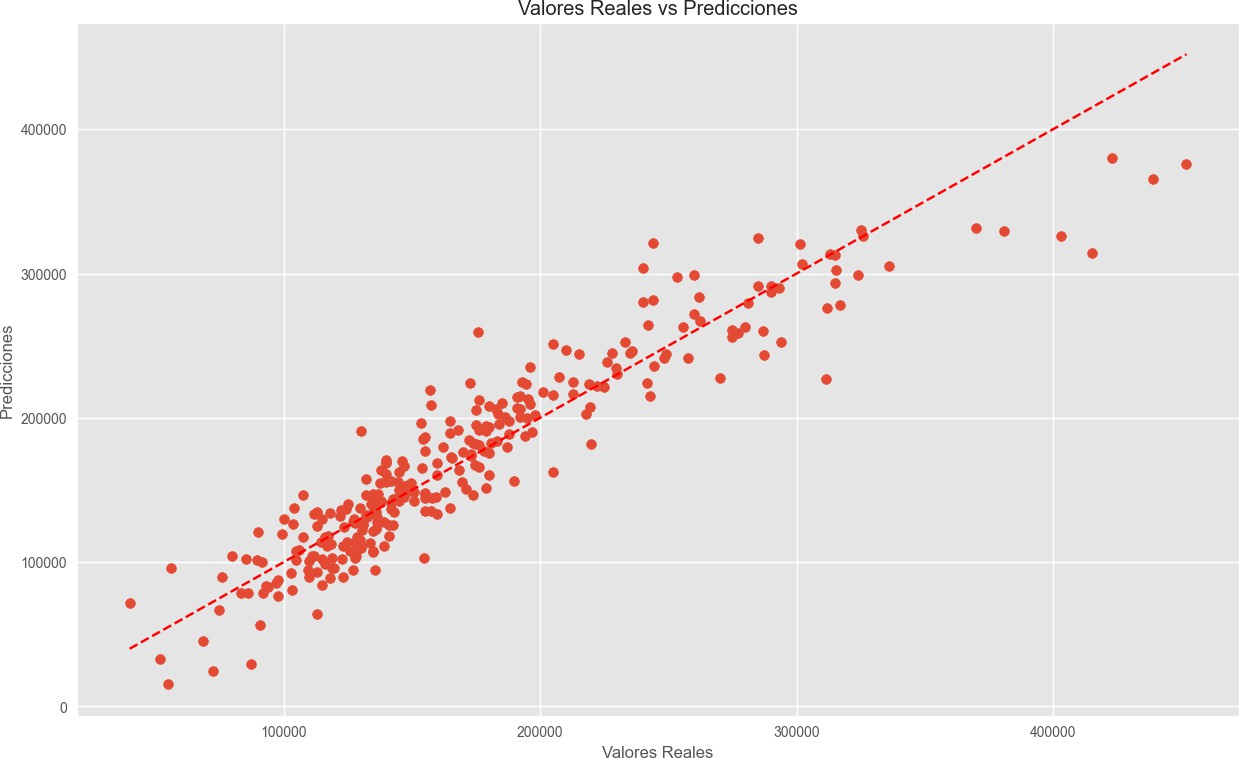
Modelo Lineal con todas las variables (especificadas en el inciso 5):

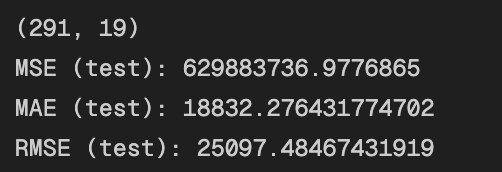
Métricas utilizadas:

MSE (Mean Squared Error): Es el promedio de los cuadrados de los errores.

MAE (Mean Absolute Error): Es el promedio de los valores absolutos de los errores.

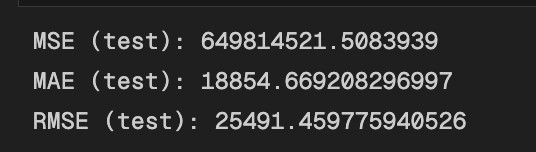
RMSE (Root Mean Squared Error): Es la raíz cuadrada del MSE

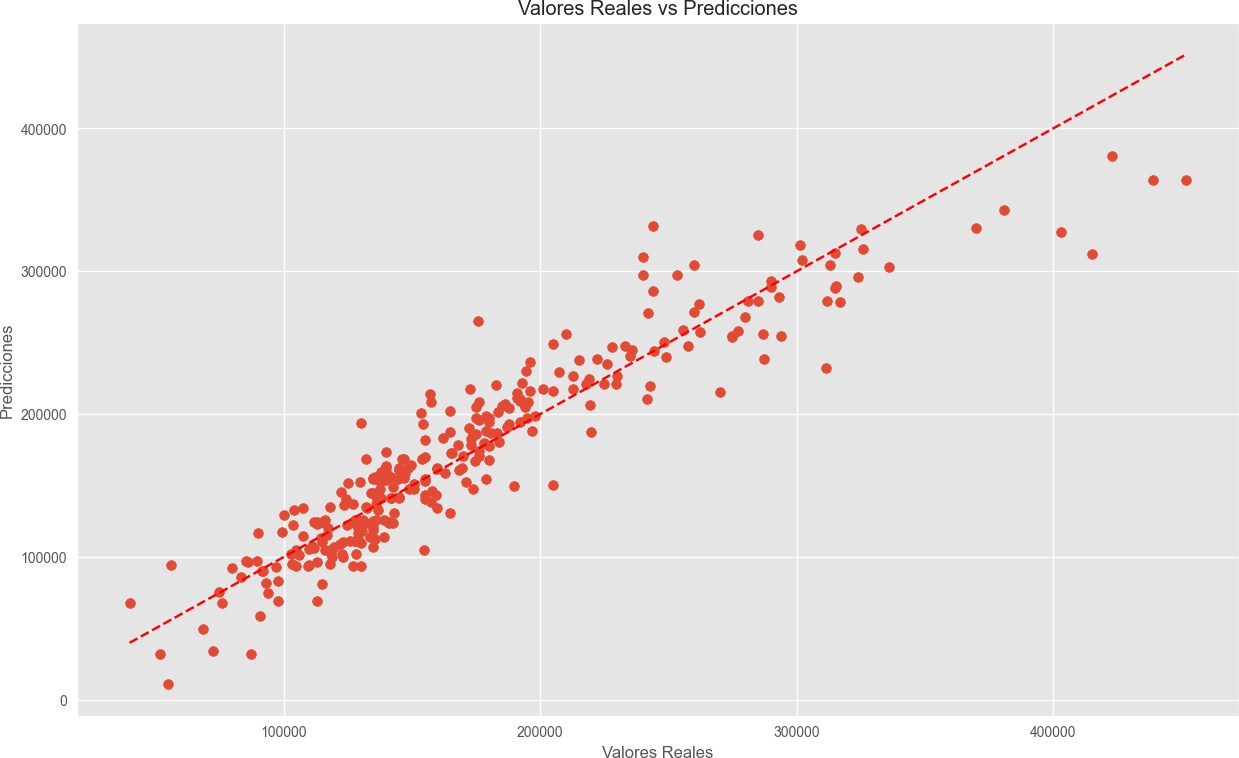




El MAE es aproximadamente 18,832.27. Esto significa que, en promedio, las predicciones de este modelo están a $18,832.27 del valor real.

Model Lineal ajustado:





El MAE es aproximadamente 18,854.67. Esto significa que, en promedio, las predicciones de este modelo están a $18,854.67 del valor real.

A pesar de que el modelo ajustado haya mejorado overfitting, el original sigue siendo mejor prediciendo el precio de las casas, aunque no por mucho. Ambas gráficas son bastante similares, y lastimosamente los dos modelos que realizamos se quedan cortos cuando los valores tienden a los extremos, lo que indica que hay áreas de mejora en la selección inicial de los datos, o bien, podríamos haber realizado normalización de los mismos para lidiar con los valores extremos.