DataRitz Technologies
Enhancing Technology Experience

Topics to be covered

- Data Importing
- Data Understanding
- Data Manipulation

Data Importing

In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
house_data = pd.read_csv('House_Data.csv')
```

Data Understanding

In [3]:
```
house_data.head(5)
```

Out[3]:

	area_type	availability	location	size	society	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	NaN	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	NaN	1200	2.0	1.0	51.00

In [4]:
```
house_data.shape
```

Out[4]:
```
(13320, 9)
```

In [5]:
```
house_data.columns
```

Out[5]:
```
Index(['area_type', 'availability', 'location', 'size', 'society',
       'total_sqft', 'bath', 'balcony', 'price'],
      dtype='object')
```

In [6]:
```
house_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13320 entries, 0 to 13319
Data columns (total 9 columns):
area_type       13320 non-null object
availability    13320 non-null object
location        13319 non-null object
size            13304 non-null object
society         7821 non-null object
total_sqft      13320 non-null object
bath            13247 non-null float64
balcony         12711 non-null float64
price           13320 non-null float64
dtypes: float64(3), object(6)
memory usage: 936.6+ KB
```

In [7]:
```
#Return an int representing the number of elements in this object
house_data.size
```

Out[7]:
```
119880
```

In [8]:
```
# Check Datatype of each attributes
house_data.dtypes
```

Out[8]:
```
area_type        object
availability     object
location         object
size             object
society          object
total_sqft       object
bath            float64
balcony         float64
price           float64
dtype: object
```

In [9]:
```
house_data.get_dtype_counts()
```

Out[9]:
```
float64    3
object     6
dtype: int64
```

In [10]:
```
# Count no of non null values per column
house_data.count()
```

Out[10]:
```
area_type       13320
availability    13320
location        13319
size            13304
society          7821
total_sqft      13320
bath            13247
balcony         12711
price           13320
dtype: int64
```

In [11]:
```
#Getting specific list of data types
# .select_dtypes(include=None, exclude=None)
#Return a subset of a DataFrame including/excluding columns based
# on their ``dtype``. ('category','number' , 'floating' , 'object')
float_data = house_data.select_dtypes(include=['floating'])
```

In [12]:
```
float_data.shape
```

Out[12]:
```
(13320, 3)
```

In [13]:
```
# using Exclude
object_data = house_data.select_dtypes(exclude=['number'])
object_data.head(5)
```

Out[13]:

	area_type	availability	location	size	society	total_sqft
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	NaN	1440
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	NaN	1200

In [ ]:
```
```

In [14]:
```
pip install pandas-profiling
```

```
The following command must be run outside of the IPython shell:

    $ pip install pandas-profiling

The Python package manager (pip) can only be used from outside of IPython.
Please reissue the `pip` command in a separate terminal or command prompt.

See the Python documentation for more information on how to install packages:

    https://docs.python.org/3/installing/ (https://docs.python.org/3/installing/)
```

In [15]:
```
import pandas_profiling
```

In [19]:
```
profile = pandas_profiling.ProfileReport(house_data)
```

In [21]:
```
profile.to_file('house.html')
```