Universidad del Rosario

Analisis Avanzado de Datos

W10. Métodos de Suavización: Kernel multiples caracteristicas

FERNEY ALBERTO BELTRAN MOLINA
Escuela de Ingeniería, Ciencia y Tecnología
Matemáticas Aplicadas y Ciencias de la Computación

En el laboratorio

Objetivo:

Encontrar la casa más similar en el conjunto de datos y usar su valor como predicción.

Pasos:

- Especificar un punto de consulta (e.g., una casa).
- Proporcionar un conjunto de datos de entrenamiento.
- Definir una métrica de distancia para medir similitud.
- Devolver la casa más similar como resultado.

Observaciones:

- Funciona bien con datos densos.
- Problemas al interpolar en regiones con pocos datos.
- Sensible al ruido en los datos, lo que puede llevar a sobreajuste.

En el laboratorio

K-Vecinos Más Cercanos (K-NN) Formalmente:

- Se buscan los k puntos más cercanos al punto de consulta.
- La predicción se basa en el promedio de las salidas de estos k vecinos.

Implementación:

- Se mantiene una cola ordenada de los k vecinos más cercanos.
- Se recorren todas las observaciones y se compara la distancia al punto de consulta.
- Si una nueva casa es más cercana que el vecino más lejano actual, se actualiza la cola.

Predicción:

- Una vez obtenidos los k vecinos más cercanos, se promedian sus valores para obtener la predicción..

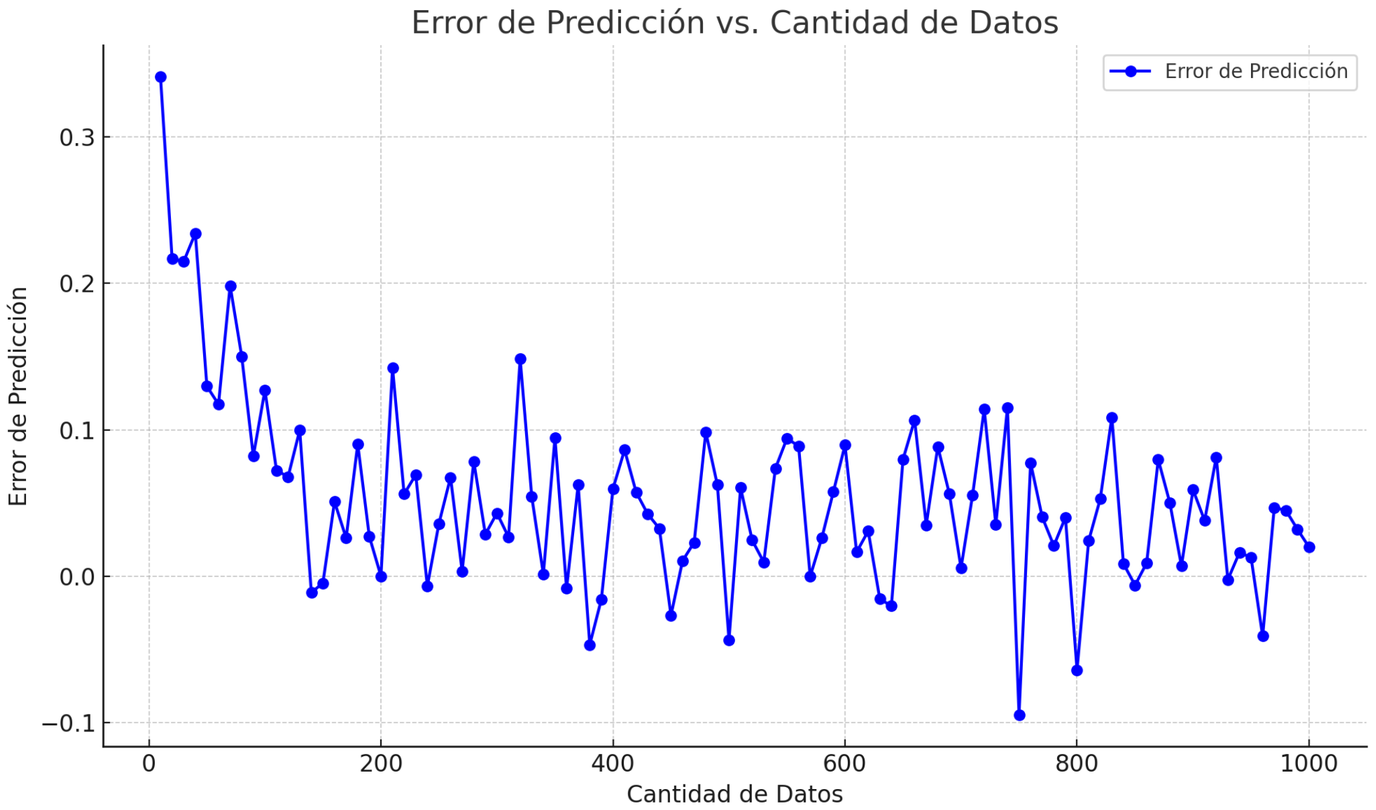
Kernel

Sin ruido:

- Para KNN, el error cuadrático medio tiende a cero con más datos.
- Esto significa que el sesgo y la varianza disminuyen.

Con ruido:

- KNN con k pequeño es sensible al ruido.
- KNN con k grande puede suavizar el ruido.

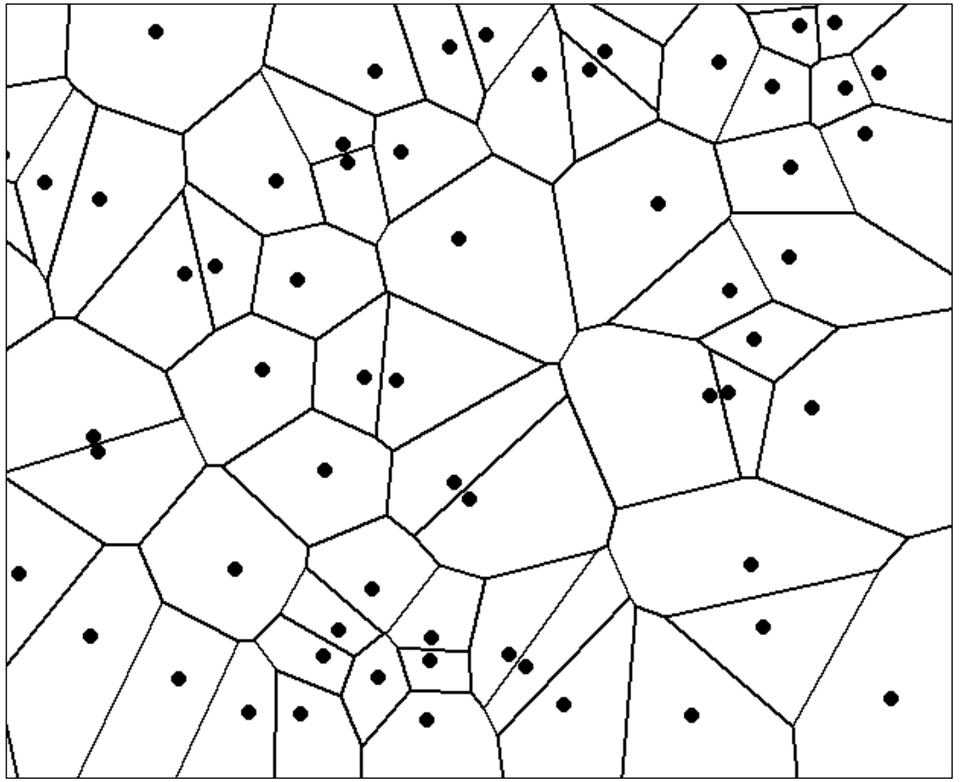


Distancia

Distancia en 1D:

- Uso de la distancia euclidiana.
- Diferencia absoluta entre dos puntos.

$$d = |x_1 - x_2|$$



Distancia en Dimensiones Superiores, Euclidiana Ponderada:

- Similar a la distancia euclidiana, pero con pesos para cada dimensión.
- Permite dar más importancia a ciertas características sobre otras.
- Si todos los pesos son 1, se reduce a la distancia euclidiana estándar.

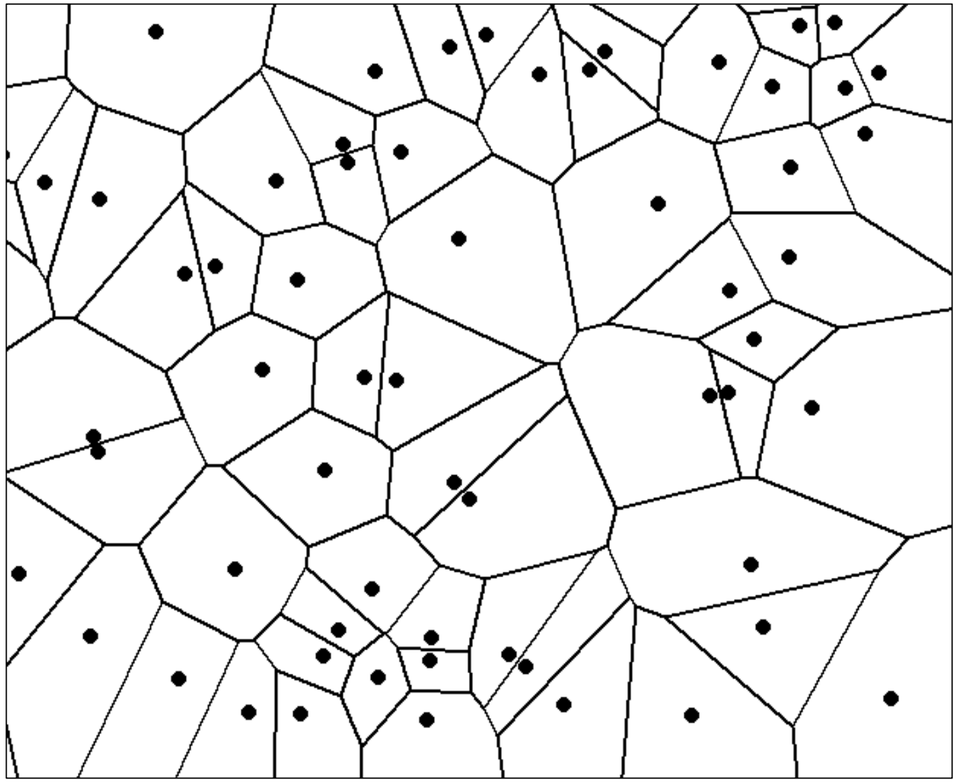
$$d = \sqrt{\sum_{i=1}^{n} a_i (x_i - y_i)^2}$$

Dado dos puntos, **x**=(*x*1,*x*2,…,*xn*) y **y**=(*y*1,*y*2,…,*yn*), en un espacio n-dimensional, y un vector de pesos **a**=(*a*1,*a*2,…,*an*), la distancia euclidiana ponderada *d* entre **x** y **y**

Distancia

Impacto de la Métrica de Distancia

- Diferentes métricas definen diferentes regiones de similitud.
- Ejemplo: Distancia de Manhattan vs. Distancia Euclidiana.
- Las regiones de similitud varían según la métrica.



En aplicaciones reales, algunas características pueden ser más importantes que otras.

Ejemplo: En viviendas, el número de habitaciones puede ser más relevante que el año de renovación.

Ajuste Global vs. Ajuste Local en Regresión

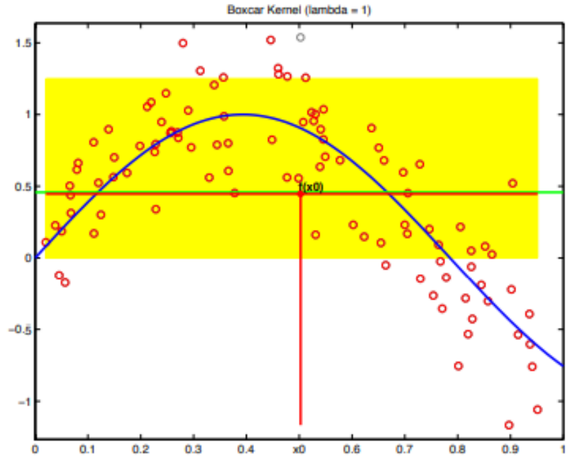
Ajuste Global:
- Se ajusta a todos los datos simultáneamente.
- Por ejemplo, la media global de todos los datos.

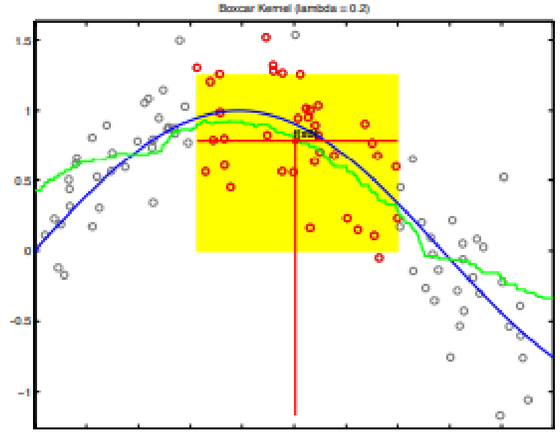
Ajuste Local (Kernel Regression):
- Se ajusta a los datos en una región local alrededor de un punto objetivo.
- Utiliza "kernels" (funciones de peso) para determinar qué datos son relevantes localmente.
- Ejemplos de kernels: Box, Epanechnikov.

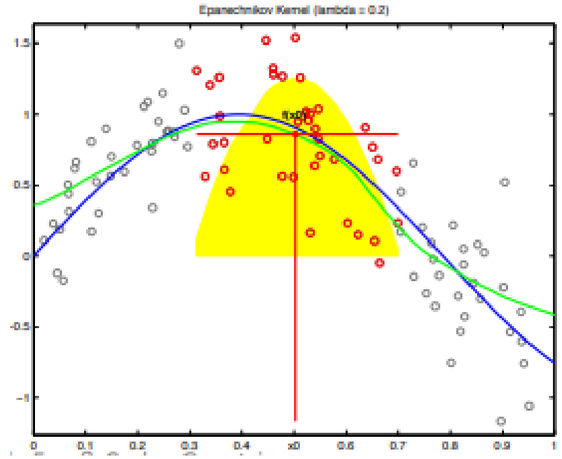
Regresión Lineal Ponderada Localmente:
- Ajuste local pero con modelos más complejos (líneas o polinomios) en lugar de constantes.
- Ayuda a reducir el sesgo en los límites y en puntos de alta curvatura.

Recomendación:
- Generalmente, se sugiere usar regresión lineal ponderada localmente para equilibrar sesgo y varianza.

- El ajuste lineal local reduce el sesgo en los límites con un mínimo aumento en la varianza, y el ajuste cuadrático local no ayuda en los límites y aumenta la varianza, pero sí ayuda a capturar la curvatura en el interior.







Resumiendo

1.Métodos No Paramétricos:

- Regresión con kernel y splines son ejemplos de estos métodos.
- Su complejidad crece con el número de observaciones.
- Objetivo: Flexibilidad en la definición de $f(x)$ y hacer mínimas suposiciones.

2.Splines:

- Los splines son funciones pieza por pieza definidas por polinomios y son útiles para modelar relaciones no lineales.
- Aportan flexibilidad local adaptándose a las características de los datos.

3.Comportamiento de la regresión con kernel:

1. En datos sin ruido, el error cuadrático medio tiende a cero.
2. En datos ruidosos, se necesita un k (número de vecinos) creciente para reducir el error.

4.Comparación con la regresión estándar:

1. Los métodos no paramétricos se adaptan localmente al ruido y pueden mejorar con más datos.
2. La regresión estándar puede tener sesgos inherentes, independientemente de la cantidad de datos

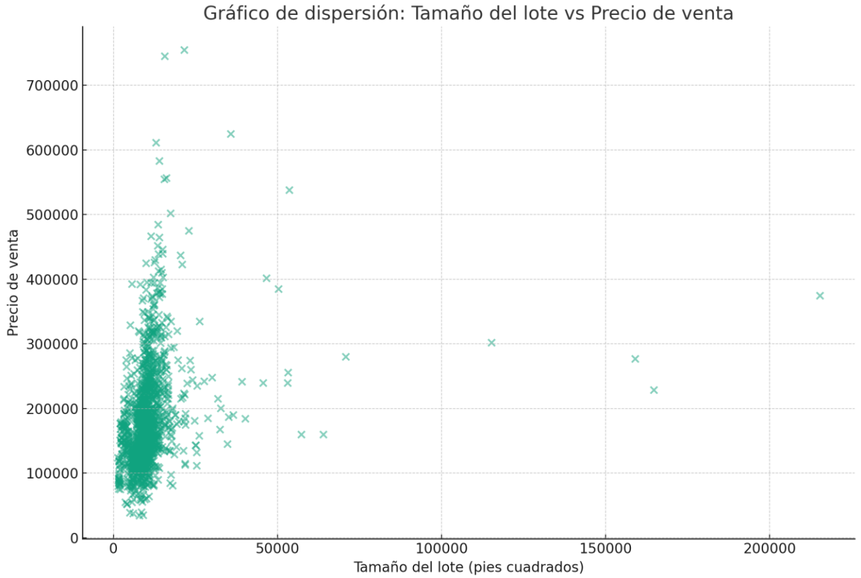
.

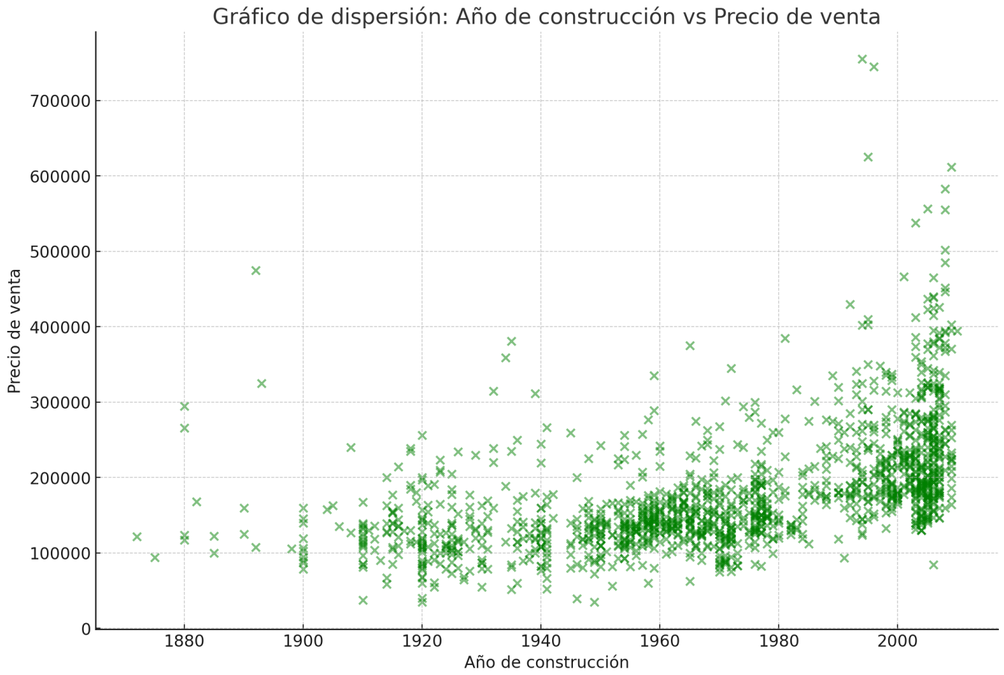
Pruebas de linealidad

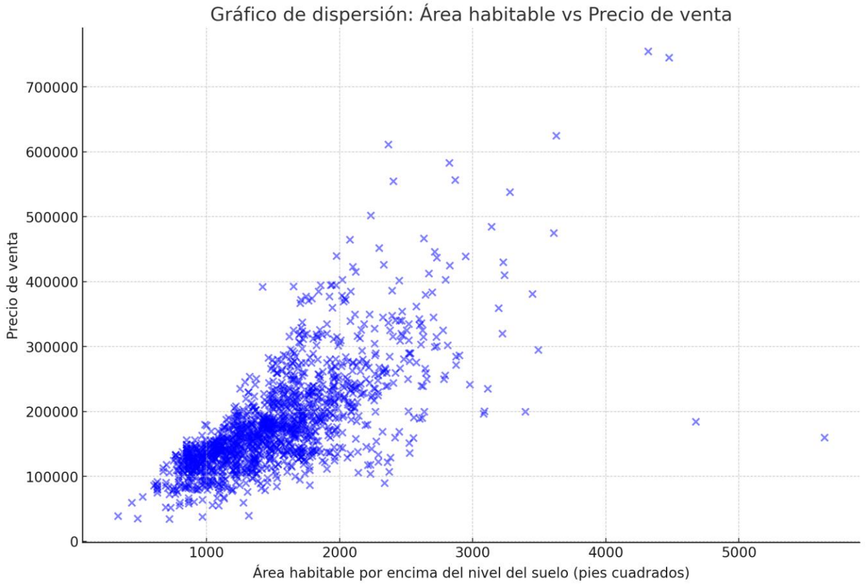
- Gráfico de dispersión (scatter plot)
- Coeficiente de correlación de Pearson
- Regresión lineal (R2 o residuos de regresión)
- Pruebas estadísticas

Pruebas de linealidad

- **Gráfico de dispersión (scatter plot)**







Pruebas de linealidad

- **Coeficiente de correlación de Pearson**

El coeficiente de correlación de Pearson solo mide relaciones lineales. Si hay una relación no lineal entre dos variables, *r* puede ser cercano a 0, incluso si hay una fuerte relación no lineal.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

r = 1: Correlación lineal positiva perfecta.
r = −1: Correlación lineal negativa perfecta.
r = 0: No hay correlación lineal.

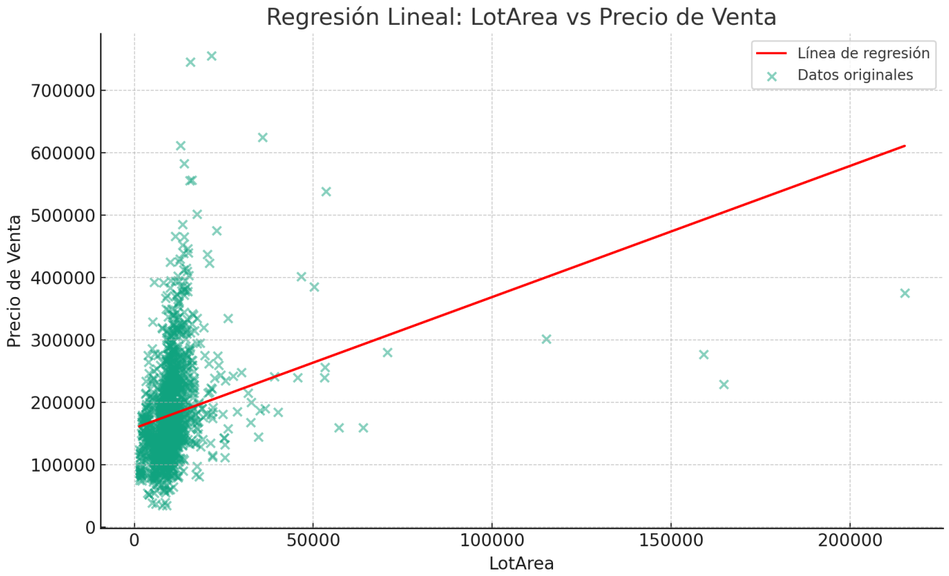
LotArea tiene una correlación débil con el precio de venta.
GrLivArea tiene una correlación fuerte con el precio de venta.
YearBuilt tiene una correlación moderada con el precio de venta

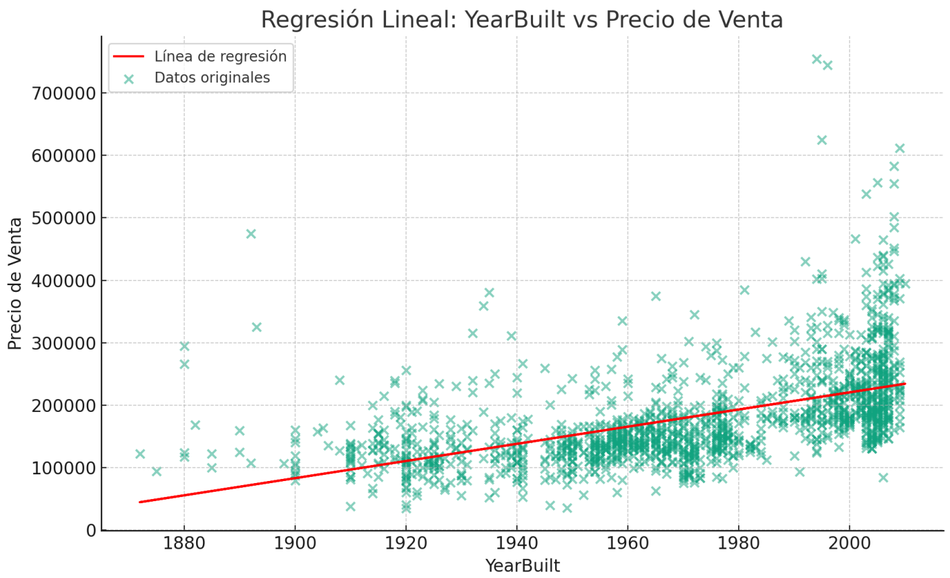
Variable	Coeficiente de Pearson
LotArea	0.264
GrLivArea	0.709
YearBuilt	0.523

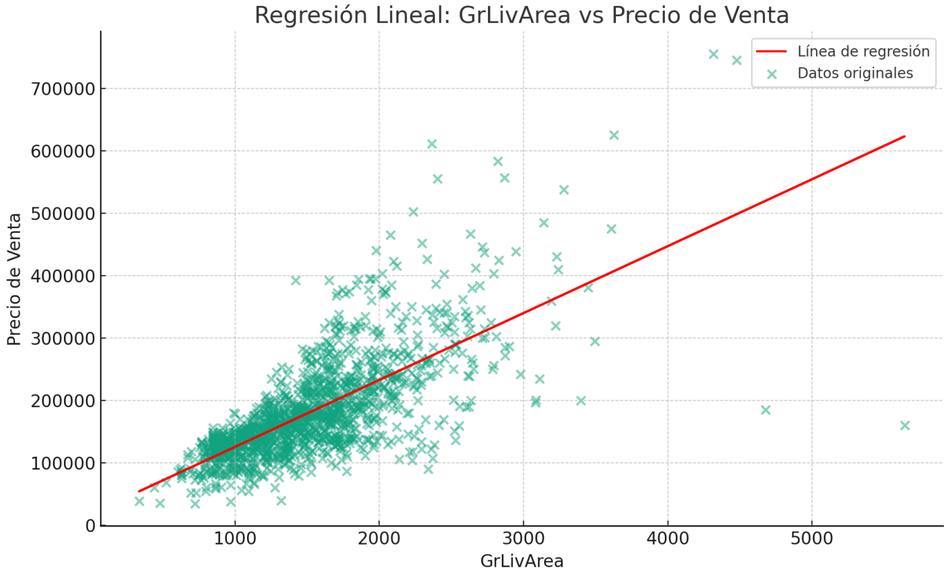
Pruebas de linealidad

- **Regresión lineal**

Variable	R2 (Regresión Lineal Simple)
LotArea	0.070
GrLivArea	0.502
YearBuilt	0.273



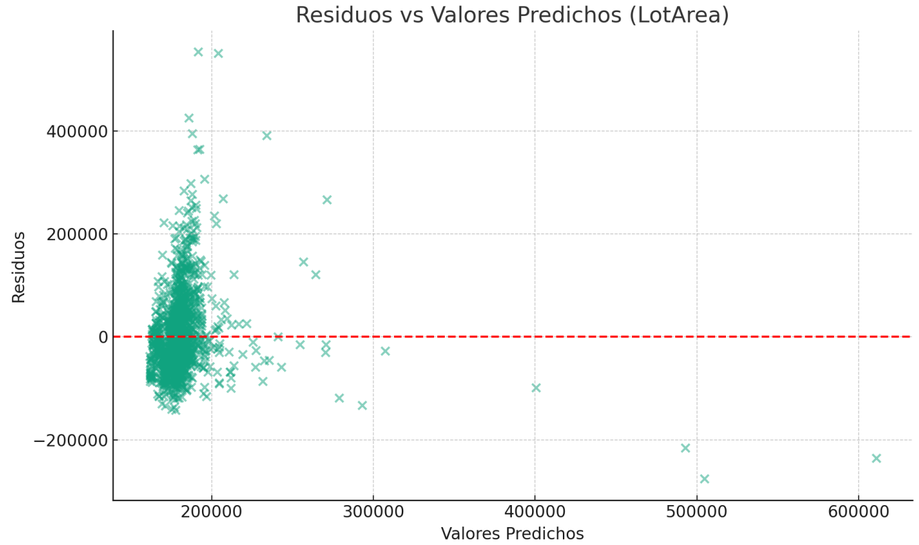


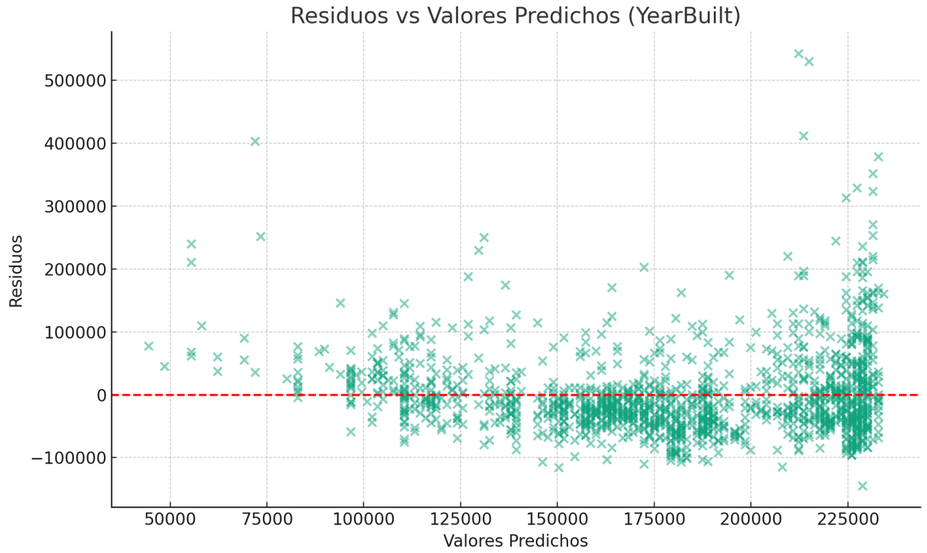


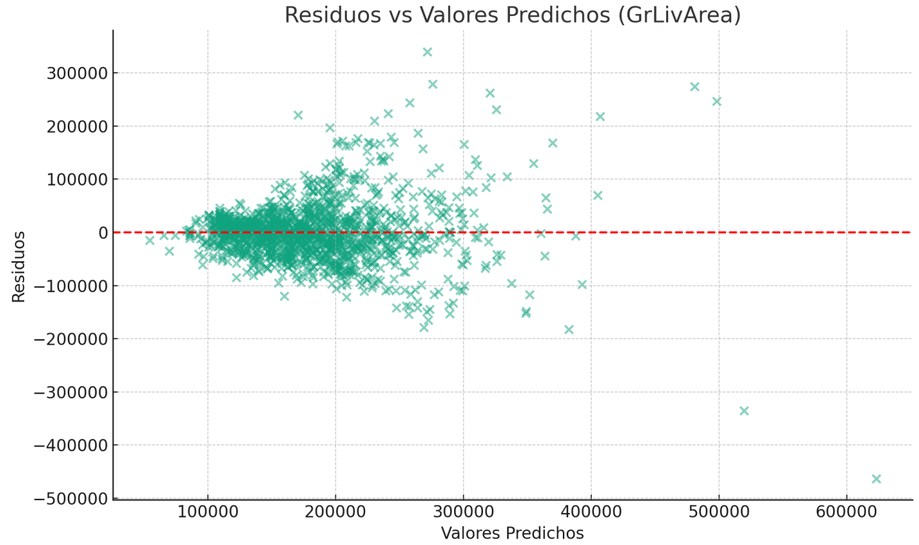
Pruebas de linealidad

- **Regresión lineal Residuos**

Residuo = valor observado − valor predicho
Si un modelo lineal es adecuado para los datos, esperaríamos que las diferencias (residuos) entre los valores observados y los valores predichos fueran simplemente errores aleatorios sin ningún patrón discernible.
los residuos deben distribuirse aleatoriamente alrededor de cero, sin mostrar ningún patrón específico







Pruebas de linealidad

- **Pruebas estadísticas**

Prueba de Harvey-Collier: Transformación recursiva de los residuos. Esta prueba evalúa la linealidad en regresiones múltiples

Prueba de Rainbow: Compara la varianza de los residuos de un modelo lineal completo con modelos lineales de subconjuntos de datos. Si estas varianzas difieren significativamente, puede ser una indicación de no linealidad.

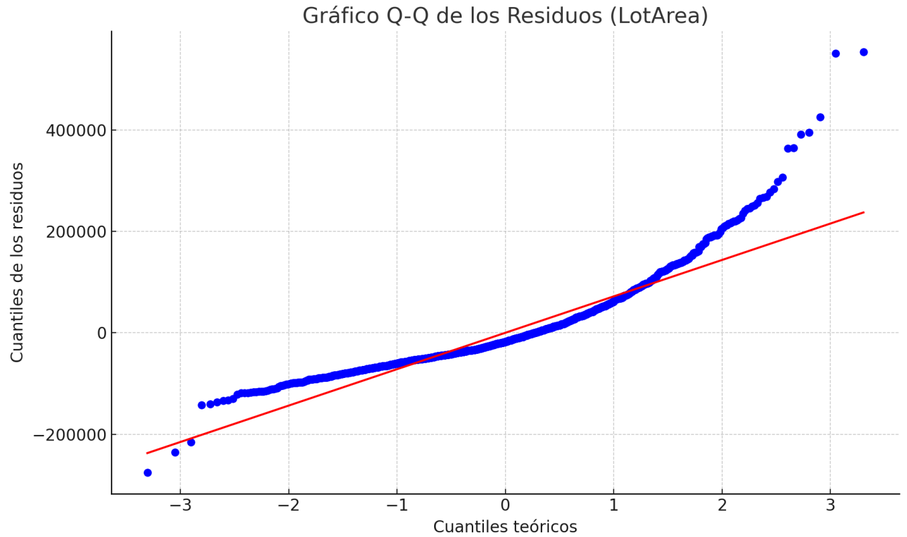
Prueba de Breusch-Pagan / Cook-Weisberg: Esta prueba está diseñada para detectar heterocedasticidad (varianza no constante de los residuos) en un modelo de regresión, pero también puede indicar no linealidad.

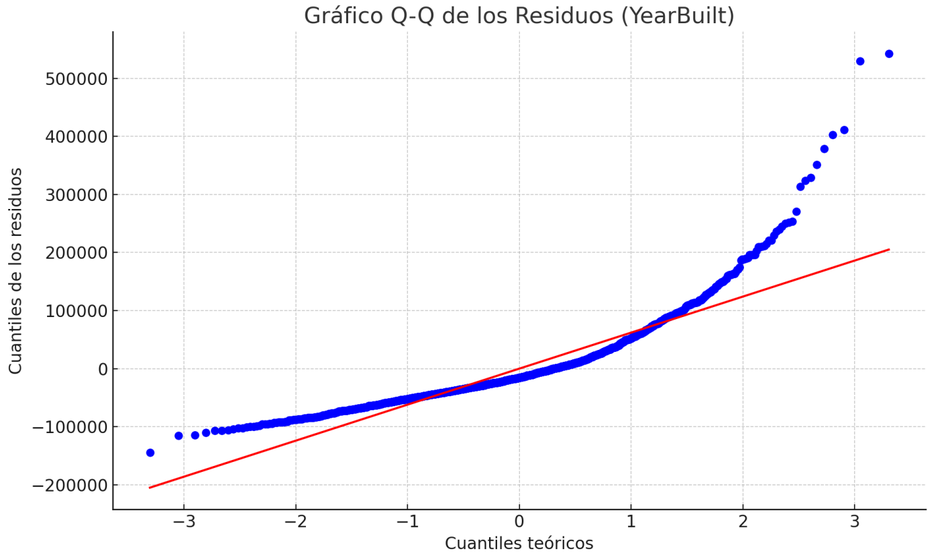
gráficos Q-Q (quantile-quantile) de residuos: Herramienta gráfica para ayudar a evaluar si un conjunto de datos sigue una distribución específica

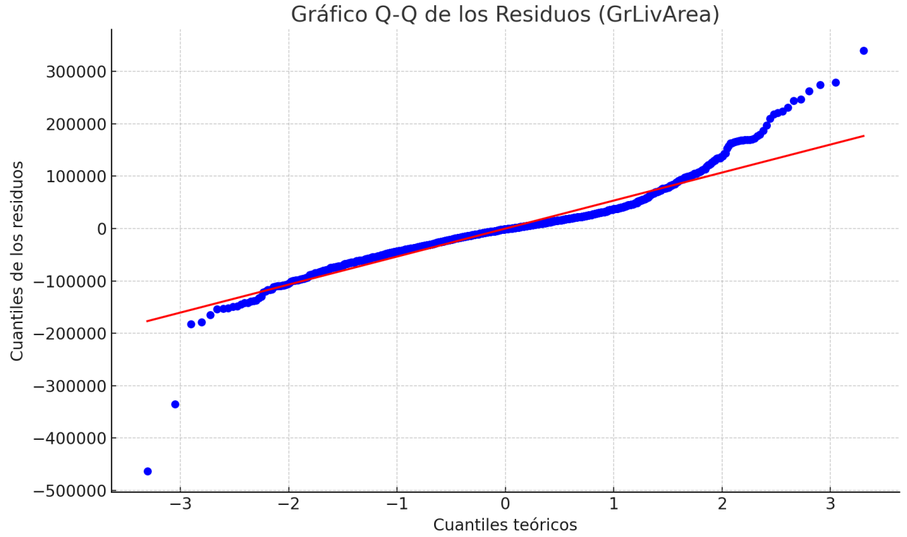
Pruebas de linealidad

Gráfico Q-Q (quantile-quantile)

En el contexto de la regresión lineal, a menudo usamos gráficos Q-Q para verificar si los residuos siguen una distribución normal.
Un gráfico Q-Q en el que los puntos caen aproximadamente en una línea recta sugiere que los residuos tienen una distribución que es aproximadamente normal. Si los puntos se desvían significativamente de esta línea, sugiere desviaciones de la normalidad.





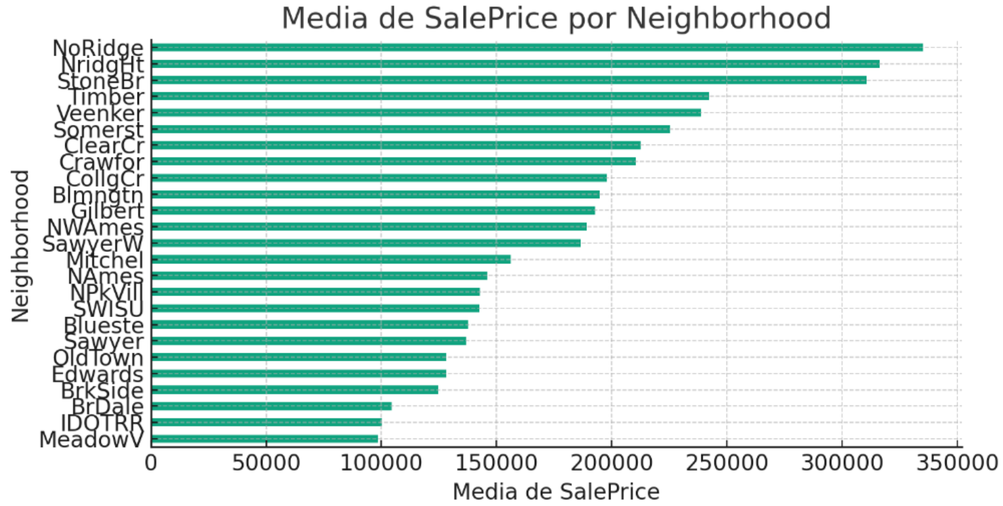


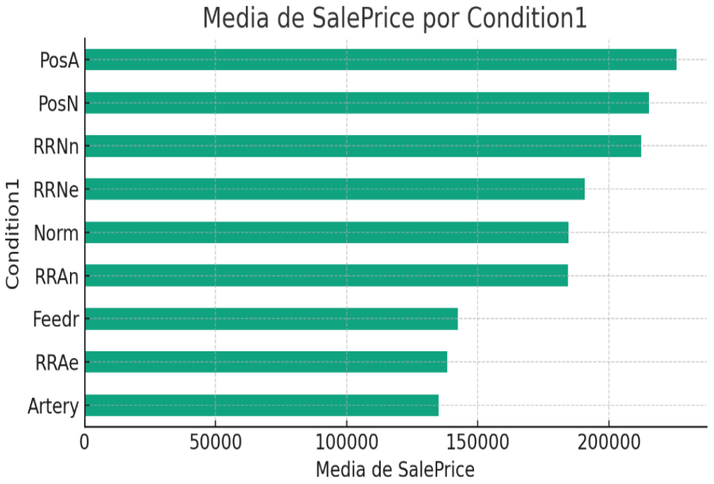
Características categóricas

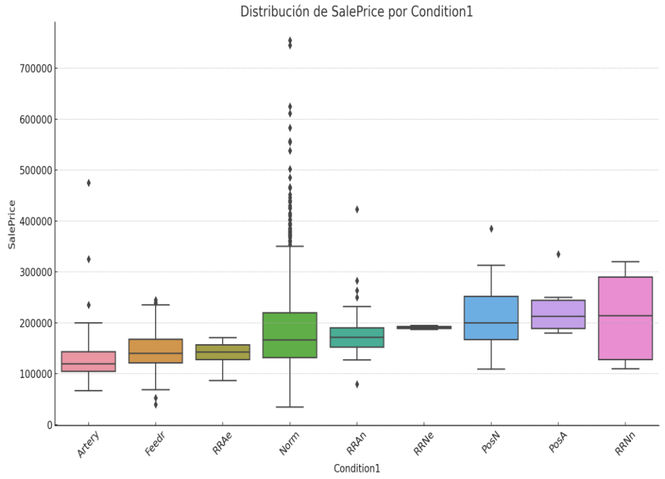
- **Codificación One-Hot (Dummy Encoding)**: Consiste en convertir cada categoría de una variable categórica en una nueva columna binaria (0 o 1)
- **Codificación Ordinal**: Si las categorías de una variable tienen un orden natural (por ejemplo, "Bajo", "Medio", "Alto"), puedes asignar un número a cada categoría basado en ese orden.
- **Codificación Binaria**: Convierte cada categoría en una combinación binaria. Por ejemplo, si hay 4 categorías, se podrían representar como 00, 01, 10 y 11.
- **Codificación de Frecuencia**: Se codifica cada categoría según su frecuencia o proporción en el conjunto de datos.

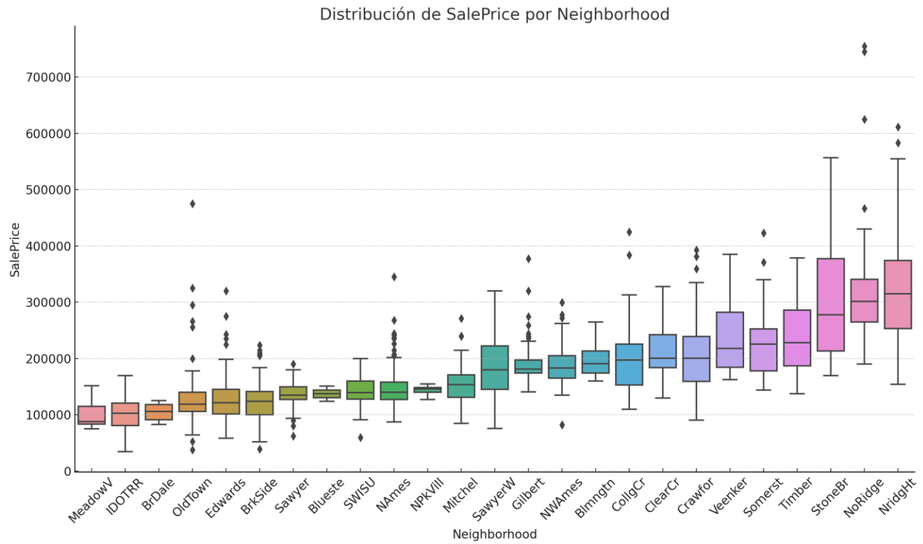
Características categóricas análisis previo

- **Análisis Descriptivo**:
- **Gráficos de barras**: Si observas diferencias significativas entre las categorías, es probable que esa variable categórica sea importante.
- **Boxplots**: Son útiles para identificar diferencias entre categorías y la presencia de valores atípicos.
- **Tabla de frecuencias**









Características categóricas análisis previo

- En el contexto de las variables categóricas una prueba comúnmente utilizada es el Análisis de Varianza (ANOVA).

- **F-statistic**: medida que indica cuánta variabilidad hay entre los grupos en comparación con la variabilidad dentro de los grupos. Un valor F más alto sugiere que las medias de las categorías son diferentes entre sí.

- **p-value**: Si el p-value es pequeño (por lo general, menor que 0.05), sugiere que al menos una de las categorías tiene una media significativamente diferente de las otras.

Variable	F-statistic	p-value
Neighborhood	71.78	1.56×10−225
Utilities	0.30	0.585

Estos resultados muestran la diferencia en la importancia de estas dos variables en relación con el precio de venta. "Neighborhood" parece ser crucial, mientras que "Utilities" parece no serlo

A partir del ANOVA podemos considerar solo aquellas características cuyos p-valores sean menores que un cierto umbral (por ejemplo, 0.05)

Puedes utilizar la biblioteca scipy.stats, que proporciona la función f_oneway