PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL Y DE SISTEMAS
ICS2122 - TALLER DE INVESTIGACIÓN OPERATIVA (CAPSTONE)

TASACIÓN DE VIVIENDAS Y DISEÑO DE LA CASA ÓPTIMA

GRUPO 05

INTEGRANTES:

JOSEFA ABETT DE LA TORRE
IGNACIO CUEVAS
VALENTINA DÍAZ
ARANTXA SALAS
ROCÍO TOLEDO
SEBASTIÁN VALDÉS
ANTONIA ZUMARÁN

PROFESOR GUÍA: MATÍAS DE GEYTER
AYUDANTE TUTOR: ALBERTO URETA

Santiago de Chile, Septiembre 2025

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

RESUMEN

La tasación de una vivienda constituye un proceso complejo debido a la gran cantidad de factores que inciden en el valor de un inmueble, tales como características estructurales, ubicación, entorno social, entre otros. Los métodos tradicionales de valoración basados en gran parte de juicios subjetivos suelen generar discrepancias significativas respecto al valor real de mercado, lo que provoca desconfianza e ineficiencia en el sector inmobiliario. A ello se suma la dificultad que enfrentan los propietarios al momento de decidir sobre el diseño y la remodelación de una vivienda, ya que deben hacerlo bajo restricciones presupuestarias y con el desafío de optimizar la inversión para lograr mayor rentabilidad. Estas problemáticas repercuten directamente en el funcionamiento del mercado inmobiliario y en la experiencia tanto de compradores como de vendedores.

La relevancia de contar con una tasación precisa y con una adecuada toma de decisiones en materia de diseño y remodelación de viviendas radica en que ello permite a compradores y vendedores actuar de manera más justa e informada, evitando tanto la sobrevaloración como la subvaloración de las propiedades. De este modo, se promueve un mercado inmobiliario más transparente y eficiente, en el cual las inversiones reflejan de mejor forma el valor real de los inmuebles. En el caso particular de la ciudad de Ames, Iowa, el desafío está presente por sus características demográficas y sociales, dado la fuerte presencia de población joven y la influencia de la universidad local generan dinámicas de alta demanda habitacional.

En este contexto, surge la necesidad de contar con instrumentos precisos que permitan estimar el precio de las viviendas con mayor exactitud y, a la vez, recomendar decisiones de diseño y remodelación que respondan a criterios de eficiencia y rentabilidad. Es por ello por lo que en el presente informe se plantea como objetivo diseñar un modelo de decisión para el mercado inmobiliario que estime el valor justo de una vivienda y proponga soluciones óptimas de diseño y remodelación orientadas a alcanzar una casa óptima. De esta forma, se

busca no solo corregir las limitaciones de las prácticas de tasación actuales, sino también aportarle un enfoque innovador para el diseño de viviendas eficientes y alineadas con las preferencias del mercado actual.

Para abordar este problema, se utiliza la base de datos *Ames Housing Dataset*, que recopila más de 2.900 transacciones de viviendas realizadas entre 2006 y 2010 en la ciudad de Ames, Iowa, e incluye 80 variables explicativas de tipo continuo, discreto y categórico. En una primera etapa, se realizó un análisis exploratorio de la base de datos empleando la matriz de correlación, de la cual se determinó la correlación entre variables. Luego, se determinó la correlación de variables con la variable *SalePrice* y en conjunto con el estudio anterior se determinaron las variables relevantes para el problema. Posteriormente, se construyó un modelo de regresión lineal multivariable para analizar el comportamiento y la significancia de las variables.

A continuación, se compararon tres metodologías principales de predicción: regresión lineal y sus derivados como *Lasso* y *Ridge*, *Random Forest* y *XGBoost*. El análisis concluyó que *XGBoost* presenta ventajas significativas, debido a que tiene documentación que respalda un buen desempeño en la base de datos a trabajar, tiene capacidad de escalabilidad y presenta aprendizaje iterativo, por lo que se selecciona como el método predictivo a emplear en este estudio. Finalmente, se plantean los pasos a trabajar en un futuro respecto al modelo y su rol operativo, y se determina incorporar herramientas de optimización, como *Gurobi*, con el fin de complementar el modelo predictivo y determinar configuraciones óptimas de diseño y remodelación bajo restricciones presupuestarias y técnicas.

1. INTRODUCCIÓN

El mercado inmobiliario es caracterizado por su complejidad y dinamismo, donde el darle un determinado valor a una vivienda representa un desafío central tanto para compradores como vendedores. Los métodos tradicionales de tasación, sustentados principalmente en referencias de ventas anteriores y criterios subjetivos, suelen generar gran discrepancia respecto al valor real del inmueble, lo que se traduce en ineficiencias, desconfianza y grandes pérdidas económicas para los agentes involucrados (Evans, Lausberg, & Sui Sang How, 2019). Este problema constituye un dolor estructural del sector, ya que afecta la transparencia del mercado y limita la capacidad de tomar decisiones informadas.

La tasación de un inmueble constituye un desafío de alta dificultad, ya que una valoración realista depende de múltiples factores, entre ellos las características técnicas de la vivienda, su localización, la percepción social de la vecindad, entre otros. Según Geerts et al. (2023), el precio de una propiedad está determinado por la disposición a pagar de los compradores, quienes buscan según sus necesidades ciertas características técnicas de una casa, junto con una buena localización y un vecindario adecuado. En consecuencia, la tarea de obtener una tasación objetiva de una vivienda se torna en un desafío particularmente complejo de resolver.

En esta misma línea, el mercado inmobiliario también impone una amplia responsabilidad en la toma de decisiones por parte de los propietarios. Con frecuencia, quienes buscan vender su vivienda realizan remodelaciones con el propósito de incrementar su rentabilidad. No obstante, dichas intervenciones no siempre cumplen las expectativas planteadas. Según Dunaway (2024), un 30% de las renovaciones efectuadas en Estados Unidos tienen como objetivo aumentar el valor de la propiedad; sin embargo, un 24% de los propietarios se arrepiente por el elevado gasto que implicaron y un 16% por haber incurrido en deudas. Estos datos reflejan que los propietarios suelen invertir en remodelaciones sin contar con claridad respecto al impacto real de dichas mejoras en el valor de la vivienda, lo que conlleva riesgos financieros y genera frustración.

Un caso de estudio interesante a analizar es el de la ciudad de Ames, ubicada en el estado de Iowa, Estados Unidos. En Ames la valoración óptima de una vivienda y la toma de decisiones en diseño y remodelación representa un desafío considerable debido a la influencia de factores políticos, económicos y sociales. Según Ye (2024), dichos fenómenos son causados por la marcada presencia de la universidad Iowa State University, que convierte a esta ciudad en un centro de atracción para estudiantes y profesionales. Esto genera una dinámica de alta demanda de inmuebles en la zona, lo que acentúa la necesidad de tasaciones precisas y decisiones óptimas de diseño y remodelación de una vivienda, lo que resalta la importancia de contar con información adecuada sobre las preferencias de los compradores a fin de satisfacer sus requerimientos habitacionales.

Considerando el desafío que implica lograr una tasación óptima de viviendas en Ames, junto con la necesidad de tomar decisiones acertadas en materia de diseño y remodelación, se plantea el desarrollo de un proyecto orientado a construir un modelo robusto de soporte para la toma de decisiones en el ámbito inmobiliario. Cuyo objetivo principal es predecir con precisión el valor de una vivienda y, adicionalmente, recomendar remodelaciones óptimas bajo restricciones presupuestarias. Con ello, se busca establecer un precio justo que contemple las necesidades y disposición de pago de los compradores, maximizando al mismo tiempo el valor y la rentabilidad de la propiedad.

De este modo, se desarrollará una herramienta que permitirá a los compradores encontrar su vivienda óptima según sus necesidades, a un precio justo y accesible, y que al mismo tiempo brindará a los vendedores orientación en decisiones de diseño y remodelación para incrementar la rentabilidad de sus propiedades.

Para continuar con esta tarea, resulta imprescindible la definición de indicadores clave de desempeño (KPI), ya que son fundamentales para evaluar el grado en que los objetivos específicos del proyecto se cumplen de manera efectiva. En este caso, los KPI seleccionados permiten medir tanto la capacidad predictiva de los modelos, como la coherencia de las soluciones propuestas y la eficiencia en el uso de recursos.

En primer lugar, se consideró medir el error de predicción como un KPI central para el objetivo de construir un modelo de tasación confiable. Para ello se utilizan diversos indicadores complementarios que permiten capturar distintas dimensiones del desempeño predictivo. Entre ellos se incluye el R^2, que refleja la proporción de la variabilidad del precio explicada por el modelo (Chicco, 2021); el *MAPE: Mean Absolute Percentage Error*, que mide la precisión relativa de las predicciones en términos porcentuales (Kim & Kim, 2016); y métricas adicionales como el *RMSE: Root Mean Squared Error*, que penaliza fuertemente los errores grandes, y el *MAE: Mean Absolute Error*, que cuantifica el error absoluto promedio en unidades monetarias (Hodson et al., 2022).

En segundo lugar, se definió cuantificar el porcentaje de soluciones que cumplen las restricciones como el KPI asociado al objetivo de evaluar consistencia y restricciones de diseño. Este indicador refleja el grado en que las soluciones generadas respetan las condiciones impuestas por el problema, por ejemplo, restricciones estructurales o normativas. De esta manera, no solo se mide la factibilidad técnica de las soluciones, sino también su validez dentro de los marcos establecidos.

Finalmente, se incorporó medir el *ROI: Return on Investment*, considerando el porcentaje de presupuesto usado, con el objetivo de diseñar un modelo de optimización de bajo presupuesto que asegure remodelaciones rentables. El ROI es una métrica estándar para expresar la relación entre beneficios e inversión (Peppard & Ward, 2016) y, en proyectos inmobiliarios, resulta especialmente relevante al comparar alternativas no solo por su costo, sino también por el valor adicional que aportan al activo (Geltner, Miller, Clayton, & Eichholtz, 2014). Monitorear este indicador permite asignar recursos de manera óptima, priorizando soluciones que maximicen la rentabilidad y minimicen el riesgo de sobreinversión.

En conjunto, estos KPI ofrecen una visión integral del desempeño del proyecto, al cubrir aspectos de precisión, consistencia y rentabilidad. Su monitoreo permite asegurar que los objetivos específicos trazados no solo se cumplan en la teoría, sino también en la práctica aplicada del modelo desarrollado.

2. DESCRIPCIÓN DEL PROBLEMA

2.1. Problema

Uno de los principales problemas en el mercado inmobiliario global es la inexactitud en la tasación de viviendas, lo que genera brechas significativas entre el valor real de un inmueble y el valor estimado en los procesos de compraventa. La literatura señala que las metodologías tradicionales de valoración, como el enfoque comparativo de ventas, dependen excesivamente de información limitada o de variables poco precisas que de vez en cuando pueden ser cualitativas. Esto puedo llevar a que la tasación de un inmueble se vuelva una "opinión de precio" en vez de reflejar un valor objetivo de mercado (AppraisersBlogs, 2015).

Junto con esto, se suman los sesgos cognitivos presentes en la práctica, como el *anchoring*[1], o la sobre dependencia en valores previos, que distorsionan la objetividad del cálculo y producen resultados inconsistentes (Evans et al., 2019). Esta falta de precisión afecta directamente a compradores como a vendedores: para los primeros, implica pagar de más o enfrentar dificultades en el financiamiento de una propiedad, mientras que para los segundos puede significar la pérdida de oportunidades de venta o una subvaloración injusta de su patrimonio. En consecuencia, la mala tasación constituye un dolor estructural del mercado inmobiliario, pues obstaculiza transacciones justas, limita la rentabilidad y disminuye la confianza de los agentes involucrados.

En esta misma línea, una tasación óptima de una vivienda debe considera la disposición al pago de las personas, la cual es definida según las necesidades y expectativas que presentan. En el caso particular de la ciudad de Ames, según datos del Censo de Estados Unidos del año 2022 analizados por el Portal de Noticias Univisión, Iowa se ubica entre los estados más jóvenes del país, ocupando la posición 20 con una media de edad de 31,4 años,

[1]Anchoring: En un contexto de tasación, hace referencia a cuando el evaluador se ancla o fija en un valor de referencia previo del inmueble.

lo que indica una presencia significativa de población *millennials*[2]. Este factor demográfico es de mucha relevancia al analizar la demanda habitacional, ya que de acuerdo con National Association of Realtors (2025), los *millennials*[2] representan el 29% de los compradores a nivel nacional, proporción que en Iowa podría ser incluso mayor debido a la estructura poblacional más joven del estado.

Si bien la edad mediana de compra de vivienda en Estados Unidos se sitúa en los 38 años, en Iowa, tanto la juventud relativa de la población como los precios más accesibles del mercado facilitan a que los *millennials*[2] ingresen al proceso de adquisición más temprano. No obstante, sus decisiones de compra se ven condicionadas por la necesidad de financiamiento, la preferencia por viviendas familiares o multigeneracionales, y la dependencia de agentes inmobiliarios para guiar el proceso. En consecuencia, el diseño de un modelo predictivo y de optimización para el mercado inmobiliario en Ames, requiere considerar las particularidades de cada segmento etario: por ejemplo, familias jóvenes con hijos suelen priorizar atributos como un patio amplio o presencia de un cerco, mientras que compradores solteros o parejas sin hijos podrían valorar más cercanía a servicios o la eficiencia del espacio.

Por otro lado, se evidencia también el desafío que enfrentan los propietarios al momento de configurar un diseño de vivienda óptima para habitarla o realizar remodelaciones con el fin de venderla y obtener ganancias. El concepto de *'casa óptima'* alude a un equilibrio entre funcionalidad, confort y un valor económico congruente con el mercado, lo que obliga a los propietarios a adaptarse a restricciones presupuestarias y a tomar decisiones informadas que maximicen la rentabilidad de sus inmuebles o bien que aseguren un espacio técnicamente adecuado a sus necesidades habitacionales. No obstante, como señala Yun and Lautz (2025), la rentabilidad real de una remodelación varía ampliamente según el tipo de proyecto. Esto refleja la complejidad de decidir en qué invertir y el riesgo de no alcanzar el retorno esperado.

[2]Millennial: Grupo de personas nacidas entre los años 1981 y 1996.

Las decisiones de diseño y remodelación resultan cada vez más difíciles en un entorno inmobiliario altamente competitivo, donde el aumento de los precios de materiales, mano de obra y construcción condiciona las alternativas disponibles para los propietarios dentro de estrictas restricciones presupuestarias. Lo desafiante no es solo ejecutar las obras, sino decidir correctamente en qué invertir. Según Yun and Lautz (2025) aunque ciertos proyectos obtienen altas tasas de recuperación del costo, otros como remodelaciones mayores de interiores o ampliaciones apenas recuperan parcialmente la inversión, lo que puede dejar al propietario en deuda o con pérdidas de valor efectivo.

Los riesgos asociados a la inversión en diseño y remodelación de viviendas constituyen un desafío significativo para el sector inmobiliario. A esta complejidad se suman factores como las condiciones del mercado, las normativas de construcción y las necesidades cambiantes de los compradores, los cuales influyen directamente en la rentabilidad de las mejoras. En consecuencia, la toma de decisiones respecto al diseño y remodelación se torna altamente difícil y, si no se gestiona adecuadamente, puede derivar en pérdidas financieras considerables. Como señala Macek and Vitásek (2024), los proyectos de construcción y renovación están expuestos a múltiples riesgos financieros, técnicos y regulatorios, lo que refuerza la necesidad de adoptar enfoques más analíticos y predictivos en la planificación de estas inversiones. Así, se vuelve indispensable contar con herramientas capaces de predecir el impacto que tendrán determinadas intervenciones sobre el valor de reventa de la propiedad y sobre el cumplimiento del presupuesto disponible.

De esta forma, se puede observar que los requerimientos de los compradores presentan una gran diversidad, ya que dependen directamente de sus expectativas y necesidades la decisión de adquirir una vivienda. Para abordar esta complejidad, el modelo de tasación inmobiliaria a optimizar debe partir de una base técnicamente coherente de diseño habitacional, lo que implica asegurar que la estructura de la vivienda sea congruente en términos funcionales, por ejemplo, que el número de habitaciones guarde una relación lógica con la cantidad de baños y con la superficie total disponible.

El modelo de tasación y diseño de viviendas debe construirse sobre una base que considere restricciones presupuestarias, técnicas y geométricas, garantizando que la vivienda propuesta sea factible tanto en términos económicos como constructivos. Sobre esta base se incorporan luego las necesidades específicas de cada cliente, las cuales definen un conjunto particular de requisitos a satisfacer. Por ejemplo, una familia de tres integrantes podría requerir tres dormitorios y dos baños, mientras que un estudiante probablemente priorizará una sola habitación y un baño. Estas diferencias en las preferencias y demandas generan una ramificación natural del problema en múltiples subproblemas, cada uno de los cuales busca determinar la configuración habitacional óptima para un perfil de comprador específico.

En esta etapa de ramificación surge otro desafío a la tasación, en donde se debe integrar múltiples variables de carácter multifactorial, como ubicación, superficie, distribución interna, restricción de habitaciones, percepción del vecindario, entre otras. Para que el modelo sea efectivo, resulta indispensable distinguir entre aquellas variables que son verdaderamente relevantes en la determinación del valor y la satisfacción del comprador, y aquellas que tienen escasa o nula incidencia, de manera de simplificar el problema sin sacrificar precisión en la búsqueda de la solución óptima. En dicho filtro se requiere un estudio exhaustivo de variables a analizar según su relevancia, lo que aumenta la dificultad del modelo a optimizar.

2.2. Análisis de datos: Ames Housing

La base de datos utilizada en este proyecto corresponde al Ames Housing Dataset, recopilado originalmente por la Oficina de Tasación de la ciudad de Ames, Iowa, y organizada por De Cock (2011) como una alternativa moderna y más completa al clásico *Boston Housing Dataset*. El conjunto de datos contiene información detallada de las ventas residenciales realizadas entre 2006 y 2010 en la ciudad de Ames, que está conformada por 2930 observaciones, cada una representando una transacción de vivienda con su respectiva información estructural, de entorno y de mercado.

La base de datos se compone de 80 variables explicativas y una variable dependiente, correspondiente al precio de venta. Estas variables abarcan aspectos físicos, cualitativos y contextuales, lo que refleja la complejidad del mercado inmobiliario. Dentro de ellas se incluyen variables continuas, que representan medidas físicas de la vivienda, como la superficie del terreno y la superficie habitable. Por otro lado, se encuentran las variables discretas, que contabilizan elementos enteros de la vivienda, como el número de habitaciones o baños. Y finalmente hay variables categóricas, tanto ordinales como nominales, que entregan información sobre la calidad o condiciones de ciertos atributos y el contexto en el que se ubica la propiedad, como la condición de la piscina en una escala jerárquica, la calidad de la cocina, el tipo de vecindario o el tipo de calle en que se encuentra la vivienda.

Este nivel de detalle introduce dificultades técnicas relevantes. Entre ellas destacan la existencia de viviendas con dimensiones inusualmente superiores al promedio, la presencia de multicolinealidad entre variables altamente correlacionadas y la asimetría en la distribución del precio de venta, cuyos valores fluctúan entre 34.900 y más de 755.000 dólares estadounidenses (Özdemir, 2022).

Asimismo, se observan fenómenos de heterogeneidad espacial, en los cuales ciertos vecindarios incrementan significativamente el valor de las propiedades, mientras que otros lo reducen. Adicionalmente, la base de datos presenta una elevada correlación entre múltiples variables, lo que da lugar a relaciones complejas y no lineales entre atributos de la vivienda y sus precios (De Cock, 2011). Estas características hacen que la predicción de los precios no pueda resolverse mediante modelos lineales simples, requiriendo metodologías más robustas capaces de capturar relaciones no lineales y dependencias espaciales.

El tratamiento de los datos consistió en diferentes etapas. En primer lugar, se realizó una depuración de los datos para entender cada variable e identificar datos faltantes, datos erróneos e inusuales. Por otro lado, se analizaron correlaciones, colinealidades y datos poco importantes, con el objetivo de desarrollar un caso base a través de una regresión lineal y así entender de mejor manera el comportamiento de cada variable sobre el precio;

además de poder ver la precisión que alcanza una regresión para estimar el precio de una vivienda.

Al iniciar el análisis, se detectó que una gran cantidad de variables contenían valores codificados como "NA" o *"None"*. En la práctica, estos valores no representaban datos faltantes reales, sino la ausencia de una característica en la vivienda. Para evitar confusiones en el proceso de modelado donde estos atributos son interpretados como valores faltantes, se procedió a diferenciarlos explícitamente de los nulos reales, reemplazándolos por "No aplica" en variables cualitativas y por cero en variables cuantitativas. Un ejemplo es la variable *PoolQC*, que aparece codificada como 'NA', lo cual en realidad significa 'No aplica'. Por lo tanto, fue necesario realizar este ajuste.

Una vez realizada la separación, se abordaron los nulos verdaderos mediante un análisis caso a caso para decidir entre imputar o eliminar registros. En particular, en LotFrontage (490 nulos), se aplicó la metodología de House Price Prediction Using Machine Learning: A Case in Iowa Ozdemir (2022), imputando los valores faltantes con la mediana correspondiente a cada vecindario (Neighborhood).

Para la eliminación de determinadas filas y corrección de datos erróneos, se siguieron recomendaciones expuestas en la literatura con el fin de depurar datos inconsistentes. En particular, como señala De Cock (2011), es recomendable eliminar de la base de datos aquellas casas que tienen más de 4000 pies cuadrados debido a que es posible que representen ventas que no reflejan los valores reales del mercado o son ventas inusuales. Así, se eliminaron cinco observaciones de la base de datos original.

En cuanto a Marcelino (2017), indica que hay un error en la variable *Garage Year Blt* debido a que se identificó una observación con el año 2207, lo cual es incorrecto dado que la base de datos abarca información entre los años 2006 y 2010. Por lo tanto, esta celda fue reemplazada por el año 2007 considerando que fue un error de input.

Finalmente, otras imputaciones puntuales y eliminaciones menores de filas se detallan en la Tabla .1 en el Anexo 1, donde se presenta un resumen general de todas las decisiones

tomadas. A continuación, para armar el caso base se analizan las variables cuantitativas y categóricas de forma separada, pues el método usado depende del tipo de variable.

2.2.1. Variables Cuantitativas

En cuanto al tratamiento de variables cuantitativas, lo primero que se analizó fue la correlación entre variables mediante una matriz de correlación de Spearman. La elección de este método se debe principalmente a que no requiere de suposiciones sobre la distribución de los datos, mientras que la correlación de Pearson asume linealidad y normalidad (Data Science & Beyond, 2023). Por lo tanto, esta matriz fue utilizada para identificar aquellas variables que tienen una correlación mayor a 0,7, debido a que queremos evitar multicolinealidad y redundancia en los datos.

En segundo lugar, las variables que están altamente correlacionadas se evaluaron de manera individual en relación con la variable *SalePrice*, como se muestra en el gráfico en el Anexo 2, .1. Las variables con mayor grado de correlación con el precio son *Gr Liv Area*, Total *Bsmt SF* y *Garage Cars* donde tienen una correlación mayor a 0,65.

Considerando ambos aspectos, se eliminaron aquellas variables que tenían alta correlación con otras variables y que además presentaban una baja correlación con *SalePrice*, dejando de esta manera solo aquellas que eran más relevantes para nuestro estudio.

Un caso en específico son las variables *Garage Yr Blt* y *Year Build*, que tienen una correlación alta igual a 0,86 y una correlación con *SalePrice* de 0,25 y 0,56 respectivamente. En consecuencia, la variable eliminada fue *Garage Yr build*, dado que presenta una menor asociación con precio de la vivienda y está altamente correlacionada con el año que fue construida la casa.

Adicionalmente, se llevó a cabo un análisis de redundancia entre variables, como en el caso de *Bsmt Fin SF 1*, *Fin SF 2* y *Bsmt Unf* , la suma de estas tres variables da como resultado la variable *Total Bsmt SF*. Por esta razón, se conservó solo esta última, debido a que representa de mejor manera al sótano y su correlación con la variable *SalePrice* es

mayor a 0,6 mientras que las otras variables presentan correlaciones menores. De manera similar ocurre con las variables *Screen Porch*, *3Ssn Porch*, *Open Porch SF* y *Enclose Porch*. Estás variables tienen baja correlación con *SalePrice* de manera individual y dividen la información sobre el espacio que tiene el porch de la vivienda. Por esta razón, decidimos sumar las tres variables y crear una nueva variable llamada *Total Porch*, debido a que no influyen de manera significativa en el precio final y es más relevante tener solo el total que indicara si hay o no *porch* y no cuanto mide cada parte por separado.

Asimismo, se eliminaron ciertas variables que, además de tener una baja correlación con la variable *SalePrice*, presentan valores atípicos, por lo que fueron consideradas poco relevantes para nuestro caso. Un ejemplo es la variable es *Misc Val*, la cual tiene una correlación de –0,01 con *SalePrice* y presenta 101 outliers. La decisión de eliminarla viene dada a que aporta poco valor explicativo en relación con el precio de venta y contiene valores atípicos que afectan su comportamiento estadístico. Sin embargo, decidimos dejar su variable categórica equivalente que es *Misc Feature* que representa las características misceláneas de la vivienda.

El resto de las variables cuantitativas eliminadas y redundantes se encuentran detalladas en la Tabla .2 en el Anexo 1.

2.2.2. Variables Categóricas

Con el fin de estandarizar el análisis, las variables categóricas se clasificaron en nominales y ordinales. Posteriormente, se aplicó la misma lógica utilizada con las variables numéricas: realizar una comparación con *SalePrice* y ver la correlación entre variables dentro de un mismo grupo.

Previo a hacer la separación definitiva, en variables con una distribución extremadamente desbalanceada (ej. *PoolQC*, donde 99,6% corresponde a “No aplica”), la escala de calidad completa o presencia de múltiples atributos no era representativa. Por lo tanto, se optó por simplificarlas a variables binarias (presencia/ausencia o atributo/otro). En la

práctica, este tratamiento no altera la decisión final de mantener o descartar la variable, sino que busca mejorar la estabilidad del modelo evitando categorías con muy pocos casos.

Categóricas Nominales

Para el análisis de las variables categóricas nominales se usaron dos modelos. En primer lugar, se aplicó ANOVA (η^2) para conocer la proporción de varianza en *SalePrice*, lo que permite establecer un *ranking* de relevancia individual. Luego se calculó Cramer's V entre pares, para conocer la relación entre ellas, detectar redundancias y posibles problemas de multicolinealidad. Los resultados se pueden visualizar en los gráficos en el Anexo 2.2 y 2.4.

Con base en Cramér's V, se consideraron redundantes aquellos pares con valores V $\geq$ 0,7, dado que este rango indica alta superposición de información. En tales casos, se retuvo la variable con mayor relevancia frente a *SalePrice* según η^2 y se descartó la otra. Un ejemplo de esto es el par *MSSubClass–BldgType*, que presentó V = 0,88. Ambas variables describen características similares de la vivienda, pero se decidió conservar *MSSubClass*, ya que mostró mayor poder explicativo en relación con el precio.

Categóricas Ordinales

Las variables categóricas ordinales fueron mapeadas a escalas numéricas de acuerdo con el diccionario del *dataset*, por ejemplo, Po=1, Fa=2, TA=3, Gd=4, Ex=5; "No aplica"=0. De esta forma, estas variables pudieron analizarse mediante correlación de Spearman, tanto frente a *SalePrice* como entre sí, siguiendo la misma lógica aplicada a las variables numéricas que se ve en en el Anexo 2.5.

Considerando ambos aspectos, se eliminaron aquellas variables que tenían alta correlación con otras variables y que además presentaban una baja correlación con *SalePrice*, dejando de esta manera solo aquellas que eran más relevantes para nuestro estudio.

Un ejemplo de este proceso se observa en *ExterQual*, que presentó una correlación superior a 0,7 tanto con *OverallQual* como con *KitchenQual*. Aunque *ExterQual* mostraba

una mayor asociación con el precio que *KitchenQual*, se optó por eliminarla para evitar redundancia, dado que *OverallQual* y *KitchenQual* en conjunto capturan mejor la variabilidad de la calidad de la vivienda.

Para finalizar el análisis, se revisaron nuevamente aquellas variables categóricas identificadas como extremadamente desbalanceadas. En estos casos, la falta de variabilidad reducía su valor explicativo, por lo que se optó por eliminarlas. Un ejemplo es *Utilities*, donde el atributo *"AllPub"* representaba aproximadamente el 99% de los registros, convirtiéndola en una variable prácticamente constante. Las variables descartadas por redundancia u otras razones se detallan en la Tabla .3 del Anexo 1; tras la depuración, el caso base quedó con 2914 observaciones y 53 variables.

2.2.3. Caso Base

Luego del análisis de la base de datos, se trabajó en la construcción de un caso base para el estudio del comportamiento de los precios de las viviendas.

En primer lugar, para llevar los valores de las viviendas a precios actuales, se ajustó la variable *SalePrice* utilizando el Índice de Precios al Consumidor (IPC) obtenido de la página web del *Federal Reserve Bank of St. Louis*. De esta forma, se construyó la variable *SalePrice_present*, que refleja el valor de las casas actuales. La razón de esta decisión es que nuestro objetivo consiste en desarrollar un modelo que se ajuste a las condiciones de mercado vigentes, de esta manera los resultados obtenidos pueden ser comparados con los precios observados en distintas plataformas de venta y arriendo de viviendas, lo que otorga mayor utilidad y aplicabilidad práctica al modelo. La evolución de los precios originales y ajustados puede observarse en la Figura .9 en el Anexo 2.

Posteriormente, se construyó un modelo de regresión lineal multivariable para analizar el comportamiento y la significancia de las variables, cuyo gráfico se puede observar en la Figura .10 en el Anexo 2. Dado el comportamiento y la distribución de los datos, se

aplicó una transformación logarítmica a la variable *SalePrice_present* con el fin de mejorar el ajuste del modelo.

A continuación, el gráfico de la regresión lineal multivariable de *log(Sale Price_present)*.

Luego, se evaluó la relevancia de las variables en la regresión lineal. Esto con el objetivo de determinar si es necesario eliminar variables que no aportan significativamente al modelo. Para ello, se calcularon los *p-values* de cada variable y su contribución al ajuste del modelo. Se eliminaron todas aquellas variables con un *p-value* mayor a 0,05, lo que indica que carecían de significancia estadística en la regresión. En total, se eliminaron 4 variables: *Lot Config* con *p-value* igual a 0,18, *Roof Style* con *p-value* de 0.095, *Sale Type* con un *p-value* de 0,183 y *Lot Shape* con un *p-value* de 0.275.

A continuación, se presenta una tabla comparativa de los resultados obtenidos de la regresión lineal antes y después de eliminar las variables menos significativas.

Table 2.1. Resultados de la Regresión Lineal Multivariable con y sin eliminación de variables.

Métrica	**Regresión Lineal Multivariable**	**Regresión Lineal Multivariable al eliminar variables**
R^2	0,925	0,924
MAPE (%)	7,98	8,04
RMSE	29.419,15	29.871,62
Skewness	−1,444	−1,417
curtosis	18,453	14,982

Se observa que la regresión lineal multivariable logró un R^2 de 0,925, lo que evidencia que el modelo explica un 92,5% de la variabilidad de los precios presentes. El MAPE de

7,98% refleja un error porcentual promedio de aproximadamente 8%. El RMSE de aproximadamente 29.419, indica en promedio cuánto se equivocó el modelo en las predicciones de los precios de las viviendas respecto a los valores reales en dólares.

Respecto de los residuos, se observó un sesgo negativo con una asimetría igual a –1,444, lo que significa que los errores no se distribuyen de manera simétrica, lo cual implica que el modelo tiende a sobreestimar los precios de las viviendas de menor valor y a subestimar las de mayor valor. Esto sugiere que la regresión lineal presenta dificultades para capturar de manera adecuada los valores extremos. Asimismo, la Curtosis con un valor de 18,453 refleja la existencia de colas pesadas en la distribución de los errores, es decir, el modelo presenta una mayor frecuencia de errores extremos en comparación con una distribución normal, como se puede ver en el gráfico de residuos en el Anexo 2, Figura .11.

En cambio, al hacer nuevamente la regresión lineal eliminando aquellas variables que eran insignificantes se observa que, si bien R^2, MAPE y RMSE empeoraron ligeramente, esto se debe a que algunas variables eliminadas podrían haber aportado información marginal al ajuste global. No obstante, el beneficio de eliminar variables no significativas es obtener un modelo menos complejo y más fácil de interpretar.

De esta manera, el análisis de eliminación de variables no significativas resalta la importancia de seleccionar cuidadosamente las variables que aportan información relevante y simplicidad para una mejor interpretabilidad o más información y mayor complejidad.

3. DISCUSIÓN METODOLÓGICA

El problema de la tasación de viviendas es altamente complejo debido a la gran cantidad y heterogeneidad de variables, la multicolinealidad y la presencia de outliers[1], entre otros factores. Estas condiciones limitan el rendimiento de métodos tradicionales, como las regresiones lineales múltiples. Por ello, para ciertos casos resulta necesario aplicar metodologías más robustas, capaces de adaptarse y mejorar la precisión frente a interacciones complejas entre variables. Una de las alternativas son las técnicas de *machine learning*, los cuales son métodos computacionales que a partir de patrones y relaciones permiten a los modelos aprender sin exclusivamente depender de supuestos estadísticos.

Para comenzar a abordar este problema, se presentarán las alternativas candidatas y luego se verá cuál de ellas tiene mejor desempeño, basado en la literatura y nuestros propios hallazgos. El primer método es una regresión lineal, la elección de esta alternativa es debido a la facilidad de interpretación econométrica que entrega y, adicionalmente, servirá como línea de base para los pasos futuros del proyecto. Otra razón que sustenta la elección de este método es que según De Cock (2011) se encontró que "cerca del 80 % de la variación en el precio de venta de las viviendas residenciales puede explicarse simplemente considerando el vecindario y la superficie total de la vivienda, variación que se podría identificar y analizar con una regresión lineal" (p. 13). Por lo que bibliográficamente según la data se justifica el uso de la regresión lineal como método.

Además, la regresión lineal está en línea con la tradición econométrica en el análisis de precios hedónicos (Harrison and Rubinfeld (1978)). El estudio de Poeta, Gerhardt, and Stumpf Gonzalez (2019) sobre análisis de precios hedónicos en ciudades de tamaño medio en Brasil, muestra que los modelos de regresión lineal clásica constituyen una herramienta válida y ampliamente utilizada en la valoración de viviendas. Los autores señalan que este tipo de modelos permiten identificar de manera clara y directa el efecto marginal de atributos específicos de la vivienda.

[1]Outlier: Hace referencia a un valor atípico, que se aleja de un patrón general.

Sin embargo, este método presenta limitaciones y desventajas. En primer lugar, el supuesto de linealidad restringe la capacidad del modelo para capturar relaciones no lineales entre las variables. Por ejemplo, el efecto de la superficie habitable sobre el precio de venta no necesariamente es proporcional, ya que un aumento de metros cuadrados en viviendas pequeñas puede tener un impacto mayor en el precio que el mismo aumento en viviendas de gran tamaño. Este tipo de comportamiento no lineal ha sido identificado en el set de datos de Ames, lo que justificaría posteriormente la aplicación de métodos más flexibles como *Gradient Boosting* o *XGBoost* (Özdemir (2022)). Además, variables ordinales como la calidad general presentan saltos discretos en el valor de la vivienda, lo que introduce comportamientos no lineales que el método de regresión no puede capturar adecuadamente (Ye (2024)).

En segundo lugar, la regresión lineal requiere el cumplimiento de independencia y homocedasticidad en los errores, condiciones que no siempre se cumplen en este caso. De Cock (2011) advierte que las viviendas más grandes tienden a presentar una varianza creciente en los precios de venta, lo cual genera un problema de heterocedasticidad que afecta la validez de los intervalos de confianza y las pruebas de significancia. En el caso específico de Ames, existen patrones espaciales y temporales que afectan los residuos. Las casas de un mismo vecindario tienden a presentar errores similares debido a características no observadas compartidas, y el periodo de ventas introduce posibles efectos cíclicos del mercado. Esto genera correlación en los errores, lo que viola la independencia y conlleva a que los residuos no cumplan con el supuesto de normalidad (Poeta et al. (2019)). Un ejemplo claro de esto se percibe al analizar la variable dependiente *SalePrice*, ya que presenta una distribución asimétrica a la derecha, lo que arrastra a los errores hacia distribuciones no normales.

En tercer lugar, la regresión clásica exige la creación de variables binarias para su incorporación al modelo, lo que incrementa la dimensionalidad y hace al modelo más sensible a problemas de multicolinealidad. Este fenómeno complica la interpretación de los coeficientes y puede afectar la estabilidad de las estimaciones, un desafío que también

ha sido destacado en la literatura más reciente al aplicar modelos de *machine learning* al conjunto de Ames (Özdemir (2022)).

Finalmente, la ausencia de multicolinealidad tampoco se cumple estrictamente como ya se discutió previamente en análisis de datos. Hay variables altamente correlacionadas como las estudiadas *GrLivArea*, *TotalBsmtSF*, *1stFlrSF* y *GarageArea*, lo que incrementa la varianza de los coeficientes estimados e introduce inestabilidad en el modelo. Este problema de multicolinealidad en bases inmobiliarias ha sido documentado también en la literatura reciente (Geerts, vanden Broucke, S., and De Weerdt, J. (2023)). Esto hace surgir la necesidad de buscar otro método que prediga con mayor precisión los precios de venta de las viviendas, por esto "Los algoritmos basados en árboles de decisión como *Random Forest*, *Gradient Boosting*, *XGBoost* y *CatBoost* superan la regresión lineal de referencia en la predicción de los precios de la vivienda en Ames, Iowa" (Özdemir (2022), p. 12).

Los métodos mencionados anteriormente se agrupan en tres categorías. Los más básicos y populares son los árboles de decisión, los cuales son algoritmo de aprendizaje supervisado no paramétrico con una estructura en forma de árbol, que se utilizan para tareas de regresión y clasificación (Ibm-b (2025)). Sin embargo, entre sus principales desventajas destacan la propensión al sobreajuste y la alta varianza de sus estimadores. Luego, los modelos de bosques aleatorios o *random forest*, son aquellos que combinan múltiples árboles entrenados de forma independiente sobre subconjuntos de datos aleatorios, logrando predicciones más precisas y estables. No obstante, presentan limitaciones como la pérdida de interpretabilidad y el elevado costo computacional y de tiempo de entrenamiento (Ibm-b (2025)). Finalmente, los modelos de *boosting* como lo son *XGBoost*, *Gradient Boost* o *CatBoost* emplean una técnica de ensamblaje que combina múltiples modelos débiles que aprenden de sus errores en cada iteración. Entre sus ventajas se encuentran la fácil implementación y la reducción secuencial del sesgo mediante la selección de variables útiles, lo que mejora su eficiencia. Sin embargo, pueden enfrentar problemas de sobreajuste si no se controlan adecuadamente, además de que su entrenamiento suele ser más largo que modelos como *Random Forest* (Ibm-a (2021)).

Dentro de las alternativas exploradas, notamos las regresiones penalizadas como *Lasso* y *Ridge*. Estos métodos surgen como extensiones de la regresión lineal tradicional mediante el uso de penalizaciones. En particular, *Lasso* se caracteriza por su capacidad de realizar selección de variables y regularización, lo que ayuda a prevenir el sobreajuste y facilita el trabajo con datos de alta dimensionalidad. En contraste, *Ridge* actúa reduciendo la magnitud de los coeficientes para enfrentar la multicolinealidad y simplificar el modelo, aunque sin eliminar predictores como lo hace *Lasso* (Chandra (2023)). Si bien estas técnicas presentan ventajas en cuanto a la regularización y reducción de complejidad, comparten limitaciones con la regresión lineal tradicional, ya que dependen de la suposición de linealidad, muestran dificultades para capturar interacciones no lineales entre variables y mantienen cierta sensibilidad frente a *outliers*.

Según múltiples investigaciones realizadas con el propio *set* de datos, tanto como lo hacen Sharma, Harsora, and Ogunleye (2024) y Ye (2024), concluyen que el modelo predictivo con mejores resultados comparado con otros modelos como los mencionados anteriormente es el algoritmo *XGBoost*, llegando a tener hasta $R^2 \approx 0.92$, superando otros modelos como regresiones lineales o *Random Forest*. Según lo que mencionan Kavlakoglu and Russi (2024), este algoritmo destaca por su velocidad, eficiencia y su capacidad de escalar a grandes volúmenes de datos.

Por ello, la metodología finalmente seleccionada en este proyecto es *XGBoost* (*Extreme Gradient Boosting*), un algoritmo de ensamble basado en *boosting* que combina múltiples árboles de decisión entrenados secuencialmente para minimizar los errores. A diferencia de los métodos lineales, *XGBoost* no requiere supuestos de distribución, puede manejar datos heterogéneos como numéricos, categóricos y ordinales, y es capaz de capturar relaciones altamente no lineales entre las variables. En el contexto específico de Ames, Iowa, investigaciones recientes han mostrado de que este algoritmo alcanza un desempeño superior en predicción de precios, llegando a valores de R^2 cercanos a 0,89 (Özdemir (2022)) y reduciendo el error en comparación con regresiones lineales y métodos penalizados.

Entre sus ventajas, destacan su alta precisión, capacidad de escalar a grandes volúmenes de datos y manejo de un gran número de predictores sin perder eficiencia. Sin embargo, como cualquier otro método, *XGBoost* no es perfecto. Entre sus desventajas se encuentra que presenta una menor interpretabilidad en comparación con modelos econométricos tradicionales, un mayor riesgo de sobreajuste si no se regulan parámetros como la profundidad de los árboles o la tasa de aprendizaje y un entrenamiento computacionalmente más costoso. Según Nadarajah (2025) una calibración inadecuada de sus hiperparámetros puede conducir al sobreajuste, generando modelos con buen desempeño en el entrenamiento, pero con una menor capacidad de generalización. Por ello, el tratamiento adecuado de los datos resulta vital para obtener un algoritmo robusto y adaptable, evitando que, en lugar de aprender, termine memorizando patrones específicos de ciertos grupos.

Aun así, este método se posiciona como una de las metodologías más adecuadas para abordar el problema de tasación de viviendas en Ames, al permitir superar las limitaciones de los modelos lineales y penalizados, y adaptarse a la complejidad de los datos inmobiliarios. La literatura reciente ha demostrado que la implementación de modelos *boosting*, además de ser útiles en la fase de predicción, también ha mostrado su valor en problemas de optimización. Mistry, Letsios, Krennrich, Lee, and Misener (2018) desarrollaron un enfoque basado en este tipo de modelos dentro de problemas mixtos enteros no lineales convexos, logrando resultados superiores frente a métodos más tradicionales que tratan al predictor como una "caja negra", es decir un sistema que produce resultados sin que el usuario pueda ver o comprender cómo funciona (Cambridge University Press (n.d.)). Al aprovechar la estructura interna de los árboles de decisión, se obtuvieron soluciones más precisas y con mejores cotas de optimalidad, incluso en escenarios de gran escala. Esto respalda la viabilidad de utilizar algoritmos como *XGBoost* tanto en la fase predictiva como en la formulación matemática para seleccionar configuraciones óptimas bajo restricciones.

En síntesis, la elección de *XGBoost* como metodología principal se fundamenta en tres razones principales. En primer lugar, es uno de los algoritmos con mayor documentación aplicada a la misma base de datos, con desempeños ejemplares en la predicción, lo que

respalda su validez empírica. En segundo lugar, destaca por su capacidad de escalabilidad frente a volúmenes crecientes de datos, un aspecto crítico en un contexto donde la oferta de información tiende a expandirse cada vez más. Finalmente, su aprendizaje iterativo le permite mejorar progresivamente su precisión al reducir errores y sesgos en cada fase de entrenamiento. Estas características lo posicionan como un modelo robusto y adaptable, capaz de superar los enfoques tradicionales y responder a la complejidad que implica la tasación y optimización de una vivienda.

4. CONCLUSIONES

4.1. Pasos futuros

En los pasos futuros, detallados en la carta Gantt adjunta en los anexos, se contempla la posibilidad de personalizar las variables del modelo según el perfil del cliente, considerando las diferencias generacionales y las preferencias habitacionales que estas implican. Esta perspectiva permitiría ofrecer recomendaciones más ajustadas a las necesidades de distintos segmentos, como familias jóvenes, parejas sin hijos o estudiantes, facilitando además la creación de restricciones dentro del modelo. Complementariamente, se proyecta la incorporación de herramientas de optimización con *Gurobi*, que dispone de bibliotecas integradas como "*Gurobi-ML*", las cuales permiten incluir modelos de *machine learning*. Esto contribuirá a generar configuraciones de diseño y remodelación más eficientes bajo restricciones presupuestarias y técnicas. Con esta información, se espera perfeccionar el modelo y avanzar hacia el objetivo de consolidar una metodología más adaptable y precisa, capaz de maximizar la rentabilidad y la satisfacción de los usuarios en el mercado de viviendas.

4.2. Conclusión

El problema de la tasación de viviendas tiene una alta relevancia debido a que la inexactitud en la valoración genera desconfianza, ineficiencias y pérdidas económicas tanto para los compradores como para vendedores. Junto con ello, la toma de decisiones frente a diseño y remodelación de las viviendas presentan un desafío importante debido a restricciones presupuestarias y el desafío de optimizar la inversión para lograr mayor rentabilidad. En escenarios como el de Ames, Iowa, donde la alta demanda habitacional y diversidad de perfiles de compradores amplifican la complejidad del mercado, se vuelve crítico contar con modelos predictivos y de optimización que superen las limitaciones de los métodos tradicionales.

La etapa de depuración y análisis de los datos resultó ser fundamental para asegurar un respaldo sólido al estudio del caso presentado. A través de un riguroso proceso de limpieza, imputación de valores y ajuste de variables, se logró conformar una base de datos coherente y libre de inconsistencias, lo que permitió aplicar una regresión lineal ajustada de manera fiel a la realidad. Este trabajo garantizó un conjunto de datos más representativo y confiable, en el cual las relaciones entre variables se manifestaron con mayor claridad y sentido práctico. Además, el proceso permitió reducir la influencia de sesgos y valores atípicos que podrían haber distorsionado los resultados. Así, el caso base definido constituye un punto de partida consolidado para el modelamiento futuro.

El uso de técnicas avanzadas como *XGBoost*, no solo nos permitirá mejorar la precisión en la estimación de precios, sino también orientar decisiones de diseño y remodelación bajo criterios de eficiencia y rentabilidad. De este modo, se fomenta un mercado más transparente, equitativo y alineado con las necesidades reales de los actores involucrados, posicionando este desafío como un problema de gran impacto social y económico cuya solución es capaz de aportar a la sostenibilidad y modernización del sector inmobiliario. Es por ello por lo que avanzar hacia herramientas de predicción y optimización más robustas no es solo una necesidad actual, sino una condición indispensable para el desarrollo sostenible del mercado inmobiliario en el mediano y largo plazo.

BIBLIOGRAFÍA

AppraisersBlogs. (2015, mar 6). *Appraisal bias and appraiser pressure: Why all appraisals are always wrong.* AppraisersBlogs. Retrieved from https://appraisersblogs.com/appraisal/appraisal-bias-and-appraiser-pressure-why-all-appraisals-are-always-wrong/

Cambridge University Press. (n.d.). *Black box.* In Cambridge Dictionary. Retrieved from https://dictionary.cambridge.org/dictionary/english/black-box

Chandra, R. (2023). *Regresión lasso vs regresión ridge en r - ¡explicado!* Kanaries. Retrieved from https://docs.kanaries.net/es/topics/R/lasso-regression-r

Chicco, D. (2021). The coefficient of determination r² is more informative than smape, etc. *Frontiers in Oncology*, *11*, 720963. Retrieved from https://doi.org/10.3389/fonc.2021.720963 doi: 10.3389/fonc.2021.720963

Data Science & Beyond. (2023, oct 3). *Choosing the right correlation: Pearson vs. spearman vs. kendall's tau.* https://ishanjainoffical.medium.com/choosing-the-right-correlation-pearson-vs-spearman-vs-kendalls-tau-02dc7d7dd01d. (Recuperado el 12 de septiembre de 2025)

De Cock, D. (2011). Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, *19*(3). Retrieved from https://doi.org/10.1080/10691898.2011.11889627 doi: 10.1080/10691898.2011.11889627

Dunaway, J. (2024, oct 7). *New data: Home renovation trends in 2024.* Clever. Retrieved from https://listwithclever.com/research/home-renovation-trends/

Evans, K., Lausberg, C., & Sui Sang How, J. (2019). Reducing property appraisal

bias with decision support systems: An experimental investigation in the south african property market. *Journal of African Real Estate Research*, *4*(1), 108-138. Retrieved from http://dx.doi.org/10.15641/jarer.v4i1.729 doi: 10.15641/jarer.v4i1.729

Geerts, M., vanden Broucke, S., & De Weerdt, J. (2023). A survey of methods and input data types for house price prediction. *ISPRS International Journal of Geo-Information*, *12*(5), 200. Retrieved from https://doi.org/10.3390/ijgi12050200 doi: 10.3390/ijgi12050200

Geltner, D., Miller, N. G., Clayton, J., & Eichholtz, P. (2014). *Commercial real estate: Analysis and investments* (3rd ed.). OnCourse Learning. Retrieved from https://books.google.com/books/about/Commercial_Real_Estate.html?id=3z2onzivNuAC

Harrison, J., D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81–102. Retrieved from https://doi.org/10.1016/0095-0696(78)90006-2 doi: 10.1016/0095-0696(78)90006-2

Hodson, T. O., Archfield, S. A., Kiang, J. E., Over, T. M., Farmer, W. H., & Hay, L. E. (2022). Root-mean-square error (rmse) or mean absolute error (mae): When to use them or not. *Geoscientific Model Development*, *15*(15), 5481–5487. Retrieved from https://doi.org/10.5194/gmd-15-5481-2022 doi: 10.5194/gmd-15-5481-2022

Ibm-a. (2021, sep 28). *What is boosting?* ibm. Retrieved from https://www.ibm.com/think/topics/boosting

Ibm-b. (2025, jan 30). *Arboles de decisión.* ibm. Retrieved from https://www.ibm.com/es-es/think/topics/decision-trees

Kavlakoglu, E., & Russi, E. (2024, may 9). *What is xgboost?* IBM. Retrieved from https://www.ibm.com/think/topics/xgboost?mhsrc=ibmsearch_a&mhq=xgboost

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent

demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679. Retrieved from https://doi.org/10.1016/j.ijforecast.2015.12.003 doi: 10.1016/j.ijforecast.2015.12.003

Macek, D., & Vitásek, S. (2024). Risk analysis in building renovations: Strategies for risk mitigation in construction engineering. *Buildings*, *14*(7), 2219. Retrieved from https://doi.org/10.3390/buildings14072219 doi: 10.3390/buildings14072219

Marcelino, P. (2017). *Comprehensive data exploration with python.* https://www.kaggle.com/code/pmarcelino/comprehensive-data-exploration-with-python. (Accedido el 12 de septiembre de 2025)

Mistry, M., Letsios, D., Krennrich, G., Lee, R. M., & Misener, R. (2018). *Mixed-integer convex nonlinear optimization with gradient-boosted trees embedded.* arXiv. Retrieved from https://arxiv.org/abs/1803.00952

Nadarajah, S. (2025). Empirical calibration of xgboost model hyperparameters. *Journal of Risk and Financial Management*, *18*(9), 487. Retrieved from https://doi.org/10.3390/jrfm18090487 doi: 10.3390/jrfm18090487

National Association of Realtors. (2025). *2025 home buyers and sellers generational trends report.* National Association of Realtors. Retrieved from https://www.nar.realtor/research-and-statistics/research-reports/home-buyer-and-seller-generational-trends

Peppard, J., & Ward, J. (2016). *The strategic management of information systems: Building a digital strategy* (4th ed.). Wiley. Retrieved from https://vdoc.pub/download/the-strategic-management-of-information-systems-building-a-digital-strategy-1oh6pds1ltg8

Poeta, S., Gerhardt, T., & Stumpf Gonzalez, M. (2019). Análisis de precios hedónicos de viviendas. *Revista ingeniería de construcción*, *34*(2), 215–220. Retrieved from https://dx.doi.org/10.4067/S0718-50732019000200215 doi: 10.4067/S0718-50732019000200215

Sharma, H., Harsora, H., & Ogunleye, B. O. (2024). An optimal house price prediction algorithm: Xgboost. *Analytics*, *3*(1), 3045. Retrieved from https://doi.org/10.3390/analytics3010003 doi: 10.3390/analytics3010003

Ye, Q. (2024). House price prediction using machine learning for ames, iowa. *Applied and Computational Engineering*, *55*, 44–54. Retrieved from http://dx.doi.org/10.54254/2755-2721/55/20241483 doi: 10.54254/2755-2721/55/20241483

Yun, L., & Lautz, J. (2025). *2025 remodeling impact report.* National Association of REALTORS® Research Group. Retrieved from https://www.nari.org/NARI/media/Assets/2025-Remodeling-Impact-Report_Final-4-9-25.pdf

Özdemir, O. (2022). *House price prediction using machine learning: A case in iowa.* ResearchGate. Retrieved from https://doi.org/10.13140/RG.2.2.19846.86086 doi: 10.13140/RG.2.2.19846.86086

ANEXO

ANEXO 1. TABLAS

Table .1. Resumen de valores nulos y correcciones aplicadas.

Variable / Par de variables	**Problema**	**Decisión**	**Justificación**
GarageYrBlt	Datos faltantes cuando no hay *garage*.	Reemplazar por 0.	Consistente: si no hay *garage*, nunca se construyó.
GarageYrBlt	Error de *input*: 2207 en vez de 2007.	Corregido a 2007.	Basado en año de construcción/remodelación reportado en la literatura.
LotFrontage	490 datos faltantes.	Imputación por mediana de vecindario; 3 filas sin información $\rightarrow$ eliminadas.	En viviendas contiguas la línea de frente es similar; la imputación por mediana de *neighborhood* está respaldada en la literatura.
Electrical	1 fila con dato faltante.	Fila eliminada.	Impacto mínimo en el *dataset*.
MasVnrArea y *MasVnrType*	23 datos faltantes y alta colinealidad.	Eliminadas.	Información contenida por otras variables (p. ej., *YearBuilt* y *OverallQual*); se reduce colinealidad.
GrLivArea > 4000	5 *outliers* extremos.	Filas eliminadas.	No representativos; generan distorsión en el ajuste.

Variable / Par de variables	**Problema**	**Decisión**	**Justificación**
Bsmt Half Bath y *Bsmt Full Bath*	NA que significa "no tiene".	Reemplazo por 0.	Consistente con la definición del *dataset* (0 cuando no existe el ítem).
Garage (1 fila)	Falta de información en variables sobre el *garage*.	Fila eliminada.	No hay información suficiente (excepto *GarageType*); no es posible imputar de forma confiable.
Bsmt (5 filas)	Variables del sótano con datos faltantes imposibles de inferir.	Filas eliminadas.	Impacto en el precio promedio $\approx$ 30 USD; efecto insignificante y evita sesgo de imputación.

Table .2. Depuración y decisiones sobre variables numéricas.

Variable / Par de variables	**Problema**	**Correlación con *SalePrice***	**Decisión**	**Justificación**
Bsmt Fin SF 2	Muchos *outliers* (347) y baja correlación con *SalePrice*.	0,006	Eliminarla	Aporta poco valor explicativo y agrega ruido. En el cuerpo ya se muestra que de todas formas se elimina por redundancia.

Variable / Par de variables	**Problema**	**Correlación con *SalePrice***	**Decisión**	**Justificación**
Total Porch = *Screen Porch* + *3Ssn Porch* + *Open Porch SF* + *Enclosed Porch*	Baja correlación individual con *SalePrice* y la información queda dividida entre variables del mismo espacio.	Screen Porch: 0,12 3Ssn Porch: 0,03 Open Porch SF: 0,32 Enclosed Porch: −0,12	Sumar y crear **Total Porch**	El total es más relevante; se reduce multicolinealidad y doble conteo entre porches.
Gr Liv Area vs *1st Flr SF* + *2nd Flr SF*	Combinación lineal; ambas suman *Gr Liv Area*.	Gr Liv: 0,72 1st Flr: 0,64 2nd Flr: 0,26	Mantener *Gr Liv Area*	Resumen más completo; mayor correlación con el precio. Disminuye redundancia.
Gr Liv Area vs *TotRms AbvGrd*	Alta correlación (0,81).	TotRms: 0,49	Mantener *Gr Liv Area*	Es la variable numérica más correlacionada con *SalePrice*; *TotRms* es menos informativa.
Mas Vnr Area y *Mas Vnr Type*	Alta correlación con otras variables; además 23 faltantes.	—	Eliminar ambas	No son esenciales y se solapan con *Year Built* y *Overall Qual*; no se pierde información relevante.

Variable / Par de variables	**Problema**	**Correlación con *SalePrice***	**Decisión**	**Justificación**
Garage Area vs *Garage Cars*	Correlación alta (0,86).	Area: 0,64 Cars: 0,65	Mantener *Garage Cars*; eliminar *Garage Area*	Misma información práctica; *Garage Cars* es más interpretable y correlaciona levemente mejor.
Misc Val	Baja correlación y 101 *outliers*.	$-0{,}01$	Eliminarla	Poco valor explicativo; nos quedamos con la categórica *Misc Feature*.
Pool Area	Baja correlación; ligada a *Pool QC*.	0,04	Eliminarla	*Pool QC* ya capta presencia/calidad; *Pool Area*=0 suele implicar "no tiene piscina".
Mo Sold y *Yr Sold*	Baja correlación con *SalePrice*; poca relevancia para el objetivo.	Mo Sold: 0,04 Yr Sold: $-0{,}03$	Eliminar ambas	Mes y año de venta no son variables influyentes en este proyecto.

Table .3. Resumen de decisiones de depuración para variables cualitativas.

Par / Variable	**Problema detectado**	**Decisión**	**Justificación**
Street	99,6% con *Pave*	Eliminada	Variable prácticamente constante; no aporta información.
Utilities	99% *AllPub*	Eliminada	Sin variabilidad; literatura recomienda descartar marcelino2016.
Condition2	99% *Norm*	Eliminada	Constante en casi todo el *dataset*.
RoofMatl	98% *CompShg*	Simplificada a binaria (0/1)	Consistencia; evita categorías con muy baja frecuencia.
PoolQC	99,6% "No tiene"	Simplificada a binaria (0/1)	Mantiene presencia/ausencia de piscina; la escala de calidad no es representativa.
Alley	Muchos NA (sin *alley*)	Simplificada a binaria (0/1)	Presencia/ausencia mantiene información esencial.
Fence	Muchos NA (sin *fence*)	Simplificada a binaria (0/1)	Consistencia; evita categorías muy infrecuentes.
MiscFeature	Muchos NA (sin *misc*)	Simplificada a binaria (0/1)	Misma razón anterior.
PavedDrive	Codificada Y/N	Simplificada a binaria (0/1)	Versión binaria más interpretable.

Continúa en la página siguiente

(Continúa de la página anterior)

Par / Variable	**Problema detectado**	**Decisión**	**Justificación**
MSSubClass vs *BldgType*	Cramér's $V = 0{,}88$	Mantener `MSSubClass`	Mayor relevancia para *SalePrice*.
MSSubClass vs *HouseStyle*	Cramér's $V = 0{,}83$	Mantener `MSSubClass`	Literatura respalda eliminar `HouseStyle` cock2011,marcelino2016.
Exterior1st vs *Exterior2nd*	Cramér's $V = 0{,}74$	Mantener `Exterior1st`	En Kaggle y literatura se elimina `Exterior2nd` marcelino2016.
GarageCond vs *GarageQual*	*Spearman* $\rho \approx 0{,}77$	Mantener `GarageQual`	`GarageQual` presenta mayor correlación con *SalePrice*.
ExterQual vs *OverallQual* / *KitchenQual*	*Spearman* $\rho > 0{,}7$ con ambas	Eliminar `ExterQual`	Redundante; `OverallQual` y `KitchenQual` capturan mejor la información.

ANEXO 2. FIGURAS

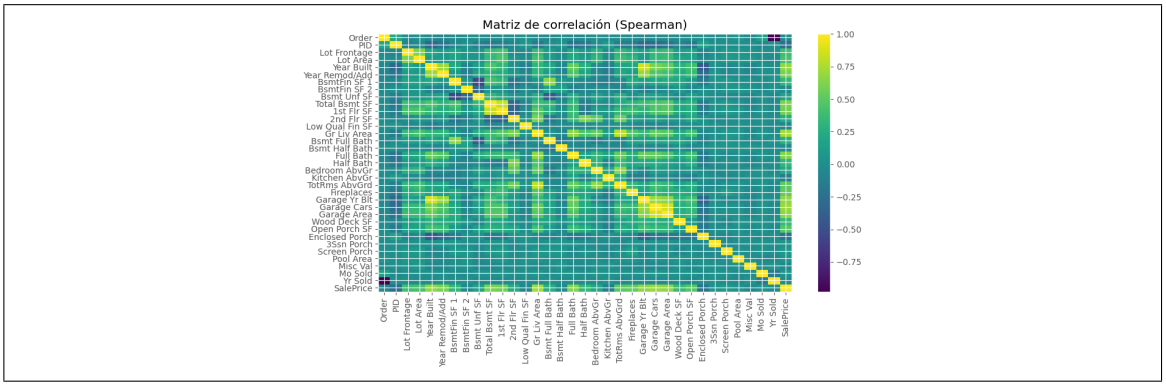

Figure .1. Matriz de correlación de *Spearman* de variables cuantitativas.

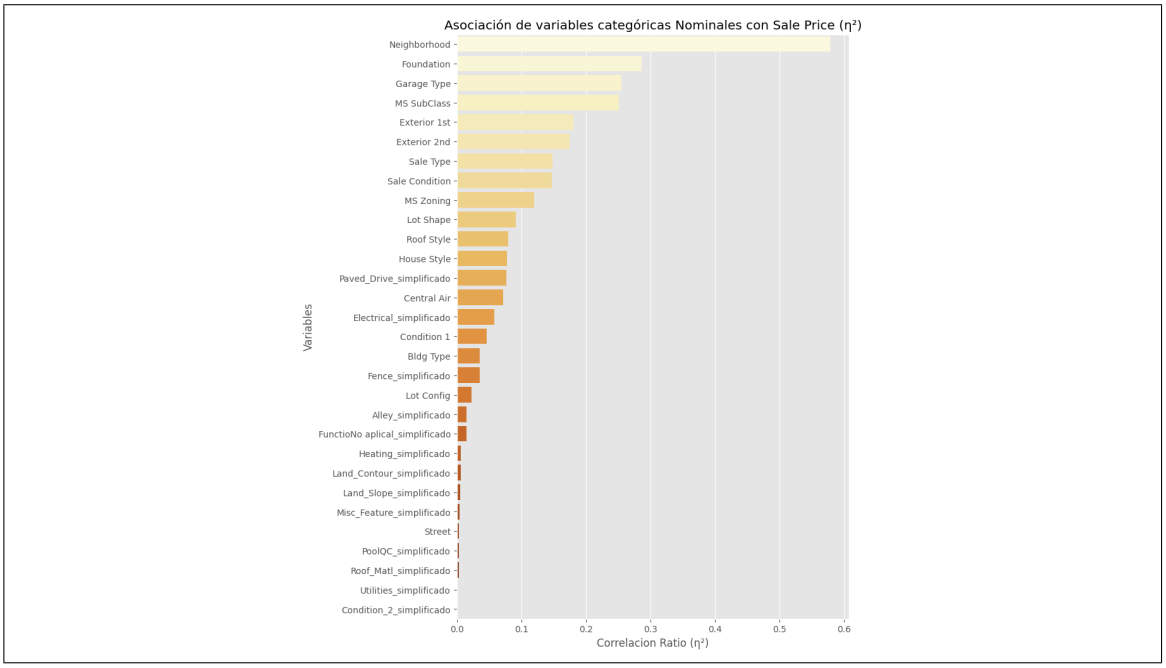

Figure .2. Asociación de variables categóricas nominales con *Sale Price*.

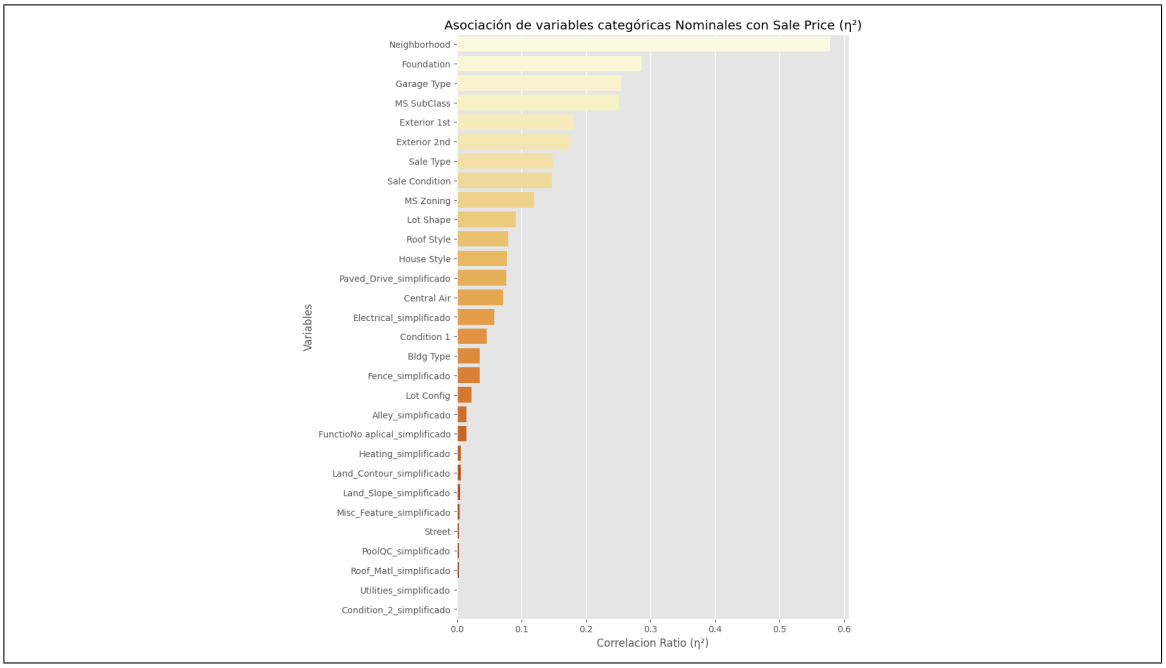

Figure .3. Asociación de variables categóricas nominales con *Sale Price*.

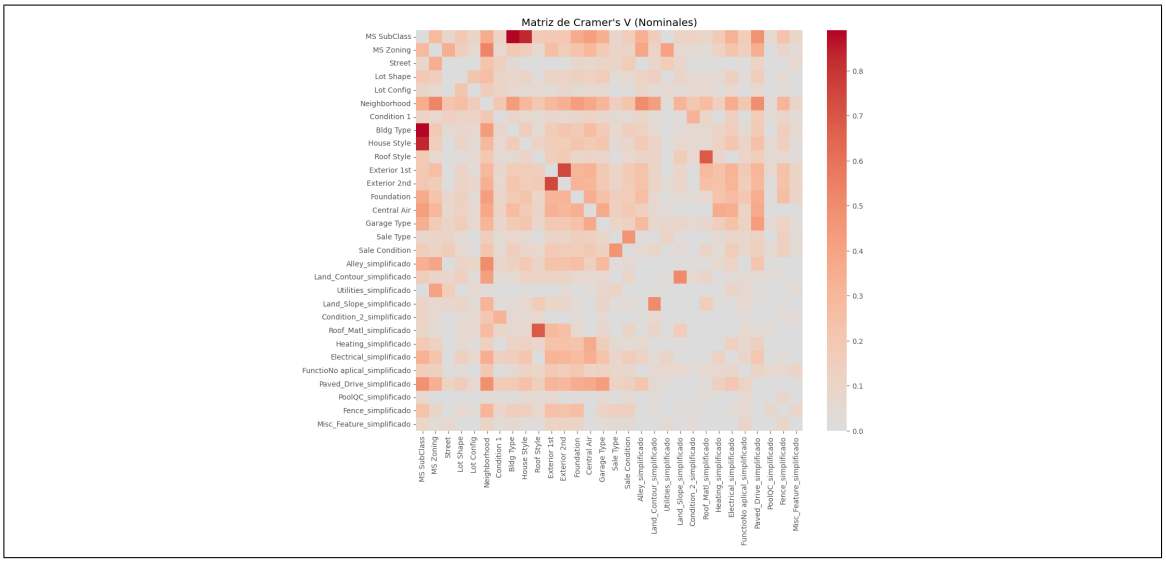

Figure .4. Matriz de Cramer's V de variables categóricas nominales.

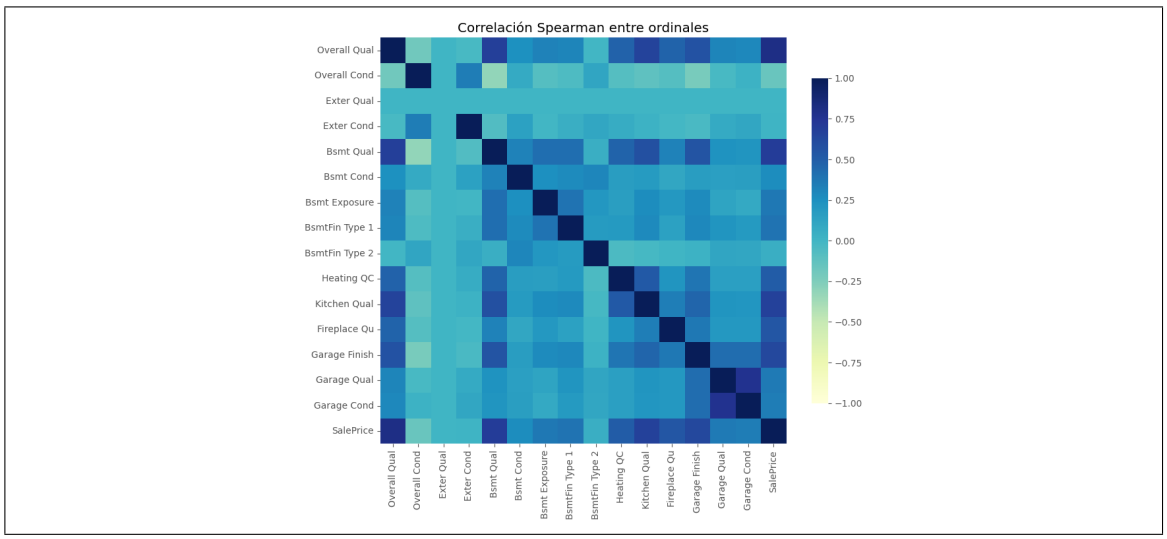

Figure .5. Matriz de correlación de Spearman de variables categóricas ordinales codificadas.

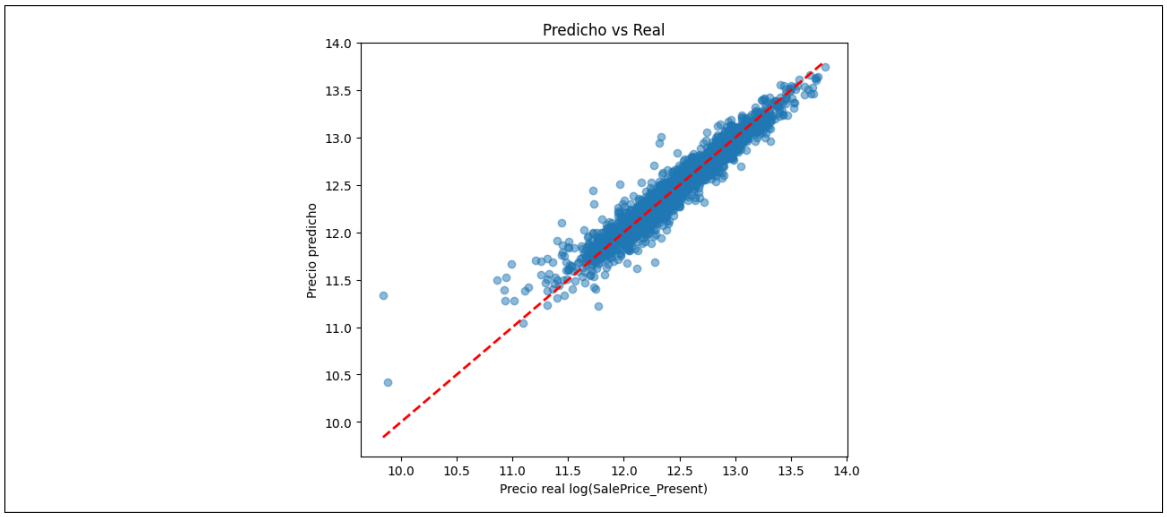

Figure .6. Regresión Lineal *log(SalePrice_present)*.

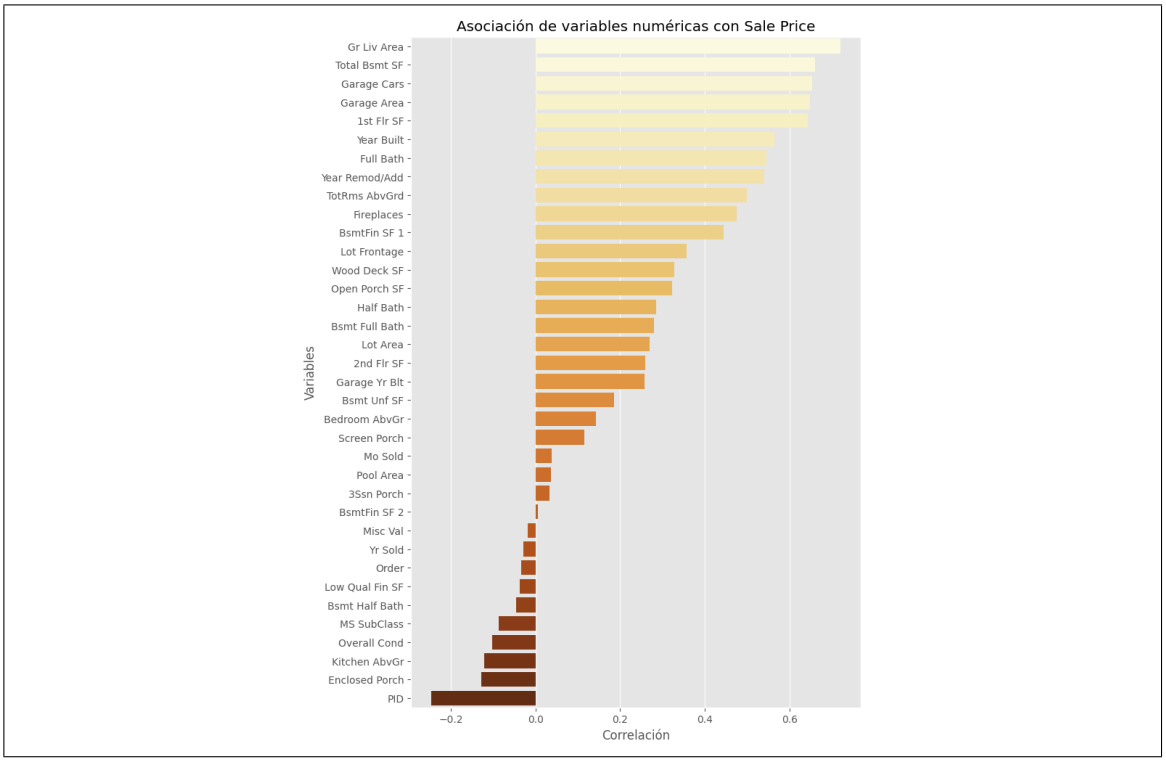

Figure .7. Asociación de variables numéricas con *SalePrice*.

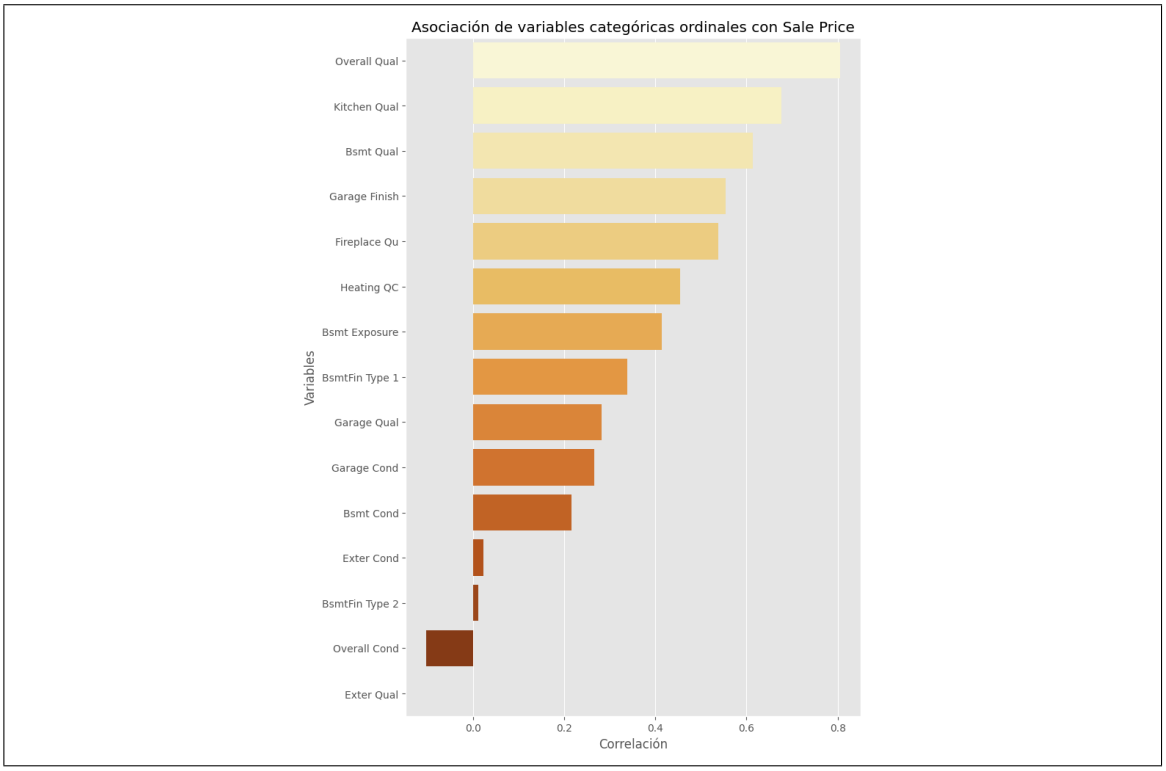

Figure .8. Asociación de variables categóricas Ordinales con *Sale Price*.

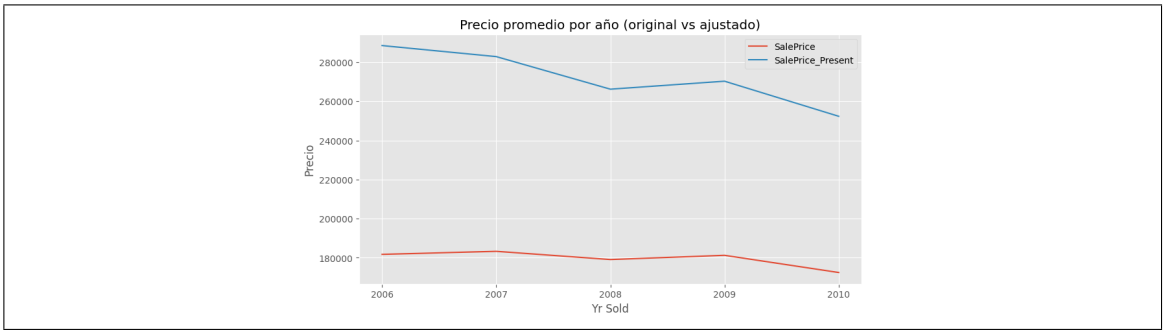

Figure .9. Gráfico promedio de *SalePrice* por año, ajustado por IPC al precio presente

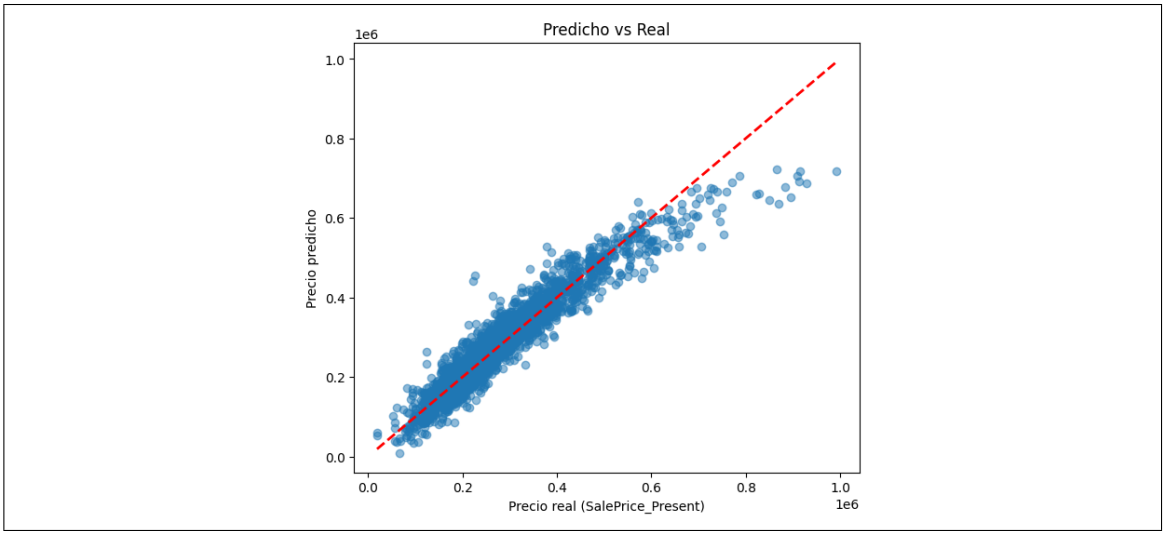

Figure .10. Regresión Lineal *SalePrice_Present*

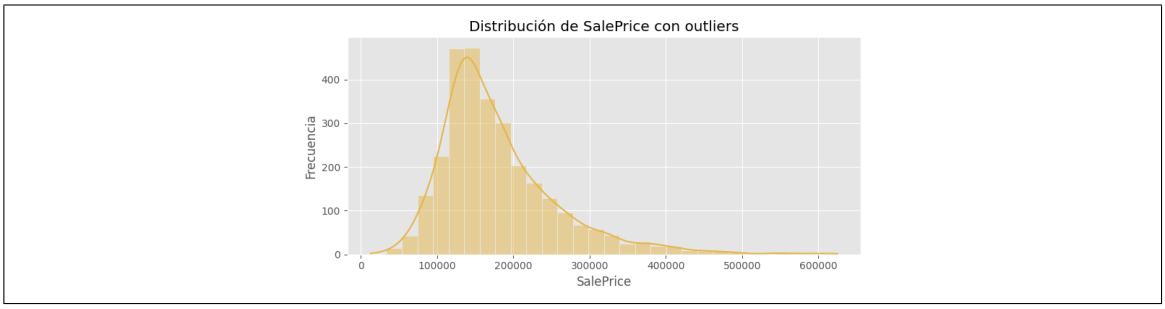

Figure .13. Distribución de *SalePrice* con *outliers.*

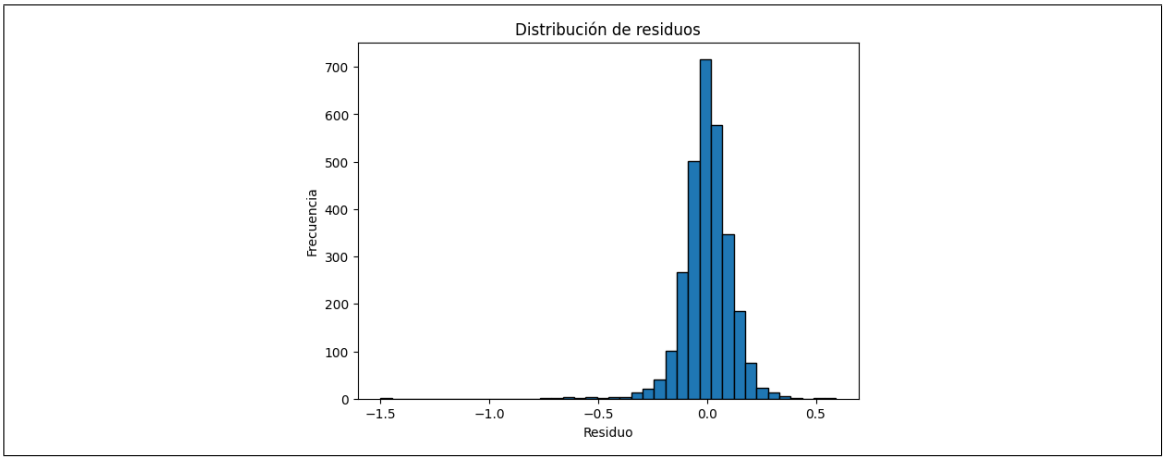

Figure .11. Distribución de residuos al aplicar *log(SalePrice_Present)*

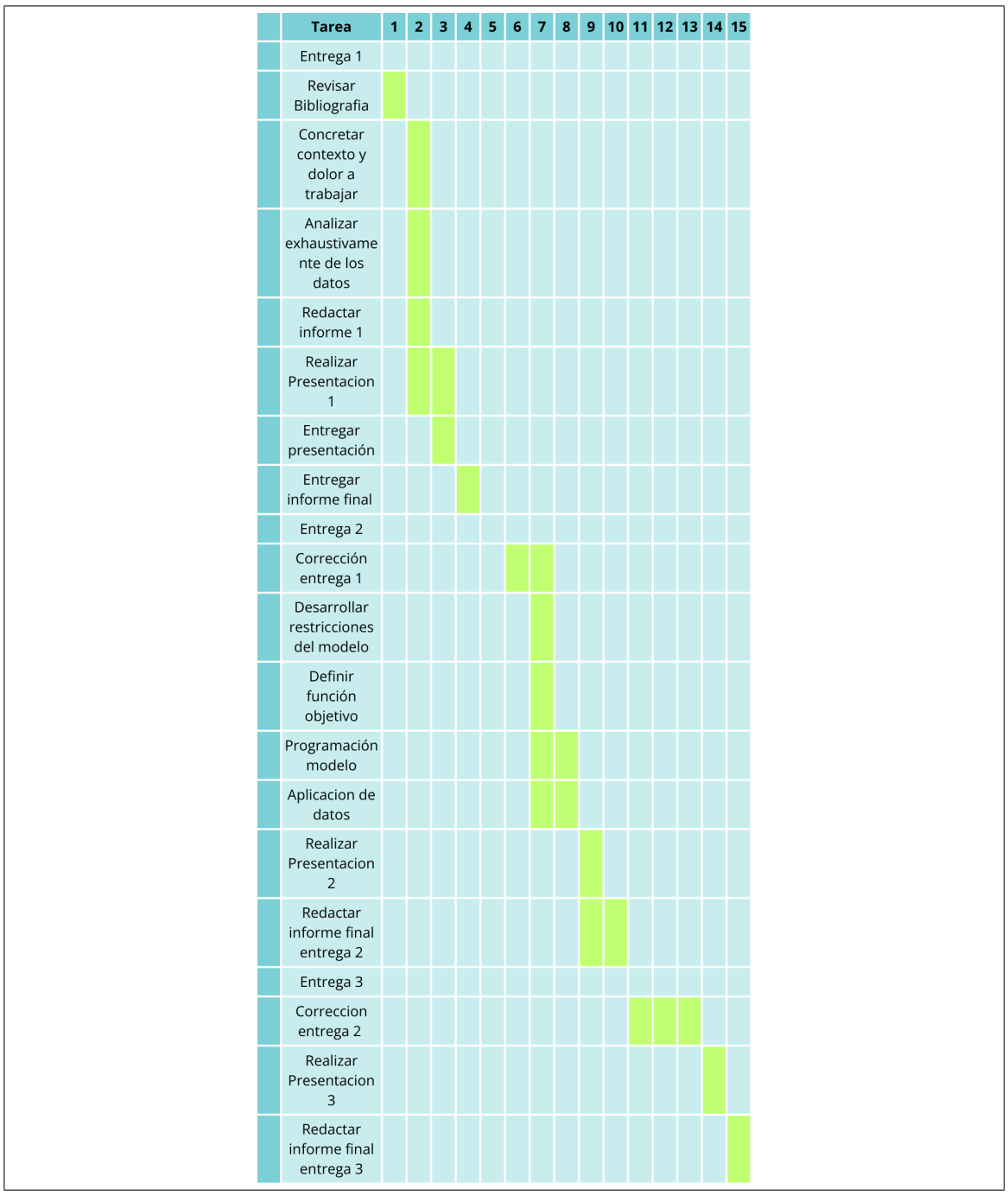

Figure .12. Carta Gantt