Pandas Profiling in Python

- To Know More About - https://pandas-profiling.github.io/pandas-profiling/docs/master/rtd/index.html (https://pandas-profiling.github.io/pandas-profiling/docs/master/rtd/index.html)
- When importing a new data set for the very first time, the first thing to do is to get an understanding of the data. This includes steps like determining the range of specific predictors, identifying each predictor's data type, as well as computing the number or percentage of missing values for each predictor.
- The pandas_profiling library in Python include a method named as ProfileReport() which generate a basic report on the input DataFrame.

The report consist of the following:

1. Overview
2. Variables
3. Interactions
4. Co-relations
5. Missing Values
6. Sample

1. Overview Section :

- This section provides overall data set information. Dataset statistics and Variable types.
- Dataset statistics display columns, rows, missing values, etc.
- Variable Types shows data types of the attributes of the data set.
- It also shows “Warnings”, where it shows which feature(s) are highly correlated to others.

2. Variables :

- This section provides information about every feature individually in detail.
- When we click on the Toggle details option as shown in the above image, the new section shows up.

3. Interactions :

- This section shows statistics, histograms, common values, and extreme values of features.
- here we intract various feature each other.

4. Co-relations :

- This Section shows how features are co-related with each other with the help of Seaborn's Heatmap.
- We can easily toggle between the different types of correlations like Pearson, Spearman, Kendall, and phik.

5. Missing Values :

- here we can see whether our dataset contain missing value or not.
- here it use 4 different function ex- Count,matrix,heatmap,dendrogram

6. Sample :

- This section displays the First 10 Rows and the Last 10 rows of the dataset.

Installation of Pandas Profiling:

Installation with the pip package

```
!pip install pandas-profiling
```

Installation with the conda package

```
conda install -c conda-forge pandas-profiling
```

Implement Using Python

In [2]:

```
import pandas as pd
df=pd.read_csv("https://raw.githubusercontent.com/agconti/kaggle-titanic/master/data/train.
df.head()
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500

In [7]:

```
import pandas_profiling as pp
# run the profile report
profile = df.profile_report(title='Pandas Profiling Report')
profile
```

```
Summarize dataset: 100%                    26/26 [00:11<00:00, 2.20it/s, Completed]

Generate report structure: 100%            1/1 [00:06<00:00, 6.68s/it]

Render HTML: 100%                          1/1 [02:32<00:00, 152.30s/it]
```

In [8]:

```
# save the report as html file
### It generate a sepaarte html file
profile.to_file(output_file="pandas_profiling1.html")
```

```
Export report to file: 100%                1/1 [01:32<00:00, 92.03s/it]
```

In [ ]:

```
# save the report as json file
profile.to_file(output_file="pandas_profiling2.json")
```