UNIVERSIDAD DEL VALLE DE GUATEMALA

Minería de Datos

Avances
Proyecto No. 2

Isabella Miralles #22293

Guatemala, 2025

Descripción de las variables

Los datos para explorar tienen 81 variables que describen diferentes características de las viviendas. Las mas importantes para este análisis son,

- SalePrice
- GrLivArea
- OverallQual
- YearBuilt
- TotalBsmtSF
- LotArea
- GarageCars
- GarageArea

Estas variables son consideradas para evaluar la influencia que tienen en el precio de las viviendas.

Análisis Exploratorio de los datos

Este análisis se realizo con el objetivo de entender la distribución de las variables, identificar los posibles valores atípicos y examinar las relaciones que puedan influir en el precio de las casas.

Revisión general y estadísticas descriptivas

- Dimensiones y tipos de datos
  El dataset tiene 1460 registros y 81 columnas. Se vieron los tipos de datos para diferenciar las variables numéricas de las categorías.
- Resumen estadístico
  Calcular medidas de tendencia central y dispersión para las variables numéricas. Esto ayuda a identificar rangos y detectar posibles inconsistencias en los datos.
- Datos faltantes
  Se identifican las columnas con valores ausentes. La cantidad de datos faltantes varían y se decide imputar o eliminar dichas variables si no aportan información significativa.

Visualización de distribuciones y detección de outliers

- Histogramas y diagramas de densidad
  Generar gráficos de distribución para variables clave
- Boxplots
  Se utilizarán para detectar valores atípicos en variables

Pruebas de Normalidad

Se utilizarán para evaluar la normalidad de la variable respuesta y otras variables importantes.

Análisis de correlación y relaciones con la variable respuesta

- Matriz de correlación

Se calculará la matriz de correlación entre las variables numéricas para identificar aquellas que tienen una relación con SalePrice.

- Diagramas de dispersión
  Se realizarán para visualizar la relación entre SalePrice y las variables clave.

Análisis de variables a incluir en el modelo

- **Correlación**
  Se priorizan las variables que muestran una correlación alta con el precio.
- **Distribución y normalidad**
  Se realizarán transformaciones a variables que presentan sesgos, así se garantizara que los supuestos de normalidad de los modelos de regresión se cumplan.
- **Relación entre variables**
  Se evaluará la relación con una matriz de correlación, esto para evitar incluir variables que aporten informacion redundante.
- **Análisis gráfico y estadístico**
  Se emplearán técnicas como el análisis de componentes principales para identificar patrones y agrupar variables.

Enlace del repositorio

https://github.com/Isabella-22293/Proyecto1-MD.git