



Instituto Tecnológico
de Buenos Aires

TRABAJO FINAL

ANÁLISIS PREDICTIVO

NICOLÁS RODRIGUES DA CRUZ

01 Caso de negocio

Objetivos del trabajo y caso de negocio

02 Análisis exploratorio

Análisis de la variable objetivo y composición de la base

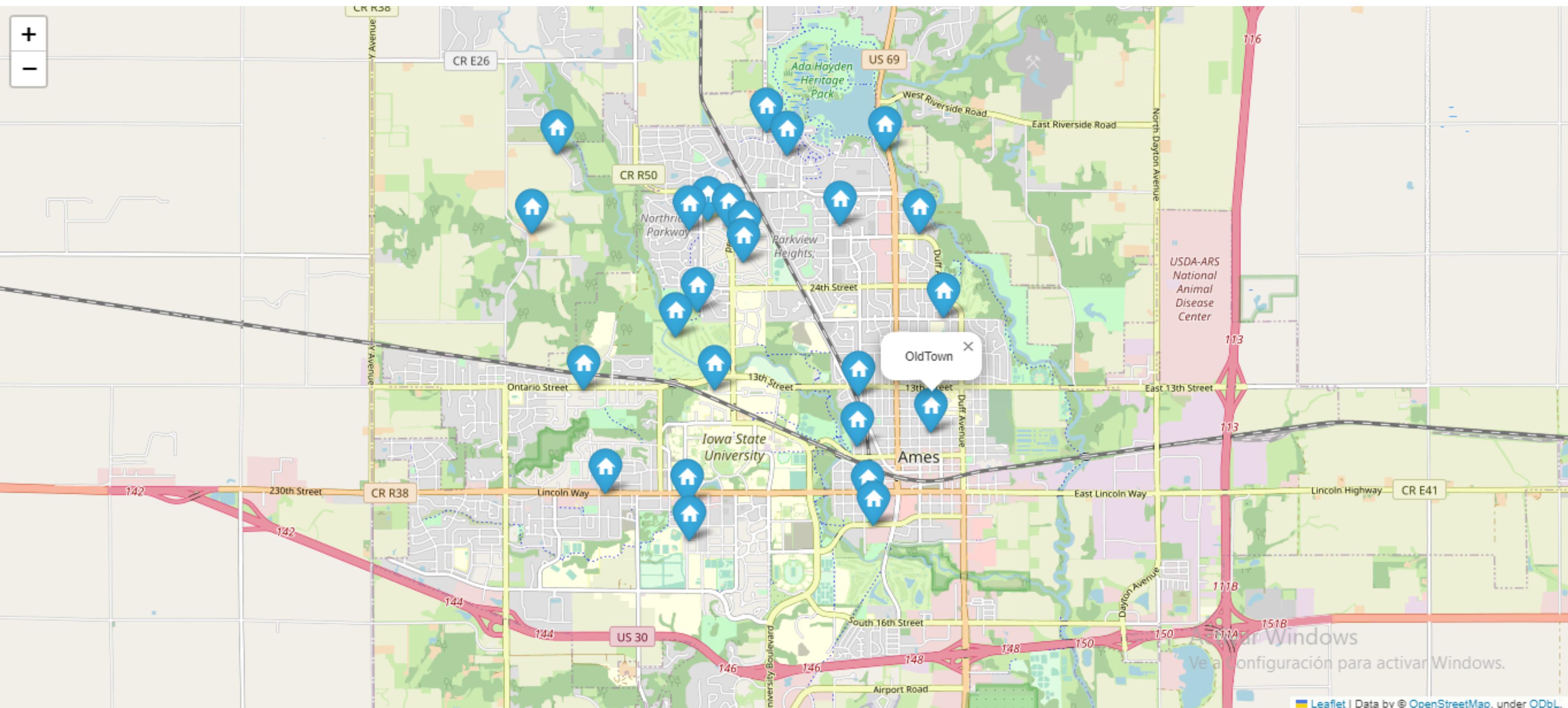
03 Tratamiento del dataset

Transformaciones realizadas sobre la base, imputación de nulls, missings y outliers

04 Modelos

Modelos predictivos utilizados

05 Conclusiones y resultados



Caso de negocio

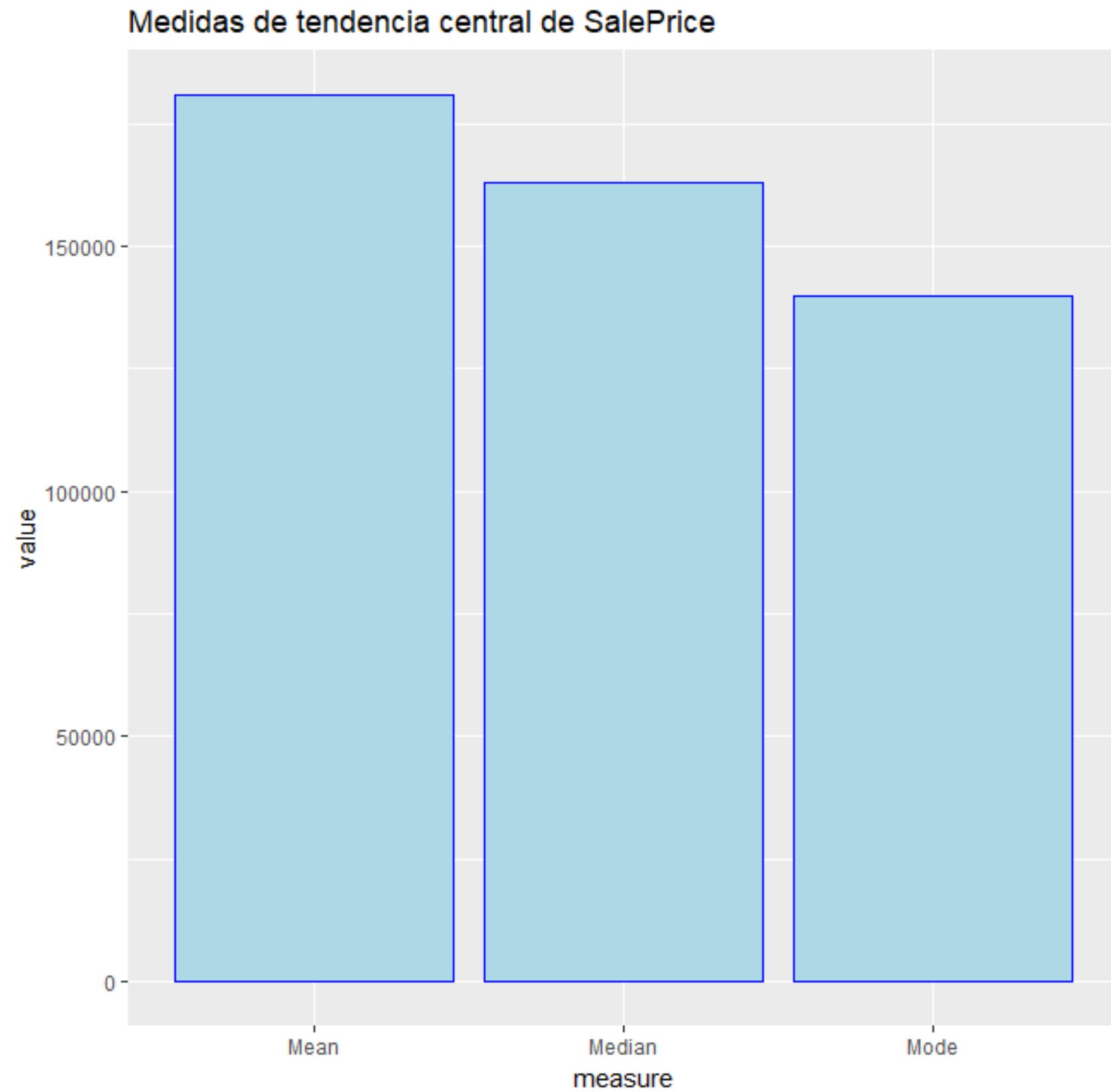
Como objetivo se busca poder predecir el precio de las casas en la ciudad de Ames. De esta manera, se mejorará la tasa de estos inmuebles y al mismo tiempo permitirá democratizar la valuación de ellos ya que cualquier persona será capaz de tasar su propiedad, sin la necesidad de pasar por las agencias inmobiliarias.

Análisis exploratorio

<code>Id</code>	<code>MSSubClass</code>	<code>MSZoning</code>	<code>LotFrontage</code>	<code>LotArea</code>	<code>Street</code>	<code>Alley</code>	<code>LotShape</code>	<code>LandContour</code>	<code>Utilities</code>
1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub
2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub
3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub
4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub
5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub
6	50	RL	85.0	14115	Pave	NaN	IR1	Lvl	AllPub
7	20	RL	75.0	10084	Pave	NaN	Reg	Lvl	AllPub

La base cuenta con **1460** filas y **81** columnas (43 variables numéricas y 38 categóricas)

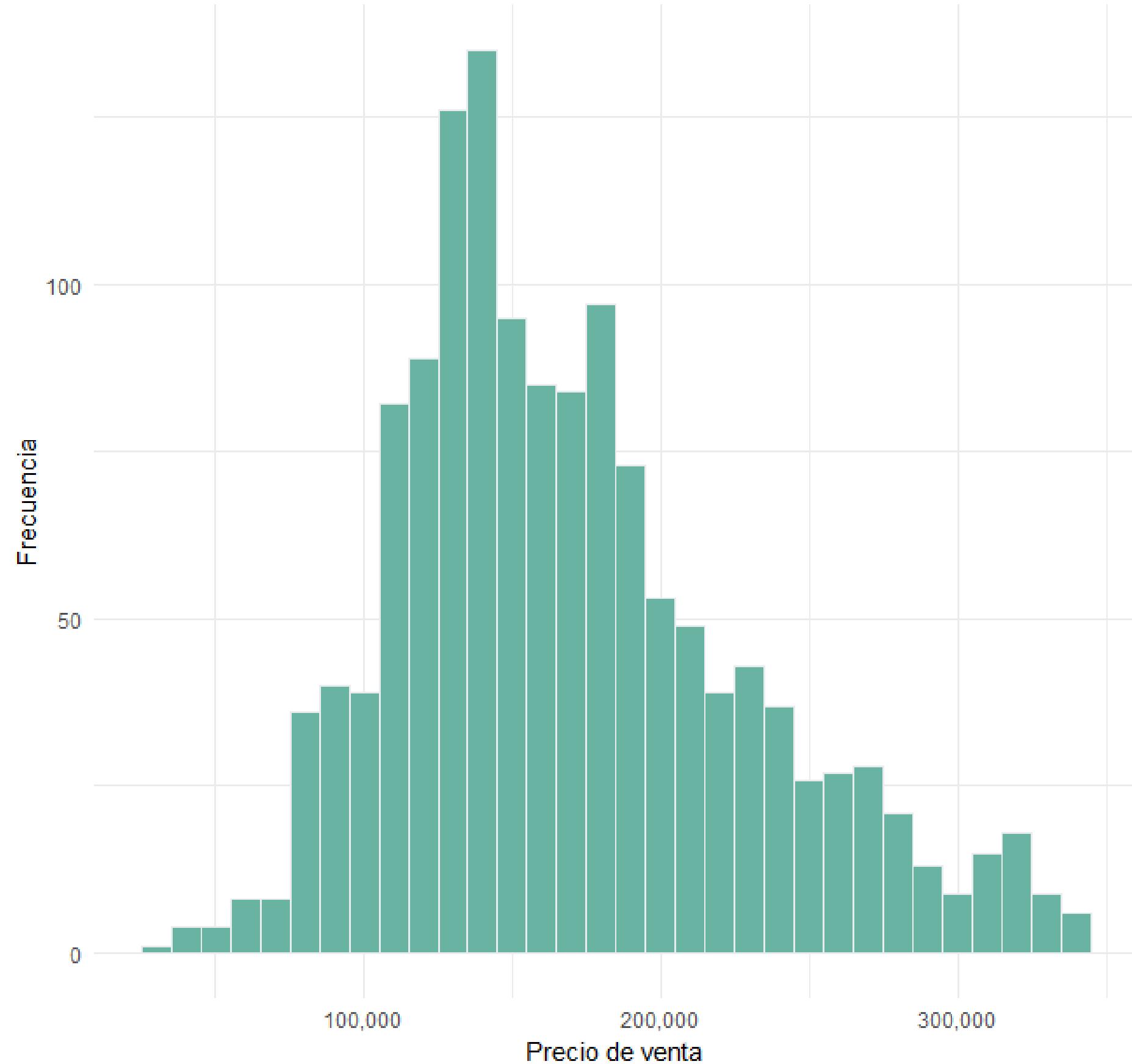
Variable objetivo: SalePrice



Media: 180921
Mediana: 163000.0
Moda: 140000

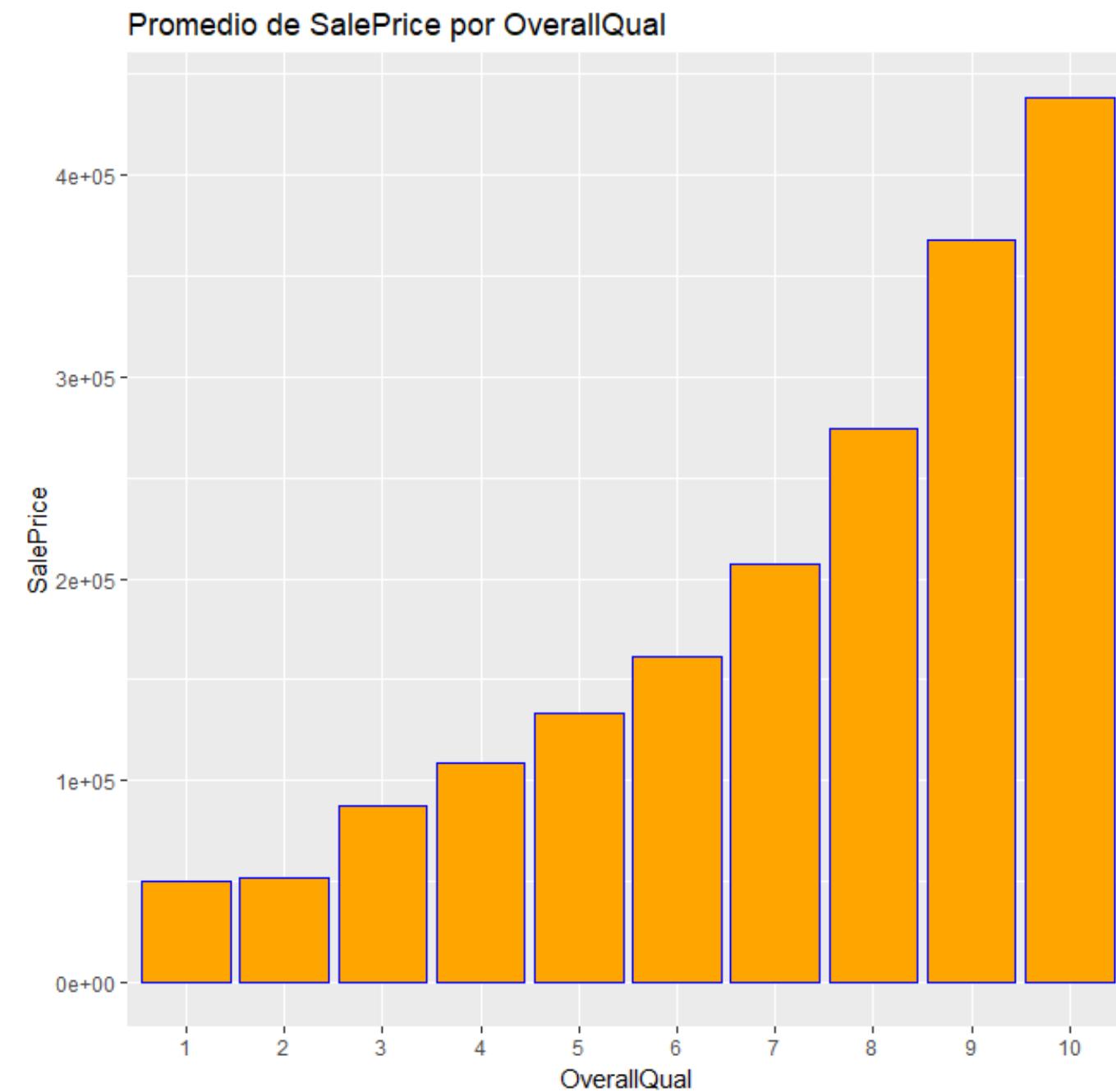
Variable objetivo: SalePrice

Distribución de los precios de las casas

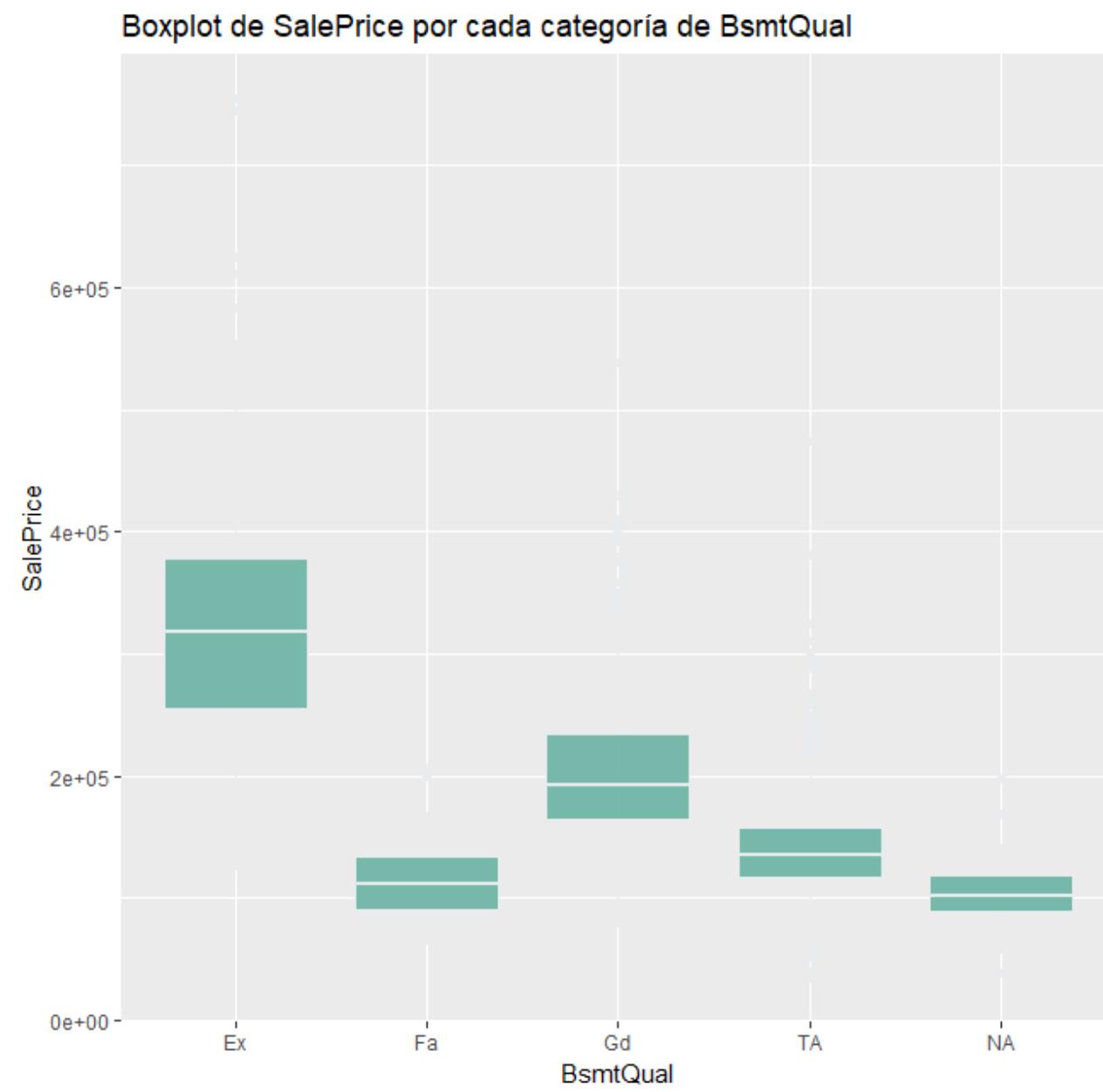


Variable objetivo: SalePrice

Tamaño del efecto con variables categóricas:



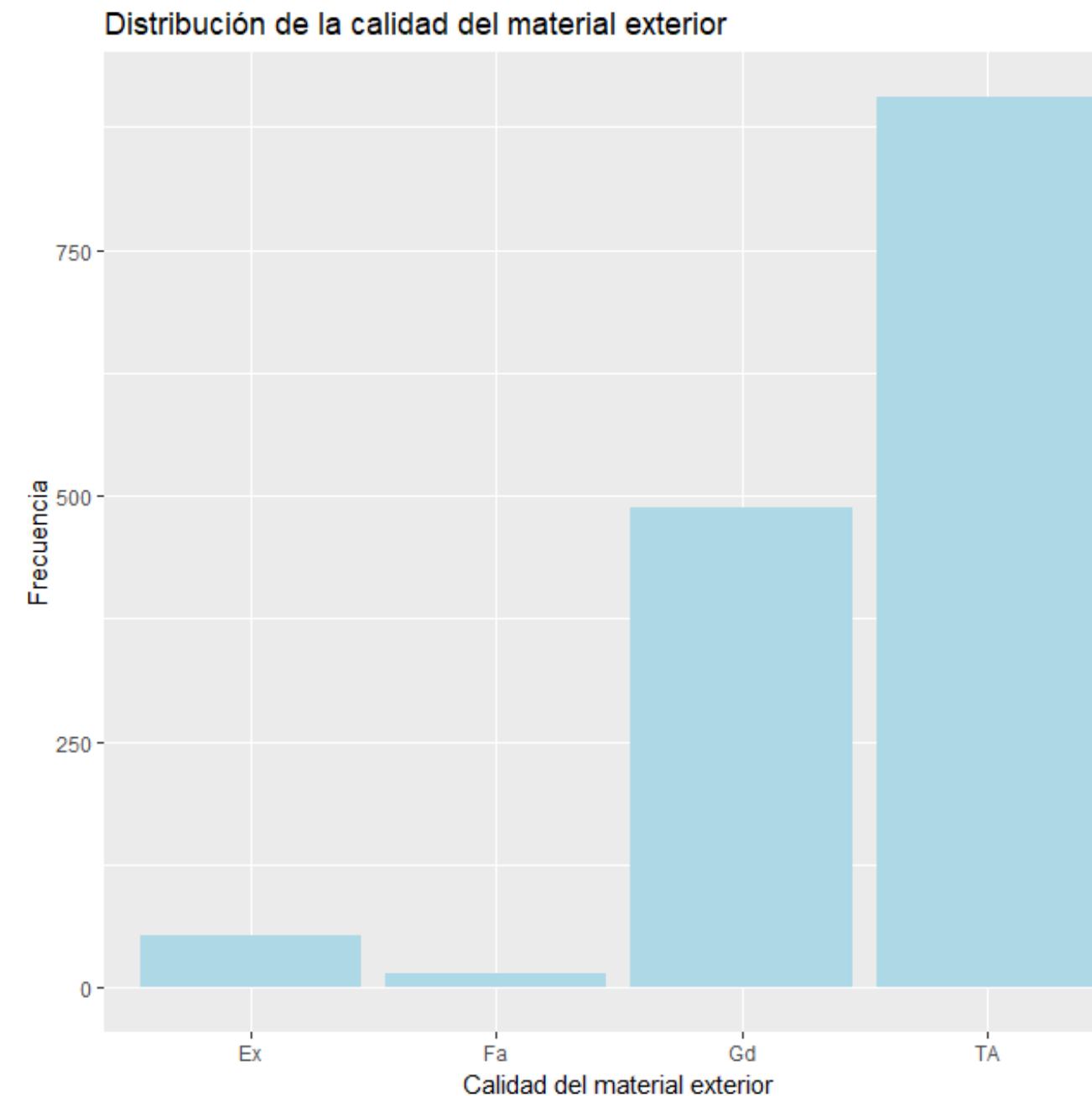
0.682



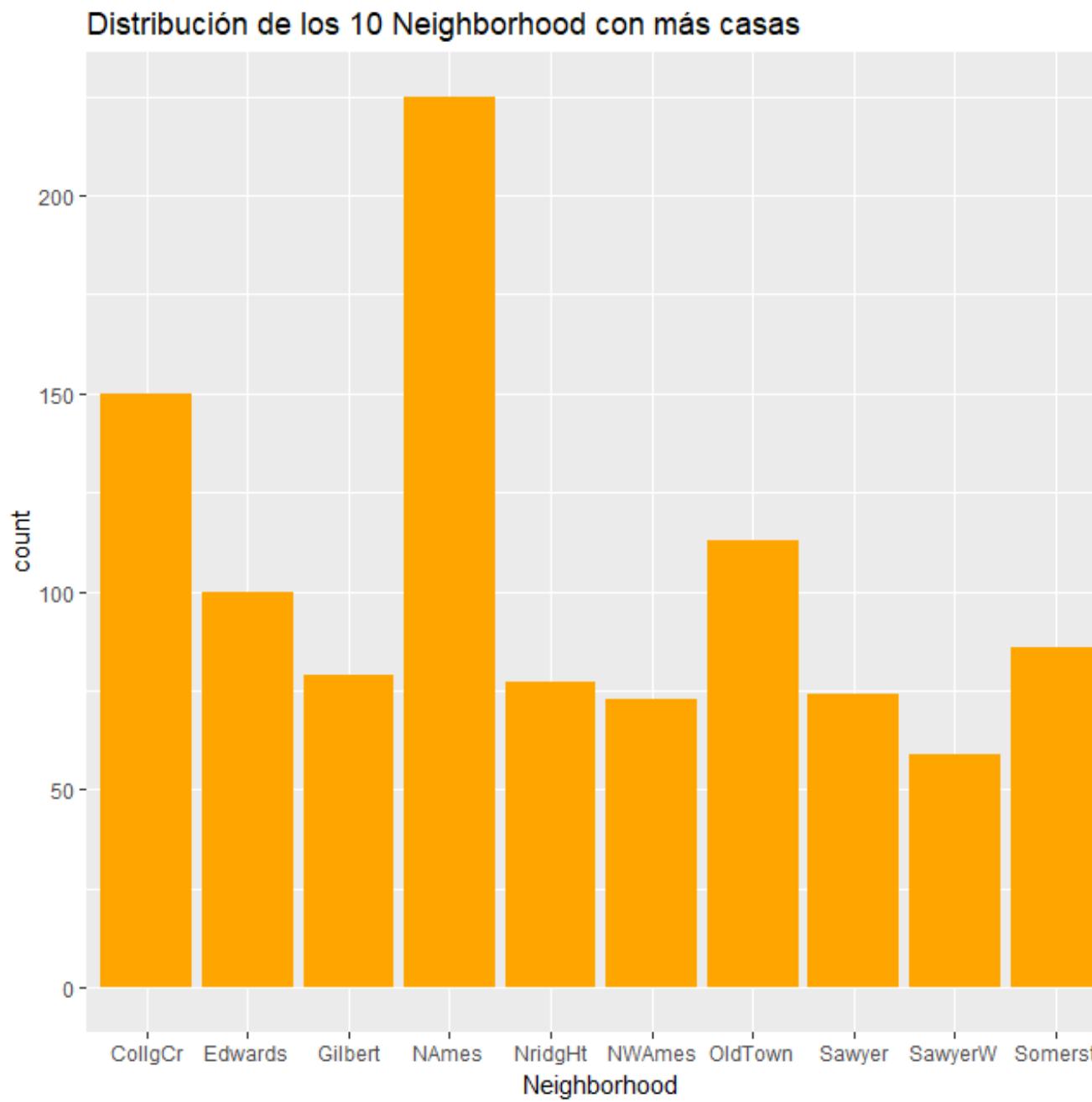
0.452

Variable objetivo: SalePrice

Tamaño del efecto con variables categóricas:



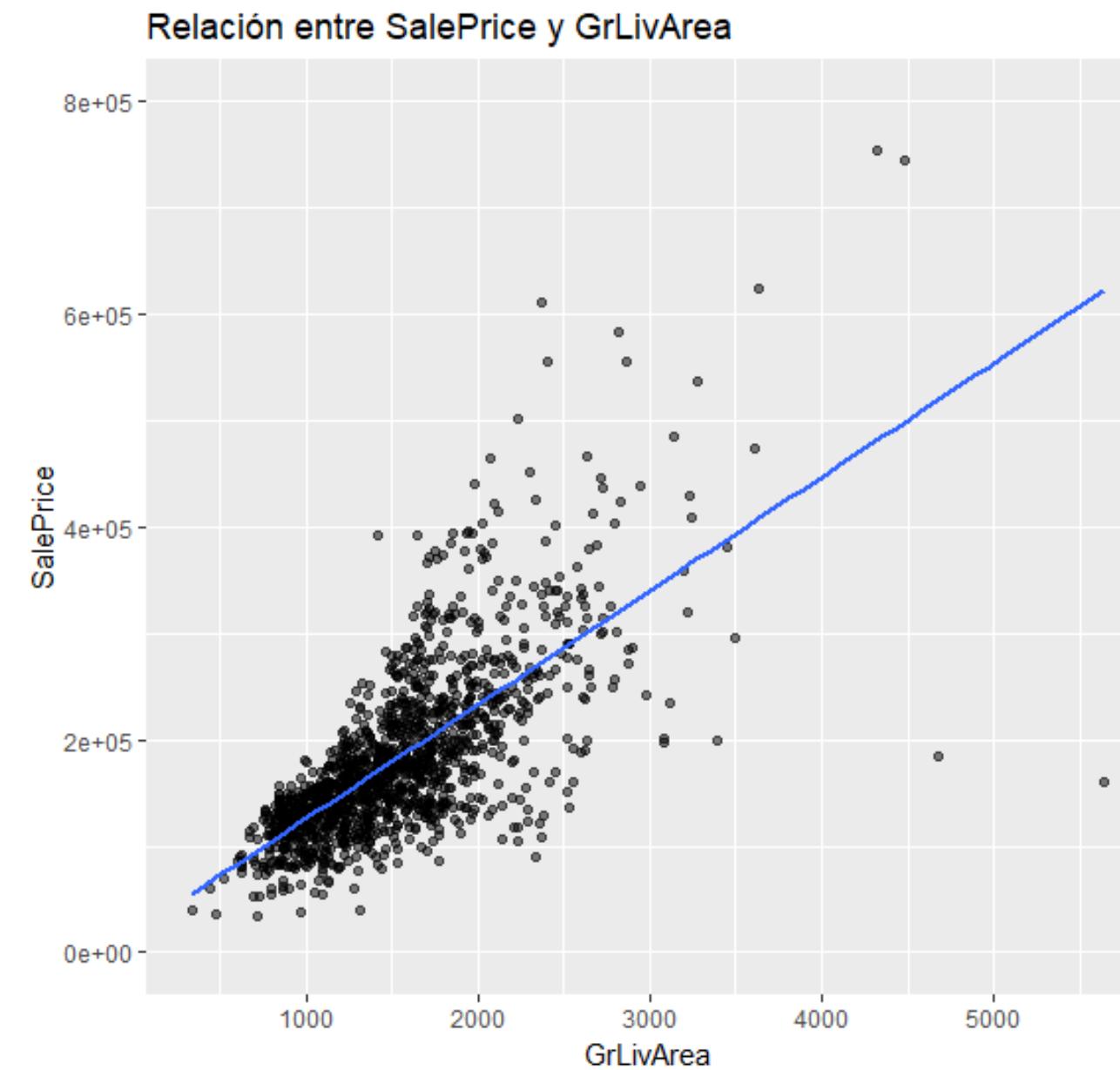
0.476



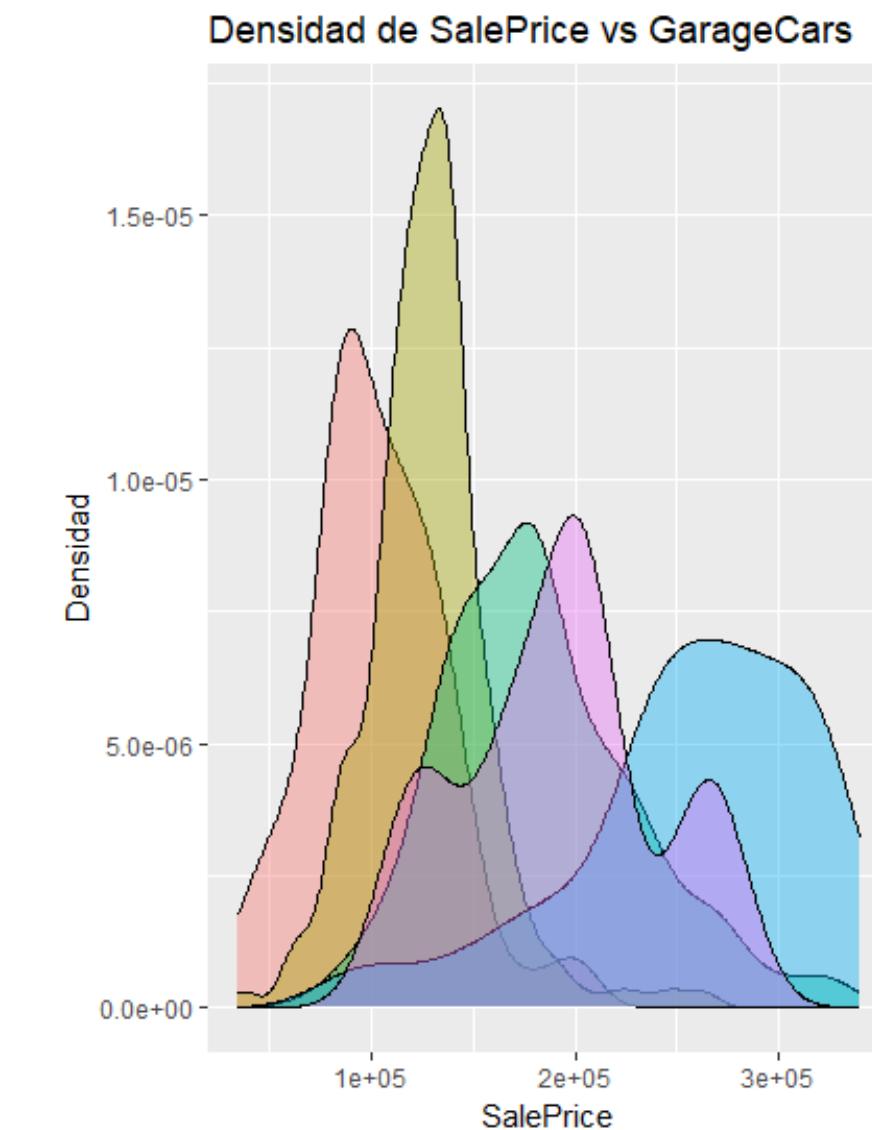
0.538

Variable objetivo: SalePrice

Correlación con variables numéricas:



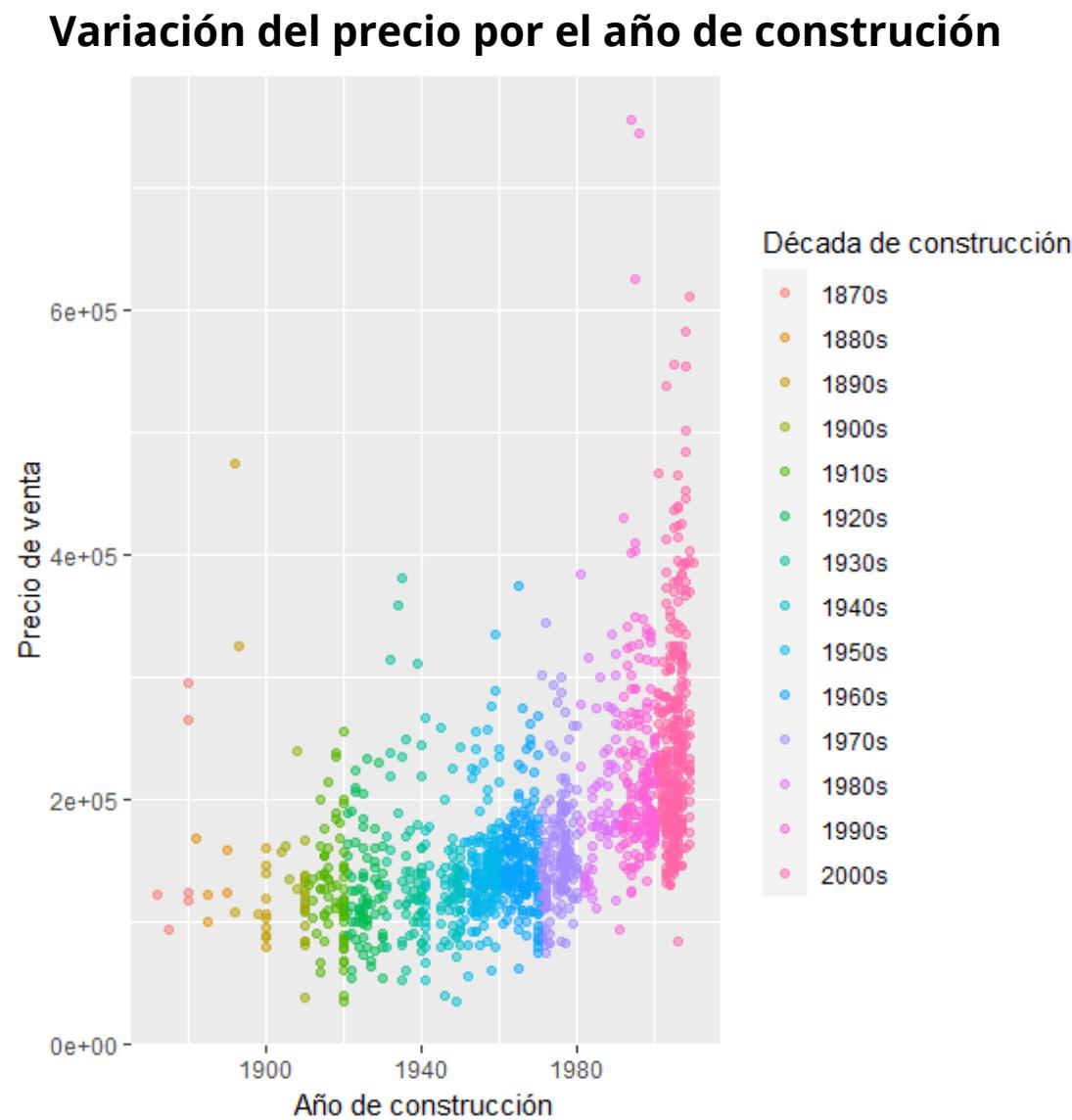
0.69



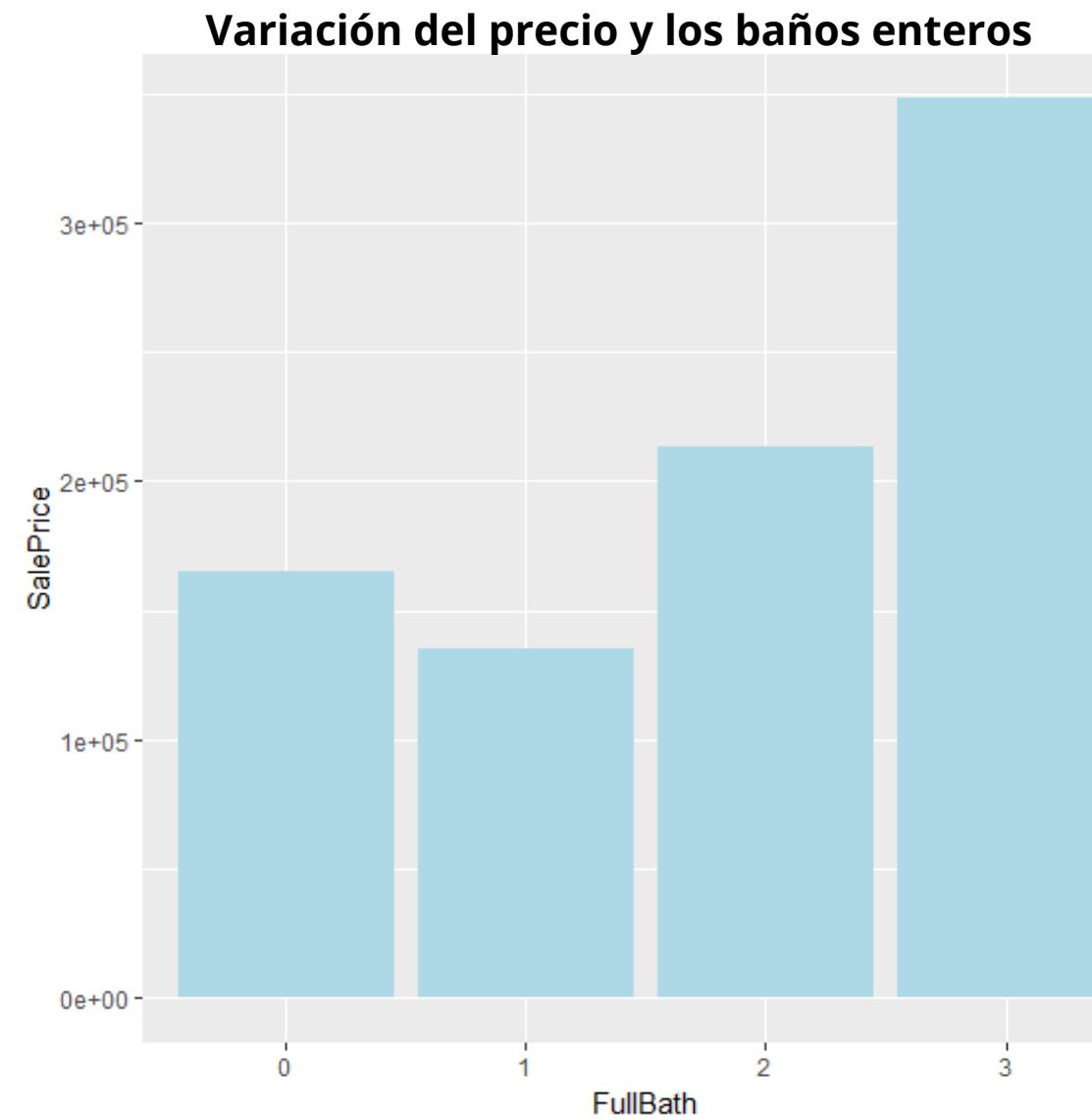
0.731

Variable objetivo: SalePrice

Correlación con variables numéricas:

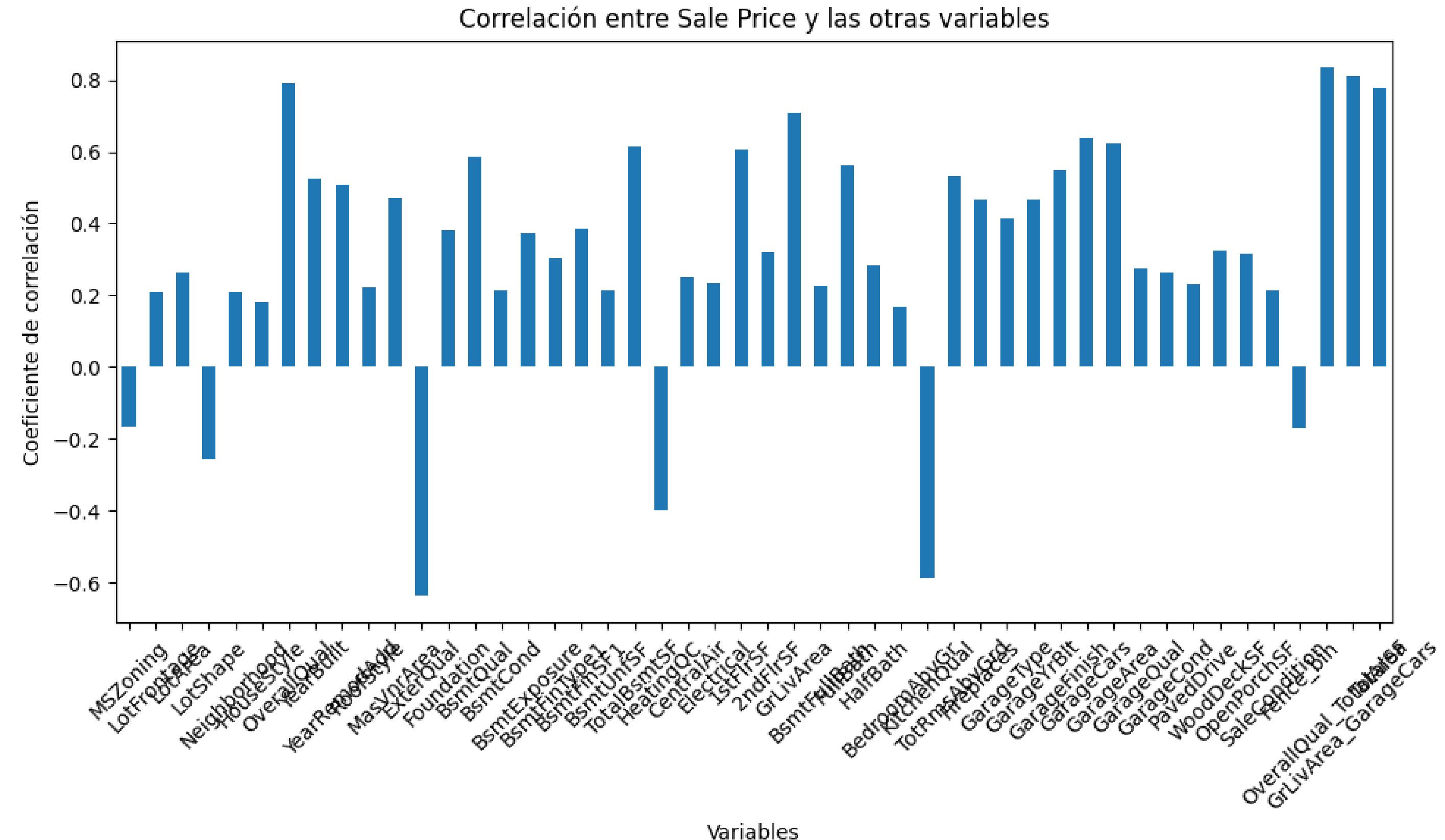


0.652

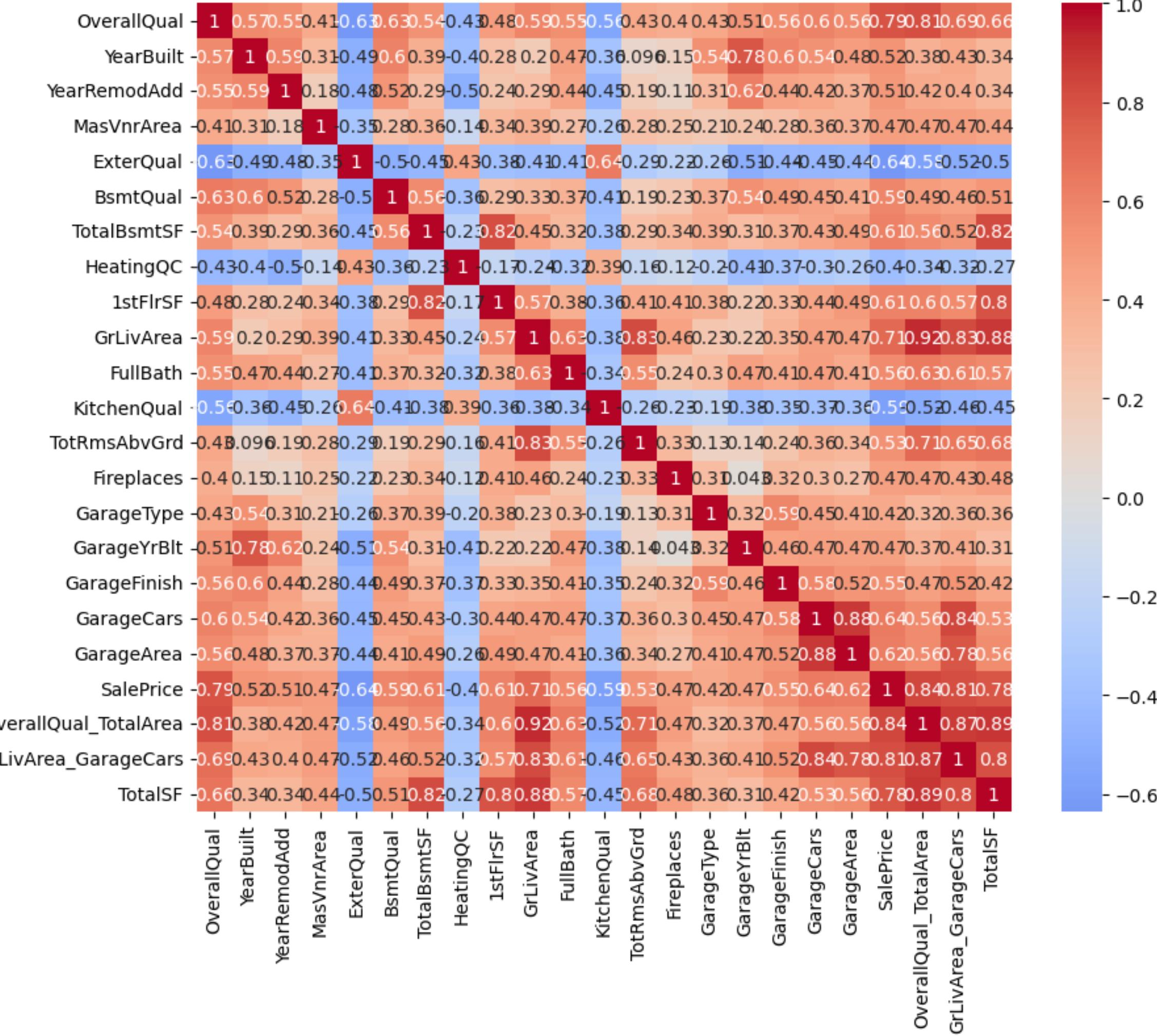


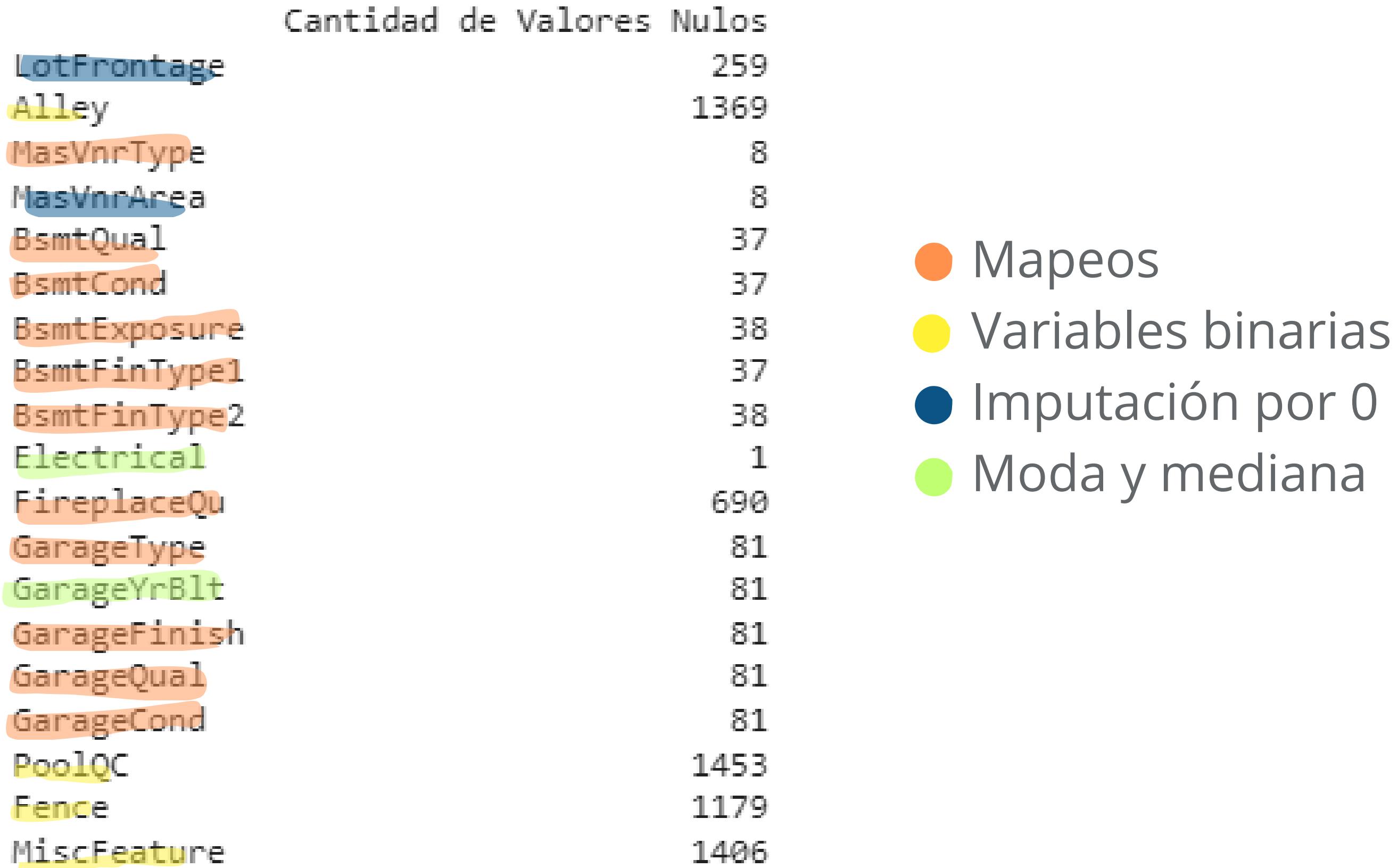
0.635

Variable objetivo: SalePrice



Matriz de Correlación





Mapeos

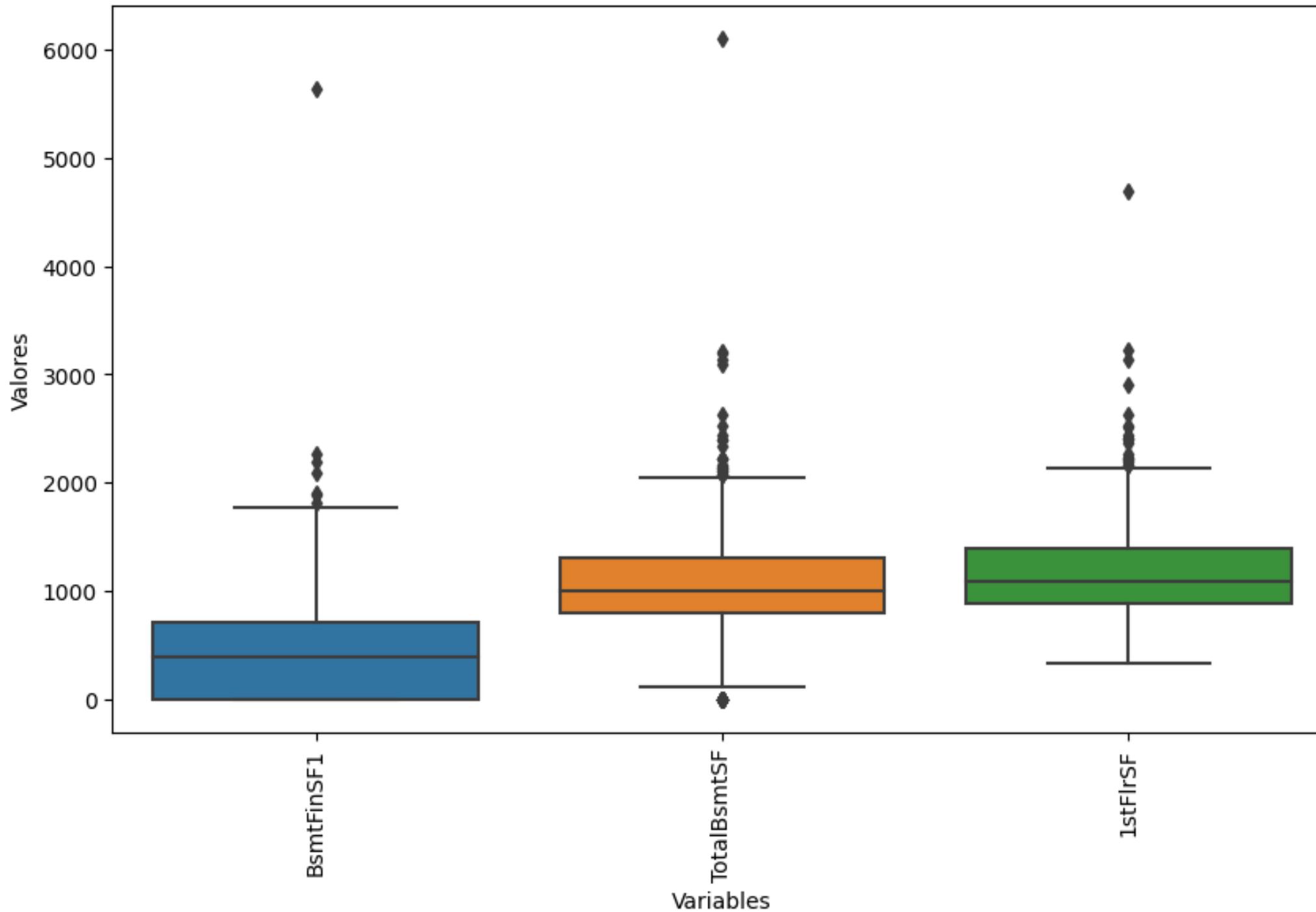
FireplaceQu: Fireplace quality

Ex	Excellent - Exceptional Masonry Fireplace
Gd	Good - Masonry Fireplace in main level
TA	Average - Prefabricated Fireplace in main living area or Masonry Firep
Fa	Fair - Prefabricated Fireplace in basement
Po	Poor - Ben Franklin Stove
NA	No Fireplace



```
mapping = {'Ex': 5, 'Gd': 4, 'TA': 3, 'Fa': 2, 'Po': 1}
train['FireplaceQu'] = train['FireplaceQu'].map(mapping)
train['FireplaceQu'] = train['FireplaceQu'].fillna(0)
```

Tratamiento del dataset: outliers



Si bien reconozco la presencia de este valor outlier no lo elimino ya que asumo que es posible que exista una casa con estas características

Tratamiento del dataset: creación de variables y encoding

→ Se crearon variables como:

- Antigüedad
- m2_techados
- TotalFloorsArea
- Año_remodelacion
- Relacion_Calidad_AreaTotal
- Relacion_GrLivArea_GarageCars

→ Para las variables categoricas como **Neighborhood** se utilizó LabelEncoder() de la librería sklearn.preprocessing

Modelos

Se probaron los modelos:

- XGBoost,
- Random Forest,
- Cat boost,
- Regresion Ridge,
- LinearRegression

Para la prueba de los modelos se utilizó `train_test_split` de la biblioteca `sklearn.model_selection`. Se destinó un 15% al conjunto de testeo.

Modelos

- El modelo que consiguió el mejor score fue **XGBoost**.
- Para la elección de los hiperparámetros se utilizó Grid Search y Cross Validation de la librería `sklearn.model_selection`

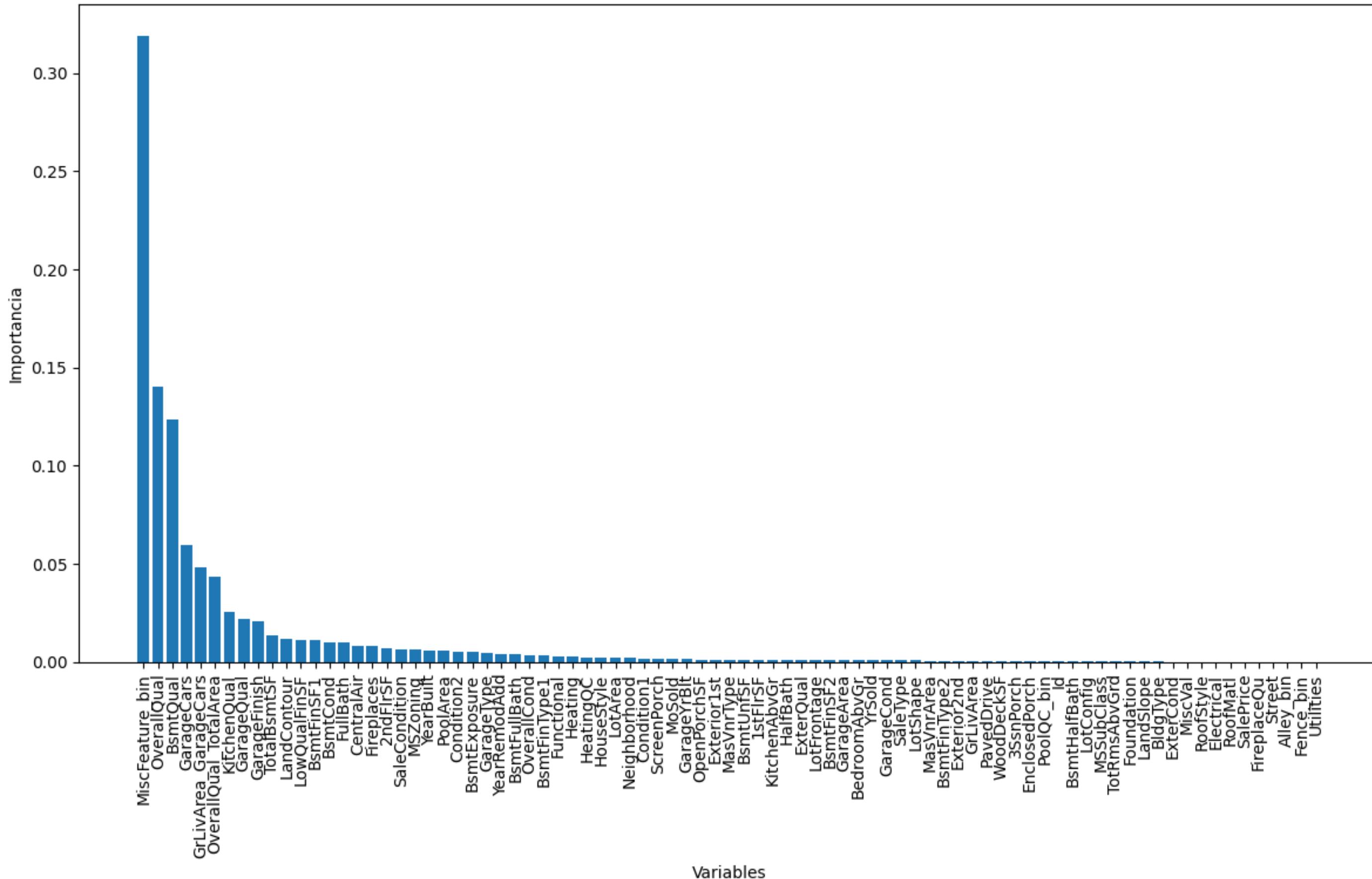


```
parametros = {
    'nthread': [4],
    'learning_rate': [0.1, 0.5],
    'n_estimators': [100],
    'max_depth': [2, 3, 4, 5],
    'min_child_weight': [1],
    'subsample': [0.8],
    'colsample_bytree': [0.8]
}
model = XGBRegressor()
grid_search = GridSearchCV(model, parametros, cv=5, scoring='r2', return_train_score=True)
```

Modelos

Para mejorar la performance del modelo se eliminaron las variables que no tenían importancia y se volvió a entrenar el modelo

Importancia de las variables



Resultado obtenido en prueba:

Puntajes de validación cruzada: [0.88547049 0.88330435 0.87087988 0.86074525 0.88284911]

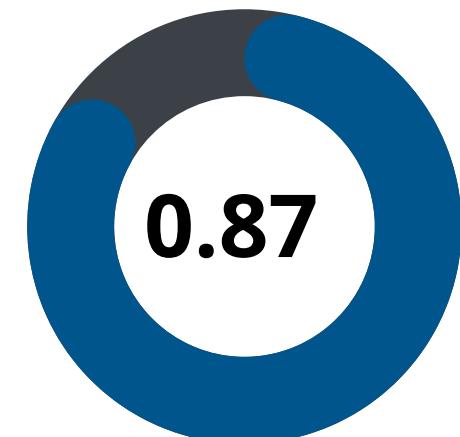
Score: 0.876649816114883

Desviación estándar: 0.009452680130450438

Kaggle:

881	NICOLAS RODRIGUES DA CRUZ		0.12805	26	2h
-----	---------------------------	---	---------	----	----

Posicion 881 de 4700



Conclusiones y resultados