Imagen que contiene Código QR

Descripción generada automáticamente

[Introducción 3](#_Toc113804567)

[EDA – Análisis Exploratorio de Datos 3](#_Toc113804568)

[Proceso básico de EDA 4](#_Toc113804569)

[Variables 4](#_Toc113804570)

[Contexto 4](#_Toc113804571)

[Objetivo 4](#_Toc113804572)

[Diccionario de datos 5](#_Toc113804573)

[0. Preparando el entorno 7](#_Toc113804574)

[Importar las bibliotecas y datos 7](#_Toc113804575)

[Bibliotecas 7](#_Toc113804576)

[Datos 7](#_Toc113804577)

[1. Descripción de la estructura de los datos 8](#_Toc113804578)

[Forma (dimensiones) del DataFrame 8](#_Toc113804579)

[Tipos de datos 8](#_Toc113804580)

[2. Identificación de datos faltantes 9](#_Toc113804581)

[3. Detección de valores atípicos 10](#_Toc113804582)

[Distribución de variables numéricas 10](#_Toc113804583)

[Resumen estadístico de variables numéricas 12](#_Toc113804584)

[Diagramas para detectar posibles valores atípicos 12](#_Toc113804585)

[Distribución de variables categóricas 14](#_Toc113804586)

[Plot 14](#_Toc113804587)

[Agrupación por variables categóricas 15](#_Toc113804588)

[4. Identificación de relaciones entre variables 16](#_Toc113804589)

[Conclusiones 17](#_Toc113804590)

**Análisis exploratorio de datos (EDA)**

# Introducción

## EDA – Análisis Exploratorio de Datos

* Una buena práctica, antes de mirar los datos, es hacer un análisis de éstos para resumir sus principales características, a menudo con métodos visuales.
* El análisis exploratorio de datos, o EDA, implica conocer los datos.

Diagrama

Descripción generada automáticamente

Es útil también revisar la descripción de los datos para comprender lo que significa cada característica.

Entre las actividades a realizar destacan:

* **Paso 1:** Descripción de la estructura de los datos.
* **Paso 2:** Identificación de datos faltantes.
* **Paso 3:** Detección de valores atípicos.
* **Paso 4:** Identificación de relaciones entre pares variables.

## Proceso básico de EDA

Diagrama

Descripción generada automáticamente

## Variables

Es un atributo (característica) que concentra valores que pueden variar de una o más maneras.

Diagrama

Descripción generada automáticamente

# Contexto

* El sector inmobiliario de Melbourne, Australia continúa en auge desde hace algunos años.
* Es de interés 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.

## Objetivo

Encontrar información de interés para predecir la próxima tendencia inmobiliaria en Melbourne.

Interfaz de usuario gráfica, Sitio web

Descripción generada automáticamente

Fuente: <https://www.kaggle.com/dansbecker/melbourne-housing-snapshot>

## Diccionario de datos

Tabla

Descripción generada automáticamente

|  |  |  |
| --- | --- | --- |
| **Item** | **Column name** | **Definición** |
| **1** | Rooms | Número de habitaciones |
| **2** | Price | Precio en dólares |
| **3** | Method | * **S** –propiedad vendida * **SP** – propiedad vendida antes * **PI** – propiedad transferida * **PN** – vendida antes no revelada * **SN** – vendida no revelada * **NB** – sin oferta * **VB** – oferta del proveedor * **W** – retirada antes de la subasta * **SA** – vendida después de subasta * **SS** – vendida después del precio de subasta no revelado. * **N/A** – precio u oferta más alta no disponible. |
| **4** | Type | * **br** –dormitorio (s) * **h** – casa, cabaña, villa, semi, terraza * **u** –unidad, dúplex * **t** –casa adosada * **dev site** – en desarrollo * **res** – otro residencial. |
| **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 ...) |
| **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) |

# 0. Preparando el entorno

## Importar las bibliotecas y datos

**Objetivo:**

* Revisar la descripción y significado de cada columna.
* Una buena práctica es observar los datos para obtener una imagen clara de éstos.
* Si se quiere ver solo las primeras filas se usa head(). Por ejemplo: DatosMelbourne.head()

### Bibliotecas

import pandas as pd               # Para la manipulación y análisis de datos

import numpy as np                # Para crear vectores y matrices n dimensionales

import matplotlib.pyplot as plt   # Para la generación de gráficas a partir de los datos

import seaborn as sns             # Para la visualización de datos basado en matplotlib

%matplotlib inline

# Para generar y almacenar los gráficos dentro del cuaderno

### Datos

El conjunto de datos corresponde a **Melbourne Housing Snapshot de Kaggle**. Este conjunto de datos incluye: dirección, tipo de inmueble, suburbio, método de venta, habitaciones, precio, agente inmobiliario, fecha de venta y Distancia desde C.B.D. (Distrito Central de Negocios).

**Fuente:** <https://www.kaggle.com/dansbecker/melbourne-housing-snapshot>

Captura de pantalla de computadora

Descripción generada automáticamente

# 1. Descripción de la estructura de los datos

## Forma (dimensiones) del DataFrame

El atributo **shape** de Pandas proporciona una estructura general de los datos. Devuelve la cantidad de filas y columnas que tiene el conjunto de datos.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

## Tipos de datos

|  |  |
| --- | --- |
| Tabla, Escala de tiempo  Descripción generada automáticamente con confianza media | El atributo **dtypes** muestra los tipos de datos de las variables.  Se observa que el conjunto de datos tiene una combinación de variables categóricas (objeto) y numéricas (flotante e int). |

**Observaciones:**

* Se observa que el conjunto de datos tiene una **combinación de variables categóricas** (objeto) y numéricas (flotante e int).
* Si se quiere predecir el precio de una vivienda, se podría **implementar un Árbol de decisión** (ya que se tienen distintos tipos de datos)

# 2. Identificación de datos faltantes

|  |  |
| --- | --- |
| Tabla  Descripción generada automáticamente | Una función útil de Pandas es **isnull().sum()** que regresa la suma de todos los valores nulos en cada variable.  **Observaciones:**   * Se podrían eliminar los datos faltantes si es que es una variable significativa. * Otra solución sería completar los datos faltantes, los cuales se pueden completar con distintas técnicas. |
|  | También se puede usar **info()** para obtener el tipo de datos y la suma de valores nulos. |

# 3. Detección de valores atípicos

* Se pueden utilizar gráficos para tener una idea general de las distribuciones de los datos, y se sacan estadísticas para resumir los datos. Estas dos estrategias son recomendables y se complementan.
* La distribución se refiere a cómo se distribuyen los valores en una variable 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.

## Distribución de variables numéricas

* Se utilizan histogramas que agrupan los números en rangos.
* La altura de una barra muestra cuántos números caen en ese rango.
* Se emplea **hist()** para trazar el histograma de las variables numéricas. También se pueden usar los parámetros: **figsize** y **xrot** para aumentar el tamaño de la cuadrícula y rotar el eje x 45 grados.

Texto

Descripción generada automáticamente

Imagen que contiene juego, biombo, edificio, tren

Descripción generada automáticamente

Qué buscar:

* Posibles valores atípicos, que pueden ser errores de medición.
* Límites que no tienen sentido, como valores porcentuales > 100.

Gráfico

Descripción generada automáticamente

* En el histograma, se observa que **BuildingArea** y **Landsize** tienen valores sesgados a la izquierda.
* La variable **Price** también está sesgada hacia la izquierda.
* **YearBuilt** está sesgado hacia la derecha y el límite comienza en 1200, lo cual es extraño.

**Cómo solucionarlo:**

* Eliminar los valores atípicos, pero antes se debe verificar que los datos son verdaderos.
* Modificar los datos atípicos, si es que hubo un error.

En resumen, se debe preguntar al proveedor de los datos y validar si es que los datos proporcionados son verdaderos.

## Resumen estadístico de variables numéricas

Se sacan estadísticas usando **describe()** que muestra un resumen estadístico de las variables numéricas.

Captura de pantalla de computadora

Descripción generada automáticamente

* Se incluye un recuento, media, desviación, valor mínimo, valor máximo, percentil inferior (25%), 50% y percentil superior (75%).
* Por defecto, el percentil 50 es lo mismo que la mediana.
* Se observa que para cada variable, el recuento también ayuda a identificar variables con valores nulos o perdidos. Estos son: Car, Landsize y YearBuilt.

## Diagramas para detectar posibles valores atípicos

Para este tipo de gráficas se utiliza Seaborn, que permite generar diagramas de cajas para detectar valores atípicos.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Gráfico, Gráfico de burbujas

Descripción generada automáticamente

Gráfico de burbujas

Descripción generada automáticamente con confianza baja

Imagen que contiene Gráfico

Descripción generada automáticamente

**Observaciones:**

Se observa que Landsize, BuildingArea y YearBuilt tienen valores atípicos que están lejos de los otros valores.

## Distribución de variables categóricas

* Se refiere a la observación de las clases de cada columna (variable) y su frecuencia.
* Aquí, las gráficas ayudan para tener una idea general de las distribuciones, mientras que las estadísticas dan números reales.

Pantalla de un video juego

Descripción generada automáticamente con confianza media

* En esta tabla se muestra el recuento de los valores de cada variable, el número de clases únicas, la clase más frecuente y con qué frecuencia ocurre esa clase en el conjunto de datos.
* Se observa que algunas clases tienen demasiados valores únicos, como Address, seguida de Suburb y SellerG.
* A partir de estos hallazgos, se puede graficar las variables con 10 o menos clases únicas.

### Plot

* Para este tipo de gráficas se utiliza Seaborn, que permite generar un histograma para variables categóricas. Cada barra en la gráfica representa una clase.
* Se crea un **For** para el conteo y distribución de las clases.
* La sentencia **select\_dtypes(include = ’object’)** selecciona las columnas categóricas con sus valores y las muestra.
* Se incluye **If** para elegir solo las columnas con 10 o menos clases usando **nunique() < 10**.

Texto

Descripción generada automáticamente

Gráfico, Gráfico de embudo

Descripción generada automáticamente Gráfico, Gráfico de embudo

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente

## Agrupación por variables categóricas

Pantalla de computadora con letras

Descripción generada automáticamente con confianza media

Texto

Descripción generada automáticamente

# 4. Identificación de relaciones entre variables

* Una matriz de correlaciones es útil para analizar la relación entre las variables numéricas.
* Se emplea la función **corr()**

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

* Se puede trazar un mapa de calor a través de la biblioteca de Seaborn.

Texto

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

* Una correlación es un valor entre -1 y 1 que equivale a qué tan cerca se mueven simultáneamente los valores de dos variables.
* **Una correlación positiva significa que a medida que una característica aumenta, la otra también aumenta.**
* **Una correlación negativa significa que a medida que una característica disminuye, la otra también disminuye.**
* Las correlaciones cercanas a 0 indican una relación débil, mientras que las más cercanas a -1 o 1 significan una relación fuerte.

# Conclusiones

Durante esta práctica, se analizaron los datos obtenidos del sector inmobiliario de Melbourne, Australia, con el objetivo de encontrar información de interés para predecir la próxima tendencia inmobiliaria de esta localidad.

Para poder cumplir con este objetivo, se realizó el análisis exploratorio de datos o EDA. Este análisis es parte de las buenas prácticas que se deben tener, ya que uno de sus objetivos es hacer un análisis de los datos para resumir sus principales características, a menudo y de preferencia con métodos visuales. En resumen, **el análisis exploratorio de datos implica conocer los datos** con los que se están trabajando.

De manera general, este análisis exploratorio de datos consta de 4 pasos, que de preferencia, se deben seguir en orden:

* **Paso 1:** Descripción de la estructura de los datos (para comprender lo que significa cada característica)
* **Paso 2:** Identificación de datos faltantes.
* **Paso 3:** Detección de valores atípicos.
* **Paso 4:** Identificación de relaciones entre pares variables.

Con esta práctica pude familiarizarme con este análisis exploratorio de datos, las distintas librerías y funciones con las que Python viene integrado, y cómo se pueden utilizar estas herramientas para cumplir con los 4 pasos del EDA.