



16 DE MARZO DE 2023

# PRÁCTICA 1

Análisis Exploratorio de Datos

## Objetivo de la práctica

Por Angel Damian Monroy Mendoza

No. de Cuenta: 316040707

Tener una idea de la estructura del conjunto de datos, identificar la variable objetivo y posibles técnicas de modelado. En este caso en particular, encontrar información de interés para predecir la próxima tendencia inmobiliaria en Melbourne.



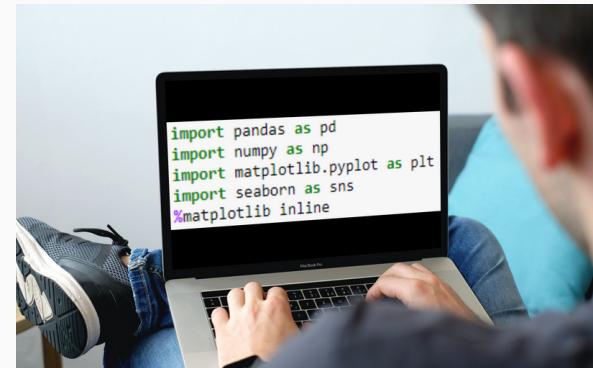
## Características:

- El sector inmobiliario de Melbourne, Australia continúa en auge desde hace algunos años. Con esto se vuelve un aspecto de interés el conocer la tendencia inmobiliaria en dicha ciudad debido a que cada vez es más difícil adquirir una unidad de 2 dormitorios a un precio razonable.
  - A continuación, se muestra un diccionario de las variables relevantes para esta práctica
1. Rooms: Número de habitaciones.
  2. Price: Precio en dólares.
  3. Method: S (propiedad vendida), SP (propiedad vendida antes), PI (propiedad transferida), VB (oferta del proveedor) y SA (vendida después de subasta).
  4. Type: h (casa, cabaña, villa, semi, terraza), u (unidad, dúplex), t (casa adosada).
  5. SellerG: Agente de bienes raíces.
  6. Date: Fecha de venta.
  7. Distance: Distancia del CBD (Centro de negocios).
  8. Regionname: Región general (oeste, noroeste, norte, noreste, etc.).
  9. Propertycount: Número de propiedades que existen en el suburbio.
  10. Bedroom2: Número de dormitorios (de otra fuente).
  11. Bathroom :Cantidad de baños.
  12. Car: Número de estacionamientos.
  13. Landsize: Tamaño del terreno.
  14. BuildingArea: Tamaño del edificio.
  15. CouncilArea: Consejo de gobierno de la zona (Municipio).

# Desarrollo

Para el desarrollo de esta práctica dividimos el proceso en cinco secciones:

1. Importación de las bibliotecas y datos.
2. Paso 1: Descripción de la estructura de los datos
3. Paso 2: Identificación de datos faltantes.
4. Paso 3: Detección de valores atípicos.
5. Paso 4: Identificación de relaciones entre pares variables.
6. Preparación de la data y conclusiones.



## 1. Bibliotecas y Datos

En este primer apartado se importó la biblioteca de pandas para la manipulación y análisis de los datos, la biblioteca de numpy para crear vectores y matrices de n dimensiones, la biblioteca de matplotlib.pyplot para la generación de gráficas a partir de los datos y la biblioteca de seaborn que, de igual forma, corresponde a la visualización de los datos.

Posteriormente, se cargaron los datos correspondientes a Melbourne Housing Snapshot de Kaggle, los cuales se leyeron como archivo '.csv' con la biblioteca de pandas. Aquí, podemos observar la descripción y el significado de cada columna, así como observar los datos a priori con ayuda de la función '.head()' o '.tail()', las cuales nos muestran los primeros o últimos 10 elementos, respectivamente .

## 2. Paso Uno: Descripción de la Estructura de los Datos

En este primer paso, para poder describir los datos de una forma correcta, necesitamos conocer la dimensión de nuestro conjunto así como los tipos de datos con los que se cuenta, pues de esta forma podemos establecer un punto de partida que nos ayudará a comprender el comportamiento de los datos.

	Suburb	Address	Rooms	Type	Price	Method
0	Abbotsford	85 Turner St	2	h	1480000.0	
1	Abbotsford	Bloomberg St	2	h	1035000.0	
2	Abbotsford	5 Charles St	3	h	1465000.0	
3	Abbotsford	40 Federation La	3	h	850000.0	
4	Abbotsford	55a Park St	4	h	1600000.0	
5	Abbotsford	129 Charles St	2	h	941000.0	
6	Abbotsford	124 Yarra St	3	h	1876000.0	
7	Abbotsford	98 Charles St	2	h	1636000.0	
8	Abbotsford	6/241 Nicholson St	1	u	300000.0	
9	Abbotsford	10 Valiant St	2	h	1097000.0	

# Desarrollo

## 2. Paso Uno: Descripción de la Estructura de los Datos

Ahora bien, para saber la dimensión de nuestro conjunto de datos bastó con utilizar el atributo shape de Pandas, que proporciona la estructura general de los datos debido a que devuelve la cantidad de filas y columnas que tiene el conjunto de datos, que en este caso fue de 13580 filas por 21 columnas.

En el caso de los tipos de datos que se tienen, el atributo dtypes muestra los tipos de datos de las variables, en donde se observó que el conjunto de datos tiene una combinación de variables categóricas (objeto) y numéricas (flotante e int).

## 3. Paso Dos: Identificación de Datos faltantes

Para este paso también se tienen dos formas para poder realizarlo. La primera de ellas es mediante la función de Pandas isnull().sum() regresa la suma de todos los valores nulos (faltantes) de cada variable. Aquí podemos observar que la suma de los valores nulos en las características 'Car', 'BuildingArea', 'YearBuilt' y 'CouncilArea' dieron valores diferentes de cero.

La otra forma es a través de la función info() para obtener el tipo de datos y la suma de valores nulos. En pocas palabras, con esta función podemos saber el tipo de datos que tenemos (Paso Uno) y si es que existen valores nulos (Paso Dos), sin embargo, la información que nos aporta es muy superficial, pues no nos dice la cantidad de datos faltantes y no podemos realizar análisis como el siguiente.

En especial, 'BuildingArea' y 'YearBuilt', presentan valores nulos arriba de 5000, los cuales pueden representar un problema para el análisis de los datos, pues estamos hablando que, para estas variables, casi el 50% de sus datos son datos faltantes, lo que puede desembocar en resultados de modelos sesgados por la falta de información, encontrándonos con uno de los tres grandes retos de la minería de datos en la actualidad.

DatosMelbourne.shape  
(13580, 21)  
DatosMelbourne.dtypes

Suburb	object
Address	object
Rooms	int64
Type	object

DatosMelbourne.isnull().sum()

Suburb	0
Address	0
Rooms	0
Type	0
Price	0
Method	0
SellerG	0
Date	0
Distance	0
Postcode	0
Bedroom2	0
Bathroom	0
Car	62
Landsize	0
BuildingArea	6450
YearBuilt	5375
CouncilArea	1369
Latitude	0
Longitude	0
Regionname	0
Propertycount	0
dtype:	int64



# Desarrollo

## 4. Paso Tres: Detección de Valores Atípicos

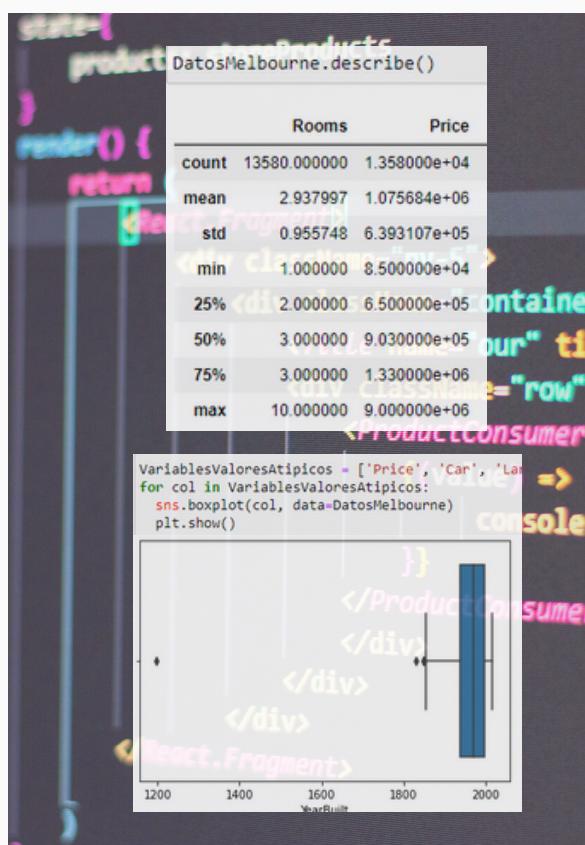
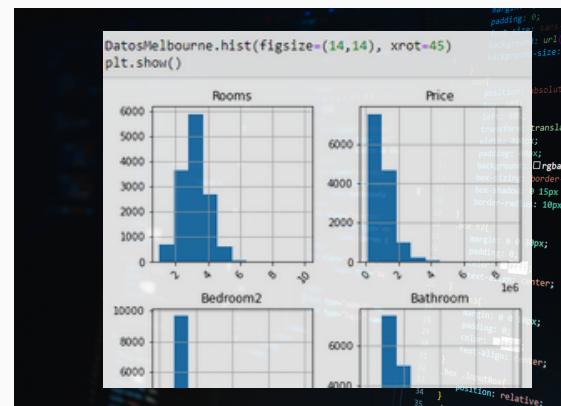
En este paso, tenemos varias formas de identificar valores atípicos, pues podemos hacerlo mediante gráficas de distribución, gráficas de bigote y mediante resúmenes estadísticos, tanto para las variables numéricas, como para las categóricas.

La razón por la que se utilizan las gráficas de distribución es para tener una idea de cómo se están comportando los datos o con qué frecuencia ocurren. Para las variables numéricas, se observa cuántas veces aparecen grupos de números en una columna. Mientras que, para las variables categóricas, son las clases de cada columna y su frecuencia.

Específicamente, se utilizan histogramas porque agrupan los números en rangos acotados que permiten realizar un análisis previo; aquí la altura de una barra muestra cuántos números caen en ese rango, en donde se emplea la función `hist()` para trazar el histograma de las variables numéricas.

Aquí logramos observar que 'BuildingArea', 'LandSize' y 'Price' son variables que están sesgadas a la izquierda, lo cual podría indicar que hay valores atípicos en estas características. Asimismo, la variable 'YearBuilt' está sesgada a la derecha y tiene la particularidad de comenzar con un límite en 1200, por lo que lo más probable es que ese dato sea un dato erróneo y que si se deja, afectará considerablemente los resultados.

Una forma de poder investigar estos datos atípicos es mediante un resumen estadístico que se obtiene mediante la función `'.describe()'`. Este resumen se muestra mediante una tabla en donde se muestra un recuento, la media, la desviación, el valor mínimo, el valor máximo y los percentiles 25%, 50% (mediana) y 75% y aquí observamos que nos ayuda a identificar las variables con valores nulos como el caso de 'Car', 'LandSize' y 'YearBuilt'.



# Desarrollo

## 4. Paso Tres: Detección de Valores Atípicos

Otra herramienta que nos permite observar con más detalle nuestras sospechas de valores atípicos son los diagramas de cajas o de bigote, pues aquí se muestran explícitamente todos los valores atípicos que salen de la varianza esperada de la variable, por lo que podemos confirmar que existen valores atípicos en las variables de 'LandSize', 'BuildingArea', 'YearBuilt' y 'Price', no obstante, en este último sólo se presumen 3 valores fuera de rango a pesar de que hay más, pero como éstos no tienen una separación notable, se toman como valores que son coherentes y significativos para nuestro conjunto de datos.

Por último, este proceso también se hace para las variables categóricas, en las cuales los resúmenes estadísticos muestran números reales que ayudan a obtener información puntual de estas variables, mientras que las gráficas sí nos ayuda a tener una idea general de las distribuciones categóricas.

Para obtener estos resúmenes basta con darle a la función '.describe()' el parámetro 'include='object'' y así nos dará como resultado una tabla que muestra el recuento de los valores, el número de clases únicas, la clase más frecuente y la frecuencia de esta clase en los datos. Aquí observamos que algunas clases cuentan con demasiados valores únicos como 'Address', seguida de 'Suburb' y luego 'SellerG'.

Finalizando con las gráficas, se utiliza Seaborn que permite representar cada clase en una barra de histograma y la forma en cómo se programa es mediante un for para el control y distribución de las clases con valores únicos menores a 10.

## 5. Paso Cuatro: Identificación de Relaciones entre Variables

En este paso se analiza la correlación que existe entre las variables. Esta correlación es un valor que está entre -1 y 1 y explica qué tanto aumenta (o disminuye) una variable mientras que la otra aumenta (o disminuye). En dado caso que el valor de la correlación sea cercano a 0, es que no varían de la misma forma dichas variables.

# Desarrollo

## 5. Paso Cuatro: Identificación de Relaciones entre Variables

En este sentido, una forma de saber las relaciones que hay es mediante una matriz de correlación que se basa en el método de Pearson y se obtiene con la función 'corr()'. A pesar de que esta matriz nos ayuda a saber las relaciones que hay entre variables numéricas, se vuelve confuso identificar estas relaciones cuando se tienen muchas variables, por lo que se auxilia del mapa de calor de Seaborn, mostrando la diagonal inferior de la matriz con la función 'sns.heatmap()' y pasando como parámetros la matriz de correlación de nuestro conjunto de datos, los colores que se quieren utilizar, la diagonal que se quiere mostrar y los valores de la correlación en cada celda.

## 6. Preparación de los Datos\*

Como parte final del desarrollo de la práctica se realizó una preparación de los datos que consistió en limpiar los valores nulos con la función '.dropna()', luego una limpieza de los valores atípicos con ayuda de los diagramas de caja y, por último, una limpieza de las correlaciones entre pares de variables, en donde nos apoyamos de los resultados del mapa de calor, terminando con un dataframe de nombre 'DatosLimpios', que guarda el resultado de los procesos anteriores.

# Conclusiones

De esta práctica podemos concluir que el *Análisis Exploratorio de Datos* es una parte fundamental tanto en la minería de datos, como en la aplicación de algoritmos de inteligencia artificial. Aquí podemos observar el comportamiento que tienen los datos, pues se nos permite identificar valores de interés en cada variable, distribuciones de las mismas, identificar valores atípicos, valores nulos y correlaciones entre pares de variables que pudieran afectar los resultados. Asimismo, podemos darnos cuenta de que se nos presentan los retos de la información redundante (correlaciones) y de la falta de información (datos nulos), pues vemos los efectos negativos tan solo en el resultado de limpiar los datos, ya que obtuvimos una matriz de (6193x20) cuando teníamos una de (13580x21).

\*Nota: no se presentó una explicación detallada del proceso de limpieza de los datos, debido a que ésta se encuentra en el cuaderno de python.