Satyam Kumar Follow

Oct 3, 2020 · 5 min read · ✨ · Listen

Save

4 Libraries that can perform EDA in one line of python code

Exploratory data analysis using Pandas Profiling, Sweetviz, Autoviz, and D-Tale

Image by Mudassar Iqbal from Pixabay, Edited using Pixlr

Exploratory data analysis (EDA) is an approach to analyze the data and find patterns, visual insights, etc. that the data set is having, before proceeding to model. One spends a lot of time doing EDA to get a better understanding of data, that can be minimized by using auto visualizations tools such as **Pandas-profiling, Sweetviz, Autoviz,** or **D-Tale**

EDA involves a lot of steps including some statistical tests, visualization of data using different kinds of plots, and many more. Some of the steps of EDA are discussed below:

- **Data Quality Check:** Can be done using pandas library functions like `describe()`, `info()`, `dtypes()`, etc. It is used to find several features, its datatypes, duplicate values, missing value, etc.
- **Statistical Test:** Some statistical tests like Pearson correlation, Spearman correlation, Kendall test, etc are done to get a correlation between the features. It can be implemented in python using the *stats* library.
- **Quantitative Test:** Some quantitative test is used to find the spread of numerical features, count of categorical features. It can be implemented in python using the functions of the pandas library.
- **Visualization:** Feature visualization is very essential to get an understanding of the data. Graphical techniques like bar plots, pie charts are used to get an understanding of categorical features, whereas scatter plots, histograms are used for numerical features.

To perform the above-mentioned tasks we need to type several lines of code. Here **auto-visualization** library comes into the play, which can perform all these tasks using just 1 line of code. Some of these auto-visualization tools we will discuss in this article:

- **Pandas-Profiling**
- **Sweetviz**
- **Autoviz**
- **D-Tale**

The dataset used for exploratory data analysis using the pandas-profiling library is the Titanic dataset downloaded from Kaggle.

Pandas-Profiling:

Pandas profiling is an open-source python library that automates the EDA process and creates a detailed report. Pandas Profiling can be used easily for large datasets as it is blazingly fast and creates reports in a few seconds.

Installation:

You can install pandas-profiling using PyPl:

```
pip install pandas-profiling
```

> *GitHub repository for pandas profiling.*

Usage:

```
#Install the below libaries before importing
import pandas as pd
from pandas_profiling import ProfileReport

#EDA using pandas-profiling
profile = ProfileReport(pd.read_csv('titanic.csv'), explorative=True)

#Saving results to a HTML file
profile.to_file("output.html")
```

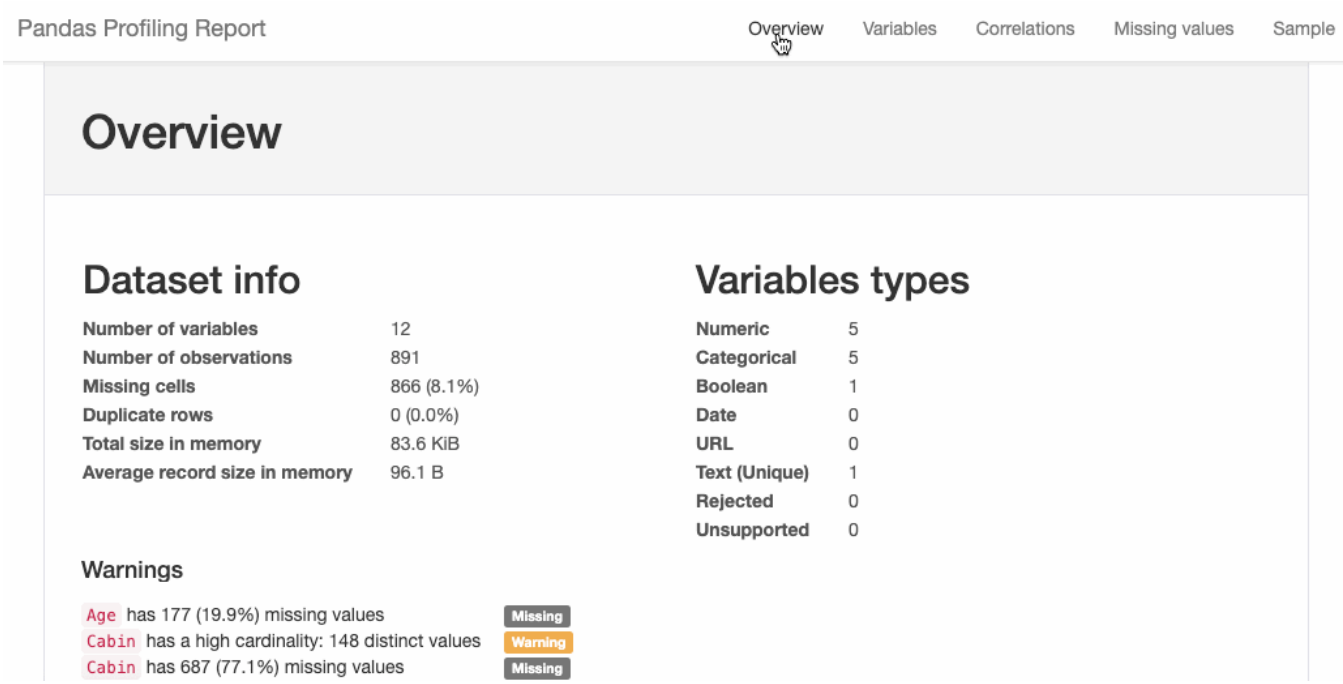
pandas-profiling.py hosted with ❤ by GitHub view raw

Report:

The **pandas-profiling** library generates a report having:

- An overview of the dataset
- Variable properties
- Interaction of variables

- Correlation of variables
- Missing values
- Sample data



(GIF by Author)

Sweetviz:

Sweetviz is an open-source python auto-visualization library that generates a report, exploring the data with the help of high-density plots. It not only automates the EDA but is also used for comparing datasets and drawing inferences from it. A comparison of two datasets can be done by treating one as training and the other as testing.

Installation:

You can install Sweetviz using PyPl:

GitHub repository for Sweetviz.

Usage:

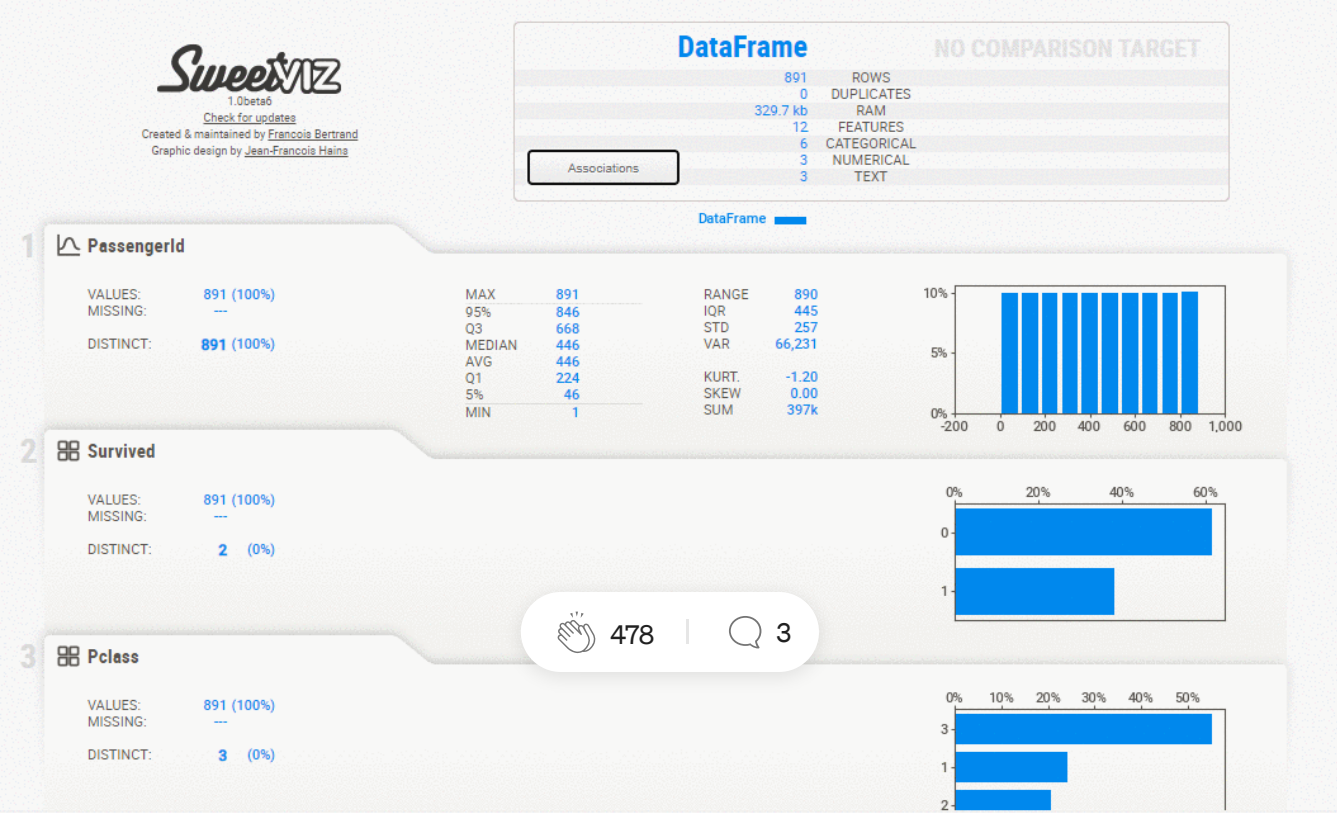
```
import pandas as pd
import sweetviz as sv

#EDA using Autoviz
sweet_report = sv.analyze(pd.read_csv("titanic.csv"))

#Saving results to HTML file
sweet_report.show_html('sweet_report.html')
```

sweetviz.py hosted with ❤ by GitHub view raw

Report:

The **Sweetviz** library generates a report having:

- An overview of the dataset
- Variable properties
- Categorical associations
- Numerical associations
- Most frequent, smallest, largest values for numerical features



(GIF by Author)

Autoviz:

Autoviz is an open-source python auto visualization library that mainly focuses on visualizing the relationship of the data by generating different types of plot.

Installation:

You can install Autoviz using PyPl:

```
pip install autoviz
```

> *GitHub repository for Autoviz.*

Usage:

Report:

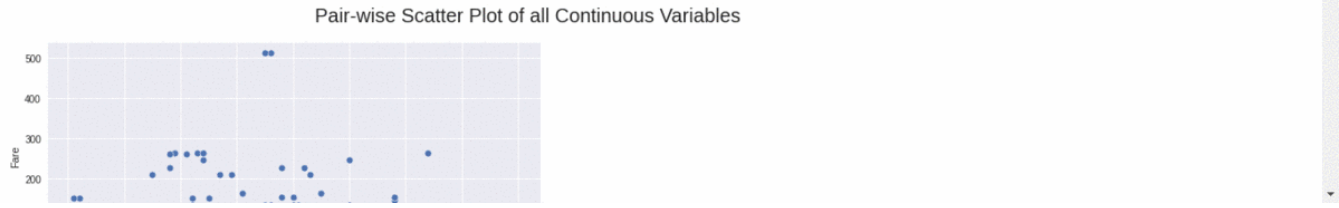
The **Autoviz** library generates a report having:

- An overview of the dataset
- Pairwise scatter plot of continuous variables
- Distribution of categorical variables
- Heatmaps of continuous variables
- Average numerical variable by each categorical variable

```
Shape of your Data Set: (891, 12)
Classifying variables in data set...
    12 Predictors classified...
        This does not include the Target column(s)
    4 variables removed since they were ID or low-information variables
Number of All Scatter Plots = 3
Could not draw Distribution Plot
Could not draw Violin Plot
Time to run AutoViz (in seconds) = 2.068
```

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22.0	1	0	7.2500	S
1	1	1	female	38.0	1	0	71.2833	C
2	1	3	female	26.0	0	0	7.9250	S
3	1	1	female	35.0	1	0	53.1000	S
4	0	3	male	35.0	0	0	8.0500	S
...	...	...	...	...	...	...	...	...
886	0	2	male	27.0	0	0	13.0000	S
887	1	1	female	19.0	0	0	30.0000	S
888	0	3	female	NaN	1	2	23.4500	S
889	1	1	male	26.0	0	0	30.0000	C
890	0	3	male	32.0	0	0	7.7500	Q

891 rows × 8 columns



(GIF by Author)

D-Tale:

D-Tale is an open-source python auto-visualization library. It is one of the best auto data-visualization libraries. D-Tale helps you to get a detailed EDA of the data. It also has a feature of **code export**, for every plot or analysis in the report.

Installation:

You can install D-Tale using PyPl:

```
pip install dtale
```

> *GitHub repository for D-Tale.*

Usage:

Report:

The **dtale** library generates a report having:

- An overview of the dataset
- Custom filters
- Correlation, Charts, and Heatmaps
- Highlight datatypes, missing values, ranges
- Code export

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
		0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.25	nan	S
		1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.28	C85	C
		1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.93	nan	S
		1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.10	C123	S
		0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.05	nan	S
		0	3	Moran, Mr. James	male	nan	0	0	330877	8.46	nan	Q
		0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	51.86	E46	S
		0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.08	nan	S
		1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	11.13	nan	S
		1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	30.07	nan	C
		1	3	Sandstrom, Miss. Marguerite Rut	female	4.00	1	1	PP 9549	16.70	G6	S
		1	1	Bonnell, Miss. Elizabeth	female	58.00	0	0	113783	26.55	C103	S
		0	3	Saundercock, Mr. William Henry	male	20.00	0	0	A/5. 2151	8.05	nan	S
		0	3	Andersson, Mr. Anders Johan	male	39.00	1	5	347082	31.28	nan	S
		0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.00	0	0	350406	7.85	nan	S
		1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.00	0	0	248706	16.00	nan	S
		0	3	Rice, Master. Eugene	male	2.00	4	1	382652	29.13	nan	Q
		1	2	Williams, Mr. Charles Eugene	male	nan	0	0	244373	13.00	nan	S
		0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31.00	1	0	345763	18.00	nan	S
		1	3	Masselmani, Mrs. Fatima	female	nan	0	0	2649	7.23	nan	C
		0	2	Fynney, Mr. Joseph J	male	35.00	0	0	239865	26.00	nan	S
		1	2	Beesley, Mr. Lawrence	male	34.00	0	0	248698	13.00	D56	S
		1	3	McGowan, Miss. Anna "Annie"	female	15.00	0	0	330923	8.03	nan	Q
23	24	1	1	Sloper, Mr. William Thompson	male	28.00	0	0	113788	35.50	A6	S
24	25	0	3	Palsson, Miss. Torborg Danira	female	8.00	3	1	349909	21.08	nan	S
25	26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)	female	38.00	1	5	347077	31.39	nan	S
26	27	0	3	Emir, Mr. Farred Chehab	male	nan	0	0	2631	7.23	nan	C
27	28	0	1	Fortune, Mr. Charles Alexander	male	19.00	3	2	19950	263.00	C23 C25 C27	S
28	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	nan	0	0	330959	7.88	nan	Q

(GIF by Author)

Conclusion:

I prefer to do my EDA with self-defined functions using several python libraries. The above-discussed libraries should be used to speed up your work.

For beginners, it is good to start doing EDA using the pandas' library and writing python code before trying these libraries, as it is more important to be equipped with the fundamental knowledge and programming practices.

The best data auto-visualization amongst the above discussed is the **DTale** library, as it reports with detailed EDA, custom filters, and code export. **Code export is the main highlight** of this library that makes it better than others.

References:

[1] Towards Data Science (Aug 30, 2020): EDA with 1 line of python code

Loved the article? Become a Medium member to continue learning without limits. I'll receive a small portion of your membership fee if you use the following link, with no extra

cost to you.