International Journal of Data Science and Analysis
2021; 7(3): 89-97
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20210703.16
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Separation of Data Cleansing Concept from EDA

Khanjan Purohit

Data Science and Analytics, Jain University, Bangalore, India

Email address:
khnjn.purohit@gmail.com

To cite this article:
Khanjan Purohit. Separation of Data Cleansing Concept from EDA. *International Journal of Data Science and Analysis*. Vol. 7, No. 3, 2021, pp. 89-97. doi: 10.11648/j.ijdsa.20210703.16

Received: May 25, 2021; **Accepted**: June 8, 2021; **Published**: June 22, 2021

Abstract: Available dataset whether it is structured, semi structured or unstructured data, is used for various purposes. These data sets are mostly used for solving an issue using different kinds of techniques like visualization, descriptive, algorithms etc. This data process includes many levels, two of those steps are exploratory data analysis (EDA) and data cleansing. Data cleansing and exploratory data analysis (EDA) are two major operations of any data mining or machine learning study. After collecting the data from various sources, Data cleansing is done to make the data set more accurate, useful and less redundant. Data cleansing is useful to get the accurate information from the dataset and It is used to deal with null values, duplicate values, multiple values, inconsistent value, inaccurate value etc, Which are existing in that data set and It can make our data set filled with error which also affects the analysis and decision making process. By performing data cleansing, we can get rid of many types of misleadings like getting inaccurate output, inaccurate model of machine learning, not getting the hidden patterns behind that data set etc. The purpose of this paper is to study existing research of Data cleansing and EDA and state why Data cleansing process is not part of exploratory data analysis (EDA).

Keywords: Data Cleansing, Exploratory Data Analysis (EDA), Data Mining, Normalization, Visualization, Big Data

1. Introduction

Data is increasing swiftly and has become very difficult to gather and operate the data. To analyze the data, get the information and make a decision from the concluded information, one has to go through a big process. There are many tools and programming languages available in the market. Some of them are free and easy to use such as R programming, Python, SaS etc. These are the languages which are very useful because of their english commands and easy to use syntax [6].

Making data useful for visualization, model implementation and decision making, need to perform data cleansing. Data cleansing is a technique of abolishing and fixing the significant problems and outliers present there in the data set [9]. Usually data scientists get the data from various sources such as internet, organization and they use several kinds of programming languages for data cleansing [14]. Data cleansing is an essential step in the process of an analysis project. The issue of dirty data can make an impact in other departments of a firm therefore we use data cleansing [15]. There are various kinds of errors present there in the data set and making it dirty data such as null / incomplete values, duplicate values, inconsistent values, multiple values, inaccurate values etc. Many approaches used in solving these errors are discussed further in the paper.

Exploratory Data Analysis (EDA) is an operation which is performed to gain the information from the cleaned and errorless data set. Exploratory data analysis includes various kinds of approaches to construe information such as descriptive statistics, visualization techniques, correlation between the variables etc. If EDA is performed without cleaning the data, Gained information can be wrongly interpreted and It can cause a decision making process as well [6].

2. Problem Statement

Because of increasing data in uncountable amounts, Firms / Businesses are totally dependent on data. Huge amount of data is gathered within seconds. It is easy for the organizations to make better decisions by using that data. Data with redundancy are always impactful in interpreting useful information and decision making further. So data cleansing is performed to make the data set less redundant and more accurate for the purposes of businesses [8]. EDA is

an operation performed for extracting the beneficial information from the data set using several techniques [2].

According to Ronald D. Snee [11], EDA is used to explore data and better understand the process that generated data but data cleansing is a separate process than EDA, so data cleansing is performed before EDA and for EDA. Data cleansing process is not included in EDA. Data cleansing is about dealing with the challenges present there in the data and making the data set errorless for the EDA and decision making. Methods which are used for data cleansing has no connection with the techniques which are used for performing EDA.

3. Data Cleansing Problems

In comprehensive data cleansing, [9] syntactical anomalies such as lexical error, domain format errors, irregularities etc. are considered as errors for data format and values for representation. Lexical error is the difference in the framework of the values. Domain format error, values do not fit in expected format G (*dom* (A)) of an attribute A. Irregularities is when the observable value is entered with a different format which is not a regular format of an attribute.

About the data which are organized in tabular / structured form, time series and big data, Challenges like incompatible data format, non-aligned data structure, data which are not consistent can affect the analyzed outcomes. The better correlation between the attributes and the accuracy and relevancy of data can be beneficial for the business from their competitors [3].

Duplicate values are the values which are present multiple times which should be existing only once. Duplicate values can occur in several ways while gathering data from different sources such as spelling mistakes, difference in formats, inconsistency in formats etc. Duplicate record deduction and data redundancy are the most essential and important concepts of data mining and data integration [1].

Missing values are the values which are not entered while collecting the data. These are the values which are omitted while gathering the data. In the place of missing values, the NOT NULL constraint exists [5].

4. Methodology: Data Cleansing

Many techniques are used to improve the data quality. And there are several algorithms which are used for the data cleansing purpose. Various kind of approaches are also used for data cleansing.

4.1. Techniques

Parsing is used for detecting the syntactical errors present there in data set while performing data cleansing. We can correct these errors by edit distance function by choosing possible correction with minimal edit distance [5].

Data transformation is a process for mapping from the given format into the format expected. To make these values correct, Normalization and standardization techniques are used [5].

Integrity constraint enforcement asserts the regularity of integrity constraints after making changes in the transaction of collecting datasets happen. By maintaining the integrity constraints, We can make sure that errors cannot violate the integrity constraint after the additional updates in data collection [5].

Duplicate elimination methods can be used for detection of duplicate values and removing them from the data set. Various algorithms are used for determining whether two or more tuples present there are of the same entity [5].

Statistical methods like analyzing data by mean values, standard deviation, range and clustering algorithms. Complex errors which are not uncovered by integrity constraints can be solved by statistical methods [5].

4.2. Algorithms

For the detection of duplicate values in the collected data, There are three types of algorithms used i.e probabilistic algorithms are about probability and statistical methods like data clustering, bayesian networks and expectation maximization etc. knowledge based algorithms are about training and the use of training for detection. and empirical algorithms are elaborated below [1].

4.2.1. Blocking

By applying the sorting key to each tuple of the data set, We can have each tuple in each partition and duplicate tuples are assigned to their different partitions so no more than one tuple is present in the partition. We can compare those values and find the duplicate values [1].

4.2.2. Windowing

Merge two data sets and sort in lexicon order according to the primary key. By comparing values in a fixed size slide window, the first record will be released to select the next record [1].

Blocking and windowing algorithms are used for performing record comparison reduction. First, Tuples are sorted and assumed that after sorting tuples are close to each other in both the algorithms. Selection procedure of tuples for comparison is different. Tuples are blocked in disjoint partitions in blocking algorithms where in windowing algorithms it is used by sliding a window over the tuples [1].

4.3. Current Approaches

There are several current approaches used for detecting the errors and fixing the errors in data cleansing operation. In this paper, most popular approaches are mentioned with their advantages and weaknesses. These approaches are,

AJAX is an extension as well as modified structure which is trying to segregate the logical and physical level of data cleansing. It's used for the transformation of the collected data from multiple data collections into the target schema and deducting duplicate values in AJAX process. It is beneficial for customization because of its extensible idiosyncrasy. But, the approach of data cleansing is very complex and time consuming [1].

Potter's wheel is a system which integrates data

transformation and error detection. Transforms are made for supporting the similar schema transformation which is done by not performing any kind of programming. These transformations are assigned to each value present there in columns. The advantage of Potter's wheel is, By building a transformation we can detect discrepancies through graphical operations or through examples. The disadvantage is, Suppose if programmers do not scroll down to that value though there can be differences, It will not be detected [1].

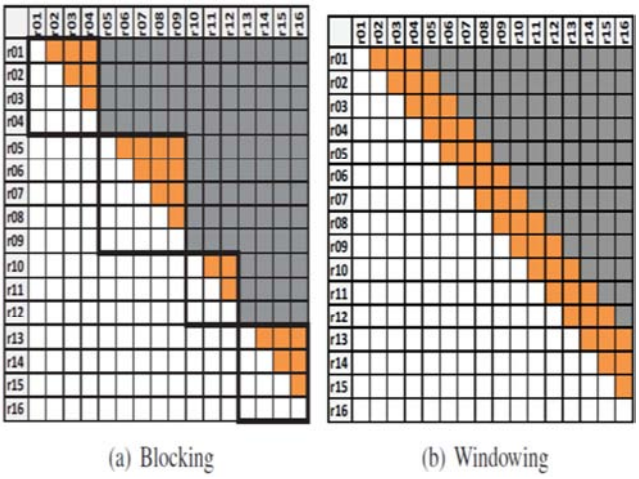

Figure 1. *Selection of elements for comparison in Windowing and Blocking [1].*

Intelliclean hires a new procedure for computation of expiring transformation of values inaccurately and it controls dealing with the exact duplicate records. The best factor is, Intelliclean framework provides effectual methods for debugging the problem of accurate repetition which is very frequent in other structures. Worst quality about intelliclean is, Data source from web is not compatible with this system i.e images, videos, audios and other file format are not compatible with intelliclean [1].

The differences between these approaches are explained regarding many factors like Porter's wheel is very interactive so easy to use where AJAX is very complex and hard to use for non-technical people and Intelliclean is interactive to the end user but It needs a little bit input by end user. Talking about human dependency, Porter's wheel requires too much user involvement for unusual errors. AJAX also requires user involvement i.e calculation and validation of outliers or inaccurate values are totally based on user experts where human dependency is very less in Intelliclean because of the special module implanted in it [8].

4.4. Normalization

Normalization is a technique used for making the dataset less redundant. It is used to analyze the relational schemas which are based on primary key constraint and functional dependencies are used for making a dataset less redundant. In the development of any software system, Normalization process is performed to transform the dataset and make the dataset errorless. Normalization is a technique that comes under one of the most used processes of data cleansing.

Normalization clarifies the connections and deducts a problem of outliers which otherwise might happen while manipulating a relation in the relational database. There are three initial and essential steps performed for the normalization which are elaborated below:

4.4.1. First Normal Form (1NF)

First normal form or 1NF is a normalization process to insure that each cell of the table consist only one value. It should not have composite or atomic values. If there is any NULL values present there in the dataset than it would be removed and replaced by any other corresponding data type values. Figure 2 is a flow chart of first normal form (1NF) [4].

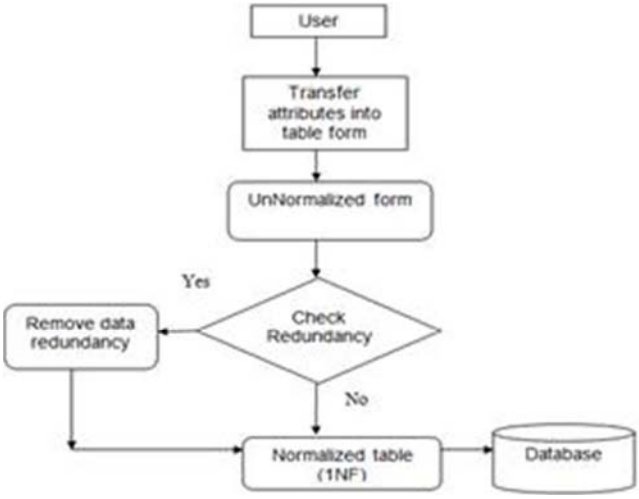

Figure 2. *Flow Chart of First Normal Form (1NF) [4].*

4.4.2. Second Normal Form (2NF)

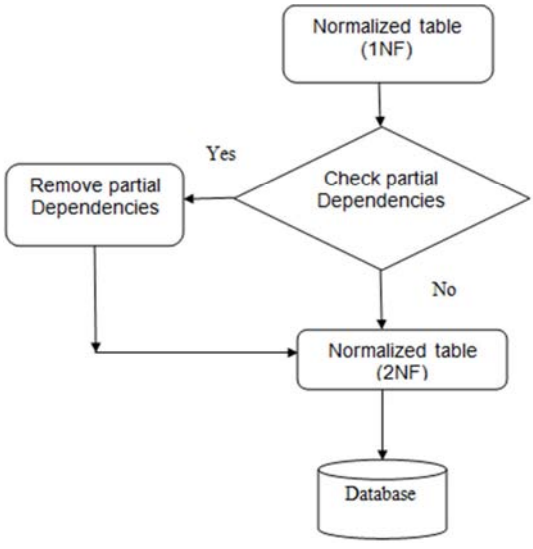

Figure 3. *Flow Chart of 2NF [4].*

Second normal form or 2NF is a next step performed in the process of normalization. 2NF is performed to satisfy the functional dependency. Functional dependency means each value of the dataset has to be integrated to the primary key or candidate key of the dataset. If a value is not integrated or

grouped with a part of primary key or candidate key then it is called partial functional dependency. Below diagram is a flow chart of the 2NF [4].

4.4.3. Third Normal Form (3NF)

Third normal form or 3NF is a part of normalization process to make sure that the transitive dependencies are not there in the dataset by taking out the independent variables of the dataset. To perform the third normal form, we need to have our dataset in second normal form. Figure 4 is a flow chart of the third normal form.

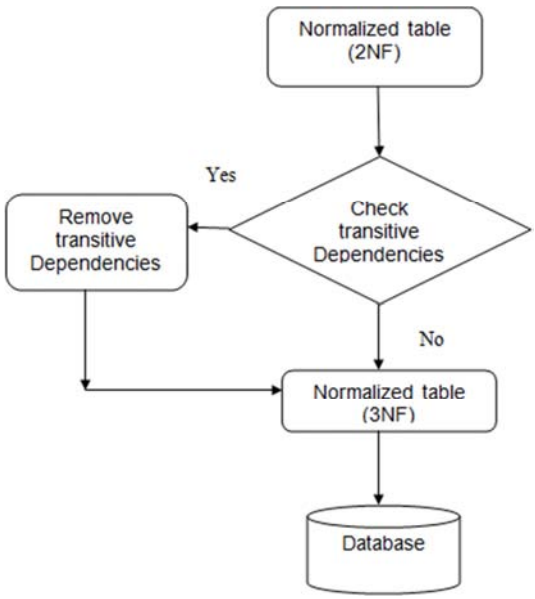

Figure 4. *Flow Chart of Third Normal Form [4].*

5. Data Cleansing for Big Data

Big data is a high level data which is extreme in size, rapidity, diversity of data source or various kinds of data which makes analysis more complex by doing it in statistical methods and computation. It is almost impossible to delete errors present there in the big data and it becomes even more hard when its metrics of distinguishing between dirty data and original data is lacking [13]. The goal of big data analysis is to give support to make real-time decision or to detect the hidden patterns which can be evaluated to improve the outcomes. Challenges faced in big data are problems of data structure, security, data standardization, storage and transfers and skills of managing like data governance [7].

[3] About data cleansing in big data, there is a system called Cleanix which is used to tackle the data quality issues such as outliers detection, imputing NULL values, solution of duplicate values. This is evolved by usability, unification, scalability which authorize Cleanix for reporting data quality and data cleansing task along. It connects several automated data cleansing task into single dataflow which has four stages,

i. Read data, detect and correct data
ii. Fill missing data
iii. Broadcast the updated value in local gram
iv. Solve deduplication and solve resolution

6. Exploratory Data Analysis (Eda)

Exploratory Data Analysis (EDA) is a popular technique or an operation which is a compulsion in any kind of analytical project done by using already collected data sets. EDA is used for many purposes like to find hidden patterns between two variables present there in data set, for conducting a hypothesis by using least possible structure. EDA is performed by using several techniques like descriptive statistics, visualization techniques, extraction of information, interpretation and then the decision making process is performed.

Exploratory data analysis (EDA) can be used for betraying certain statistical properties present there in the data set. EDA is used to establish a specific answer or collection of answers opposite to the noticed data [12]. Exploratory data analysis (EDA) reduces assumptions which are made and suggests for the selection of model for further examinations. According to the Tukey (1977), He suggests to adopt the five number summaries which is about counting and sorting in descriptive statistics for summarizing the data set [2].

Exploratory Data Analysis (EDA) is a technique used for the summarization of the data by by considering data sets' main characteristics and representing it. EDA considers more narrowly on checking the presumptions needed to deal with the NULL / missing values, to fit the model, hypothesis testing, and making changes in attributes. Data analytics address organizations to understand their efficiency and performance and eventually helps business to make more informed decisions [6].

6.1. Eda Techniques

For performing the process of EDA, There are many techniques used and they are efficient and useful as well. The available data set has to be normalized or errorless so the output of EDA process is the most accurate and can be utilized further for decision making process. If the data set is not accurate and filled with lots amount of errors, it might impact the output of EDA and decision making process.

In this paper, the steps performed for ML and data analysis and visualization techniques are explained first. The steps and techniques are,

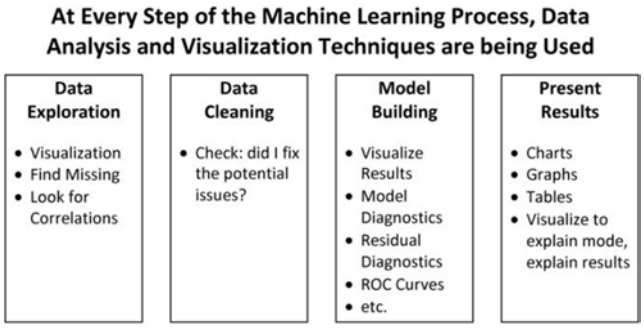

Figure 5. *The steps for machine learning process [6].*

Data Exploration

In the first step of analyzing the data set, We can get the

information about the attributes / characteristics of the data set. It represents the size of the data set. It shows the hidden patterns between the variables in the data set. Understanding of characteristics and tabular data are essentially required for the data visualization [6].

Data Cleansing

Data cleansing is the operation to indicate the erroneous data and corrupt data. Unimportant elements of the data set has to be removed and It should be replaced by corrected data. Validation of the data set is done in this process of data cleansing. Data has to be verified for eliminating the corruption and present issues may be solved by checking the data [6].

Model Building

There are various machine learning models like linear regression, decision trees, kNN clustering, support vector machine (SVM) etc. They all are used for different purposes and for different kind of data sets and each one of them belongs to whether supervised learning or unsupervised learning. We have to visualize the model before the evaluation of the model [6].

Present Result:

Charts, graphs and tables are used for the visualization of the large amount of the complex data set. We can process data using charts and other plots and It's also a simple way to convey the meaning. It can be used to know the areas where improvization is needed and It can clarify the factors as well [6].

There are two types of exploratory data analysis. First one is non-graphical exploratory data analysis and another one is graphical exploratory data analysis. Each one of them are further divided into univariate exploratory data analysis and multivariate exploratory data analysis.

6.2. Non-Graphical Exploratory Data Analysis

Following non-graphical methods gives the detailed information of the characteristics of the data set and the distribution of the variable (s). The types of non-graphical exploratory data analysis are,

Univariate Non-Graphical EDA –

There are several methods used for the univariate non-graphical exploratory data analysis which are,

Tabulation of Categorical Data –

Tabulation technique for categorical variable is about creating a table of the number of data and frequency of data of each variables. An example of tabulation method is shown below [10].

Table 1. *Tabulation of Categorical Data.*

	Group Count	Frequency
Green Ball	15	75
Red Ball	5	25
Total	20	100

Characteristics of Quantitative Data

These characteristics are considered as the predictions of the related population parameters from where the sample is taken. It expresses central tendency of the data like mean, median, mode, variances, standard deviation, interquartile range, maximum and minimum value etc. It also represents some measures of its distribution like skewness, kurtosis etc. These characteristics are only applicable for quantitative data, not categorical data. Some of these characteristics can be represented on histogram as shown below [10].

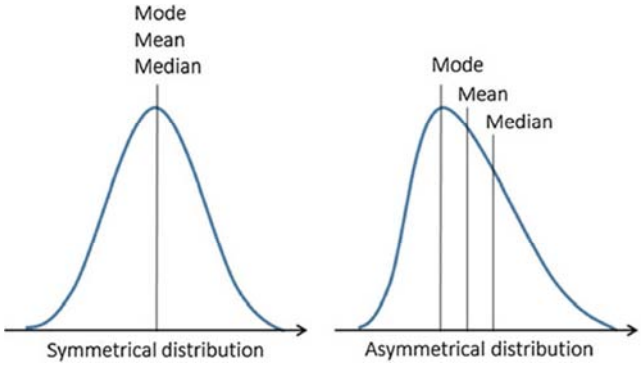

Figure 6. *Symmetrical versus Asymmetrical / Skewed Distribution Showing Mode, Mean and Median [10].*

Central Tendency Parameters

The Arithmetic mean or mean is the value calculated by the total (sum) of the all data and divided by the number of values. Median is the middle value in a list of containing all the values sorted and median values are affected little by extreme values or outliers [10].

Variance

In rare cases, we calculate on the entire data set. The variance (σ^2) is calculated by dividing the total of squares by total number of values (n). The formula for the variance has denominator as n-1 instead of n just because to make it unbiased which means when we calculate for calculate for random samples belonging the same population, The mean has to be matched to the corresponding population quantity. The square root of the variance is called standard deviation (s) therefore it has same values as the actual data which is useful for making it very interpretable. s^2 is an unbiased predictor of population variance σ^2 [10].

$$S^2=\sum (X - X) / (n - 1)$$

Interquartile Range (IQR)

The IQR is evaluated by the ranges of the data set which are established between the 1st quartile and 3rd quartile. IQR is very robust spread measure than variance and standard deviation therefore it should be effective for small and asymmetrical distributions.

$$IQR=Q3 - Q1$$

Skewness / Kurtosis

Skewness is a feature of the asymmetry of an ideally symmetric probability distribution and it is given by the third standardized moment. There are two types of skewness, positive skewness and negative skewness. Lightness or heaviness in the tails which means that data seems flatter

compared to the normal distribution. There are two types of kurtosis, positive kurtosis shows that we have heavy tails and negative kurtosis shows that we have light tails.

Finding the Outliers

Outliers are the ones which are completely odds in the data set. There are many outliers detection techniques used in the EDA to deal with these outliers like Tukey's method, Z-score Studentized Residuals etc [10].

Multivariate Non-Graphical EDA –

Multivariate non-graphical EDA methods generally represent the connection between two or more attributes with each other. For each combination one variable is continuous variable and another one is categorical variable. Continuous variable is usually the outcome and categorical variable is the explanatory. Methods for multivariate non-graphical EDA are explained below.

Cross Tabulation

Cross tabulation is a fundamental non-graphical EDA method and it is an extended version of tabulation which deals with only categorical and quantitative data with only few variables. To create a two-way table with columns headings matching the levels of one variable and row headings matching the level of other variable and then filled the numbers of all subjects that share a pair of levels [10].

Covariance and Correlation

Covariance and correlation are the measures of the degree of the connection between two random variables and expresses how much they modify together. The covariance is calculated as below,

$$\text{Cov}(x, y)=[\sum (Xi - X)*(Yi – \bar{Y})] / N-1$$

Where, X and Y are attributes, N is the count of values present there in the population. X□ is the mean (average) value of variable X and is the average value of the variable Y. A covariance which is positive means that the variables are positively related and they move together in the same direction, where a covariance which is negative means the variables are inversely related. An issue with covariance is that it is dependent on the scale of the values of the random variables. The higher the values of x and y, the higher the covariance. It makes it impossible for the comparison from the data sets with different scales. It can be solved by dividing the covariance by the multiplication of standard deviation of each random variable and that is called as Pearson's correlation coefficient.

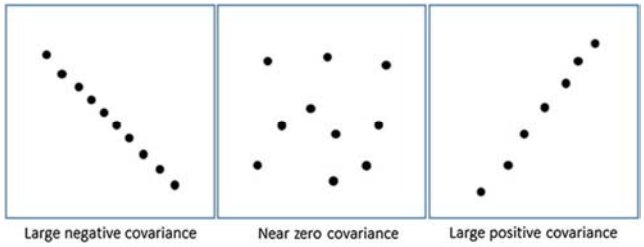

Figure 7. *Examples of covariance of three different data sets [10].*

Correlation is a scaled version of covariance and It is used to assert the linear conection between two attributes and it is evaluated using the formula shown below:

$$\text{Cor}(x, y) \text{ or } r=\text{Cov}(x, y) / Sx * Sy$$

Where Cov (x, y) is the covariance between x and y and Sx and Sy are the sample standard deviations of x and y. Fisher's z transformation is used to evaluate the accuracy of the correlation coefficient between both normally distributed variables.

6.3. Graphical Exploratory Data Analysis

Graphical techniques are fundamental of EDA. They are mostly used for exploratory data analysis operation. Various graphs and plots are used for different purposes. Graphical EDA helps to visualize the statistical characteristics of the data set with the help of several graphs and charts. The techniques are categorized. The first category is univariate graphical EDA, The second one is bivariate graphical EDA and the other one is multivariate graphical EDA. The techniques which are used for graphical EDA are as following.

Univariate Graphical EDA –

Univariate graphical EDA gives the statistical summary of each attribute in the raw data set or gives the summary on one variable. Example of these types of EDA are cumulative distribution function (CDF), probability density function (PDF) and the other charts and graphs. Some of them are discussed below [6].

Histogram

A histogram is about summarizing discrete or continuous data. In other words, it provides a visual representation of continuous data by representing the number of values that belong to a particular range of values (called “bins”). It is the same to a vertical bar graph. However, a histogram, not like a vertical bar graph, represents no gaps between the bars.

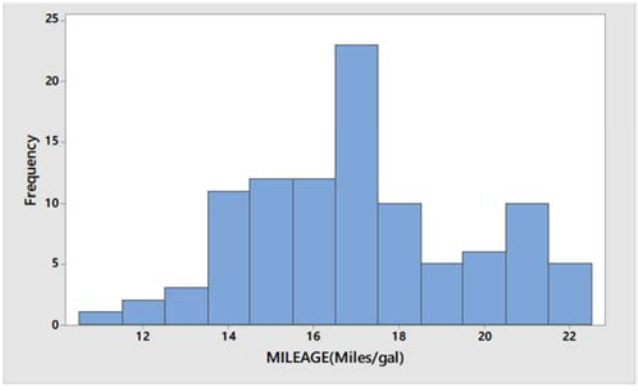

Figure 8. *Histogram of Gasoline Mileage Data [11].*

The Histogram of the gasoline data (Figure 9) shows that the MPG varies from approximately 11 – 22 with a centre location around 16 – 17. The shape of the distribution appears to have two modes (humps) around 16 and 21. The gasoline mileage data set is taken from the reference paper [11].

Stem Plots

The stem plots splits a data set into two sections. The first digit shows the stems and the last digit shows the leaves. Stem plots are mostly used for the comparison purpose.

Box Plots

By using the box plots, we can represent a better graphical image of the characteristics and measures of the data. It represents the central tendency, summary, skew and outliers. The minimum value, maximum value, median value, first quartile and third quartile are used to construct a box plot. These values are compared to represent how closer the other values are to the measures.

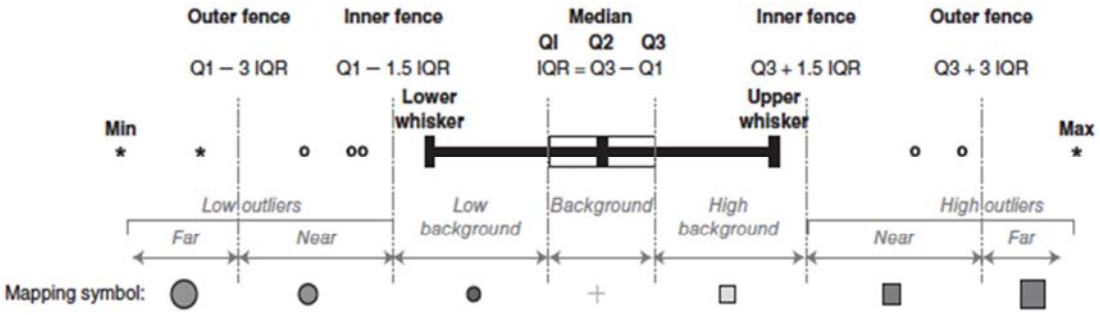

Figure 9. *Tukey box plot used to describe a dataset with five-number summaries (Min, Q1, Q2, Q3 and Max) [2].*

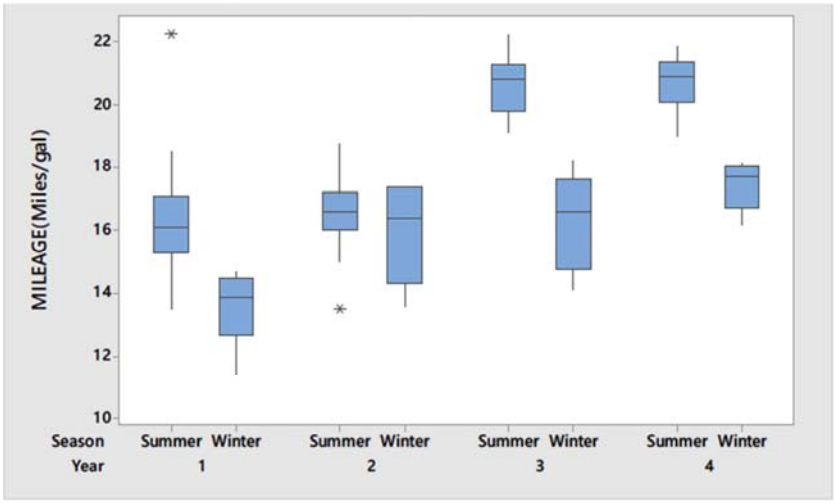

Figure 10. *Example of Box Plot Using Gasoline Mileage Data [11].*

2D – Line Plot

Line plots are the types of plots that represent as a series of points called markers which are joined through a straight line segment.

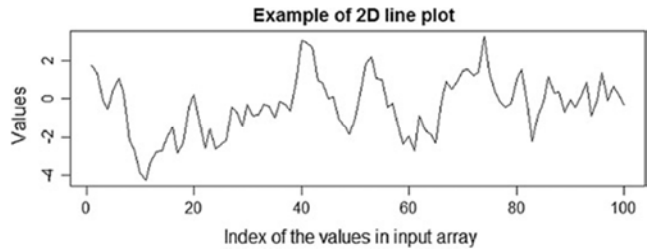

Figure 11. *An Example of 2D Line Plot [10].*

Bivariate Graphical EDA

Bivariate graphical EDA is performed to know the relation of each attribute in the data set and the target attribute or with two attributes and exploring relation amongst them. Violin plots are used for bivariate graphical EDA. It is very analogous to a box plot but It is an extension of a rotated kernel density plot on each side.

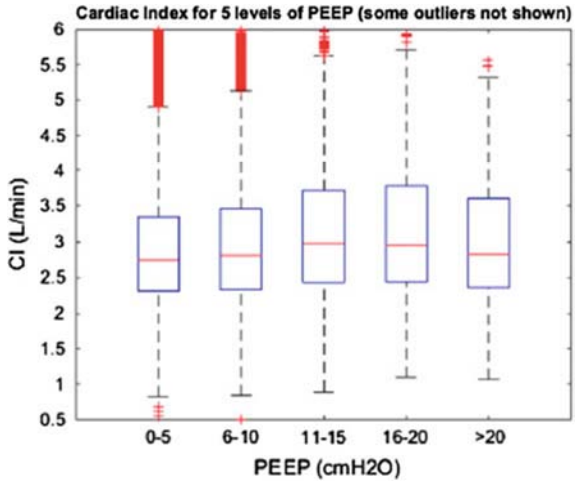

Figure 12. *An Example of Side-by-side Box Plot [10].*

Multivariate Graphical EDA

Multivariate graphical EDA is very useful to know the connection between different variables in the dataset or

exploring the relation between two or more variables. The graphs and charts used for multivariate graphical EDA are explained below.

Bar Graph

Bar graphs are the most used visualization plot for showing the relationship between two or more variables and It is the easiest to understand and interpret the information.

Side-by-side Boxplot

Side-by-side box plot is used for comparing the levels of all possible values. It is used to compare two datasets. It summarizes the data for each instant of categorical variable.

Heat Maps and 3D Surface Plots

Heat maps are generated by taking the entire feature variable. Feature variables are considered as row and column header and the attributes opposite itself on the diagonal. It is used for representing the connection between two variables in high dimensional space. Heat maps are a 2D grid and made from a 2D array, whose colour is dependent on the value of each cell's value. The dataset which is being used, has to be corresponded to a 2D array whose cell consists the value of the outcome attributes.

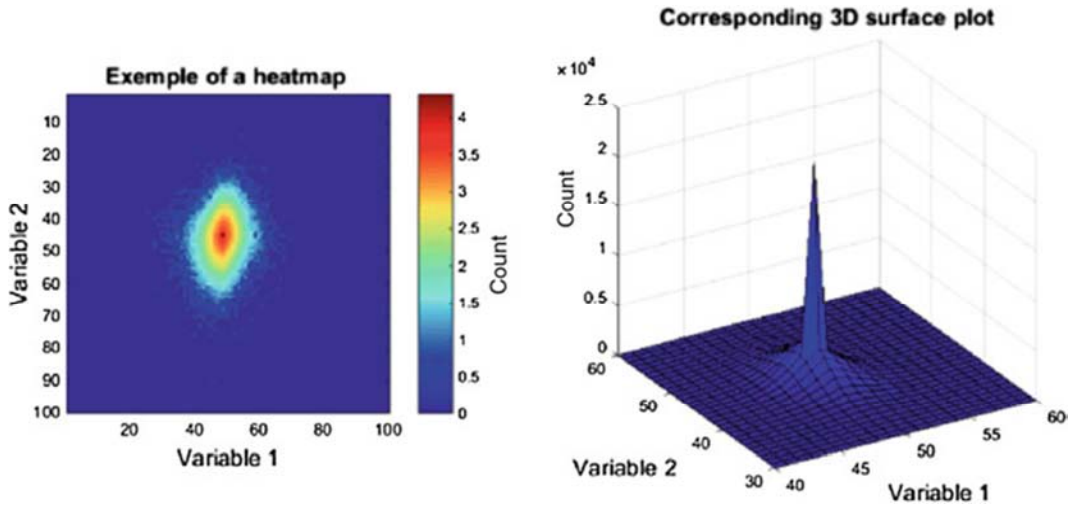

Figure 13. *An Example of Heat Map and 3D Surface Plots [10].*

Scatter Plot

Scatter plot is a plot which is represented on a Cartesian coordinate to show the data points between two variables of a dataset. It is constructed by two continuous, ordinal and discrete quantitative variables. It can be constructed by taking the variable value in X-axis and Y-axis.

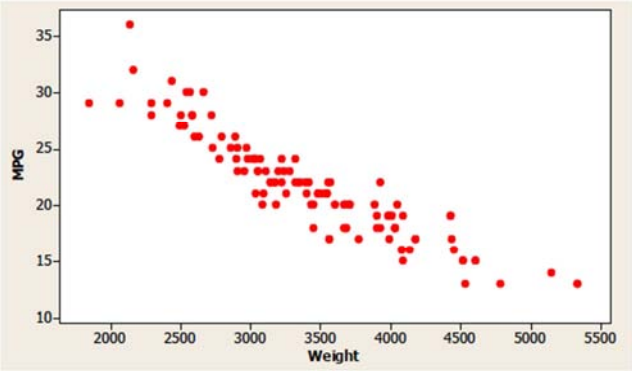

Figure 14. *An Example of Scatter Plot Using Gasoline Mileage Data [11].*

7. Discussion

This study aims to state that data cleansing concept is not included in exploratory data analysis (EDA) concept. They both are different concepts and used for different purposes. From the details explained in details in previous sections, It is clear that the data cleansing concept is not a part of exploratory data analysis (EDA), It is a totally different process. Data cleansing is an essential technique to make the dataset errorless. The errors such as missing values, duplicate values, composite values and other errors might be there in the dataset which is going to affect exploratory data analysis in visualization, statistical information and in the decision making process ultimately. These kinds of problems have to be solved and made a dataset in normalization form. Exploratory data analysis is another equally essential process for any data science project, machine learning task, decision making process. Various types of statistical techniques such as central tendency, variances, covariances and correlation, visualization techniques like plots, graphs and charts are used to represent the relations of variables and hidden patterns between the variables to understand the important information and make decisions with that information.

8. Conclusion

The firms are highly dependent on data-driven decision making and the information system is very much integrated with the business process management and used for various competitive advantages. Because of increasing the amount of data, the quality of the collected data is very low. Data cleansing is a process which improves the quality of the datasets and make the dataset errorless for the upcoming analytical operations. Exploratory data analysis (EDA) is an operation performed to get the important information which

is useful for any data science tasks by using various statistical and visualization techniques. Data cleansing concept is totally a different concept from exploratory data analysis, It is not included in EDA. Data cleansing is performed to make the dataset without consisting any error and in normalization form to perform exploratory data analysis correctly.

References

[1] Arfa Skandar, Mariam Rehman, Maria Anjum (October 2015). An Efficient Duplication Record Detection Algorithm for Data Cleansing. In International Journal of Computer Applications (0975 – 8887) Volume 127–No. 6.

[2] Estelle Camizuli, Emmanuel John Carranza (2018). Exploratory Data Analysis (EDA), In the Encyclopedia of Archaeological Sciences. Edited by Sandra L. López Varela. © 2018 JohnWiley & Sons, Inc. Published 2018 by John Wiley & Sons, Inc.

[3] Fakhitah Ridzuan, Wan Mohd Nazmee Wan Zainon (2019). A Review on Data Cleansing Methods for Big Data. In The Fifth Information Systems International Conference 2019, Procedia Computer Science 161 (2019) 731–738.

[4] G. Sunitha, Dr. A. Jaya (May 2013). A Knowledge Based Approach for Automatic Database Normalization. In International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, No 5, May 2013.

[5] Heiko Müller, Johann-Christoph Freytag (January 2003). Problems, Methods, and Challenges in Comprehensive Data Cleansing.

[6] Kabita Sahoo, Abhaya Kumar Samal, Jitendra Pramanik, Subhendu Kumar Pani (October 2019). Exploratory Data Analysis using Python. In International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-12.

[7] Karen A. Monsen. Intervention Effectiveness Research: Quality Improvement and Program Evaluation. © Springer International Publishing AG 2018.

[8] Kofi Adu-Manu Sarpong, John Kingsley Arthur (August 2013). Analysis of Data Cleansing Approaches regarding Dirty Data – A Comparative Study. In International Journal of Computer Applications (0975–8887) Volume 76–No. 7, August 2013.

[9] Kofi Adu-Manu Sarpong, John Kingsley Arthur (August 2013). A Review of Data Cleansing Concepts Achievable Goals and Limitations. In International Journal of Computer Applications (0975–8887) Volume 76–No. 7.

[10] Matthieu Komorowski, Dominic C. Marshall, Justin D. Salciccioli, Yves Crutain (2016). Exploratory Data Analysis. © The Author (s) 2016 in MIT Critical Data, Secondary Analysis of Electronic Health Records.

[11] Ronald D. Snee (2020). Using Exploratory Data Analysis. In Statistical Engineering Handbook, Chapter 3 - Section 3.

[12] Rory M. Leith, Keith W. Hipel & Herman Goertz (1991). Exploratory Data Analysis, Canadian Water resources journal, 16: 1, 81-92.

[13] Hiroyuki Konno, Naoshi Uchihira, Michitaka Kosaka (December 2018). Effective Data Cleansing Method Based on Metadata. International Journal of Japan Association for Management Systems Vol. 10 No. 1, December 2018, pp. 53-58

[14] Sardjono, R Yadi Rakhman Alamsyah, Marwondo3, Elia Setiana (2020). Data Cleansing Strategies on Data Sets Become Data Science. International Journal of Quantitative Research and Modeling Vol. 1, No. 3, pp. 145-156, 2020.

[15] Otmane Azeroual1, 2, 3, Gunter Saake2, Mohammad Abuosba (February 2018). Data Quality Measures and Data Cleansing for Research Information Systems. Journal of Digital Information Management Volume 16 Number 1 February 2018.