# Descripción del Dataset

Este dataset se denomina “Ames Housing” y mediante 79 variables, describe todos los aspectos que influyen en la valoración de compra/venta de viviendas residenciales en Ames ( Iowa, EEUU). Mediante estas variables podemos determinar el precio futuro de una vivienda.

El dataset ha sido elaborado por Dean De Cock con el objetivo de educar en ciencia de datos, siendo una alternativa al famoso dataset “Boston Housing”. Por ello, consideramos este dataset una excelente elección para la elaboración de esta práctica.

El dataset se compone de 4 ficheros, de los cuales, únicamente emplearemos train.csv

* Train.csv : destinado a crear el modelo
* Test.csv : destinado a testear el modelo
* Data description: descripción de cada atributo.
* Sample\_submision.csv : ejemplo de referencia de una regresión lineal por año y mes de venta, superficie y numero de dormitorios.

## Atributos

A continuación, se describe cada atributo para una mejor comprensión de los mismos y poder elaborar las preguntas a responder. Adicionalmente, en el Anexo 1.A, se adjunta una tabla, con cada atributo, el tipo de dato y un ejemplo.

1. MSSubClass: identifica el tipo de vivienda involucrada en la venta.
2. MSZoning: identifica la clasificación general de zona
3. LoteFrontage: pies lineales de calle conectados a la propiedad.
4. **LotArea: Tamaño del lote en pies cuadrados.**
5. Street: Tipo de vía de acceso a la propiedad
6. Alley: Tipo de callejón de acceso a la propiedad.
7. LotShape: forma general de la propiedad.
8. LandContour: Planitud de la propiedad.
9. Utilities: tipo de instalaciónes disponibles.
10. LotConfig: configuración del lote.
11. LandSlope: Pendiente de la propiedad.
12. **Neighborhood: ubicaciones físicas dentro de los límites de la ciudad de Ames.**
13. Condition 1: proximidad a varias condiciones.
14. Condition 2: Proximidad a varias condiciones (si más de una está presente).
15. BldgType: Tipo de vivienda.
16. HouseStyle: estilo de vivienda.
17. **OverallQual: Califica el material general y el acabado de la casa.**
18. **OverallCond: Califica el estado general de la casa.**
19. **YearBuilt: fecha de construcción original.**
20. RoofStyle: Tipo de techo.
21. RoofMatl: material del techo.
22. Exterior1st: Revestimiento exterior de la casa.
23. Exterior2nd: Revestimiento exterior de la casa (si hay más de un material).
24. MasVnrType: tipo de chapa de mampostería.
25. MasVnrArea: Área de revestimiento de mampostería en pies cuadrados.
26. ExterQual: Evalúa la calidad del material en el exterior.
27. ExterCond: Evalúa el estado actual del material en el exterior.
28. Foundation: Tipo de fundación.
29. BsmtQual: Evalúa la altura del sótano.
30. BsmtCond: Evalúa el estado general del sótano.
31. BsmtExposure: se refiere a muros a nivel de jardín o de huelga.
32. BsmtFinType1: Calificación del área terminada del sótano.
33. BsmtFinSF1: pies cuadrados terminados tipo 1
34. BsmtFinType2: Calificación del área terminada del sótano (si hay varios tipos).
35. BsmtFinSF2: pies cuadrados terminados tipo 2.
36. BsmtUnfSF: pies cuadrados sin terminar de área del sótano
37. TotalBsmtSF: Total de pies cuadrados de área del sótano.
38. **Heating: Tipo de calefacción.**
39. HeatingQC: calidad y estado de la calefacción.
40. CentralAir: aire acondicionado central.
41. Electrical: sistema eléctrico.
42. 1stFlrSF: pies cuadrados del primer piso.
43. 2ndFlrSF: pies cuadrados del segundo piso.
44. LowQualFinSF: pies cuadrados con acabado de baja calidad (todos los pisos).
45. GrLivArea: pies cuadrados de área habitable sobre el nivel (suelo).
46. BsmtFullBath: Baños completos en el sótano.
47. BsmtHalfBath: Medios baños del sótano.
48. **FullBath: baños completos sobre rasante.**
49. HalfBath: Medios baños sobre el nivel del suelo.
50. **Bedroom: dormitorios sobre el nivel del suelo (NO incluye dormitorios en el sótano).**
51. **Kitchen: Cocinas sobre rasante.**
52. KitchenQual: calidad de la cocina.
53. **TotRmsAbvGrd: Total de habitaciones sobre rasante (no incluye baños).**
54. Functional: funcionalidad del hogar (suponga que es típico a menos que se justifiquen las deducciones).
55. Fireplaces: Número de chimeneas.
56. FireplaceQu: Calidad de chimenea.
57. GarageType: Ubicación del garaje.
58. GarageYrBlt: año en que se construyó el garaje.
59. GarageFinish: Acabado interior del garaje.
60. **GarageCars: Tamaño del garaje en capacidad de automóvil.**
61. GarageArea: Tamaño del garaje en pies cuadrados.
62. GarageQual: Calidad de garaje.
63. GarageCond: Estado del garaje.
64. PavedDrive: entrada pavimentada.
65. WoodDeckSF: Área de la plataforma de madera en pies cuadrados.
66. OpenPorchSF: Área de porche abierto en pies cuadrados.
67. EnclosedPorch: área de porche cerrado en pies cuadrados.
68. 3SsnPorch: área de porche de tres estaciones en pies cuadrados.
69. ScreenPorch: Área del porche de la pantalla en pies cuadrados.
70. PoolArea: Área de la piscina en pies cuadrados.
71. PoolQC: Calidad de la piscina.
72. Fence: calidad de la valla.
73. MiscFeature: característica miscelánea no cubierta en otras categorías.
74. MiscVal: $ Valor de la característica miscelánea.
75. MoSold: Mes vendido (MM).
76. YrSold: año vendido (YYYY).
77. SaleType: Tipo de venta.
78. **SaleCondition: Condición de venta.**
79. **SalePrice. Precio de venta**

## Objetivo

Para establecer los objetivos del análisis, extrapolamos el ejercicio a un entorno del mundo, tomando la hipótesis de que el interés del análisis es el de una inmobiliaria. Con ello, dado el gran volumen de atributos, realizaremos una selección previa de los mismos, seleccionando los 12 atributos más relevantes para el estudio

En primer objetivo es realizar es realizar un estudio descriptiva para obtener un resumen de las características de la muestra y obtener una visión general de los datos necesarios previos antes de realizar la limpieza de los datos. También pretendemos dar una visión general de la consistencia del dataset, visualizando valores extremos, nulos, incongruentes o mal formateados.

Tras ello, se construirá un modelo de regresión múltiple que permita determinar el precio de la vivienda, a partir de los atributos seleccionados.

Finalmente, nos planteamos las siguientes preguntas a responder:

* ¿Cuál es el precio de la propiedad en función de las variables OverallQual, OverallCond, LotArea, Yearbuilt?
* ¿Se puede asumir que las casas del barrio CollgCr tienen un mayor lote de área que las de BrkSide?
* ¿La varianza del precio de venta es diferente de 70442^2?
* ¿Cuál es el grado de correlación de OverallQual, OverallCond, LotArea, Yearbuilt respecto a SalePrice?
* Dados los valores lotArea = 8650, overallCond = 6, overallQual = 5, yearBuilt = 2004, cuál es el valor de venta de la propiedad?

# Integración y selección de los datos a analizar

En este punto se describen los primeros pasos que se deben realizar antes de la limpieza de los datos y su posterior análisis, es decir, la lectura y visionado previo del dataset. Se han seguido los siguientes pasos:

1. Librerias

Lo primero será siempre cargar las librerías que emplearemos en la practica, siendo las siguientes:

Texto

Descripción generada automáticamente

1. Carga del dataset

Para ello, se ha empleado la función read\_csv() de la librería pandas, para obtener un dataframe con el que trabajaremos en toda la practica. Dado que el fichero se encuentra en la misma carpeta que el archivo .ipynb, no es necesario especificar ruta.

Diagrama

Descripción generada automáticamente con confianza baja

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

1. Visualización de las cabeceras

Para hacernos una idea de los datos que componen el dataset, se ha empleado la función head(), para visualizar las cabeceras.

Tabla

Descripción generada automáticamente con confianza baja

Posteriormente, para guardar esta información en un CSV (comprimido ) y elaborar el Anexo 1.A, se ha empleado el siguiente código:

Texto

Descripción generada automáticamente

1. Tipología de los datos

El siguiente paso es ver el tipo de datos que tiene el dataset, esta información es de gran utilidad para el análisis. Esta información la obtenemos mediante dtypes(). Dado el gran número de atributos, en el Anexo 1.A, se pueden ver todos los atributos, a continuación se muestran solo algunos.

Tabla

Descripción generada automáticamente con confianza media

Para elaborar el Anexo 1.A se ha guardado esta información en un CSV mediante el siguiente código:

Texto

Descripción generada automáticamente

## Selección de datos

Una vez que tenemos una imagen general del dataset y como se ha mencionado en el punto anterior, se va a realizar una preselección de los atributos más relevantes para la elaboración de la práctica. Estos atributos serán los siguientes:

|  |  |  |
| --- | --- | --- |
| Id | Atributo | Descripcion |
| 4 | LotArea | Tamaño del lote en pies cuadrados. |
| 12 | Neighborhood | ubicaciones físicas dentro de los límites de la ciudad de Ames. |
| 17 | OverallQual | Califica el material general y el acabado de la casa. |
| 18 | OverallCond | Califica el estado general de la casa. |
| 19 | YearBuilt | fecha de construcción original. |
| 38 | Heating | Tipo de calefacción. |
| 48 | FullBath | baños completos sobre rasante. |
| 50 | Bedroom | dormitorios sobre el nivel del suelo (NO incluye dormitorios en el sótano). |
| 51 | Kitchen | Cocinas sobre rasante. |
| 53 | TotRmsAbvGrd | Total, de habitaciones sobre rasante (no incluye baños). |
| 60 | GarageCars | Tamaño del garaje en capacidad de automóvil. |
| 78 | SaleCondition | Condición de venta. |
| 79 | **SalePrice** | Precio de venta |

**SalePrice**

Una vez que tenemos los atributos que nos interesan, eliminados del dataset el resto.

Para ello, empleamos el comando drop, después de una sentencia if para evitar errores. La siguiente imagen es un ejemplo:

Texto

Descripción generada automáticamente

Finalmente, empleamos el método describe para obtener un resumen estadístico de los atributos que vamos a emplear

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

# 3 Limpieza de Datos

## 3.1. Valores nulos

En primar lugar empleamos la función isnull para evaluar el dataframe y determinar si el valor que se le pasa como argumento no coincide ( “True”) o si coincide (“False”)

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Después empleamos un “loop for” para determinar el numero de valores que faltan para cada tributo; True significa que el valor falta y False que el valor está.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Como vemos en los resultados, no hay ningún valor nulo

|  |  |
| --- | --- |
|  |  |

Lo anterior podemos verlo visualmente de la siguiente manera; comprobando que no hay ningún valor nulo:

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Tabla

Descripción generada automáticamente

Lo anterior, también sirve para comprobar que el numero de filas es 1400

## 3.2. Valores Extremos

**ANEXO 1.A.**

|  |  |
| --- | --- |
|  |  |