Journal of Science and Technology
ISSN: 2456-5660 Volume 7, Issue 09 (November2022)
www.jst.org.in *DOI:https://doi.org/10.46243/jst.2022.v7.i09.pp01-11*

Graphical Exploratory Data Analysis (GEDA): A Case Study on Employee Attrition

Dr. Ayesha Banu, Dr. Sharmila Reddy, M. Rama
Associate Professor & Head, CSE(Data Science).Vaagdevi College of Engineering
headcsd@vaagdevi.edu.in.
Associate Professor & Head, CSE(Data Science).Vaagdevi Engineering College
sharmilakreddy@vecw.edu.in.
Assistant Professor,CSE,Vaagdevi College of Engineering
rama_m@vaagdevi.edu.in

To Cite this Article
Dr. Ayesha Banu, Dr. Sharmila Reddy, M. Rama, **Graphical Exploratory Data Analysis (GEDA): A Case Study on Employee Attrition""**,*Journal of Science and Technology, Vol. 07, Issue 09,-November 2022, pp*01-11
Article Info
Received: 28-09-2022 Revised: 17-09-2022 Accepted: 25-10-2022 Published: 5-11-2022

Abstract
Exploratory Data Analysis (EDA) popularly performs some preliminary investigations on the dataset to understand its content and structure. EDA is a mandatory step in the complete process of data analysis, since its mandatory to analyze the data in order to produce good results and in turn help in decision making. There are several Graphical EDA techniques which not only analyze the data but also present the results in graphical form. This paper uses the Python programming language for both data analysis and visualization of results. The rich set of python libraries including pandas, numpy, matplotlib, seaborn etc greatly supports the process of GEDA. This paper works on the "Employee Performance and Attrition" dataset to analyze and extract potential information and present results in graphical form.

Keywords: *Graphical Exploratory Data Analysis, Python, Data Visualization,matplotlib, seaborn.*

1. INTRODUCTION

In today's world of technology data is growing very fast in both volumes and variety and it has become highly impossible to understand and analyze the data manually. Data analysis is collection of different processes to inspect, clean, transform, and model the data with an objective of discovering potentially useful information, drawing several conclusions, and finally supporting decision-making.Exploratory Data Analysis (EDA)evaluates or comprehends data and is a significant component of any process in data science or machine learning.It helps in exploring the data; understanding the structure and relationships between variables andbuilds a consistent and valuable output.

Python is a very popular programming language today due its flexibility andwide collection of inbuilt libraries, which are very essential to performdata analytics and complex computations.Pythonsupports multiple libraries for data analytics like NumPy for mathematical and statistical calculations and PandasthePython Data Analysis Library.Data visualization plays a vital role in representingthe data and also complex data relationships graphically such that it is easy to understand. Python has many libraries that support for displaying data in the form of charts, graphs , plots and animations. Two such popular libraries used in this work are Matplotlib and Seaborn.

This paper works on the "Employee Performance and Attrition" dataset to perform Exploratory Data Analysispresent results in graphical form using python. The paper is organized in to four sections. Section 2 briefs the review of literatureand section 3 explains the differenttechniques for EDA both non graphical and graphical. In section 4, the graphical exploratory data analysis is studied on the Employee Attrition dataset using python.

2. LITERATURE SURVEY

Aindrila et al. [1] made a study on thetools for data visualization with respect to their efficacy in theEDA process.They also examined the scalability of the exploration tools for analyzing large datasets. Matthew NtowGyamfi et al. [2] investigated the commercial banks practices regarding credit risk and loan default to find the causes of nonperformingloans.X.FrancisJency et al. [5]have performed EDA on bank data to understand the nature of clients who apply for loans in banks. Based on the results they applied machine learning algorithms for loan prediction and classified the clients as good customer and bad customer.

K. Ulaga Priya1 et al. [4] have done EDA on bank dataset using random forest algorithm to predictcustomers loan privilege in R programming for analysis.Kiranbala&Deepika [5] have performed EDA both numerical and graphical on the World Happiness report 2021 to understand the various aspects of data analysis. KabitaSahoo et.al [6] have done EDA using python to understand the different libraries of python for data analysis and graphical representation of results.

3. TECHNIQUES FOR EDA

In the complete process of data analysis, after collecting the data and pre processing it,EDA is the very important step for data manipulation, plotting and visualization.Most of the EDA techniques are graphical and few are quantitative which help in analyzing the data sets with respect to their statistical characteristics. The techniques available for Exploratory Data Analysis (EDA) are broadly classified in to Non-Graphical EDA and Graphical EDA where in both the techniques are classified in to two types namelyunivariate and multivariate [7]. Some of the EDA techniques depend on the type of data on which they are applied and some depend on the purpose of the analysis. Table 1 shows the preferable EDA technique that can be adopted for a given type of data and purpose of analysis.

3.1. Non-Graphical Exploratory Data Analysis: NGEDA

Thistechniques help in providing an ideaabout the description anddistribution of the variable(s). There are two methods under this category namely univariate and multivariate.

3.1.1 Univariate NGEDA: This is a principal form of data analysis that involves only one variable to identify underlying data distribution and the characteristics of population distribution.This analysis also covers outlier detection.For any quantitative variableUnivariate EDA helps making initial assessments on the variable distributionusing the data sample.

Type of data	***Preferable EDA techniques***	***Purpose***	***Preferable EDA techniques***
Categorical	Descriptive statistics	distribution of a variable	Histogram
continuous Univariate	Histograms ,Line plot	Outlier detection	Histogram, scatter plots, box-and-whisker plots
continuous Bivariate	Heatmap,2D arrays and scatter plots	Quantify the relationship betweentwo variables	2D scatter plot , Covariance and correlation
trivariate	3D scatter plot	Visualize the relationship between two exposure variables	Heatmap
Multiple groups	Side-by-side box plot	Visualizinghigh-dimensional data	2D or 3D scatter plot

Table 1:EDA techniques preferable based on data type and analysis

The fundamental description of the distribution include:
A. Central tendency: For any population distribution the central tendencymeasures mean, median, and mode where median is preferred for skewed distribution or when there are outliers.
B.Spread: Spread indicates how far from the centrecan wefind the data values.Variance, interquartile range and standard deviation are the commonly used measuresfor finding spread of any distribution.The variance is computed by taking the mean of the squares of all the individualdeviations andwe get the standard deviationby taking the square root of the variance.
C. Skewnessand kurtosis: These areextradescriptors for any distribution where themeasure of asymmetry is called skewnessandmeasure of peakedness is called kurtosis [8].

3.1.2. Multivariate NGEDA: This shows the relationship between multiplevariables as a cross-tabulation or statistics. Cross-tabulation will be of great usefor categorical datawhich is a simple extension of tabulation.This is a two-way table with columnsrepresentingheadingswhich match with one of the variables andthe row headings match withthe other variable. The subject count thatshare commonpair ofvalues are filled in to the table.This is also called as thebivariate non-graphical EDA technique.For categorical variables we can also calculate the correlation and covariance[8].Table2 shows three columns where column1 contains the course, column2 holds the age of the person pursuing the course and column3 shows the gender. Table 3 shows the cross tabulation for the data of table2.

Course	**Age**	**Gender**
CS121	youth	F
CS222	Middle age	F
CS431	youth	M
CS506	youth	M
CS222	Middle age	F
CS121	Middle age	F
CS431	Senior	F
CS222	Senior	F

CS431	youth	M
CS506	Senior	F
CS121	youth	F
CS222	Middle age	M

Age/Gender	**Female**	**Male**	**Total**
youth	2	3	5
Middle age	3	1	4
Senior	3	0	3
Total	8	4	12

Table 3: Cross Tabulation for Course Data set
Table 2: Course Data Set

3.2. Graphical Exploratory Data Analysis

This isa graphical method of NGEDA.Non-graphical methods mostly are objective and quantitative in nature. They fail to givecompleterepresentation of the data. GEDAisfound to be more qualitative. This data analysis is also divided in to univariate and multivariate.

3.2.1. Univariate Graphical EDA: The primary focus of thisanalysis is on the data from asingle variable values on n subjects and graphically represents the distribution of the data. Some of the common forms of univariate graphics include:
A. Histogram: This is thefirst fundamental graph also called as bar plot where every barrepresents the frequency or proportion for agiven range of values.Histograms helpto learn about theshape, spread, central tendency and outliers of the given data.
B. Boxplots: This is another graphical technique which is very useful to represent the data proportions and information related to the skew, symmetry, central tendency and outliers. These are excellent techniquesas theydepend on powerful statistical measures includingmedian and IQR instead of mean and standard deviation. Distribution comparisons are easily done usingboxplots.
C. Quantile-Normal plots: These plots are also called as QN plot or QQ plotquantile-quantile. These are considered to be more complicated plots. QQ plots are best suitable to observe which theoretical distribution does the data particularly follow[8].

3.2.2. Multivariate graphical EDA: Thesetechniques represent the association between two or more knowledge sets graphically. Some primaryways techniques of multivariate graphics include:

A. Scatter Plot:this is called as aessential graphical EDA technique when the variables are quantitative and it plots variable 1 on x-axis, and variable 2on y-axis with one point corresponding to every casein the given dataset.If anytwo variable are explanatory and outcome, then it is always recommended to plot the outcome variable on y axis.

B.Run chart: To plot the data over time we can use the Run Chart.

C.Heat map: The graphical representation which depicts the data values usingcolours.

D.Bubble chart: It displays bubbles- multiple circles in two-dimensional plot.

4. Graphical Exploratory Data Analysis (GEDA) Using Python

A. Why Python:

Python is an interpretedprogramming language with a very rich set of libraries supporting both procedural and object-oriented programming paradigms. Some of the essential features of python are its free open source which is portable with support of numerous IDE [9].

B.Packages:

The packages of python used in this study include
• Pandas
• Numpy
• Matplotlib
•Seaborn

C. Dataset :

IBM HR Analytics Employee Attrition & Performance data set downloaded from *https://www.kaggle.com/code/faressayah/ibm-hr-analytics-employee-attrition-performance/data*.

This data set has a total of 27 attributes describing the employee with respect to age, gender, job role, job satisfaction and many more. In this work we consider only 16 attributes which show the employee performance and attrition. A part of the dataset is given in figure 1.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	EmpNuml	Age	Gender	EducationBackg	MaritalSta	EmpDepai	EmpJobRo	EmpEduca	EmpEnvir	EmpJobIn	EmpJobLe	EmpJobSa	EmpRelati	EmpWork	Attrition	PerformanceRating	
2	E1001000	32	Male	Marketing	Single	Sales	Sales Exec	3	4	3	2	4	4	2	No	3	
3	E1001006	47	Male	Marketing	Single	Sales	Sales Exec	4	4	3	2	1	4	3	No	3	
4	E1001007	40	Male	Life Sciences	Married	Sales	Sales Exec	4	4	2	3	1	3	3	No	4	
5	E1001009	41	Male	HR	Divorced	HR	Manager	4	2	2	5	4	2	2	No	3	
6	E1001010	60	Male	Marketing	Single	Sales	Sales Exec	4	1	3	2	1	4	3	No	3	
7	E1001011	27	Male	Life Sciences	Divorced	Dev	Develope	2	4	3	3	1	3	2	No	4	
8	E1001016	50	Male	Marketing	Married	Sales	Sales Repı	4	4	3	1	2	4	3	No	3	
9	E1001019	28	Female	Life Sciences	Single	Dev	Develope	2	1	1	1	2	4	3	Yes	3	
10	E1001020	36	Female	Life Sciences	Married	Dev	Develope	3	1	4	3	1	1	3	No	3	
11	E1001021	38	Female	Life Sciences	Single	Dev	Develope	3	3	3	3	3	4	4	No	3	
12	E1001022	44	Male	Medical	Single	Dev	Develope	3	1	1	1	3	3	3	No	3	
13	E1001024	47	Female	Medical	Divorced	Sales	Sales Exec	3	4	3	4	3	4	2	No	3	
14	E1001025	30	Male	Marketing	Divorced	Sales	Sales Exec	5	3	3	2	4	4	2	No	4	
15	E1001027	29	Male	Life Sciences	Single	Sales	Sales Repı	3	3	3	1	3	3	3	No	3	
16	E1001030	42	Male	Medical	Divorced	Dev	Develope	3	3	4	1	3	4	3	Yes	3	
17	E1001035	34	Female	Medical	Single	Dev	Develope	2	2	3	2	3	4	3	No	3	
18	E1001038	39	Female	HR	Married	HR	HR	3	3	4	2	2	3	1	No	3	
19	E1001040	56	Male	Medical	Married	Dev	Develope	3	3	3	4	4	3	2	No	3	
20	E1001041	40	Female	Medical	Single	Dev	Develope	1	4	2	1	4	4	2	No	4	
21	E1001042	27	Female	Medical	Single	Dev	Develope	3	4	2	2	1	1	1	No	3	
22	E1001044	29	Male	Marketing	Divorced	Sales	Sales Repı	3	4	3	1	2	4	3	No	3	
23	E1001047	53	Male	Life Sciences	Single	Dev	Develope	3	4	3	2	4	4	3	No	3	
24	E1001049	35	Female	Life Sciences	Divorced	Dev	Senior De	4	4	3	2	1	1	4	No	3	
25	E1001050	32	Male	Life Sciences	Married	Dev	Develope	4	1	3	2	4	4	3	No	3	

Fig 1: A Snippet of the Employee-Attrition Dataset

D. Using Python and Working with the dataset

- *Importing libraries:*To start the analysis work on the data set we first need to import all the required python libraries necessary for the analysis process.

import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import io

import seaborn as sns

- *Importing dataset:* after importing all the necessary python libraries we need to import the dataset. The datasetis imported in tojupyter notebook using following code.

  mydata=pd.read_csv("Employee.csv")
  mydata

  If we use Google Colab then the code to import the data set is

  fromgoogle.colab import files
  uploaded = files.upload()
  df = pd.read_csv(io.BytesIO(uploaded['Employee.csv']))

- *Cleaning Data:* Before we start using the data we need to check if there are any missing values or null values for which we use the isnull( ) method which returns true wherever there is no value in the dataset. We can also use the sum( ) method which returns total number of null values in each column. Zero indicates that there are no null values in the columns.

```
df = pd.read_csv(io.BytesIO(uploaded['Employee.csv']))
df.isnull().sum()

Choose Files  Employee.csv
• Employee.csv(text/csv) - 95858 bytes, last modified: 9/3/2022 - 100% done
Saving Employee.csv to Employee (2).csv
EmpNumber                       0
Age                             0
Gender                          0
EducationBackground             0
MaritalStatus                   0
EmpDepartment                   0
EmpJobRole                      0
EmpEducationLevel               0
EmpEnvironmentSatisfaction      0
EmpJobInvolvement               0
EmpJobLevel                     0
EmpJobSatisfaction              0
EmpRelationshipSatisfaction     0
EmpWorkLifeBalance              0
Attrition                       0
PerformanceRating               0
Age1                            0
dtype: int64
```

Fig 2: Checking for null values in the dataset

We can also selectonly particular rows from the entire datasetfor analysis using head(n) function which extracts top n rows and tail(n) function which extracts n rows from the bottom of the dataset.

df = pd.read_csv(io.BytesIO(uploaded['Employee.csv']))
top=df.head(100)
top=df.tail(50)

E. Exploratory Data Analysis

This method analyzethe data sets and summarize the important characteristics of the data using data visualization toolsand statistical graphics.Even if any statistical model is used or not, EDA primarily aims at seeing what the data shows beyond the hypothesis testing task or

Journal of Science and Technology
ISSN: 2456-5660 Volume 7, Issue 09 (November2022)
www.jst.org.in *DOI:https://doi.org/10.46243/jst.2022.v7.i09.pp01-11*

formal modeling. John Tukey was thefirst to promote this EDA to encourage statisticiansto collect new data ,explore it, formulate hypotheses, and perform experiments [10].

All the column data types in a given dataset are printed using dtypes.The statistical summary such as mean, count, min, max, etcof the given dataframe can be extractedusing describe() function. To show the relation between any two we use correlationcorr( ) function which also helps in measuring the linear relation strength of any two variables.

```
df = pd.read_csv(io.BytesIO(uploaded['Employee.csv']))
df.dtypes
df.describe()
df.corr()
```

Choose Files Employee.csv

- **Employee.csv**(text/csv) - 95858 bytes, last modified: 9/3/2022 - 100% done

```
Saving Employee.csv to Employee (4).csv
```

	Age	EmpEducationLevel	EmpEnvironmentSatisfaction	EmpJobInvolvement	EmpJobLevel	EmpJobSatisfaction	EmpRelationshipSatisfaction	EmpWorkLifeBala
count	1200.000000	1200.00000	1200.000000	1200.000000	1200.000000	1200.000000	1200.000000	1200.000
mean	36.918333	2.89250	2.715833	2.731667	2.067500	2.732500	2.725000	2.744
std	9.087289	1.04412	1.090599	0.707164	1.107836	1.100888	1.075642	0.699
min	18.000000	1.00000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000
25%	30.000000	2.00000	2.000000	2.000000	1.000000	2.000000	2.000000	2.000
50%	36.000000	3.00000	3.000000	3.000000	2.000000	3.000000	3.000000	3.000
75%	43.000000	4.00000	4.000000	3.000000	3.000000	4.000000	4.000000	3.000
max	60.000000	5.00000	4.000000	4.000000	5.000000	4.000000	4.000000	4.000

	Age	EmpEducationLevel	EmpEnvironmentSatisfaction	EmpJobInvolvement	EmpJobLevel	EmpJobSatisfaction	EmpRelationshipSatisfacti
Age	1.000000	0.207313	0.013814	0.027216	0.509139	-0.002436	0.0497
EmpEducationLevel	0.207313	1.000000	-0.037103	0.027544	0.100734	0.000357	-0.0166
EmpEnvironmentSatisfaction	0.013814	-0.037103	1.000000	0.004865	-0.008272	-0.004319	-0.0105
EmpJobInvolvement	0.027216	0.027544	0.004865	1.000000	-0.034349	-0.005501	0.0180
EmpJobLevel	0.509139	0.100734	-0.008272	-0.034349	1.000000	-0.011853	0.0029
EmpJobSatisfaction	-0.002436	0.000357	-0.004319	-0.005501	-0.011853	1.000000	-0.0220
EmpRelationshipSatisfaction	0.049749	-0.016690	-0.010504	0.018037	0.002992	-0.022028	1.0000
EmpWorkLifeBalance	-0.019563	0.010276	-0.000262	-0.014129	0.049218	-0.018548	-0.0049
PerformanceRating	-0.040164	0.020529	0.395561	-0.010539	-0.076632	0.000606	-0.0195
Age1	-0.055529	-0.001805	-0.042738	0.017483	0.008584	0.045835	0.0238

Fig 3: Statistical Description and Correlation Values of Employee Attrition dataset

F. Graphical Exploratory Data Analysis (GEDA)

Graphical techniques can be used to identify the most important properties of a dataset. GEDA is further classified in to univariate and multivariate based on number of variables considered for analysis and also the type of data.

1. Univariate GEDA

This analysis givesthe statistical summary ofeverycolumn in the given data set. There are many examples for this analysis which include:

- *Histogram:*histogram provides the most intuitive visualizations of any distribution. It is also called as the graphical representation of data organized in to specified range of points. It is like a bar graph where range of data is represented as columns across x axis and y axis represents the respectivedata count for each column[11]. Considering the top 150 employees of the dataset the bar graph for age on x axis and employee performance rating on y axis shown in figure 4.

- *Stem Plot:* Stem plot is a popular statistical tool that helps in graphical exploratory data analysis which separates the digits in data points in to two columns. A stem plot is drawn as all set of y values plotted against common values on x axis. The digit with higher value forms the left column – called stem and the digit with lower value forms theright column – called leafs.The data is ordered in a stem plot.A stem plot helps visualizing the shape of the distribution[7].*matplotlib.pyplot.stem()* function can be used to draw the Stem plot.The age values of Employee dataset represented as stem plot are shown in figure 5. For a total of 1200 employees the lower value starts from 18 and higher value is 60.

Fig 4: Histogram

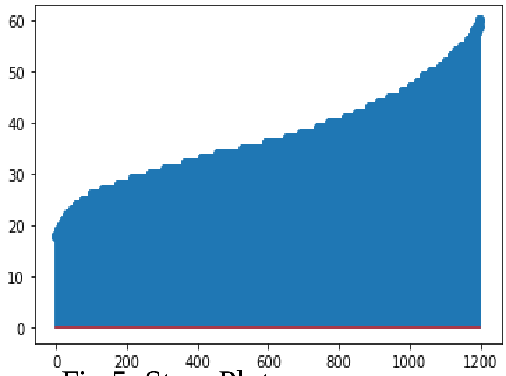

Fig 5: Stem Plot

- *Box Plot:*It is a graphical representation that shows comparison between groups of data. It shows the spread of data its statistical components like central tendency, symmetry, skew and also helps to identify the outliers. The box plot is built using a 5-value summary of the given data set (minimum, Q1, median, Q3,maximum value).These values show the closeness of data values. During the comparison the values which do not fit in to the boundary of the box will become the outliers whose features do not comply with other values in the dataset [12]. The box plot on the Age attribute of Employee data set with the clear outliers can be shown in figure 6.

- *Count Plot:*This represents the frequency or number of occurrences for categorical data using bars using thecountplot() function [5].

```
import seaborn as sns
df = pd.read_csv(io.BytesIO(uploaded['Employee.csv']))
ax=sns.countplot(df.EmpDepartment)
for p in ax.patches:
    ax.annotate('{:.1f}'.format(p.get_height()), (p.get_x()+0.25, p.get_height()+0.01))
plt.show()
```

The count plot on Employee data set showing the different departments in which they work and respective counts is shown in figure 7.

Journal of Science and Technology
ISSN: 2456-5660 Volume 7, Issue 09 (November2022)
www s://doi.org

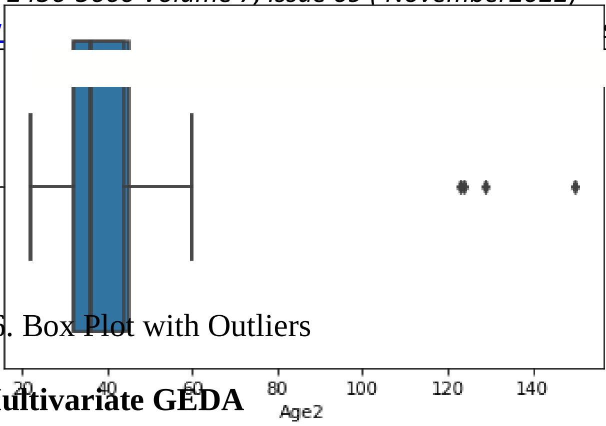

Fig 6. Box Plot with Outliers

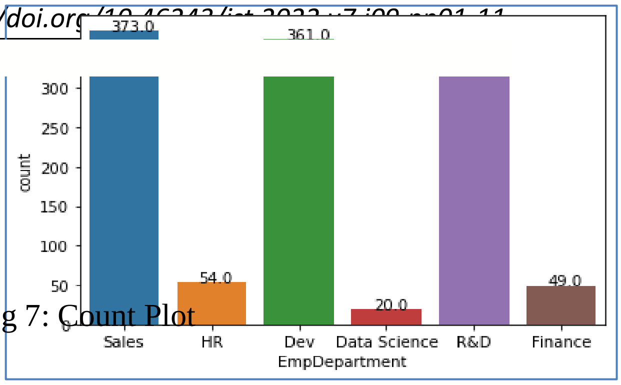

Fig 7: Count Plot

2. Multivariate GEDA

This analysis is used to recognize the associations between different values or variables in the dataset and display the relationship graphically. Some common forms of multivariate graphics include:

- *Scatter Plot:*A scatter plot is a two-dimensional chart showing the comparison of two variables scattered across two axes. The scatter plot is also known as the XY chart as two variables are scattered across X and Y axes. A scatter plot can be displayed without connecting lines or being displayed with smooth curved connectors or connecting lines [12]. For the Employee-Attrition data set the scatter plot between the variables daily and monthly wage of employees can be shown in figure 8.
- *Violin Plot:*This is similar to box and whisker plot. It shows the quantitative data distribution for more than one categorical variableacross different levels.In a box plot, all the components represent the actual datapoints, where as in the violin plot they represent the estimation of thekernel densityfor the given distribution.This is an effective way to show multiple distributions of data at once. For the Employee dataset the violin plot for the performance rating of the employee with respect to their gender is shown in figure 9.

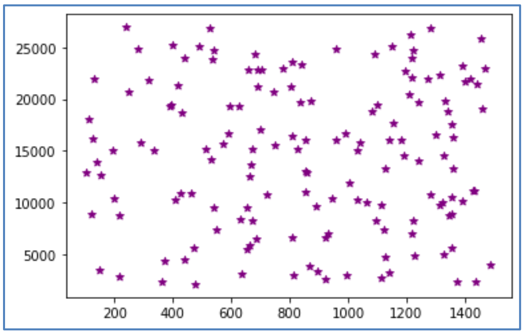

Fig 8: Scatter Plot

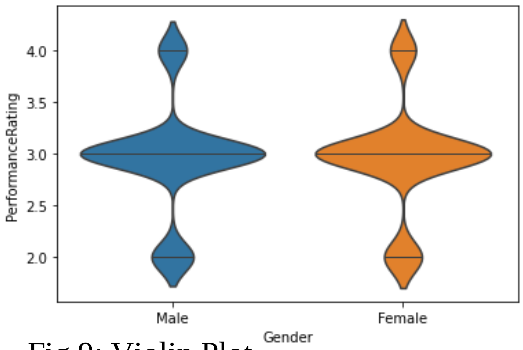

Fig 9: Violin Plot

- *Pair Plot:*This plot shows multiple pairwisebivariaterelationship for (n, 2)variable combinations in a single DataFrame as a matrix of plots where the diagonal plots are the univariate.It is apairwise relationships thatcreate a grid of Axes where each variable shares y-axis across one row and x-axis across one column.

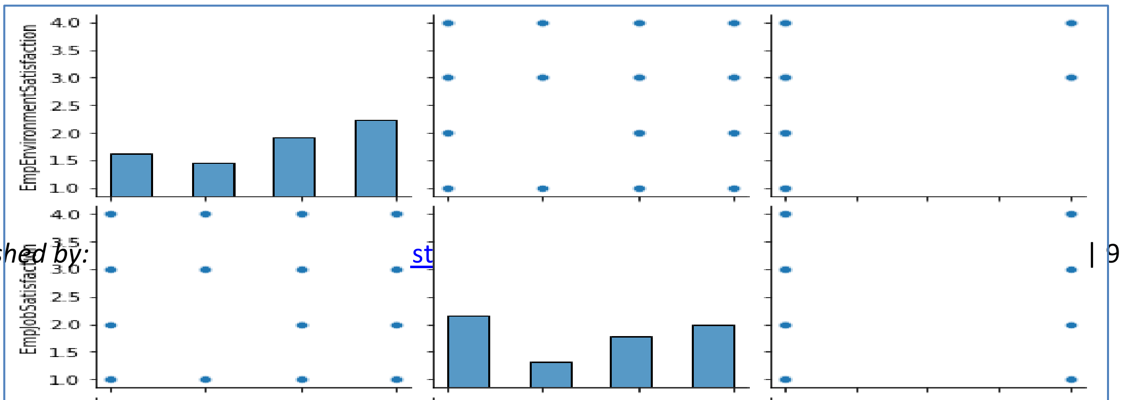

Fig 10: Seaborn Pair Plot.

4. CONCLUSION

In this paper, the different techniques for exploratory data analytics are discussed briefly. It includes both non graphical and graphical methods of analysis where in both univariate and multivariate are also explained. Python is used for implementation purpose importing major libraries and modules necessary for the graphical data analysis. The "Employee Attrition" data set is used in this work and numerous results are extracted and visualized. This work studies the dependence between attributes and effect of one variable on another for employee performance and attrition. Different graphs are been plotted using several attributes in the dataset to show the results in an easy way.

REFERENCES

1. AindrilaGhosh, Mona Nashaat, James Miller, ShaikhQuader, and Chad Marston, "A Comprehensive Review of Tools for Exploratory Analysis of Tabular Industrial Datasets," Visual Informatics, Volume 2, Issue 4, December 2018, pp. 235-253.
2. Matthew Ntow-Gyamfi and Sarah SerwaaBoateng, "Credit Risk and Loan Default among Ghanaian Banks: An Exploratory Study,"Management Science Letters, Vol. 3, 2013, pp.753–762.
3. X. Francis Jency, V. P. Sumathi, Janani Shiva Sri, "An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients," International Journal of Recent Technology and Engineering (IJRTE), Volume-7 Issue-4S, November 2018, pp.176-179.
4. K. UlagaPriya, S. Pushp, K. Kalaivani, A. Sartiha, "Exploratory Analysis on Prediction of Loan Privilege for Customers using Random Forest," International Journal of Engineering & Technology, Vol. 7, Issue 2.21, 2018, pp. 339-341.
5. KiranbalaNongthombam, Deepika Sharma. "Data Analysis using Python".International Journal of Engineering Research & Technology (IJERT),Vol. 10 Issue 07, July-2021, pp. 463-468.
6. KabitaSahoo, Abhaya Kumar Samal, JitendraPramanik, Subhendu Kumar Pani. "Exploratory Data Analysis using Python".International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8, Issue-12, October 2019, pp. 4727-4735.
7. Komorowski, M., Marshall, D.C., Salciccioli, J.D., Crutain, Y. (2016). Exploratory Data Analysis.Chapter 15. In: Secondary Analysis of Electronic Health Records. Springer, Cham.https://doi.org/10.1007/978-3-319-43742-2_15.
8. Steve Midway. "Exploratory Data Analysis – A first look at the data".Chapter 4-Data Analysis in R.
9. Guido Van Rossum et al. Python programming language. In USENIX annual technical conference, 2007.
10. KabitaSahoo, Abhaya Kumar Samal, JitendraPramanik, and Subhendu Kumar Pani. Exploratory data analysis using python.International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2019.

11. Claus O. Wilke. “Fundamentals of DataVisualizationA Primer on Making Informativeand Compelling Figures”. O’Reilly.978-1-492-03108-6. March 2019
12. KalilurRahman. “PythonData VisualizationEssentials Guide”.ISBN: 978-93-91030-063. FIRST EDITION 2021.BPB Publications.