Adrian Ramirez

tRABAJO FIN DE MASTER

Automatización de predicciones de precios de ventas y gestión y almacenamiento en base de datos SQLServer



Contenido

[Resumen 1](#_Toc176725298)

[Introducción 2](#_Toc176725299)

[Contexto y Justificación del proyecto 2](#_Toc176725300)

[Objetivos y preguntas del proyecto 2](#_Toc176725301)

[Modelos de Aprendizaje Supervisado del Proyecto 3](#_Toc176725302)

[Random Forest 3](#_Toc176725303)

[XGBoost 3](#_Toc176725304)

[LightGBM 4](#_Toc176725305)

[Comparación de modelos 5](#_Toc176725306)

[Análisis descriptivo de los datos 6](#_Toc176725307)

[Variables 6](#_Toc176725308)

[Preprocesamiento de Datos para el Modelado Predictivo 8](#_Toc176725309)

[Análisis visual y conclusiones observadas sobre los datos 9](#_Toc176725310)

[Análisis boxplots 9](#_Toc176725311)

[Análisis histogramas 10](#_Toc176725312)

[Análisis scatter plots 11](#_Toc176725313)

[Mapa de calor 12](#_Toc176725314)

[Creación del pipeline de procesado de datos 12](#_Toc176725315)

[Normalización de variables 12](#_Toc176725316)

[PCA 12](#_Toc176725317)

[Pipeline final 13](#_Toc176725318)

[Creación y evaluación de los modelos 14](#_Toc176725319)

[Procesos ETL y Base de datos 14](#_Toc176725320)

[Descripción de la Base de Datos "HouseSales" 14](#_Toc176725321)

[Estructura de las Tablas de Dimensiones 14](#_Toc176725322)

[Estructura de las Tablas de Hechos 15](#_Toc176725323)

[Tabla Central de la Base de Datos 16](#_Toc176725324)

[Relaciones entre las Tablas 16](#_Toc176725325)

[Análisis y Uso de los Datos 16](#_Toc176725326)

[Ventajas de Utilizar Tablas de Dimensiones y Hechos 16](#_Toc176725327)

[Proceso ETL 17](#_Toc176725328)

[Anexos 18](#_Toc176725329)

[Anexo A: Matriz de Correlación de Variables 18](#_Toc176725330)

[Anexo B: Bloxplots respecto a SalePrice 19](#_Toc176725331)

[Anexo C: Histogramas 23](#_Toc176725332)

[Anexo D: Scatter plots 27](#_Toc176725333)

[Anexo E: Mapa de calor 29](#_Toc176725334)

[Anexo F: Scree Plot 30](#_Toc176725335)

[Anexo G: Varianza acumulada por componentes principales 30](#_Toc176725336)

[Anexo E: Diagrama Base de datos 31](#_Toc176725337)

# Resumen

Este Trabajo de Fin de Máster se centra en la predicción de precios de viviendas utilizando el dataset de Ames Housing, un conjunto de datos extenso y detallado disponible públicamente. El objetivo principal es desarrollar un modelo predictivo robusto para estimar el precio de venta de las viviendas, integrando características estructurales y cualitativas del inmueble.

**Metodología:**

1. **Análisis Exploratorio de Datos (EDA):** Se realiza una exploración inicial para entender la estructura y distribución de los datos. Se llevan a cabo análisis estadísticos descriptivos y se generan visualizaciones, como mapas de calor, para identificar patrones y correlaciones significativas.
2. **Mejora del Dataset:** Se enriquece el dataset original mediante la fusión con el de California Housing Prices, otro dataset de Kaggle. Esta integración permite asignar coordenadas de latitud y longitud a las viviendas. Además, se generan direcciones aleatorias para cada propiedad, creando un entorno comercial más realista.
3. **Preprocesamiento de Datos:** Las variables categóricas se convierten en formatos adecuados mediante codificación one-hot. Se manejan los valores faltantes mediante imputación y se normalizan las variables numéricas.
4. **Modelado Predictivo:** Se aplican técnicas avanzadas de modelización, incluyendo Random Forest, XGBoost y LightGBM. Los modelos se comparan en términos de precisión y rendimiento utilizando métricas como RMSE y MAE. Se selecciona el modelo más adecuado en base a su capacidad predictiva y se realizan las predicciones de precios de las viviendas.
5. **Interpretación de Resultados:** Se analiza la importancia de las características en el modelo seleccionado, proporcionando insights sobre las variables con mayor impacto en la predicción de precios.
6. **Integración en Base de Datos:** Opcionalmente, los resultados y las predicciones se almacenan en una base de datos SQL Server. Se crean tablas de hechos y dimensiones, y se utiliza SSIS para realizar la carga, transformación y carga (ETLT) de los datos, facilitando su integración con aplicaciones empresariales y el análisis continuo.

Este proyecto no solo mejora la precisión en la estimación de precios de viviendas, sino que también proporciona una infraestructura avanzada para análisis y visualización, reflejando un entorno comercial práctico y aportando valor adicional en el campo del análisis de datos y modelización predictiva.

# Introducción

## Contexto y Justificación del proyecto

En un entorno económico globalizado y altamente competitivo, la valoración precisa de propiedades inmobiliarias se ha convertido en una necesidad crítica para una variedad de actores, desde compradores y vendedores hasta inversores y agentes inmobiliarios. La capacidad de predecir con precisión los precios de las viviendas puede ofrecer ventajas significativas en la toma de decisiones, influir en las estrategias de inversión y contribuir a la estabilidad del mercado inmobiliario.

El dataset de Ames Housing, ampliamente utilizado en el ámbito del análisis de datos y la ciencia de datos, proporciona una rica colección de información sobre características residenciales, lo que lo convierte en un recurso valioso para desarrollar modelos predictivos. A pesar de su popularidad, la aplicación de técnicas avanzadas de modelización y el enriquecimiento de los datos con información adicional pueden ofrecer nuevas perspectivas y mejoras en la precisión de las predicciones.

En este contexto, la integración del dataset de Ames Housing con el dataset de California Housing Prices y la generación de direcciones aleatorias permiten crear un entorno comercial más realista y completo. Estas mejoras en el dataset original buscan no solo aumentar la precisión del modelo predictivo sino también ofrecer una solución práctica que pueda ser utilizada en aplicaciones empresariales reales.

## Objetivos y preguntas del proyecto

**Objetivo Principal:** Desarrollar un modelo predictivo robusto que estime con precisión el precio de venta de viviendas utilizando el dataset de Ames Housing, enriquecido con datos adicionales de California Housing Prices y direcciones aleatorias.

**Objetivos Específicos:**

1. Realizar un análisis exploratorio de datos (EDA) para entender la estructura y las características del dataset.
2. Implementar técnicas avanzadas de modelización predictiva, incluyendo Random Forest, XGBoost y LightGBM, y seleccionar el modelo más adecuado basado en su rendimiento.
3. Integrar datos adicionales para mejorar el contexto de las propiedades y evaluar el impacto en las predicciones.
4. Desarrollar una infraestructura para almacenar y gestionar los resultados en una base de datos SQL Server, facilitando su uso en aplicaciones empresariales.

# Modelos de Aprendizaje Supervisado del Proyecto

## Random Forest

**Concepto:** Random Forest es un algoritmo de ensamblaje basado en árboles de decisión, desarrollado por Leo Breiman en 2001. Su enfoque se centra en la construcción de múltiples árboles de decisión durante el proceso de entrenamiento, con el objetivo de realizar predicciones mediante la combinación de las predicciones individuales de cada árbol. Este método utiliza el bootstrap aggregating (bagging) y la selección aleatoria de características para reducir la varianza y mejorar la precisión general del modelo.

**Teoría:**

* **Bootstrap Aggregating (Bagging):** Cada árbol en el bosque se entrena utilizando una muestra aleatoria de los datos de entrenamiento. Este proceso de muestreo aleatorio ayuda a reducir el sobreajuste al promediar las predicciones de múltiples árboles, proporcionando una estimación más robusta y estable.
* **Random Subspace Method:** En cada nodo del árbol, se selecciona aleatoriamente un subconjunto de características para realizar la división de los datos. Esta técnica promueve la diversidad entre los árboles, lo cual es crucial para mejorar la generalización del modelo.

**Ventajas:**

* **Robustez:** Random Forest es menos susceptible al sobreajuste en comparación con un solo árbol de decisión debido a la agregación de múltiples árboles.
* **Manejo de Datos Faltantes:** Puede gestionar datos faltantes sin necesidad de imputación adicional, ya que cada árbol puede utilizar diferentes subconjuntos de datos.

**Desventajas:**

* **Complejidad:** El modelo puede ser menos interpretable debido a la cantidad de árboles y al proceso de ensamblaje, lo que dificulta la comprensión de las decisiones del modelo.
* **Computacionalmente Costoso:** La necesidad de entrenar y hacer predicciones con un gran número de árboles puede requerir un tiempo considerable y mayores recursos computacionales.

## XGBoost

**Concepto:** XGBoost (Extreme Gradient Boosting) es una técnica de boosting basada en árboles de decisión que ha ganado popularidad por su alto rendimiento en competiciones de Machine Learning y su eficiencia computacional. Desarrollado por Tianqi Chen, XGBoost optimiza el método de gradient boosting al incorporar técnicas avanzadas que mejoran tanto la velocidad como la precisión del modelo.

**Teoría:**

* **Gradient Boosting:** XGBoost construye árboles de forma secuencial, donde cada nuevo árbol corrige los errores residuales del modelo anterior. Utiliza un enfoque de optimización basado en el descenso del gradiente para ajustar los árboles, lo que permite mejorar el rendimiento del modelo iterativamente.
* **Regularización:** Incluye términos de regularización L1 y L2 en la función objetivo para controlar el sobreajuste. Esto proporciona mayor flexibilidad y control sobre el ajuste del modelo, contribuyendo a un mejor rendimiento general.
* **Técnicas de Optimización:** Emplea técnicas avanzadas como el approximate tree learning y histogram-based algorithms para acelerar el proceso de entrenamiento y reducir el uso de memoria.

**Ventajas:**

* **Rendimiento Superior:** Generalmente ofrece una mayor precisión y rendimiento en comparación con otros métodos de boosting, especialmente en competiciones de Machine Learning.
* **Eficiencia:** Es rápido en el entrenamiento y la predicción, y maneja eficazmente conjuntos de datos grandes y características de alta dimensión.

**Desventajas:**

* **Complejidad:** Requiere una cuidadosa sintonización de hiperparámetros para alcanzar el mejor rendimiento, lo cual puede ser un proceso complejo y demandante.
* **Menor Interpretabilidad:** Al igual que otros métodos de boosting, XGBoost puede ser menos interpretable que modelos más simples, dificultando la comprensión de sus predicciones.

## LightGBM

**Concepto:** LightGBM (Light Gradient Boosting Machine) es otro algoritmo de boosting optimizado para manejar grandes volúmenes de datos de manera eficiente. Desarrollado por Microsoft, LightGBM utiliza técnicas avanzadas como histogramas y aprendizaje por gradiente para ofrecer un entrenamiento rápido y un menor uso de memoria, haciendo que sea particularmente adecuado para conjuntos de datos grandes.

**Teoría:**

* **Gradient Boosting:** Al igual que XGBoost, LightGBM construye árboles secuencialmente para corregir errores residuales y utiliza el descenso del gradiente para mejorar el modelo de manera iterativa.
* **Histogram-Based Method:** Utiliza histogramas para agrupar características en bins, lo que acelera el proceso de entrenamiento y reduce el uso de memoria en comparación con los métodos tradicionales de boosting.

**Ventajas:**

* **Velocidad:** LightGBM es extremadamente eficiente en el entrenamiento y la predicción, especialmente en grandes conjuntos de datos, gracias a su enfoque optimizado.
* **Escalabilidad:** Está diseñado para manejar grandes volúmenes de datos y alta dimensionalidad de manera efectiva, lo que lo hace adecuado para aplicaciones en el mundo real con grandes cantidades de información.

**Desventajas:**

* **Sensibilidad a Hiperparámetros:** Requiere una cuidadosa configuración de hiperparámetros para optimizar el rendimiento, lo que puede ser un desafío.
* **Menor Interpretabilidad:** Similar a otros métodos de boosting, puede ser menos interpretable que modelos basados en árboles individuales, complicando la interpretación de sus predicciones.

## Comparación de modelos

Al comparar Random Forest, XGBoost y LightGBM, es esencial comprender cómo cada uno aborda el problema del aprendizaje automático y sus respectivas ventajas y desventajas.

Random Forest se basa en la combinación de múltiples árboles de decisión construidos con muestras aleatorias de datos y subconjuntos de características. Este enfoque proporciona un modelo robusto que reduce el riesgo de sobreajuste y maneja bien los datos faltantes. Sin embargo, puede ser computacionalmente costoso y menos interpretable debido a su complejidad y al número de árboles involucrados.

XGBoost emplea el método de boosting, donde cada árbol sucesivo corrige los errores del anterior mediante el descenso del gradiente. Este enfoque le permite obtener un alto rendimiento en términos de precisión y eficiencia, convirtiéndolo en una opción popular en competiciones de Machine Learning. No obstante, XGBoost requiere una cuidadosa sintonización de hiperparámetros y puede ser menos interpretable en comparación con modelos más simples debido a su complejidad.

LightGBM también utiliza boosting, pero está optimizado para manejar grandes volúmenes de datos de manera más eficiente mediante el uso de histogramas para agrupar características y reducir el uso de memoria. Esto le confiere una ventaja en términos de velocidad y escalabilidad en comparación con XGBoost. Sin embargo, al igual que XGBoost, LightGBM requiere una adecuada configuración de hiperparámetros y puede ser menos interpretable debido a la complejidad del modelo y los ajustes necesarios.

En términos de interpretabilidad, Random Forest ofrece una mayor transparencia debido a que permite un análisis más claro de cada árbol individual y la evaluación de la importancia de las características. En contraste, LightGBM resulta menos interpretable debido a la interacción compleja entre los árboles y la necesidad de ajustar numerosos hiperparámetros. Por lo tanto, Random Forest brinda una mayor comprensión del modelo, mientras que LightGBM proporciona una mejor precisión y eficiencia a costa de una menor claridad en su funcionamiento.

En resumen, Random Forest destaca por su robustez y simplicidad en la implementación, mientras que XGBoost y LightGBM ofrecen mejoras en precisión y eficiencia. LightGBM se destaca particularmente en velocidad y manejo de grandes volúmenes de datos.

# Análisis descriptivo de los datos

## Variables

Este conjunto de datos contiene información detallada sobre diversas características de propiedades vendidas en Ames, Iowa. Cada variable proporciona un aspecto específico de las propiedades, desde su tipo de construcción hasta la calidad de los materiales. A continuación, se detallan las principales variables incluidas en el conjunto de datos:

1. **MSSubClass**: Identifica el tipo de vivienda involucrada en la venta. Esta variable tiene diferentes valores numéricos, cada uno representando un estilo de construcción específico, como casas de una planta construidas después de 1946 (valor 20) o viviendas de dos familias convertidas (valor 190).
2. **MSZoning**: Clasificación general de zonificación de la propiedad en el momento de la venta. Los valores incluyen categorías como "A" para Agricultura, "C" para Comercial, y "RL" para Residencial de Baja Densidad.
3. **LotFrontage**: Representa la cantidad de pies lineales de calle conectada a la propiedad, una medida importante para evaluar la accesibilidad y potencial uso del terreno.
4. **LotArea**: Superficie total del lote en pies cuadrados, proporcionando una medida del tamaño del terreno de la propiedad.
5. **Street**: Tipo de acceso a la propiedad. Puede ser "Grvl" para grava o "Pave" para pavimentado, lo cual afecta tanto el valor como la accesibilidad de la vivienda.
6. **Alley**: Tipo de acceso a la propiedad a través de un callejón, con opciones como "Grvl" para grava, "Pave" para pavimentado, o "None" si no tiene acceso a callejón.
7. **LotShape**: Forma general de la propiedad, que puede influir en la facilidad de construcción y el uso del terreno. Los valores varían desde "Reg" para regular hasta "IR3" para muy irregular.
8. **LandContour**: Describe la planitud del terreno, con opciones como "Lvl" para nivelado, "Bnk" para inclinado, y "HLS" para ladera.
9. **Utilities**: Tipo de servicios públicos disponibles para la propiedad, como "AllPub" para todos los servicios públicos o "NoSeWa" para electricidad y gas solamente.
10. **LotConfig**: Configuración del lote, indicando si es un lote interno, en esquina, o en un callejón sin salida, entre otros.
11. **LandSlope**: Describe la pendiente del terreno, lo cual puede impactar en el drenaje y la construcción, con opciones como "Gtl" para pendiente suave o "Sev" para pendiente severa.
12. **Neighborhood**: Ubicación física dentro de los límites de la ciudad de Ames, como "CollgCr" para College Creek o "NridgHt" para Northridge Heights, que puede afectar el precio y la demanda de la vivienda.
13. **Condition1 y Condition2**: Proximidad a diversas condiciones o características, como calles arteriales o vías del tren, que pueden influir en el valor de la propiedad debido al ruido o la accesibilidad.
14. **BldgType**: Tipo de vivienda, desde "1Fam" para viviendas unifamiliares hasta "Duplx" para dúplex, afectando el uso del suelo y el mercado objetivo.
15. **HouseStyle**: Estilo de construcción de la vivienda, con opciones como "1Story" para casas de una planta o "SFoyer" para viviendas con entrada dividida.
16. **OverallQual y OverallCond**: Calificación general de la calidad del material y el acabado de la casa, así como de su estado general, en una escala del 1 (muy pobre) al 10 (muy excelente).
17. **YearBuilt y YearRemodAdd**: Año de construcción original y año de remodelación o adición, proporcionando contexto sobre la antigüedad y la modernización de la propiedad.
18. **RoofStyle y RoofMatl**: Estilo y material del techo, como "Gable" para techos a dos aguas o "CompShg" para tejas compuestas, que afectan tanto el estilo como la durabilidad de la vivienda.
19. **Exterior1st y Exterior2nd**: Material de revestimiento exterior de la casa, que puede influir en el atractivo estético y la resistencia a las inclemencias del tiempo.
20. **MasVnrType y MasVnrArea**: Tipo y área de revestimiento de mampostería, detalles que aportan al valor estético y estructural de la propiedad.
21. **ExterQual y ExterCond**: Calidad y condición del material en el exterior de la casa, evaluando la durabilidad y el mantenimiento requerido.
22. **Foundation**: Tipo de cimiento de la propiedad, como "PConc" para concreto vertido, influye en la estabilidad y los costos de construcción.
23. **BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinSF1, BsmtFinType2, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF**: Describen diferentes características del sótano, como la calidad, el acabado y la exposición, importantes para determinar el espacio habitable adicional y la funcionalidad de la vivienda.
24. **Heating y HeatingQC**: Tipo de calefacción y su calidad, factores que afectan el confort y la eficiencia energética del hogar.
25. **CentralAir**: Indica si la propiedad cuenta con aire acondicionado central, una característica deseable en muchas áreas.
26. **Electrical**: Tipo de sistema eléctrico instalado en la vivienda, relevante para la seguridad y la modernización.
27. **1stFlrSF y 2ndFlrSF**: Superficie de la primera y segunda planta, respectivamente, que afectan la distribución del espacio y la funcionalidad.
28. **LowQualFinSF y GrLivArea**: Metros cuadrados acabados de baja calidad y área habitable sobre el suelo, utilizados para evaluar la calidad y cantidad del espacio habitable.
29. **BsmtFullBath, BsmtHalfBath, FullBath, HalfBath**: Número de baños completos y medios tanto en el sótano como sobre el suelo, características clave en la habitabilidad de la casa.
30. **Bedroom y Kitchen**: Número de dormitorios y cocinas por encima del suelo, elementos fundamentales en la funcionalidad y el valor de una vivienda.
31. **KitchenQual**: Calidad de la cocina, un factor importante en la evaluación del atractivo interior de la propiedad.
32. **TotRmsAbvGrd**: Total de habitaciones por encima del suelo, excluyendo baños, utilizado para describir el tamaño general de la vivienda.
33. **Functional**: Evalúa la funcionalidad del hogar, desde típico hasta condiciones de salvamento, lo cual puede influir en la habitabilidad y los costos de reparación.
34. **Fireplaces y FireplaceQu**: Número de chimeneas y su calidad, características que aportan tanto a la comodidad como al valor estético.
35. **GarageType, GarageYrBlt, GarageFinish, GarageCars, GarageArea, GarageQual, GarageCond**: Detalles del garaje, incluidos su tipo, tamaño, calidad y año de construcción, elementos que añaden valor y funcionalidad.
36. **PavedDrive**: Estado del camino de entrada, ya sea pavimentado o de grava, que afecta la estética y la accesibilidad.
37. **WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch**: Superficie de diferentes tipos de porches y terrazas, características que aumentan el espacio habitable y el atractivo exterior.
38. **PoolArea y PoolQC**: Área y calidad de la piscina, factores que pueden influir significativamente en el valor de una propiedad en mercados específicos.
39. **Fence**: Calidad de la cerca, relevante para la privacidad y la seguridad.
40. **MiscFeature y MiscVal**: Características misceláneas no cubiertas en otras categorías y su valor monetario, que pueden incluir elementos únicos como elevadores o canchas de tenis.
41. **MoSold y YrSold**: Mes y año de la venta, información útil para analizar tendencias de mercado y estacionalidad.
42. **SaleType y SaleCondition**: Tipo y condición de la venta, que pueden proporcionar contexto sobre el estado del mercado y la motivación del vendedor.

Estas variables proporcionan una imagen completa de las características físicas, el estado y el contexto de cada propiedad, permitiendo un análisis detallado para la valoración y comparación de viviendas.

## Preprocesamiento de Datos para el Modelado Predictivo

El preprocesamiento de datos es una etapa fundamental en cualquier proyecto de análisis de datos y modelado predictivo, ya que garantiza que los datos estén en la forma más adecuada para alimentar a los modelos de machine learning. En el contexto de un análisis de mercado inmobiliario, es vital transformar y escalar correctamente las variables del conjunto de datos para mejorar la precisión de las predicciones sobre el precio de venta de las viviendas.

**Definición y Ordenación de Variables Categóricas y Ordinales**

Para abordar correctamente las variables categóricas y ordinales presentes en el conjunto de datos, primero se establecen estructuras de datos específicas para estas variables. Se define una lista de nombres de variables categóricas (nominales), denominada nom\_var, que contiene todas las variables que no tienen un orden inherente pero que son relevantes para el análisis, como el tipo de vecindario o el estilo arquitectónico de la vivienda. El tratamiento adecuado de estas variables es crucial para captar relaciones no lineales que podrían influir en el precio de venta.

Además, se crea una lista de niveles de calidad ordinal (grade\_lvls) para variables que representan categorías con un orden natural, como la condición exterior o la calidad de los acabados interiores de la vivienda. Esta lista puede incluir niveles como "Po" (Poor), "Fa" (Fair), "TA" (Typical/Average), "Gd" (Good), y "Ex" (Excellent), que permiten representar el estado de la propiedad de manera jerárquica.

Para variables ordinales con valores numéricos explícitos, se utiliza una lista de niveles numéricos del 1 al 10 (number\_lvls). Este enfoque permite mantener el orden natural de estas variables durante el análisis y el modelado.

Un diccionario llamado ordered\_levels se utiliza para definir el orden de los niveles de diferentes variables ordinales clave, como OverallQual (calidad general de la vivienda), OverallCond (condición general de la vivienda), ExterQual (calidad exterior), entre otros. La correcta ordenación de estas variables es fundamental, ya que permite al modelo interpretar correctamente las diferencias en calidad o condición. Además, se añade un nivel adicional, "None", al principio de cada lista de niveles para manejar casos donde pueda faltar información, asegurando así que el modelo no se vea afectado por datos ausentes.

## Análisis visual y conclusiones observadas sobre los datos

**Matriz de correlación**

Esta matriz permite identificar rápidamente qué características de una propiedad, como el tamaño total del área habitable o la calidad de la construcción, tienen un mayor impacto en el precio de venta. Esto ayuda a priorizar las variables que deberían recibir más atención durante el modelado y en la toma de decisiones de inversión, como focalizar esfuerzos en propiedades con características que maximicen el retorno sobre la inversión.

Ver [Anexo A](#_Anexo_A:_Matriz) para un análisis más detallado.

### Análisis boxplots

Para cada variable categórica relevante, se generan **boxplots** que muestran la distribución del precio de venta (SalePrice) para cada categoría. Esta visualización es útil para entender cómo distintas categorías dentro de una variable, como el tipo de vecindario o el estilo arquitectónico, afectan el precio de venta. Los boxplots permiten visualizar la mediana, los cuartiles y los posibles valores atípicos, lo que proporciona una visión detallada de cómo varía el precio de las propiedades dentro de diferentes categorías.

En el análisis del precio de venta de las propiedades, los datos revelan variaciones significativas según diferentes características y vecindarios. Primero, el vecindario **NridgHt** destaca con la mediana más alta de precios de venta, sugiriendo que es una zona percibida como más valiosa o demandada en comparación con otras áreas. De manera similar, **NoRidge** presenta los valores atípicos más altos en precios de venta y la segunda mediana más alta, lo que indica que es un barrio atractivo para compradores con alto poder adquisitivo.

En cuanto a los tipos de construcción, **1Fam** (casas unifamiliares) y **Twnhs** (townhouses) muestran medias de precios de venta más altas, siendo **1Fam** notable por su amplia variabilidad en precios debido a numerosos valores atípicos elevados. Esto sugiere que las casas unifamiliares tienen el potencial de generar ingresos superiores en ciertas condiciones del mercado.

El análisis de los estilos de vivienda, como **1.5Fin**, **1Story**, y **2Story**, revela una mayor cantidad de valores atípicos altos, lo que implica una significativa variabilidad en los precios de venta dentro de estos estilos. En contraste, **SFoyer** presenta valores atípicos bajos, lo que podría indicar que este estilo de vivienda tiende a tener precios de venta más bajos en algunos casos.

En el aspecto de zonificación, la categoría **RL** (Residential Low Density) muestra una concentración de valores atípicos altos, indicando que existen propiedades con precios de venta excepcionalmente altos dentro de esta zona residencial. Esto podría reflejar características especiales que elevan el valor de estas propiedades.

Con respecto a la fundación, **PConc** (Poured Concrete) destaca con la mediana más alta y numerosos valores atípicos altos, sugiriendo que las propiedades con esta fundación tienden a tener precios de venta superiores. **CBlock** (Cinder Block) también presenta una mediana elevada y valores atípicos altos, aunque con menor intensidad. En contraste, **Stone** (Stone) tiene una amplia caja de precios, indicando una gran variabilidad en los precios de venta.

En el tipo de garaje, las propiedades con **CarPort** y **None** generalmente tienen los precios de venta más bajos, mientras que las propiedades con garajes **BuiltIn** y **Attchd** (Attached) muestran medianas más altas y una mayor cantidad de valores atípicos elevados. Esto sugiere que los garajes adjuntos están asociados con precios más altos y mayor variabilidad en los precios.

Finalmente, el análisis de la condición de venta muestra que **Partial** tiene la caja y la mediana más grandes, lo que indica precios de venta más altos y mayor variabilidad en esta categoría. Sin embargo, **Normal** presenta la mayor cantidad de valores atípicos, reflejando que, a pesar de tener precios de venta altos en promedio, también existen propiedades con precios excepcionalmente elevados.

En términos de calidad general (**OverallQual**), el nivel 10 tiene la media y la caja más grandes, destacando precios de venta más altos y una amplia variabilidad. El nivel 5 muestra numerosos valores atípicos tanto inferiores como superiores, indicando una considerable dispersión en los precios para propiedades con esta calificación. Los niveles 6 y 7 también presentan una cantidad significativa de valores atípicos altos, sugiriendo una gama amplia de precios en estas categorías de calidad.

En resumen, estos análisis destacan cómo las características de las propiedades y sus ubicaciones impactan significativamente en los precios de venta, proporcionando una visión integral de los factores que contribuyen a las variaciones en el mercado inmobiliario.

Ver [Anexo B](#_Anexo_B:_Bloxplots) para los gráficos de los boxplots.

### Análisis histogramas

El análisis de histogramas proporciona una visión profunda sobre la distribución de varias características clave en el conjunto de datos. En el caso de GarageArea (Área del Garaje), el histograma revela que los valores más altos se concentran en el rango de aproximadamente 120 a 160 pies cuadrados, mientras que el resto de los valores se sitúan por debajo de 80 pies cuadrados. Esto sugiere que, aunque hay algunas propiedades con garajes de tamaño considerable, la mayoría de los garajes son relativamente pequeños, con una tendencia a concentrarse en el rango inferior de la distribución.

Para GrLivArea (Área de Vivienda en Pisos Superiores), el histograma muestra una notable concentración de datos en el rango de 1000 a 2000 pies cuadrados, indicando que la mayoría de las propiedades tienen un área de vivienda en este intervalo. Este patrón puede reflejar un estándar común de tamaño para las viviendas en el mercado representado por el conjunto de datos.

En el análisis de YearBuilt (Año de Construcción), el histograma destaca que el año 2000 tiene la barra más alta, lo que indica que muchas propiedades fueron construidas en ese período. Esto puede sugerir que el mercado en ese año estuvo particularmente activo o que se han construido muchas viviendas nuevas en torno a esa fecha.

Para YearRemodAdd (Año de Renovación), el histograma muestra que tanto el primer año como el penúltimo año tienen las barras más altas, sugiriendo que se realizaron renovaciones importantes en esos años específicos. Esto podría reflejar picos en la actividad de remodelación durante esos períodos.

El histograma de TotalBsmtSF (Área del Sótano Total) presenta una distribución en forma de montaña, con una barra muy alta cerca de 1000 pies cuadrados, lo que indica que muchas propiedades tienen sótanos de tamaño en este rango. Los tamaños significativamente mayores o menores son menos frecuentes, sugiriendo una concentración en el tamaño de sótano común.

Para GarageCars (Número de Coches en el Garaje), el histograma indica que la opción más común es tener espacio para dos coches, con aproximadamente 800 propiedades en esta categoría. Esto refleja que tener capacidad para dos coches es un estándar frecuente en el mercado.

Finalmente, el histograma de LotArea (Área del Lote) muestra una mayor concentración de datos en el rango inferior, con valores que no superan los 50,000 pies cuadrados. Esto indica que la mayoría de los lotes son de tamaño moderado, con una frecuencia notablemente menor a medida que el tamaño del lote aumenta.

El histograma de MasVnrArea (Área de Revestimiento de Masilla) destaca que la mayoría de las propiedades tienen un área de revestimiento de masilla relativamente pequeña, con la barra correspondiente al área más pequeña siendo la más alta. Las áreas más grandes son menos frecuentes, indicando que pocos lotes tienen grandes áreas de revestimiento de masilla.

Ver [Anexo C](#_Anexo_C:_Histogramas) para los gráficos de los histogramas

### Análisis scatter plots

En el análisis de los gráficos de dispersión entre el **precio de venta** y otras variables, se observan patrones significativos. En el gráfico que compara el precio de venta con el **área de vivienda en pisos superiores (GrLivArea)**, se nota que la dispersión de precios aumenta considerablemente a partir de un área de 2,000 pies cuadrados. Esto indica que, para viviendas más grandes, los precios tienden a variar más ampliamente, mostrando una mayor variabilidad en los precios a medida que el tamaño de la vivienda crece. En contraste, antes de alcanzar los 2,000 pies cuadrados, la dispersión es más contenida, con menos variabilidad en los precios para tamaños de vivienda menores.

En el gráfico que compara el precio de venta con el **área del lote (LotArea)**, se observa un incremento significativo en el precio con respecto a un pequeño aumento en el tamaño del lote. Esto sugiere que incluso un pequeño aumento en el área del lote puede resultar en un aumento notable en el precio de venta. Sin embargo, la dispersión de precios en relación con el área del lote es menor en comparación con la dispersión observada en el área de vivienda, pero más abrupta. Esto significa que, aunque hay menos variabilidad en los precios para lotes de tamaño similar, los precios pueden ser considerablemente altos incluso para áreas de lote cercanas, indicando que el valor del terreno puede ser considerablemente alto en comparación con el tamaño del lote.

Ver [Anexo D](#_Anexo_D:_Scatter) para los gráficos de los scatter plots

## Mapa de calor

El mapa de calor de precios de venta (**SalePrice**) proporciona una visualización detallada de la distribución espacial de los valores de las propiedades en el área analizada. Este análisis revela patrones significativos en la concentración de viviendas de alto y bajo valor. Áreas con una alta densidad de precios elevados indican vecindarios más exclusivos y deseables, probablemente debido a factores como mejores infraestructuras, servicios o características socioeconómicas favorables. En contraste, zonas con menor densidad de precios altos sugieren vecindarios con menor demanda o atributos menos atractivos para los compradores.

Ver [Anexo E](#_Anexo_E:_Mapa) para observar el mapa de calor

# Creación del pipeline de procesado de datos

## Normalización de variables

Este proceso de normalización es una evolución del preprocesamiento inicial de limpieza de datos, al incorporar nuevas etapas que mejoran significativamente la preparación del conjunto de datos para el modelado predictivo. Mientras que la fase inicial se centraba principalmente en la limpieza básica, como el manejo de valores faltantes y la eliminación de valores atípicos, esta etapa avanzada amplía las transformaciones para incluir la estandarización, imputación y codificación detallada de variables, con el objetivo de garantizar la integridad y consistencia de los datos.

En esta etapa avanzada, se utiliza un ColumnTransformer que aplica transformaciones específicas basadas en el tipo de variable: las columnas numéricas se estandarizan con StandardScaler para que tengan una media de cero y una desviación estándar de uno, asegurando que los datos numéricos sean comparables entre sí; las variables categóricas nominales se codifican con OneHotEncoder para convertirlas en una representación binaria sin introducir colinealidad; y las variables categóricas ordinales se transforman con un LabelEncodingTransformer para reflejar su orden inherente en los modelos de forma adecuada.

Además, se ha desarrollado un transformador personalizado, ColumnNameTransformer, que maneja de manera dinámica los nombres de las columnas resultantes después de la transformación, facilitando la comprensión y el análisis de los datos procesados. Este enfoque no solo perfecciona el proceso de limpieza inicial, sino que también mejora la coherencia, calidad y preparación de los datos, estableciendo así una base sólida para la construcción de modelos predictivos más precisos y robustos.

## PCA

En esta fase, se aplica el Análisis de Componentes Principales (PCA) al conjunto de datos estandarizado para reducir la dimensionalidad y comprender la varianza explicada por cada componente principal. Este paso sigue la evolución del preprocesamiento de datos, que previamente incluyó la limpieza, codificación y estandarización de variables.

Primero, se ajusta el PCA al conjunto de datos estandarizado X\_scaled para calcular la varianza explicada por cada componente principal. Un Scree Plot ([Ver Anexo F](#_Anexo_F:_Scree)) se utiliza para visualizar cómo la varianza se distribuye entre los componentes, ayudando a identificar cuántos de ellos son necesarios para captar la mayor parte de la variabilidad en los datos. Además, se calcula la varianza acumulada y se crea un gráfico que muestra el número de componentes necesarios para alcanzar un umbral del 95% de varianza explicada. [Ver Anexo G](#_Anexo_G:_Varianza)

Para facilitar la interpretación de los componentes principales, se define un transformador personalizado, PCAToDataFrameTransformer, que aplica PCA y transforma los datos en un DataFrame con nombres de columnas claros para cada componente. Este proceso es una evolución del preprocesamiento inicial que se centró en la limpieza y estandarización, añadiendo ahora una reducción efectiva de la dimensionalidad. De las 80 variables originales, se obtienen 56 componentes principales que aseguran un 95% de la varianza explicada.

## Pipeline final

El pipeline final es una evolución natural y consolidada de los enfoques de preprocesamiento previos, integrando y refinando las etapas para una preparación de datos más robusta y completa. En las fases iniciales, los pipelines se centraron principalmente en la limpieza básica de datos, como el manejo de valores faltantes y la eliminación de valores atípicos, estableciendo así una base sólida. Con el tiempo, se incorporaron etapas adicionales, como la estandarización de variables, utilizando StandardScaler para normalizar los datos numéricos, y técnicas de codificación avanzadas para adaptar las variables categóricas a los requerimientos de los modelos predictivos.

Este pipeline avanzado integra de manera eficiente todos estos pasos: primero, aplica CleanTransformer para asegurar una limpieza exhaustiva, seguido por EncodeTransformer para gestionar la codificación de variables categóricas y ordinales, y ImputeTransformer para manejar valores faltantes. Posteriormente, utiliza un ColumnTransformer para aplicar transformaciones específicas basadas en el tipo de variable: estandarización de columnas numéricas, codificación one-hot para variables categóricas nominales, y codificación de etiquetas para variables categóricas ordinales. Además, se incluye un transformador personalizado, ColumnNameTransformer, que facilita la comprensión de los datos al manejar dinámicamente los nombres de las columnas resultantes.

Una innovación clave en este pipeline es la incorporación del PCAToDataFrameTransformer, que aplica el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos. Este transformador convierte los datos estandarizados en un DataFrame con componentes principales claramente etiquetados, facilitando una representación compacta y eficaz de la información. A partir de las 80 variables originales, se obtienen 56 componentes principales que explican el 95% de la varianza en los datos.

El pipeline final optimiza la preparación de datos y facilita la transición hacia el modelado predictivo, consolidando y perfeccionando las técnicas desarrolladas en los pipelines anteriores para lograr una preparación de datos más completa y eficiente.

# Creación y evaluación de los modelos

En la fase de creación y evaluación de modelos, se han utilizado tres enfoques distintos: Random Forest, XGBoost y LightGBM, cada uno con sus parámetros específicos para abordar el problema de aprendizaje automático.

Para **Random Forest**, se configuró el modelo con parámetros estándar, incluyendo random\_state para garantizar la reproducibilidad. El modelo mostró un MSE de 0.303 y un R² de 0.704 en el conjunto de prueba. Estos resultados reflejan la robustez de Random Forest en manejar datos con alta dimensionalidad y su capacidad para generalizar, aunque la precisión es ligeramente inferior en comparación con los modelos de boosting.

**XGBoost** se entrenó con parámetros como n\_estimators, learning\_rate, y max\_depth ajustados para optimizar el rendimiento. El modelo alcanzó un MSE de 0.353 y un R² de 0.654 en el conjunto de prueba. La capacidad de XGBoost para manejar datos complejos y corregir errores de árboles anteriores se traduce en un rendimiento sólido, aunque requiere una cuidadosa sintonización de hiperparámetros para maximizar su precisión.

**LightGBM** se configuró con parámetros optimizados para manejar grandes volúmenes de datos, como num\_leaves, learning\_rate, y n\_estimators. Este modelo demostró una ventaja en velocidad y eficiencia, logrando un MSE de 0.300 y un R² de 0.707 en el conjunto de prueba. Los resultados de LightGBM confirman su capacidad para ofrecer una precisión superior y un rendimiento más rápido en comparación con XGBoost, especialmente en conjuntos de datos grandes.

La coherencia entre las características esperadas de cada modelo y los resultados obtenidos es evidente. Random Forest ofrece robustez y una buena interpretación de los datos, XGBoost proporciona precisión con un alto rendimiento en contextos complejos, y LightGBM destaca por su velocidad y manejo eficiente de grandes volúmenes de datos. Estos resultados alinean con las fortalezas inherentes de cada modelo, destacando cómo sus características afectan directamente su desempeño en la práctica.

# Procesos ETL y Base de datos

## Descripción de la Base de Datos "HouseSales"

La base de datos "HouseSales" está diseñada para almacenar y gestionar información detallada relacionada con la venta de propiedades inmobiliarias. A continuación, se expone una descripción exhaustiva de la estructura de las tablas principales, las relaciones entre ellas y cómo estas interacciones permiten realizar análisis comprensivos sobre las propiedades y sus características en el contexto del mercado inmobiliario.

### Estructura de las Tablas de Dimensiones

Las tablas de dimensiones en la base de datos "HouseSales" contienen datos cualitativos que describen aspectos específicos de las propiedades. Estas tablas proporcionan una descripción detallada de las características de las propiedades, permitiendo un análisis en profundidad de las diferencias en los precios de venta según estas características.

* **DimBasement**: Esta tabla almacena información sobre las características del sótano de cada propiedad, incluyendo la calidad del sótano (BsmtQual), su estado general (BsmtCond), el nivel de exposición (BsmtExposure), y los tipos de acabado interior (BsmtFinType1 y BsmtFinType2). La clave primaria de esta tabla es BasementID.
* **DimBuildingType**: Contiene información sobre el tipo de edificio (BldgType) y el estilo de la casa (HouseStyle), lo que permite clasificar las propiedades según su estructura y diseño arquitectónico. La clave primaria es BuildingTypeID.
* **DimExterior**: Almacena detalles sobre las características exteriores de las propiedades, incluyendo los materiales de las paredes exteriores (Exterior1st y Exterior2nd), el tipo de revestimiento de mampostería (MasVnrType), y las calidades exteriores (ExterQual y ExterCond). La clave primaria es ExteriorID.
* **DimFoundation**: Describe el tipo de fundación de cada propiedad (Foundation), con FoundationID como clave primaria.
* **DimLotCharacteristics**: Esta tabla detalla características del terreno, como la forma del lote (LotShape) y la configuración del mismo (LotConfig). La clave primaria es LotCharacteristicsID.
* **DimNeighborhood**: Contiene información sobre el vecindario donde se encuentra la propiedad (Neighborhood), así como las condiciones circundantes (Condition1 y Condition2). La clave primaria es NeighborhoodID.
* **DimPropertyType**: Proporciona información sobre el tipo de propiedad, incluyendo la subclase de la propiedad (MSSubClass) y la zonificación municipal (MSZoning). La clave primaria es PropertyTypeID.
* **DimRoof**: Esta tabla describe el estilo del techo (RoofStyle) y el material del techo (RoofMatl). La clave primaria es RoofID.
* **DimStreetAccess**: Contiene detalles sobre el acceso a la propiedad, como el tipo de calle que da acceso (Street), la presencia de callejones (Alley), el contorno del terreno (LandContour), y la pendiente del terreno (LandSlope). La clave primaria es StreetAccessID.
* **DimUtilities**: Almacena información sobre los servicios públicos disponibles (Utilities) y la funcionalidad general de la propiedad (Functional). La clave primaria es UtilitiesID.

### Estructura de las Tablas de Hechos

Las tablas de hechos registran eventos transaccionales y métricas cuantitativas. Estas tablas permiten realizar un seguimiento detallado de las características físicas de las propiedades y los eventos de venta.

* **FactLocation**: Esta tabla incluye datos geoespaciales de las propiedades, como latitud y longitud (Latitude, Longitude), así como la dirección exacta (Address). Es fundamental para el análisis espacial de las ventas de propiedades. La clave primaria de esta tabla es compuesta por LocationID y PropertyID.
* **FactProperty**: Contiene información detallada sobre cada propiedad, como la fecha de predicción de la venta (PredictionDate), el precio de venta previsto (PredictedSalePrice), la superficie total construida (GrLivArea), el número de baños y habitaciones, el área del garaje, entre otros. Cada registro en esta tabla representa una propiedad única con sus características específicas. La clave primaria es compuesta por PropertyID y Address.
* **FactSale**: Esta tabla registra las ventas de propiedades, incluyendo el precio de venta predicho (PredictedSalePrice) y la fecha de predicción de la venta (PredictionDate). Es crucial para el análisis histórico de ventas y la proyección de tendencias futuras. La clave primaria es compuesta por SaleID y PropertyID.

### Tabla Central de la Base de Datos

* **Property**: Esta tabla actúa como el eje central que conecta las demás tablas de dimensiones y hechos. Contiene información detallada de cada propiedad, incluyendo dirección, fecha de predicción, características físicas, y múltiples métricas cualitativas y cuantitativas que describen la propiedad. La clave primaria de esta tabla es PropertyID.

### Relaciones entre las Tablas

La estructura relacional de la base de datos "HouseSales" está diseñada para maximizar la eficiencia en la consulta de datos y en el análisis de la información sobre ventas de propiedades. Las tablas de hechos (**FactProperty**, **FactLocation**, y **FactSale**) están conectadas con las tablas de dimensiones a través de claves foráneas que corresponden a las claves primarias de las tablas de dimensiones. Por ejemplo, en la tabla **FactProperty**, las columnas como PropertyTypeID, NeighborhoodID, y StreetAccessID actúan como claves foráneas que referencian a las tablas de dimensión correspondientes, permitiendo una integración coherente y eficaz de la información.

### Análisis y Uso de los Datos

El diseño de esta base de datos permite realizar una variedad de análisis complejos relacionados con el mercado inmobiliario, tales como:

* **Análisis de Precios por Características de Propiedad**: Utilizando la tabla **FactProperty** junto con las tablas de dimensiones como **DimExterior** o **DimLotCharacteristics**, se pueden efectuar análisis detallados sobre cómo diversas características específicas influyen en el precio de venta de una propiedad.
* **Análisis Espacial**: A través de la tabla **FactLocation**, es posible realizar análisis espaciales para comprender cómo la ubicación afecta a los precios de las propiedades y a las tendencias de ventas.
* **Predicción y Tendencias de Ventas**: La combinación de datos en **FactSale** y **FactProperty** facilita la creación de modelos predictivos y el análisis de tendencias para prever precios de ventas futuras basados en datos históricos y características específicas de las propiedades.

## Ventajas de Utilizar Tablas de Dimensiones y Hechos

El uso de tablas de dimensiones (dims) y tablas de hechos (facts) en el diseño de la base de datos "HouseSales" proporciona una estructura altamente optimizada para el análisis de grandes volúmenes de datos. Las tablas de dimensiones contienen atributos descriptivos que permiten categorizar y segmentar los datos, facilitando un análisis detallado y granular. Por otro lado, las tablas de hechos registran eventos cuantitativos y métricas numéricas, lo que permite realizar análisis de tendencias y mediciones precisas. Esta separación entre datos descriptivos y transaccionales no solo mejora la eficiencia en la consulta y el almacenamiento de datos, sino que también proporciona una mayor flexibilidad para realizar análisis complejos, integrando múltiples perspectivas y dimensiones de la información disponible.

Ver [Anexo E](#_Anexo_E:_Diagrama) para observar el diagrama de la base de datos.

## Proceso ETL

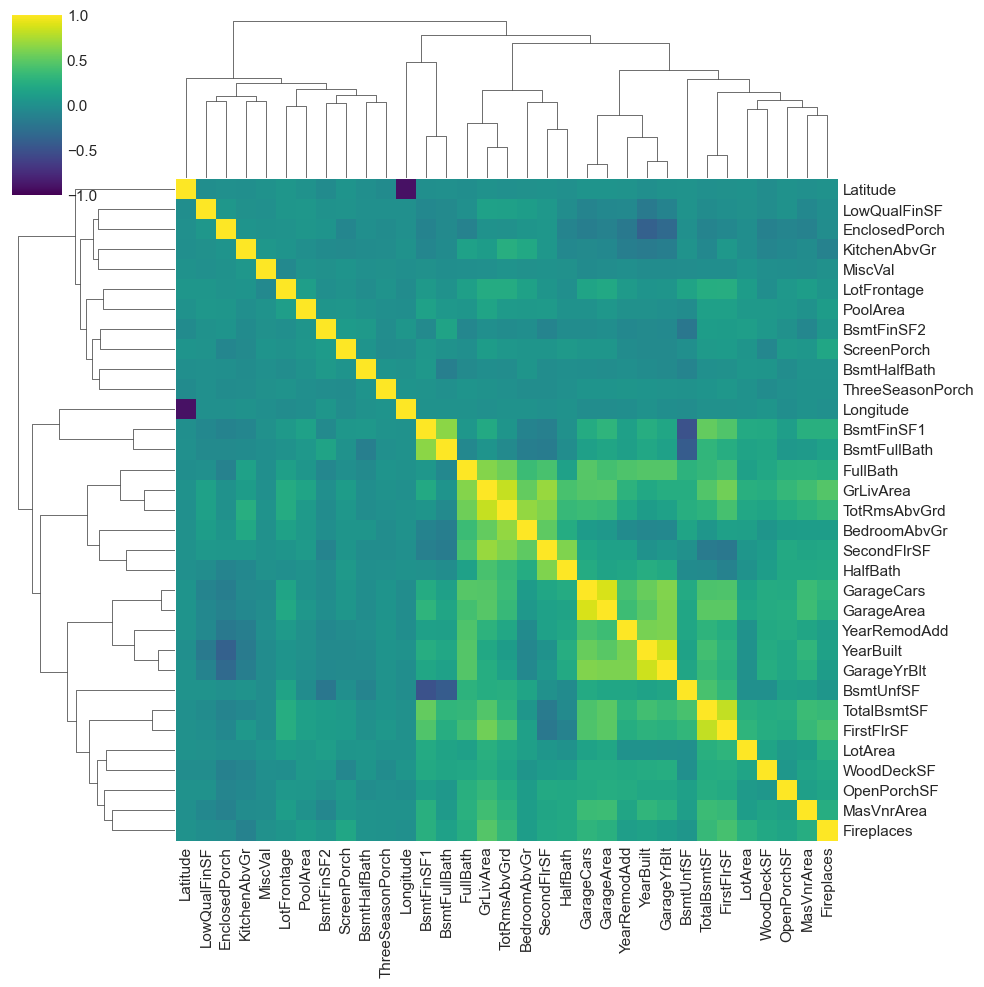
Para facilitar la carga y transformación de datos en la base de datos "HouseSales", se han implementado procesos utilizando SQL Server Integration Services (SSIS). SSIS es una tecnología de Microsoft que forma parte de SQL Server y está diseñada para realizar tareas de extracción, transformación y carga (ETL) de datos. Mediante SSIS, se crean flujos de trabajo que permiten extraer datos desde diversas fuentes, transformarlos de acuerdo con las necesidades del modelo de datos y, finalmente, cargarlos en las tablas correspondientes de la base de datos. En el caso de "HouseSales", los procesos SSIS incluyen la importación de datos desde archivos CSV y otras fuentes externas, la limpieza y normalización de datos para asegurar su calidad y consistencia, y la carga de estos datos en las tablas de hechos y dimensiones. La capacidad de SSIS para automatizar y optimizar estos procesos asegura una integración eficiente y efectiva de grandes volúmenes de datos, facilitando así un análisis más preciso y oportuno.

En el proceso de carga de datos para la base de datos "HouseSales", se sigue un orden específico para asegurar la integridad y consistencia de los datos. Primero, se carga la tabla Property, que contiene información básica sobre las propiedades inmobiliarias, como la dirección, el tipo de propiedad y el número de habitaciones. Esta tabla es fundamental ya que proporciona los datos base que se utilizarán en las tablas de dimensiones. Una vez que la tabla Property está completa, se procede a cargar las tablas de dimensiones (Dim\_), que incluyen detalles adicionales que describen aspectos específicos de las propiedades, como las características de los barrios y los tipos de propiedades. Finalmente, se cargan las tablas de hechos (Fact\_), que contienen las transacciones de ventas y otras métricas relacionadas con las propiedades. Este orden asegura que todas las relaciones entre las tablas estén correctamente establecidas y que las tablas de hechos puedan referirse a las dimensiones ya cargadas, garantizando así la integridad referencial y la precisión de los datos en el análisis.

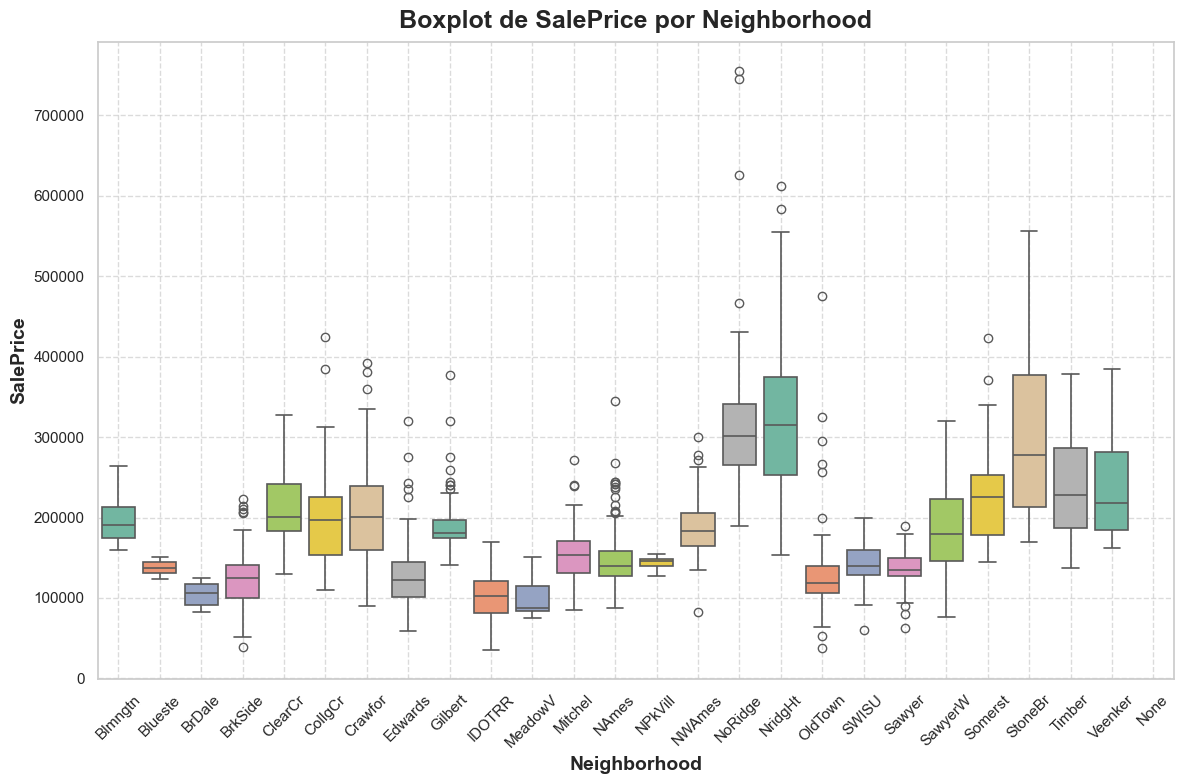
La carga de la tabla Property se realiza a partir de un archivo Excel generado mediante un script en Python. Este script aplica el modelo predictivo generado anteriormente, consultar el capítulo ‘[Creación y evaluación de modelos’](#_Creación_y_evaluación), para calcular el precio estimado de venta de las propiedades. El archivo Excel resultante incluye tanto la información básica de cada vivienda, como la dirección, tipo de propiedad y número de habitaciones, así como la predicción del precio de venta. Esta metodología no solo automatiza el proceso de generación de datos, sino que también asegura que la información sobre las propiedades esté actualizada y sea precisa antes de su carga en la base de datos. Una vez generado el archivo Excel, los datos se importan a la tabla Property utilizando un proceso ETL en SSIS, estableciendo así la base para las posteriores cargas de las tablas de dimensiones y hechos. Para asegurar un proceso continuo y sin interrupciones, se ha creado un job que ejecuta automáticamente el script en Python a intervalos regulares, gestionando la lectura de los datos, la generación de la predicción y la creación del archivo Excel, facilitando así la actualización continua de la base de datos.

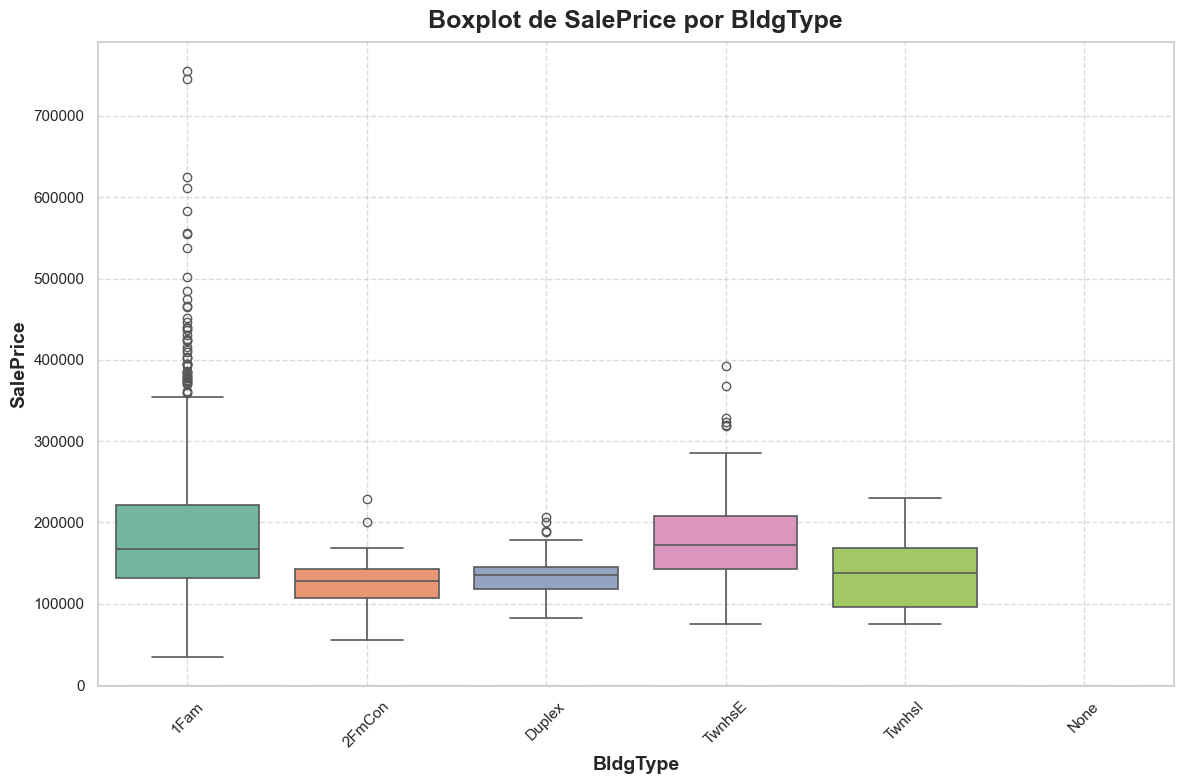
# Anexos

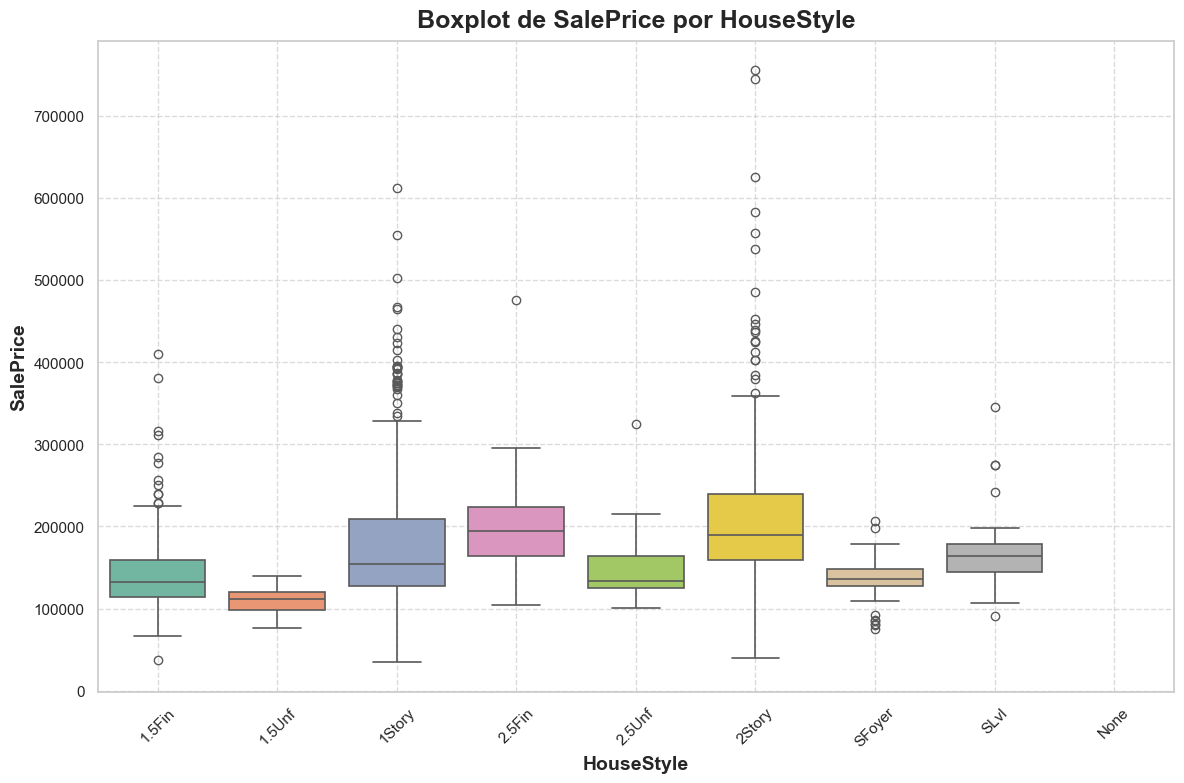
## Anexo A: Matriz de Correlación de Variables

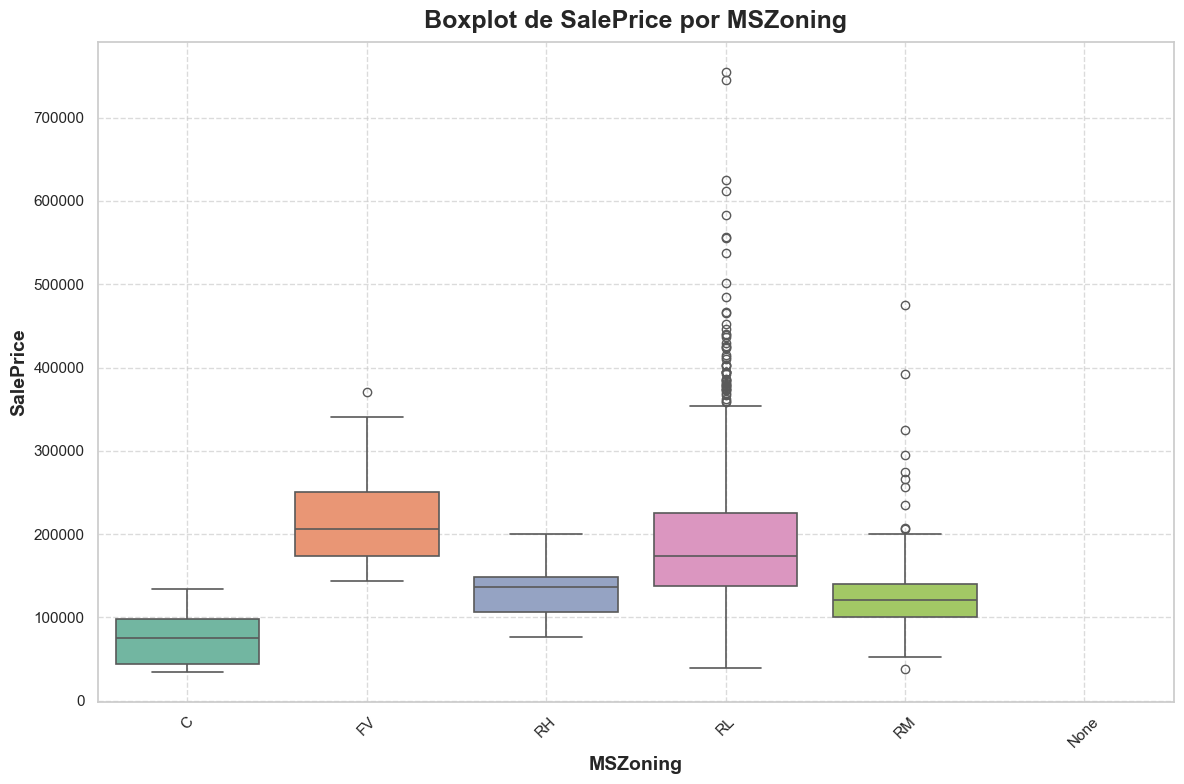


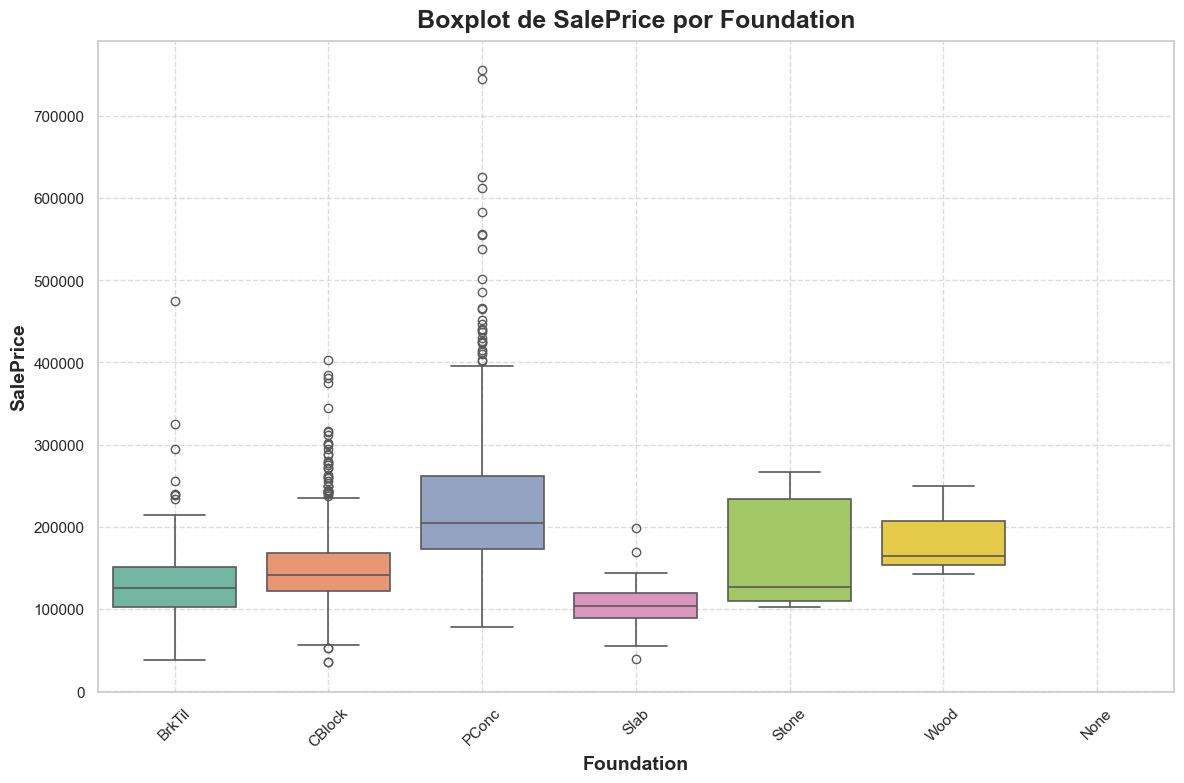
## Anexo B: Bloxplots respecto a SalePrice

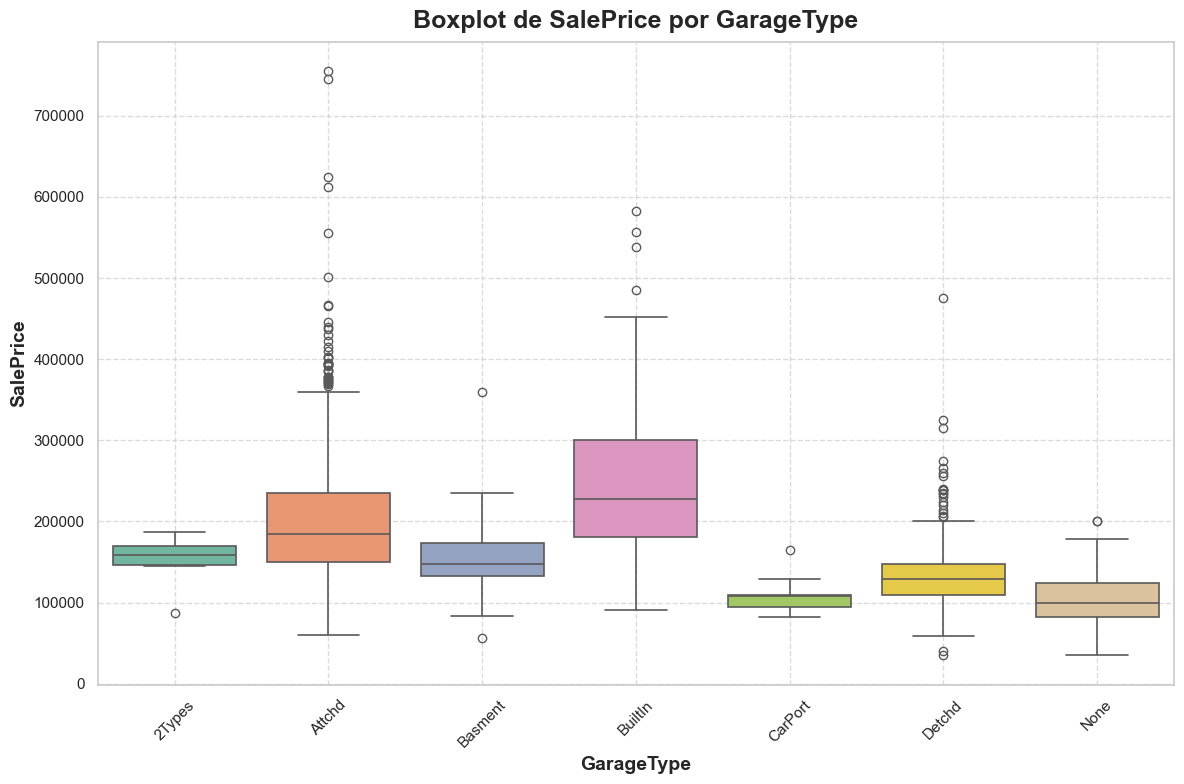


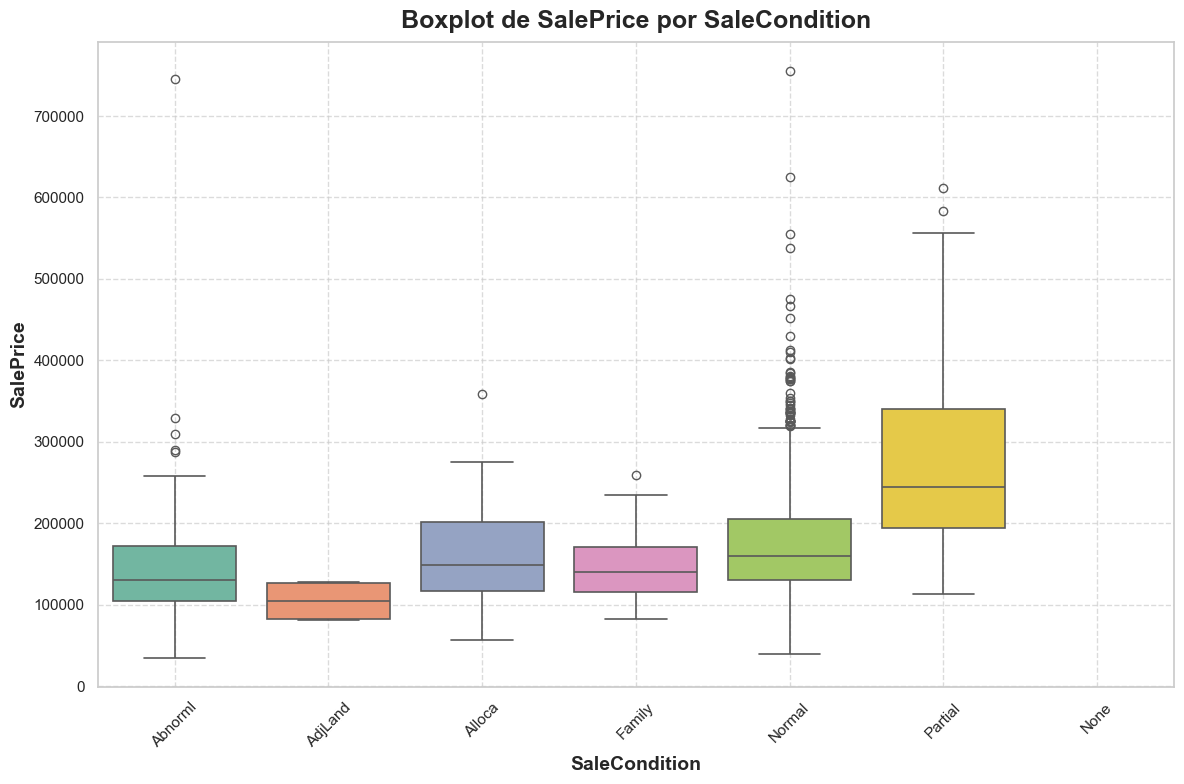


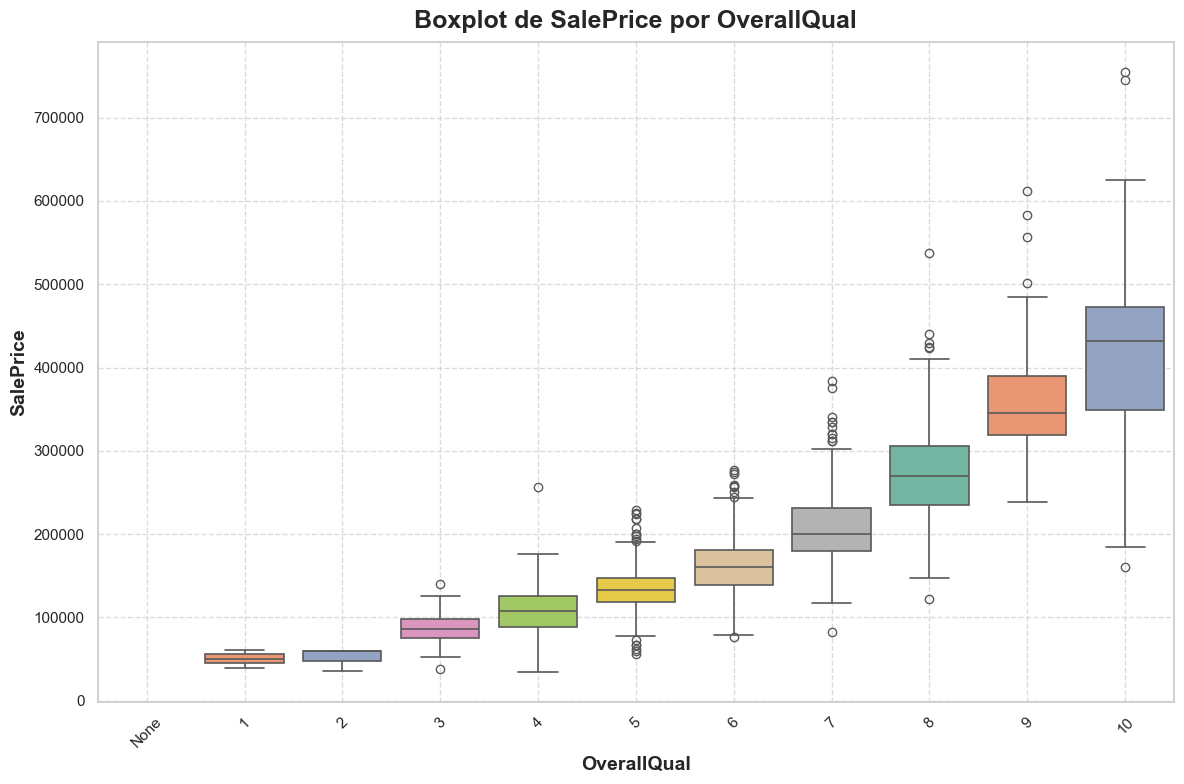




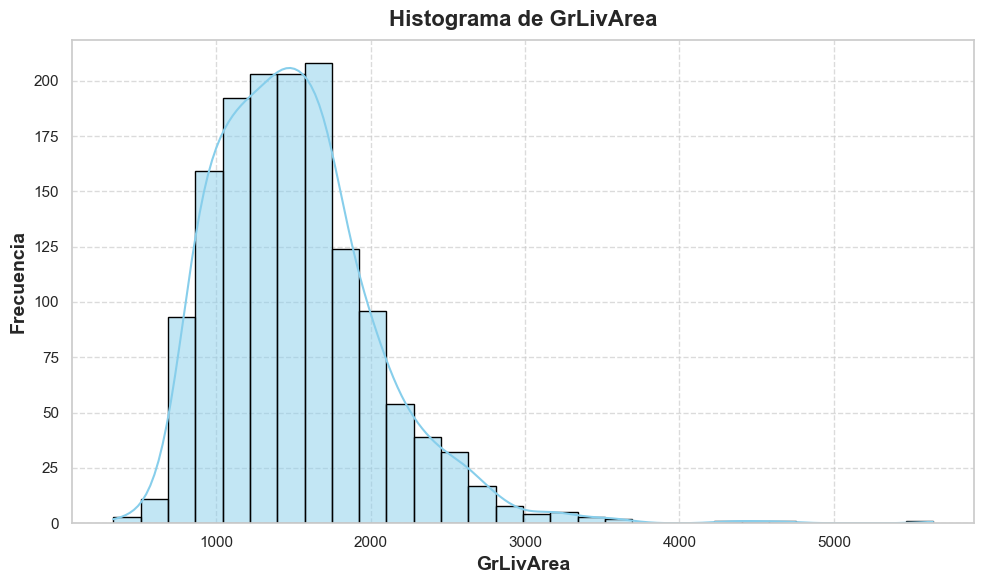


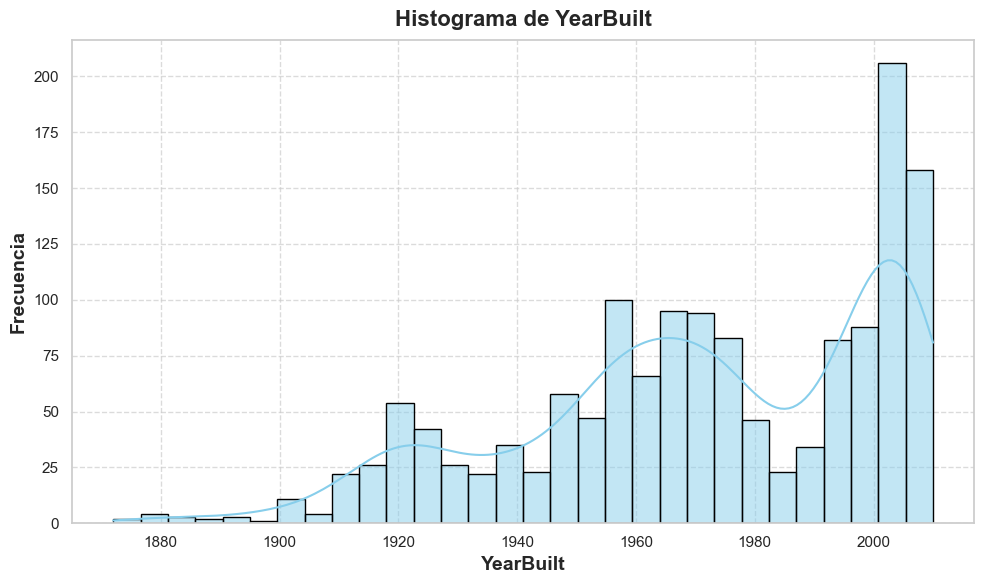


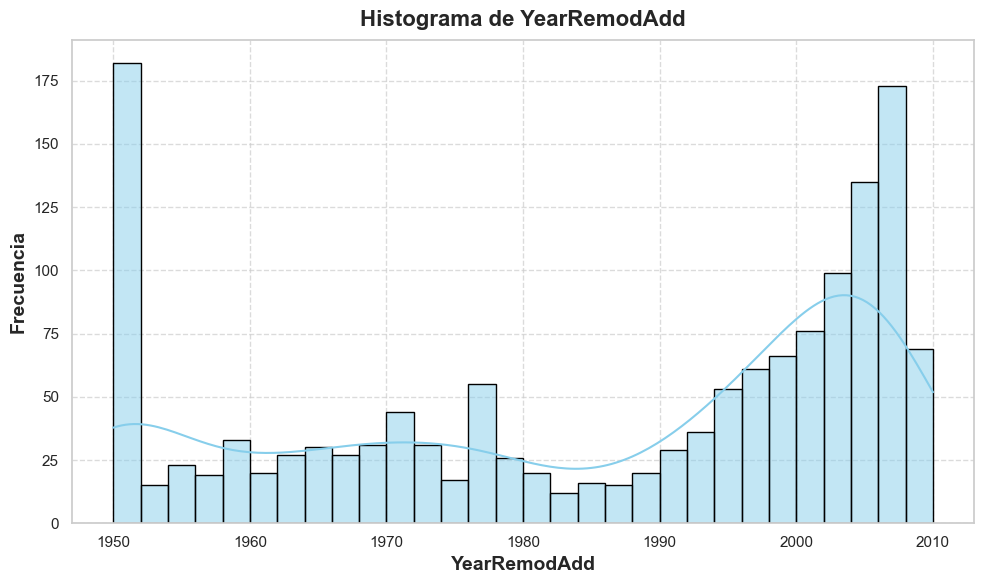


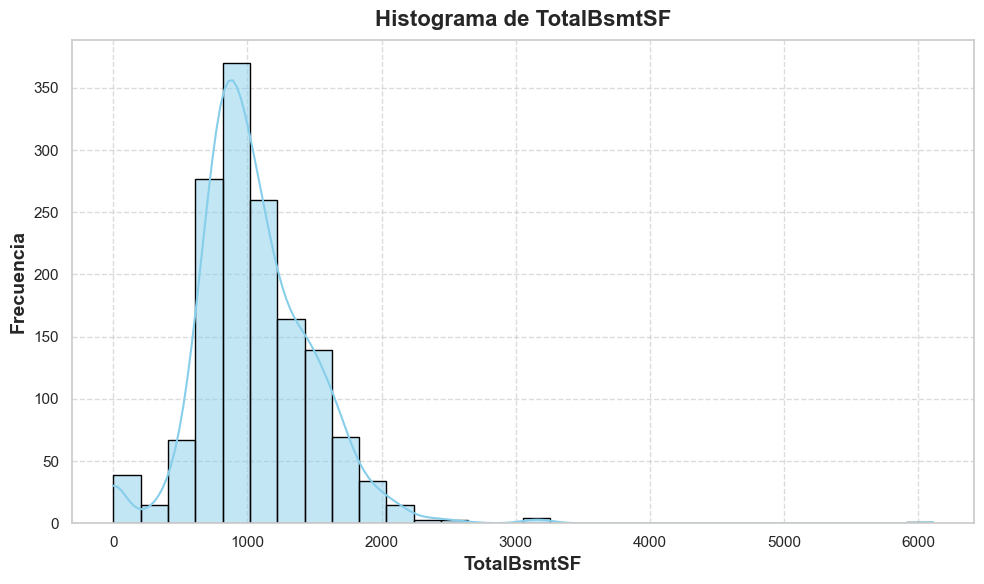


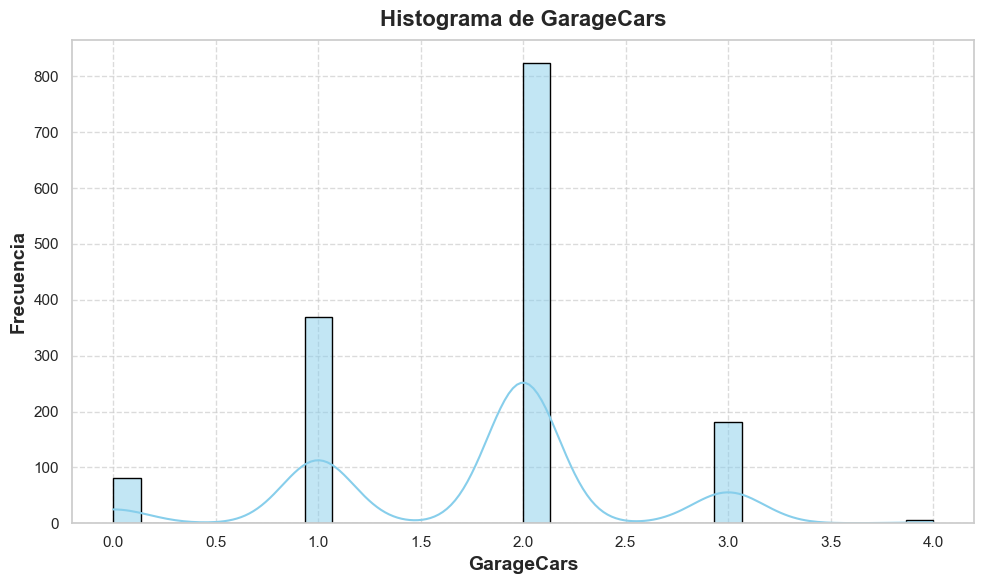
## Anexo C: Histogramas

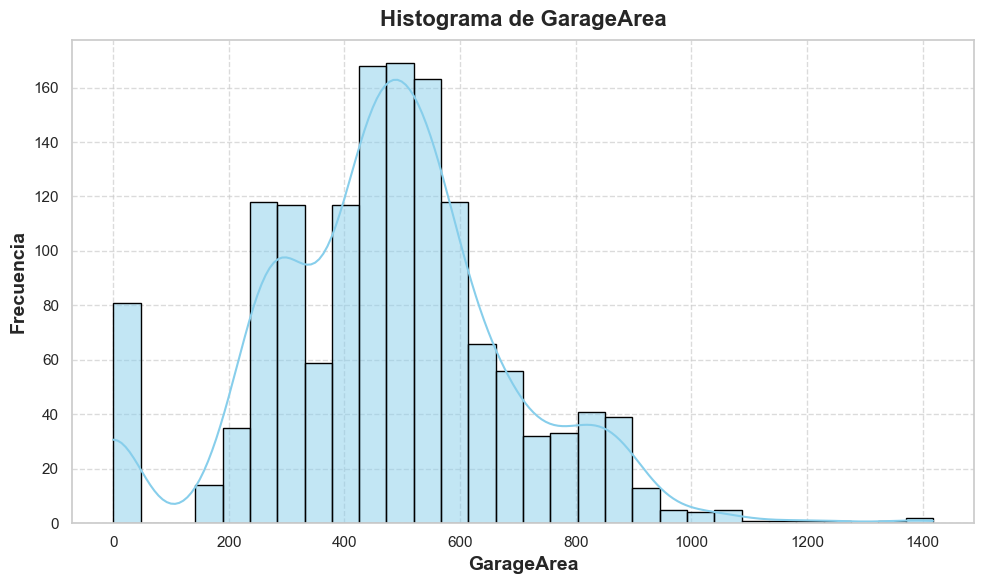


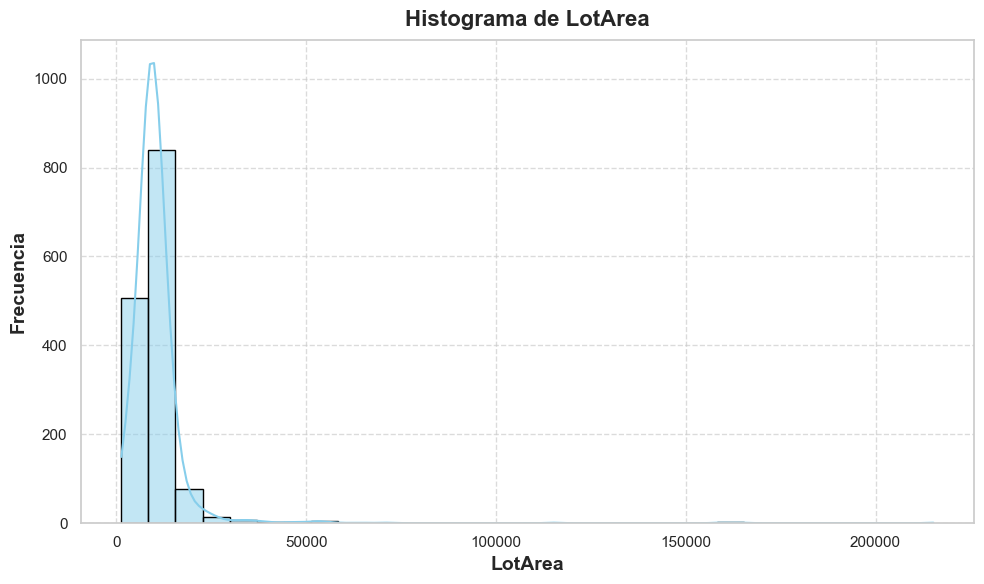


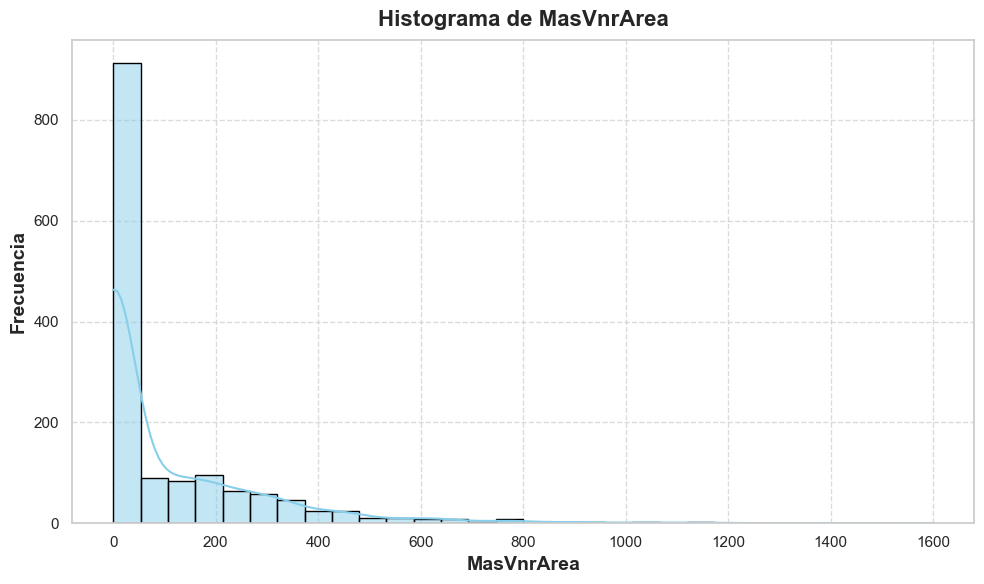




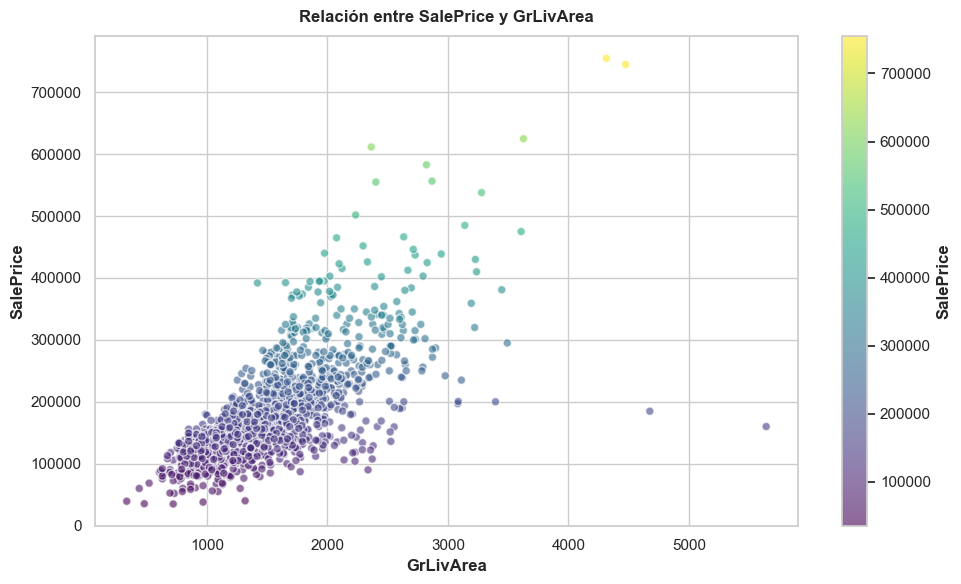


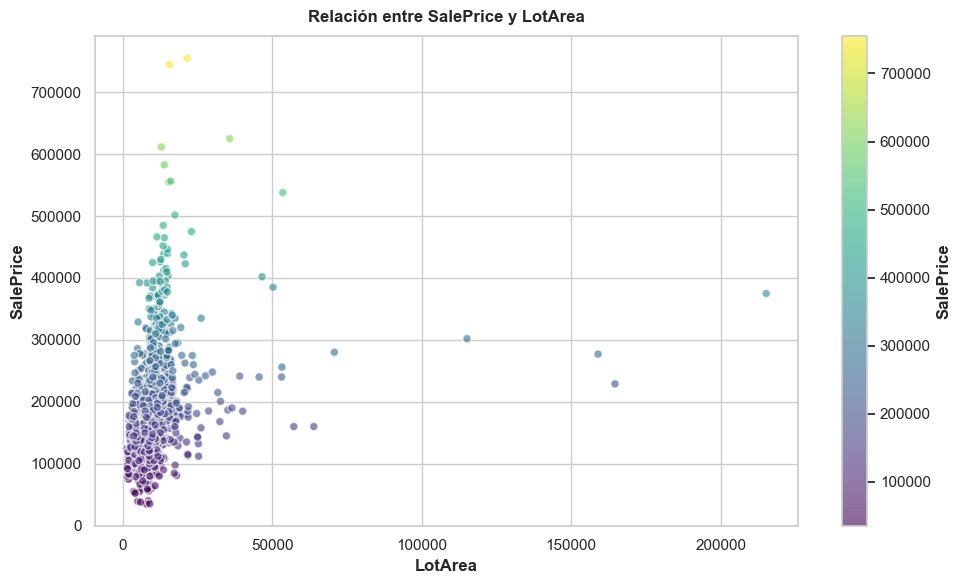


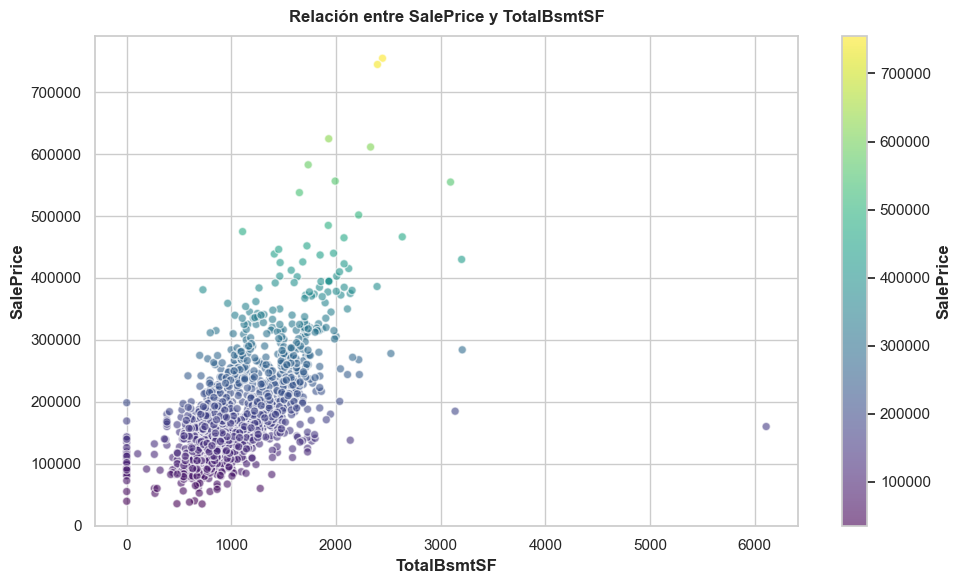


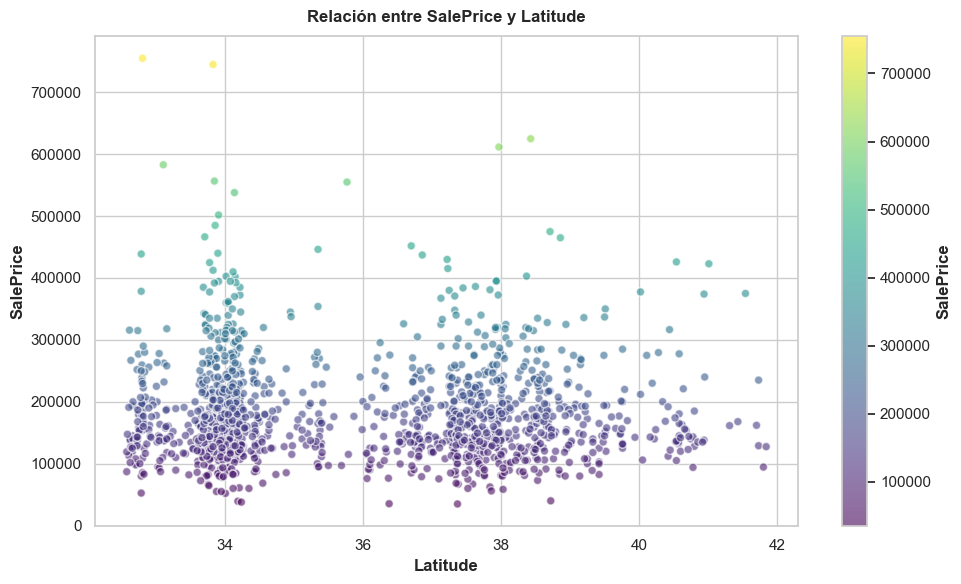


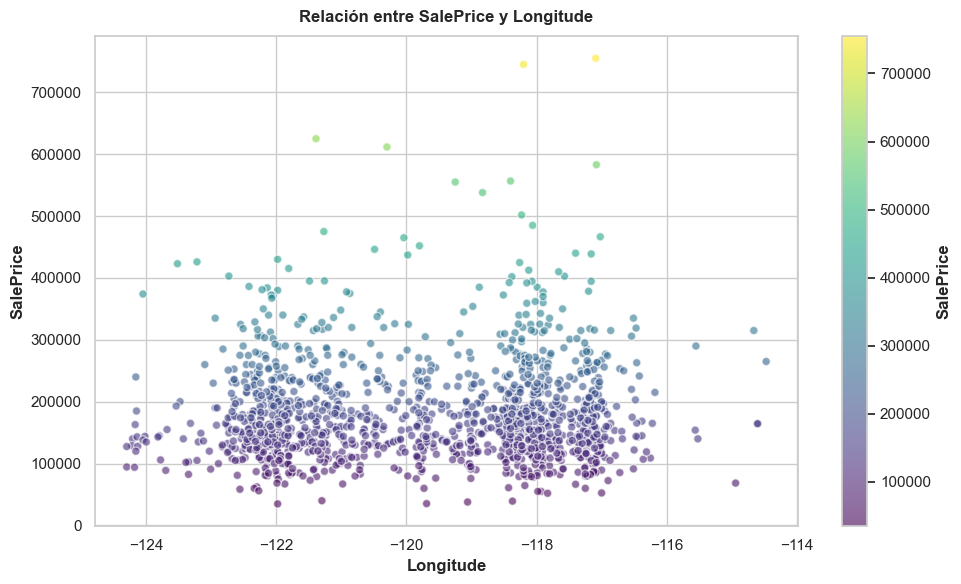
## Anexo D: Scatter plots



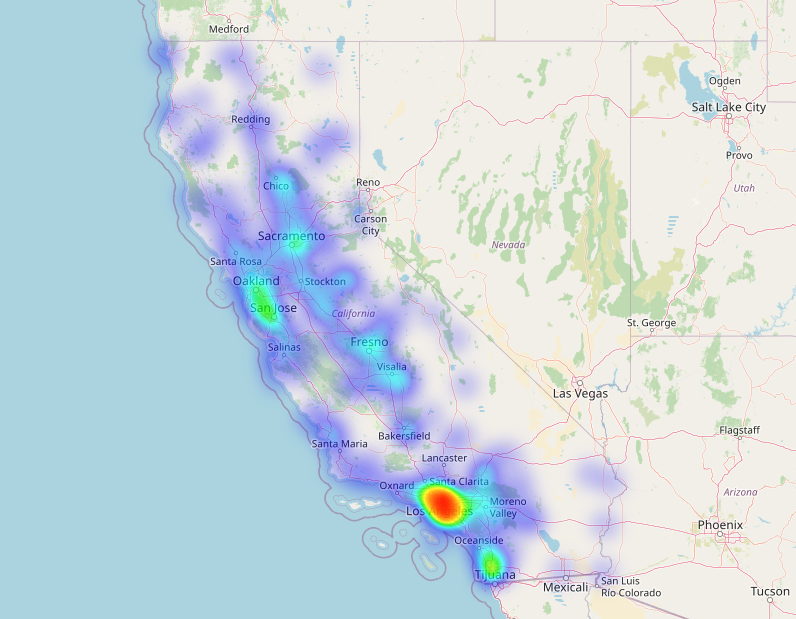






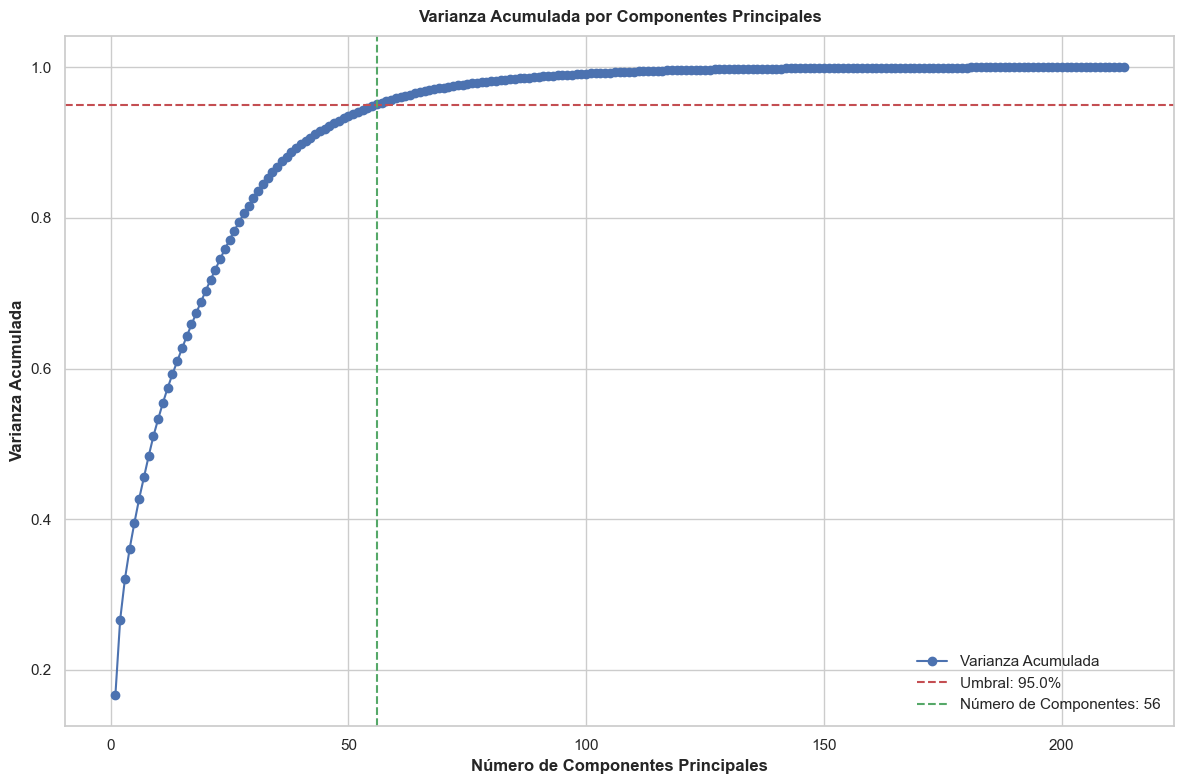


## Anexo E: Mapa de calor



## Anexo F: Scree Plot

## Anexo G: Varianza acumulada por componentes principales



## Anexo E: Diagrama Base de datos

