autoeda-benguluru-housing-pandasprofiling

April 26, 2022

0.1 Objective: To explore various AutoEDA capabilities and perform analysis on a given dataset

0.1.1 This notebook will focus on pandas profiling

0.2 1. AutoEDA - Pandas Profiling

0.2.1 Dataset Reference: Bengaluru Housing dataset

0.2.2 Features:

- General Overview - Quick insights of all variables in the dataset
- Details about each variables / features in the dataset
- Interactions between numeric variables
- Correlations between variables - Pearson's Correlation Coefficient, Spearman's Rank Correlation Coefficient, Kendall's Rank Correlation Coefficient, Phik Correlation Coefficient, Cramer's V for displaying association measure for nominal random variables
- Missing Values - Count, Matrix, Heatmap, Dendogram representations
- Sample data - first and last 10 rows

0.2.3 When To Use?

- Dataset size is not very large
- Need some quick insights about an unknown dataset
- Use this as a basis for your further EDA analysis on top of it

[1]:

```python
# Import libraries
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import pandas_profiling as pp
```

[2]:

```python
# Read data
df = pd.read_csv("C:/Users/Karthik.Iyer/Downloads/AccelerateAI/DV_EDA/
 ↪DV_and_EDA-main/data/Bengaluru_House_Data.csv")

df.head()
```

[2]:

```
              area_type   availability                  location       size  \
0  Super built-up  Area         19-Dec  Electronic City Phase II      2 BHK   
1            Plot  Area  Ready To Move          Chikka Tirupathi  4 Bedroom   
2        Built-up  Area  Ready To Move               Uttarahalli      3 BHK   
3  Super built-up  Area  Ready To Move        Lingadheeranahalli      3 BHK   
```

```
4  Super built-up  Area  Ready To Move                  Kothanur      2 BHK

   society total_sqft  bath  balcony   price
0  Coomee        1056   2.0      1.0   39.07
1  Theanmp       2600   5.0      3.0  120.00
2      NaN       1440   2.0      3.0   62.00
3  Soiewre       1521   3.0      1.0   95.00
4      NaN       1200   2.0      1.0   51.00
```

```
[3]: # Check shape
     df.shape
```

```
[3]: (13320, 9)
```

```
[4]: pp.ProfileReport(df)
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]

Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]

<IPython.core.display.HTML object>
```

```
[4]:
```

0.2.4 Interpretation:

- The profiling report gives quick insights of all variables in the dataset.
- The details about each variables / features in the dataset are also captured.
- Quick insights around the following:
  - Interactions between numeric variables
  - Correlations between variables - Pearson's Correlation Coefficient, Spearman's Rank Correlation Coefficient, Kendall's Rank Correlation Coefficient, Phik Correlation Coefficient, Cramer's V for displaying association measure for nominal random variables
  - Missing Values - Count, Matrix, Heatmap, Dendogram representations
  - Sample data - first and last 10 rows