Exploratory Data Analysis

Ankita Singh
C74 Batch

INTRODUCTION

(Exploratory Data Analysis) used to analysis of dataset using different liberaries (Pandas,Numpy,matplotlib,seaborn).

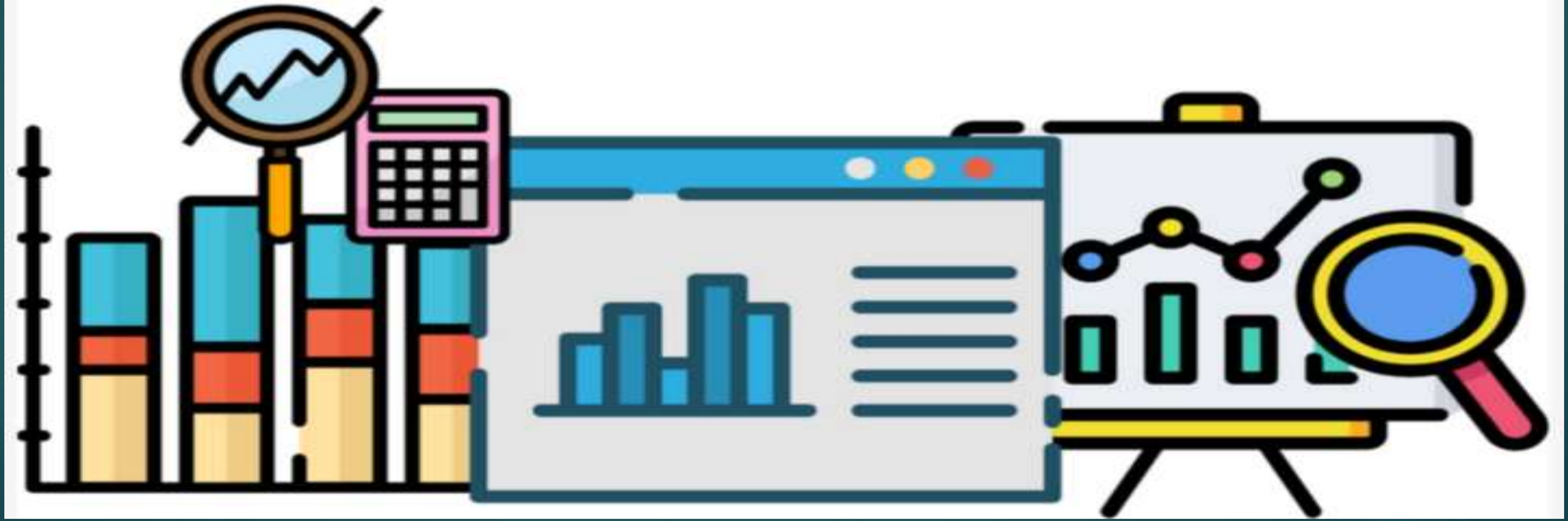
EDA can have the insights of dataset, missing values,outliers(anomalies) further analysis methods.

Used tools box plot, Histogram, bar plot, pie chart more other tools to analyse the data variables.

Sourcing the dataset

- Excel/csv data import to jupyter using pandas liberary

```
pd.read_csv(r'C:\Users\akku\Downloads\application_data.csv')
```



Cleaning of dataset

- Checking the missing values(NAN) in the dataset.
- df.isnull().sum().sum()
- Dropping all the null values from rows and columns
- Drop the unusual duplicates from the dataset
- Remove the space in the dataset

Handling the missing values

- Handling missing values with (0,1),mean,median,mode or any other categorical values depending on the dataset.
- In columns (/) removes using replace(“/”,””,regex).
- Check print df again,i.e., dataset is cleaned.
- Print column df['NAME_EDUCATION_TYPE'] to ckeck no unusual format or data present in the dataset.

Handling the outliers

- Outliers are values beyond normal range.
- Outliers are treated by imputation,dropping,binning,capping.
- Outliers analysis are two types,i.e,. Univariate analysis, Multivariate analysis.
- Outliers are visualize by Box plot, scattered plot, Histogram,Pair plot.

Visualization Of Outliers

- Visualize the data by using the boxplot.
- Use Target variable with other variable to know about the correlation(corr()).
- Heatmap to know about the correlation based on annotation and cmap given.
- Use groupby method to categorize the data and fine mean & median of variable.

Multivariate Analysis

- By using pivot table we can have multivariate analysis.
- By using Heatmap we can plot the variable.
- Using annot & cmap we can have annotation of data and color according to the data, also using center value we can have center value of variable.

Other CSV dataset

- Upload other csv dataset using same method of liberary in python.

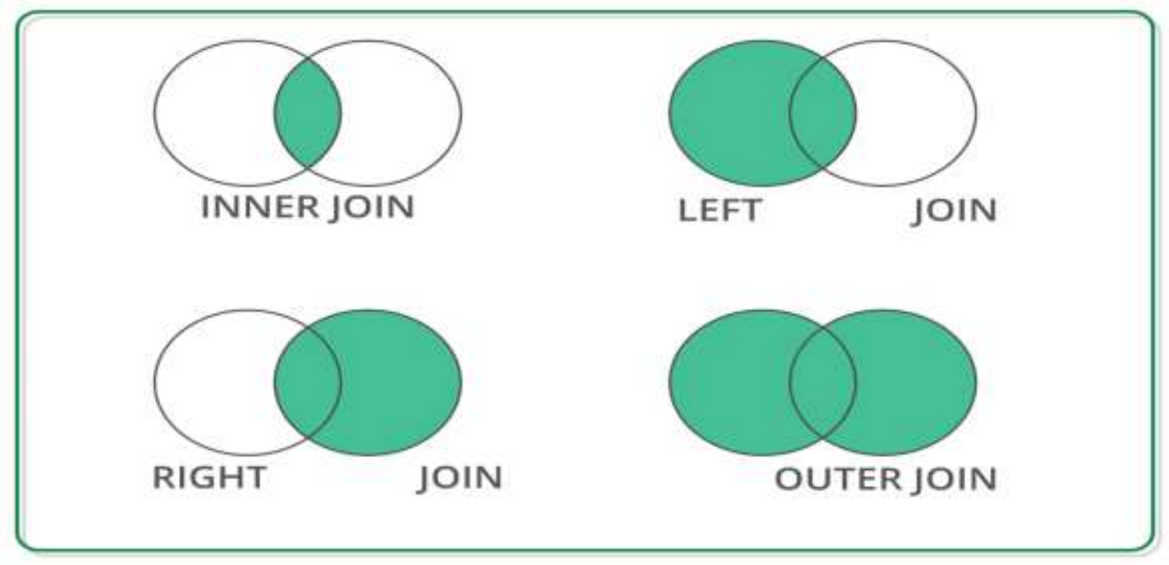
```
pd.read_csv(r'C:\Users\akku\Downloads\application_data.csv')
```

- Cleaning od dataset, dropping missing values (NAN),Dropping duplicates & plotting outliers in the dataset.
- We can have analysis (univariate,bivariate & multivariate) and plot dataset for visualization.

Merge & concatenation

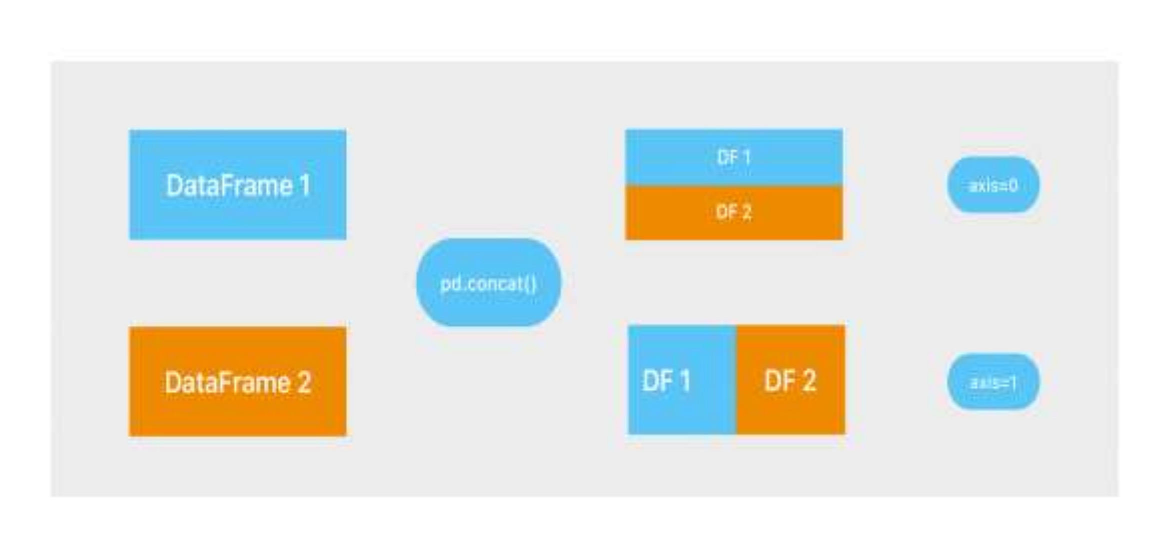
- Merge the datasets

```
merge_df=pd.merge(df,df1,on="NAME_CONTRACT_TYPE",how="left")
```



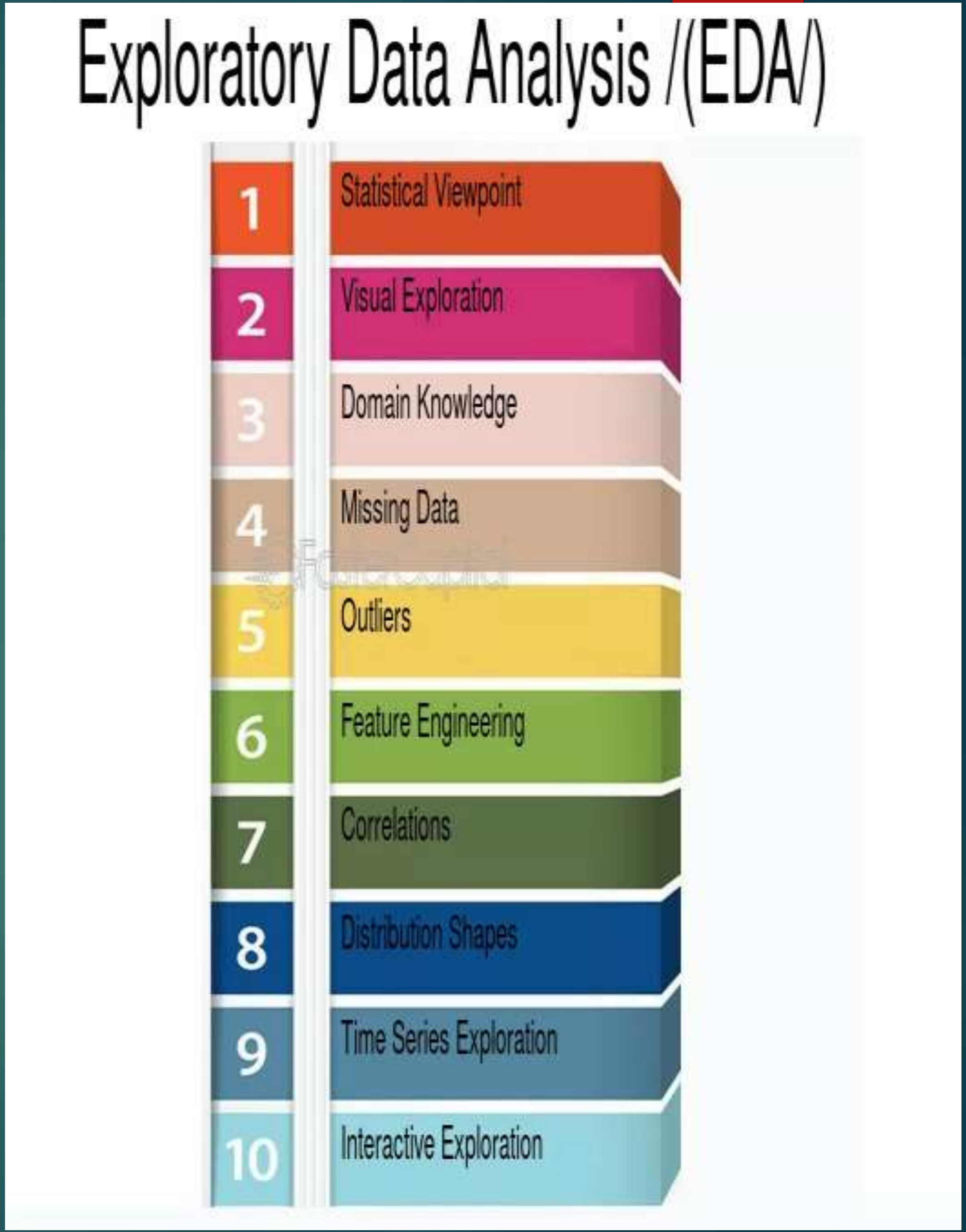
- Contatenation of datasets

```
df_con1=pd.concat([df,df1],axis=1)
```



Conclusion

- We have the insights & trends in the dataset(distribution of values,correlation,missing values)
- We can have the skewness in the data and outliers present in data.
- We have the statistical values in the dataset
- By Cleaning, EDA we can improve our data and can use for further analysis.



Exploratory Data Analysis /(EDA/)

1 Statistical Viewpoint

2 Visual Exploration

3 Domain Knowledge

4 Missing Data

5 Outliers

6 Feature Engineering

7 Correlations

8 Distribution Shapes

9 Time Series Exploration

10 Interactive Exploration

THANK YOU