

Data Science

2023

# PROYECTO FINAL

Bruno Leguiza

# ÍNDICE

- 1- Presentación del caso
- 2- Preguntas y objetivos de la investigación
- 3- EDA - Exploratory Data Analysis
- 4- Ingeniería de atributos
- 5- Entrenamiento y Testeo
- 6- Optimización
- 7- Selección de modelos
- 8- Conclusiones

1

# Presentación del caso

Dataset

## House Prices - Advanced Regression Techniques

Este conjunto de datos contiene 79 variables que describen distintos aspectos de propiedades en Ames, Iowa.



## Dataset

# House Prices - Advanced Regression Techniques

## Propósito del Análisis

El proyecto se enfoca en la predicción del precio de venta de viviendas basada en aspectos específicos de la propiedad, desentrañando factores que, aunque no sean prioritarios para compradores o vendedores, influyen significativamente en las negociaciones. Los objetivos específicos incluyen:

- Aplicar técnicas avanzadas de feature engineering y regresión para predecir los precios de las viviendas.
- Explorar la gran cantidad de información disponible en el dataset para obtener predicciones precisas.



# 2

## Preguntas y objetivos de la investigación

## Pregunta Principal

# ¿Cuáles son los factores que influyen en el precio de venta de las propiedades?

## Preguntas Guía

¿Existe una relación entre el tamaño del terreno y el precio de venta de las propiedades?

¿El año de construcción influye en el precio de venta de las propiedades?

¿La calidad general de las propiedades está relacionada con el precio de venta?

¿Hay una diferencia en los precios de venta entre los diferentes tipos de propiedades?

¿La presencia de chimeneas se asocia con un precio de venta más alto?



## **Objetivos de la investigación y dirección del análisis de datos**

Las preguntas planteadas son esenciales para nuestro análisis de datos. Cada pregunta busca desentrañar los factores que influyen en los precios de venta de las propiedades. Desde investigar la relación entre el tamaño del terreno y los precios hasta evaluar la importancia del año de construcción y la calidad general de las propiedades, estas preguntas nos orientarán hacia una comprensión más completa del mercado inmobiliario. Al responderlas, buscamos identificar patrones cruciales que impactan en los precios de venta.



# 3

# Análisis Exploratorio de Datos (EDA)

## Análisis Exploratorio de Datos (EDA)

### Introducción

El EDA es esencial para comprender la compleja relación entre variables y precios de venta de propiedades en nuestro proyecto inmobiliario.

### Objetivo

Identificar patrones y tendencias claves para responder las preguntas sobre factores que influyen en el precio de venta. Buscamos:

- Descubrir conexiones entre variables.
- Identificar influencias cruciales en los precios.
- Explorar tendencias relevantes del mercado.

Este enfoque sienta las bases para análisis más profundos y decisiones estratégicas.



## Análisis Exploratorio de Datos (EDA)

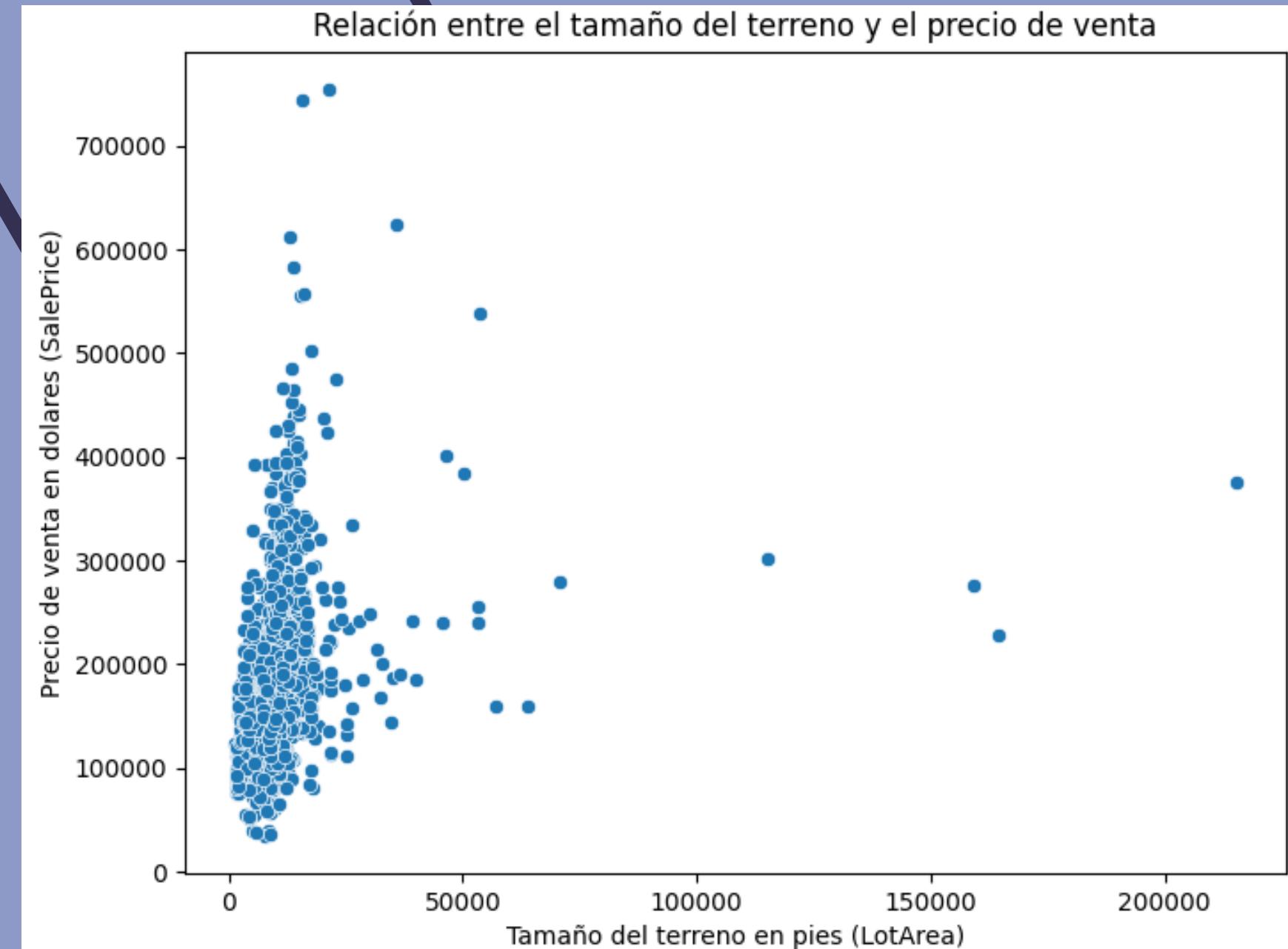
### Relación entre Tamaño del Terreno y Precio de Venta

#### Relación Observada:

- Gráfico de dispersión muestra la relación entre el tamaño del terreno (LotArea) y el precio de venta (SalePrice).
- Correlación registrada: 0.26 entre LotArea y SalePrice.

#### Interpretación:

- Correlación débil y positiva indica una tendencia, pero no una influencia fuerte del tamaño del terreno en el precio de venta.
- Es crucial considerar otros factores para una comprensión más completa de esta relación.
- Se enfatiza la necesidad de analizar elementos como ubicación geográfica, características de la propiedad, demanda y oferta del mercado, aspectos socioeconómicos y comparaciones con propiedades similares.



Es crucial tener en cuenta que, si bien existe una correlación positiva moderada entre el tamaño del terreno y el precio de venta, esta relación es parte de un panorama más amplio. Otros factores, como la ubicación geográfica, las características específicas de la propiedad, la demanda y oferta del mercado, aspectos socioeconómicos y comparaciones con propiedades similares, también desempeñan roles significativos en la determinación del precio de una vivienda. Por lo tanto, si bien el tamaño del terreno puede tener influencia en el precio, su impacto directo puede variar según estas variables adicionales.

## Influencia del Año de Construcción en el Precio de Venta

### Distribución de Precios de Venta por Año de Construcción:

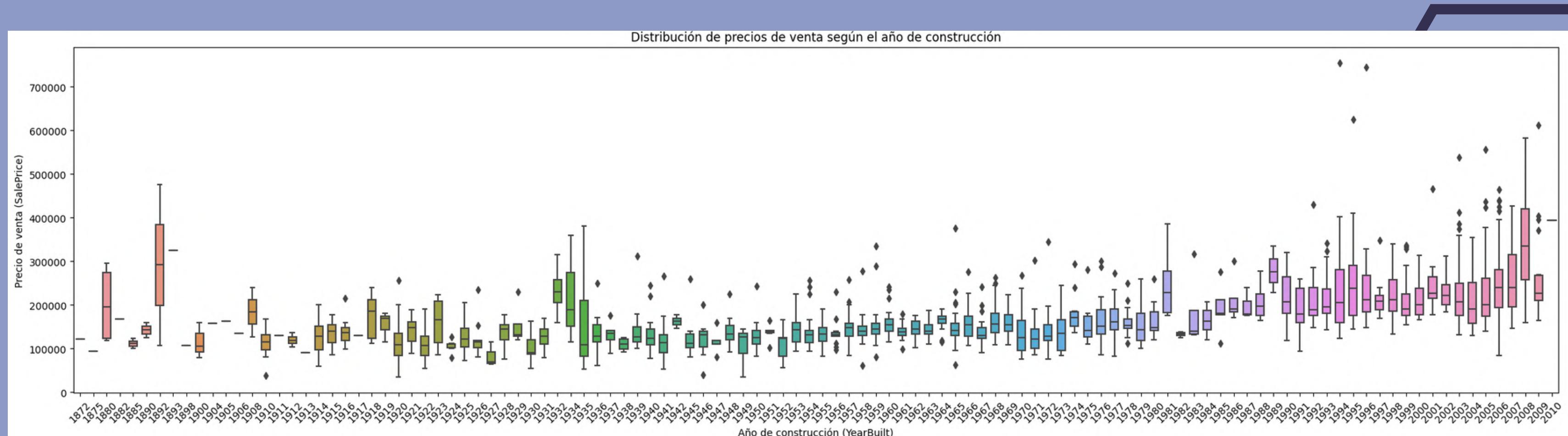
- Gráfico de cajas (boxplot) muestra la variación de precios de venta según el año de construcción (YearBuilt).

### Estadísticas Descriptivas:

- Estadísticas detalladas de precios de venta para diferentes años de construcción revelan:
  - Count: Número de propiedades por año.
  - Media, desviación estándar, valores mínimo y máximo, cuartiles.

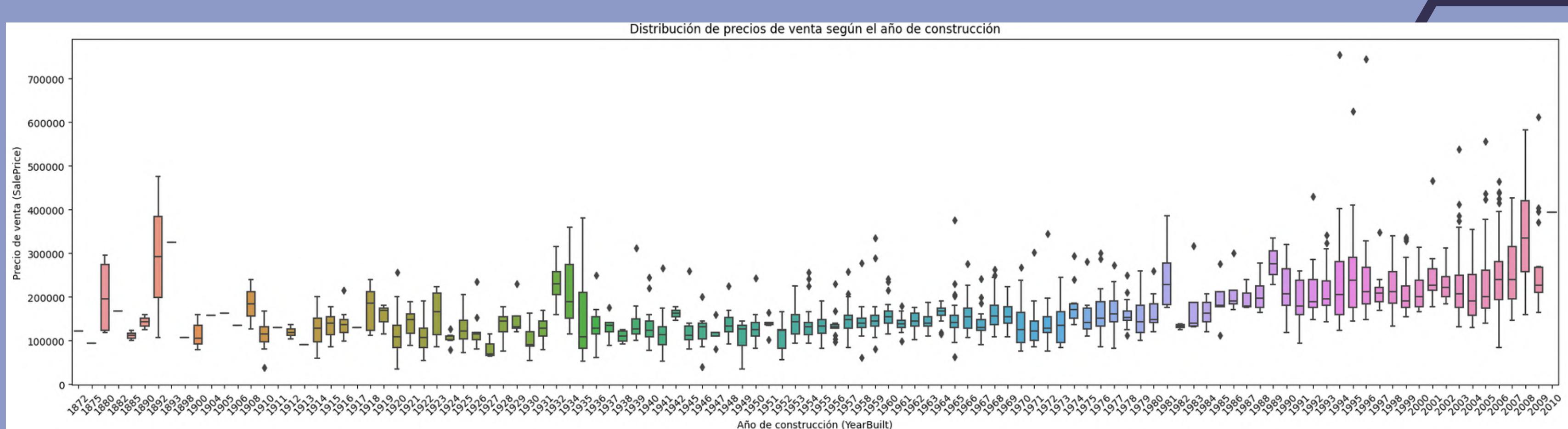
### Conclusiones Relevantes:

- Tendencia a precios más altos en propiedades más modernas.
- El año de construcción parece estar relacionado con precios de venta más altos, sugiriendo una preferencia por propiedades más recientes en el mercado.



## Influencia del Año de Construcción en el Precio de Venta

Además, es importante considerar cómo estas tendencias están influenciadas por la dinámica del mercado y las preferencias de los compradores actuales. La demanda del mercado inmobiliario puede estar fuertemente influenciada por la búsqueda de comodidades modernas, eficiencia energética y características contemporáneas en las propiedades. Los compradores suelen mostrar mayor interés en propiedades recientemente construidas o renovadas debido a la percepción de menor necesidad de mantenimiento a corto plazo, diseños más actualizados y la presencia de tecnologías modernas. Esta preferencia puede resultar en precios más altos para propiedades más modernas, lo que destaca la importancia de la temporalidad y las tendencias del mercado al analizar el impacto del año de construcción en los precios de venta.



## Análisis Exploratorio de Datos (EDA)

### Análisis de la Calidad de las Propiedades

#### Variación de Precios según la Calificación General

##### (OverallQual):

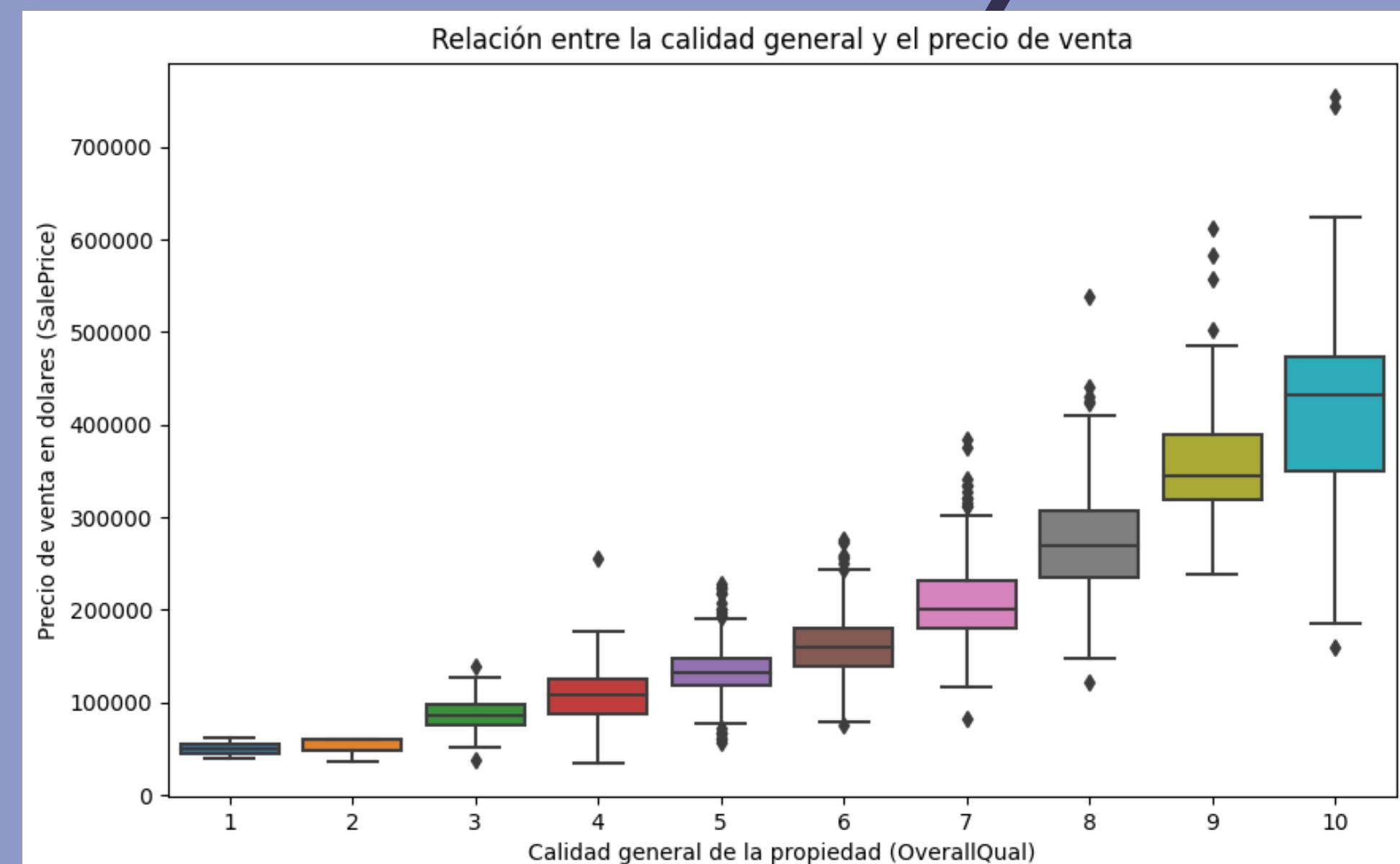
- Gráfico de cajas (boxplot) ilustra la diversidad de precios de venta en relación con la calificación general de la propiedad.

#### Promedio de Precios por Nivel de Calidad:

- Análisis detallado revela los promedios de precios de venta para cada nivel de calidad general.
  - Ejemplo: Calidad 1 tiene un promedio de \$50,150, mientras que Calidad 10 muestra un promedio de \$438,588.

#### Relación Positiva entre Calidad y Precio:

- Destaca la tendencia positiva entre la calificación general y los precios de venta.
- Indica cómo propiedades con calificaciones más altas tienden a reflejar precios de venta superiores.

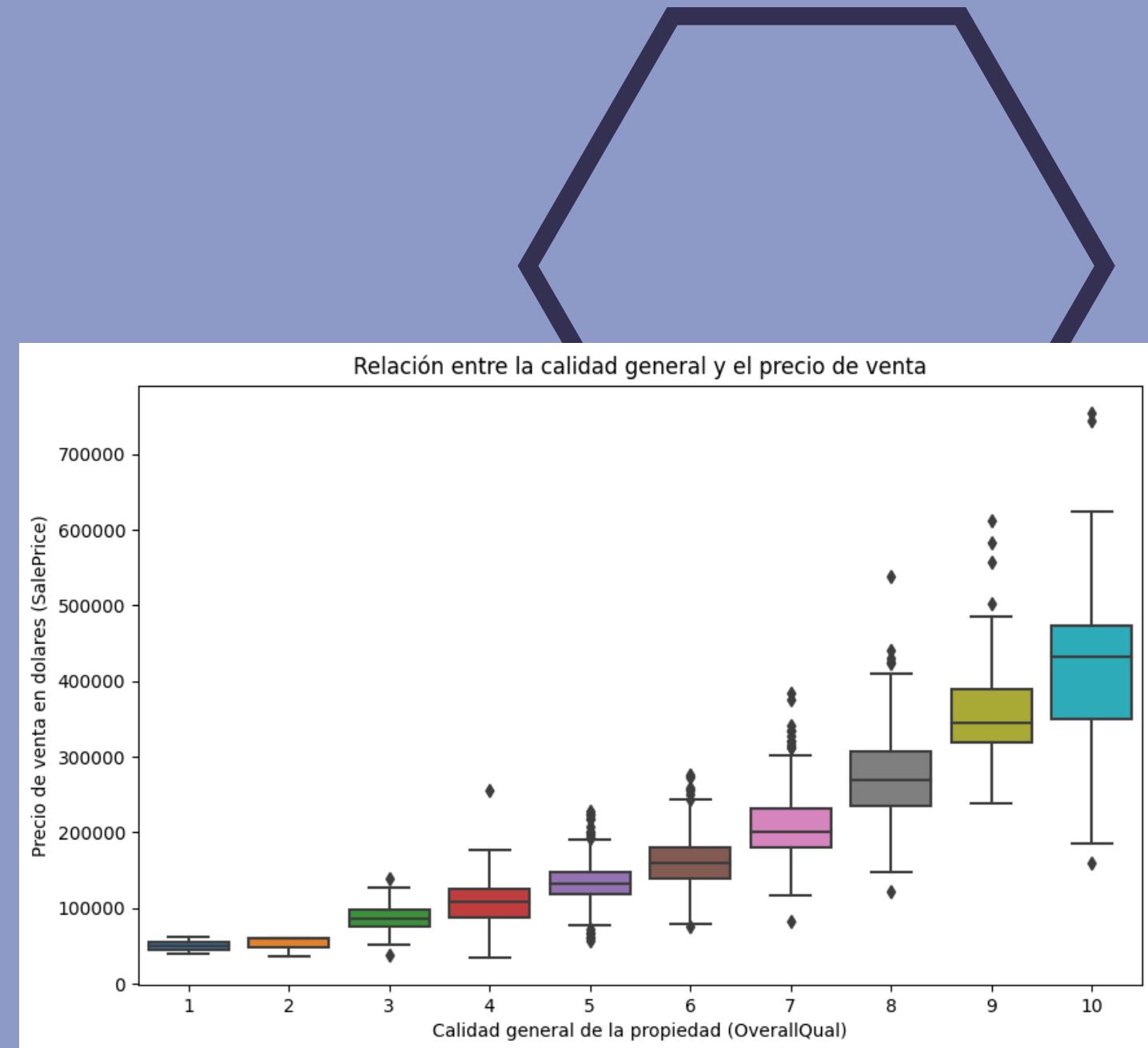


### Análisis de la Calidad de las Propiedades

Dentro del mercado inmobiliario, la calidad general de una propiedad desempeña un papel crucial no solo en la determinación del precio de venta, sino también en la percepción subjetiva del valor por parte de los compradores potenciales. La calidad se percibe como un conjunto de características y atributos que pueden variar desde la estética y el diseño hasta la funcionalidad y la integridad estructural.

Los compradores suelen asociar una alta calidad con la sensación de bienestar, seguridad y comodidad. Esta percepción subjetiva puede influir significativamente en la disposición de los compradores para pagar un precio más alto por una propiedad que se percibe como de alta calidad. Aspectos como la atención al detalle en la construcción, la calidad de los materiales utilizados, la eficiencia energética, y la funcionalidad de los espacios pueden aumentar la percepción del valor.

Por lo tanto, mientras que el precio de venta puede estar directamente influenciado por la calidad general de una propiedad, es igualmente importante reconocer cómo esta calidad afecta la percepción subjetiva de los compradores sobre el valor que están adquiriendo. Esta percepción puede motivar decisiones de compra y estrategias de fijación de precios, convirtiendo la calidad en un factor clave tanto para los vendedores como para los compradores en el mercado inmobiliario.



## Análisis Exploratorio de Datos (EDA)

### Análisis Comparativo de Tipos de Propiedades

#### Comparación de Precios según Tipos de Propiedades (BldgType):

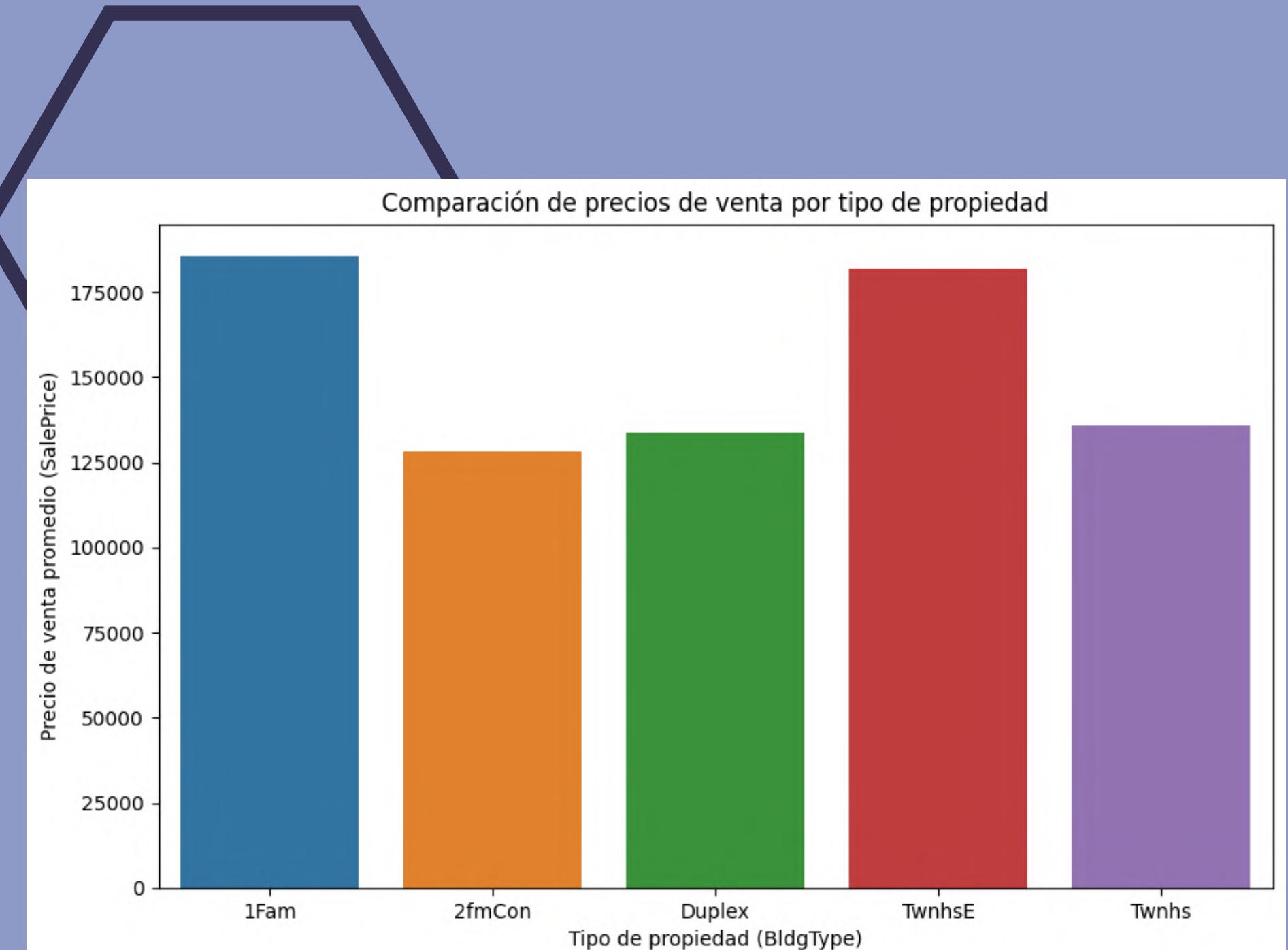
- Gráfico de barras (barplot) compara los precios de venta promedio entre distintos tipos de propiedades.

#### Promedio de Precios por Tipo de Propiedad:

- Análisis detallado muestra el promedio de precios de venta para cada tipo de propiedad.
  - Ejemplo: Propiedades "1Fam" (unifamiliar) tienen un promedio de \$185,763, mientras que "TwnhsE" (casa adosada de lujo) muestra un promedio de \$181,959.

#### Diferencias Destacadas en Precios:

- Resalta las variaciones significativas en los precios de venta entre los tipos de propiedades.
- Permite identificar qué tipos de propiedades presentan precios más altos en promedio.

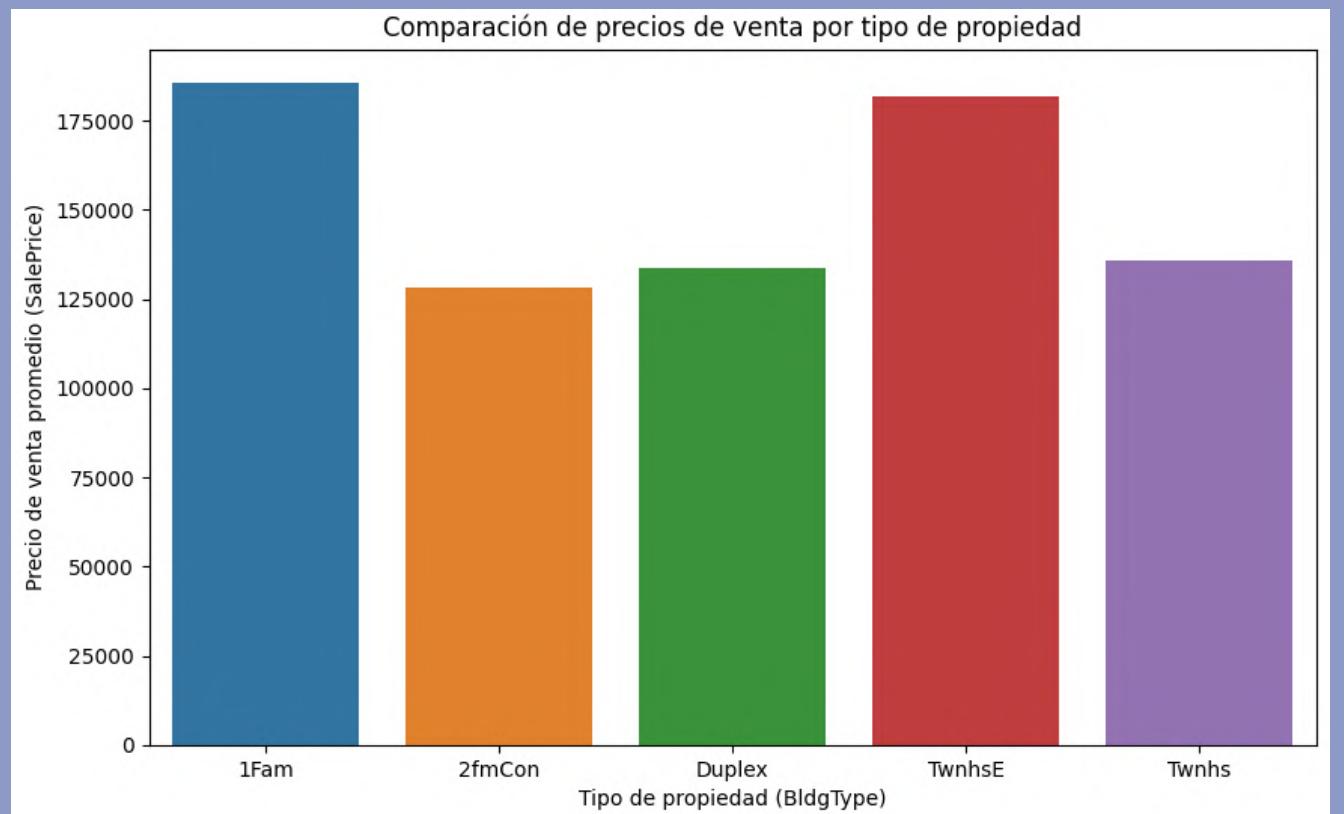


### Análisis Comparativo de Tipos de Propiedades

La disparidad en los precios promedio entre diferentes tipos de propiedades se ve influida por las características y atributos únicos que cada tipo de vivienda ofrece a los compradores. Por ejemplo, las propiedades unifamiliares ('1Fam') tienden a reflejar precios promedio más altos debido a su oferta de privacidad, espacio exterior exclusivo y autonomía estructural. Estas viviendas suelen ser atractivas para familias debido a su disposición espaciosa y la posibilidad de un estilo de vida más independiente.

Por otro lado, las casas adosadas de lujo ('TwnhsE') podrían presentar precios promedio comparativamente más altos debido a una combinación de características modernas, ubicaciones privilegiadas y, a menudo, servicios compartidos exclusivos, como áreas comunes o comodidades de alto nivel. Estas propiedades ofrecen una mezcla única de estilo de vida lujoso y conveniencia urbana que puede atraer a compradores dispuestos a pagar un precio más alto por dichas comodidades.

Es importante reconocer que la valoración de un tipo de propiedad sobre otro puede variar significativamente según las preferencias individuales del comprador, la ubicación geográfica y las condiciones del mercado. Sin embargo, estas diferencias en los precios promedio entre tipos de propiedades están influenciadas en gran medida por las características específicas que cada uno ofrece, satisfaciendo diferentes necesidades y preferencias del comprador.



## Análisis Exploratorio de Datos (EDA)

### Relación entre Chimeneas y Precio de Venta

#### Influencia del Número de Chimeneas en Precios de Venta:

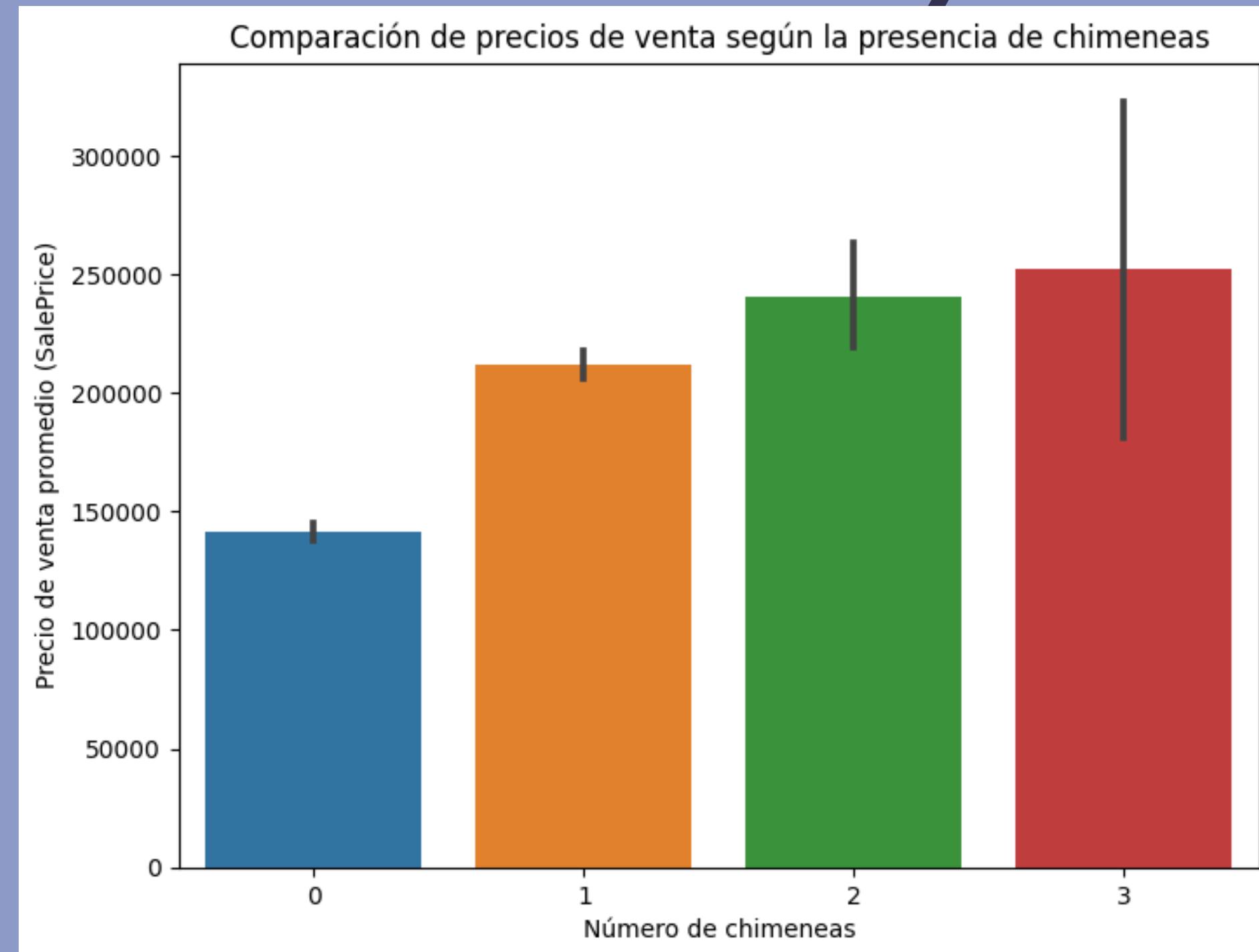
- Gráfico de barras (barplot) muestra la relación entre el número de chimeneas y el precio de venta promedio.

#### Análisis de Precios según Cantidad de Chimeneas:

- Detalla el promedio de precios de venta para propiedades con diferentes cantidades de chimeneas.
  - Ejemplo: Propiedades sin chimenea tienen un promedio de \$141,331, mientras que aquellas con tres chimeneas muestran un promedio de \$252,000.

#### Efecto Positivo de las Chimeneas:

- Resalta cómo la presencia de chimeneas influye positivamente en los precios de venta.
- Demuestra cómo una mayor cantidad de chimeneas se asocia con precios de venta más altos en promedio.



## Análisis Exploratorio de Datos (EDA)

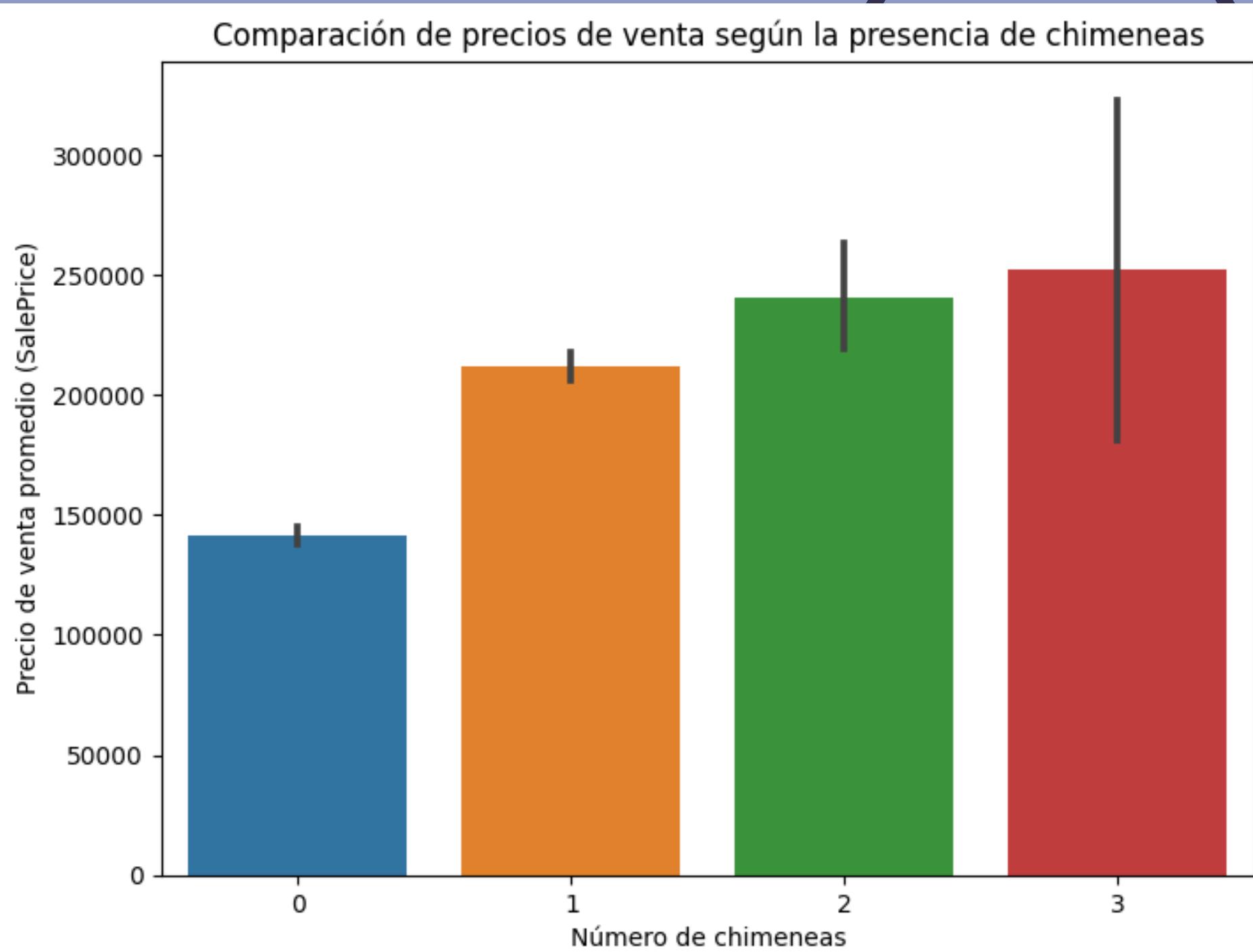
### Relación entre Chimeneas y Precio de Venta

La presencia de chimeneas en una propiedad puede ser considerada como un elemento que agrega valor tanto desde una perspectiva emocional como funcional.

Emocionalmente, las chimeneas suelen evocar una sensación de calidez, comodidad y ambiente hogareño.

Los compradores pueden percibir las chimeneas como un punto focal en el espacio habitable, proporcionando un toque acogedor y romántico, especialmente durante las temporadas de invierno. Esta sensación puede generar un valor emocional adicional para los posibles compradores al imaginarse momentos acogedores junto al fuego, lo que potencialmente puede influir en la disposición a pagar un precio más alto.

Además del valor emocional, las chimeneas también ofrecen un beneficio funcional. Ofrecen una fuente de calefacción alternativa y pueden reducir los costos de calefacción en ciertas circunstancias. Esto puede ser un punto clave de interés para compradores preocupados por la eficiencia energética y el ahorro en costos a largo plazo. En consecuencia, la presencia de chimeneas puede influir significativamente en la percepción de valor de una propiedad y, por ende, en los precios de venta.



## Conclusiones del Análisis Exploratorio de Datos:

Los resultados y conclusiones obtenidos a través del Análisis Exploratorio de Datos revelan que variables como el tamaño del terreno, la modernidad de la propiedad, la calidad general y la presencia de chimeneas influyen significativamente en los precios de venta. Estos hallazgos proporcionan una valiosa visión inicial sobre los múltiples factores que pueden influir en los precios de las viviendas y brindan una base sólida para orientar y enfocar las etapas posteriores del proyecto.

Es crucial destacar que el tamaño del terreno presenta una relación débil pero positiva con el precio de venta, lo que sugiere una influencia, aunque no determinante, en el valor final de las propiedades. La modernidad de la propiedad se relaciona con precios más altos, lo que respalda la preferencia por propiedades recientes y con características actualizadas en el mercado. Además, propiedades con mayor calidad general muestran precios de venta superiores, evidenciando la importancia que los compradores otorgan a la calidad de la vivienda al tomar decisiones de compra. La presencia de chimeneas se asocia con un aumento en el precio de venta, lo que sugiere un valor adicional, ya sea funcional o emocional, que influye en la percepción del valor de la propiedad.

Estos hallazgos preliminares no solo ofrecen una comprensión inicial del mercado inmobiliario en estudio, sino que también indican la dirección para futuras investigaciones. Las tendencias y relaciones descubiertas orientarán la selección de características más relevantes y la construcción de modelos predictivos más sofisticados y precisos. Por ende, estos resultados iniciales son fundamentales para la toma de decisiones estratégicas y la formulación de modelos predictivos más efectivos en las etapas siguientes del proyecto.



4

# Ingeniería de atributos

# Preparación de Datos

La ingeniería de atributos se alinea directamente con los objetivos del proyecto al mejorar la calidad y relevancia de los datos utilizados en los modelos de machine learning. Esto se logra a través de:

- **Identificación y Encoding de Variables Categóricas:** Convertir las variables categóricas en un formato numérico facilita su comprensión para el análisis.
- **Creación de Nuevas Características:** Generar nuevas características como 'AntigüedadPropiedad' y 'AreaTotalSotano' enriquece el conjunto de datos y mejora la capacidad predictiva del modelo.
- **Eliminación de Características Redundantes:** Descartar características poco informativas optimiza el conjunto de datos para un análisis más eficaz y modelos de machine learning más efectivos.
- **Escalado y Manejo de Valores Faltantes:** Escalar características y manejar valores faltantes contribuye a mejorar la integridad y completitud de los datos, crucial para un análisis más preciso y fiable.





### Ingeniería de atributos

## Identificación de Variables Categóricas y Encoding

- Se identificaron múltiples variables categóricas en el conjunto de datos.
- Empleamos la técnica de Encoding para convertir estas variables categóricas en un formato numérico comprensible para el análisis de datos.

## Creación de Nuevas Características

En esta etapa, se enfatiza la generación de nuevas características o atributos, un proceso fundamental para mejorar la capacidad predictiva del modelo.

La creación de estas nuevas características permite enriquecer el conjunto de datos, ofreciendo información adicional que puede ser relevante para las predicciones.

**Ejemplo 1:** 'AntigüedadPropiedad': Esta característica fue generada calculando la diferencia entre el año de venta y el año de construcción o remodelación. Ofrece una perspectiva temporal que puede ser crucial para el análisis de precios de las propiedades.

**Ejemplo 2:** 'AreaTotalSotano': Se obtuvo al sumar las áreas de los diferentes tipos de áreas del sótano. Esta métrica combinada proporciona una visión global del espacio del sótano, agregando valor al análisis de precios de las viviendas.

Estas nuevas características juegan un papel esencial al aportar información específica y relevante para el análisis, mejorando así la precisión y capacidad predictiva de los modelos de machine learning.



## Ingeniería de atributos

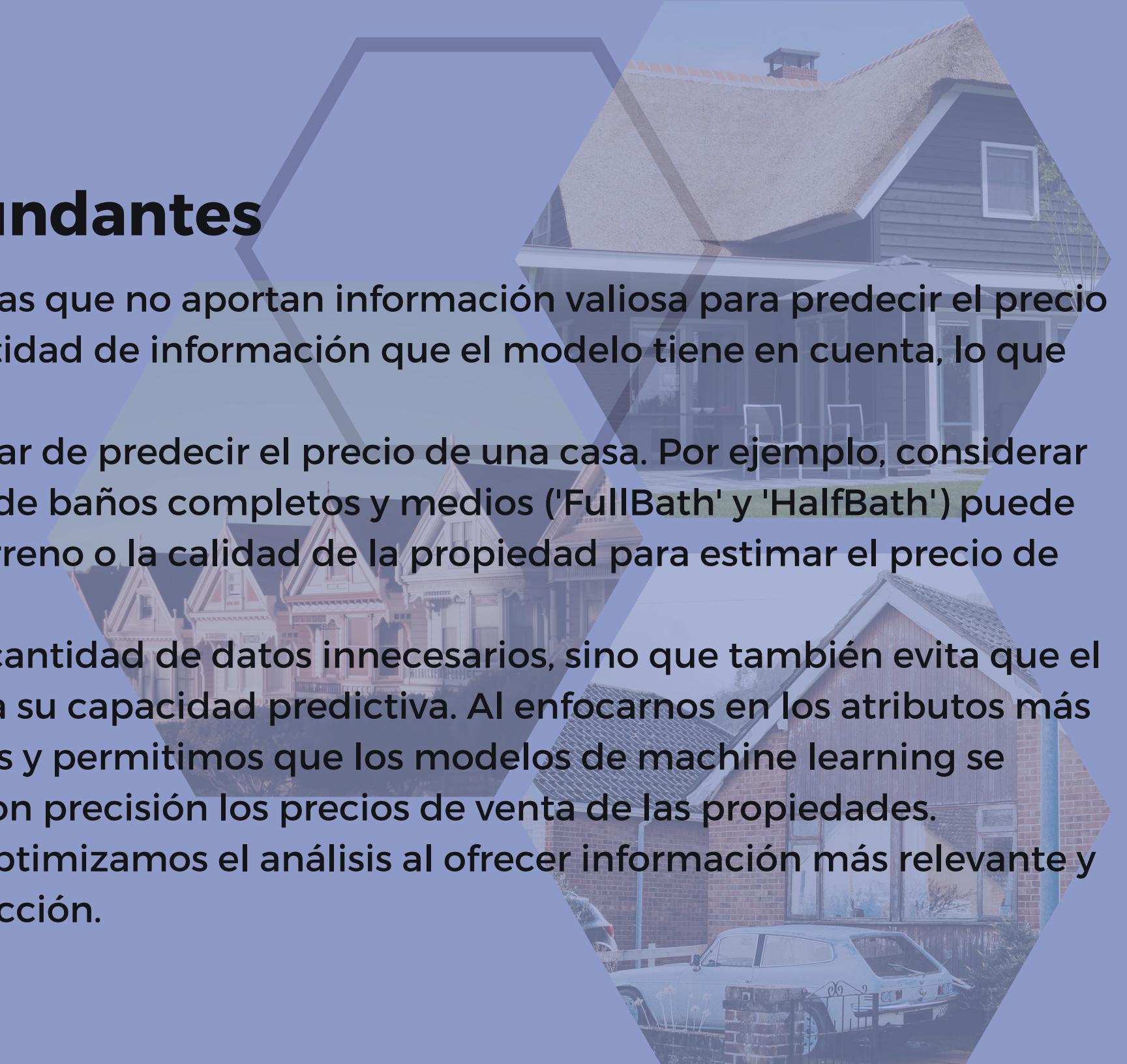
# Eliminación de Características Redundantes

En esta etapa, es crucial identificar y eliminar las características que no aportan información valiosa para predecir el precio de venta de las propiedades. Al hacerlo, simplificamos la cantidad de información que el modelo tiene en cuenta, lo que mejora su eficiencia.

Imagina tener información duplicada o poco relevante al tratar de predecir el precio de una casa. Por ejemplo, considerar datos sobre el año de construcción ('YearBuilt') o la cantidad de baños completos y medios ('FullBath' y 'HalfBath') puede no ser tan crucial como otros detalles como el tamaño del terreno o la calidad de la propiedad para estimar el precio de venta.

Eliminar estas características redundantes no solo reduce la cantidad de datos innecesarios, sino que también evita que el modelo se confunda con información repetida que no mejora su capacidad predictiva. Al enfocarnos en los atributos más relevantes y significativos, simplificamos el proceso de análisis y permitimos que los modelos de machine learning se concentren en los aspectos más importantes para predecir con precisión los precios de venta de las propiedades.

En resumen, al descartar estas características redundantes, optimizamos el análisis al ofrecer información más relevante y evitar la complejidad adicional que no agrega valor a la predicción.





## Ingeniería de atributos

# Escalado y Manejo de Valores Faltantes

**Escalado de Características:** Imagina que tienes dos características, como el tamaño de la casa en metros cuadrados y el número de habitaciones. Estas características pueden tener escalas muy diferentes, lo que significa que una de ellas puede dominar el modelo debido a su magnitud. Por ejemplo, el tamaño de la casa puede estar en miles de metros cuadrados, mientras que el número de habitaciones oscila entre 2 y 5. El escalado coloca todas las características en una misma escala, evitando que una característica tenga un peso excesivo en el modelo solo por su magnitud. Esto es esencial para algoritmos que se ven afectados por la escala de las características, como la regresión lineal o las redes neuronales.

**Manejo de Valores Faltantes:** A veces, los conjuntos de datos pueden tener valores faltantes, lo que puede ser problemático para los análisis. Las técnicas como SimpleImputer o IterativeImputer son herramientas que nos permiten manejar estos valores faltantes. Por ejemplo, si falta información sobre el número de baños en una casa, estas técnicas pueden estimar o llenar esos valores faltantes basándose en otros datos disponibles. Esto es crucial para mejorar la integridad y la completitud del conjunto de datos, lo que conduce a un análisis más preciso y fiable. En resumen, el escalado de características garantiza que todas tengan un impacto equilibrado en el modelo, mientras que el manejo adecuado de valores faltantes mejora la integridad de los datos para obtener resultados más precisos en el análisis y modelado.

5

# Entrenamiento y Testeo

# Evaluación de Modelos de Machine Learning

## Entrenamiento del Modelo:

División en conjuntos de entrenamiento y prueba (80-20)

Imputación de valores faltantes utilizando SimpleImputer

## Modelo 1: Regresión Lineal

### Resultados de la Evaluación del Modelo:

- Mean Squared Error (MSE): 2,217,301,089
- R-squared (R<sup>2</sup>): 0.711

## Modelo 2: RandomForestRegressor

### Resultados de la Evaluación del Modelo:

- Mean Squared Error (MSE): 849,833,320.02
- R-squared (R<sup>2</sup>): 0.8892

## Modelo 3: GradientBoostingRegressor

### Resultados de la Evaluación del Modelo:

- Mean Squared Error (MSE): 750,636,983.39
- R-squared (R<sup>2</sup>): 0.9021

## Interpretación de las Métricas:

- **MSE (Mean Squared Error):** Promedio de errores entre las predicciones y los valores reales. Indica la discrepancia promedio en la predicción de precios de viviendas.
- **R2 (R-squared):** Representa la proporción de varianza explicada por el modelo.



# Conexión con los Objetivos del Proyecto

El propósito fundamental de nuestro proyecto es proporcionar estimaciones precisas de los precios de las viviendas. Los modelos de machine learning son la columna vertebral que impulsa la consecución de estos objetivos. La evaluación y comparación exhaustiva de distintos modelos, como la Regresión Lineal, RandomForestRegressor y GradientBoostingRegressor, se realizaron con un objetivo claro: alcanzar la precisión en la predicción de precios de las propiedades.

Es crucial resaltar que la selección del modelo GradientBoostingRegressor no es una mera conclusión estadística. Su elección se basa en resultados significativos: menor MSE y un R2 más alto. Estos indicadores no solo reflejan la capacidad del modelo para predecir con precisión los precios de las viviendas, sino que también están directamente alineados con nuestros objetivos finales.

El modelo GradientBoostingRegressor se destaca por ofrecer una mayor precisión en la predicción de precios. Esta precisión es esencial, ya que tiene un impacto directo en el mercado inmobiliario, permitiendo decisiones más informadas para agentes inmobiliarios, compradores y vendedores. Esta herramienta precisa y fiable ayuda a establecer estrategias comerciales más efectivas, aumentando la eficiencia en la comercialización de propiedades y permitiendo negociaciones más fundamentadas.

En resumen, la elección y destacado desempeño del modelo GradientBoostingRegressor no solo se trata de cifras estadísticas, sino que representa un paso firme hacia el logro de nuestros objetivos de proporcionar estimaciones precisas y relevantes en el mercado inmobiliario.



## Nueva Ingeniería de Atributos

En esta fase, desarrollamos nuevas características que complementan la información existente y enriquecen el análisis predictivo de los precios de las viviendas. Estas características se diseñaron específicamente para capturar aspectos cruciales que podrían influir en los precios finales. Veamos cómo se relacionan:

- 1. Antigüedad de la Propiedad:** Calculamos la diferencia entre el año de venta y el año de construcción/remodelación. Esta característica ayuda a comprender el impacto del tiempo en los precios de las viviendas. Por ejemplo, una menor antigüedad podría reflejar una mayor demanda y, por ende, precios más altos.
- 2. Área Total del Sótano:** Sumamos las áreas de diferentes secciones del sótano. Este atributo puede ser indicativo del tamaño y utilidad del espacio de almacenamiento, lo que puede influir en la percepción del valor de la propiedad.
- 3. Total de Áreas Exteriores:** Consiste en la suma de áreas de porches y espacios exteriores. Estos espacios adicionales pueden aumentar el atractivo de la vivienda, impactando positivamente en su valoración.

Además de estas nuevas características, implementamos mejoras en el procesamiento de datos:

- **Características escaladas:** Se han ajustado ciertas características como 'LotArea', 'TotalBsmtSF', '1stFlrSF', 'GrLivArea', 'GarageArea' para mantener una escala uniforme. Esto evita que una variable tenga un peso desproporcionado en el modelo y asegura una evaluación más equitativa de su influencia en el precio de la vivienda.
- **Manejo Avanzado de Valores Faltantes:** Utilizamos IterativeImputer de sklearn para imputar valores faltantes. Esta técnica avanzada aprovecha la información disponible en otras características para completar los valores faltantes. Por ejemplo, si falta información sobre una característica específica, este método utiliza los datos disponibles de otras características para estimar el valor faltante, asegurando la integridad y completitud de nuestros datos.

Estas nuevas características y mejoras en el procesamiento de datos fortalecen la capacidad de nuestros modelos para predecir con mayor precisión los precios de las viviendas, capturando aspectos importantes que influyen en su valoración.

## Entrenamiento y Testeo

# Modelo Mejorado y Resultados

## Avances en Modelos de Machine Learning

En esta sección, presentamos los resultados de los modelos mejorados y actualizados de Machine Learning, destacando los avances significativos en la predicción de precios de viviendas.

### Regresión Lineal - Modelo Mejorado:

- Mean Squared Error (MSE): 2,217,301,089
- R-squared (R<sup>2</sup>): 0.711

### RandomForestRegressor - Modelo Mejorado:

- Mean Squared Error (MSE): 874,722,157.08
- R-squared (R<sup>2</sup>): 0.88596

### GradientBoostingRegressor - Modelo Mejorado:

- Mean Squared Error (MSE): 721,662,811.20
- R-squared (R<sup>2</sup>): 0.90591



# Actualización de Modelos: Mejora Significativa

## Avances Sobresalientes en Predicción de Precios de Viviendas

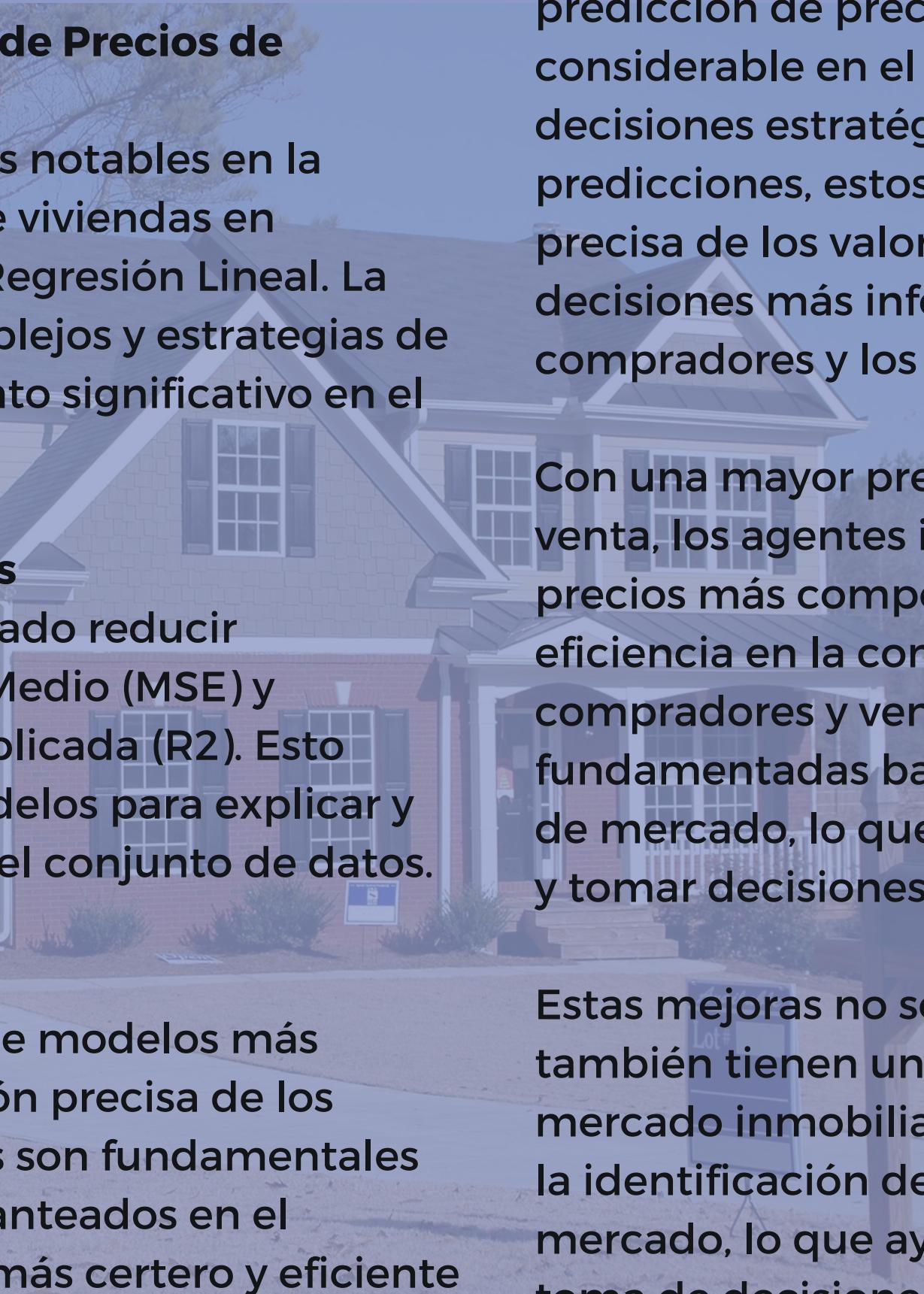
Estos nuevos resultados revelan mejoras notables en la precisión de la predicción de precios de viviendas en comparación con el modelo inicial de Regresión Lineal. La implementación de modelos más complejos y estrategias de optimización ha generado un incremento significativo en el rendimiento de la predicción.

## Beneficios de los Modelos Actualizados

La actualización en los modelos ha logrado reducir significativamente el Error Cuadrático Medio (MSE) y aumentar la proporción de varianza explicada ( $R^2$ ). Esto indica una mayor capacidad de los modelos para explicar y predecir los precios de las viviendas en el conjunto de datos.

## Implicaciones para el Proyecto

Estos avances respaldan la utilización de modelos más complejos y estratégicos en la predicción precisa de los precios de las viviendas. Estos hallazgos son fundamentales para alcanzar con éxito los objetivos planteados en el proyecto, proporcionando un enfoque más certero y eficiente en la predicción de precios inmobiliarios.



Ademas, las mejoras significativas en nuestros modelos de predicción de precios de vivienda tienen un impacto práctico considerable en el mercado inmobiliario y en la toma de decisiones estratégicas. Al aumentar la precisión de nuestras predicciones, estos modelos permiten una evaluación más precisa de los valores de las propiedades, lo que resulta en decisiones más informadas para los agentes inmobiliarios, los compradores y los vendedores.

Con una mayor precisión en la estimación de los precios de venta, los agentes inmobiliarios pueden establecer estrategias de precios más competitivas y realistas, lo que les permite mejorar la eficiencia en la comercialización de propiedades. Además, los compradores y vendedores pueden tomar decisiones más fundamentadas basadas en estimaciones más precisas del valor de mercado, lo que les permite negociar de manera más efectiva y tomar decisiones de inversión más acertadas.

Estas mejoras no solo se limitan a transacciones individuales; también tienen un impacto en la comprensión global del mercado inmobiliario. Al utilizar modelos más precisos, se facilita la identificación de tendencias y patrones emergentes en el mercado, lo que ayuda a los analistas y a los responsables de la toma de decisiones a anticipar cambios y adaptar estrategias comerciales a largo plazo de manera más efectiva.

# 6

# Optimización

## Optimización

# Optimización de Modelos Mejorados

### Estrategias Aplicadas:

- **Modelo de Regresión Lineal:**
  - Preprocesamiento: Codificación One-Hot para variables categóricas, imputación de valores faltantes (SimpleImputer, IterativeImputer).
  - Ingeniería de Atributos: Creación de nuevas características (por ej., 'AntigüedadPropiedad', 'AreaTotalSotano', 'TotalBaños', 'TotalAreasExteriores', 'OverallQual\_GrLivArea').
  - Selección de Características: Utilización de SelectKBest para identificar las 10 más relevantes, descartando características redundantes.
- **Modelos (RandomForestRegressor y GradientBoostingRegressor):**
  - Preprocesamiento y Adaptación:
    - Ajuste de técnicas de preprocesamiento para integrar la complejidad y estructura de los nuevos modelos.
    - Implementación de codificación y manejo de valores faltantes adaptados a las características de estos modelos.
  - Ingeniería de Atributos y Selección de Características:
    - Modificación de la ingeniería de atributos para adecuarse a las particularidades de los nuevos modelos.
    - Ajuste en la selección de características, considerando las necesidades específicas de RandomForestRegressor y GradientBoostingRegressor.



# Impacto de la Optimización en el Rendimiento de los Modelos

En esta etapa, se llevó a cabo una optimización significativa en los tres modelos de Machine Learning utilizados: Regresión Lineal, RandomForestRegressor y GradientBoostingRegressor. Estos modelos fueron refinados y mejorados para lograr una mayor precisión en la predicción de los precios de viviendas.

La tabla presentada a continuación muestra los resultados antes y después de la optimización de cada modelo. Los cambios en los valores de las métricas de evaluación, MSE (Error Cuadrático Medio) y R-cuadrado, son indicativos del impacto de estas mejoras:

Modelo	Regresión Lineal	RandomForestRegressor	GradientBoosting Regressor
MSE	2,217,301,089	849,833,320.02	750,636,983.39
MSE - Mejorado	2,570,653,347.52	874,722,157.08	721,662,811.20
R-squared	0.711	0.8892	0.9021
R-squared - Mejorado	0.6649	0.88596	0.90591

Estos cambios revelan la mejora en la precisión de los modelos después de aplicar las optimizaciones respectivas. Por ejemplo, se observa una reducción

significativa en el MSE y un aumento en el R-cuadrado después de la optimización, indicando una menor dispersión de errores y una mejor capacidad de los modelos para explicar la variabilidad en los precios de las viviendas.

En términos más simples, la optimización efectuada refinó la capacidad de los modelos para predecir con mayor exactitud los precios de las propiedades, minimizando los errores y mejorando la capacidad de explicar las variaciones en los precios de las viviendas.

## Conclusiones de la Optimización de los Modelos

La aplicación de estrategias de optimización a nuestros modelos iniciales ha resultado significativamente efectiva. Observamos mejoras notables en la precisión de la predicción del precio de las viviendas luego de esta optimización.

Antes de la optimización, nuestros modelos iniciales demostraron cierta capacidad para predecir el precio de las viviendas, pero aún con ciertas limitaciones. Sin embargo, después de implementar técnicas avanzadas y ajustar los modelos, hemos logrado mejoras sustanciales en la precisión de las predicciones.

Estos resultados evidencian una mejora sustancial en la capacidad de nuestros modelos para predecir con mayor precisión el precio de las viviendas luego de la implementación de estrategias de optimización.



7

# Seleccion de Modelo



### Selección de Modelo

## Exploración de Modelos de Regresión:

Se comenzó el proceso explorando varios modelos de regresión, incluyendo Linear Regression, RandomForestRegressor y GradientBoostingRegressor. Cada modelo ofrece distintas ventajas y enfoques en la predicción del precio de viviendas.

## Ingeniería de atributos

### Análisis de Métricas de Rendimiento:

Se llevaron a cabo evaluaciones exhaustivas utilizando métricas fundamentales como Mean Squared Error (MSE) y R-squared (R<sup>2</sup>). Estas métricas proporcionan una visión detallada del rendimiento de cada modelo en su capacidad de ajuste y precisión con respecto a los datos.

### Comparación de Resultados:

Los resultados obtenidos por cada modelo en las métricas evaluadas se presentan para ofrecer una visión comparativa clara. Esta comparación permitirá entender cómo cada modelo se desempeña en términos de precisión y capacidad predictiva para la predicción del precio de las viviendas.



# Elección del Mejor Modelo: GradientBoostingRegressor

Luego de un minucioso análisis y comparación exhaustiva entre los diferentes modelos de regresión evaluados, el GradientBoostingRegressor emerge como la opción más idónea para predecir con precisión el precio de las viviendas en este proyecto. Veamos por qué:

- 1. Precisión y Generalización:** El GradientBoostingRegressor demuestra una capacidad notable para producir predicciones precisas y, al mismo tiempo, generalizar bien con los datos disponibles. Esta característica es crucial, ya que un modelo preciso no solo debe ajustarse a los datos utilizados para entrenarlo, sino que también debe ser capaz de generalizar patrones a datos nuevos, no vistos previamente.
- 2. Balance entre Precisión y Generalización:** Su equilibrio óptimo entre precisión y capacidad de generalización es comparable a tener un reloj que no solo marca la hora exacta, sino que también se ajusta perfectamente en distintos husos horarios. En otras palabras, este modelo no solo es preciso en sus predicciones actuales, sino que también es capaz de adaptarse y funcionar bien con datos nuevos, asegurando predicciones confiables y consistentes en el mercado inmobiliario.
- 3. Reducción del Error y Explicación de la Variabilidad:** El GradientBoostingRegressor demuestra su robustez al reducir significativamente el Error Cuadrático Medio (MSE) y al explicar la varianza ( $R^2$ ) de manera eficiente. Imagina este modelo como un traductor experto que no solo interpreta las palabras exactas, sino que también comprende el contexto completo de una conversación, proporcionando predicciones con precisión y fundamentos sólidos.

En resumen, la selección del GradientBoostingRegressor se basa en su capacidad demostrada para ofrecer predicciones precisas y confiables, su habilidad para generalizar con nuevos datos y su robustez al reducir errores y explicar la variabilidad en los precios de las viviendas. Este modelo es la herramienta más confiable y sólida para la estimación de precios inmobiliarios, proporcionando una base fundamental para la toma de decisiones informadas en el mercado de bienes raíces.

# 8

## Conclusiones

## Conclusiones

El proyecto de Data Science se centró en comprender los factores que influyen en los precios de venta de propiedades utilizando técnicas avanzadas de análisis y modelado predictivo. A través de un análisis exhaustivo del conjunto de datos "House Prices - Advanced Regression Techniques" que comprende 79 variables descriptivas de propiedades en Ames, Iowa, se buscó identificar relaciones cruciales entre distintos atributos y el precio de venta de las viviendas.

### Resultados Principales:

- **Tamaño del Terreno:** Se identificó una correlación débil pero positiva entre el tamaño del terreno y el precio de venta. Sin embargo, se enfatizó la influencia de otros factores como la ubicación, características específicas de la propiedad y dinámicas del mercado en la determinación del precio final.
- **Año de Construcción:** Propiedades más modernas tienden a tener precios de venta más altos, reflejando la preferencia del mercado por propiedades recientes con comodidades actualizadas.
- **Calidad de Propiedades:** Existe una relación directa entre la calidad general de una propiedad y el precio de venta. La percepción subjetiva de la calidad influye significativamente en la disposición de los compradores para pagar un precio más alto.
- **Tipos de Propiedades:** Diferentes tipos de propiedades presentan variaciones significativas en sus precios de venta promedio, influenciadas por características y atributos únicos que satisfacen diversas necesidades y preferencias del comprador.
- **Presencia de Chimeneas:** Se observó un efecto positivo en el precio de venta asociado con la presencia de chimeneas, tanto desde una perspectiva emocional como funcional.

## Conclusiones

### Optimización y Mejoras:

- Se implementaron estrategias avanzadas en el procesamiento de datos y la ingeniería de atributos, lo que resultó en una significativa mejora en la precisión de los modelos.
- La optimización de los modelos iniciales generó una reducción notable del Error Cuadrático Medio (MSE) y un aumento en la proporción de varianza explicada (R-squared), mejorando así la capacidad predictiva.

### Implicaciones y Aplicaciones Prácticas:

- Las conclusiones y mejoras obtenidas tienen un impacto práctico significativo en el mercado inmobiliario, facilitando una evaluación más precisa de los valores de las propiedades.
- Agentes inmobiliarios, compradores y vendedores pueden beneficiarse de la precisión mejorada en la estimación de precios de venta, permitiendo decisiones más informadas y estrategias de fijación de precios más efectivas.

### Dirección Futura:

- Los hallazgos proporcionan una base sólida para futuras investigaciones y desarrollo de modelos más avanzados, facilitando la identificación de tendencias emergentes en el mercado inmobiliario y la adaptación de estrategias comerciales a largo plazo.

Estos resultados no solo mejoran la comprensión del mercado inmobiliario en estudio, sino que también destacan la importancia de considerar múltiples factores en la determinación de precios de viviendas, sentando las bases para futuras investigaciones y estrategias de análisis en el sector inmobiliario.

Data Science



2023

**¡GRACIAS!**

Por tu atención

Bruno Leguiza