



Universidad Nacional Autónoma de México
Facultad de Ingeniería

Análisis exploratorio de datos (EDA)

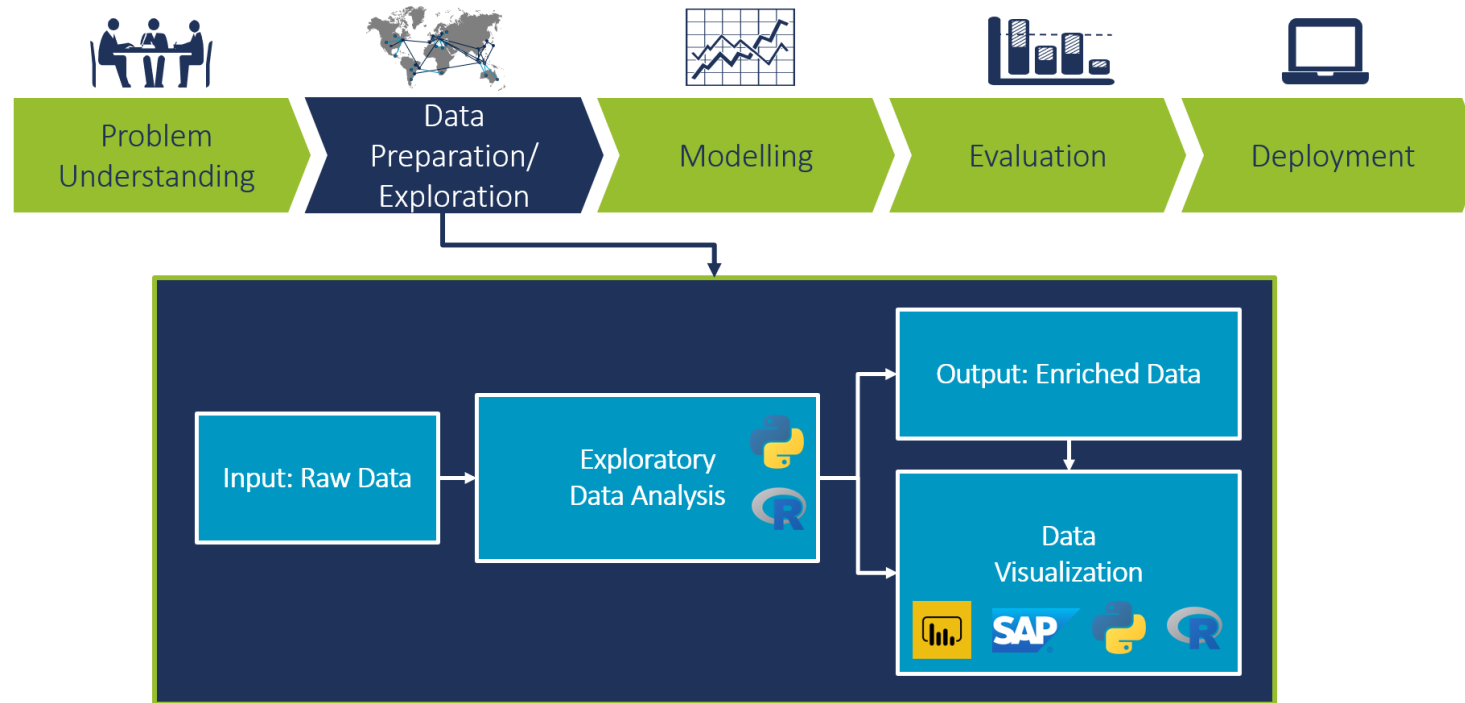
Guillermo Molero-Castillo

guillermo.molero@ingenieria.unam.edu

Septiembre, 2022

Análisis exploratorio de datos

EDA

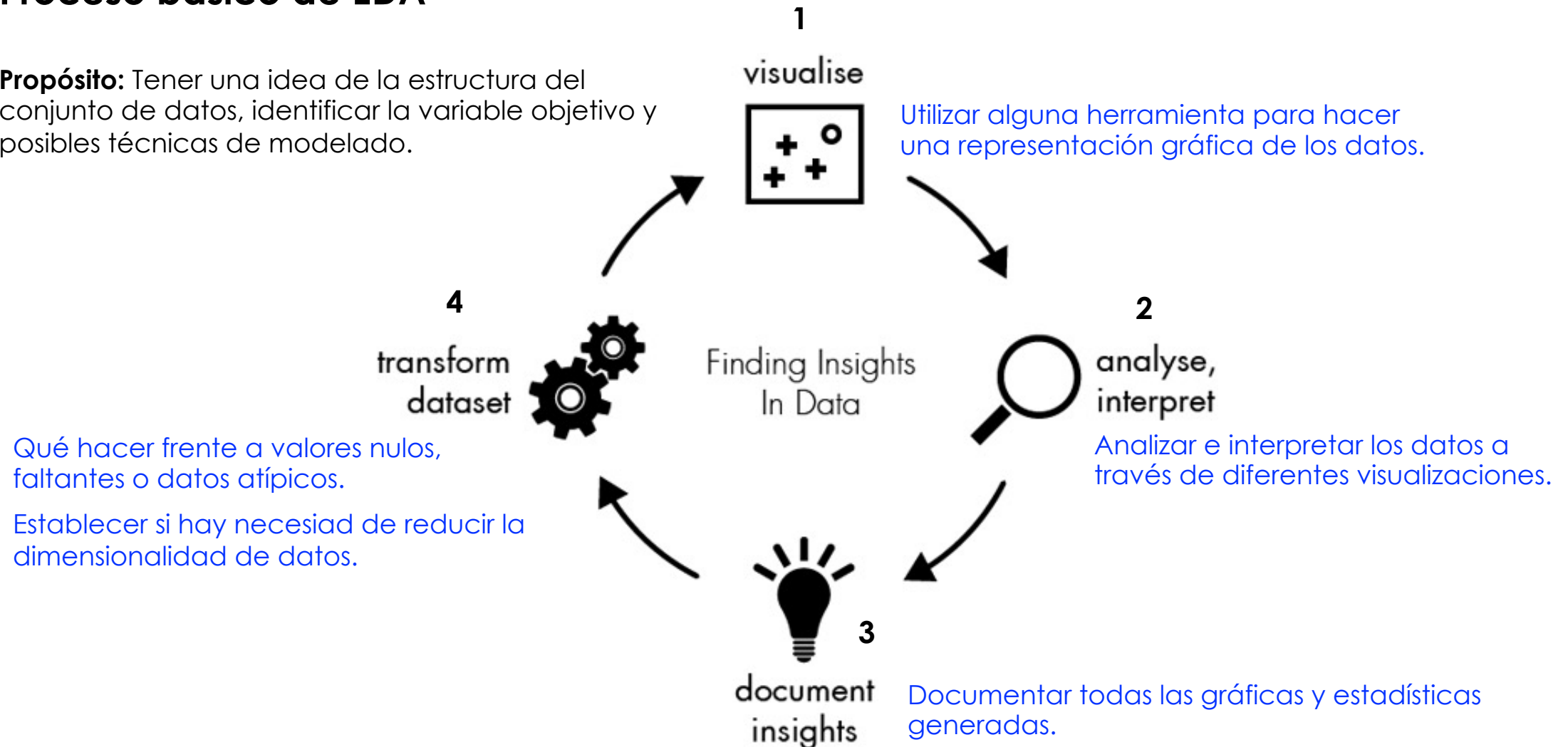


- Una buena práctica, antes de mirar los datos, es hacer un análisis de éstos para resumir sus principales características, a menudo con métodos visuales.
- El análisis exploratorio de datos, o EDA, implica conocer los datos.

Análisis exploratorio de datos

Proceso básico de EDA

Propósito: Tener una idea de la estructura del conjunto de datos, identificar la variable objetivo y posibles técnicas de modelado.



Análisis exploratorio de datos

Variables

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



Análisis exploratorio de datos

EDA

- 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.

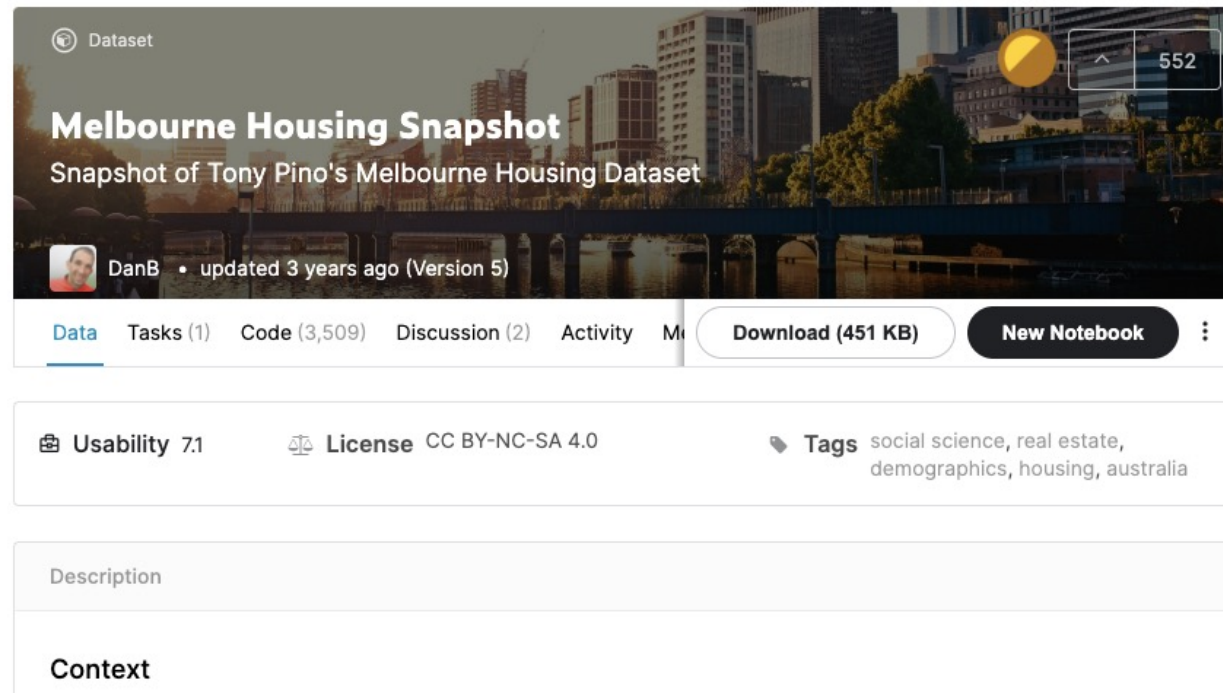
Práctica 1

Análisis exploratorio de datos

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.



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

Análisis exploratorio de datos

Diccionario de datos

	Column name	Definition
0	Rooms:	Number of rooms
1	Price:	Price in dollars
2	Method:	S - property sold; SP - property sold prior; PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; W - withdrawn prior to auction; SA - sold after auction; SS - sold after auction price not disclosed. N/A - price or highest bid not available.
3	Type:	br - bedroom(s); h - house,cottage,villa, semi,terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential.
4	SellerG:	Real Estate Agent
5	Date:	Date sold
6	Distance:	Distance from CBD
7	Regionname:	General Region (West, North West, North, North east ...etc)
8	Propertycount:	Number of properties that exist in the suburb.
9	Bedroom2 :	Scraped # of Bedrooms (from different source)
10	Bathroom:	Number of Bathrooms
11	Car:	Number of carspots
12	Landsize:	Land Size
13	BuildingArea:	Building Size
14	CouncilArea:	Governing council for the area

Análisis exploratorio de datos

Diccionario de datos

Item	Column name	Definición
1	Rooms	Número de habitaciones
2	Price	Precio en dolares
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; o 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)

Análisis exploratorio de datos

Importar las bibliotecas y datos

```
▶ 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

▶ from google.colab import files
files.upload()
```

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>

Análisis exploratorio de datos

Importar las bibliotecas y datos

```
[3] DatosMelbourne = pd.read_csv('melb_data.csv')
```

```
[4] DatosMelbourne
```

	Suburb	Address	Rooms	Type	Price	Method	SellerG	Date	Distance	Postcode	Bedroom2
0	Abbotsford	85 Turner St	2	h	1480000.0	S	Biggin	3/12/2016	2.5	3067.0	2.0
1	Abbotsford	25 Bloomburg St	2	h	1035000.0	S	Biggin	4/02/2016	2.5	3067.0	2.0
2	Abbotsford	5 Charles St	3	h	1465000.0	SP	Biggin	4/03/2017	2.5	3067.0	3.0

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()`

Análisis exploratorio de datos

Paso 1. Descripción de la estructura de los datos

1) 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.

```
DatosMelbourne.shape  
(13580, 21)
```

Análisis exploratorio de datos

Paso 1. Descripción de la estructura de los datos

DatosMelbourne.dtypes	
Suburb	object
Address	object
Rooms	int64
Type	object
Price	float64
Method	object
SellerG	object
Date	object
Distance	float64
Postcode	float64
Bedroom2	float64
Bathroom	float64
Car	float64
Landsize	float64
BuildingArea	float64
YearBuilt	float64
CouncilArea	object
Lattitude	float64
Longtitude	float64
Regionname	object
Propertycount	float64
dtype:	object

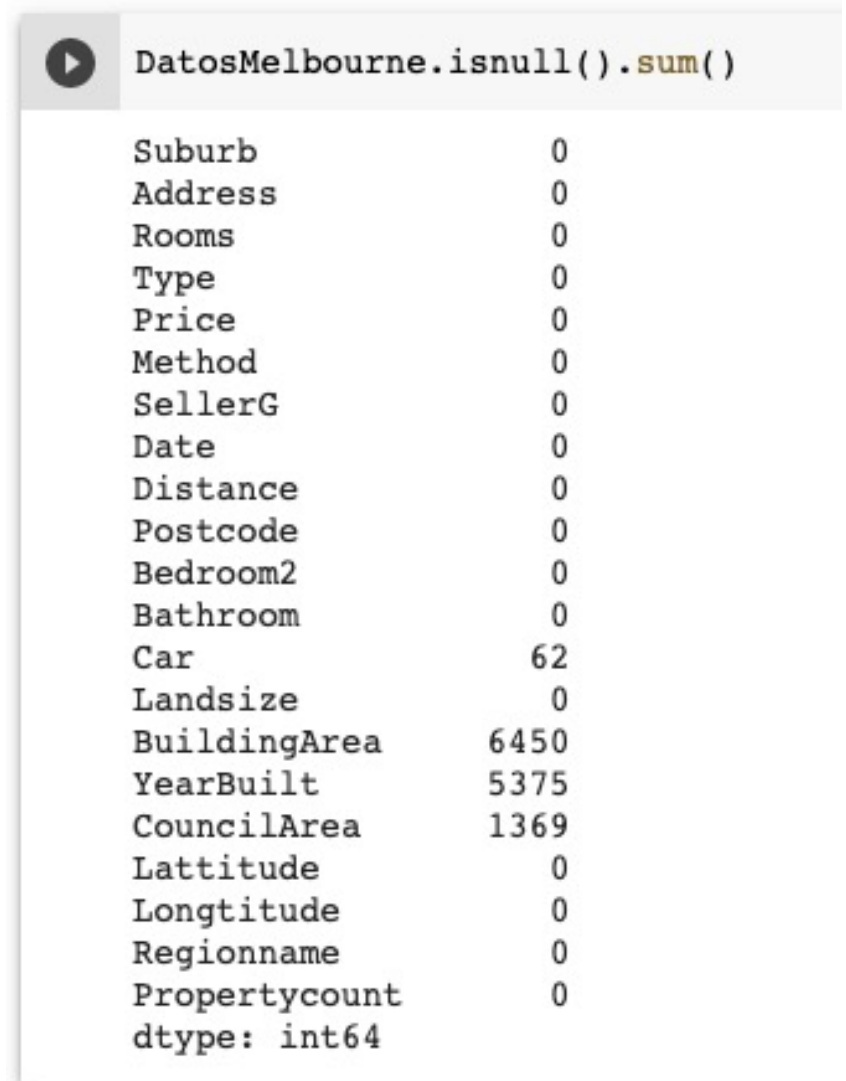
2) Tipos de datos

- 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).

Análisis exploratorio de datos

Paso 2. Identificación de datos faltantes

Una función útil de Pandas es `isnull().sum()` que regresa la suma de todos los valores nulos en cada variable.



```
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
Lattitude	0
Longtitude	0
Regionname	0
Propertycount	0
dtype:	int64

Análisis exploratorio de datos

Paso 2. Identificación de datos faltantes

También se puede usar `info()` para obtener el tipo de datos y la suma de valores nulos.

```
DatosMelbourne.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13580 entries, 0 to 13579
Data columns (total 21 columns):
#   Column              Non-Null Count  Dtype  
---  -
0   Suburb              13580 non-null  object 
1   Address             13580 non-null  object 
2   Rooms               13580 non-null  int64  
3   Type                13580 non-null  object 
4   Price               13580 non-null  float64 
5   Method              13580 non-null  object 
6   SellerG             13580 non-null  object 
7   Date                13580 non-null  object 
8   Distance             13580 non-null  float64 
9   Postcode            13580 non-null  float64 
10  Bedroom2            13580 non-null  float64 
11  Bathroom            13580 non-null  float64 
12  Car                  13518 non-null  float64 
13  Landsize            13580 non-null  float64 
14  BuildingArea        7130 non-null   float64 
15  YearBuilt            8205 non-null   float64 
16  CouncilArea         12211 non-null  object 
17  Lattitude            13580 non-null  float64 
18  Longtitude          13580 non-null  float64 
19  Regionname          13580 non-null  object 
20  Propertycount       13580 non-null  float64 
dtypes: float64(12), int64(1), object(8)
memory usage: 2.2+ MB
```

Análisis exploratorio de datos

Paso 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.

Análisis exploratorio de datos

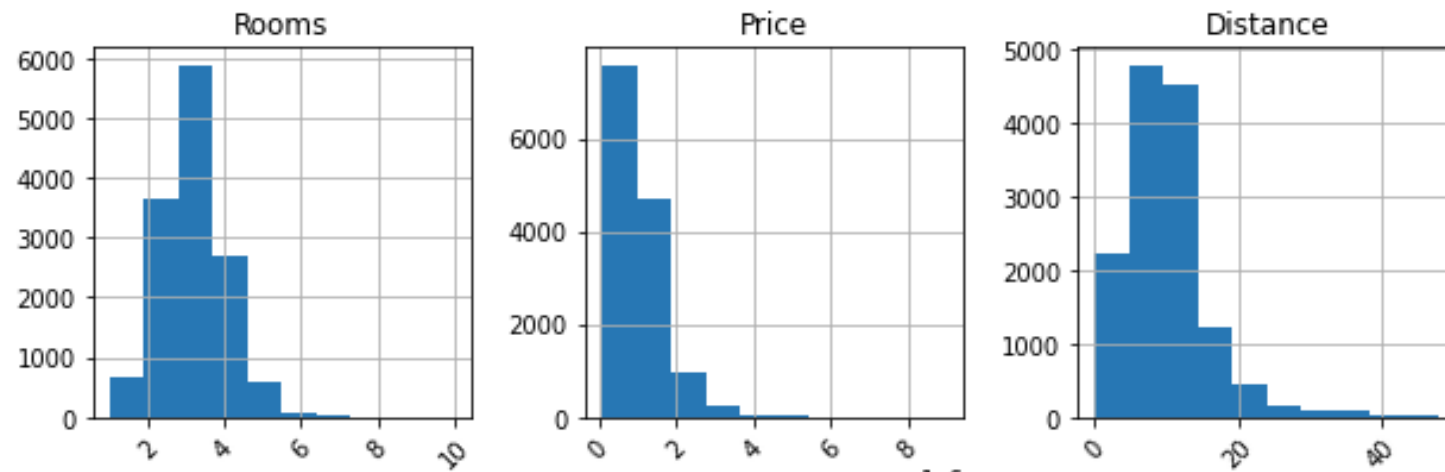
Paso 3. Detección de valores atípicos

1) 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.



```
DatosMelbourne.hist(figsize=(14,14), xrot=45)  
plt.show()
```



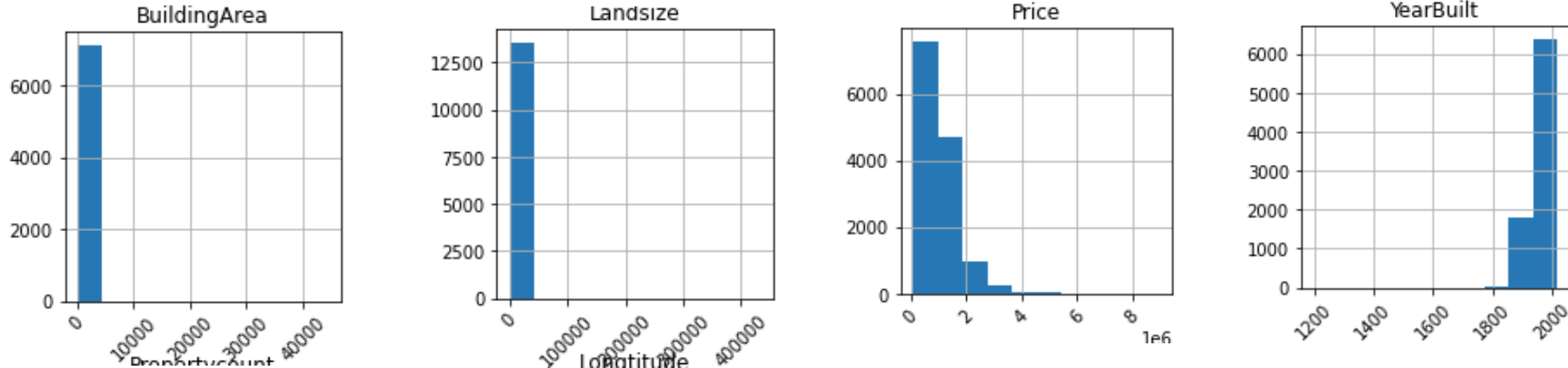
Análisis exploratorio de datos

Paso 3. Detección de valores atípicos

1) Distribución de variables numéricas

Qué buscar:

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



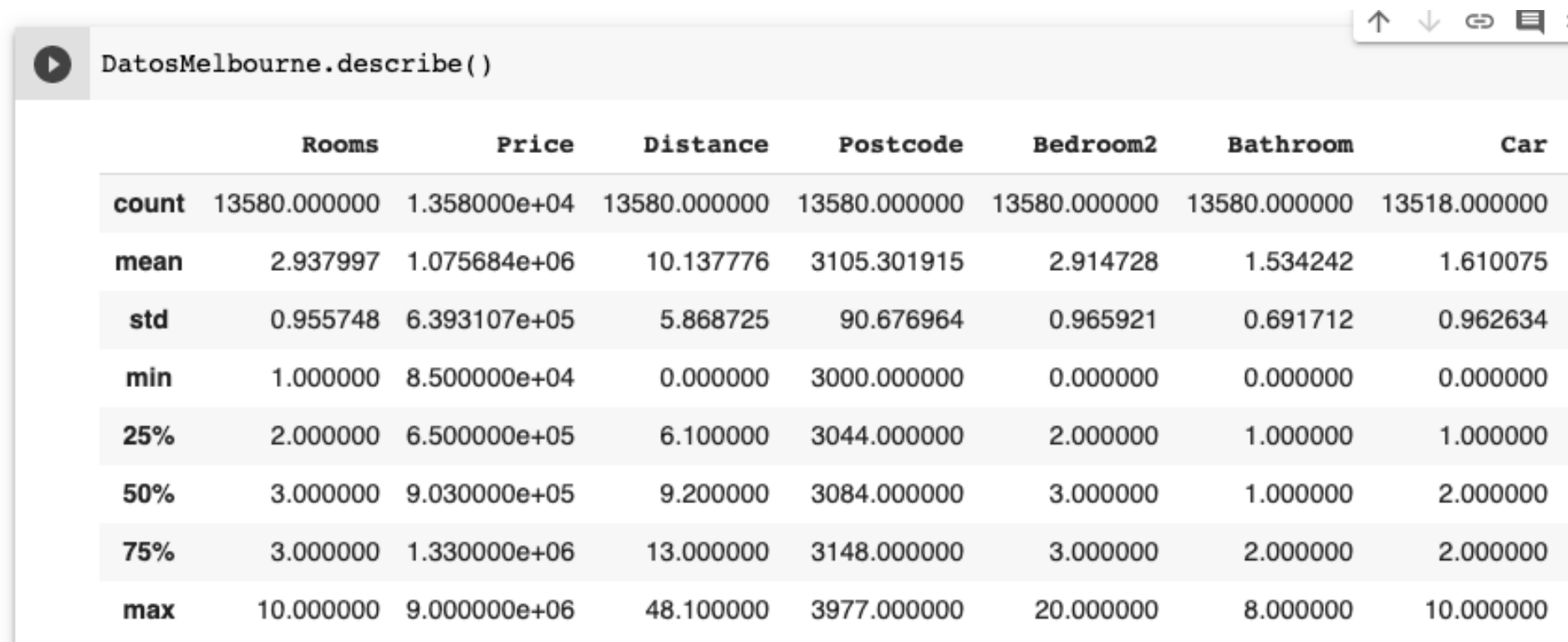
En el histograma, se observa que **BuildingArea** y **Landsize** tienen valores sesgados a la izquierda. La variable **Price** también está sesgada hacia la izquierda. **YearBuilt** está sesgado hacia la derecha y el límite comienza en 1200, lo cual es extraño.

Análisis exploratorio de datos

Paso 3. Detección de valores atípicos

2) Resumen estadístico de variables numéricas

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



The image shows a Jupyter Notebook interface with a code cell containing `DatosMelbourne.describe()` and its output. The output is a summary statistics table for the variables: Rooms, Price, Distance, Postcode, Bedroom2, Bathroom, and Car. The table includes rows for count, mean, std, min, 25%, 50%, 75%, and max.

	Rooms	Price	Distance	Postcode	Bedroom2	Bathroom	Car
count	13580.000000	1.358000e+04	13580.000000	13580.000000	13580.000000	13580.000000	13518.000000
mean	2.937997	1.075684e+06	10.137776	3105.301915	2.914728	1.534242	1.610075
std	0.955748	6.393107e+05	5.868725	90.676964	0.965921	0.691712	0.962634
min	1.000000	8.500000e+04	0.000000	3000.000000	0.000000	0.000000	0.000000
25%	2.000000	6.500000e+05	6.100000	3044.000000	2.000000	1.000000	1.000000
50%	3.000000	9.030000e+05	9.200000	3084.000000	3.000000	1.000000	2.000000
75%	3.000000	1.330000e+06	13.000000	3148.000000	3.000000	2.000000	2.000000
max	10.000000	9.000000e+06	48.100000	3977.000000	20.000000	8.000000	10.000000

Análisis exploratorio de datos

Paso 3. Detección de valores atípicos

2) Resumen estadístico de variables numéricas

	Rooms	Price	Distance
count	13580.000000	1.358000e+04	13580.000000
mean	2.937997	1.075684e+06	10.137776
std	0.955748	6.393107e+05	5.868725
min	1.000000	8.500000e+04	0.000000
25%	2.000000	6.500000e+05	6.100000
50%	3.000000	9.030000e+05	9.200000
75%	3.000000	1.330000e+06	13.000000
max	10.000000	9.000000e+06	48.100000

- 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.

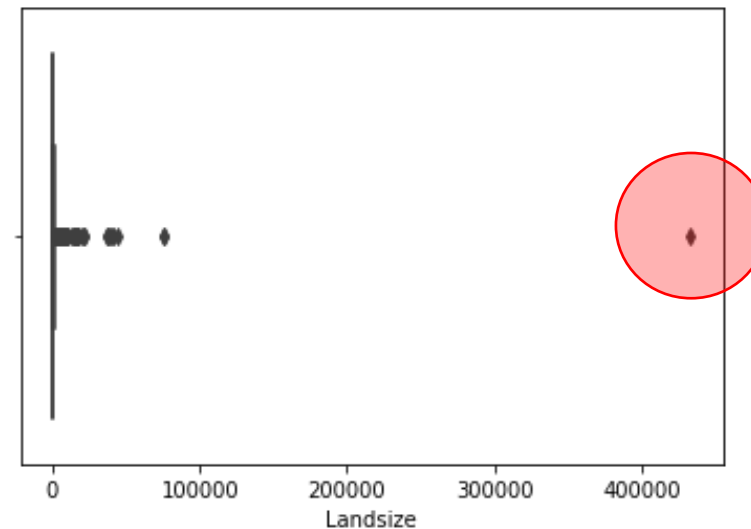
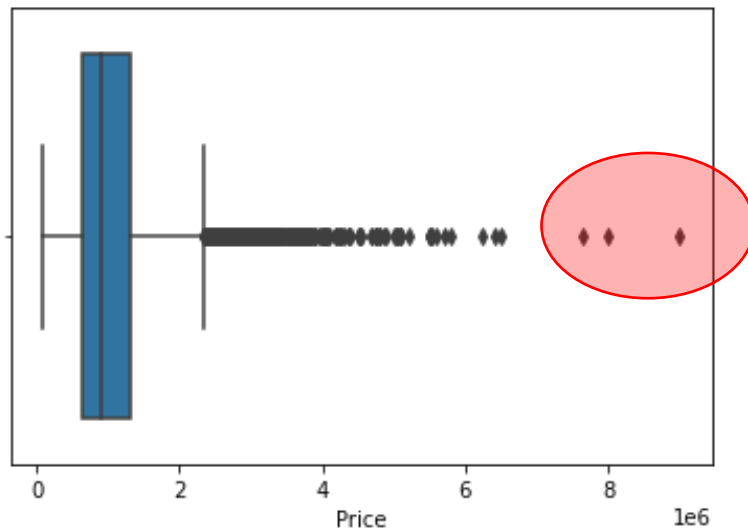
Análisis exploratorio de datos

Paso 3. Detección de valores atípicos

3) Diagramas para detectar posibles valores atípicos

Para este tipo de gráficas se utiliza Seaborn, que permite generar diagramas de cajas.

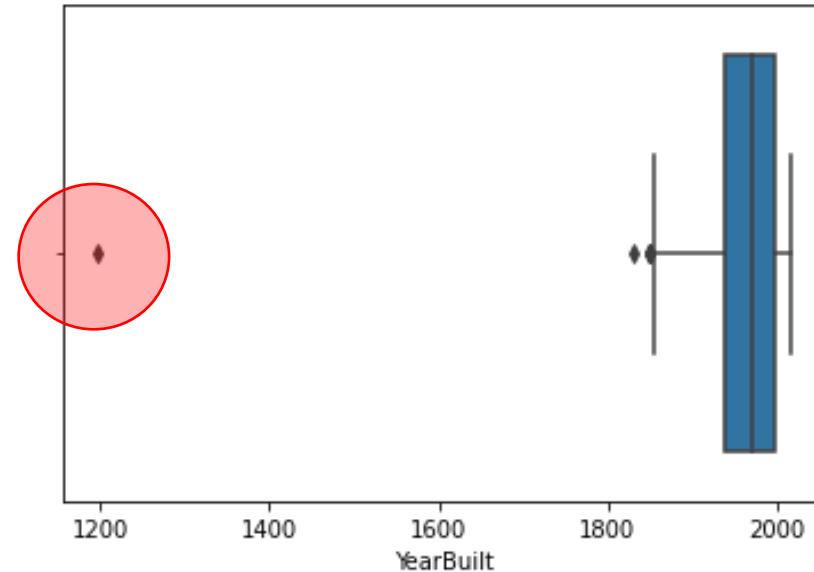
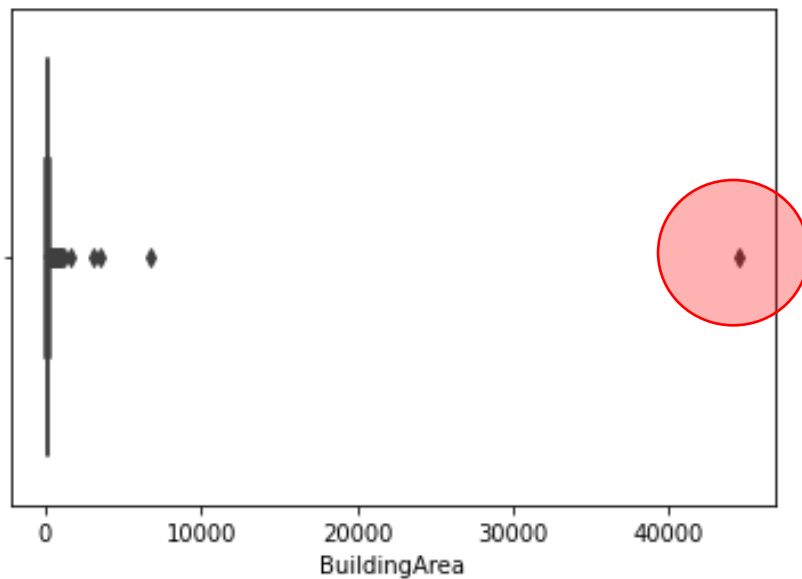
```
variables = ['Price', 'Landsize', 'BuildingArea', 'YearBuilt']  
for col in variables:  
    sns.boxplot(col, data=DatosMelbourne)  
    plt.show()
```



Análisis exploratorio de datos

Paso 3. Detección de valores atípicos

3) Diagramas para detectar posibles valores atípicos



Análisis exploratorio de datos

Paso 3. Detección de valores atípicos

4) 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.



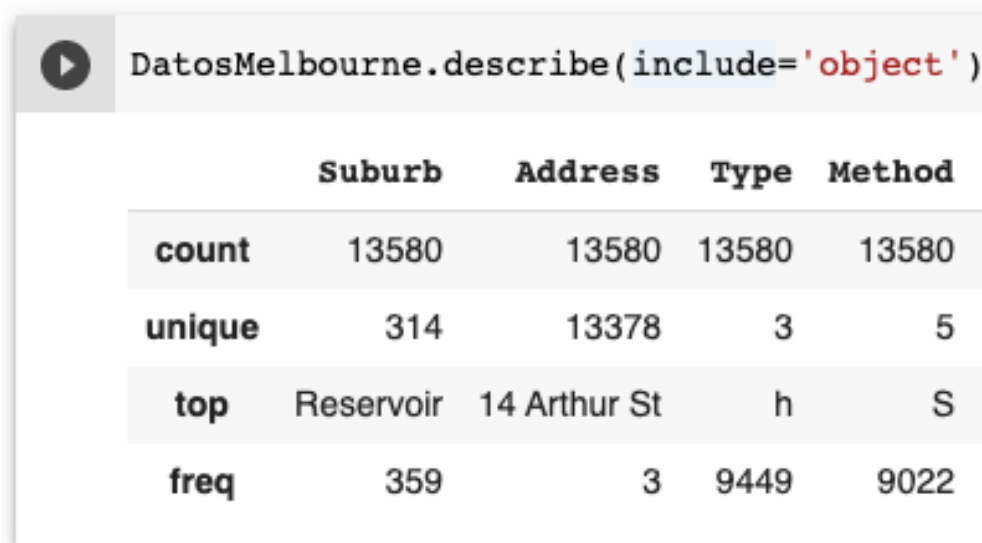
```
DatosMelbourne.describe(include='object')
```

	Suburb	Address	Type	Method	SellerG	Date	CouncilArea
count	13580	13580	13580	13580	13580	13580	12211
unique	314	13378	3	5	268	58	33
top	Reservoir	14 Arthur St	h	S	Nelson	27/05/2017	Moreland
freq	359	3	9449	9022	1565	473	1163

Análisis exploratorio de datos

Paso 3. Detección de valores atípicos

4) Distribución de variables categóricas



```
DatosMelbourne.describe(include='object')
```

	Suburb	Address	Type	Method
count	13580	13580	13580	13580
unique	314	13378	3	5
top	Reservoir	14 Arthur St	h	S
freq	359	3	9449	9022

- 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.

Paso 3. Detección de valores atípicos

4) Distribución de variables categóricas

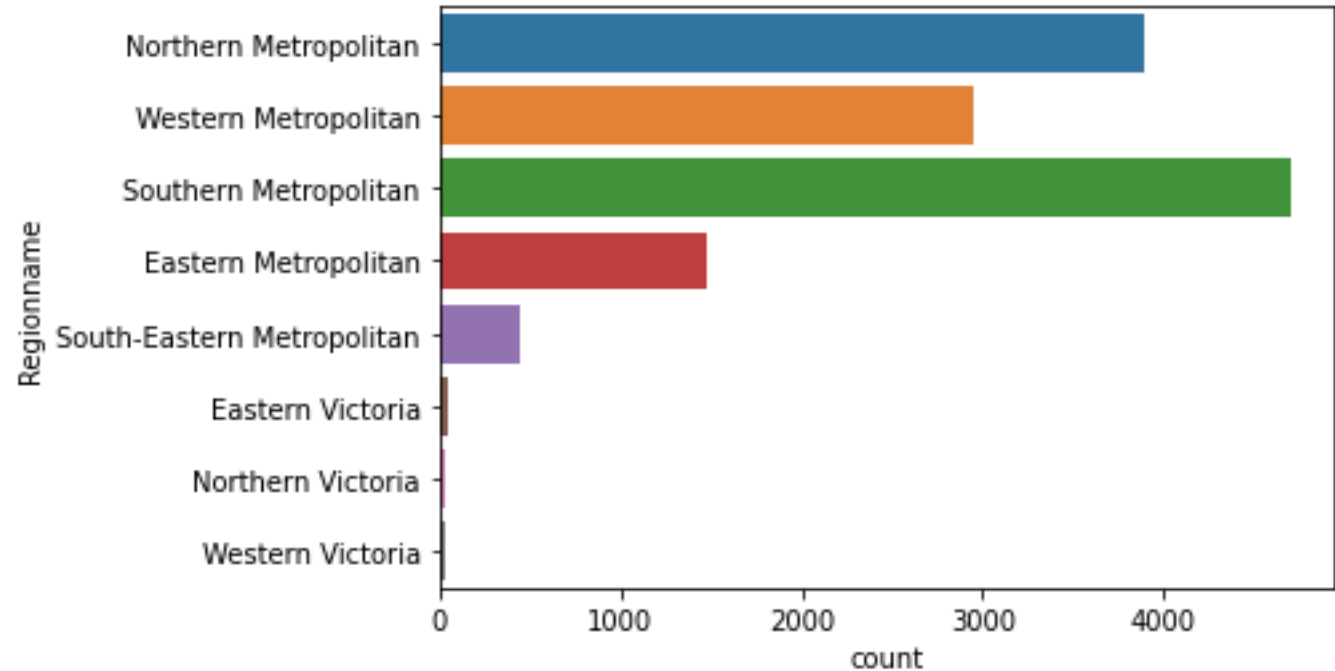
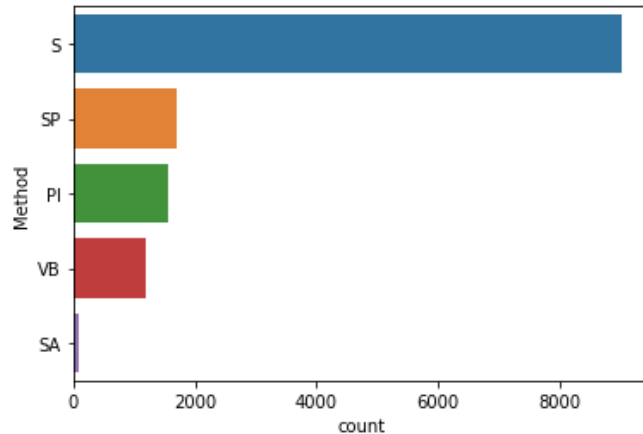
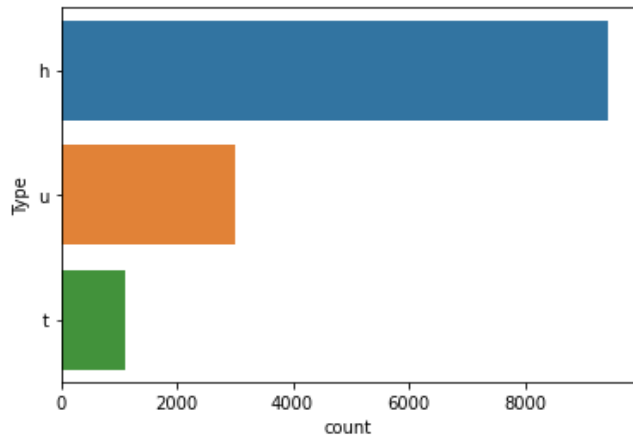
- Para este tipo de gráficas se utiliza Seaborn, que permite generar un histograma para variables categóricas. Cada barra en la gráfica representa una clase.
- Se crea un **For** para el conteo y distribución de las clases.
- La sentencia `select_dtypes(include = 'object')` selecciona las columnas categóricas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

```
for col in DatosMelbourne.select_dtypes(include='object'):  
    if DatosMelbourne[col].nunique() < 10: sns.countplot(y=col, data=DatosMelbourne)  
    plt.show()
```

Análisis exploratorio de datos

Paso 3. Detección de valores atípicos

4) Distribución de variables categóricas



Paso 3. Detección de valores atípicos

5) Agrupación por variables categóricas

```
▶ for col in DatosMelbourne.select_dtypes(include='object'):  
    if DatosMelbourne[col].nunique() < 10:  
        display(DatosMelbourne.groupby(col).agg(['mean']))
```

	Rooms	Price	Distance	Postcode	Bedroom2	Bathroom
	mean	mean	mean	mean	mean	mean
Type						
h	3.260874	1.242665e+06	10.979479	3104.080643	3.229336	1.613822
t	2.837522	9.337351e+05	9.851346	3100.777379	2.814183	1.809695
u	1.963871	6.051275e+05	7.607391	3110.797481	1.966523	1.183295

Análisis exploratorio de datos

Paso 3. Detección de valores atípicos

5) Agrupación por variables categóricas

	Rooms	Price	Distance	Postcode	Bedroom2	Bathroom							
	mean	mean	mean	mean	mean	mean							
Method													
PI	3.077366	1.133242e+06	9.482097	3106.742327	3.062660	1.714194							
S	2.941809	1.087327e+06	10.431523	3106.171359	2.914875		Rooms	Price	Distance	Postcode	Bedroom2	Bathroom	
SA	3.010870	1.025772e+06	12.385870	3132.304348	3.010870		mean	mean	mean	mean	mean	mean	
SP	2.795655	8.998924e+05	10.374692	3096.480916	2.785672		Regionname						
VB	2.924103	1.166510e+06	8.273728	3107.337781	2.896580		Eastern Metropolitan	3.322230	1.104080e+06	13.901088	3111.162475	3.313392	1.698844
							Eastern Victoria	3.396226	6.999808e+05	34.209434	3567.584906	3.396226	1.811321
							Northern Metropolitan	2.755527	8.981711e+05	8.078329	3071.360925	2.734190	1.367866
							Northern Victoria	3.560976	5.948293e+05	33.748780	3418.707317	3.560976	1.853659

Análisis exploratorio de datos

Paso 4. Identificación de relaciones entre variables

- 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.

Análisis exploratorio de datos

Paso 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()`



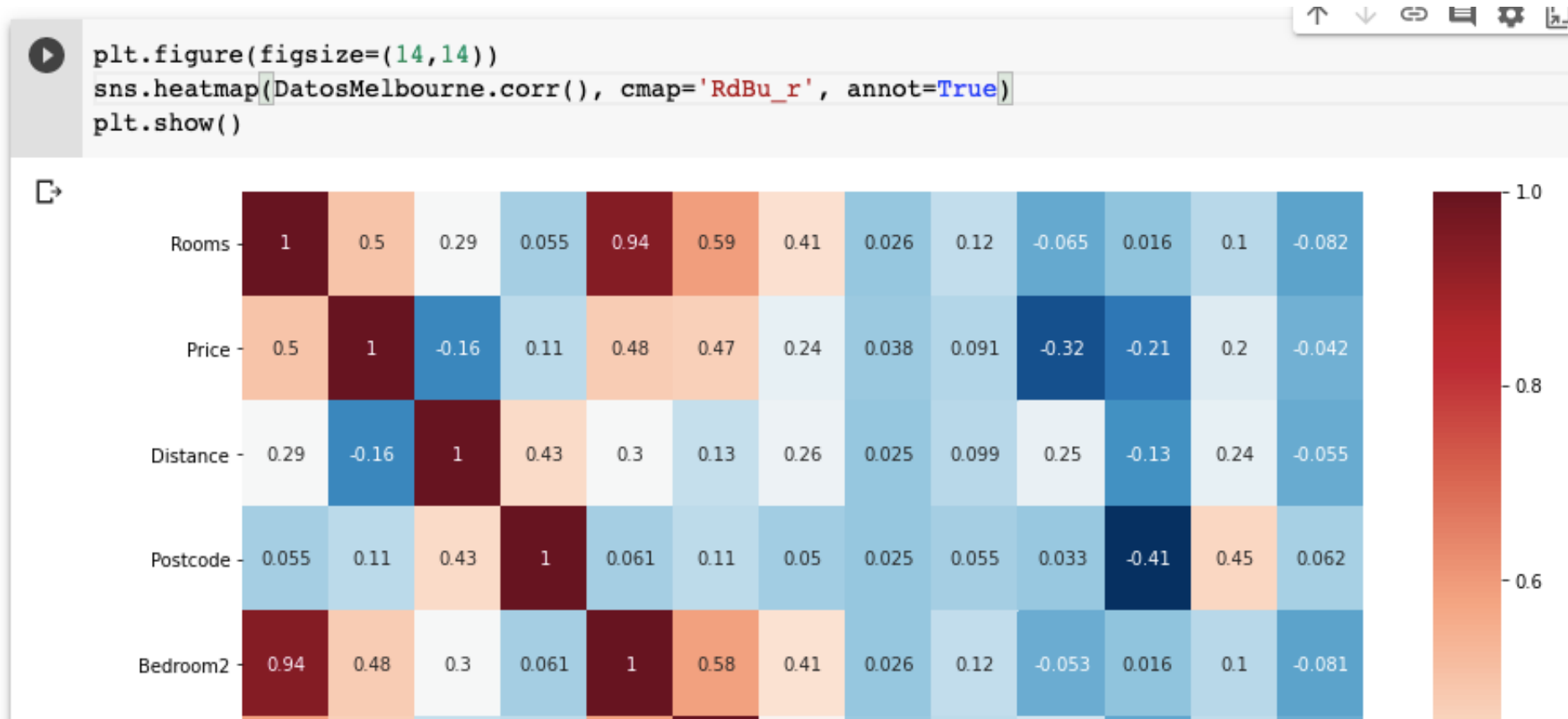
DatosMelbourne.corr()

	Rooms	Price	Distance	Postcode	Bedroom2	Bathroom	Car
Rooms	1.000000	0.496634	0.294203	0.055303	0.944190	0.592934	0.408483
Price	0.496634	1.000000	-0.162522	0.107867	0.475951	0.467038	0.238979
Distance	0.294203	-0.162522	1.000000	0.431514	0.295927	0.127155	0.262994
Postcode	0.055303	0.107867	0.431514	1.000000	0.060584	0.113664	0.050289
Bedroom2	0.944190	0.475951	0.295927	0.060584	1.000000	0.584685	0.405325
Bathroom	0.592934	0.467038	0.127155	0.113664	0.584685	1.000000	0.322246
Car	0.408483	0.238979	0.262994	0.050289	0.405325	0.322246	1.000000

Análisis exploratorio de datos

Paso 4. Identificación de relaciones entre pares variables

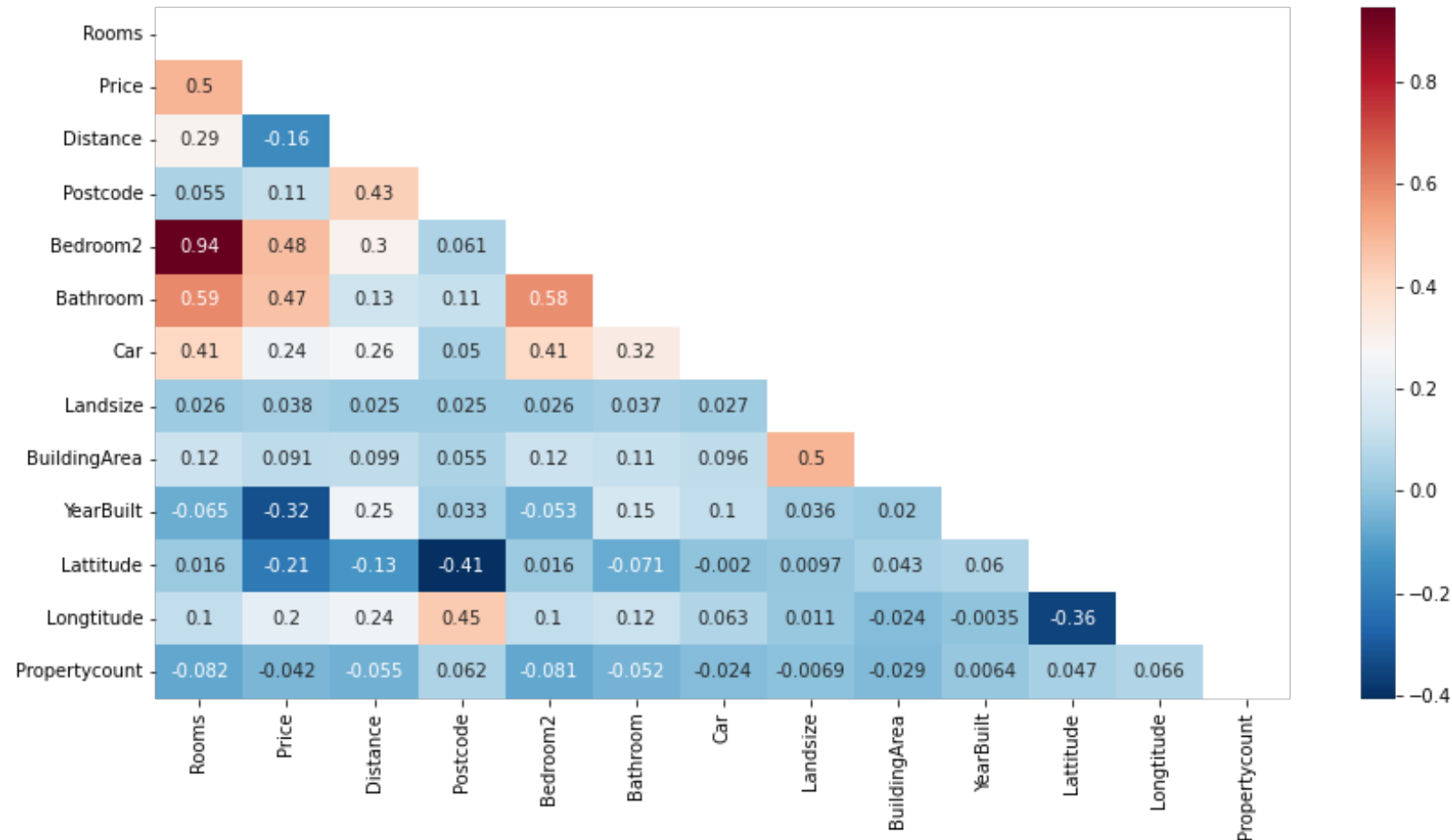
- Se puede trazar un mapa de calor a través de la biblioteca de Seaborn.



Análisis exploratorio de datos

Paso 4. Identificación de relaciones entre pares variables

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



Tarea 2

Objetivo. Identificar una fuente de datos (datos abiertos) para su importación en Google Colab o Jupyter a través de alguna plataforma Git (Github, Gitlab o algún otro).

Mostrar el procedimiento y la ejecución de la importación de datos (archivos CSV) desde un repositorio Git dentro de un cuaderno en Colab o Jupyter. Además, realizar el proceso de EDA con los datos identificados.

Fecha de entrega: Martes 20 de septiembre de 2022

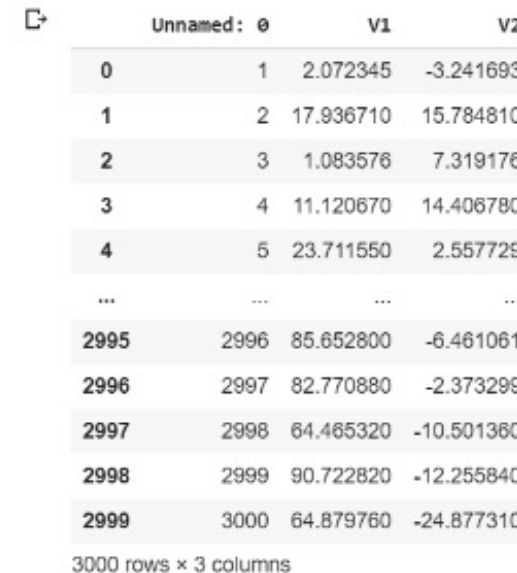
Hora: Antes de las 16:00 horas

Formato: Libre, subir a la carpeta compartida el reporte de la tarea en un archivo 'pdf'.

Se debe pasar el parámetro a `read_csv()` en pandas para obtener la matriz de datos.

```
url = 'copied_raw_github_link'  
df = pd.read_csv(url)
```

Salida:



	Unnamed: 0	V1	V2
0	1	2.072345	-3.241693
1	2	17.936710	15.784810
2	3	1.083576	7.319176
3	4	11.120670	14.406780
4	5	23.711550	2.557729
...
2995	2996	85.652800	-6.461061
2996	2997	82.770880	-2.373299
2997	2998	64.465320	-10.501360
2998	2999	90.722820	-12.255840
2999	3000	64.879760	-24.877310

3000 rows x 3 columns