Proceedings of the 6th International Conference on Communication and Electronics Systems (ICCES-2021)
IEEE Xplore Part Number: CFP21AWO-ART; ISBN: 978-0-7381-1405-7

Role of Exploratory Data Analysis in Data Science

Dr. A Suresh Rao[1]
Professor, Computer Science
& Engineering
TKR College of Engineering
& Technology
sureshrao@tkrcet.com

Dr. B. Vishnu Vardhan[2]
Professor & Vice-Principal,
Dept. of CSE
College Of Engineering
Manthani
mailvishnu@jntuh.ac.in

Hafeezuddin Shaik[3]
Assistant Professor,
Dept. of CSE
TKR College of Engineering
& Technology
hafeezuddin@tkrcet.co

2021 6th International Conference on Communication and Electronics Systems (ICCES) | 978-1-6654-3587-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCES51350.2021.9488986

Abstract—**Exploratory Data analysis (EDA) is one of the hidden and mundane tasks in analysis of Data, as a Model, Project or analysis is based on data, which is intuitive, extremely heterogenous and distorted in its form. (Data has become an integral part of every project, Model &) The analyzed data is more insightful for identifying and improving extremely critical business insights across the Organization. This paper defines the exploratory data analysis phenomenon in detail and also states its role in the data analysis domain. Tools & Packages used in EDA are also discussed. In addition, the impact of missing out the phase of EDA in data analysis has been analyzed.**

Keywords— Data Analysis; EDA; Data Visualization;)

I. Introduction

Data intuitive applications and analysis are on steep rise due to advances in technology. It is a well-known fact that we are generating 2.5 e+9 GB of Data per Day. In short, each day we are generating data that is equivalent 1 Decade of Data that we have generated. Exploratory Data analysis is a pre-requisite step developed in the year 1970 by John Tukey. [1]. The name itself indicate that it is a process of analysing the dataset so that it can be fed for deriving insights and building models. It is always a good practise to understand and gather as many insights as possible before using the data for modelling. Exploratory Data Analysis is a essential process of carrying out critical investigation on the Data to Discovery any existing anomalies, missing Data, Patterns etc.

2. Understanding EDA

To expose our self to insights and implications of EDA, dataset of Wine Quality from UCI Machine Learning Repository is used [2].

On analysis of Dataset, the following inferences were made as part of initial analysis.

1. A delimiter separates each attribute in Dataset.

2. Dataset consists of 4,898 Instances with 12 attribute or characteristics. Format

TABLE 1 Understanding the Data

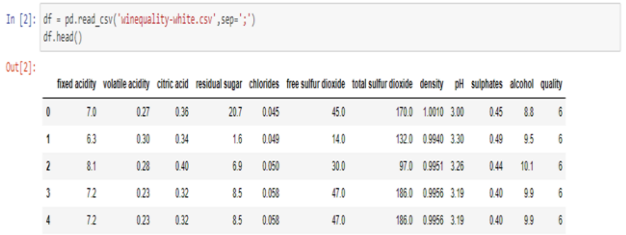
Sl. No	Characteristics of Dataset	Multi-variate
1	Total Number of Instances in Dataset	4898
2	Associated Tasks	Classification & Regression

It is a recommended procedure to first understand the Dataset by analyzing the column’s viz. attributes/characteristics along with their Datatype. This greatly helps us in clearly classifying and categorizing the Data [3] for the next phases. Data is looked across to find any Null Values, Blank/Missing values. Interestingly there are many useful and interesting libraries available in python using which Data can be easily analyzed without much hassle [4]. Seaborn, Numpy, Pandas, Matplotlib are the few to be named among them.

The following functions of pandas are used to draw the inferences from the Dataset.

TABLE 2. Exploratory Functions used for extracting the specific data from the dataset

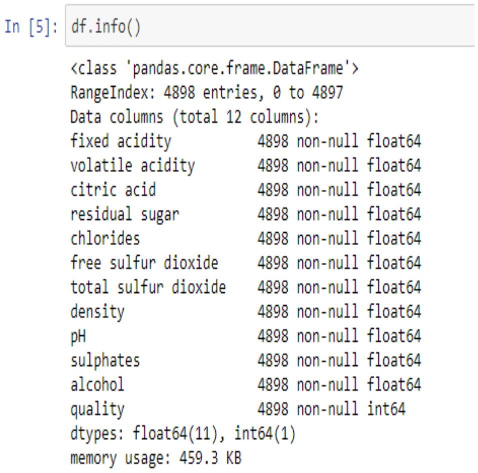
Sl. No	**Characteristics of Dataset**	**Multi-variate**
1	df.head()	Prints the top five instances from the Dataset
2	df.tail()	Prints the bottom five instances from the Dataset
3	df.shape()	Gives the information about the total number of instances and the count of attributes
	df.info ()	Gives the information about the attributes and their Datatypes



```
In [2]: df = pd.read_csv('winequality-white.csv',sep=';')
        df.head()
Out[2]:
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

Fig.1. Loading of Dataset & Reading top 5 Instances using head( ) function.



```
In [5]: df.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 4898 entries, 0 to 4897
        Data columns (total 12 columns):
        fixed acidity           4898 non-null float64
        volatile acidity        4898 non-null float64
        citric acid             4898 non-null float64
        residual sugar          4898 non-null float64
        chlorides               4898 non-null float64
        free sulfur dioxide     4898 non-null float64
        total sulfur dioxide    4898 non-null float64
        density                 4898 non-null float64
        pH                      4898 non-null float64
        sulphates               4898 non-null float64
        alcohol                 4898 non-null float64
        quality                 4898 non-null int64
        dtypes: float64(11), int64(1)
        memory usage: 459.3 KB
```

Fig: 2 Exploring the attributes of Data using data frame info() function

From the above analysis it is noted that
1. All attributes are either integers or floating values.
2. In addition, no attribute has any missing value.
3. Value of median < value of mean

As a part of EDA, we also tend to visualize the data using python. Matplotlib is used to visualize, summarize and generate statistical graphs in multi-dimensions.

These gives a much deeper insights about the data before proceeding to the modelling.

Ignoring or skipping the EDA Process may lead to Modelling of inaccurate Models, Using wrong characteristics with correct model and inefficient usage of computing resources.

3. Steps Involved in EDA:
1. Identification of Variables
2. Uni & Bi-variate Analysis
3. Handling Missing Values
4. Visualization: After cleansing and preparing the data using heatmaps, correlation matrix etc.,

1. Identification of Variables: This phase is all about sorting the variables and its data type for analysis. Input variables and the Output variables as target are identified.

2. Univariate & Bivariate Analysis: This phase helps in identifying the correlation [5] between the attributes. We also try to inspect the frequency of distribution.

3. Handling Missing values: Missing data and null data may produce inaccurate results when used with the model. One commonly used method in handling incomplete and Missing values is deleting such instances. Few more methods involve calculating Mean/Median/Mode whose objective is to fill the data with the mean or median values.

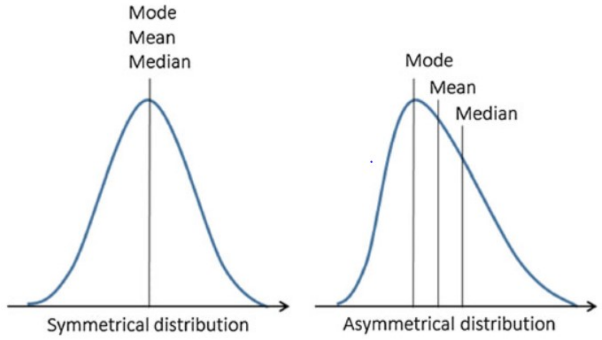
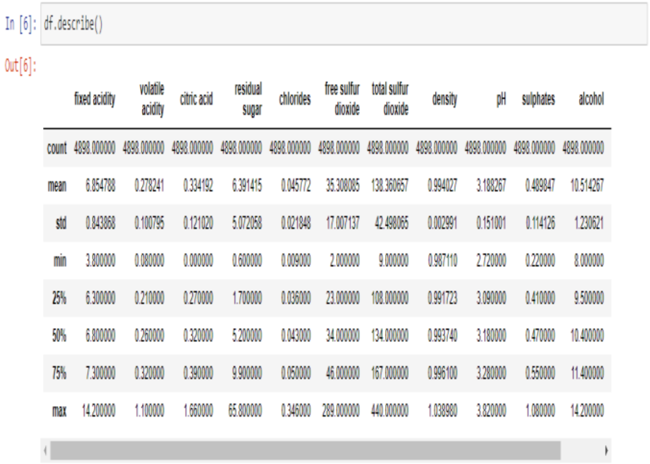


Fig.3 Showing Mean, Median, Model for Sysmmetrical and asymmetrical distribution of Data.

4. Outlier Analysis: In outlier Analysis, abnormal distance from other values in the dataset is observed in the dataset. Such distracting instances [6] are deleted or transformed using natural log function.



```
In [6]: df.describe()
Out[6]:
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267	0.489847	10.514267
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001	0.114126	1.230621
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.220000	8.000000
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991723	3.090000	0.410000	9.500000
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993740	3.180000	0.470000	10.400000
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996100	3.280000	0.550000	11.400000
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000	1.080000	14.200000

Fig: 4 Summary of the data

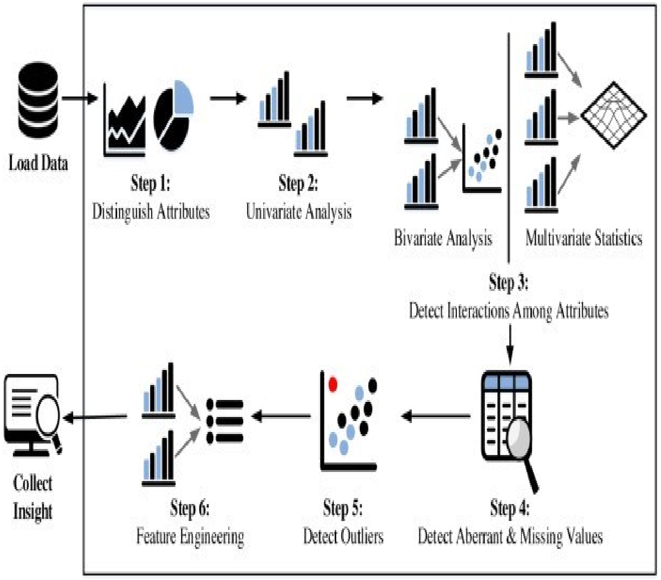


Fig: 4 Different Stages of Exploratory Data Analysis

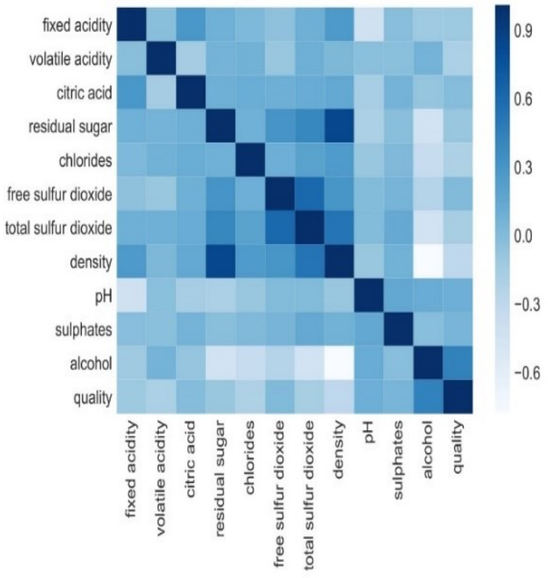


Fig: 5 Heatmap representation for relationship between variables across x and y axis.

Heatmap is a matrix representation of relationships among the given variables that are plotted across a 2 Dimensional graphs x and y axes[7]

In the above heatmap

1. The darker colour shades indicate the positive
   correlation
2. Lighter colour shades represent the negative correlation.

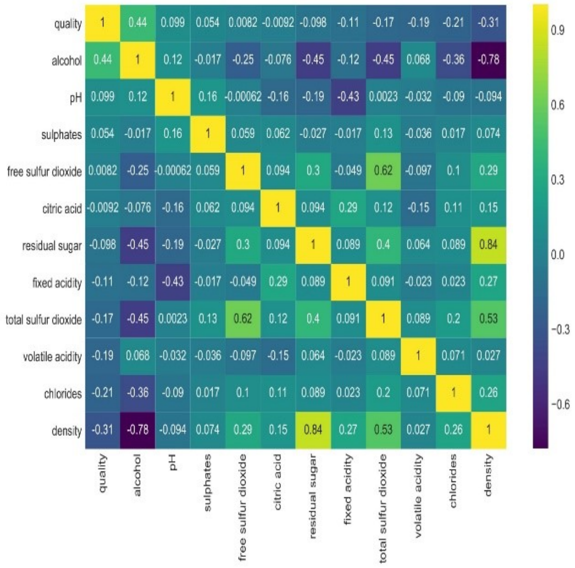


Fig: 6 Correlation Matrix

From the above derived correlation matrix, it can be inferred that there is no linear relationship between Sulphur dioxide and citric acid. In addition, density has an impactful positive correlation with sugar residual.

Benefits of EDA: The primary motive of applying EDA Techniques is to obtain the confidence [8] on the data and saturate to it until it is fit to be fed to a Machine Learning Algorithm.

1. Extraction of usable variables and eliminating the variables, which are unnecessary.

2. Understand the relationship between the variables and attributes.

3. Identification of missing values, Errors caused by Human's.

4. It helps in cleaning of Data and make it meaning and insightful.

Exploratory Data analysis is an practical, definite and operative approach [9] which includes science and statistics. EDA is a important process which is aimed at improvision the accesbility[10] and legitimatness if the Data. It is very well understood that EDA Process helps in identifying hidden artificats, unknown information and knowledge apart from visualizing the data in a presentable form. Visual graphs produced during the process of EDA has an intrinsic impact[11] in identifying the clusters, trends, differences and correlations.

Scientific researchers, scholars undeniable use data analysis rather than opting for Hypothesis. This means we are extracting the solutions from the data. Rather in EDA Process Data is allowed to exhibit the information in a figurative manner. This may include the relations between each attribute, dependencies which lies beneath the data hidden in the raw data. Tukey[12], in his work, meant to make an information examination structure, where the visual assessment of informational indexes, through measurably critical representations. As assessed, the required tools for EDA are purely based on graphical approach. Historgrams, box-plots, Distribution charts are extremely helpful in probing the Data.

The other important aspect of Analysing the Data in exploratory mode is, it helps in easing the process as it is very flexible, forgivable and higher degree of ease of computation[13]. EDA recognizes the fact that researchers and statisticans may choose regression, hierachical classificATION or principal component analysis based on the requirement as analysis of Data is a much more creative and exploratory job.

EDA can't genuinely be concentrated in separation of its Contexts, situations in particular. Its job in whole process is it continues through human cognitive cycles cycles[14]. Anscombe's[15] shows the significance of EDA as a fundamental piece of information investigation. Perception is, it is especially significant in understanding examples in large data sets.

CONCLUSION: The surveyed literatures evident that process adapted by the data science professionals, in analyzing the need of hour tend to ignore the EDA Process to stay ahead with the timelines. Hence, from the above inferences, it is evidently proven that EDA can have significant impact in modelling algorithms, ignoring the same may lead to inaccurate predictions and re-working. As a resultant increase in the overhead cost and resource utilisation is observed on whole.

Acknowledgment

The authors would like to thank Family, Friends, Collegues and TKR College of Engineering and Technology for constant support. The physical and technical contribution of JNTUH is truly appreciated. Without their support and funding, this project could not have reached its goal.

References

[1] Hoaglin, David. (2017). Exploratory Data Analysis. DOI. 10.1002/9781118445112.stat00419.pub2.

[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

[3] Kennedy, Ryan & Waggoner, Philip. (2021). Exploratory Data Analysis. 10.1201/9781003030669-6.

[4] Rajagopalan, Gayathri. (2021). A Python Data Analyst’s Toolkit: Learn Python and Python-based Libraries with Applications in Data Analysis and Statistics. 10.1007/978-1-4842-6399-0.

[5] Cozzolino D, Cynkar WU, Shah N, Dambergs RG, Smith PA. A brief introduction to multivariate methods in grape and wine analysis. International Journal of Wine Research. 2009;1:123-130 https://doi.org/10.2147/IJWR.S4585

[6] 6. Gehlenborg N, Wong B (2012) Points of view: heat maps. Nat Methods 9(3):213

[7] Aggarwal, C.C. & Yu, P.S.(2001). Outlier Detection for High Dimensional Data. Proceedings of the ACM SIGMOD Conference 2001.

[8] Andrew T. Jebb, Scott Parrigon, Sang Eun Woo, Exploratory data analysis as a foundation of inductive research, Human Resource Management Review, Volume 27, Issue 2, 2017, Pages 265-276, ISSN 1053-4822, https://doi.org/10.1016/j.hrmr.2016.08.003.

[9] Mario Li Vigni, Caterina Durante, Marina Cocchi,

Chapter 3 - Exploratory Data Analysis, Editor(s): Federico Marini, Data Handling in Science and Technology, Elsevier,

Volume 28, 2013, Pages 55-126, ISSN 0922-3487, ISBN 9780444595287, https://doi.org/10.1016/B978-0-444-59528-7.00003-X.

[10] Komorowski M., Marshall D.C., Salciccioli J.D., Crutain Y. (2016) Exploratory Data Analysis. In: Secondary Analysis of Electronic Health Records. Springer, Cham. https://doi.org/10.1007/978-3-319-43742-2_15

[11] Camizuli, Estelle & Carranza, Emmanuel John. (2018). Exploratory Data Analysis (EDA). 1-7. 10.1002/9781119188230.saseas0271.

[12] Turkey JW. Exploratory data analysis. Lebanon, IN, USA: Addison-Wesley;1977.

[13] Morgenthaler, Stephan. (2009). Exploratory data analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 1. 33 - 44. 10.1002/wics.2.

[14] Mast, Jeroen & Kemper, Benjamin. (2009). Principles of Exploratory Data Analysis in Problem Solving: What Can We Learn from a Well-Known Case?. Quality Engineering. 21. 366-375. 10.1080/08982110903188276.

[15] Anscombe FJ (1973) Graphs in statistical analysis. Am Stat 27(1):17–21. doi:10.2307/2682899