**Hallazgos del Análisis Exploratorio de Datos**

**Realizado por:** Dennys Rojas Llangarí

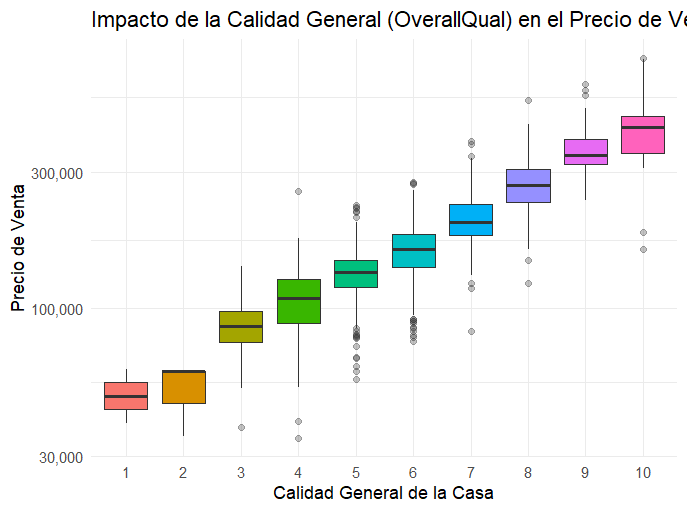
1. **¿Cuáles son las 3 características más correlacionadas con el precio de venta (SalePrice)?**

Tabla

El contenido generado por IA puede ser incorrecto.

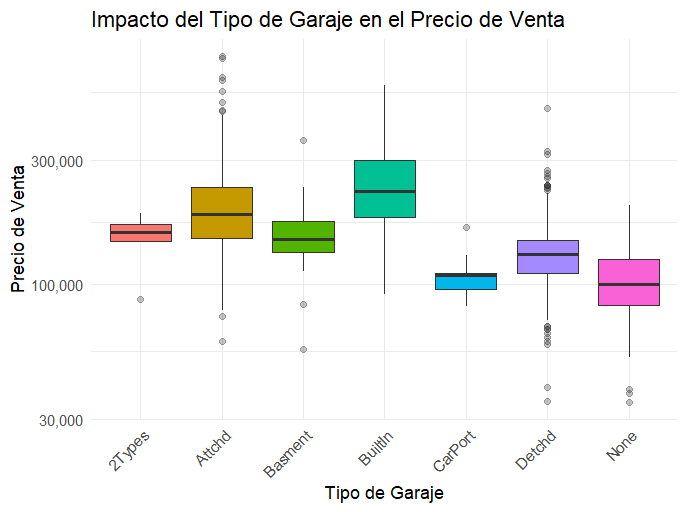
Tras calcular la matriz de correlación para todas las variables numéricas y ordenar por el valor absoluto de la corrección con SalePrice, las 3 características más correlacionadas son: OverallQual, GrLivArea, GarageCars; esto significa que estas características estás correlacionadas fuertemente con el precio de venta.

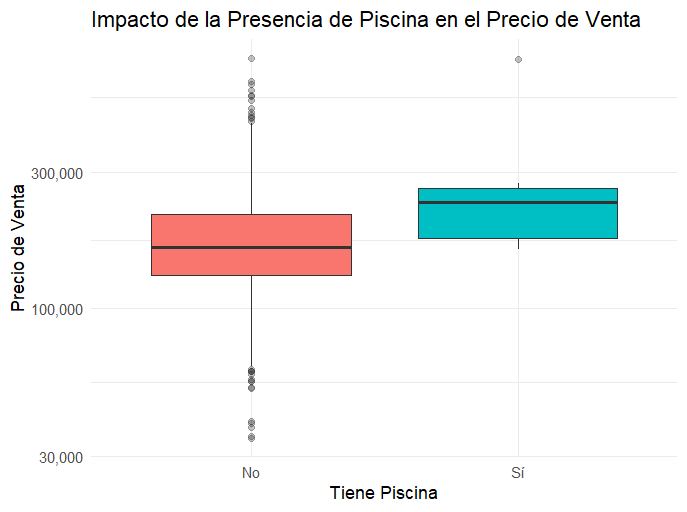
1. **¿Cómo afecta la calidad general de la casa (OverallQual) al precio? (Ideal para un boxplot)**



Se muestra una relación fuerte y clara de tendencia ascendente, es decir que a medida que OverallQual aumenta, el precio de venta también aumenta y que existe mayor variabilidad a mayor calidad, de hecho, en los últimos niveles (8-10) la dispersión de precios tiende a ser mayor.

1. **¿La presencia de un garaje o una piscina impacta significativamente en el precio?**





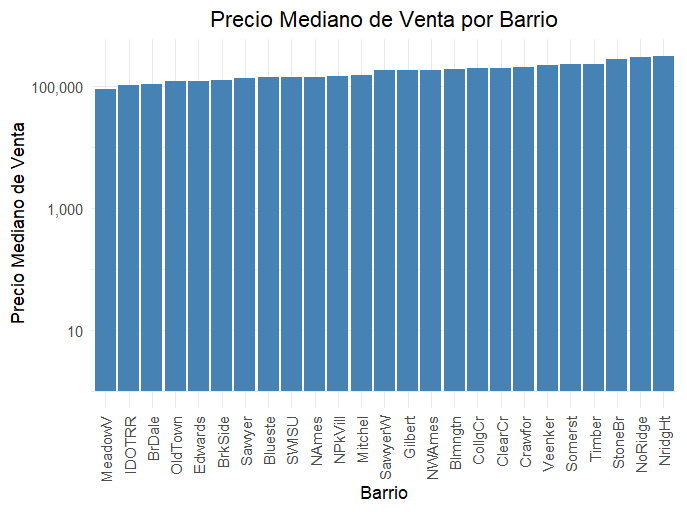
En el boxplot que compara SalePrice con GarageType se revela un impacto significativo; las casas sin garaje tienen un precio de venta considerablemente más bajo en comparación con las casas que sí.

El boxplot que compara SalePrice con la presencia de piscinas revela que los precios son más altos para casas que tienen piscina en comparación con los que no tienen.

1. **¿Qué barrios (Neighborhood) son los más caros y los más baratos?**

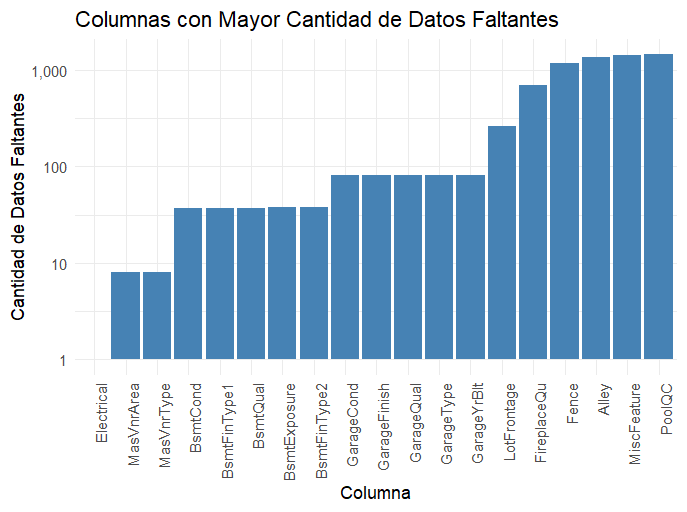
Tabla

El contenido generado por IA puede ser incorrecto.



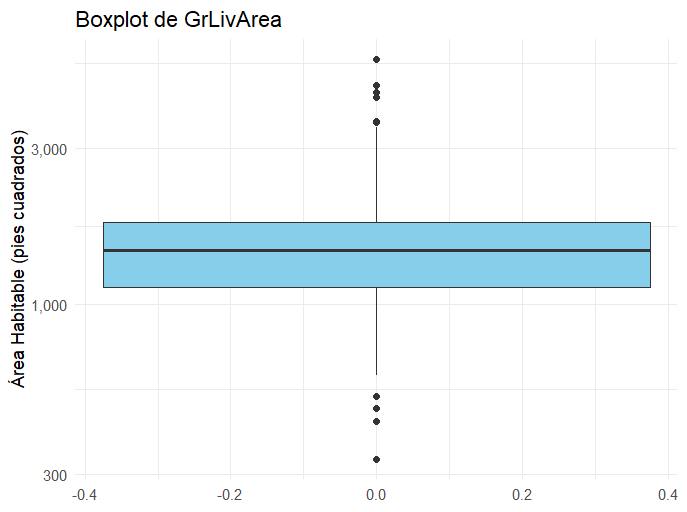
Los barrios como NridgHt, NoRidge y StoneBr son las zonas con precios de venta más alto, lo que podría indicar alguna característica de las propiedades que causen este fenómeno. Por el contrario, MeadowV, IDOTRR y BrDale son barrios con propiedades con precios más asequibles, lo que podría deberes a factores relacionados con la propiedad, pero en sentido contrario al observado en los barrios anteriores.

1. **¿Qué columnas tienen la mayor cantidad de datos faltantes y qué estrategia propones para cada una (eliminar, imputar, o re-codificar)?**



Las columnas con mayor proporción de NA son PoolQC, MiscFeature y Alley. En general la estrategia sería recodificar los NA a una variable categórica como None. Si se analiza de una manera particular, se entiende en el set de datos que los NA implican que no existe o no se presenta cierta categoría. Otra estrategia es imputar con la mediana de los valores numéricos existente o a su vez con CERO dependiendo del caso.

1. **Identifica y visualiza los outliers en el área habitable (GrLivArea). ¿Qué harías con ellos?**



La decisión sobre como manejar estos outliers depende del objetivo de análisis. Primero es necesario investigar cual es la naturaleza de estos outliers y entender si son errores de entrada de datos (para lo cual deberían corregirse o eliminar) o si son valores legítimos extremos y por ende son valores correctos la opción más válida sería mantenerlos y dependiendo del caso tratarlos en función de lo que se pretenda hacer a continuación: si es una análisis exploratorio, únicamente exponer estos datos; si se pretende construir un modelo predictivo se podría aplicar transformaciones a las variables para reducir el impacto de valores extremos o a su vez utilizar modelos más robustos a estos valores outlier, o en el caso más extremo, eliminarlos.