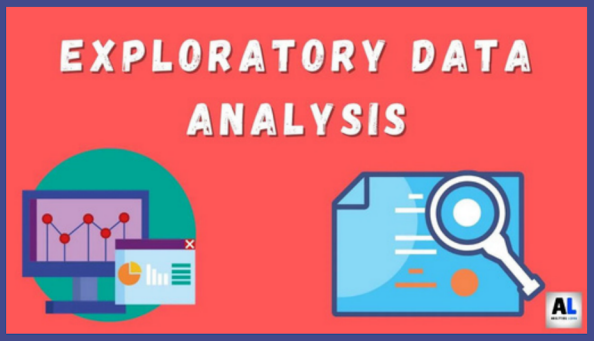
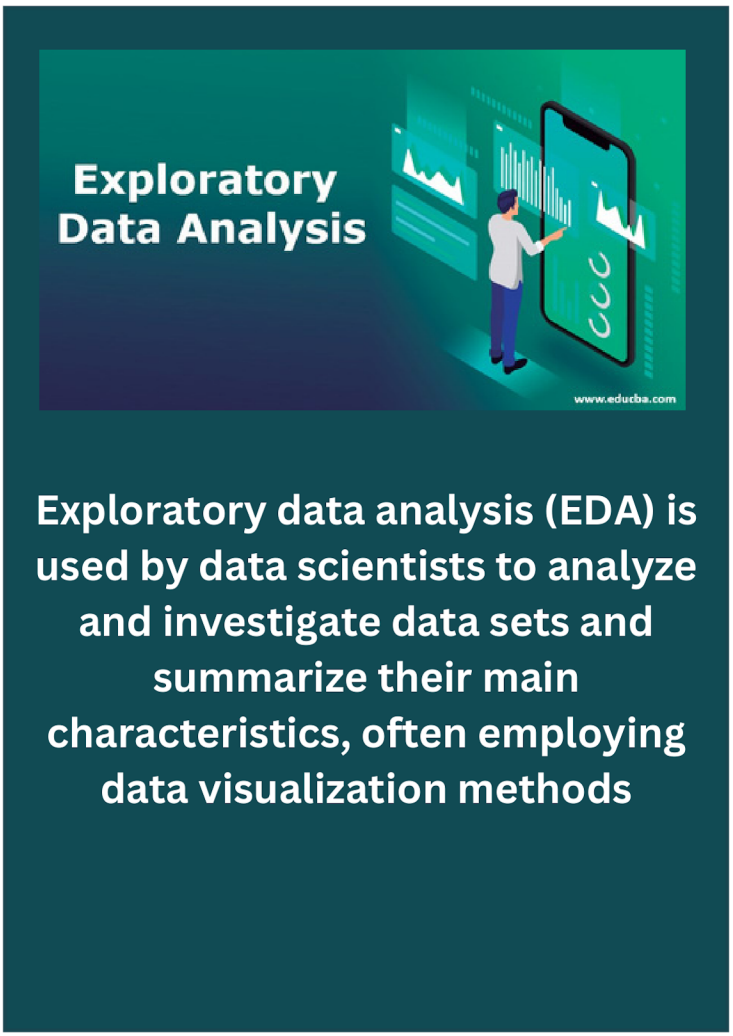
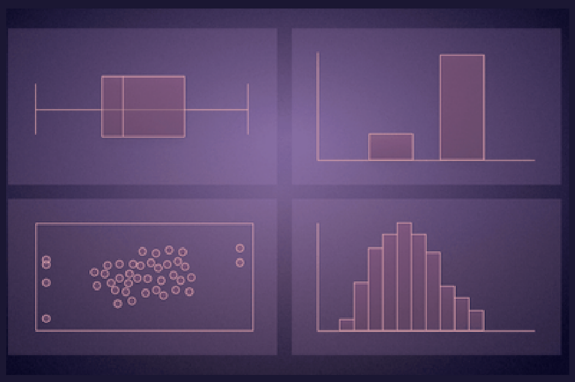
What is a EDA?



By Sachin Kumar


Exploratory
Data Analysis
www.educba.com
Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods

It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.



EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them.

Why is EDA important in data science?

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

EDA Tools

1.Python : An interpreted, object-oriented programming language with dynamic semantics. Its high-level, built-in data structures, combined with dynamic typing

2.R: An open-source programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.

Dataset

For the simplicity, we used a single dataset. We use the employee data for this. It contains 8 columns namely – First Name, Gender, Start Date, Last Login, Salary, Bonus%, Senior Management, and Team.

Github Link For Complete Dataset -

–

Code

```
import pandas as pd
import numpy as np

df = pd.read_csv('employees.csv')
df.head()
```

Start Date	Last Login Time	Salary	Bonus %	S
8/6/1993	12:42 PM	97308	6.945	
3/31/1996	6:53 AM	61933	4.170	
4/23/1993	11:17 AM	130590	11.858	
3/4/2005	1:00 PM	138705	9.340	
1/24/1998	4:47 PM	101004	1.389	

Getting insights about the dataset

1. **df.shape**

   Output:
   (1000, 8)

2. **df.info()**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   First Name         933 non-null    object
 1   Gender             855 non-null    object
 2   Start Date         1000 non-null   object
 3   Last Login Time    1000 non-null   object
 4   Salary             1000 non-null   int64
 5   Bonus %            1000 non-null   float64
 6   Senior Management  933 non-null    object
 7   Team               957 non-null    object
dtypes: float64(1), int64(1), object(6)
memory usage: 62.6+ KB
```

df.describe()

	Salary	Bonus %
count	1000.000000	1000.000000
mean	90662.181000	10.207555
std	32923.693342	5.528481
min	35013.000000	1.015000
25%	62613.000000	5.401750
50%	90428.000000	9.838500
75%	118740.250000	14.838000
max	149908.000000	19.944000

Handling Missing Values

There are several useful functions for detecting, removing, and replacing null values in Pandas DataFrame :

- isnull()
- notnull()
- dropna()
- fillna()
- replace()
- interpolate()

Describe method

The describe() function applies basic statistical computations on the dataset like extreme values, count of data points standard deviation, etc. Any missing value or NaN value is automatically skipped. describe() function gives a good picture of the distribution of data.

Data visualization

Data Visualization is the process of analyzing data in the form of graphs or maps, making it a lot easier to understand the trends or patterns in the data. There are various types of visualizations –

1. Univariate analysis
2. Bi-Variate analysis
3. Multi-Variate analysis

Handling Outliers

An Outlier is a data-item/object that deviates significantly from the rest of the (so-called normal)objects. They can be caused by measurement or execution errors. The analysis for outlier detection is referred to as outlier mining. There are many ways to detect the outliers, and the removal process is the data frame same as removing a data item from the panda’s dataframe.

Thankyou