Universidad del Valle de Guatemala
Facultad de Ingeniería
Minería de Datos

HT 3. Modelos de regresión lineal

Entrega # 1

Autor/Carnet:

María José Villafuerte 22129
Fabiola Contreras 22787

Catedrático:

Mario Barrientos

Sección 20

Fecha:

02/03/2025

Índice

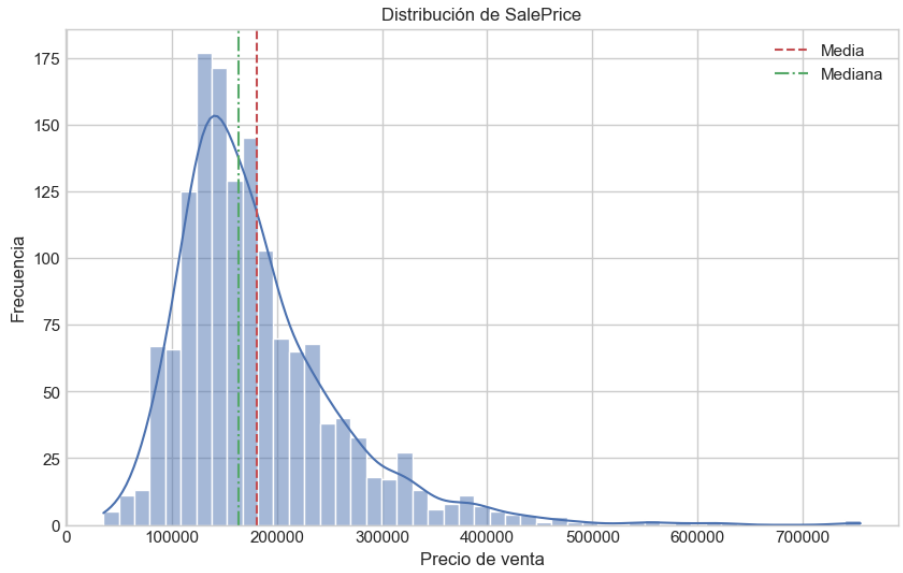
I. Introducción

El presente trabajo desarrolla un modelo predictivo de precios inmobiliarios para InmoValor S.A., una empresa especializada en valoración de propiedades. A través de análisis exploratorio, segmentación y técnicas de regresión lineal, esta investigación busca determinar con precisión los factores que influyen significativamente en el valor de mercado de las viviendas. Utilizando un conjunto de datos de 1,460 propiedades con 81 características, se identifican patrones fundamentales que permiten predecir precios con mayor exactitud y objetividad que las valoraciones tradicionales basadas en criterios subjetivos.

La metodología empleada combina análisis estadístico avanzado con la interpretación práctica del mercado inmobiliario, progresando desde modelos univariados simples hasta un modelo optimizado que equilibra complejidad y precisión. Los resultados demuestran que variables como la calidad general, el área habitable, la ubicación y ciertas características específicas como garajes y sótanos son determinantes clave del precio. Este enfoque no solo mejora la capacidad predictiva, sino que también proporciona insights valiosos sobre la interacción entre diferentes características de las propiedades y su impacto en el valor final de mercado.

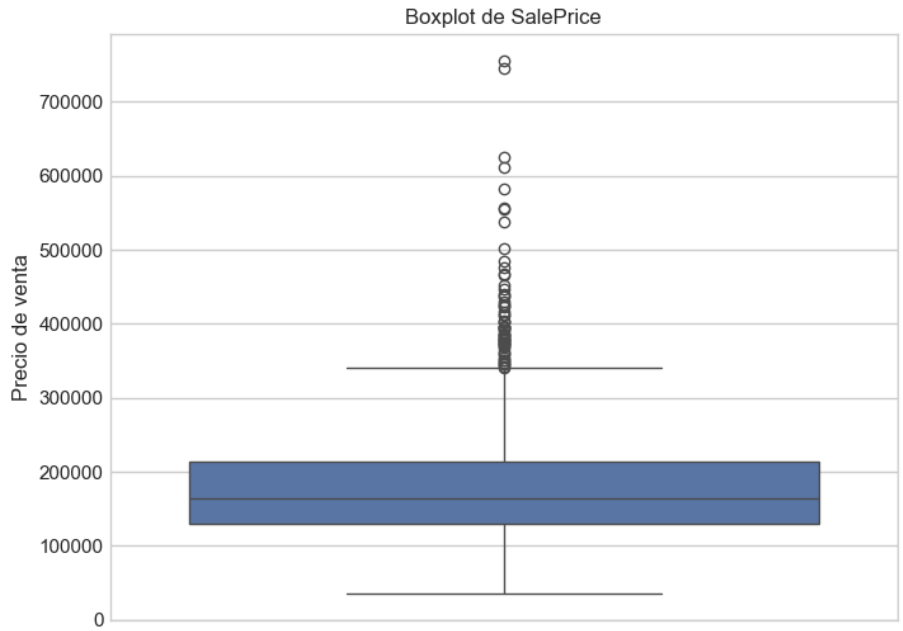
II. Análisis exploratorio

a. El primer paso es conocer la data, lo que logramos con una lectura y exploración breve
    i. El conjunto de datos de entrenamiento contiene 1460 registros de casas con 81 columnas, mientras que el conjunto de prueba tiene 1459 registros con 80 columnas (falta la columna SalePrice que es la variable objetivo).
    ii. Hay 35 variables numéricas (tipos int64) y 43 variables categóricas (tipo object)
    iii. El precio de venta (SalePrice) varía desde $34,900 hasta $755,000, con un promedio de $180,921
    iv. Las casas fueron construidas entre 1872 y 2010, con una media en 1971
    v. La mayoría de las casas tienen 2-3 dormitorios y 1-2 baños completos
    vi. Las casas tienen un área habitable promedio (GrLivArea) de 1,515 pies cuadrados
b. Evaluamos si hay algún dato faltante
    i. Existen valores faltantes en varias columnas, como LotFrontage (259 valores nulos), MasVnrType (872 valores nulos), y PoolQC (1453 valores nulos)
    ii. Los valores faltantes nos indican que estos valores corresponden a "No aplica" y no a información realmente perdida.
c. Hemos identificado nuestra variable objetivo como SalesPrices, ya que con este estudio buscamos obtener un precio estándar a los precios del mercado. El análisis de mercado revela que
    i. La variable SalePrice presenta un rango amplio desde $34,900 hasta $755,000, indicando gran diversidad en los valores de las propiedades analizadas.
    ii. El precio medio de venta es $180,921, mientras que la mediana es $163,000, sugiriendo una distribución con asimetría positiva (sesgada hacia la derecha).



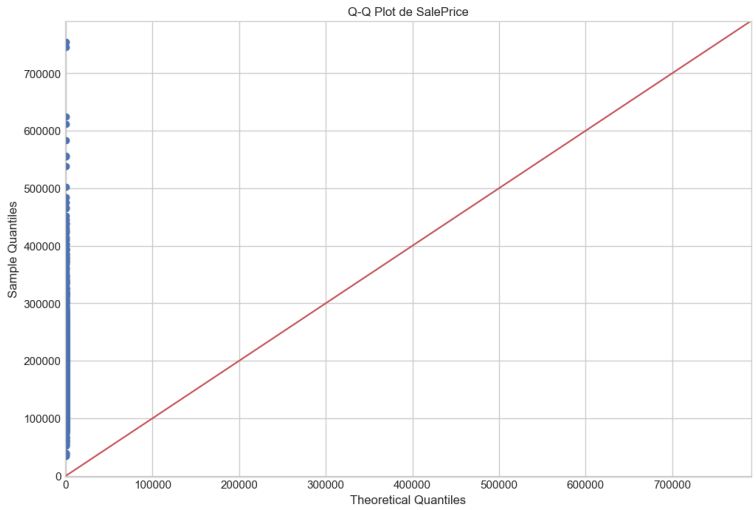
iii. La diferencia entre la media y la mediana ($17,921) confirma la presencia de valores atípicos de alto precio que están elevando el promedio.

iv. El percentil 75 ($214,000) está más cercano a la mediana que el valor máximo ($755,000), lo que refuerza la conclusión de que existen valores extremos en el extremo superior de la distribución.



d. Nos interesa estudiar la normalidad de la variable, como pudimos observar en el inciso anterior

i. El Q-Q Plot de SalePrice muestra una clara desviación de la línea de referencia, indicando una distribución no normal con fuerte asimetría positiva y colas pesadas en los valores altos.



ii. La prueba de Shapiro-Wilk rechaza formalmente la hipótesis de normalidad tanto para SalePrice (W=0.87) como para Log(SalePrice) (W=0.99), con valores p extremadamente pequeños.

1. Prueba de Shapiro-Wilk para SalePrice:
2. Estadístico W: 0.869671
3. Valor p: 0.000000
4. Conclusión: SalePrice no sigue una distribución normal (se rechaza H0)

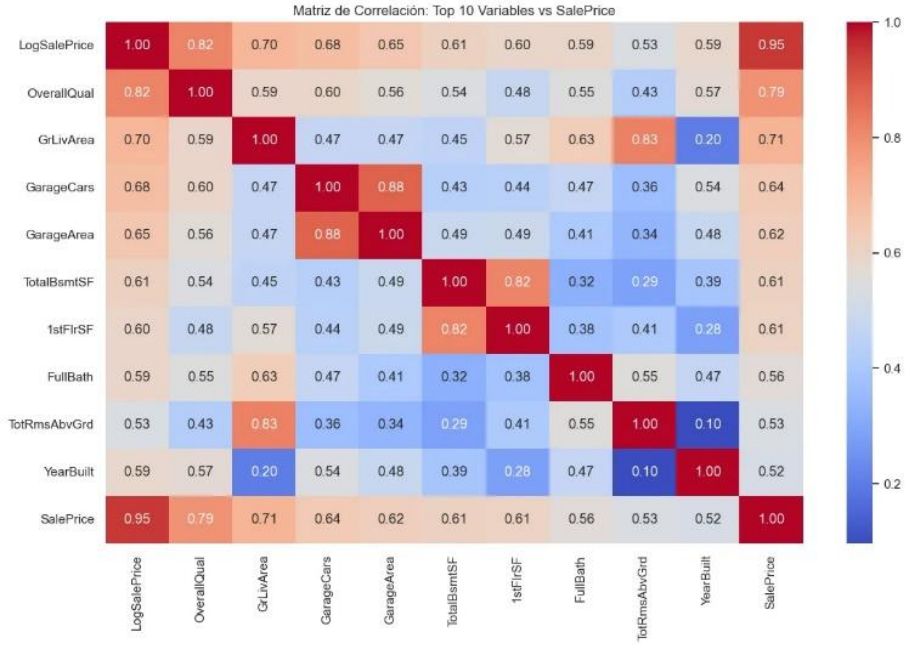
e. Pasemos a responder unas preguntas que pueden llegar a ser de utilidad para el futuro del análisis.

   i. ¿Cuál es la distribución del precio de venta (SalePrice) y muestra alguna asimetría que requiera transformación?

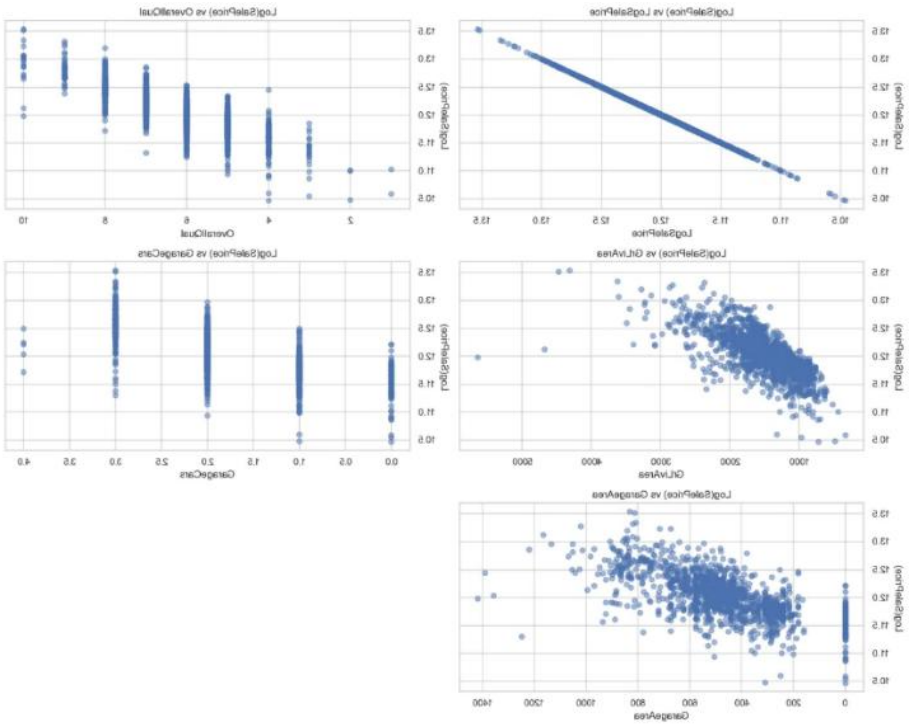
      1. El precio de venta (SalePrice) muestra una asimetría positiva, lo que significa que la distribución está sesgada hacia la derecha con una cola larga que se extiende hacia valores más altos. Esto se confirma en el análisis exploratorio donde se observa que:
      2. La media ($180,921) es mayor que la mediana, indicando esta asimetría
      3. Los gráficos de histograma muestran claramente esta cola alargada hacia la derecha
      4. La prueba de Shapiro-Wilk confirma que SalePrice no sigue una distribución normal

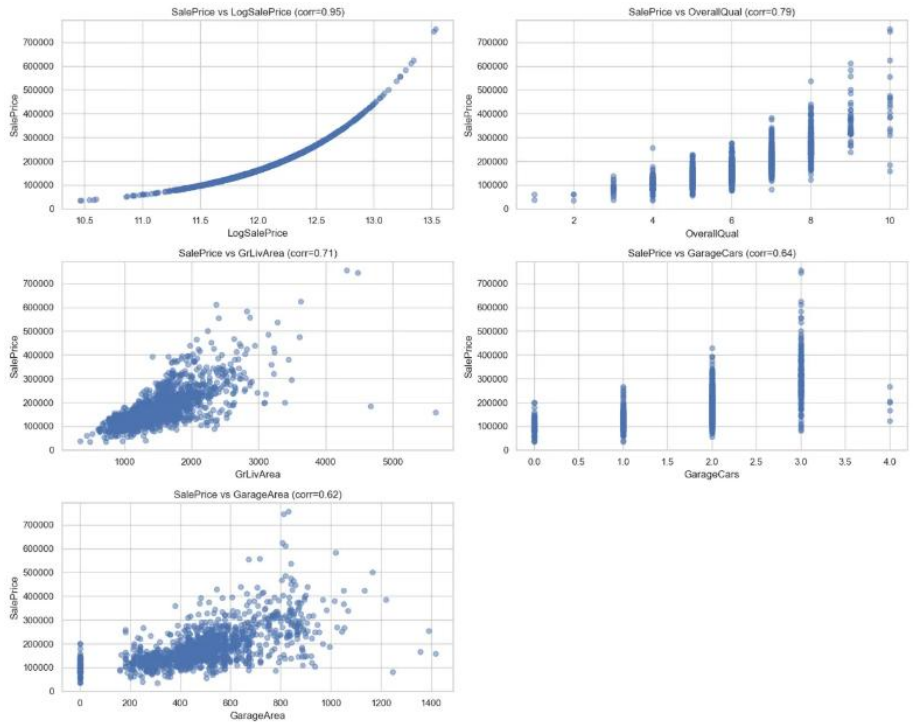
   ii. ¿Qué variables numéricas tienen mayor correlación con el precio de venta y podrían ser buenos predictores?

      1. En la imagen (matriz de correlación) se puede ver claramente que OverallQual tiene una correlación de 0.79 con SalePrice y GrLivArea tiene 0.71.



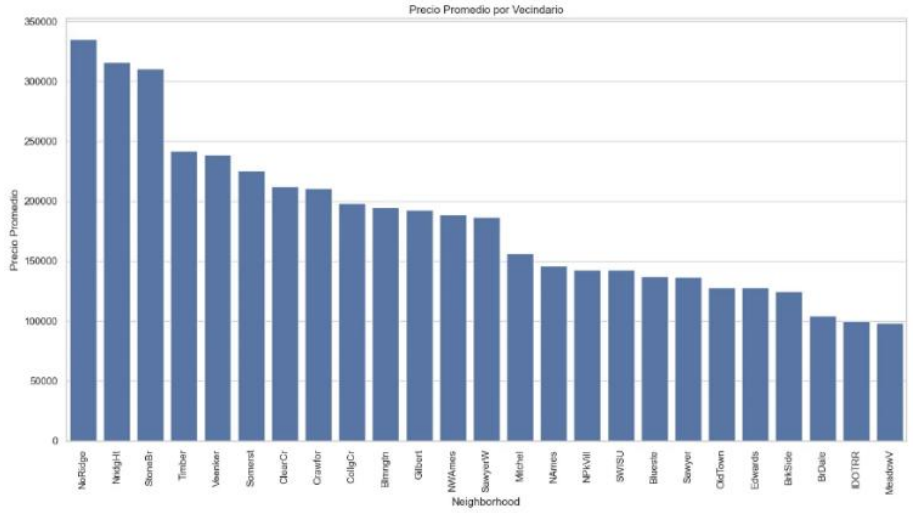
2. En las imágenes 2 y 3 se visualizan estas relaciones mediante gráficos de dispersión, donde se aprecia el patrón escalonado para OverallQual y la relación lineal para GrLivArea.



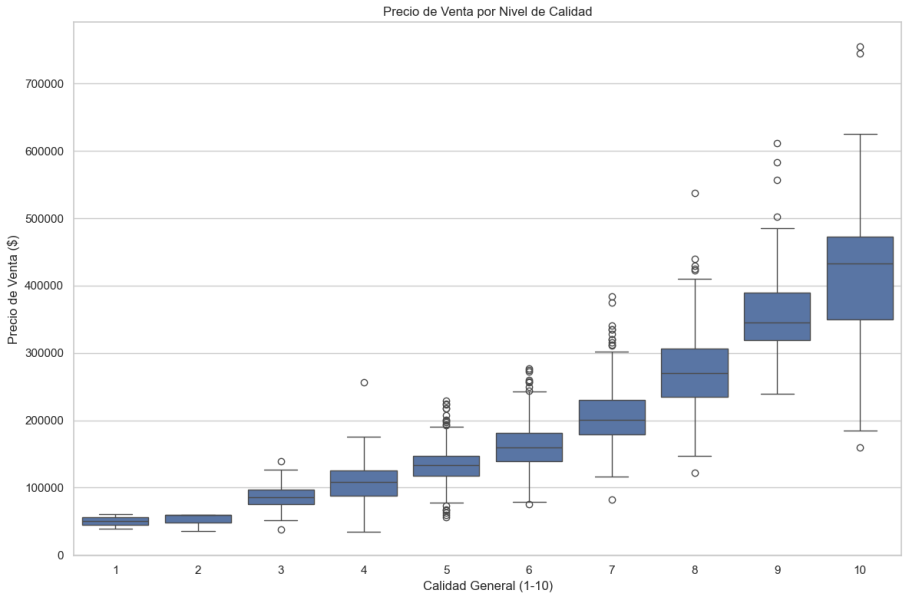


iii. ¿Cómo varía el precio de venta entre los diferentes vecindarios (Neighborhood) y qué vecindarios tienen los precios más altos y más bajos?

1. La imagen (gráficos de barras) muestra las diferencias de precio por vecindario, confirmando visualmente que NoRidge, NridgHt y StoneBr tienen precios significativamente más altos que los demás vecindarios.

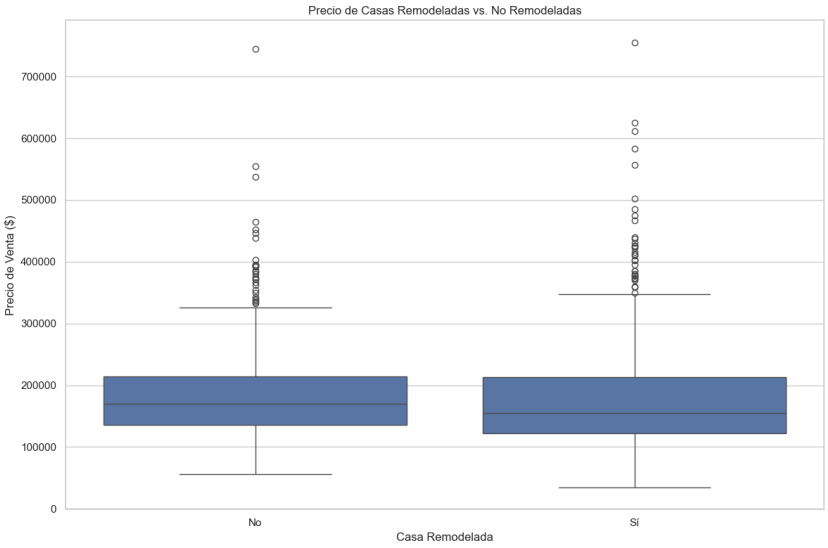


iv. ¿Qué efecto tiene la calidad general (OverallQual) de la vivienda en el precio de venta?



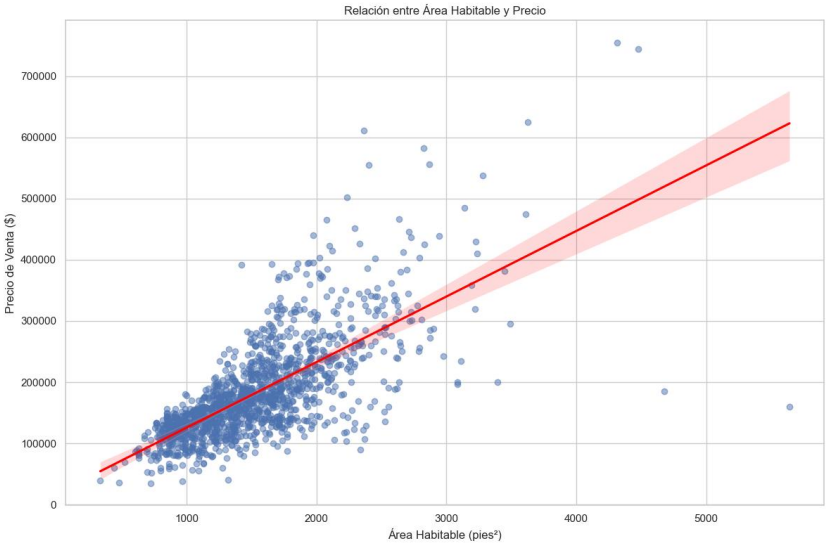
1. El gráfico muestra una clara relación positiva entre la calidad general (OverallQual) de las viviendas y su precio de venta. A medida que aumenta el nivel de calidad en la escala de 1 a 10, se observa un incremento consistente en el precio mediano de las propiedades. Las viviendas con calificaciones más bajas (1-3) tienen precios medianos por debajo de los $100,000, mientras que aquellas con calificaciones altas (9-10) superan los $400,000. La dispersión de precios también aumenta con la calidad, siendo mucho mayor en los niveles superiores, donde algunas propiedades excepcionales alcanzan valores cercanos a los $600,000. Esto confirma que la calidad general de la vivienda es un predictor muy importante del precio de venta, con una relación que parece ser exponencial más que lineal.

v. ¿Las casas remodeladas (donde YearRemodAdd > YearBuilt) tienen un precio significativamente diferente a las no remodeladas?



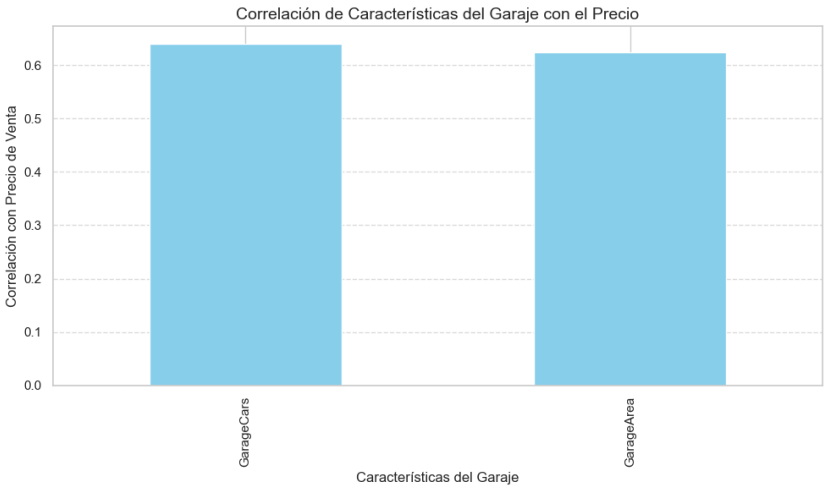
1. El gráfico boxplot que compara los precios de venta entre casas remodeladas (Sí) y no remodeladas (No) muestra diferencias sutiles pero notables. Aunque ambos grupos presentan medianas similares (línea horizontal dentro de cada caja) alrededor de los $160,000-170,000, las casas remodeladas tienden a tener una distribución más amplia en el rango intercuartílico (altura de la caja), sugiriendo mayor variabilidad en sus precios. También se observa que las casas remodeladas tienen más valores atípicos en el extremo superior, con algunas propiedades alcanzando precios superiores a $600,000. Esta distribución indica que, si bien la remodelación por sí sola no garantiza un aumento sustancial en el precio mediano, puede estar asociada con una mayor probabilidad de alcanzar valores de venta más elevados, especialmente en el segmento premium del mercado.

vi. ¿Cómo se relaciona el área habitable sobre el nivel del suelo (GrLivArea) con el precio de venta?



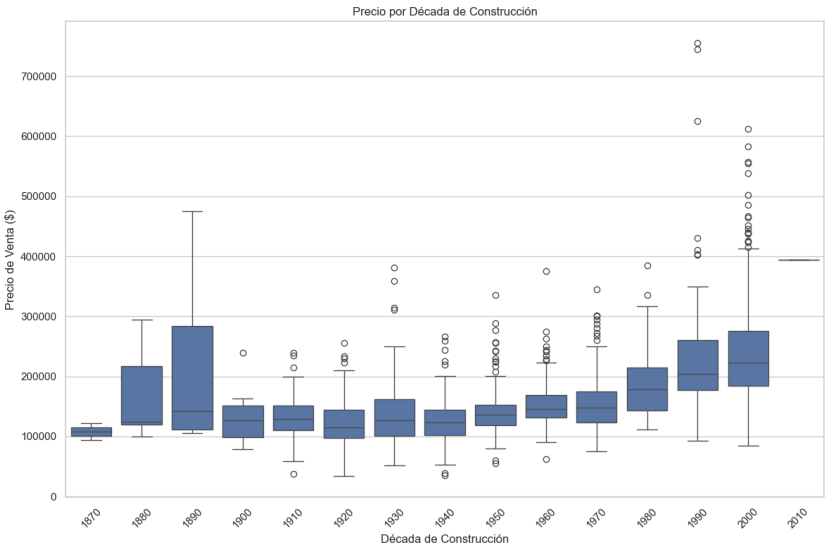
1. El gráfico de dispersión muestra una clara relación positiva entre el área habitable y el precio de venta. La línea de tendencia roja indica que a mayor superficie, mayor precio, con una concentración de propiedades entre 1,000-2,000 pies cuadrados. El intervalo de confianza se ensancha en valores superiores donde hay menos datos. Se observan algunos valores atípicos, especialmente en propiedades grandes que no siguen estrictamente la tendencia. Esta variable es claramente un predictor importante del precio de venta, con una correlación que parece ser moderadamente alta.

vii. ¿Qué características del garaje (GarageArea, GarageCars, GarageQual) tienen mayor impacto en el precio?



1. El gráfico muestra que tanto la capacidad del garaje (GarageCars) como el área del garaje (GarageArea) presentan correlaciones positivas similares con el precio de venta, ambas alrededor de 0.62. Esto indica que las dos características tienen un impacto comparable y significativo en el precio de las viviendas.

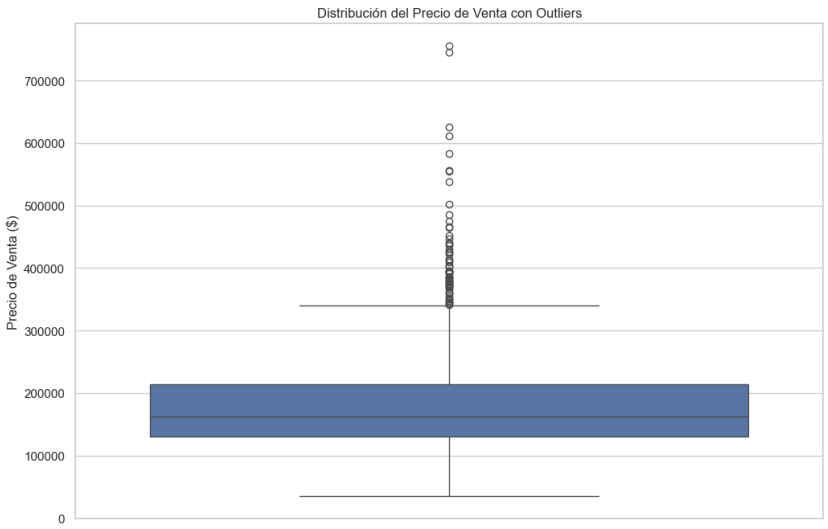
viii. ¿Cómo afecta la antigüedad de la casa (YearBuilt) al precio de venta?



1. El gráfico muestra una tendencia no lineal entre la década de construcción y el precio de venta. Las casas más recientes (1990-2010) tienen precios medianos significativamente más altos, con medianas entre $190,000 y

$280,000, y presentan mayor dispersión hacia valores superiores, con varios outliers por encima de $400,000. Se observa un patrón interesante en casas de 1890, que muestran precios relativamente altos para su antigüedad. A partir de 1950, se aprecia un incremento gradual en los precios medianos a medida que las construcciones son más recientes. Las propiedades más antiguas (1870-1920) generalmente tienen precios más bajos. Esta relación sugiere que la antigüedad influye en el precio, pero probablemente interactúa con otros factores como la calidad, el mantenimiento y posibles renovaciones.

ix. ¿Existen valores atípicos (outliers) en el precio de venta o en otras variables importantes que podrían afectar los modelos predictivos?



1. El boxplot muestra una distribución asimétrica de los precios de venta con presencia clara de valores atípicos. La caja principal (que representa el rango intercuartílico) se sitúa aproximadamente entre $130,000 y $215,000, con una mediana cercana a $160,000. Se observan numerosos valores atípicos por encima del límite superior, con algunos precios extremos llegando hasta los $620,000. Estos outliers representan propiedades de lujo o con

características excepcionales que se alejan significativamente del comportamiento típico del mercado. Para el modelado predictivo, estas observaciones extremas podrían influir desproporcionadamente en los resultados, por lo que sería recomendable considerar una transformación logarítmica de la variable objetivo o técnicas robustas que sean menos sensibles a valores extremos.

Basándonos en los análisis gráficos y estadísticos, las mejores variables predictoras del precio de las casas son: OverallQual (0.79 de correlación), que muestra una relación casi exponencial con el precio; GrLivArea (0.71), cuyo gráfico de dispersión confirma su fuerte influencia; características del garaje (GarageCars/GarageArea), con correlaciones de 0.64/0.62; superficies habitables (TotalBsmtSF, 1stFlrSF); y antigüedad (YearBuilt). Entre las variables categóricas, destacan Neighborhood, ExterQual y KitchenQual.

III. Análisis de grupos

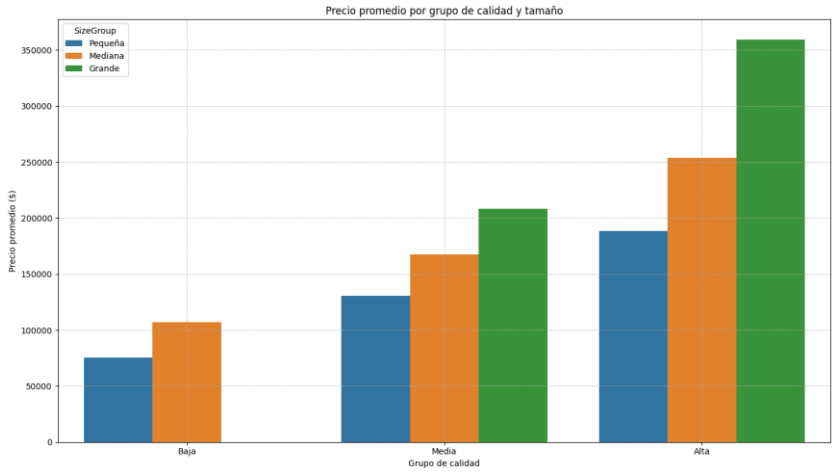
a. Agrupación por calidad y tamaño

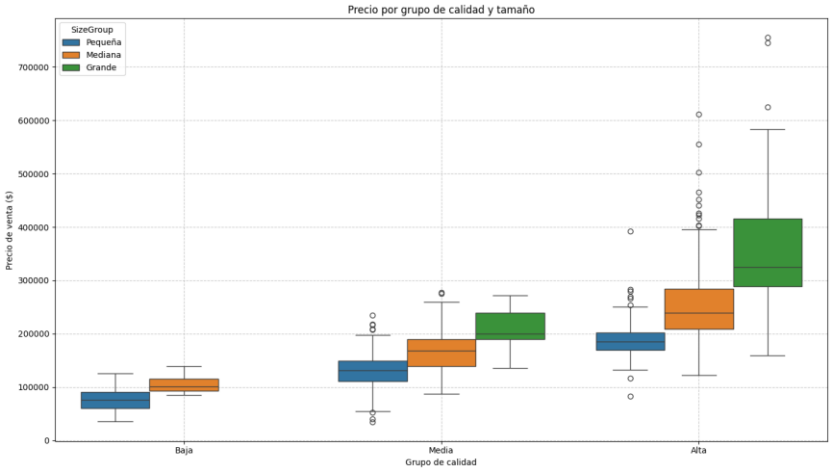
i. Variables clave: OverallQual y GrLivArea

ii. Grupos sugeridos:

1. Calidad (Baja, Media, Alta) combinado con Tamaño (Pequeña, Mediana, Grande)
2. Esto crearía 9 segmentos distintos que capturarían bien la variación de precios

iii. El análisis por calidad y tamaño es fundamental para la valoración inmobiliaria porque captura la interacción entre dos factores determinantes del precio: dimensiones y acabados. Esta segmentación permite identificar patrones específicos que mejoran significativamente la precisión de los modelos predictivos, facilitando la creación de clusters de propiedades comparables y permitiendo valoraciones más refinadas que un análisis univariado.





iv. Las gráficas revelan una clara relación positiva entre el precio y ambas variables. El gráfico de barras muestra un incremento progresivo del precio promedio al aumentar la calidad y el tamaño, con las casas grandes de alta calidad alcanzando valores cinco veces superiores a las pequeñas de baja calidad. El boxplot complementa esta información mostrando mayor dispersión en el segmento premium, lo que sugiere la influencia de factores adicionales como ubicación o características exclusivas en este grupo.

v. En conclusión, la interacción entre calidad y tamaño genera un efecto multiplicador en el precio, no meramente aditivo, lo que debe ser considerado al desarrollar modelos predictivos precisos.

b. Agrupación por época/antigüedad

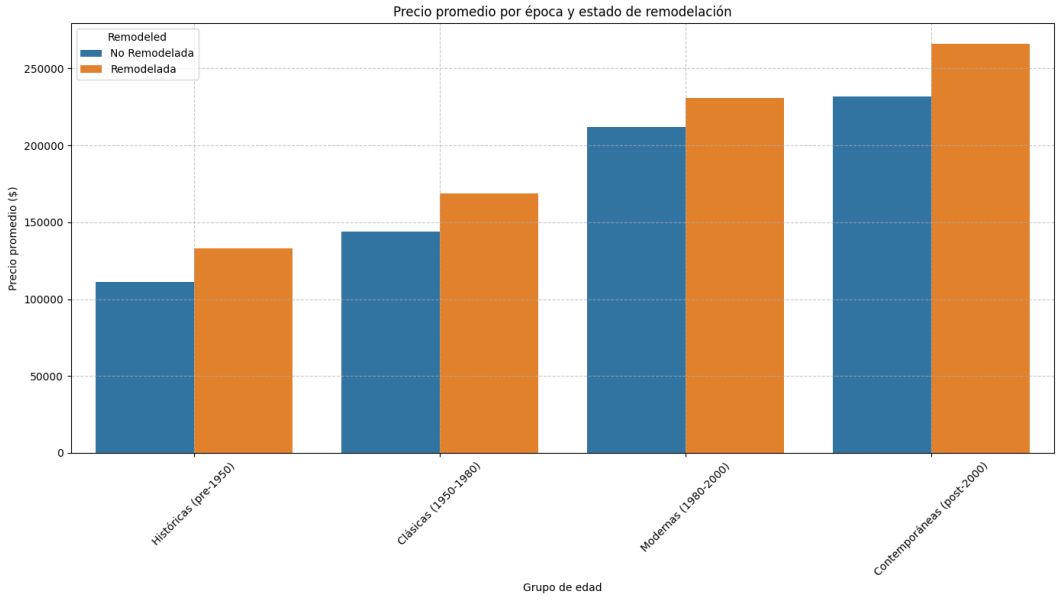
i. Variables clave: YearBuilt y YearRemodAdd

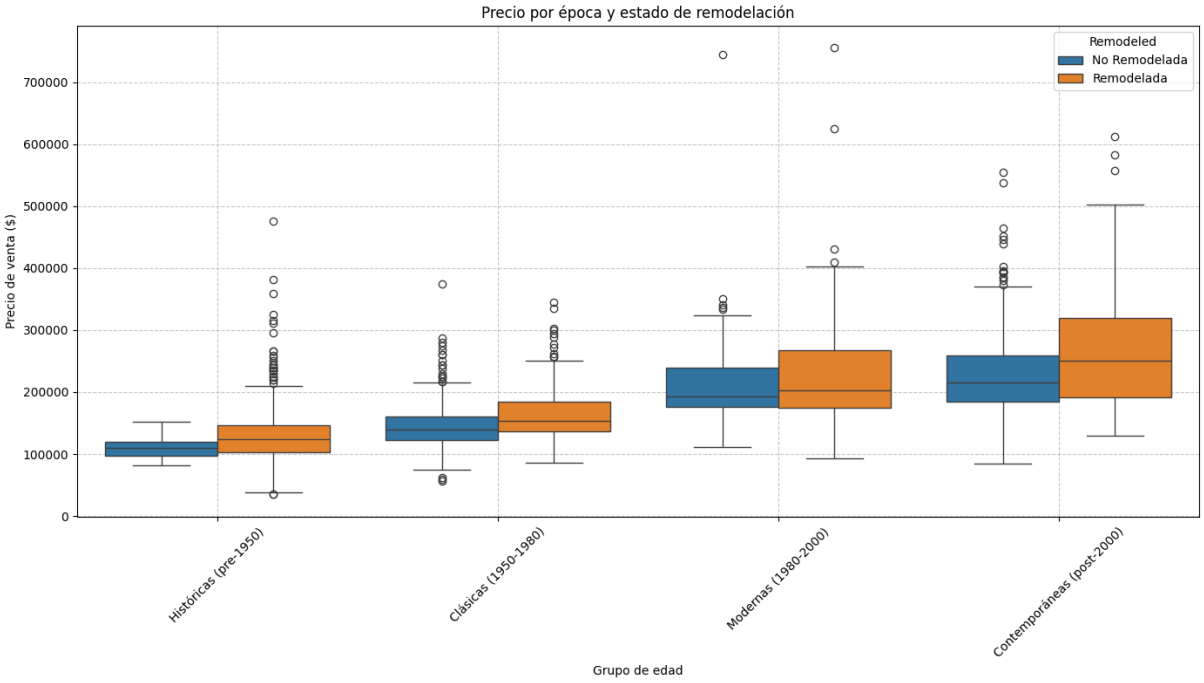
ii. Grupos sugeridos:

1. Casas históricas (pre-1950)
2. Casas clásicas (1950-1980)
3. Casas modernas (1980-2000)
4. Casas contemporáneas (post-2000)

5. Dentro de cada grupo, subgrupos según si han sido remodeladas o no

iii. El análisis por época y estado de remodelación es esencial porque captura la interacción entre la antigüedad de las viviendas y las mejoras realizadas, dos factores cruciales para la valoración inmobiliaria. Esta segmentación permite entender cómo el impacto de las remodelaciones varía según la época de construcción, proporcionando insights sobre la depreciación y apreciación de las propiedades a lo largo del tiempo. Para InmoValor S.A., este análisis resulta particularmente valioso al permitir cuantificar el retorno de inversión de las renovaciones en diferentes segmentos del mercado y evaluar si las casas antiguas remodeladas pueden alcanzar valores comparables a las construcciones más recientes.





iv. Las gráficas muestran una clara tendencia ascendente en los precios tanto por época más reciente como por estado de remodelación. En el gráfico de barras se observa que las viviendas remodeladas consistentemente superan en valor a las no remodeladas dentro de cada grupo de edad, con la mayor diferencia en las propiedades contemporáneas (post-2000), donde la remodelación eleva el precio promedio de $232,000 a $266,000. El boxplot complementa esta información mostrando que las propiedades contemporáneas remodeladas presentan no solo la mediana más alta sino también mayor variabilidad y valores atípicos superiores, alcanzando incluso los $600,000. Notablemente, las viviendas históricas (pre-1950) remodeladas muestran precios comparables a las clásicas (1950-1980) no remodeladas, evidenciando el importante impacto de las actualizaciones en propiedades antiguas.

v. En conclusión, la remodelación incrementa significativamente el valor de las propiedades en todas las épocas, con un efecto más pronunciado en viviendas contemporáneas.

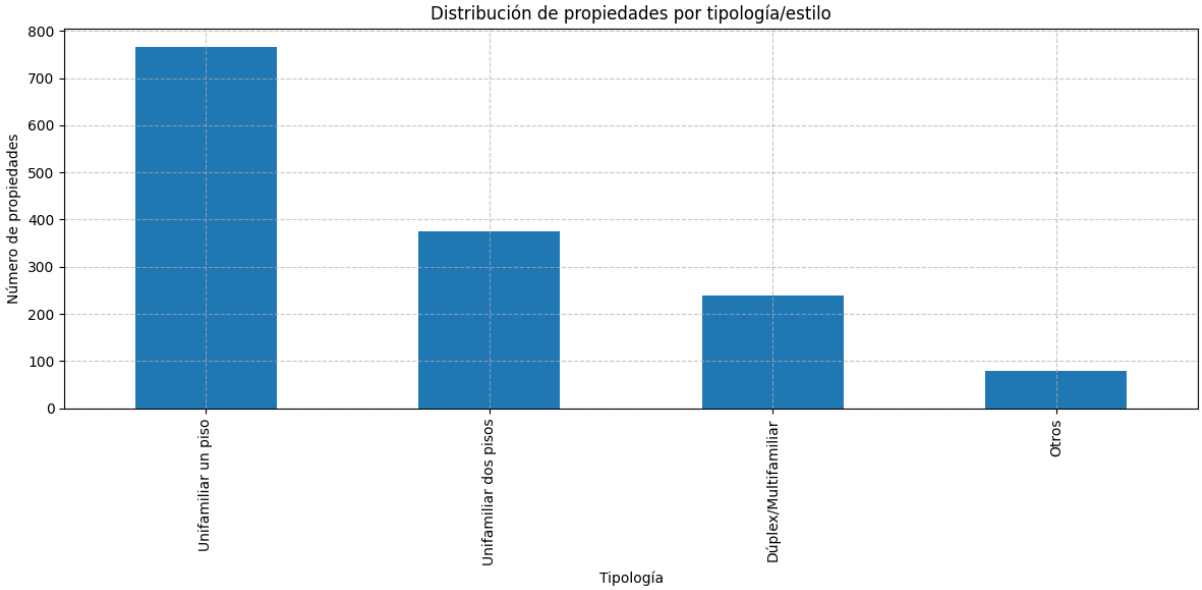
c. Agrupación por tipología/estilo

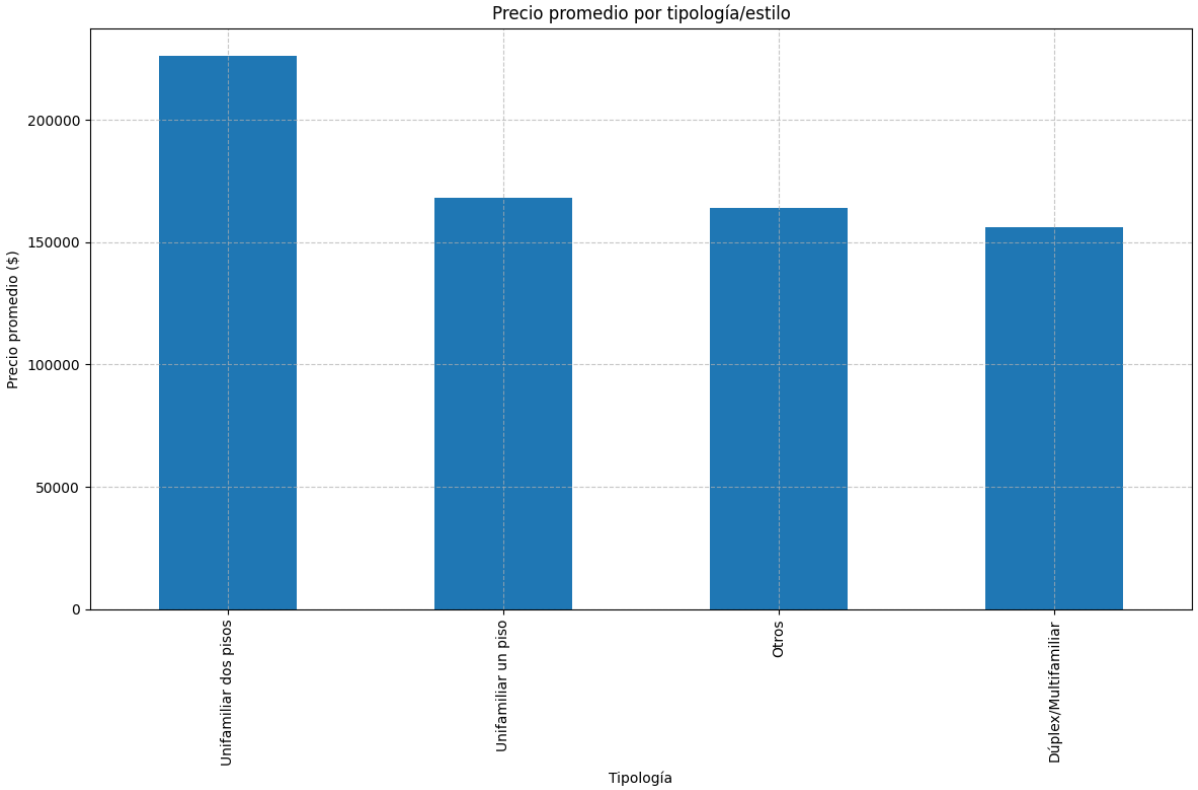
i. Variables clave: BldgType, HouseStyle y MSSubClass

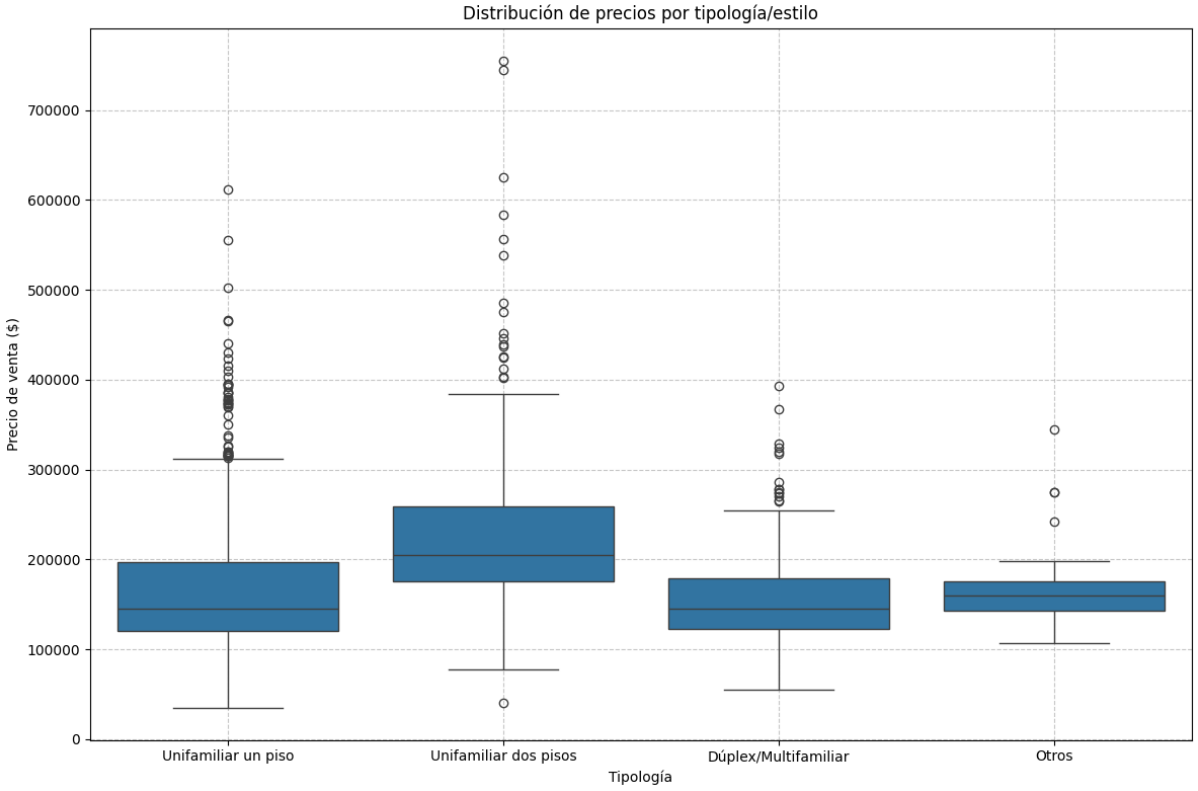
ii. Grupos sugeridos:

1. Viviendas unifamiliares de un piso
2. Viviendas unifamiliares de dos pisos
3. Dúplex y propiedades multifamiliares
4. PUD (Planned Unit Development)

iii. El análisis por tipología o estilo arquitectónico es fundamental en la valoración inmobiliaria porque representa una característica estructural que influye significativamente en las preferencias de los compradores y en el valor de mercado. Esta segmentación permite a InmoValor S.A. identificar patrones de precios específicos para diferentes configuraciones de vivienda (unifamiliar de un piso, dos pisos, dúplex o multifamiliar), lo que mejora la precisión de los modelos predictivos. Además, el conocimiento de la distribución del mercado por tipologías ayuda a contextualizar los resultados y a comprender qué segmentos tienen mayor representación, facilitando estrategias de valoración más focalizadas y precisas para diferentes tipos de inmuebles.


Distribución de propiedades por tipología/estilo
Número de propiedades
0
100
200
300
400
500
600
700
800
Unifamiliar un piso
Unifamiliar dos pisos
Dúplex/Multifamiliar
Otros
Tipología


Precio promedio por tipología/estilo
Precio promedio ($)
0
50000
100000
150000
200000
Unifamiliar dos pisos
Unifamiliar un piso
Otros
Dúplex/Multifamiliar
Tipología




```
Análisis por tipología/estilo:
                              mean  count          std    median
Typology
Unifamiliar un piso    168189.019557    767  76390.518126  144900.0
Unifamiliar dos pisos  226213.192000    375  87631.539021  205000.0
Dúplex/Multifamiliar   156304.587500    240  54394.802994  145000.0
Otros                  164114.358974     78  35609.436827  159975.0
```

iv. Las gráficas revelan hallazgos importantes sobre la distribución y precios por tipología. La primera visualización muestra que las viviendas unifamiliares de un piso dominan el mercado (aproximadamente 767 propiedades), seguidas por las unifamiliares de dos pisos (375), dúplex/multifamiliares (240) y otras tipologías (78). Sin embargo, el segundo gráfico demuestra que las unifamiliares de dos pisos alcanzan el precio promedio más alto (aproximadamente $226,000), significativamente superior al resto de categorías que se mantienen en rangos similares entre $156,000 y $168,000. El boxplot complementa esta información mostrando que las viviendas unifamiliares de

dos pisos no solo tienen la mediana más alta ($205,000), sino también mayor dispersión y valores atípicos que llegan hasta los $750,000, mientras que las otras tipologías presentan rangos más compactos de precios.

v. En conclusión, las unifamiliares de dos pisos representan el segmento premium del mercado, con valores sustancialmente más altos que las demás tipologías a pesar de no ser las más numerosas.

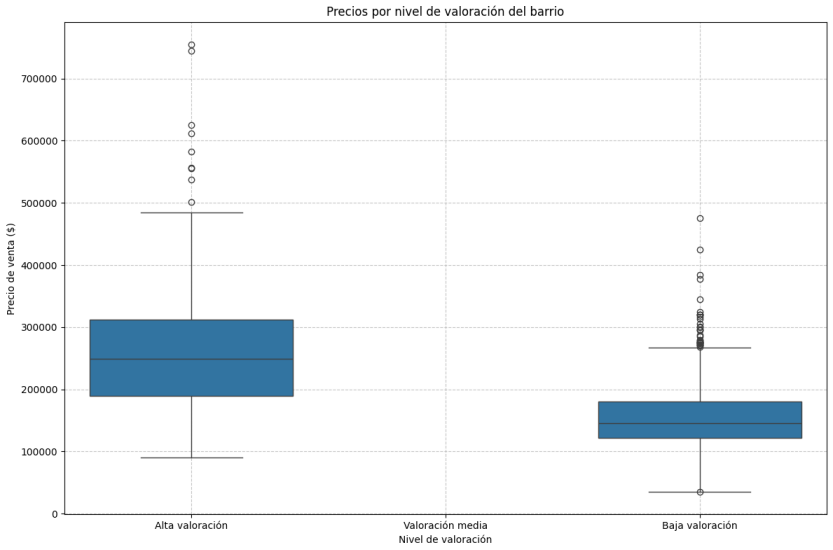
d. Agrupación por ubicación

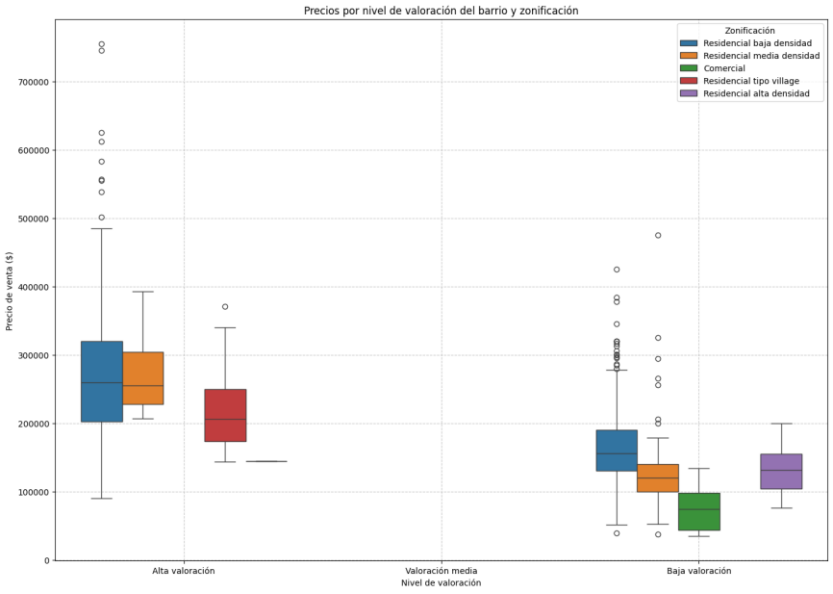
i. Variables clave: Neighborhood y MSZoning

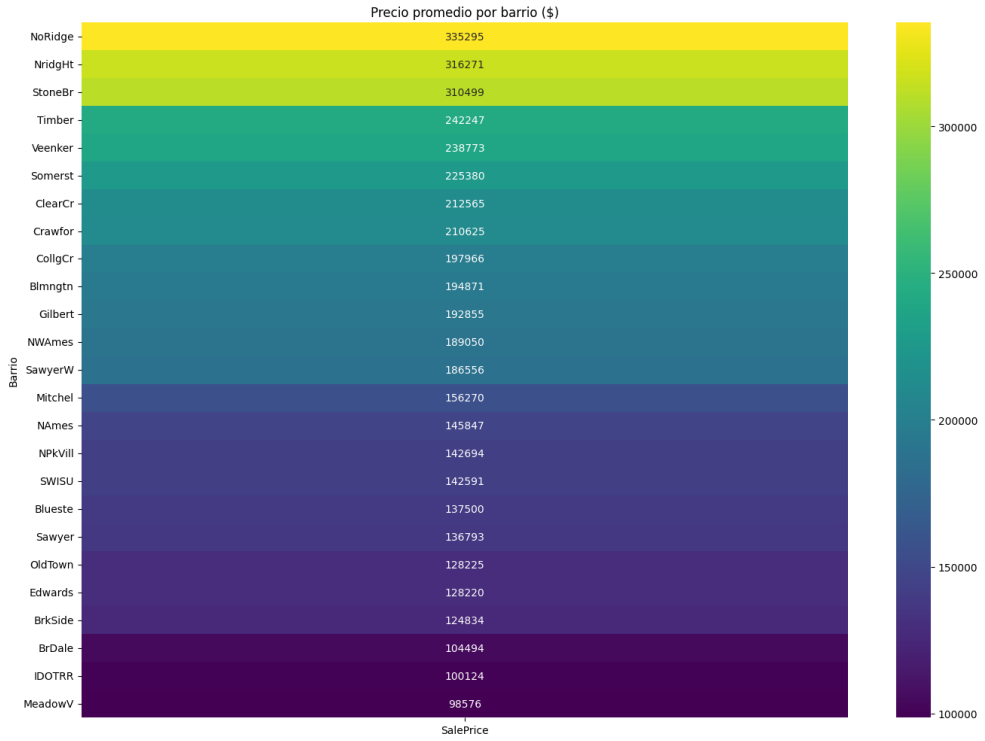
ii. Grupos sugeridos:

1. Barrios de alta valoración (como NridgHt, StoneBr, NoRidge)
2. Barrios de valoración media
3. Barrios de valoración baja
4. Complementar con la zonificación (residencial, comercial, etc.)

iii. El análisis por ubicación es crucial en la valoración inmobiliaria porque la localización es uno de los factores más determinantes del valor de una propiedad. La segmentación por niveles de valoración del barrio y zonificación permite identificar patrones de precios específicos por zona geográfica y uso del suelo, capturando la variación espacial inherente al mercado inmobiliario. Para InmoValor S.A., esta agrupación es particularmente valiosa porque proporciona una comprensión estructurada de cómo la micro-ubicación (barrio específico) y la macro-ubicación (tipo de zonificación) interactúan para determinar el valor de mercado.


Precios por nivel de valoración del barrio
Precio de venta ($)
0
100000
200000
300000
400000
500000
600000
700000
Alta valoración
Valoración media
Baja valoración
Nivel de valoración


Precios por nivel de valoración del barrio y zonificación
Zonificación
Residencial baja densidad
Residencial media densidad
Comercial
Residencial tipo village
Residencial alta densidad
Precio de venta ($)
0
100000
200000
300000
400000
500000
600000
700000
Alta valoración
Valoración media
Baja valoración
Nivel de valoración




```
Análisis por ubicación (valor de barrio + zonificación):
                                                mean  count  \
NeighborhoodValue ZoningSimple
Alta valoración   Residencial media densidad  277375.000000      4
                  Residencial baja densidad   274336.604895    286
                  Residencial tipo village    214014.061538     65
Baja valoración   Residencial baja densidad   163452.578035    865
Alta valoración   Residencial alta densidad   145000.000000      2
Baja valoración   Residencial alta densidad   129638.142857     14
                  Residencial media densidad  123493.313084    214
                  Comercial                    74528.000000     10

                                                std
NeighborhoodValue ZoningSimple
Alta valoración   Residencial media densidad  81665.348629
                  Residencial baja densidad   99321.852022
                  Residencial tipo village    52369.662067
Baja valoración   Residencial baja densidad   48680.793898
Alta valoración   Residencial alta densidad       0.000000
Baja valoración   Residencial alta densidad   37946.822293
                  Residencial media densidad  43221.607683
                  Comercial                   33791.092031
```

iv. Las gráficas revelan contrastes significativos en los precios según la ubicación. La primera visualización muestra una clara división entre los barrios de alta valoración, con medianas cercanas a los $250,000, y los de baja valoración, con medianas alrededor de $150,000. La segunda gráfica refina este análisis al incorporar la zonificación, evidenciando que dentro de los barrios de alta valoración, las propiedades en zonas residenciales de baja densidad alcanzan

los valores más elevados (mediana aproximada de $260,000), mientras que en los barrios de baja valoración, las propiedades comerciales presentan los precios más bajos (mediana aproximada de $70,000). El mapa de calor complementa esta información destacando las diferencias entre barrios específicos: NoRidge, NridgHt y StoneBr superan los $300,000 de precio promedio, mientras que MeadowV, IDOTRR y BrDale apenas alcanzan los $100,000, ilustrando una variación de más del triple entre las ubicaciones más y menos cotizadas.

v. En conclusión, la ubicación genera diferencias de precio más pronunciadas que cualquier otra característica, con variaciones de hasta un 300% entre barrios de alta y baja valoración.

e. Agrupación por características adicionales

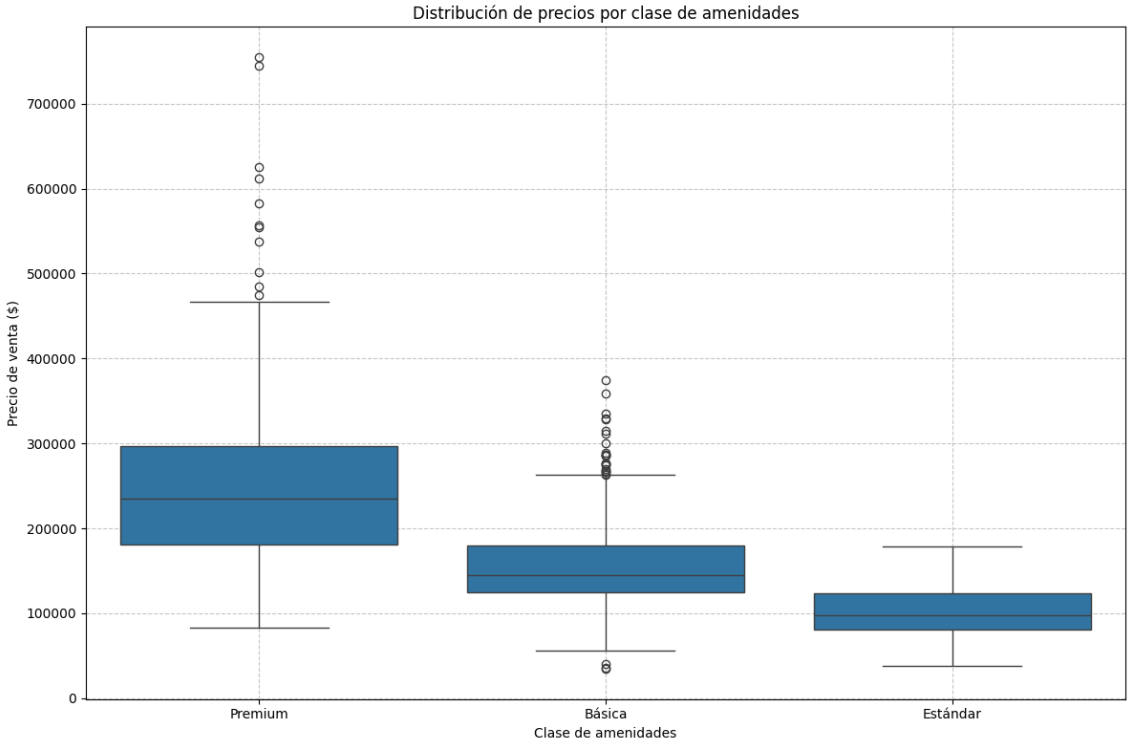
i. Variables clave: Combinación de GarageType, PoolArea, Fireplaces, etc.

ii. Grupos sugeridos:

1. Propiedades estándar (sin amenidades especiales)
2. Propiedades con amenidades básicas (garaje, porche)
3. Propiedades premium (con múltiples amenidades como piscina, chimenea, etc.)

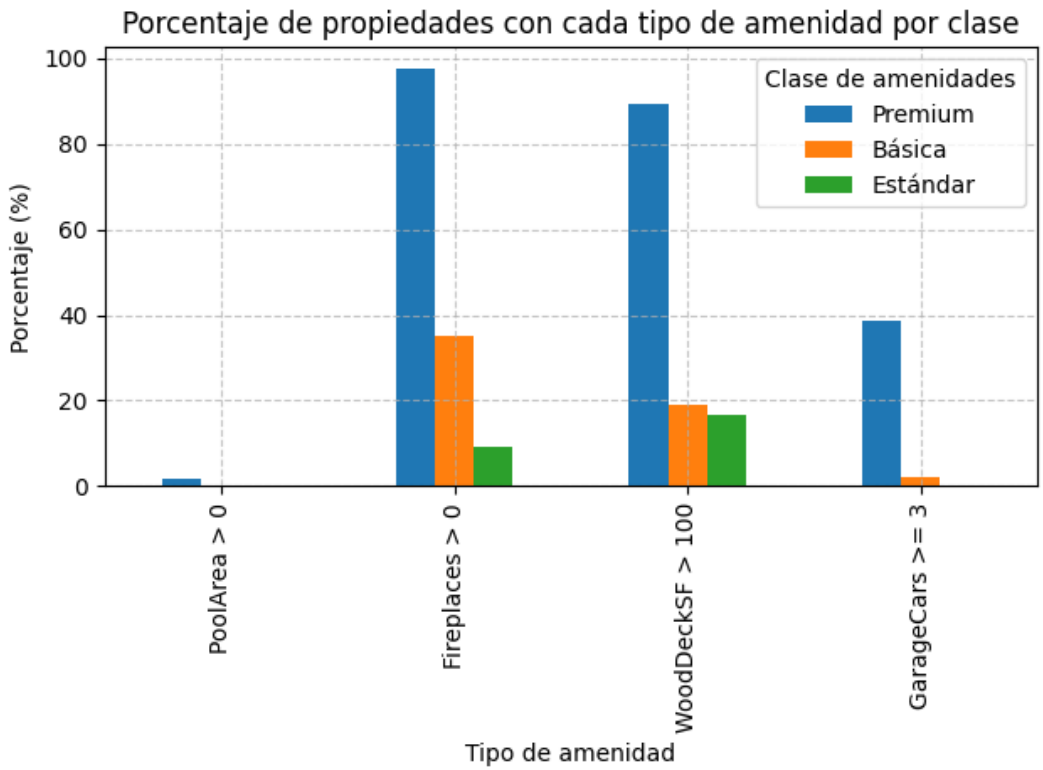
iii. El análisis por clase de amenidades es fundamental en la valoración inmobiliaria porque permite cuantificar el impacto de características adicionales que aportan confort y exclusividad a las propiedades. Esta segmentación ayuda a InmoValor S.A. a comprender cómo diferentes conjuntos de comodidades (piscinas, chimeneas, terrazas y garajes amplios) influyen en el valor de mercado y en las preferencias de los compradores. Al categorizar las viviendas según la presencia y combinación de estas características especiales, se pueden

identificar patrones de precios más específicos y desarrollar modelos predictivos que capturen adecuadamente el valor añadido de cada tipo de amenidad, lo que resulta crucial para valoraciones precisas de propiedades con diferentes niveles de equipamiento.



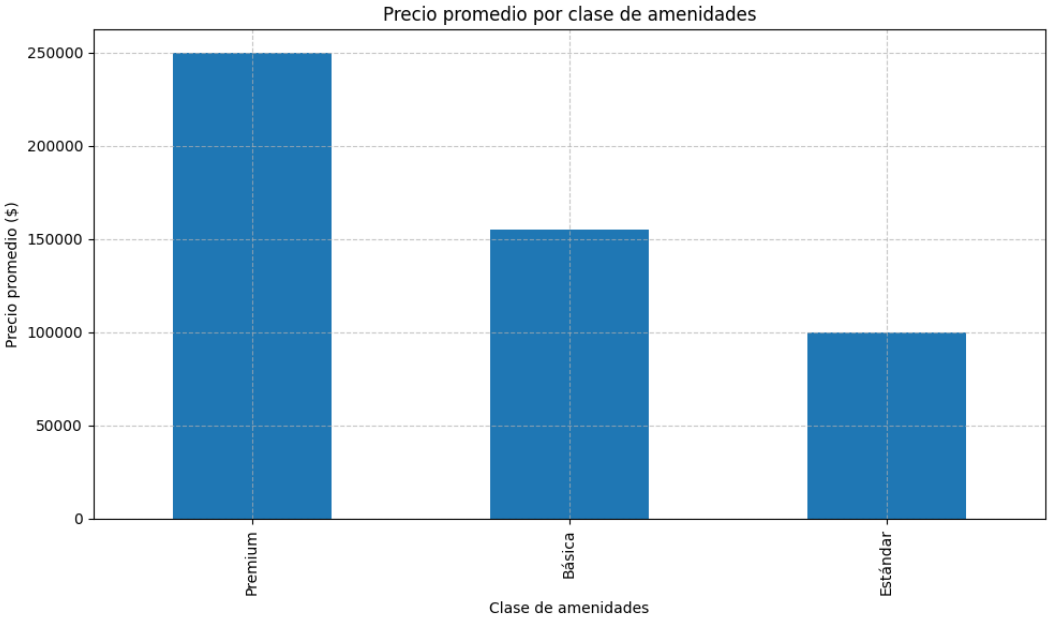
Análisis por clase de amenidades:

AmenityClass	mean	count	std	median
Premium	249779.796767	433	95665.697490	235000.0
Básica	154767.876670	973	46811.356562	145250.0
Estándar	100021.296296	54	31417.234187	97500.0



Porcentaje de propiedades con cada tipo de amenidad por clase:

	PoolArea > 0	Fireplaces > 0	WoodDeckSF > 100	GarageCars >= 3
Básica	0.0	35.149024	19.013361	1.952724
Premium	1.616628	97.690531	89.376443	38.568129
Estándar	0.0	9.259259	16.666667	0.0



iv. Las gráficas muestran una clara estratificación de precios según la clase de amenidades. El boxplot revela que las propiedades premium (con múltiples amenidades) tienen una mediana de precio aproximadamente de $235,000, más del doble que las propiedades estándar ($97,500), con las básicas en un punto intermedio ($145,250). El segundo gráfico ilustra la distribución de amenidades específicas: prácticamente todas las propiedades premium (97.7%) cuentan con chimeneas y la gran mayoría (89.4%) con terrazas de madera sustanciales, mientras que las piscinas son exclusivas de este segmento premium, aunque solo están presentes en el 1.6% de estas propiedades. El gráfico de barras confirma esta diferenciación, mostrando un precio promedio de $250,000 para propiedades premium, $155,000 para las básicas y $100,000 para las estándar, evidenciando una progresión escalonada que refleja el valor incremental que aportan las diferentes amenidades.

v. En conclusión, la presencia de amenidades premium como chimeneas, terrazas amplias y garajes para tres o más vehículos incrementa el valor de las propiedades hasta en un 150% respecto a las viviendas estándar.

f. Agrupación por tamaño del terreno (LotArea)

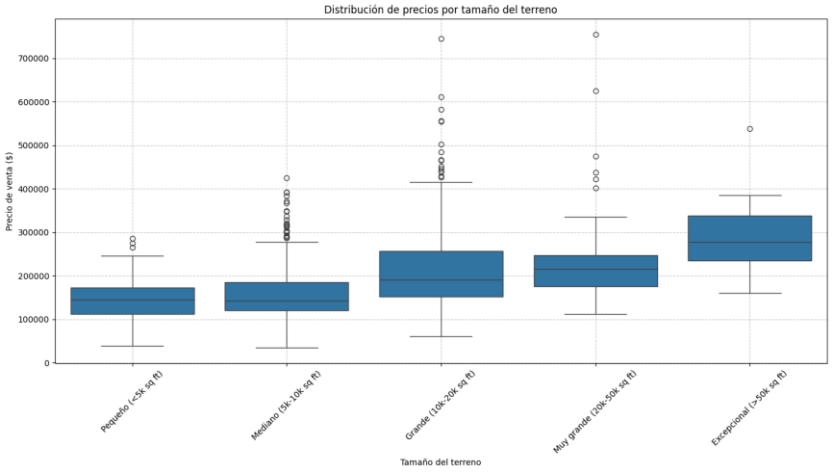
i. Variables clave: LotArea y SalePrice

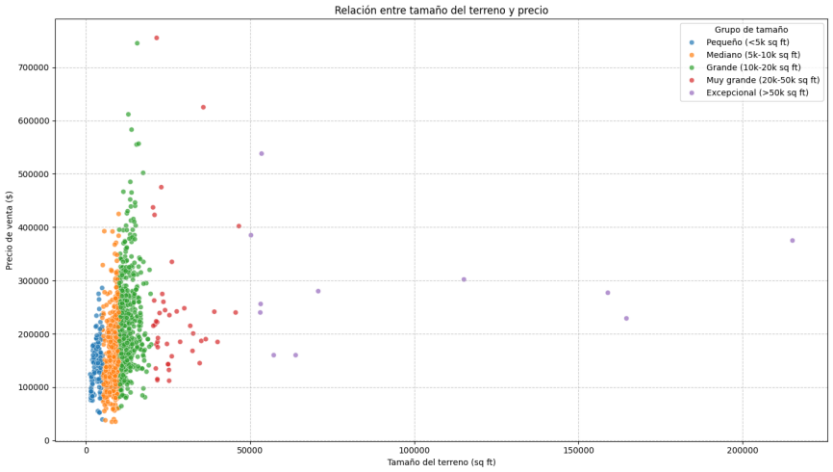
ii. Grupos sugeridos:

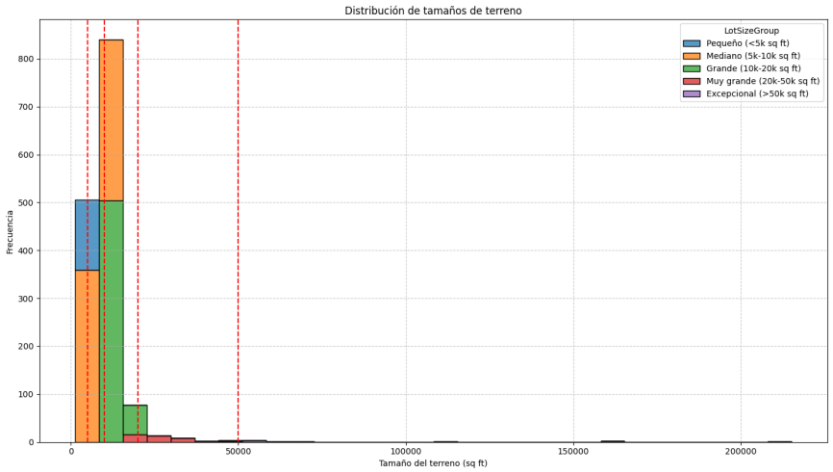
1. Terrenos pequeños (< 5,000 sq ft) - Precio generalmente económico a moderado
2. Terrenos medianos (5,000-10,000 sq ft) - Precio típicamente en el rango medio del mercado
3. Terrenos grandes (10,000-20,000 sq ft) - Precio moderadamente elevado

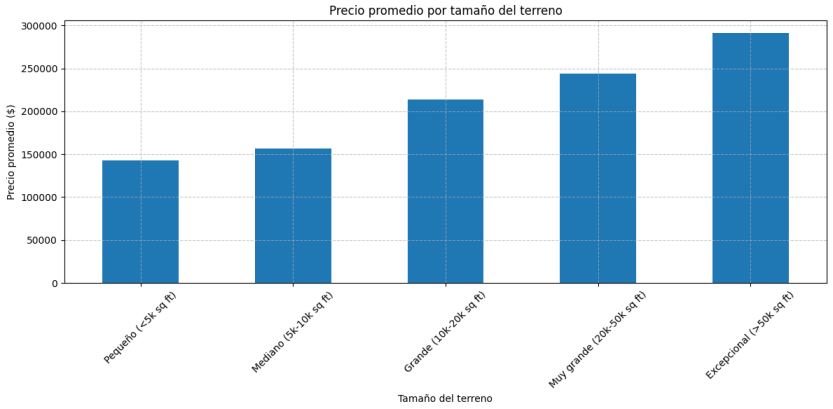
4. Terrenos muy grandes (> 20,000 sq ft) - Precios premium, con alta variabilidad
5. Terrenos excepcionalmente grandes (outliers > 50,000 sq ft) - Propiedades de muy alto valor o con características atípicas

iii. El análisis por tamaño del terreno es fundamental en la valoración inmobiliaria porque representa uno de los componentes básicos del valor de una propiedad: el suelo sobre el que se asienta. Esta segmentación permite a InmoValor S.A. cuantificar cómo la superficie del terreno influye en el precio final, identificando umbrales relevantes donde el valor marginal puede cambiar. Para los modelos predictivos, categorizar las propiedades según el tamaño del lote (pequeño, mediano, grande, muy grande o excepcional) proporciona una estructura que facilita la captura de patrones no lineales en la relación precio-tamaño, mejorando la precisión de las estimaciones, especialmente en los extremos de la distribución donde las propiedades pueden comportarse de manera atípica.









iv. Las gráficas revelan una clara correlación positiva entre el tamaño del terreno y el precio, aunque con matices importantes. El boxplot muestra un incremento progresivo de la mediana de precios a medida que aumenta el tamaño, desde aproximadamente $145,000 para terrenos pequeños (<5k sq ft) hasta $280,000

para los excepcionales (>50k sq ft). El diagrama de dispersión complementa esta información mostrando que la mayor concentración de propiedades se encuentra en terrenos menores a 20,000 sq ft, mientras que los terrenos excepcionales son escasos pero alcanzan valores atípicos elevados. Los histogramas confirman esta distribución asimétrica: la gran mayoría de las propiedades (aproximadamente 85%) tienen terrenos medianos o pequeños, con muy pocas superando los 20,000 sq ft. El gráfico de barras sintetiza la relación mostrando un incremento constante del precio promedio con cada categoría de tamaño, desde $145,000 para los pequeños hasta casi $300,000 para los excepcionales.

v. En conclusión, aunque el tamaño del terreno impacta positivamente en el precio, este efecto no es lineal, con un incremento más pronunciado al pasar de grande a excepcional que de pequeño a mediano.

IV. Data set procesado

a. Conjuntos de datos

Para nuestro análisis, hemos trabajado con la división de datos establecida por Kaggle en el conjunto "House Prices: Advanced Regression Techniques". En lugar de utilizar el archivo test.csv, que no contiene la variable objetivo 'SalePrice', optamos por dividir el archivo train.csv en un 70% para entrenamiento y 30% para validación (esto representa 1021 filas para entrenamiento y 438 para validación).

La división 70%-30% es una práctica común en machine learning, ya que permite entrenar el modelo con una cantidad suficiente de datos mientras se reserva un conjunto representativo para evaluar su capacidad de generalización. Un conjunto de validación más pequeño podría generar estimaciones inestables, mientras que uno demasiado grande reduciría la cantidad de datos disponibles para el entrenamiento, afectando el aprendizaje del modelo.

V. Ingeniería de características

a. ¿Qué variables cree que puedan ser mejores predictores para el precio de las casas?

Basándonos en los análisis exploratorios y de agrupamiento realizados anteriormente, donde examinamos distribuciones estadísticas, correlaciones y comportamientos visuales a través de boxplots, gráficos de barras y mapas de calor, hemos identificado un conjunto de variables con alta capacidad predictiva para el precio de las casas. Estas variables mostraron patrones consistentes y diferencias estadísticamente significativas entre grupos, como se evidenció en las visualizaciones de clusters por calidad-tamaño, época-antigüedad, tipología, ubicación y amenidades. A continuación, presentamos las variables que demostraron mayor influencia sobre SalePrice, ordenadas según su poder predictivo:

i. OverallQual - Calidad general de los materiales y acabados (escala 1-10). El análisis muestra una correlación muy fuerte entre la calidad general y el precio, siendo posiblemente el predictor individual más potente.

ii. Neighborhood - Ubicación dentro de los límites de la ciudad. El análisis por ubicación demuestra que el barrio es un factor determinante, con diferencias dramáticas entre zonas de alta y baja valoración.

iii. GrLivArea - Área habitable sobre rasante (pies cuadrados). El tamaño habitable es consistentemente uno de los principales predictores de precio, mostrando una fuerte relación positiva.

iv. LotArea - Tamaño del lote en pies cuadrados. Influye en el precio, con una relación a mayor tamaño mayor precio.

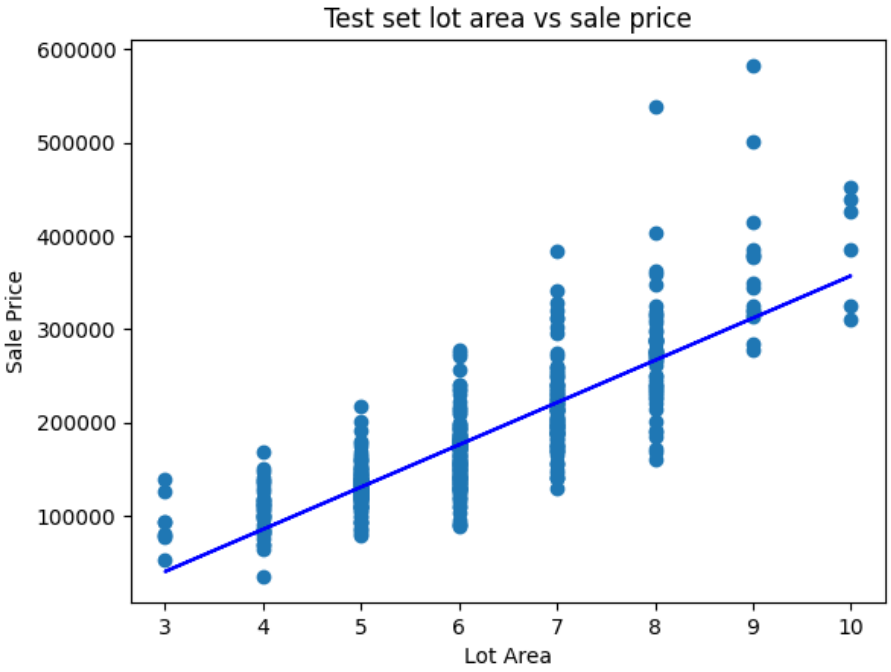
v. YearBuilt - Año de construcción original. Las propiedades más nuevas, especialmente las "Contemporáneas", muestran valores significativamente más altos.
vi. KitchenQual - Calidad de la cocina. La calidad de la cocina tiene un impacto desproporcionado en el valor total de la propiedad.
vii. ExterQual - Calidad de los materiales exteriores. El acabado exterior y su calidad representan tanto el atractivo visual como la durabilidad.
viii. TotalBsmtSF - Superficie total del sótano. Espacio adicional valioso, particularmente cuando está bien acabado.
ix. BsmtQual - Evaluación de la altura del sótano. La calidad del sótano, especialmente su altura, afecta significativamente su usabilidad.
x. GarageCars - Tamaño del garaje en capacidad de coches. El análisis muestra que garajes grandes (3+ coches) son un indicador de propiedades premium.
xi. YearRemodAdd - Año de remodelación. Las propiedades remodeladas tienen valores significativamente más altos que las no remodeladas dentro del mismo grupo de edad.

b. En conclusión, el análisis exhaustivo de las propiedades inmobiliarias reveló que los predictores más potentes del precio son la calidad general de la construcción (OverallQual), la ubicación (Neighborhood), el tamaño habitable (GrLivArea), y características de calidad específicas como cocinas, exteriores y sótanos, demostrando que tanto los aspectos estructurales como los de ubicación juegan un papel crucial en la determinación del valor de mercado.

VI. Modelo Univariado de regresión lineal para precio de casas

Para este modelo se eligió trabajar con la variable de 'OverallQual' que nos da la calidad general de materiales y acabados, en una escala del 1 al 10. Esa decisión se tomó porque según el análisis realizado la variable parece tener una gran influencia en el precio final de venta. Los resultados son los siguientes:

price_pred = 45304.8124*area -95782.2317

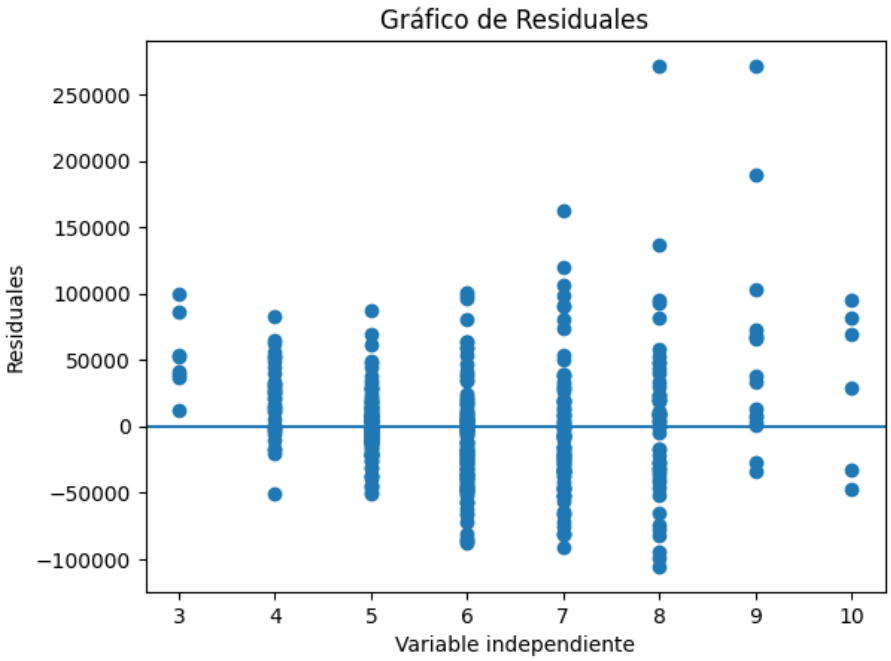


R^2: 0.64

MSE: 2350108583.96

RMSE: 48477.92

Por el resultado de los estadísticos es notorio que los resultados del modelo son completamente diferentes a los resultados esperados. En realidad, eso tiene sentido pues sabemos que el valor de una propiedad no solo depende de la calidad de sus materiales, sino que depende de más variables. Aun así, observemos el resultado del análisis de residuales:



Se observa que los residuos están distribuidos alrededor de cero. Sin embargo, hay cierta dispersión en los valores, especialmente para ciertos valores de la variable independiente. En este caso, parece haber variabilidad en la dispersión de los residuos a lo largo del eje x, lo que sugiere una posible heterocedasticidad. También se puede observar una menor cantidad de datos en los valores límite (0 a 3 y 9 a 10) lo que puede influenciar en una menor precisión en futuras predicciones con valores de este tipo.

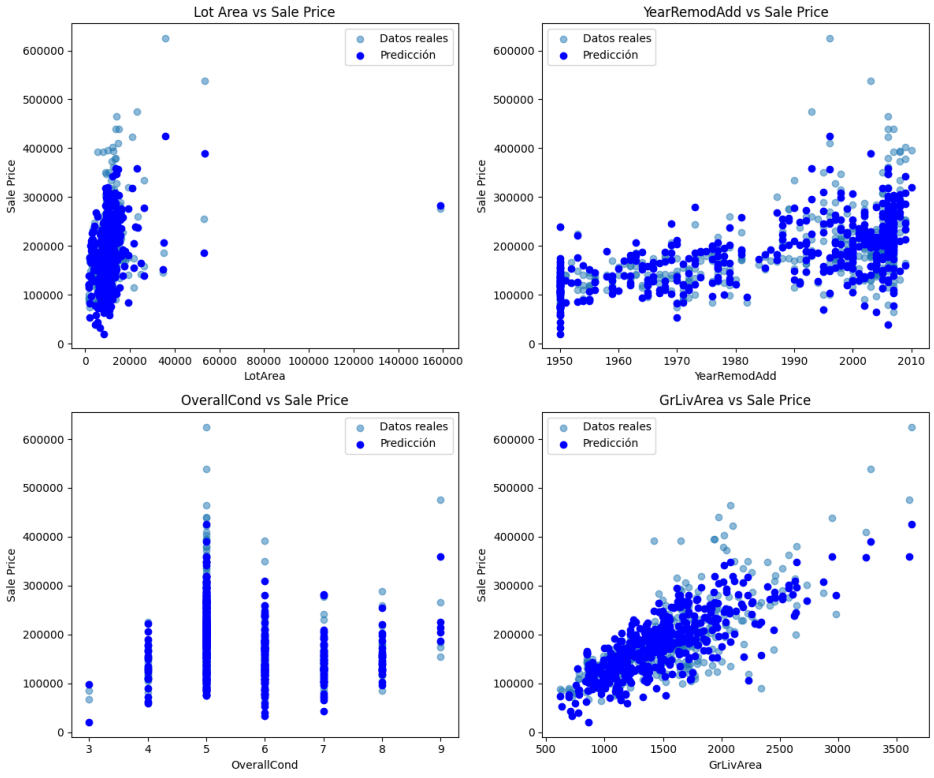
VII. Modelo de regresión lineal todas las variables numéricas

El modelo utiliza 26 variables: ['LotArea', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd', 'BsmtFinSF1', 'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'LowQualFinSF', 'GrLivArea', 'BsmtFullBath', 'BsmtHalfBath', 'GarageCars', 'GarageArea', 'WoodDeckSF', 'OpenPorchSF', 'EnclosedPorch', '3SsnPorch', 'ScreenPorch', 'PoolArea', 'MiscVal', 'MoSold', 'YrSold'] las cuales son todas numéricas. Se quitaron del listado aquellas que tienen datos ‘NaN’ pues al revisarlas notamos que no eran completamente relevantes según el análisis exploratorio. Para mejores resultados también se quitaron algunas que, según el análisis, no tenían impacto claro en el precio de venta.

Ecuación:

SalePrice = 191473.2359 + (0.4104 * LotArea) + (19794.6067 * OverallQual) + (4406.1131 * OverallCond) + (349.7480 * YearBuilt) + (187.1189 * YearRemodAdd) + (7.1507 * BsmtFinSF1) + (-2.3034 * BsmtFinSF2) + (2.5008 * BsmtUnfSF) + (7.3481 * TotalBsmtSF) + (32.5323 * 1stFlrSF) + (21.0254 * 2ndFlrSF) + (-27.4218 * LowQualFinSF) + (26.1360 * GrLivArea) + (9229.3962 * BsmtFullBath) + (3086.4745 * BsmtHalfBath) + (13344.8349 * GarageCars) + (2.4129 * GarageArea) + (22.6635 * WoodDeckSF) + (-5.7737 * OpenPorchSF) + (15.6004 * EnclosedPorch) + (-16.7895 * 3SsnPorch) + (55.6262 * ScreenPorch) + (-50.1348 * PoolArea) + (-0.6713 * MiscVal) + (-682.1948 * MoSold) + (-670.0016 * YrSold)

Al ser tantas variables elegimos algunas variables para hacer varios gráficos de dos dimensiones mostrando los resultados

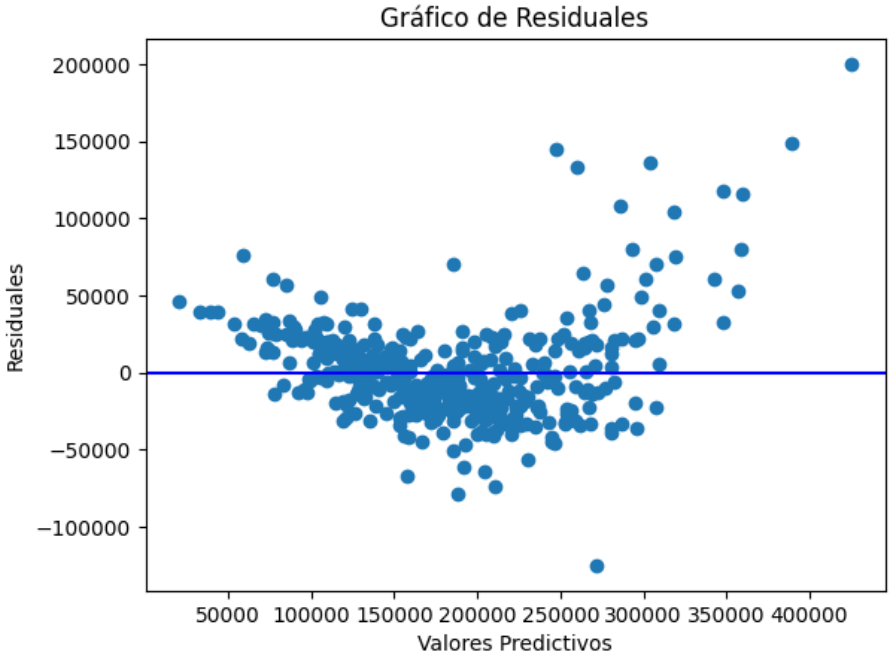


R^2: 0.7706

MSE: 1580045490.1482

RMSE: 39749.786

En cada una es notorio que el modelo es bastante certero en calcular el precio de venta de la propiedad al utilizar todas las variables numéricas que influyen en el resultado. Tanto el valor R^2 como el MSE tienen resultados más adecuados lo que nos indica que es un mejor resultado que en el modelo de una sola variable. Veamos los resultados de análisis de residuales:



Se observa una mejor distribución alrededor de cero, aunque los valores aun sugieren heterocedasticidad, es decir, la varianza de los errores no es constante. Por lo que podríamos decir que el modelo puede estar funcionando bien para valores bajos/intermedios, pero tiene errores más grandes para valores altos.

VIII. ¿Multicolinealidad?

a. Diagnóstico de Multicolinealidad

i. Variables de Superficie Relacionadas con el Sótano

- TotalBsmtSF es la suma de BsmtFinSF1, BsmtFinSF2 y BsmtUnfSF
- Evidencia en los coeficientes:
    - BsmtFinSF1: 7.1507
    - BsmtFinSF2: -2.3034
    - BsmtUnfSF: 2.5008
    - TotalBsmtSF: 7.3481
- Los signos contradictorios y la presencia del total y sus componentes simultáneamente confirman un alta multicolinealidad

ii. Variables de Superficie Habitable

- GrLivArea está directamente relacionada con 1stFlrSF, 2ndFlrSF y LowQualFinSF
- Coeficientes:
    - 1stFlrSF: 32.5323
    - 2ndFlrSF: 21.0254
    - LowQualFinSF: -27.4218
    - GrLivArea: 26.1360
- La presencia de variables que son componentes y sus totales genera redundancia

iii. Variables de Garaje

- GarageCars y GarageArea miden aspectos muy correlacionados

- Coeficientes:
    - GarageCars: 13344.8349
    - GarageArea: 2.4129
- La magnitud tan grande para GarageCars comparada con GarageArea sugiere que el modelo está asignando importancia desproporcionada debido a la multicolinealidad

iv. Variables Temporales

- YearBuilt y YearRemodAdd están relacionadas
- Coeficientes:
    - YearBuilt: 349.7480
    - YearRemodAdd: 187.1189
- La gran diferencia entre coeficiente da un indicio que multicolinealidad

b. Variables que Más Aportan al Modelo

v. Variables con Mayor Contribución Positiva

1. **OverallQual** (19,794.61): La calidad general tiene el mayor impacto individual, con casi $20,000 de incremento por cada punto adicional en la escala de 1-10
2. **GarageCars** (13,344.83): Cada espacio adicional para un automóvil añade más de $13,000
3. **BsmtFullBath** (9,229.40): Cada baño completo en el sótano añade alrededor de $9,200
4. **OverallCond** (4,406.11): La condición general también tiene un impacto significativo

vi. **Variables con Impacto Moderado por Unidad**

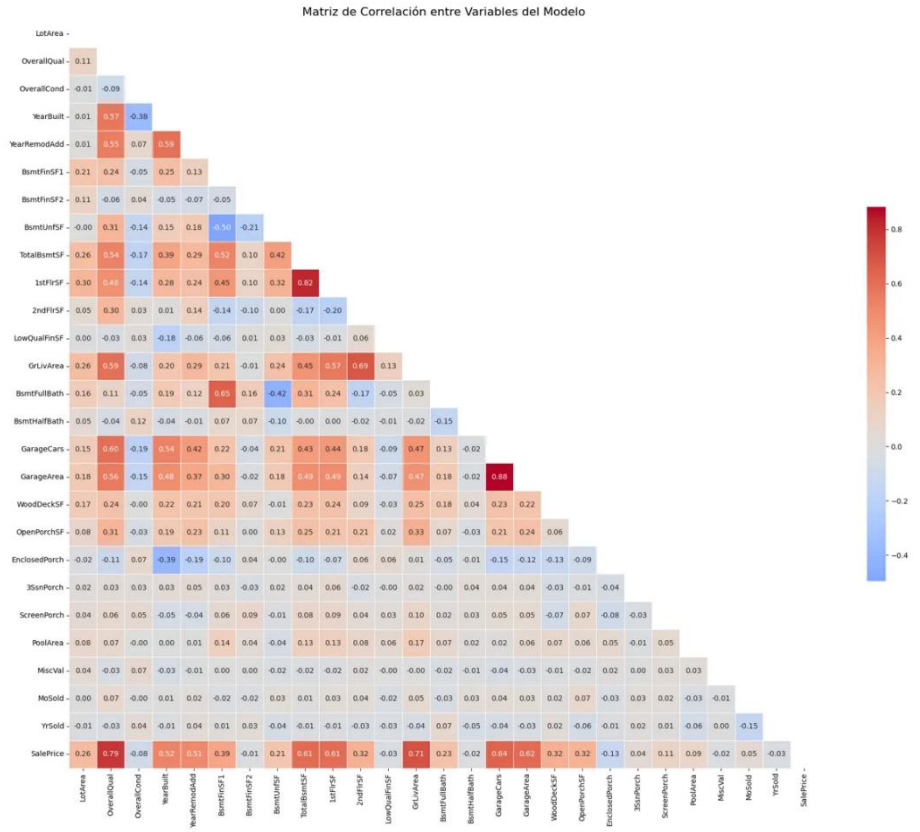
1. **1stFlrSF** (32.53): Cada pie cuadrado adicional en la primera planta
2. **GrLivArea** (26.14): Cada pie cuadrado adicional de área habitable general
3. **2ndFlrSF** (21.03): Cada pie cuadrado adicional en la segunda planta
4. **ScreenPorch** (55.63): Cada pie cuadrado de porche con pantalla

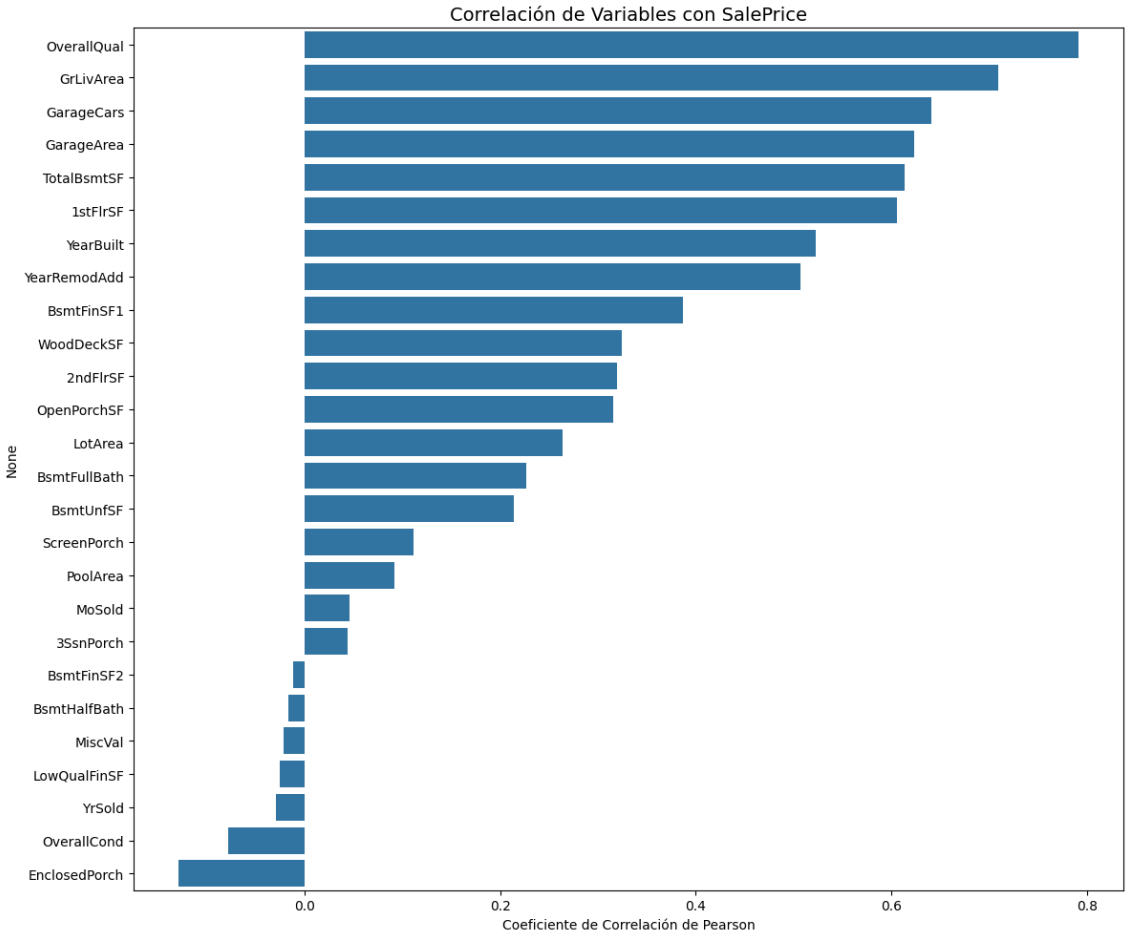
vii. **Variables con Contribución Negativa**

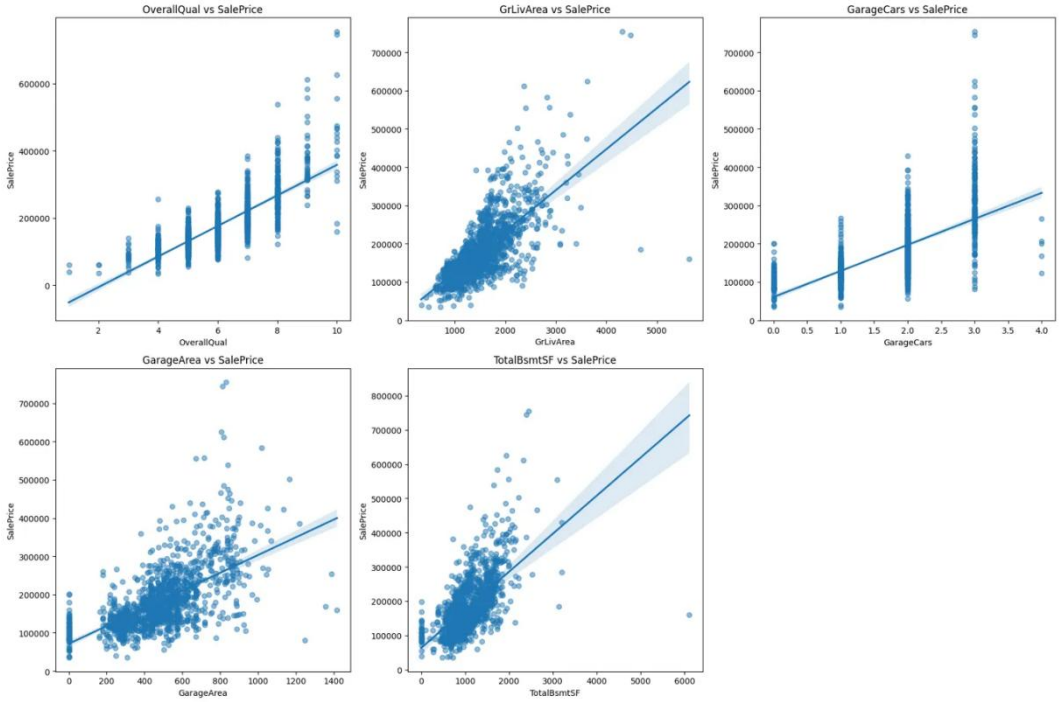
1. **LowQualFinSF** (-27.42): Los acabados de baja calidad reducen el valor
2. **PoolArea** (-50.13): Sorprendentemente, el área de piscina parece reducir el valor
3. **MoSold** (-682.19) y **YrSold** (-670.00): Sugieren tendencias de precio estacionales y anuales

c. Análisis de correlación

viii. El análisis de correlación es esencial para identificar qué características de las viviendas influyen más en su precio de venta. Este conocimiento permite desarrollar modelos predictivos más precisos y evitar problemas estadísticos como la multicolinealidad, mejorando así la capacidad de empresas inmobiliarias como InmoValor S.A para realizar valoraciones objetivas.


Matriz de Correlación entre Variables del Modelo
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF
1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF
EnclosedPorch
3SsnPorch
ScreenPorch
PoolArea
MiscVal
MoSold
YrSold
SalePrice
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Correlación de Variables con SalePrice
OverallQual
GrLivArea
GarageCars
GarageArea
TotalBsmtSF
1stFlrSF
YearBuilt
YearRemodAdd
BsmtFinSF1
WoodDeckSF
2ndFlrSF
OpenPorchSF
LotArea
BsmtFullBath
BsmtUnfSF
ScreenPorch
PoolArea
MoSold
3SsnPorch
BsmtFinSF2
BsmtHalfBath
MiscVal
LowQualFinSF
YrSold
OverallCond
EnclosedPorch
None
0.0
0.2
0.4
0.6
0.8
Coeficiente de Correlación de Pearson



ix. Las tres representaciones gráficas muestran el mismo fenómeno desde diferentes perspectivas. La matriz de correlación (Imagen 1) revela que SalePrice tiene correlaciones positivas fuertes con OverallQual (0.79), GrLivArea (0.69), GarageCars (0.64) y GarageArea (0.62), mientras señala posibles problemas de multicolinealidad entre pares como GarageCars/GarageArea (0.88) y TotalBsmtSF/1stFlrSF (0.82). El gráfico de barras (Imagen 2) confirma estos hallazgos al ordenar las variables según su correlación con SalePrice, destacando la importancia de la calidad general, área habitable y características del garaje, mientras que variables como EnclosedPorch muestran correlaciones negativas débiles. Los gráficos de dispersión (Imagen 3) complementan este análisis mostrando las relaciones específicas: OverallQual exhibe una relación casi lineal con el precio; GrLivArea presenta una tendencia positiva con algunos valores atípicos en áreas grandes; GarageCars muestra un patrón escalonado ascendente debido a su naturaleza

discreta; mientras que GarageArea y TotalBsmtSF presentan relaciones positivas con mayor dispersión a medida que aumenta el área.

x. La calidad general de la vivienda, el área habitable y las características del garaje son los predictores más influyentes del precio. El modelo óptimo debería incluir estas variables clave, evitando incluir simultáneamente variables altamente correlacionadas entre sí para prevenir problemas de multicolinealidad.

d. Overfitting? Sí

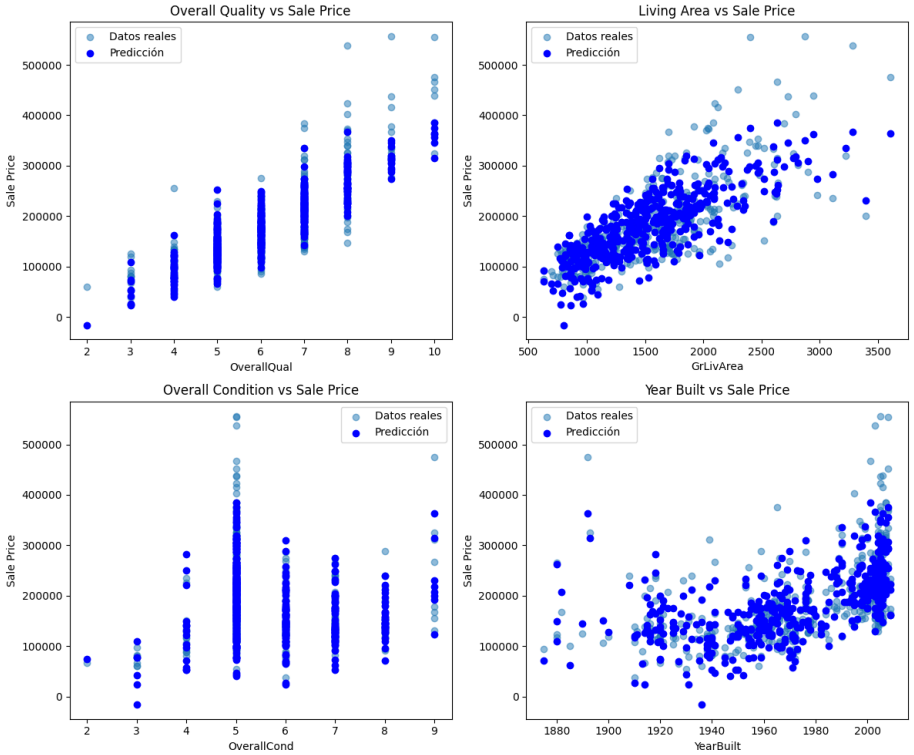
xi. Sí, existen indicios significativos de sobreajuste (overfitting) en el modelo analizado. A continuación presento un análisis detallado:

1. Señales Estadísticas de Sobreajuste: R^2 artificialmente elevado (0.8316):
    a. Este valor puede parecer positivo, pero con 26 variables predictoras, algunas redundantes, es probable que parte de este ajuste se deba a "aprender" ruido de los datos
    b. No disponemos del R^2 ajustado, que penalizaría por el número de variables y probablemente sería menor
2. Multicolinealidad como indicador de sobreajuste:
    a. La severa multicolinealidad identificada (variables redundantes como TotalBsmtSF y sus componentes) es una señal clásica de modelos sobreajustados
    b. Esta redundancia permite al modelo "explicar" variaciones aleatorias específicas del conjunto de entrenamiento

IX. Nuevo modelo

Tomando en cuenta el análisis anterior, este modelo utiliza estas 10 columnas: 'OverallQual', 'OverallCond', 'YearBuilt', 'BsmtFullBath', 'TotalBsmtSF', 'GrLivArea', 'GarageCars', 'WoodDeckSF', 'FullBath'. Eso elimina rastros de multicolinealidad y overfitting. Los resultados son los siguientes:

SalePrice = 179720.0656 + (27329.6499 * OverallQual) + (6486.0627 * OverallCond) + (11090.2987 * YearBuilt) + (6803.0109 * BsmtFullBath) + (8612.5443 * TotalBsmtSF) + (26161.5171 * GrLivArea) + (11137.4944 * GarageCars) + (4046.6524 * WoodDeckSF) + (679.4898 * FullBath)

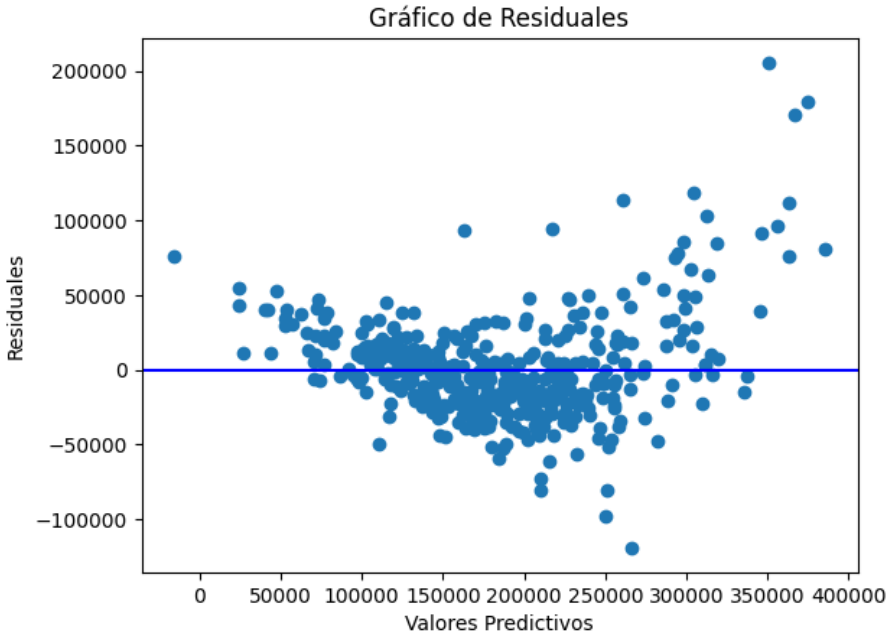


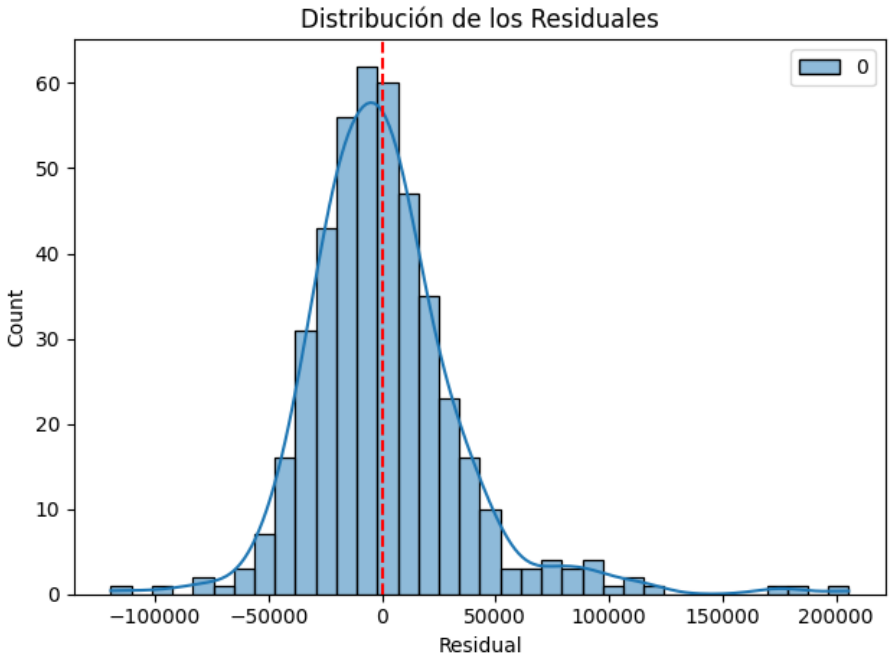
R^2: 0.8174
MSE: 1174827939.0087
RMSE: 34275.7631

El modelo de regresión lineal desarrollado para predecir los precios de las casas explica aproximadamente el 82% de la variabilidad en los datos, según el coeficiente de determinación $R^2 = 0.8174$. Esto indica que el modelo captura correctamente la mayoría de los patrones en los datos, aunque existe un margen de error. El RMSE de aproximadamente 34,276 dólares sugiere que, en promedio, el modelo puede desviarse de los valores reales en esta cantidad. Los coeficientes muestran que las variables más influyentes en el precio son OverallQual, GrLivArea y GarageCars, lo que tiene sentido en el contexto del mercado inmobiliario. Por ejemplo, un punto adicional en la calidad

de la casa aumenta el precio en aproximadamente $27,330, mientras que un metro cuadrado adicional de área habitable agrega $26,161 en promedio. A pesar de su buen desempeño global, el modelo tiende a subestimar los precios de las viviendas más costosas, lo que sugiere posibles limitaciones en la linealidad de las relaciones modeladas.

a. Análisis de los residuos





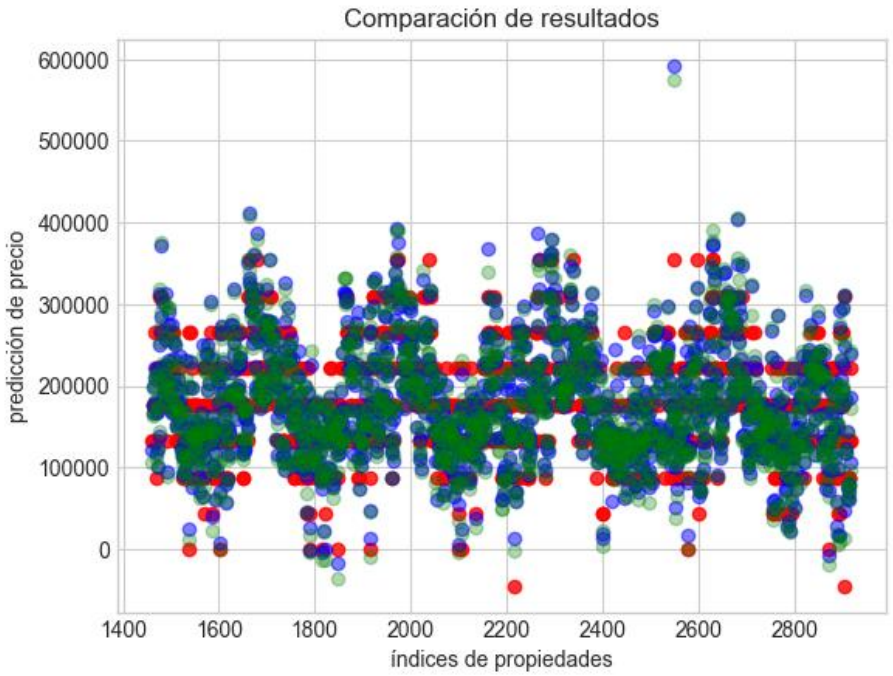
El análisis de los residuales muestra algunos patrones que pueden indicar áreas de mejora. Se observa heterocedasticidad, ya que la variabilidad de los errores aumenta con los precios predichos, lo que sugiere que el modelo predice con mayor precisión las viviendas de precio medio-bajo, pero subestima las propiedades más caras. Además, la presencia de outliers con errores superiores a $150,000 indica que el modelo tiene dificultades para capturar correctamente los extremos del mercado. La distribución de los residuales es aproximadamente normal, aunque presenta una ligera asimetría positiva, reforzando la idea de que ciertos valores altos no están bien ajustados. En general, aunque los residuales muestran una estructura razonable, la presencia de estos patrones sugiere que técnicas adicionales, como transformaciones logarítmicas o modelos con interacciones, podrían mejorar la precisión del modelo.

X. Conjunto de prueba, eficiencia del algoritmo

a. ¿Qué tan bien lo hizo? ¿Qué medidas usó para determinar la calidad de la predicción?

Para evaluar la eficiencia del algoritmo, calculamos R^2 en el conjunto de entrenamiento, obteniendo valores cercanos a 0.6, 0.75 y 0.8, lo que indica que cada modelo explica un 60%, 75% y 80% de la variabilidad del precio de las casas. Sin embargo, como el conjunto de prueba no incluye valores reales de "SalePrice", no podemos calcular métricas como RMSE en este conjunto.

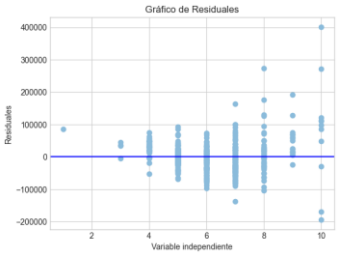
Entonces, para verificar la coherencia de las predicciones, comparamos la distribución de los precios predichos con la de los precios reales del entrenamiento y encontramos que son similares. También analizamos los coeficientes del modelo y observamos que variables como "OverallQual" tienen mayor impacto en el precio de cada modelo, lo cual tiene sentido en el contexto del problema y el análisis estadístico realizado.

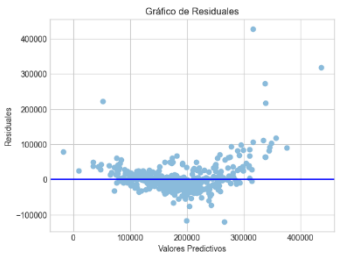


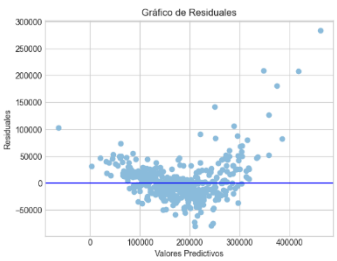
Finalmente, comparamos todos los modelos para verificar que los valores predichos sigan una lógica coherente con los precios esperados. Aunque es normal que las predicciones varíen entre modelos, observamos una tendencia consistente en los valores estimados, lo que indica que los modelos son adecuados. Si bien algunos muestran una mayor precisión, en general, cada uno captura de manera razonable el patrón real de los precios.

XI. Discusión

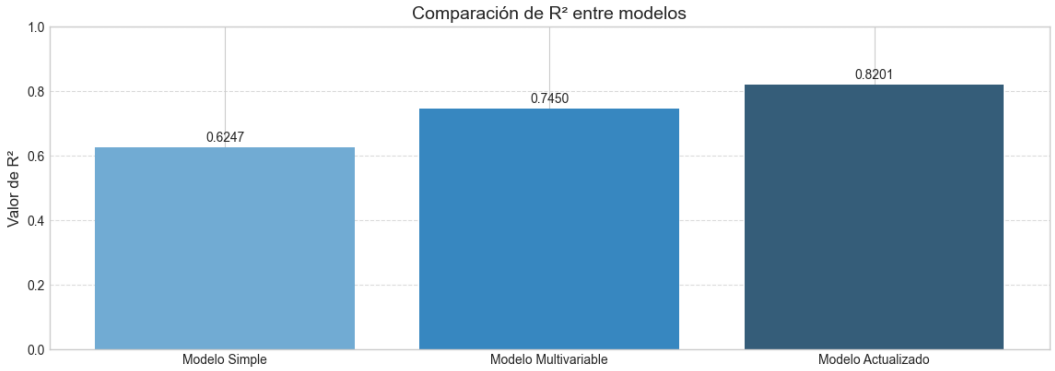
Al comparar los tres gráficos de residuales, podemos observar una clara evolución en la calidad predictiva de los modelos a medida que se incrementa la complejidad. El primer gráfico, que corresponde al modelo univariado (con 1 variable), muestra una marcada heterocedasticidad con residuales que se expanden significativamente al aumentar el valor de la variable independiente, señalando que este modelo simple es insuficiente para capturar la naturaleza de la relación con el precio de venta. El segundo gráfico, del modelo intermedio (26 variables), presenta un patrón similar de "embudo" pero con una dispersión más controlada, lo que indica una mejora sustancial en las predicciones, aunque persiste la tendencia a subestimar los valores más altos. El tercer gráfico, correspondiente al modelo multivariable optimizado (9 variables), muestra la distribución de residuales más equilibrada de los tres, con una menor varianza general y una compresión visible de los errores extremos, particularmente en el rango medio de predicciones, lo que demuestra que este modelo logra un mejor equilibrio entre simplicidad y poder predictivo, confirmando que a veces un modelo más limitado pero bien seleccionado puede superar a modelos más complejos con exceso de variables.

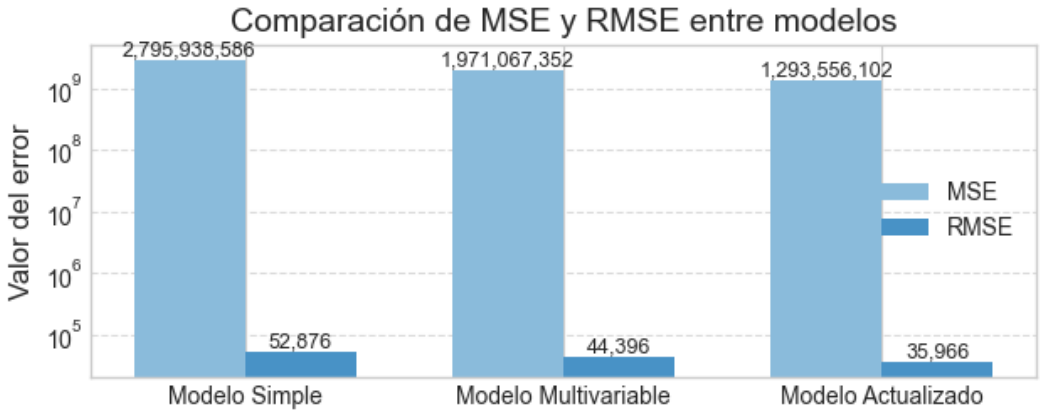






Para una mejor comparación de los estadísticos en cada modelo, realizamos gráficos que los ponen en mejor perspectiva:





El análisis comparativo de los tres modelos de regresión desarrollados revela una clara progresión en la calidad predictiva para la estimación de precios de viviendas. El Modelo Simple, con una sola variable, muestra un R^2 de 0.6247, evidenciando una capacidad explicativa limitada y los mayores errores (RMSE de 52,876). El Modelo Multivariable, que incorpora 26 variables, presenta una mejora sustancial con un R^2 de 0.7450 y un RMSE de 44,396, demostrando que la inclusión de predictores adicionales, captura mejor la complejidad del mercado inmobiliario. Sin embargo, es el Modelo Actualizado, con solo 9 variables cuidadosamente seleccionadas, el que logra el mejor desempeño con un R^2 de 0.8201 y un RMSE de 35,966, representando una mejora del 32% en precisión respecto al modelo inicial.

Estos resultados confirman la importancia de la atención a los detalles en la modelación estadística, demostrando que un conjunto óptimo de variables puede superar a modelos más complejos. La reducción consistente tanto en MSE como en RMSE, junto con el aumento progresivo del R^2, sugiere que no estamos enfrentando problemas de sobreajuste. El error promedio final de aproximadamente $36,000 debe evaluarse en el contexto del mercado inmobiliario estudiado; si consideramos un precio promedio de vivienda de $200,000, esto representa un error relativo del 18%, una precisión razonable para un sector influenciado por numerosos factores subjetivos y contextuales.

Es importante destacar que los dos modelos con múltiples variables presentan resultados similares en términos de capacidad predictiva; sin embargo, la complejidad del Modelo Actualizado se reduce significativamente al utilizar solo 9 variables en lugar de 26, lo que constituye una ventaja importante que permite afirmar que este modelo optimizado es superior. Además, según el análisis previo, el modelo con mayor número de variables muestra indicios de overfitting y problemas de multicolinealidad, lo que puede conducir a resultados imprecisos y menor robustez cuando se aplica a nuevos datos.

Esta reducción en la dimensionalidad no solo mejora la interpretabilidad del modelo, sino que también disminuye el riesgo de capturar relaciones espurias presentes únicamente en los datos de entrenamiento. El Modelo Actualizado, al seleccionar cuidadosamente las variables más influyentes y eliminar redundancias, logra un mejor equilibrio entre sesgo y varianza, lo que se traduce en estimaciones más estables y confiables para la predicción de precios inmobiliarios en diferentes contextos del mercado.

En conclusión, la comparación confirma que agregar más variables no siempre mejora un modelo. En este caso, el modelo con 9 variables es superior, ya que mantiene un alto desempeño predictivo sin la complejidad innecesaria de modelos más extensos. Su menor tamaño mejora la interpretación y la estabilidad en nuevos datos, lo que lo hace la mejor opción para predecir los precios de las viviendas.

XII. Repositorio/Documento

https://github.com/Fabiola-cc/InmoValor_SA

HDT 3. MRL.docx