Cursada: Datascience Comisión 22740

Conclusions generales Dataset Ames Housing:

Equipo de trabajo:

* Germán Hilbert
* Iván González Seguezzo
* José Nardulli

Profesor:

Luca Cittá Giordano

Tutor:

Juan Felipe González Sanmiguel

Objetivos Generales

En este trabajo, se desarrolla un análisis sobre los datos correspondientes a las viviendas pertenecientes a la ciudad de Ames situada en el condado de Story del estado de Iowa en Estados Unidos. Su población es de 66.427 habitantes según el censo de 2020 y tiene una longitud de 55,90 kilómetros cuadrados.

El “Dataset” contiene información acerca de características muy variadas sobre viviendas y donde cada una fue valuada a un determinado precio. Se entiende que las diversas cualidades como las dimensiones de terreno, las cantidad de habitaciones o la ubicación de cada casa influyen en el proceso de valuación. Teniendo esta premisa en claro, este trabajo tendrá como finalidad entender cómo se comportan las distintas variables y en qué medida influencian en el precio.

La información extraída puede ser de gran utilidad para agencias inmobiliarias y para potenciales compradores de viviendas en la zona bajo análisis.

Objetivos de la investigación:

Objetivo principal:

Realizar un análisis descriptivo que permita hallar las principales variables que pueden influir en el precio de las viviendas.

Objetivos secundarios:

Limpiar y transformar el dataset elegido para poder manipular la información de forma eficiente.

Extraer \*insights\* que permitan un mejor entendimiento de la determinación de los precios de los inmuebles.

Encontrar las correlaciones más importantes entre variables, ya sean directas o inversas.

Determinar qué variables categóricas pueden diferenciar distintos rangos de precio.

Adquisición de datos:

Fuente del dataset:

Realizamos una investigación por diferentes páginas afines a la ciencia de datos realizando una primera selección de Datasets con características necesarias para realizar el trabajo de este curso. En esta instancia cada integrante tuvo via libre a su imaginación.

Fundamentos de la elección:

De todos los encontrados realizamos una segunda selección teniendo en cuenta diferentes parámetros como:

Tema a tratar: Debe ser interesante para los 3 integrantes del grupo.

Objetivo y resultados: El Dataset debe permitir realizar un análisis completo.

Datos suficientes: El dataset debe contener muchos atributos y registros para facilitar el aprendizaje de nuestros algoritmos de Machine Learning.

Originalidad: El tema a tratar debe ser original en la comisión.

Calidad: El trabajo finalizado debe ser de calidad. Para poder formar parte de nuestro Portfolio.

Nos decidimos a trabajar sobre el tema viviendas. Un dataset que cuenta con 2930 registros y 80 atributos. Los atributos corresponden a diferentes características de cada vivienda.

https://www.kaggle.com/prevek18/ames-housing-dataset

Conclusiones obtenidas por la limpieza y manejo de datos:

Características generales del dataset:

* 2930 registros de viviendas.
* 80 columnas con atributos de cada vivienda. 1 columna adicional PID con número identificación único de vivienda.
* Tipos de dato:
  + 43 del tipo Object.
  + 11 del tipo Float.
  + 26 del tipo Int64.
* Valores nulos:
  + 53 columnas con datos completos (2930 not null).
  + 21 columnas con datos completos <95% (2771 a 2929 not null).
  + 1 columna con la mitad de datos faltantes (1508 not null).
  + 4 columnas con muchos datos null. Porcentaje de datos faltantes 80% (13 a 572 not null).

Conclusiones obtenidas en el manejo de Outliers

Existe casos que podrían contrariar la lógica inicial de la propuesta en la relación de precio y área y el precio y tamaño del garaje:

* 5 Viviendas con demasiada área y poco valor.
* 5 garajes enormes y con poco valor
* 4 viviendas con mucho valor y sin área de garaje
* 2 viviendas extremadamente económicas a pesar de poseer el comodato de garaje

Manejo de data perdida en el dataset:

* Realizamos análisis exhaustivo de los datos faltantes.
* Dependiendo cada caso con los datos faltantes podemos:
  + Mantenerlos.
  + Borrarlos.
  + Reemplazarlos por otro valor.
* Comenzamos analizando los missing <1%. Dado que son pocos los miramos en detalle para dar una solución individual.
* Se continua con los missing que comprenden de >2% y <17%. Se verifica casos particulares.
* Se finaliza con los missing >80%. Si no se pueden completar se dropean.
* Luego de verificar los PID y la documentación, podemos estimar que los valores nulos corresponden a que la vivienda no cuenta con SOTANO. Por lo que se procede a rellenar con 0 los valores numéricos y None los valores String.
* Observando la documentación deducimos que valores faltantes de Mas Vnr Type corresponden a que la vivienda no cuenta con revestimiento de mampostería. Por lo que rellenamos los valores por 0 y None respectivamente.

Conclusiones obtenidas en el EDA:

para el análisis general las correlaciones optimas son:

* Overall Qual & SalePrice
* Garage Yr Built & Year Built
* 1st Flr SF & Total BsmntSF1
* -otRms AbvGr & Gr LivArea
* Garage Cars & Garage Area

Conclusiones obtenidas en el análisis Univariado:

* El frente de lote se encuentra mayormente entre los valores 50 a 80 metros de longitud.
* En su mayoría son viviendas consideradas con calidad y condiciones en un nivel medio.
* Sólo hay algunas pocas casas con dos cocinas, más del 90% de las casas vendidas poseen una sola cocina.
* Las casas que tienen capacidad para dos autos en su garaje supera el doble de las viviendas que poseen sólo uno.
* La cantidad de habitaciones en la totalidad de las casas es mayormente de 6.
* Las casas que poseen 3 habitaciones son mayores en cantidad a las viviendas que poseen 2 y 4 habitaciones juntas.
* La mayoría de las transacciones de venta se concretaron a mitad de año, en los meses de mayo, junio y julio.

Precio de venta:

* los precios generales de las propiedades examinadas rondan entre los 120000$ y los 755000$ con la mayor parte de las propiedades ubicándose en el rango de los 213500$

Conclusiones obtenidas en el análisis Bivariado:

* se observan que hay tendencias obvias que afectan principalmente el precio de venta, el tamaño del lote la cantidad de comodatos (piscina garajes), aumentan el precio total solapando a otro tipo de data como la ubicación y se precie más el tamaño de áreas cubiertas que de áreas verdes y los comodatos más valorados seria la piscina seguida por el garaje
* propiedades más nuevas por regla general presentan un costo superior a propiedades antiguas con marcadas excepciones
* podría marcarse como una obviedad, pero la calidad general impacta directamente sobre el precio del lote, a mayor calidad mayor precio

Tipos de viviendas:

* 20: 1 PISO - 1946 Y MÁS NUEVOS TODOS LOS ESTILOS
* 30: 1 PISO - 1945 Y MÁS ANTIGUOS
* 40: 1 PISO CON ÁTICO TERMINADO TODAS LAS EDADES
* 45: 1-1/2 HISTORIA - SIN TERMINAR TODAS LAS EDADES
* 50: 1-1/2 HISTORIA TERMINADA TODAS LAS EDADES
* 60: 2 PISOS 1946 Y MÁS NUEVOS
* 70: 2 PISOS 1945 Y MÁS ANTIGUOS
* 75: 2-1/2 HISTORIA TODAS LAS EDADES
* 80: DIVIDIDOS O MULTINIVEL
* 85: VESTÍBULO DIVIDIDO
* 90: DÚPLEX - TODOS LOS ESTILOS Y EDADES
* 120: PUD DE 1 PISO (Desarrollo de unidades planificadas) - 1946 Y MÁS RECIÉN
* 150: 1-1/2 STORY PUD - TODAS LAS EDADES
* 160: PUD DE 2 PISOS - 1946 Y MÁS NUEVOS
* 180: PUD - MULTINIVEL - INCLUYE NIVEL DIVIDIDO/VESTÍBULO
* 190: 2 CONVERSIÓN FAMILIAR - TODOS LOS ESTILOS Y EDADES

Los tipos de casa 160 y 180 son las que tiene menor longitud en el frente de lote. Mientras que los tipos 20, 60, 75 y 80 tienen los frentes de lote más grandes.

Las viviendas con 2 o más baños completos en planta baja conservan precios bastante más elevados que las que tienen 1 o ninguno. A su vez, la mayoría de los datos se encuentran en las casas que tiene uno o dos baños completos.

Cuántos más autos entran en el garaje de cada vivienda, mayor es la superficie que el mismo ocupa.

Se observa una correlación positiva fuerte entre ambas variables. Cuántas más habitaciones tiene la casa, mayor es la superficie que las mismas ocupan.

Condiciones de venta:

* Normal: Venta normal
* Abnorml: Venta anormal. Comercio, ejecución hipotecaria, venta corta
* AdjLand: Adquisición de terrenos adyacentes
* Alloca: Asignación. Dos propiedades vinculadas con escrituras separadas, generalmente condominio con una unidad de garaje
* Family: Venta entre miembros de la familia
* Partial: La vivienda no se completó cuando se evaluó por última vez (asociada con viviendas nuevas)

Las ventas calificadas como parciales fueron pagadas por las casas con un precio mayor al resto. Mientras que las que se adquirieron como terrenos adyacentes, el precio pagado fue el menor.

Barrios:

* Blmngtn: Bloomington Heights
* Blueste: Bluestem
* BrDale: Briardale
* BrkSide: Brookside
* ClearCr: Clear Creek
* CollgCr: College Creek
* Crawfor: Crawford
* Edwards: Edwards
* Gilbert: Gilbert
* IDOTRR: Iowa DOT and Rail Road
* MeadowV: Meadow Village
* Mitchel: Mitchell
* Names: North Ames
* NoRidge: Northridge
* NPkVill: Northpark Villa
* NridgHt: Northridge Heights
* - NWAmes: Northwest Ames
* OldTown: Old Town
* SWISU: South & West of Iowa State University
* - Sawyer: Sawyer
* SawyerW: Sawyer West
* Somerst: Somerset
* StoneBr: Stone Brook
* - Timber: Timberland
* Veenker: Veenker

Las casas más caras su ubican mayormente en: Stone Brook, Northridge Heights y Northridge. Algunos de los barrios más accesibles para adquirir una casa son: Meadow Village, Iowa DOT and Rail Road y Briardale

Tipos de calle:

* Pave: Calle pavimentada
* Grvl: Camino de gravilla

Las casas ubicadas en calles pavimentadas tienen en general un precio más elevado que las ubicadas sobre caminos de gravilla.

Tipo de Zona:

* A: Agricultura
* C: Comercial
* FV: Residencial de pueblo flotante
* I: Industrial
* RH: Residencial Alta Densidad
* RL: Residencial Baja Densidad
* RP: Parque Residencial de Baja Densidad
* RM: Residencial Media Densidad

Las viviendas ubicadas en zona de Residencial de pueblo flotante son más caras que el resto de las otras zonas. Mientras que las casas en zonas agrícolas, industriales y comerciales son las más baratas. El resto del tipo de zonas residenciales se encuentran en el medio de ambos extremos.

Las viviendas construidas a partir del año 1980 en adelante mayormente tienen una condición media respecto a la totalidad. Por otro lado, las viviendas construidas antes del 1970 tienen condiciones muy heterogéneas. Hay casas en muy buen estado y otras en condiciones muy malas.

es una tendencia generalizada que la las casas con mejor calidad general son aquellas que tienen solo un piso o nivel mientras que en las casas con más pisos la tendencia de calidad general suele ser descendente.

Conclusiones obtenidas del análisis Multivariado:

la tendencia observada en el análisis bivariado se mantiene:

-se observan que hay tendencias obvias que afectan principalmente el precio de venta, el tamaño del lote la cantidad de comodatos (piscina garajes), aumentan el precio total solapando a otro tipo de data como la ubicación y se precie más el tamaño de áreas cubiertas que de áreas verdes y los comodatos más valorados serian la piscina seguida por el garaje

-propiedades más nuevas por regla general presentan un costo superior a propiedades antiguas con marcadas excepciones

-podría marcarse como una obviedad, pero la calidad general impacta directamente sobre el precio del lote, a mayor calidad mayor precio

precio de venta año de construcción y calidad general:

Se pueden observar algunas tendencias de agrupamiento de datos en determinadas zonas. Hay una correlación positiva débil entre el precio de venta y el año de construcción. La calidad general está más relacionada con el precio de venta que con el año de construcción, porque se pueden observar más conformaciones de agrupamientos en forma de barras verticales que horizontales. Lo que significa que a medida que el precio crece la calidad va aumentando, mientras que por el lado del año de construcción se pueden ver calidades muy distintas para viviendas construidas durante la misma fecha.

Tipo de Zona:

* A: Agricultura
* C: Comercial
* FV: Residencial de pueblo flotante
* I: Industrial
* RH: Residencial Alta Densidad
* RL: Residencial Baja Densidad
* RP: Parque Residencial de Baja Densidad
* RM: Residencial Media Densidad

Hay una correlación muy leve positiva entre la calidad general de las viviendas con su año de construcción. Las viviendas construidas en los últimos años tienden a tener una mayor calidad que las más antiguas. En esta misma zona de la gráfica las viviendas pertenecen al tipo de zona Residencial Baja Densidad. Mientras que algunas pocas casas que son antiguas y son consideradas de muy buena calidad pertenecen a la zona Residencial Media Densidad.

Estilo de vivienda:

* 1Story: Un piso
* 1.5Fin: Un piso y medio, 2do nivel terminado
* 1.5Unf: Un piso y medio, 2do nivel sin terminar
* 2Story: dos pisos
* 2.5 Fin: Dos pisos y medio, 2do nivel terminado
* 2.5Unf: Dos pisos y medio, 2do nivel sin terminar
* Foyer: Hall dividido
* SLvl: Construido sobre dos niveles

La mayor condición general asignada a las viviendas fue de 5. La mayoría de los datos con esta condición tiene capacidad en el garaje para dos autos, y en menor medida en otros casos, para uno y para tres. Los que tienen capacidad para dos autos son mayormente casas de uno o de dos pisos. Mientras que los que tienen capacidad sólo para un vehículo tienen generalmente un solo piso de vivienda.