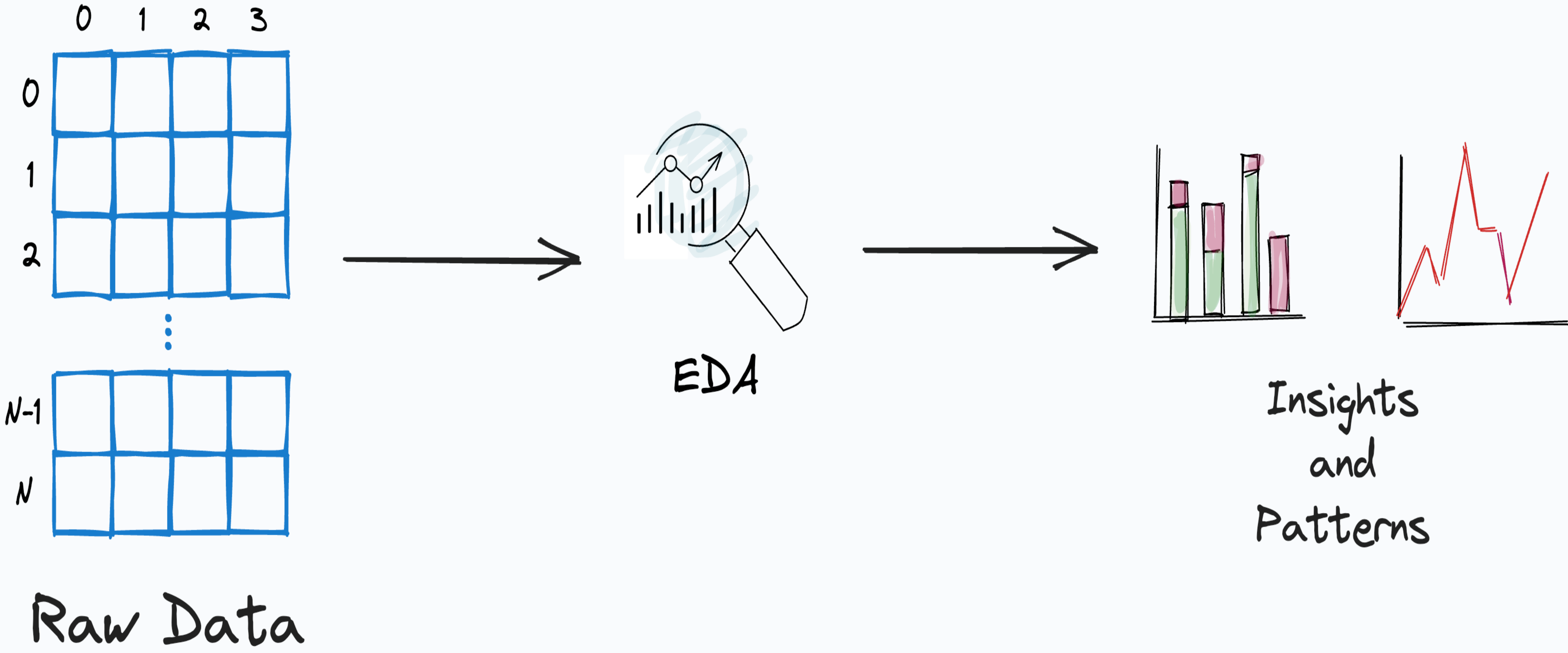
8 NEW AUTOMATED EDA TOOLS

Sridhar Padhi

INTRODUCTION

Exploratory Data Analysis is one of the popular components of a data science model development pipeline. A data scientist spends most of their time performing EDA to generate insights about the data.

Automated EDA packages can perform EDA in a few lines of Python code. In this article, we will discuss 8 automated tools that can perform EDA and generate insights about the data.


0
1
2
3
0
1
2
N-1
N
Raw Data
EDA
Insights
and
Patterns

TOOLS

1)SweetViz

2)Pandas-profiling

3)Data Prep

4)AutoViz

5)D-Tale

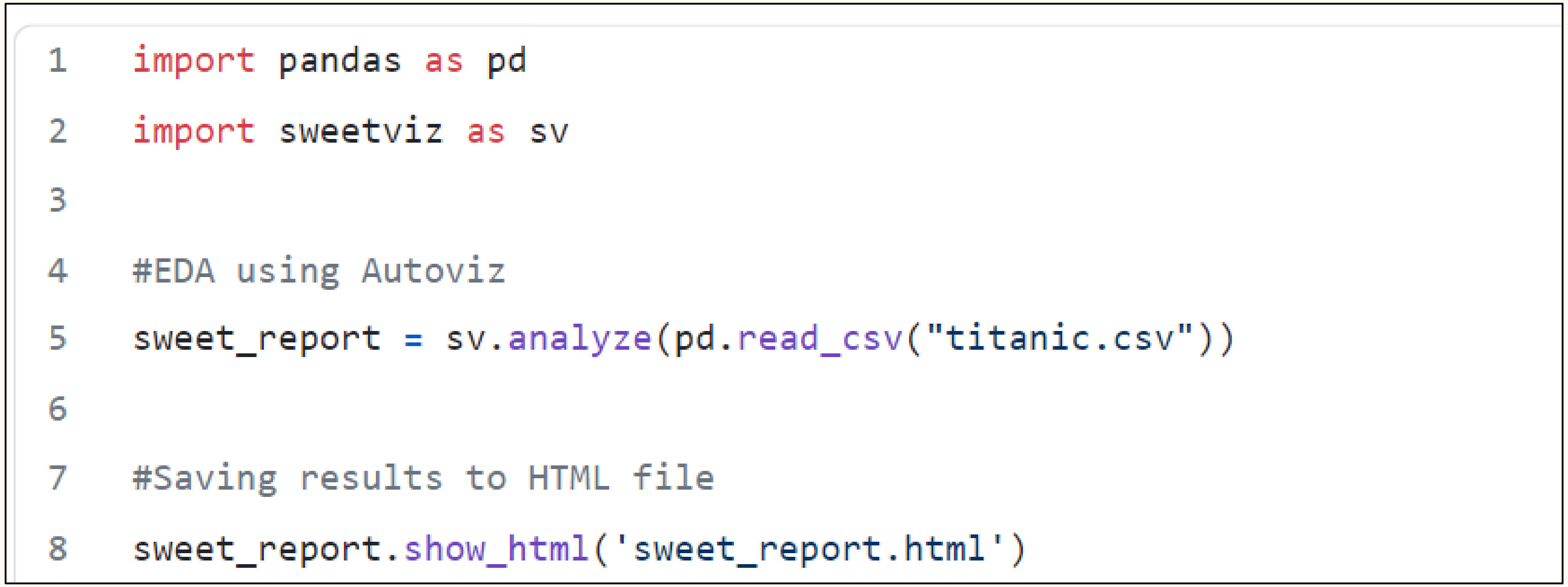
6)Dabl

7)QuickDA

8)Lux

1) Sweetviz:

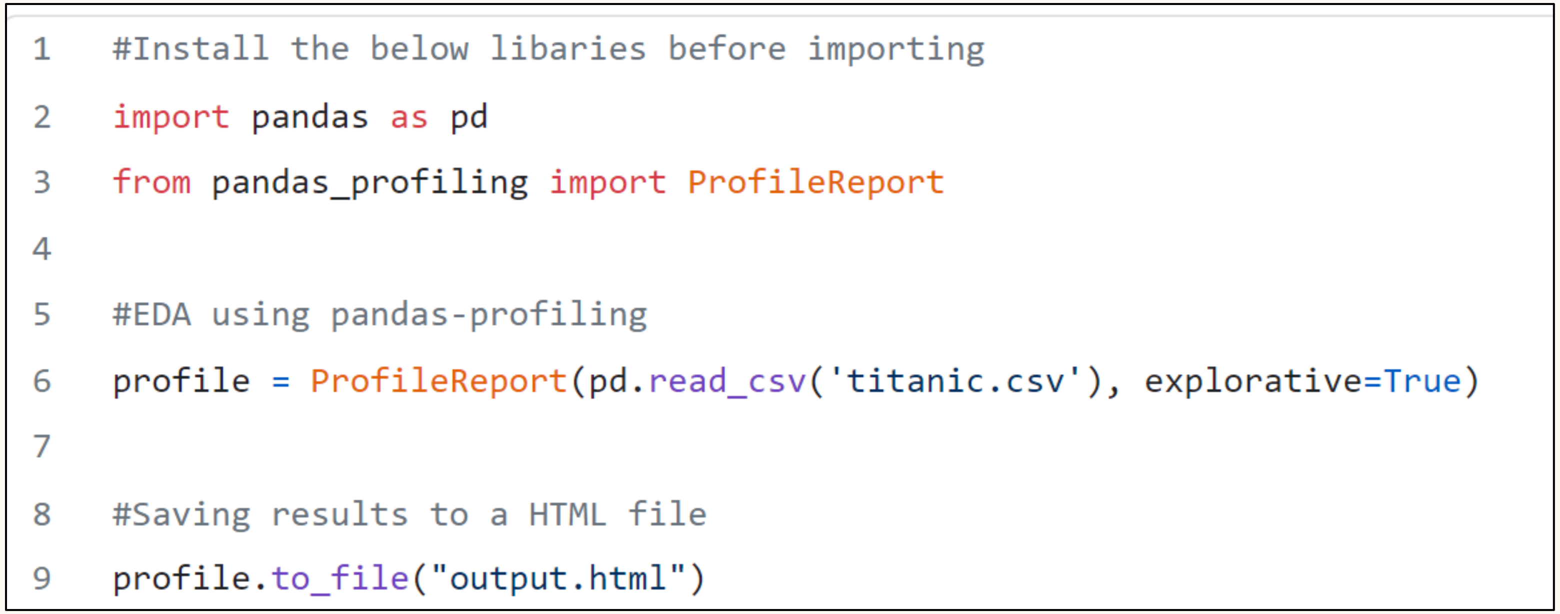
Sweetviz is an open-source Python library that generates beautiful, high-density visualizations to kickstart EDA (Exploratory Data Analysis) as an HTML application with just two lines of Python code. Sweetviz package is built around quickly visualizing target values and comparing datasets.



```
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')
```

2) Pandas-Profiling:

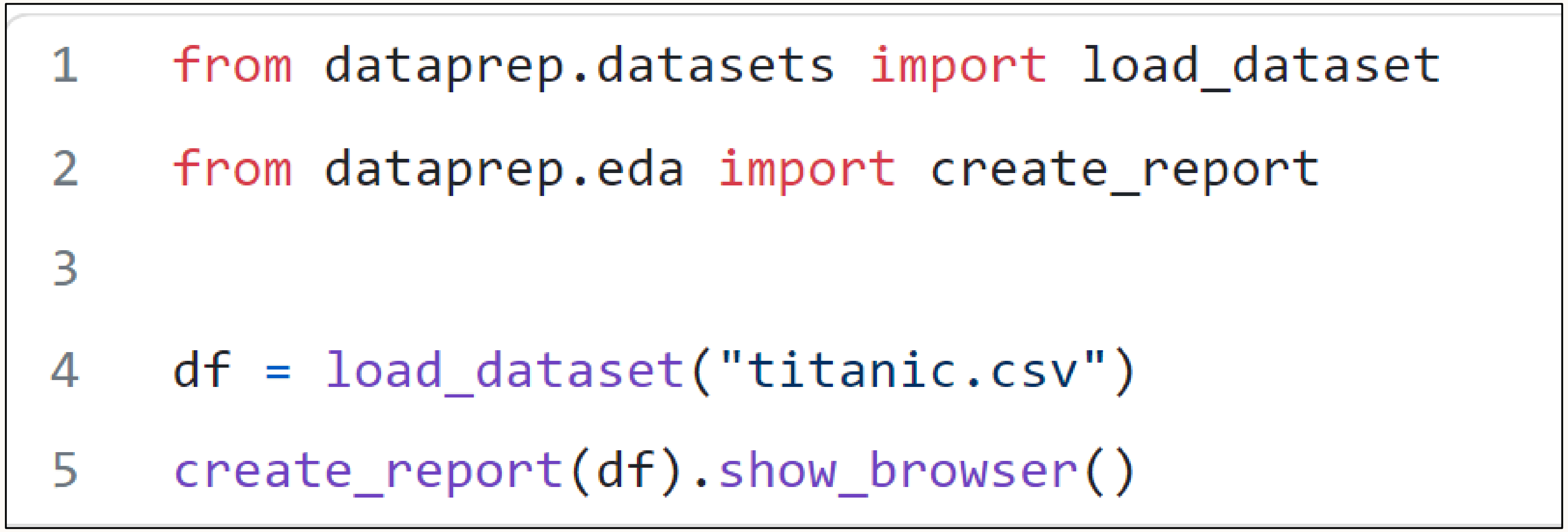
The pandas-profiling package generates profile reports of the Pandas DataFrame. Pandas-profiling extends the pandas DataFrame **df.profile_report()** for quick data analysis. Pandas-profiling works quite well with large datasets as it creates reports in a few seconds.



```python
#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")
```

3)Dataprep:

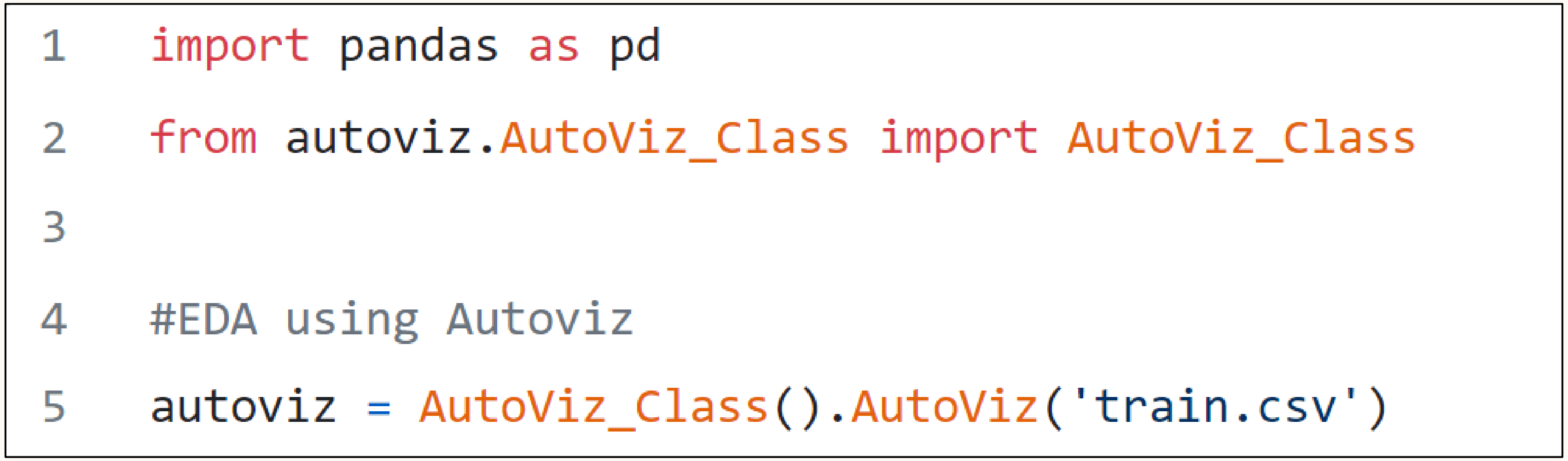
Dataprep is an open-source Python package built to analyze, prepare and process data. DataPrep is built on top of Pandas and Dask DataFrame and can be easily integrated with other Python libraries.



```
from dataprep.datasets import load_dataset
from dataprep.eda import create_report

df = load_dataset("titanic.csv")
create_report(df).show_browser()
```

4) AutoViz:

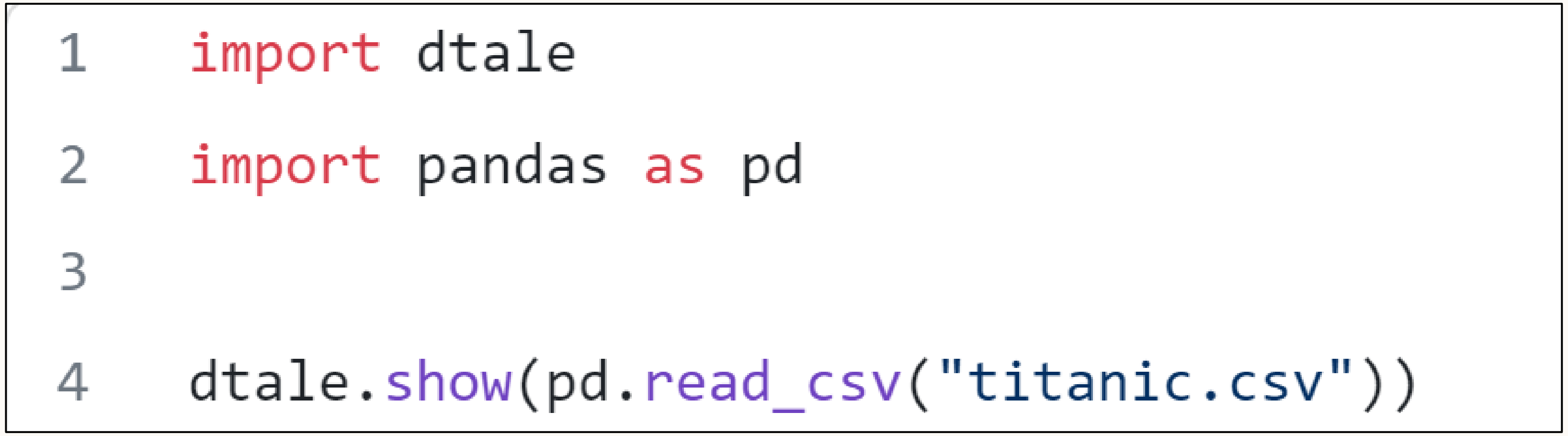
Autoviz package can automatically visualize any dataset, any size with a single line of code and automatically generate reports as HTML, bokeh, and image files. The user can interact with the HTML report generated by the AutoViz package.



```python
import pandas as pd
from autoviz.AutoViz_Class import AutoViz_Class

#EDA using Autoviz
autoviz = AutoViz_Class().AutoViz('train.csv')
```

5) D-Tale:

combines a Flask back-end and a React front-end that integrates seamlessly with ipython notebooks and terminals. Currently, D-Tale supports Pandas objects such as DataFrame, Series, MultiIndex, DatetimeIndex & RangeIndex.



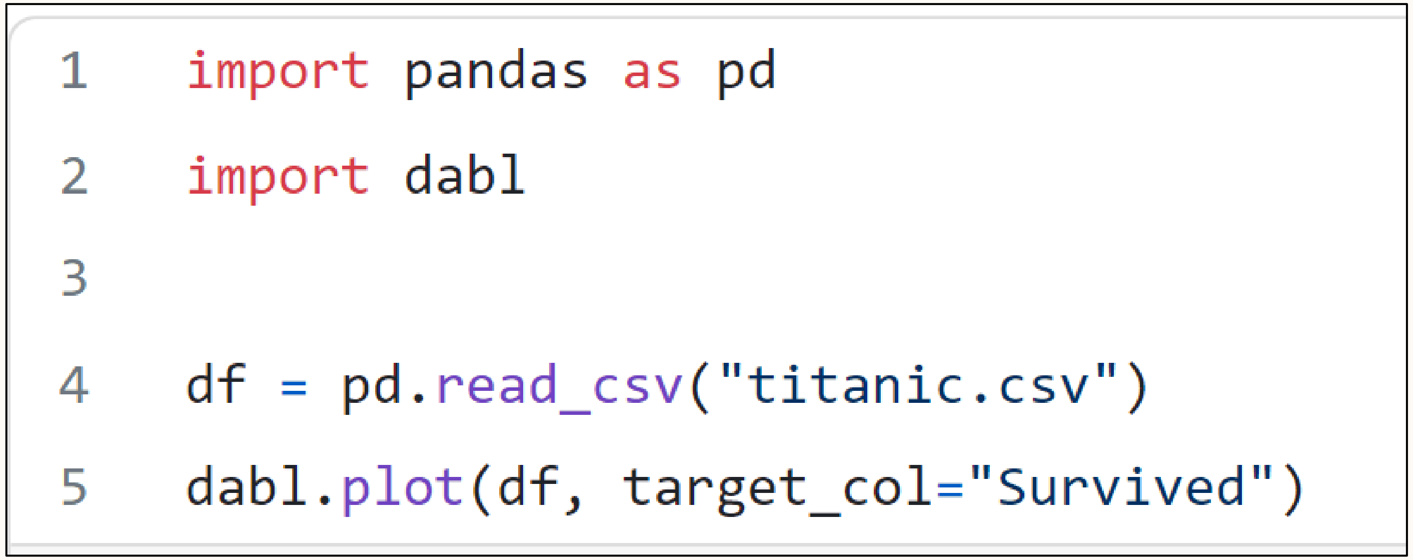
```python
import dtale
import pandas as pd

dtale.show(pd.read_csv("titanic.csv"))
```

6)dabl:

dabl focuses less on statistical measures of individual columns, and more on providing a quick overview via visualizations, as well as convenient preprocessing and model search for machine learning.

Dabl() function in dabl can feature visualization by plotting various plots including:

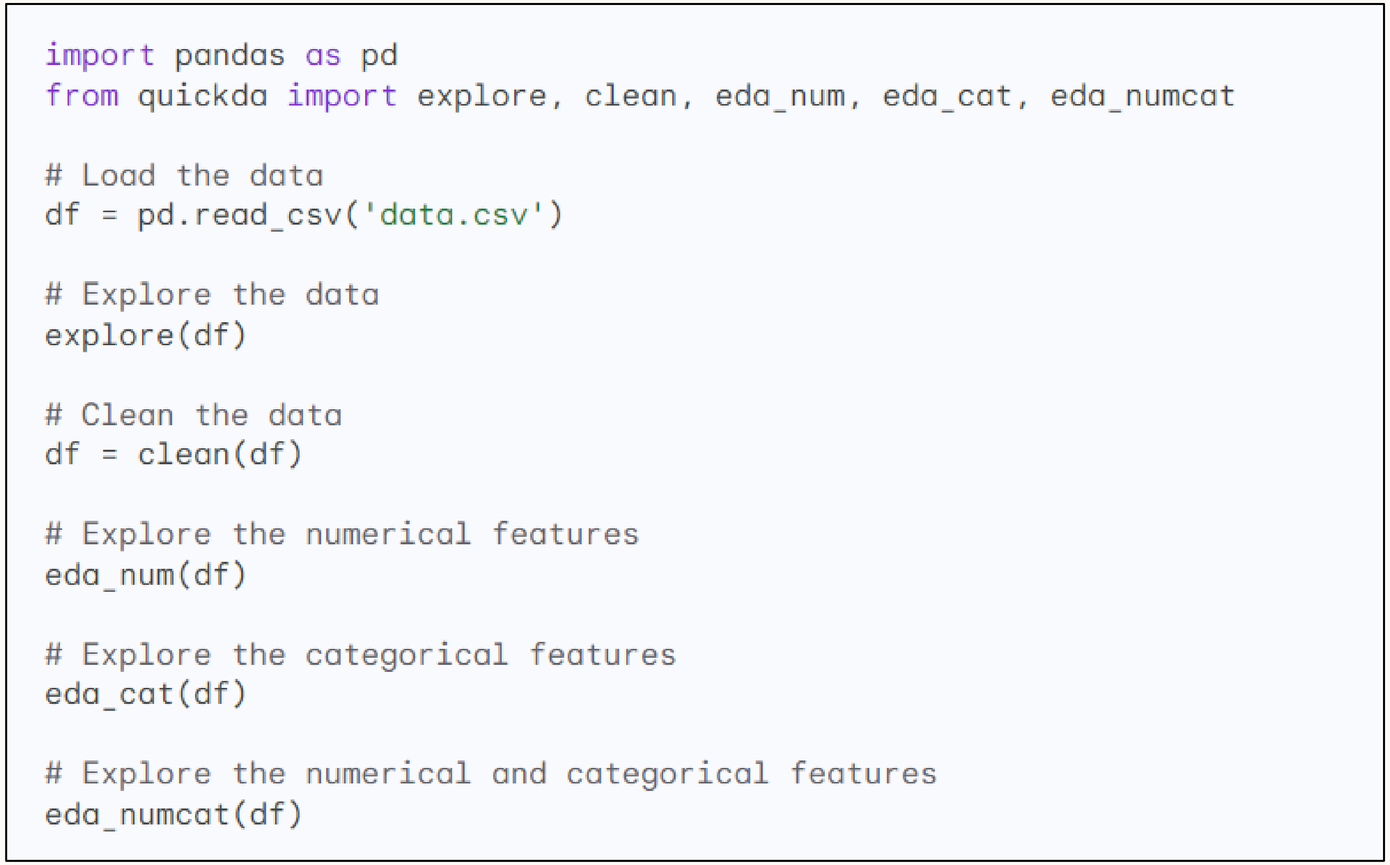
- Bar plot for target distribution
- Scatter Pair plots
- Linear Discriminant Analysis



```python
import pandas as pd
import dabl

df = pd.read_csv("titanic.csv")
dabl.plot(df, target_col="Survived")
```

7)QuickDA:

The QuickDA is an open-source EDA tool that helps to build a successful data model from the structured dataset. It covers info about missing values,data statistics,co-relation etc. It produces data alerts and plots data feature interactions.

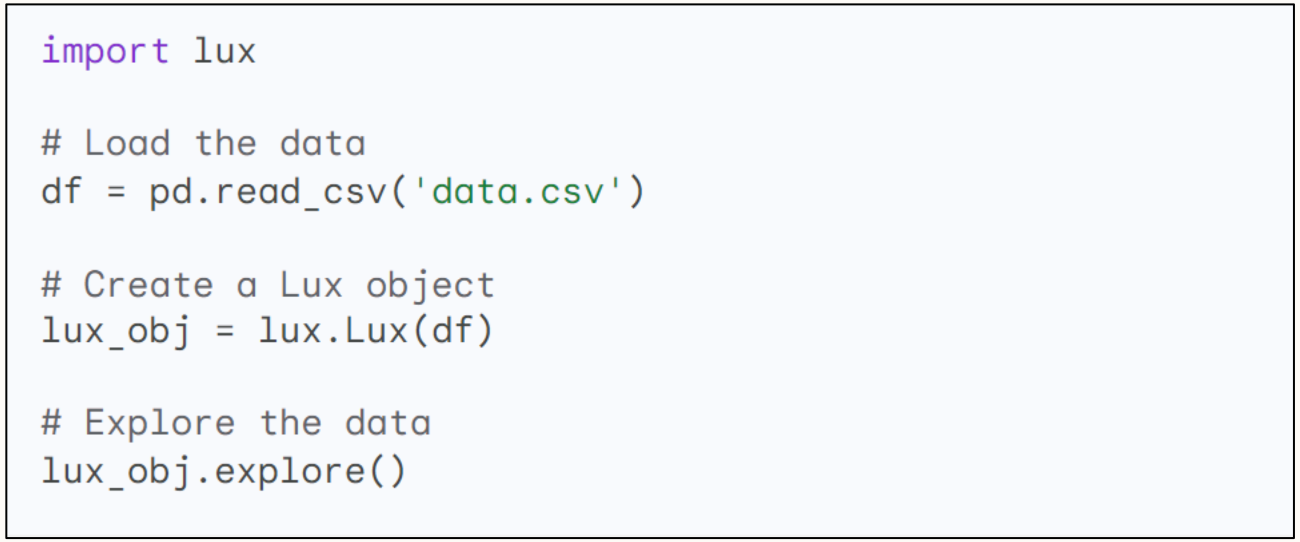


```
import pandas as pd
from quickda import explore, clean, eda_num, eda_cat, eda_numcat

# Load the data
df = pd.read_csv('data.csv')

# Explore the data
explore(df)

# Clean the data
df = clean(df)

# Explore the numerical features
eda_num(df)

# Explore the categorical features
eda_cat(df)

# Explore the numerical and categorical features
eda_numcat(df)
```

8)Lux:

Lux is a Python library that automates Exploratory Data Analysis (EDA) by generating interactive visualizations of your data. It can be used to explore data distributions, identify outliers, and find relationships between variables.

Lux is a great tool for anyone who wants to quickly and easily explore their data, especially for users with no prior experience with EDA.



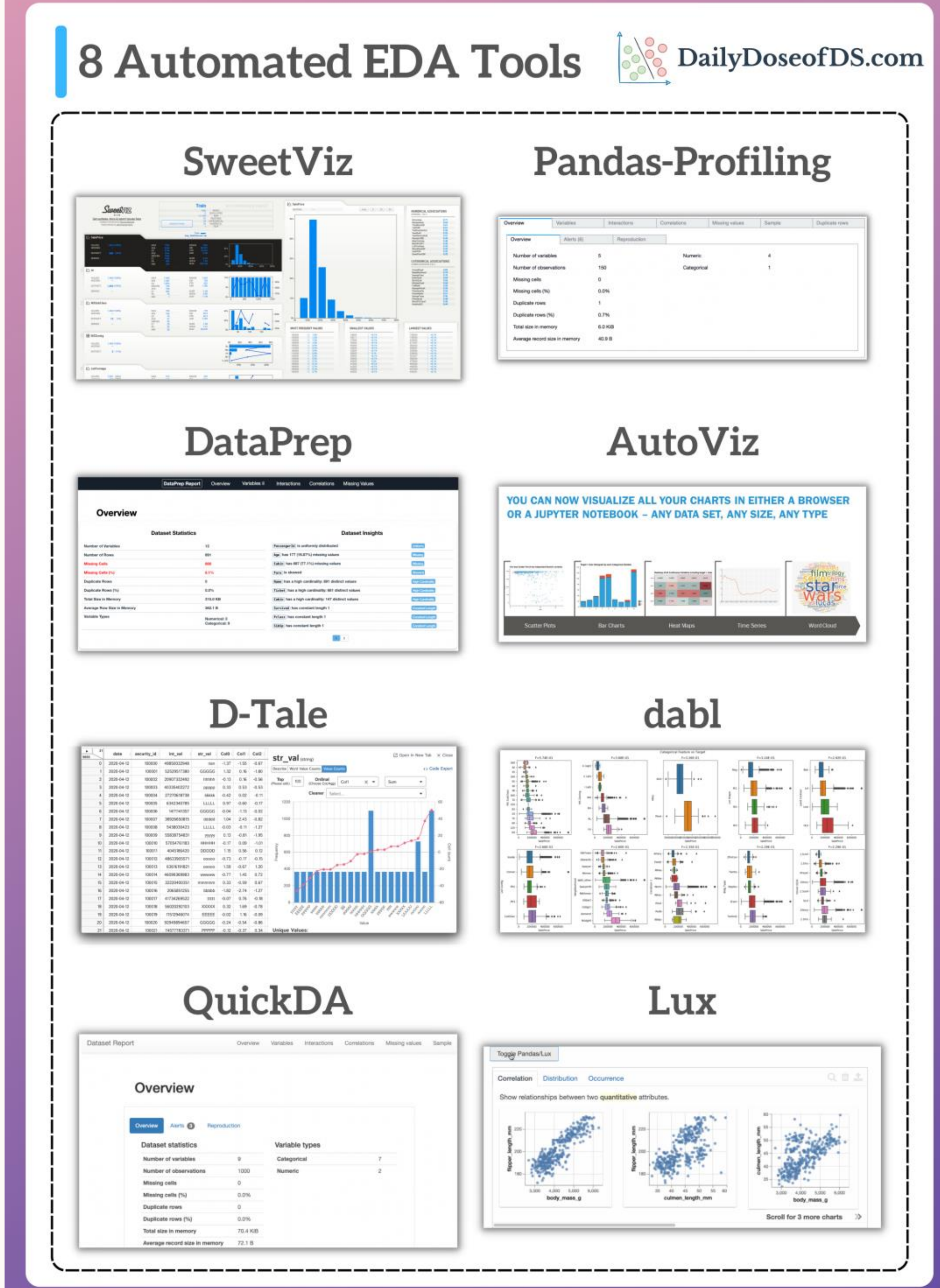
```
import lux

# Load the data
df = pd.read_csv('data.csv')

# Create a Lux object
lux_obj = lux.Lux(df)

# Explore the data
lux_obj.explore()
```

CONCLUSION

By uncovering hidden insights and patterns, one can make informed decisions about subsequent steps in the project.

Despite its importance, it is often a time-consuming and tedious task.

The above visual summarizes 8 powerful EDA tools, that automate many redundant steps of EDA and help you profile your data in quick time.



THANK YOU

Sridhar Padhi
sridharpadhi19@gmail.com