Pandas Profiling

Now Days ,Data Scientists and Analysts usually spend Most of time to get insight the data they are going to work on by doing exploratory analysis.

It's one of the first steps in their journey before making further analysis and predictions. They use methods in pandas library like:

`head()`, `describe()`, `info()`, `columns()`, `shape()`, `isnull()`, `value_counts()`, `unique()`, `duplicated()`, `corr()`

Also using libraries such as seaborn and matplotlib to get visualization.

- **Is there any way to save time?**
  What if with just two lines of code we will be able to get insights the data?
  What if there is a report with visualization?

pandas-profiling can provide us a report with exploratory insights using only two line of codes that will save a lot of time.

Overview on pandas-profiling library

pandas-profiling is an open-source Python library that allows us to quickly do exploratory analysis with just a few lines of code.
we will use this library to generate an interactive report we also can save this report to HTML file that can help us to share it with others.

Now, we will learn how to deal with pandas-profiling

Installing pandas-profiling

- **Open anaconda prompt.**
- **Write pip install pandas-profiling then press enter.**

```
C:\Users\Mansi>pip install pandas-profiling
```

- **Open your Jupyter Notebook.**
- **load your data set.**
- **here I will use house prices advanced regression data set you can download it from here.**

```
import pandas as pd
```

```
train=pd.read_csv('train_house.csv')
```

```
train.head()
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley
0	1	60	RL	65.0	8450	Pave	NaN
1	2	20	RL	80.0	9600	Pave	NaN
2	3	60	RL	68.0	11250	Pave	NaN
3	4	70	RL	60.0	9550	Pave	NaN
4	5	60	RL	84.0	14260	Pave	NaN

5 rows × 81 columns

- **Import profileReport from pandas-profiling library.**

```
from pandas_profiling import ProfileReport
```

- **let`s create the report.**

```
output_Report = ProfileReport(df)
output_Report.to_file(output_file='Report.html')
```

- **it will take a few minutes to create Reprt.html**
- **you can download the report from here**
- **you can download notebook from here**

The report is composed of a lot of information

- **Overview: we can see some general statistics of the data, information on the report and warnings, that show insights that can highly impact the analysis, such as a high number of null values in a variable, duplicated rows, and high correlation between variables.**
- **Variables: composed of descriptive and quantile statistics information for each variable. Also, it's possible to see the histogram and the common and extreme values of the variable, in the case of continuous variables, and pie chart and frequency of each value for categorical data.**
- **Interactions: allows us to see the relationship between two variables through the scatter plot visualization.**

- **Correlations: shows the heatmap of correlation matrix.**
- **Missing values: through a bar chart or matrix visualization it's possible to see the missing values for each variable.**
- **Sample: first 10 rows and last 10 rows are printed.**
- **Duplicate rows: shows the duplicated rows.**

Pandas-profiling limitation

One limitation I could see of pandas-profiling is when it's applied to large datasets because, as the dataset size increases, the report generation time increases a lot.

Conclusion

we can get an exploratory data analysis report using the pandas-profiling library. With a two lines of code, we can generate an interactive report and create an HTML file for it