****

TAREA INDIVIDUAL

Laboratorio: Árboles de Decisión

**Estudiante:** Góngora Rodríguez Nicole Patricia

**Código:** 51225

**Materia:** Tópicos Selectos de Inteligencia Artificial

**Carrera:** Ingeniería de Sistemas Computacionales

**Semestre:** Octavo

**Docente:** Ruth Chirinos

2022

# Análisis estadístico de la información:

Este informe es un resumen explicado del tratamiento de datos en un dataset para análisis de variables, correlación y aplicación de árboles de decisión en la clasificación y predicción del precio. Los resultados completos están en el Notebook adjunto.

## Estadísticas de la información numérica:

Para revisar cada característica estadística básica de las variables numéricas en la base de datos, se utiliza la función *describe()*, la cual proporciona información acerca del número de datos no nulos, la media, la desviación estándar, los cuartiles de distribución 1 (25%) 2 (50% o mediana) y 3 (%75), junto a los valores mínimos y máximos encontrados, como se observa en la Fig. 1, basándose en los resultados de la Fig. 2.

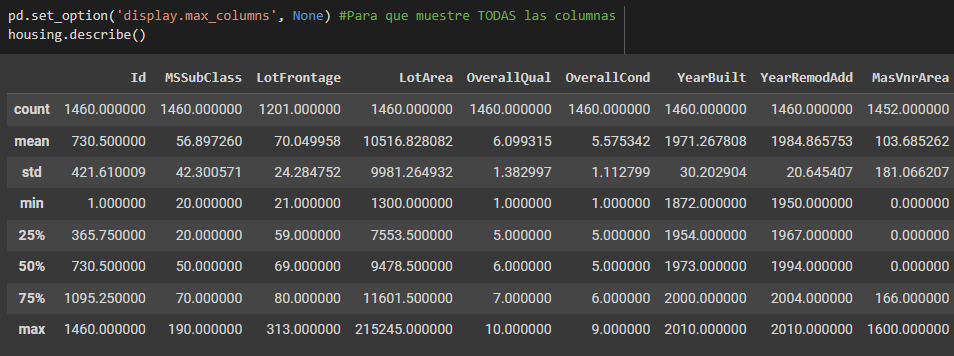


Figura 1: Resumen del resultado del comando describe(), mostrando las estadísticas esenciales

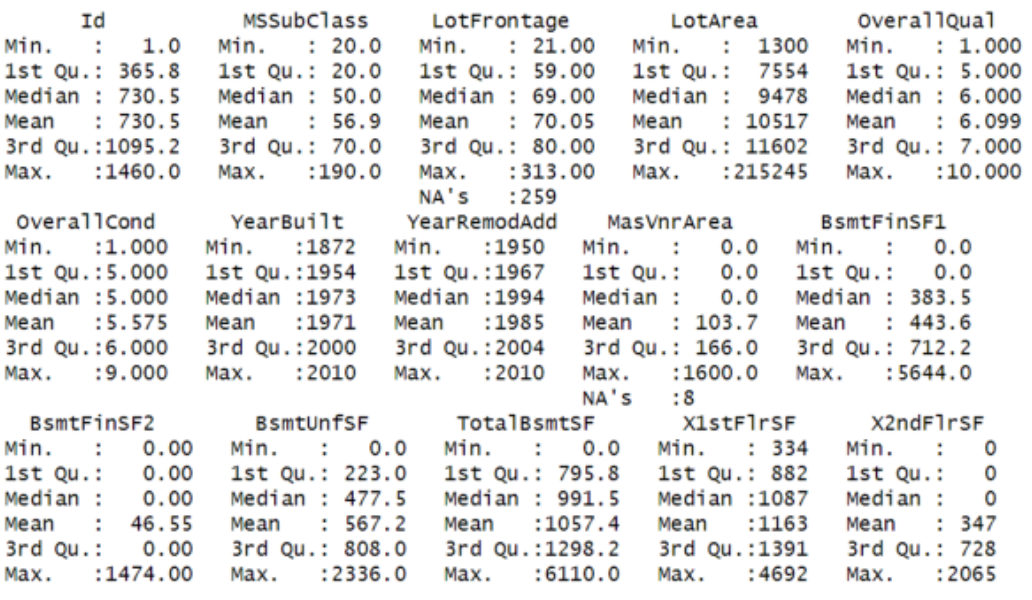


Figura 2: Estadísticas de columnas numéricas

**(Respuesta a b))** A partir de estas visualizaciones previas obtenidas, se pueden apreciar las siguientes observaciones:

* Existen variables como la superficie final de las superficies acabadas 1 y 2 del sótano, la superficie de revestimiento de mampostería, y el área del segundo piso cuyas distribuciones de datos no están muy bien distribuidas, lo que le quita las secciones de los primeros cuartiles y reduce la mediana a cero. Esto nos puede decir que estas variables o atributos no están muy presentes en las propiedades: hay más casas con solo un piso y un tipo de superficie de sótano, así con las paredes sin cobertura de mampostería externa.
* La calidad y condición de las casas se inclina a un estado medio, con una subclase de construcción 50 y construidas/remodeladas alrededor de los 80’ y 90. Así mismo, muchas casas contienen sótanos, y un área de parcela de 9478.5 pies (880 metros) cuadrados, lo cual indica un tamaño grande de casas.

Para entender mejor estas variables, se pueden crear diagramas de cajas y bigotes, describiendo de manera gráfica la dispersión de las variables, como se muestra en las Fig. 3 a Fig. 6 (hay más gráficas en el Notebook).

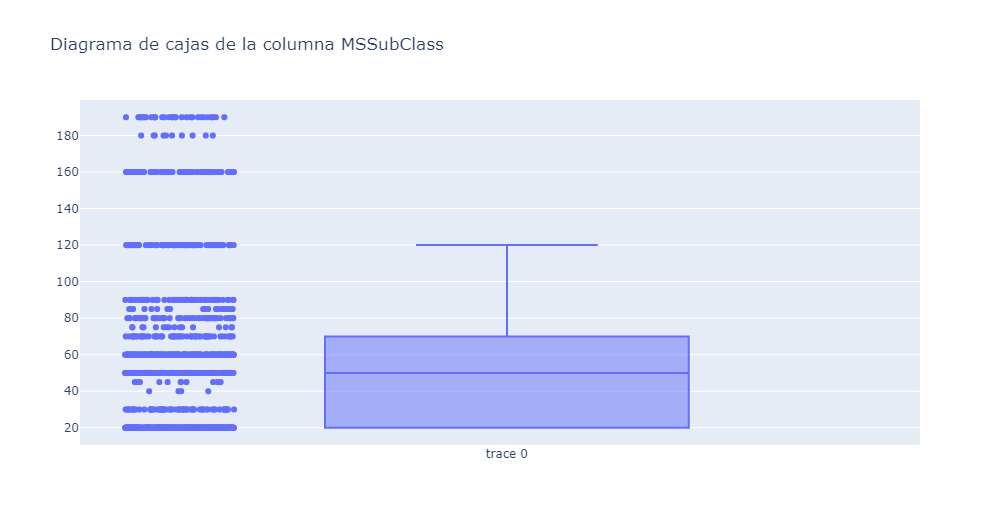


Figura 3: Boxplot de la columna MSSubClass

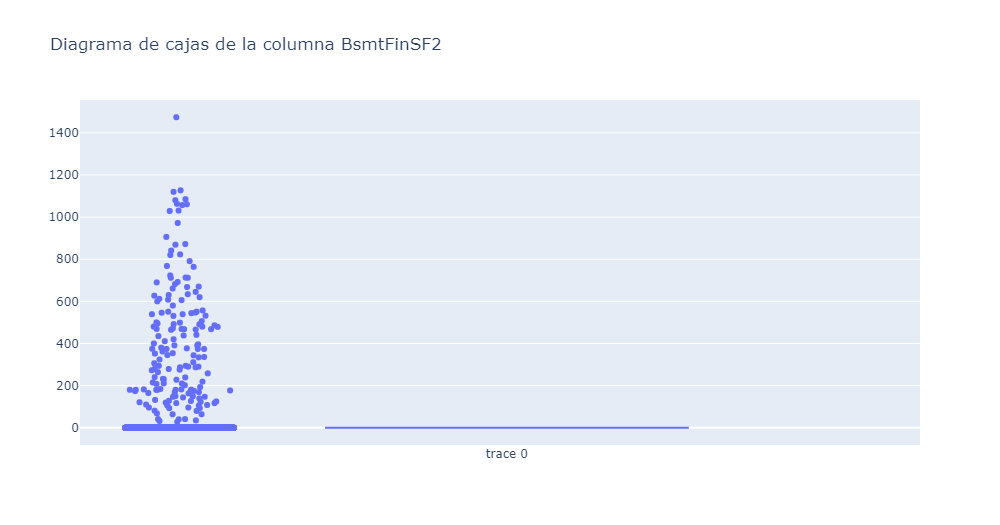


Figura 4: Boxplot de la columna BsmtFinSf2

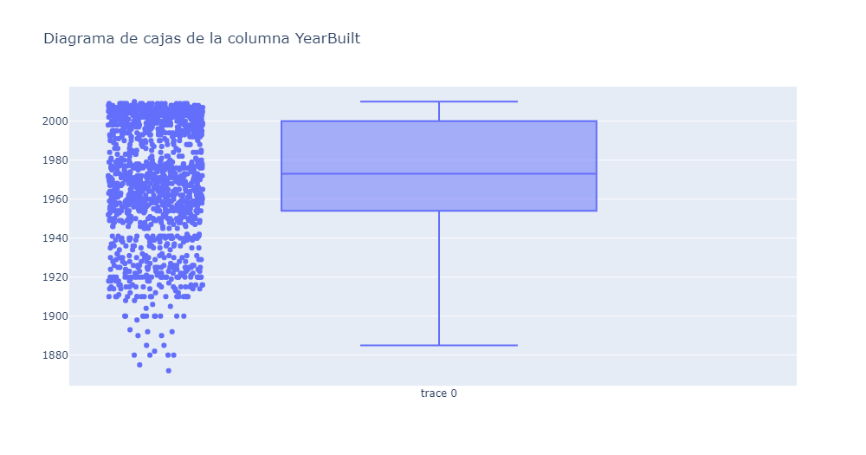


Figura 5: Boxplot de la columna YearBuilt

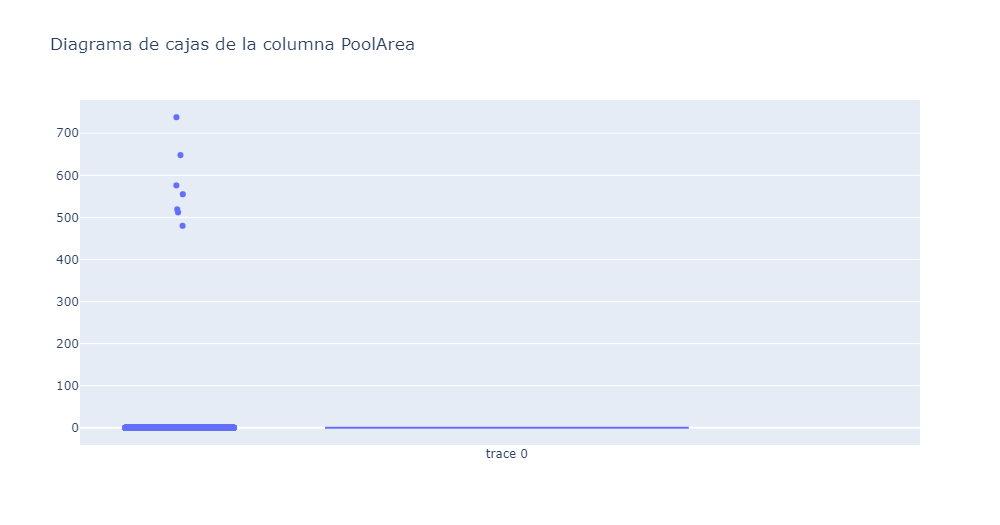


Figura 6: Boxplot de la columna PoolArea

**(Respuesta a a) y b))** Gracias a las gráficas mostradas, se puede observar las siguientes cualidades:

1. Variables numéricas como ID, MSSubClass (clase de construcción), OverallQual (calidad del material y acabado), OverallCond (condición general), todas aquellas con el prefijo Year o Yr (correspondiente al año: YrSold (año de venta), GarageYrBlt (año de construcción del garaje), YearRemodAdd (año de remodelación), YearBuilt (Año de construcción)), MoSold (el mes de venta) serían más categóricas que numéricas/estadísticas en sí, ya que se puede clasificar una casa por su calidad, su tipo de construcción, el año y mes en el que fue construida/remodelada/vendida
2. Las variables numéricas/estadísticas que quedan, siendo correspondientes, en general:
   1. Al número de objetos como chimeneas, o habitaciones, tanto normales como baños, cocinas; incluso el tamaño del garaje medido en plazas de autos
   2. A las medidas de superficie y área de las habitaciones, piscinas y las casas en sí en pies, así como el valor en dólares de características misceláneas.

están, en su mayoría, bien distribuidas, con las siguientes observaciones:

1. Algunas no tienen mediana, es decir, al tener sus valores 0, crean una distribución de cuartiles en las que el primero y segundo no existen
2. Otras tienen buenas distribuciones, pero con extremos muy distantes

Por lo tanto, luego del análisis de categorías y sus frecuencias, hay que normalizar esta información. Esto, no solamente para dar una mejor explicación antes de convertir las categorías para la regresión, sino para determinar el impacto de los datos faltantes.

Así mismo, a primera vista se pueden eliminar columnas donde no se vea mucha presencia, como el área y calidad de la piscina, el porche de tres estaciones, los baños medios en el sótano, porque se observa (véase en el Notebook) que muchas casas no contienen este atributo (visto en la mediana y los cuartiles nulos) y que la minoría se distribuye en pocas unidades en el aire, junto al Id al ser una columna de identidad más que de clasificación o calificación. Esto habría que observar con la matriz de correlación en el siguiente paso.

## Análisis de frecuencia de clases en variables categóricas:

Antes de la normalización y categorización de las variables, hay que conocerlas a fondo. Se considerarán categóricas no solo a las variables con tipo "object" identificados en sus columnas, sino también a las numéricas antes mencionadas, como se observa en la Figura 7.

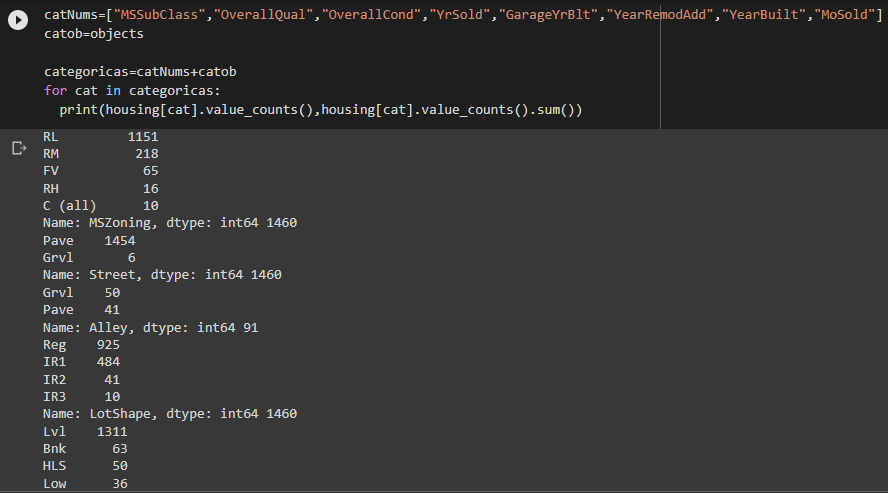


Figura 7: Resultado de la función value\_counts en cada columna categórica

**(Respuesta a a))** A partir de la anterior visualización, se puede concluir que esta información corresponde:

* A las diversas categorías que describen la calidad/condición de la construcción en general, la cocina, el sótano, el exterior y la piscina entre otros; así como la proximidad a la carretera o el ferrocarril.
* A los años y meses de venta, remodelación y construcción de la casa y el garaje
* A las clases de construcción, el tipo de casa, los materiales y categorías del techo, mampostería, chimeneas, atributos del exterior.

Para determinar el nivel de frecuencia que estas variables tienen en la base de datos y su influencia en el precio, se realizaron gráficas en cada una, desde barras a histogramas relacionadas o no a la variable objetivo, como se observa en las Figuras 8 y 9 (Más información y gráficas en el Notebook).

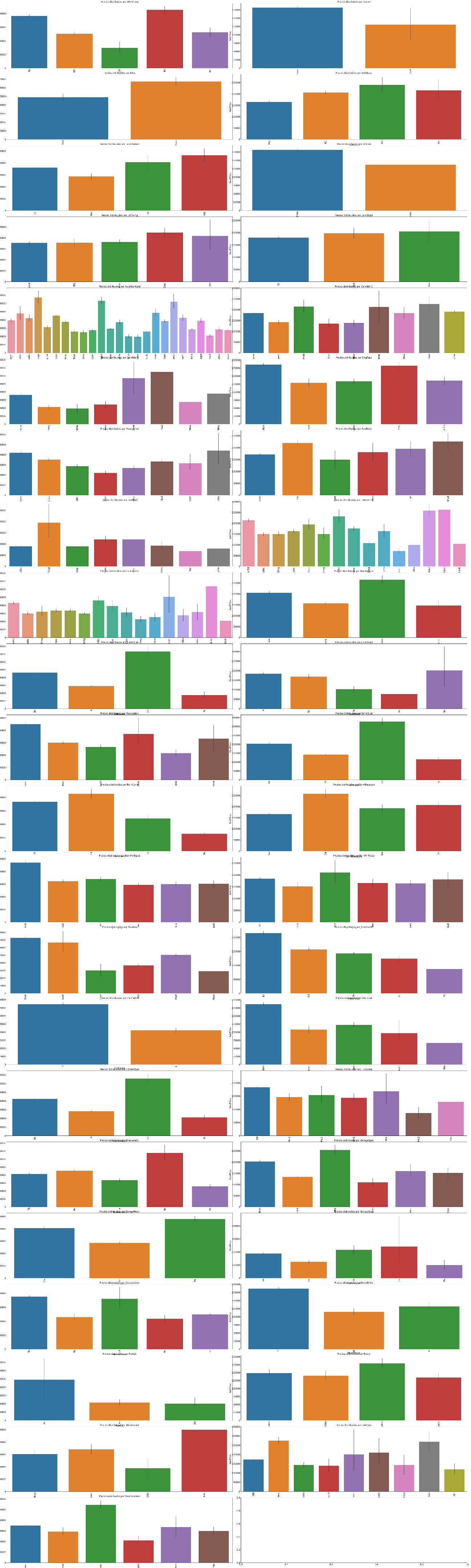


Figura 8: Diagrama de Barras de las variables numérico-categóricas con respecto al precio (Configuración de lote, Pendiente de suelo, Barrio y Proximidad a camino)

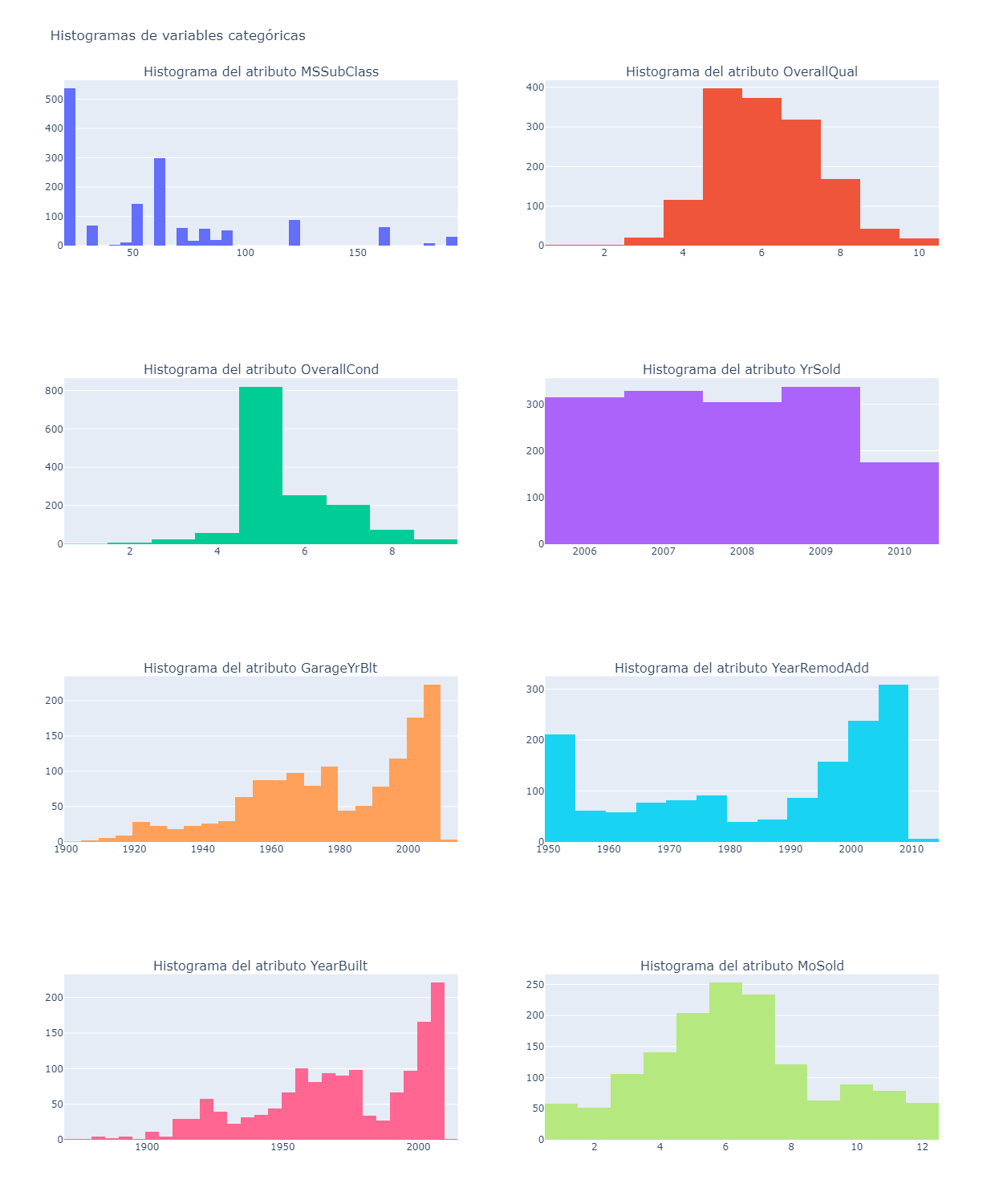
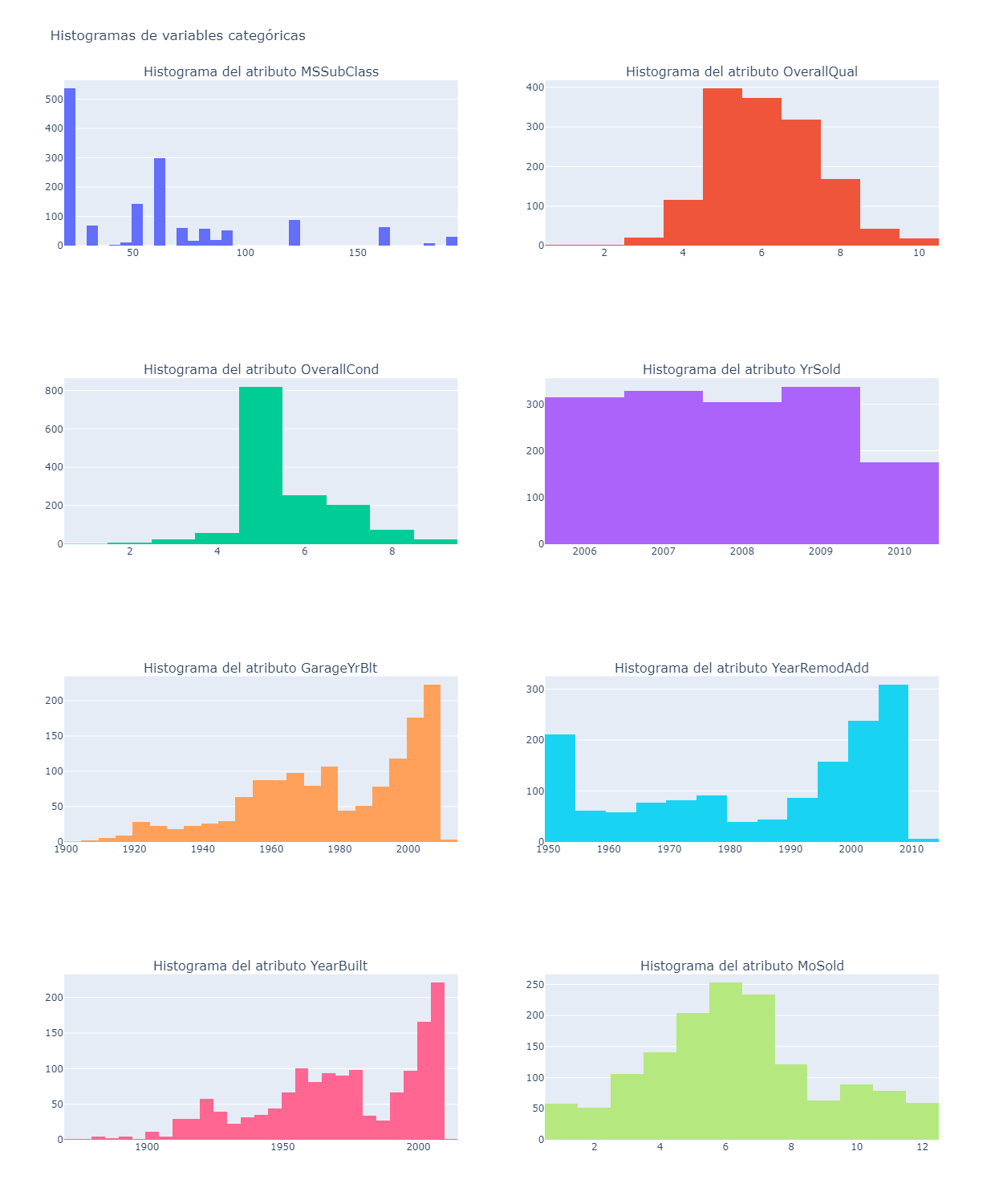


Figura 9: Histogramas de frecuencia de subclases en cada categoría (Clasificación de Zona, Calidad general, Año de Construcción, Mes de venta)

A partir de esta visualización de la cantidad de clases y su frecuencia en la base de datos, se pueden realizar conclusiones de inferencia:

* Las mejores ventas de los inmuebles se realizaron en el mes de septiembre en el rango de años 2006 al 2010; con transacciones parciales y nuevas acumulando mejor distribución. Sin embargo, hubo más ventas de mayo a julio y menos casas vendidas en el 2010
* Las casas de la categoría 20 de construcción son más frecuentes en el registro, un buen cúmulo entre los 50’ a 70’, la mayoría edificadas y/o remodeladas en el 2010. Pero aquellas con mayor acumulación de dinero fueron aquellas de categoría 60 desde los 80’ a los 2010.
* La calidad de los inmuebles es media, calificada de 4 a 5 puntos en este rubro, pero aquellas que acumularon más dinero de venta se acercan a 9 a 10 puntos.
* Las casas que aportan más valor a su precio son aquellas de clasificación FV, caminos pavimentados a la calle y empedrados al callejón, lote con una sola entrada de forma IR2 y contorno HLS con todos los servicios. Además, cuentan con mampostería de piedra, exteriores de excelente calidad formados con estuco, techos compuestos de forma diagonal uniforme (shed) y capacidad para una familia mediante dos plantas con sótano terminado (2.5 fin). Se ubican en los barrios NoRidge, StoneBr y NRidgHt, tienen sótanos de gran calidad con paredes a nivel del jardín, calefacción a gas, sistema eléctrico Sbrkr y garajes internos en la casa.
* La mayoría de las casas son de un piso, para una sola familia, de techo triangular compuesto y acceso normal al lote, en barrios fuera de Ames (NAmes) con todos los servicios y un contorno de territorio nivelado y regular.
* A diferencia de la calidad y el mes de calidad, todas las variables categóricas numéricas están distribuidas irregularmente, con pendientes hacia años recientes o separación de clases

Gracias a esto, se pueden generar varias conclusiones de entrada de los posibles aportadores al precio de una casa mediante algunos análisis exploratorios en ejercicios alternativos (disponible en el Notebook):

* Hipótesis: Las mejores ventas se realizarán en meses desde mayo a septiembre.
* Hipótesis: Hay casas que se venden el mismo año que se construyeron, es decir, que pueden agregarle el factor de novedad al precio de venta. Estas van desde el 2006 al 2010, el primer año teniendo más registros.
* El área más común del garaje es de 400 a 500 pies cuadrados, con capacidad para 2 autos. Hipótesis: El precio del inmueble más alto puede predominar en este rango.
* Hay más variedad en estilo de casas y construcciones para los inmuebles con una calidad regular de atributos como la calefacción, la cocina, el exterior y el garaje; mientras que aquellas con calidad excelente o pobre se limitan a 2 o 3 estilos de casa y 1 a 2 tipos de construcciones.
* Hipótesis: Las parcelas con 1300 a 171.000 pies de área acumulan precios desde aproximadamente 8000 a 300000 dólares, aquellas de mayor tamaño y/o precio dispersándose distantemente. Estas son las que mayor valor de precio pueden aportar.
* En todos los tipos de construcciones, el número de baños completos influye más que el de los medios, habiendo algo de equilibrio en zonas de categoría 160.

## Matriz de correlación entre variables:

Para revisar qué variables tienen un vínculo tanto entre sí mismas como con la variable objetivo del precio, se realiza una matriz de correlación, como se contempla en la Figura 10 y 11:

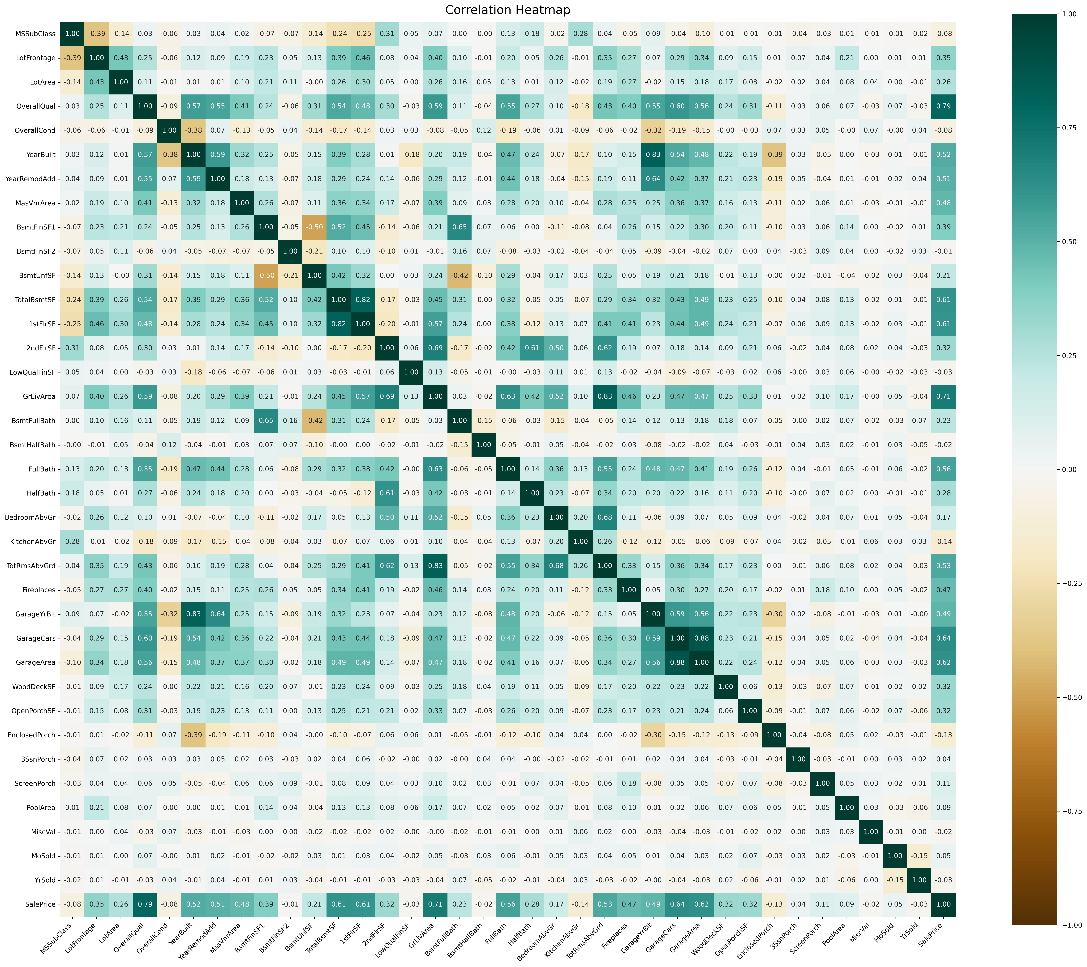


Figura 10: Matriz de Correlación inicial

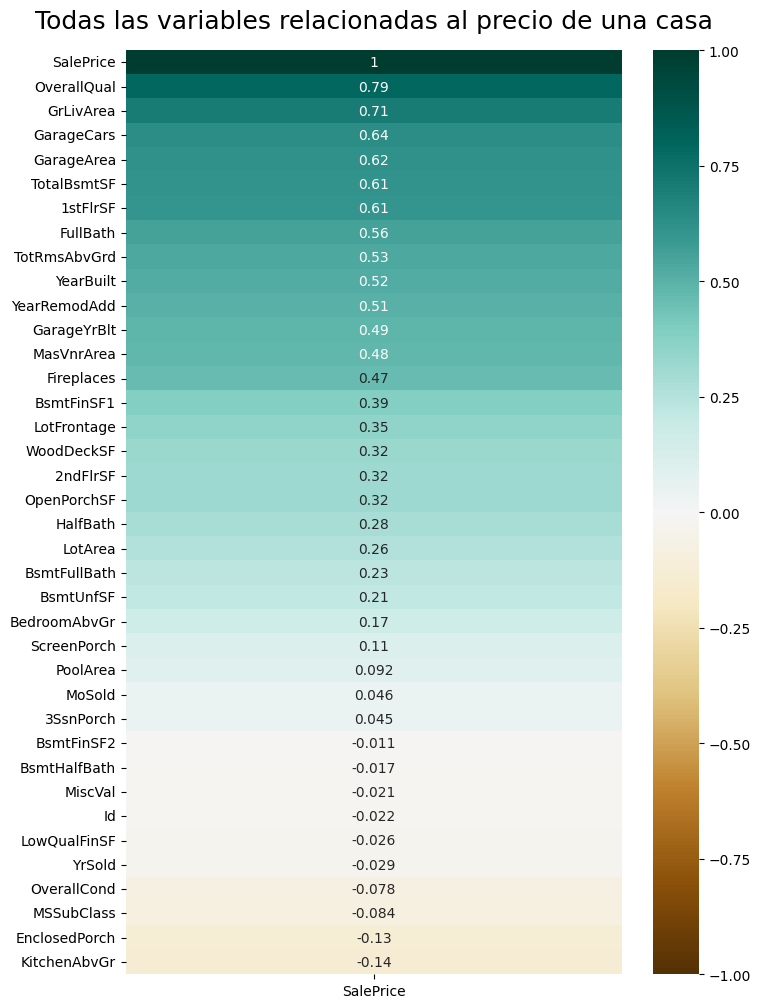


Figura 11: Barra extraída de la matriz de correlación, describiendo sólo para el precio

**(Respuesta a c))** Esta contemplación gráfica de la correlación se pueden identificar las variables con mayor correlación (al menos mayor a 50%):

1. Entre aquellas que no son el precio:
   1. El año de construcción del garaje está muy vinculado al año en el que la construcción ha sido edificada, talvez debido a que ambos hayan sido edificados en el mismo periodo como se vio en los gráficos de frecuencia de categorías.
   2. La calidad en general de la construcción tiene una mayor dependencia al año de la construcción del garaje, la cantidad de baños completos, el área disponible sobre el nivel de suelo, el área total del sótano y el primer piso, el año de construcción y de remodelación de la casa, y el área del garaje en pies y autos.
   3. El año de remodelación se vincula al año de construcción de la casa, probablemente debido a que se realice esta acción en el mismo periodo. Lo mismo con el año de construcción del garaje, por la misma razón hipotética
   4. El área total del sótano tiene una correlación intensa con la primera área, seguida de la segunda y parcial (incompleta), debido a que es una suma de estas variables, en especial de aquella con mayores registros mayores a 0; así como del área de calle de la parcela. El área incompleta del sótano, en cambio, depende de manera fuerte pero negativa del número de baños y el área completa tipo 1: es decir, mientras más aumenten estos últimos, se ocupa más espacio sin acabar, por lo que esta cantidad se reduce mucho.
   5. El área sobre el nivel del suelo depende mucho de los cuartos disponibles en este sector, seguido del área del segundo y primer piso junto al sótano, el número de baños completos y medios, el área de garaje y las chimeneas.
   6. Él área de los cuartos sobre el suelo depende del número de los dormitorios y los baños, tanto completos como parciales, así como del área del segundo piso y ligeramente del primero.
   7. Los dos tipos de área del garaje se relacionan uno al otro fuertemente, esto debido a que uno justifica al otro en una distinta medida (# de autos – pies cuadrados). Así mismo, los dos casi paralelamente dependen del año de construcción del garaje, el año de construcción/remodelación de la casa, y las áreas por encima del nivel del suelo, el sótano y el primer piso.
   8. El primer piso del área, aparte del garaje y el sótano en áreas, depende del área de pies lineales de calle, así como del número de chimeneas.
2. Con respecto al precio, de mayor a menor correlación:
   1. La calidad en general de la casa afecta mucho la proporción del precio, esto talvez sea a que, a mayor perfección del inmueble, mejor se vende.
   2. El área disponible sobre el nivel del suelo y las habitaciones encima de este sector también influyen en el precio, esto probablemente debido a que permite tener una mejor calidad de los cuartos y espacios para que se puedan distribuir.
   3. Las superficies y áreas terminadas del sótano, garaje y el primer piso le siguen, justificado en una hipótesis de que, al haber registros que confirmen la existencia de sótano, garaje y al menos una planta en la construcción, estos sean influyentes en el valor de venta según su valor suba.
   4. El año de remodelación y construcción parece tener una correlación algo liviana pero resaltable en el precio, esto explicado por los gráficos de la distribución de precios según años en los que muestran un crecimiento del valor de venta según avanzara el tiempo.

Viendo estas correlaciones, se observan posibles columnas a eliminarse:

* Si bien la calidad del material y el acabado son importantes, hay ya atributos que pueden cubrirla una vez categorizados en cifras (material del techo, calidad del material exterior, calidad de la cocina, etc.), sin acumular demasiado valor que afecte la dependencia del precio.
* El área de superficie habitable por encima del suelo puede ser mejor explicada y menos acumulativa al separarla con las áreas del primer y segundo piso, del lote, e incluso con las habitaciones por encima del suelo.
* Lo mismo con el área total del sótano, ya que existen columnas que explican mejor su valor (la altura del sótano, las superficies acabadas y sin terminar), reduciendo dependencia en la variable del primer piso
* Una de las variables del garaje, la de los autos que pueden entrar, podría quitarse, ya que el área de por sí sola explica el tamaño de este sector

## Tratamiento de valores faltantes:

### Variables numéricas faltantes:

Se deben encontrar las variables numéricas con valores faltantes que puedan afectar la regresión, como se observa en la Figura 12.

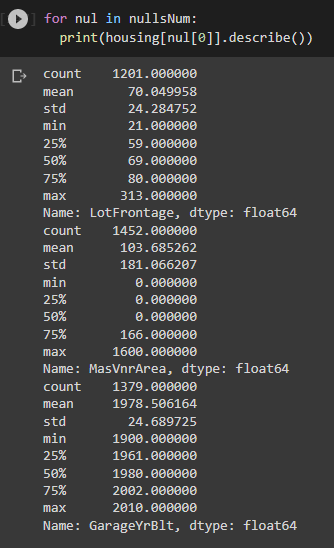


Figura 12: Función manual para la descripción de distribución de datos faltantes numéricos

**(Respuesta a d))** Para las variables numéricas con datos faltantes, se puede contemplar que, una de las tres tiene una distribución de área de mampostería (MasVnrArea) muy desigual, dos de sus cuartiles están desiguales. Este dato se puede observar también en el análisis estadístico de las variables numéricas antes realizados. Si bien llenarla con la mediana, siendo esta 0, sería una situación que puede llevar a una normalización errónea, se puede considerar que, siendo el revestimiento de mampostería una capa externa de ladrillos o piedras colocadas en la pared de una construcción, se asume que pocas casas pusieron ladrillos o piedras en la pared como decoración o refuerzo de la construcción, dejándolas planas; agregando que los datos faltantes son 8 en esta columna. Por ende, es mejor llenar los datos con la mediana podría no afectar mucho la información.

Para las otras variables numéricas, siendo estas el año de construcción del garaje y los pies lineales de calle de la parcela, observando tanto sus distribuciones en el describe como los gráficos estadísticos anteriores correspondientes, se contempla que existe una distribución muy poblada de datos cerca de la mediana y media. Así mismo, para el caso de los años, la mediana nos da 1980, un sector muy poblado entre los cuartiles, mientras que la media da 1978.50, un valor válido en este rango, pero flotante. Esto lleva a determinar que la mejor manera de llenar sus datos faltantes sea con la mediana

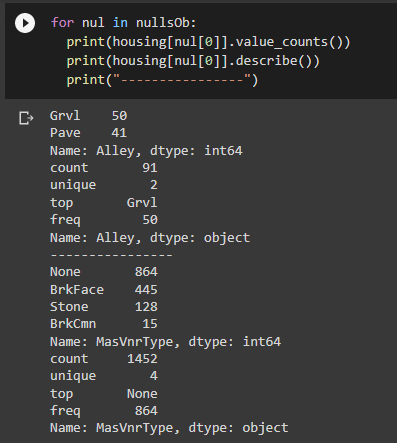


Figura 13: Función manual para la obtención de valores contados por columnas categóricas

**(Respuesta a d))** A partir del conteo de valores y la descripción estadística adaptada de las columnas categóricas con datos faltantes de la Figura 13, se pueden apreciar los siguientes datos: Las veces que aparecen cada categoría en la base de datos, el número de valores únicos o categorías en cada una, y la categoría con mayor frecuencia y el número de veces que se presenta. Si bien es tentador utilizar un llenado por la moda o la mejor variable por la cual se relacione más con el precio, esto podría generar una dependencia alta de la columna a la variable objetivo, lo cual no es deseable. Por ende, se tienen dos opciones de llenado de estas variables con datos nulos:

* Como no se pueden inventar los valores de los registros, y esto signifique que probablemente el registro nulo de las columnas indique que la propiedad no contiene al atributo, se llena con la palabra "None"
* Se llena el valor nulo de cada columna con una elección aleatoria de las variables únicas para generar una población variada y que pueda dar una regresión variada a la anterior

Sin embargo, se toman en cuenta observaciones en los datos faltantes con respecto a la piscina. Al comparar la inexistencia de los valores, se puede observar la cohesión de los datos en 0 con la inexistencia de datos. Si le damos categorías randómicas a las variables nulas, no sólo podríamos dar información equivocada, pero a la vez crear mayor dependencia del precio a estas variables, lo cual no es deseado. Se debe tomar la primera opción.

## Aplicación de árboles de decisión al problema de regresión:

Si bien se tienen ya claras las columnas candidatas a eliminarse y los datos llenados con el método escogido mediante su justificación, puede haber dos posibles caminos para obtener una buena regresión:

- Trabajando sólo con las variables numéricas

- Utilizar las variables numéricas y categóricas para determinar el nivel de correlación entre estas últimas y el precio, y ver si hay más columnas que eliminar.

### Numeración de las variables categóricas, normalización y comparación correlacional:

Para cambiar los valores de las variables categóricas, se han considerado dos modos: Mediante una librería de Python llamada OrdinalEncoder del grupo sklearn.preprocessing para el llenado automático de numerales, o de manera manual. Se realizaron las dos opciones; para luego normalizar cada dataset creado mediante el método y librería MinMax, el cual preprocesa las variables mediante una valuación con respecto a los valores mínimos y máximos de las columnas, lo cual podría funcionar en los valores dispersos de la información. Así mismo, para ver el efecto de las categóricas en la correlación con el precio, se recurrió a la elaboración de matrices de esta naturaleza, como se observa en las Figuras 14 y 15:

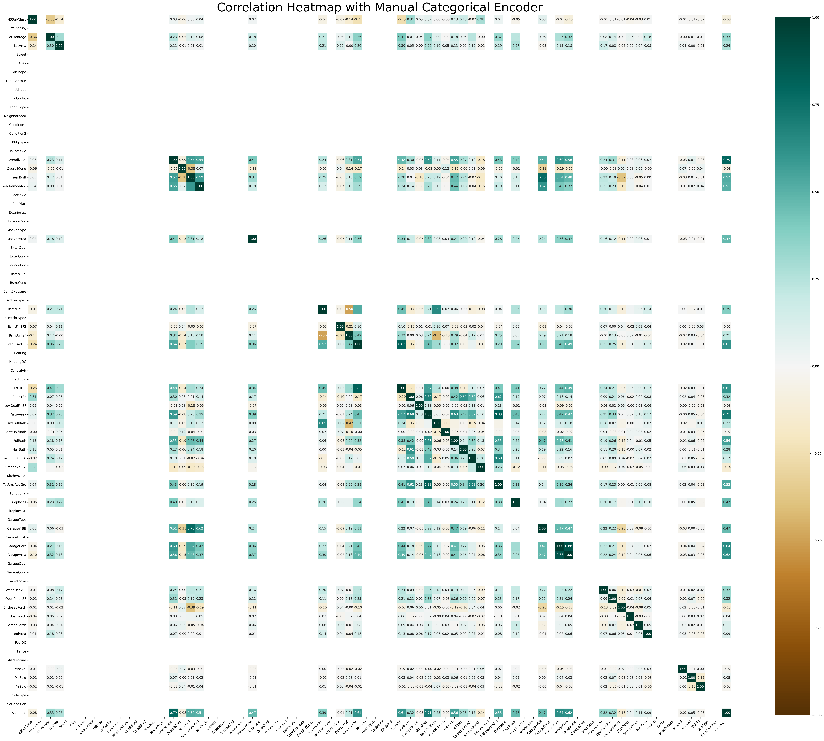


Figura 14: Matriz de correlación con categorización manual

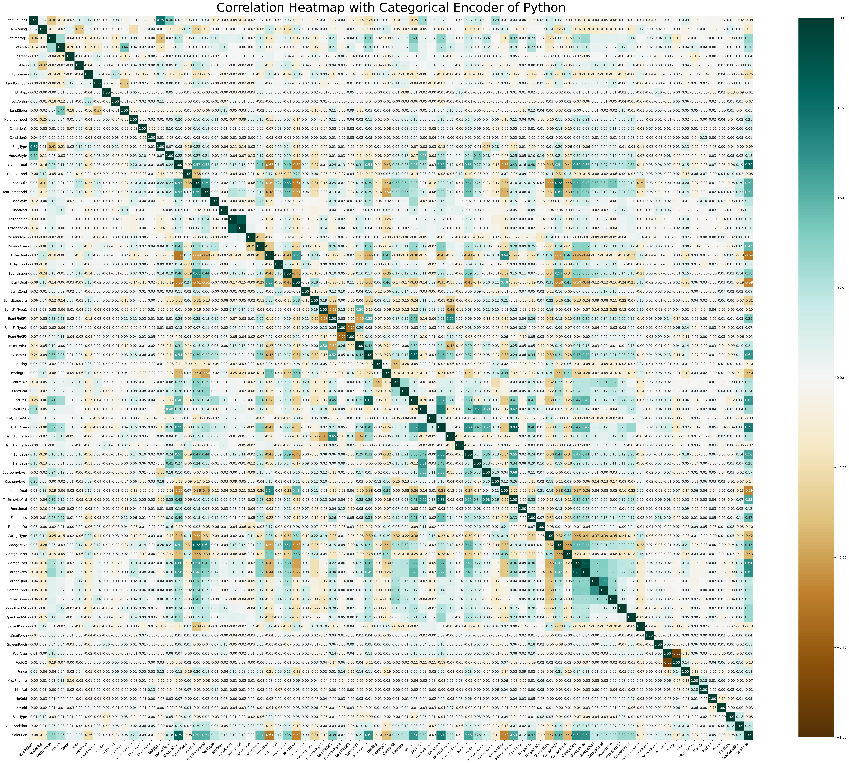


Figura 15: Matriz de correlación con categorización automática

Se puede ver que la primera matriz de correlación cuenta con espacios vacíos en TODAS las variables categóricas, mientras que la segunda tiene datos. Según la documentación de Pandas, esto puede deberse a que, al asignar de manera manual y bruta los números a las categorías, hubiera una falta de variación en estas, lo cual genera problemas en la varianza de datos y con ello el nivel de relación entre la información. En el caso de aquella matriz tratada con la librería, se puede ver que el sistema ha calculado el mejor modo de codificar las clases de cada categoría para evitar este problema, lo cual da una correlación amplia. Se comparan a continuación las correlaciones del precio con las demás variables de las matrices de sólo datos numéricos con la última realizada gracias a la librería, como se contempla en la Figura 16:

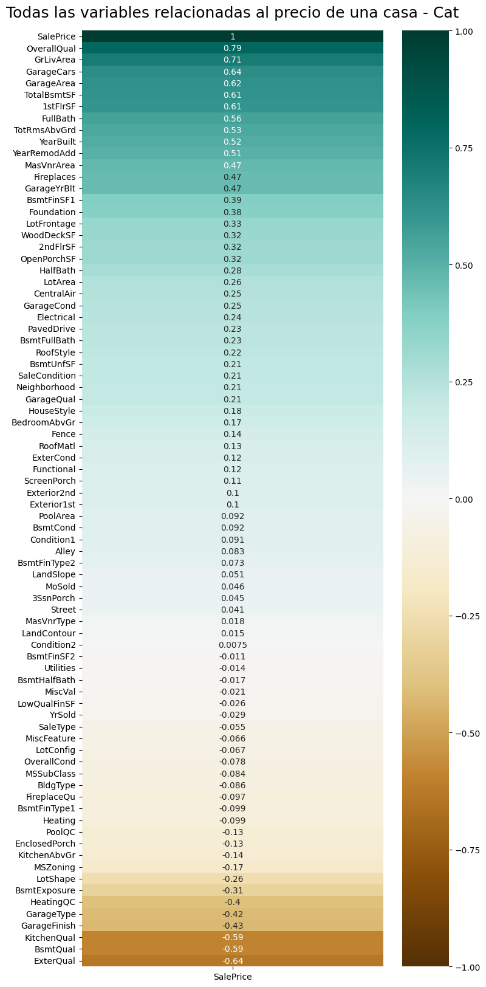
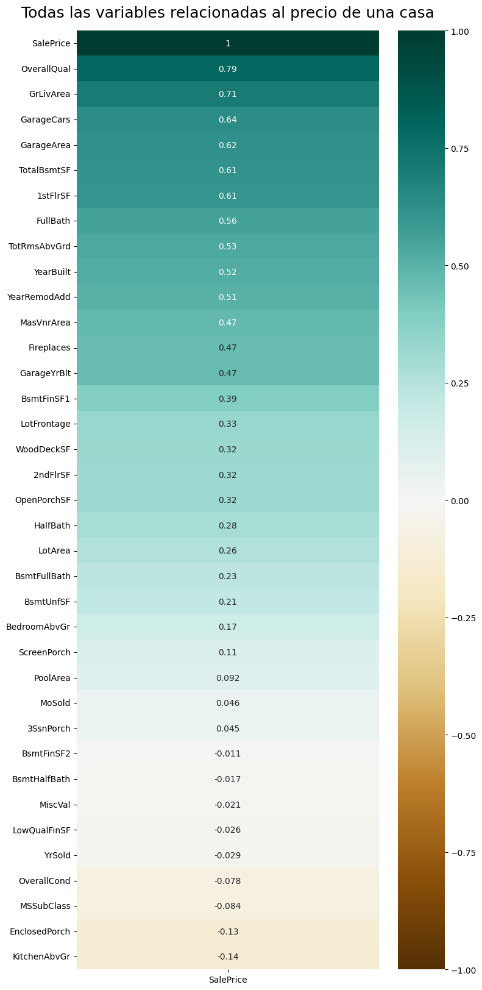


Figura 16: Correlaciones entre el precio y otras columnas, numerales y numerales-categóricas

Se puede observar que, incluso con las variables categóricas agregadas, las correlaciones mayores que tiene la columna del precio siguen siendo las mismas que las obtenidas anteriormente. Hay algunos atributos que tomaron un grado de correlación alto en el caso de aquella dataframe con variables categóricas, como la calidad del sótano, exterior, cocina y garaje, pero son más negativas y podrían llegar a contrarrestar el valor del precio, no afectando mucho la regresión.

### Realización de los árboles de regresión:

Como se estableció antes en la matriz de correlación y análisis estadísticos, se van a eliminar ciertas columnas para llegar a una regresión óptima posible: Calidad general, Área habitable encima del nivel del suelo, Total de cuartos por encima del suelo, Área total del Sótano, Número de autos, y el año de construcción del garaje. El Id ha sido eliminado antes de punto 4. Para todas las regresiones, se han creado datos de entrenamiento y prueba, tanto dependientes como objetivos a partir de la librería *train\_test\_split* de sklearn.model\_selection, de tal manera de no sólo optimizar los modelos de evaluación, sino también evitar un overfitting innecesario. Tomando en cuenta que la variable objetivo, el precio, está en una distribución de datos muy dispersa, lo cual podría arruinar la clasificación y dar imprecisiones, se utilizó la misma función automática de escala MinMax para normalizar los datos objetivo de prueba y entrenamiento antes de aplicar el ajuste del modelo del árbol de decisión.

Luego de la evaluación del mejor modelo para los árboles de decisión, se realizaron dos predicciones, una con los datos normalizados, y otra con una transformación inversa de los valores para obtener las cifras en el formato de precio ideal.

Finalmente, se calcularon tres errores:

* Error Absoluto Medio (MAE): La diferencia en valor absoluto entre el valor real y el predicho. Una cifra pequeña indica una mejor precisión del modelo. Tratamiento igual de valores, incluso los extremos
* Error Cuadrado Medio (MSE): La media de la diferencia entre el valor real y el predicho. Sensible a valores extremos, generando una penalización de error incrementado. Una cifra pequeña indica una mejor precisión del modelo.
* Logaritmo de la raíz del error cuadrático medio (RLMSE): Basado en el MSE. Penaliza más la subestimación de predicciones que la sobreestimación de estas.

Para luego graficar los resultados de cada árbol con las librerías *tree* y *graphviz*, describiendo cada posible variable en la toma de decisiones y clasificación del rango del precio.

#### Árbol de decisión con categóricas y numéricas:

El número máximo de niveles a crear en este árbol han sido 7, tomando como mínimo muestras de información 2; optimizando la clasificación, como se contempla en la Figura 17. Esto se puede crear debido a la gran variedad de rangos de precio, incluso normalizados, en los que se necesitan más de una comprobación para afirmar la clase. Los errores obtenidos de esta regresión han sido muy bajos, con valores que se acercan a 0, lo cual es muy positivo, indicando que la estrategia tomada ha sido beneficiosa, como se observa en la Figura 18.

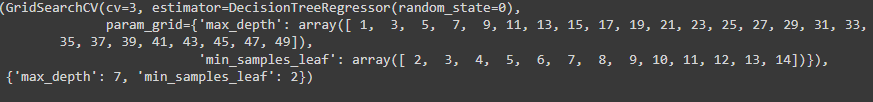


Figura 17: Resultados del estimador para el árbol de decisión categórico-numérico

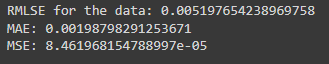


Figura 18: Los errores obtenidos de la regresión del árbol de decisión

El árbol de decisión generado, descrito en la Figura 19, da a conocer una cohesión entre valores de precio estándar en las comparaciones como los normalizados en los resultados de predicción. A partir de una predicción desde la identificación del índice de la predicción normalizada y su uso en la estándar, se pueden sacar las siguientes conclusiones:

* Aquellas casas con un área de sótano terminado tipo 1 de 282 (5%) pies llegan a un valor alrededor de 296, mientras que aquellas con un valor más grande, de 1185 (21%) pies, acumulan 345 mil dólares aproximadamente.
* Aquellas casas con electricidad fuera de las categorías de fusibles A, F, P, mixta y no identificadas tienen un valor máximo de 39 mil dólares aproximadamente.
* Las casas con un garaje completado o no registrado tienen un valor alrededor de 286 mil dólares, mientras que aquellas parcialmente completadas o sin finalizar acumulan 271 mil dólares aproximadamente.
* Las casas que pertenezcan a los barrios Timber y Veenker de la zona tienen un valor aproximado de 515 mil dólares, menos que las casas de otros barrios con 567 mil dólares. Así mismo, las casas con un área de 1360 (29%) pies para el primer piso aportan 400 mil dólares aproximadamente. Así mismo, las casas con 2 a 3 (31%) dormitorios por encima del nivel del suelo pueden aportar un valor cercano a los 377 mil dólares.

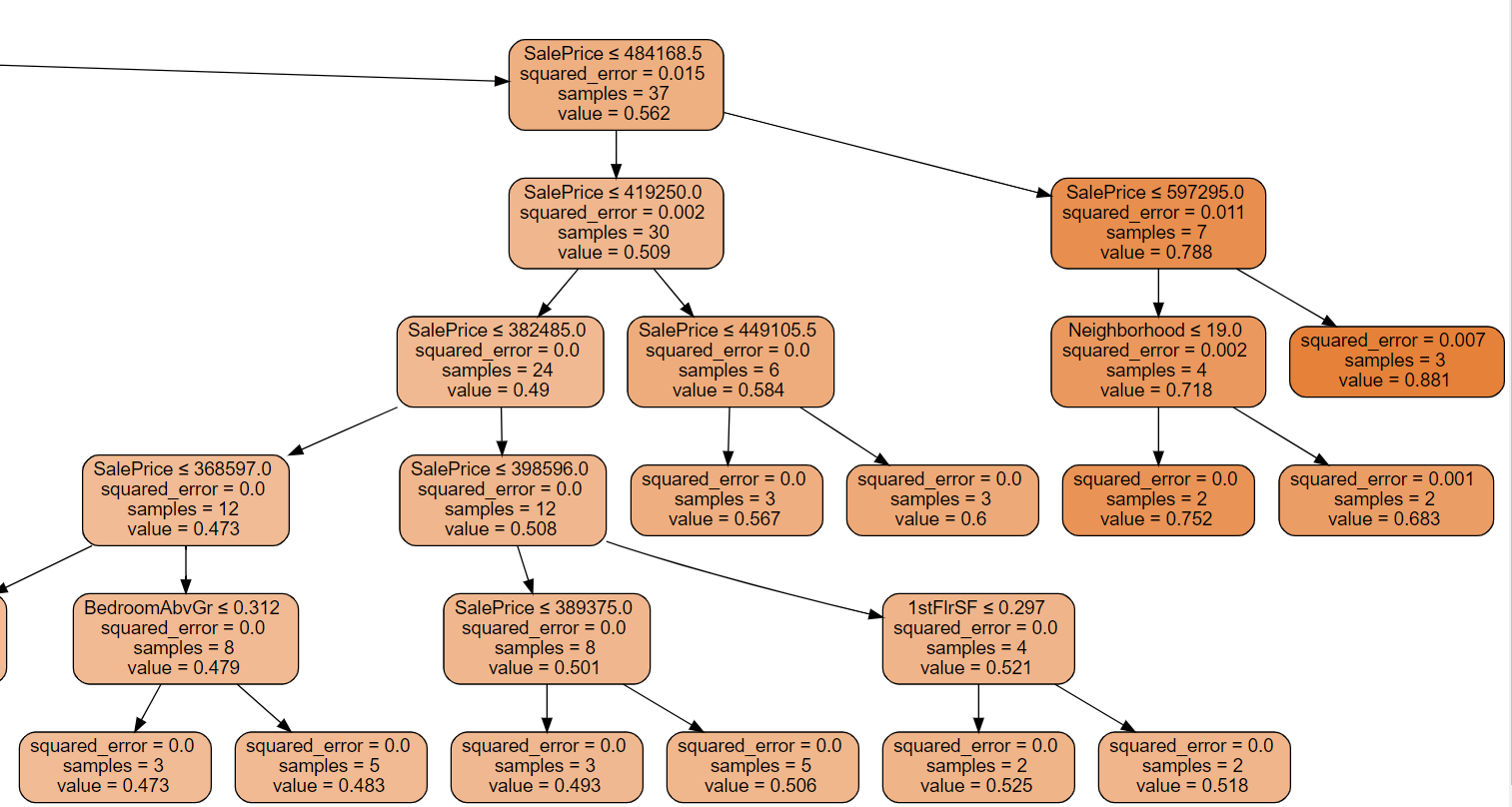


Figura 19: Porción del árbol resultante con variables categóricas y numéricas

* + 1. *Árbol de decisión con numéricas:*

El número máximo de niveles a crear en este árbol han sido 9, tomando como mínimo 3 muestras de información. Las cifras en este árbol, mayores comparadas a las del anterior, pueden deberse a que, sin la ayuda de categorías externas a las numéricas, no se le facilite el organizar los niveles, por lo que necesita ver más rangos de precio necesarios, como se contempla en la Figura 20. Los errores obtenidos de esta regresión han sido ligeramente menores a la anterior regresión, indicando que ambos caminos pueden ser válidos para la aplicación de árboles de decisión, como se observa en la Figura 21.

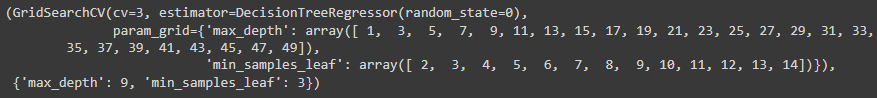


Figura 20: Resultados del estimador para el árbol de decisión numérico

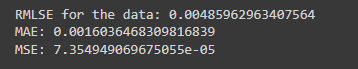


Figura 21: Los errores obtenidos de la regresión del árbol de decisión

El árbol de decisión generado, descrito en la Figura 22, da a conocer los distintos rangos de decisión y predicción del precio, así como las variables que lo complementan:

* Las casas vendidas antes de julio-agosto con un valor aproximado de 58 mil dólares llegan a los 60 - 66 mil dólares, en mayo llegando incluso de máximo a 185 dólares, mientras que en meses posteriores llegan a los 230 mil dólares. Las casas vendidas cerca del 2007 tuvieron 128 a 154 mil dólares de ganancia, mientras que las del 2008 al 2009 en 175 a 215 mil dólares. Las casas remodeladas en los 70 a 80 ganan 155 a 169 mil dólares; y las construidas en 1971 a 1993 ganan 162 a 180 mil $.
* Las casas con un área de garaje aproximados a 312 pies cuadrados (el 22% de 1416) se venden a 90 mil dólares, mientras aquellas de 425 (29%) pies cuadrados pueden ofrecerse a 120 mil $, 147 mil $ con 496 pies, 297 pies dando 155 mil dólares, 397 pies dan 190 mil dólares, 172 mil $ en 510 pies, 618 pies dan 200 $ y 794 pies dan 300 mil $.
* Los inmuebles con pies de calle en la parcela de 34 a 37 de área proporcionan 99 mil y 164 dólares del precio respectivamente, con aquellos de 53 a 60 pies de frente ganando 180 a 199 mil $. Así mismo, aquellas casas con un área de 516 a 563 pies cuadrados en el primer piso ganan 115 a 123 mil dólares, mientras que las de 938 a 1032 pies obtienen 150 a 250 mil dólares. Las de 392 a 743 pies cuadrados en el segundo piso ganan 220 a 310 mil dólares.
* Las casas con plataforma de madera de 59 a 68 pies (7 – 8%) ganan 147 a 210 mil dólares, con las de 128 pies (15%) ganando 286 mil $. Las de subclase 65 a 75 ganan 168 a 189 mil dólares. Las de condición general media (5.5) tienen 172 mil dólares de ganancia. Las casas con ninguna o al menos 1 (16%) chimenea ganan 174 mil dólares. Las casas con un área de revestimiento de mampostería de 64 pies (4%) ganan 196 dólares. Las casas con porche de 136 (25%) pies ganan 461 mil dólares respectivamente.
* Las casas con ninguno a un solo (%16) baño en el sótano ganan 259 mil dólares, las de pocos a un solo (25%) baño parcial ganan 318 mil dólares. Las casas con menos o igual a 2 cuartos aproximadamente se venden desde 82 mil a 370 mil dólares aproximadamente, y las que lo superen pueden llegar a los 380 mil dólares. Las casas con sótano finalizado de tipo 1 de 451 (8%) pies ganan 151 mil dólares; mientras que las casas con sótano sin terminar de 467 a 770 pies (20-33%) gana 162 a 233 mil dólares.

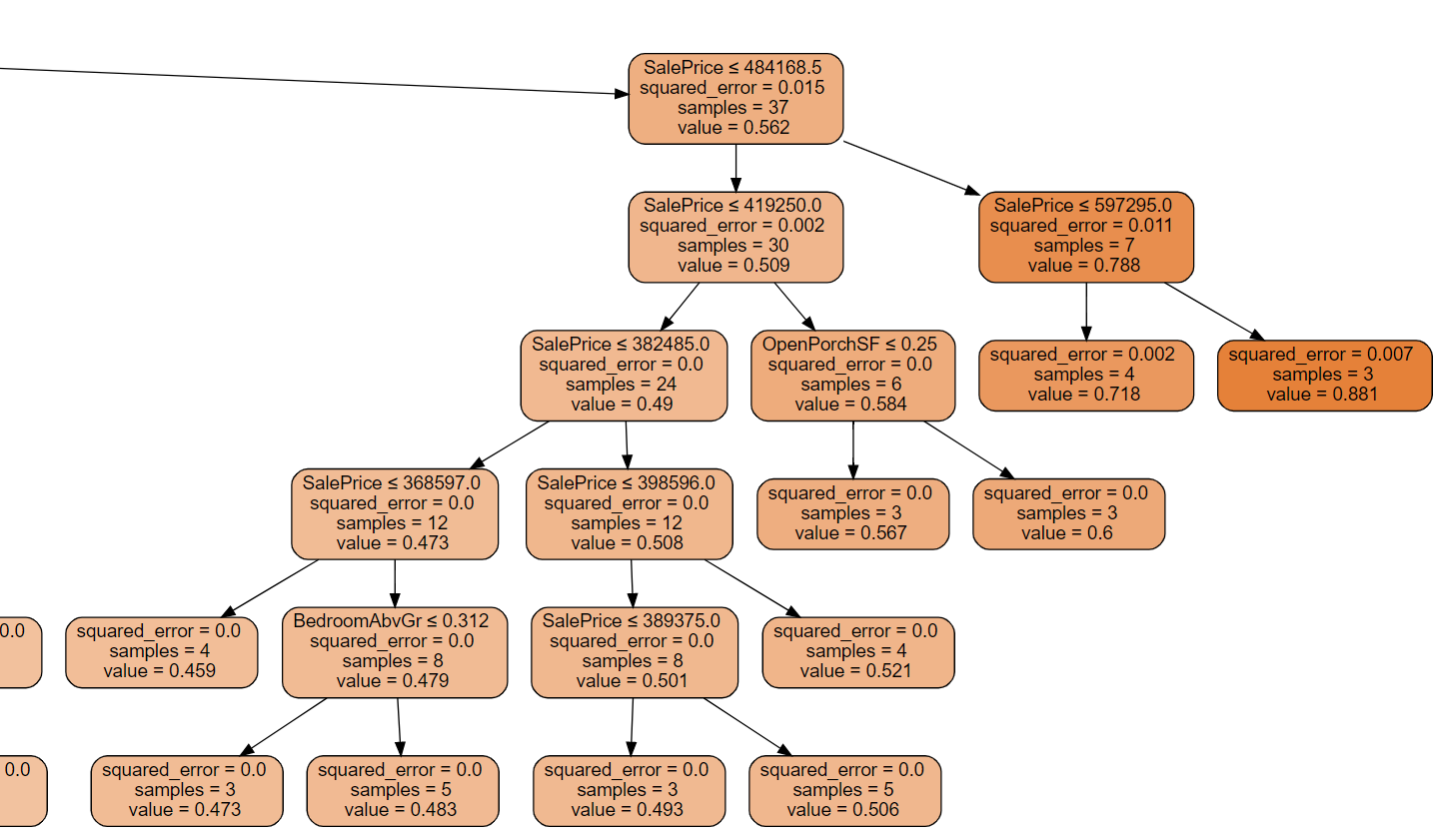


Figura 22: Porción del árbol resultante con variables numéricas

## Discusión de resultados:

A partir de las regresiones realizadas, se regresan a las hipótesis antes planteadas, y se puede observar que:

* Las casas han tenido valores más altos dentro del rango de mayo a septiembre, llegando incluso a cifras alrededor de 200 mil dólares.
* El garaje con los precios más altos predomina en aquellos con 794 pies cuadrados de área, más que en aquellos dentro del rango del 400 a 500. Sin embargo, hay más distribución de rangos de precios dentro del conjunto establecido, lo cual indica una frecuencia de áreas de garaje posible para este cálculo.
* Hay mayor precio ganado en casas que se vendieron VARIOS AÑOS luego de su construcción y/o remodelación que en aquellas que lo hacen en un lapso menor de tiempo. Esto talvez se deba al valor de antigüedad de la casa en sí, siendo muchas de estas de los 70’ a los 90’.
* El área de la parcela no influye mucho en el aumento del valor del precio, sólo en aquellos inmuebles más pequeños que el rango de 1300 a 300 mil pies cuadrados. Para los más grandes, el precio aumentaría más por los cuartos, el segundo piso, porches y el garaje.

Por lo tanto, adjunto a las comparaciones de los dos modelos, se puede determinar que, de todas las variables que pueden categorizar los precios, pueden ser más influyentes el área del garaje, los años de venta y construcción/remodelación, el área del primer piso y el área del sótano terminado de tipo 1; secundadas del barrio, el número de habitaciones dormitorio o baño, entre otras mencionadas en la interpretación de los árboles de decisión

## Conclusiones:

El precio de un inmueble en una zona específica puede ser estimado a partir de muchas posibles variables que lo acompañan con respecto a su construcción, la distribución de sus espacios y las características que le dan una mejor calidad. Como se observó en este laboratorio, es justificado tomar el área, tanto general como por sectores, como el factor más importante, ya que permite determinar no sólo el tamaño de las facilidades a construirse, sino también la capacidad de vivienda que puede tener el inmueble para los habitantes; junto a la capacidad de crear espacios en lugares aparte del suelo, siendo subterráneos (como los sótanos, en los que un área óptima permite dar muchos servicios y usos extra en la casa) como de más de un nivel; y el año de construcción: la valuación de una casa de una época pasada, si bien su calidad puede llegar a deteriorarse con el pasar de los años, puede adquirir más valor de dinero al tener una estructura que pueda ser limpiada, o incluso, remodelada dependiendo del estado del inmueble, pero sin dejar a un lado que los materiales y aspectos varios (más que todo históricos, como los muebles, o los sótanos antes construidos para ciertas situaciones de emergencia) tienen un valor.

Los árboles de regresión son una herramienta de aprendizaje muy útil para clasificar los diferentes tipos de rangos y predicciones acerca de una variable con respecto a otras, determinando el valor de esta cuando un atributo juega a su favor o contra, obteniendo así la mejor información posible acerca de este objetivo y así analizar su entorno. Gracias a estos métodos, se pudieron hacer inferencias de los mejores valores que determinan el precio de un inmueble, tomando siempre un camino que determine la mejor explicación posible de las variables que juegan a su favor, explicando de la mejor manera su impacto y posibles resultados.