Received: 30 January 2024 | Revised: 29 February 2024 | Accepted: 31 March 2024

Conceptualizing Explorative Data Analysis in Applied Statistics

Indra Malakar[1]

Bidur Nepal[2]

Abstract

The paper examines the conceptual comprehension of explorative data analysis and its application. The content analysis technique has been deployed and hypothetical examples have been used for explaining its application for the completion of this study. Generally, it is sensible to begin any statistical analysis with an informal, exploratory examination of the data, and this is often called exploratory data analysis (short form EDA). The ingredients of EDA are discussed in this paper and it is suggested that it is important to observe EDA as an integral part of statistical inference, and several examples are presented to show the usages of EDA. Many graphical techniques are used in EDA that discerns what data set can reveal further beyond formal modeling of data or hypothesis testing task. This really enables researcher to gain depth knowledge of the variables in data sets and their relationship. Similarly, the study explores about the stem and leaf display, five number summaries, box and whisker plot and outliers. Stem and leaf display helps in listing data in array and also aids in finding data as minimum and maximum values. Similarly, five number summaries are useful to determine the shape of the distribution. Furthermore, box and whiskers plot visually display data through quartiles or using five number summaries whereas outliers incorporate to uncover the values that lie beyond the whiskers.

Key Words: box plot, maximum, minimum, Stem and leaf and whisker.

Introduction

Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. EDA assists Data science professionals in various ways: getting a better understanding of data, identifying various data patterns, getting a better understanding of the problem statement. In data mining, Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. EDA is used for observing what the data can tell us before the modeling task. It is not easy to look at a column of numbers or a whole spreadsheet and determine important characteristics of the data. It may

[1] Mr. Malakar is the Assistant Professor of Population Studies, Patan Multiple Campus, TU

[2] Mr. Nepal is the Assistant Professor of Statistics, Patan Multiple Campus, TU

be tedious, boring, and/or overwhelming to derive insights by looking at plain numbers. Exploratory data analysis techniques have been devised as an aid for such situation (Chatfield, 1995). Understanding where outliers occur and how variables are related can help one design statistical analyses that yield meaningful results.

Tukey defined data analysis in 1961 as "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data (Tukey, 1961). Tukey's championing of EDA encouraged the development of statistical computing packages, especially S at Bell Labs. The S programming language inspired the systems S-PLUS and R. This family of statistical-computing environments featured vastly improved dynamic visualization capabilities, which allowed statisticians to identify outliers, trends and patterns in data that merited further study. Sentence structure Exploratory data analysis is generally cross-classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate. Univariate analysis is the simplest form of data analysis, where the data being analyzed consists of only one variable. Since it's a single variable, it does not deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

Similarly, in statistics, exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task (Good, 1983). EDA is different from initial data analysis (IDA) which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. Exploratory Data Analysis refers to the critical process of performing initial investigations on data' so, as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand, before getting them dirty with it.

Just like everything in this world, data has its imperfections. Raw data is usually skewed, may have outliers, or too many missing values. A model built on such data results in sub-optimal performance. In hurry to get to the machine learning stage, some data professionals either entirely skip the exploratory data analysis process or do a very mediocre job. This is a mistake with many implications, that includes generating inaccurate models, generating accurate models but on the wrong data, not creating the right types of variables in data preparation, and using resources inefficiently (Nabriya, 1999).

Objectives

The main objective of this study is to analyze the concept of explorative data analysis in applied statistics. The specific objectives are to provide basic concepts of explorative data analysis and to assess the application of EDA to the readers.

Methods

The information for this research study has been extracted from the secondary sources. The content analysis technique has been used as research design. The collected data and descriptive information have been presented in different tables and some examples of application of explorative data analysis as well as hypothetical examples have been used to explain the application of EDA. Furthermore, this study explores only about the stem and leaf display, five number summary, box and whisker plot and outliers.

Analysis and Interpretation of Data

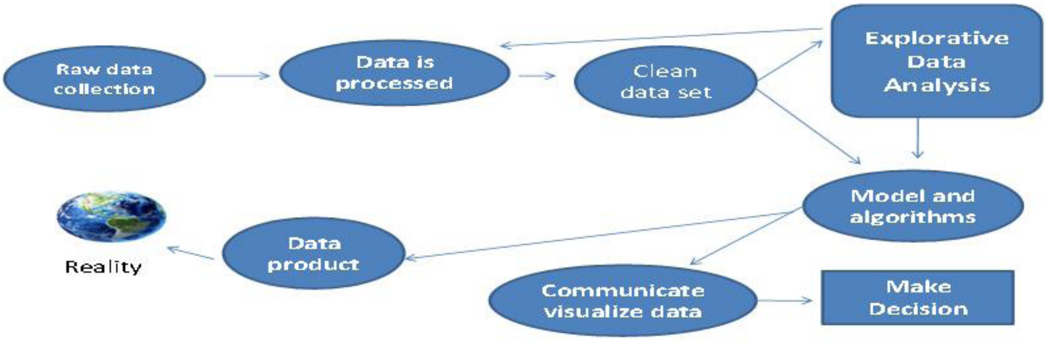
Tukey's EDA was related to two other developments in statistical theory: robust statistics and nonparametric statistics, both of which tried to reduce the sensitivity of statistical inferences to errors in formulating statistical models. Tukey promoted the use of five number summary of numerical data the two extremes (maximum and minimum), the median, and the quartiles because these median and quartile, being functions of the empirical distribution are defined for all distributions, unlike the mean and standard deviation; moreover, the quartile and median are more robust to skewed or heavy-tailed distributions than traditional summaries (the mean and standard deviation). The packages S, S-PLUS, and R included routines using resampling statistics, such as Quenouille and Tukey's jackknife and Efron's bootstrap, which are non-parametric and robust (for many problems). Exploratory data analysis, robust statistics, non-parametric statistics, and the development of statistical programming languages facilitated statisticians' work on scientific and engineering problems. These statistical developments, all championed by Tukey, were designed to complement the analytic theory of testing statistical hypotheses, particularly the Laplacian tradition's emphasis on exponential families *(*Morgenthaler et al., 2000).

The researchers can develop their own idea of EDA from denotation of its name due to the unaware of historical pattern in development of EDA. Sometimes, it is used as exploratory analysis in general. Mulaik in 1984 for example clarified the long history of generic "exploratory statistics" in response to an article concerning

EDA (Good, 1983). Sometimes the model building approach of Box (e.g., 1980) is considered exploratory, although it relies heavily on probabilistic measures than does EDA.

The box plot (box and whisker diagram) is a standardized way of displaying the distribution of data based on the five number summaries: minimum, first quartile, median, third quartile, maximum. In the simplest box plot the central rectangle spans the first quartile to the third quartile (the inter quartile range or IQR). A segment inside the rectangle shows the median and “whiskers” above and below the box show the locations of the minimum and maximum. In research the raw data is collected which is then processed and data set is cleaned. After that for modeling or making decision explorative data analysis is carried out. The process is shown below in flow chart:

Process of data manipulation in EDA



Flow chart of process of data manipulation in EDA

John W. Tukey wrote the book Exploratory Data Analysis in 1977 and held that too much emphasis in statistics was placed on statistical hypothesis testing (confirmatory data analysis); more emphasis needed to be placed on using data to suggest hypotheses to test. In particular, he held that confusing the two types of analyses and employing them on the same set of data can lead to systematic bias owing to the issues inherent in testing hypotheses suggested by the data. The objectives of EDA are to suggest hypotheses about the causes of observed phenomena, assess assumptions on which statistical inference will be based, support the selection of appropriate statistical tools and techniques, provide a basis for further data collection through surveys or experiments (Tukey, 1997).

Techniques and tools

Data specialists primarily use exploratory data analysis to discern what datasets can reveal further beyond formal modeling of data or hypothesis testing tasks. This enables them to gain in-depth knowledge of the variables in datasets and their relationships. Exploratory data analysis can help detect obvious errors, identify outliers in datasets, understand relationships, unearth important factors, find patterns within data, and provide new insights. Developed in the 1970s by American statistician John Tukey famed for his box plot techniques and the Fast Fourier Transform algorithm - EDA continues to find relevance even today in the field of statistical analysis. It allows data professionals to produce relevant and valid results that drive desired business goals. There are four exploratory data analysis techniques that data experts use, which include:

Univariate Non-Graphical

This is the simplest type of EDA, where data has a single variable. Since there are is only one variable, data professionals do not have to deal with relationships. The relationship like the increase in one variable has been caused by another variable and vice versa.

Univariate Graphical

Non-graphical techniques do not present the complete picture of data. Therefore, for comprehensive EDA, data specialists implement graphical methods, such as stem-and-leaf plots, box plots, and histograms.

Multivariate Non-Graphical

Multivariate data consists of several variables. Non-graphic multivariate EDA methods illustrate relationships between two or more data variables using statistics or cross-tabulation.

Multivariate Graphical

This EDA technique makes use of graphics to show relationships between two or more datasets. The widely-used multivariate graphics include bar chart, bar plot, heat map, bubble chart, run chart, multivariate chart, and scatter plot.

There are a number of tools that are useful for EDA, but EDA is characterized more by the attitude taken than by particular techniques (Tukey, 1980). Typical graphical techniques used in EDA are: box plot, histogram, multivariate chart, run chart, pareto chart, scatter plot, stem-and-leaf plot, parallel coordinates, odds ratio, targeted projection pursuit, glyph-based visualization methods such as phenol plot and chernoff faces, projection methods such as grand tour, guided tour and manual tour, Interactive versions of these plots (Sailem et al., 2015). Among them we will be discussing on stem and leaf display, five number summary, box plot and whiskers and outliers on the basis of hypothetical examples in this paper.

Stem and leaf Display

This method was introduced by Tukey (1977). It is the method of sorting the data in such a way that each number is divided into two parts called a Stem and a Leaf. A stem is a leading digit(s) of each number and is used in sorting. A Leaf is the rest of the number or trailing digit(s). A vertical line is used to separate the stem and Leaf (Leaves). For two digits number suppose the number is 85. This can be shown in Leaf and stem as follows:

Leading digit	Trailing digit
8 (Stem)	5 (Leaf)

Suppose we have three digits number i.e. 763. It can be displayed in stem and leaf in two ways. Either make first two digits as stem and remaining digit as leaf or make first single digit as stem and second and third two digits as leaves.

Leading digit	Trailing digit
76 (Stem)	3 (Leaf)

Or

Leading digit	Trailing digit
27 (Stem)	63 (Leaf)

For a number with decimal like 2.5 we can display this number as stem and leaf as follows:

Leading digit	Trailing digit
2 (Stem)	5 (Leaf - decimal)

It is to be noted that the stem is sorted or listed only once and that no numbers are left. To construct stem and leaf display the observation must first be arranged in ascending order. For example: Age of 30 people for voting in election were as follows; 31, 18,37, 54, 61, 64, 40, 43, 71, 51, 12, 20, 52, 65, 53, 42, 39, 62, 74, 48, 29, 67, 30, 49, 68, 35, 57, 26, 58, 27. Now we construct stem and leaf display and list the data in a array as follows: First we have to locate maximum and minimum values of the data. Here, minimum value is 12 and maximum value is 72. Since all the values are two digits so stems and leaves both will be of single digit. Now we will list all the possible stems of the numbers by observing the data. Leaves will be filled in one by one by observing the data. Then we display stem and leaf as follows:

Stem	Leaf	Leaf (ordered)
1	8 2	2 8
2	0 9 6 7	0 6 7 9
3	1 7 9 0 5	0 1 5 7 9
4	0 3 2 8 9	0 2 3 8 9
5	4 1 2 3 7 8	1 2 3 4 7 8
6	1 4 5 2 7 8	1 2 4 5 7 8
7	1 4	1 4

Now using stems and corresponding ordered leaves, the array of data is given below:

12	18	20	26	27	29	30	31	35	37	39	40
42	43	48	49	51	52	53	54	57	58	61	62
64	65	67	68	71	74						

This is the array of data in ascending order. It is easier to array data with the help of leaf which have been ordered from the above table. Some more examples of leaf and stem display of decimal numbers is given below.

For example: The heights of 18 pupils are given in centimeter: 3.5, 3.8, 3.8, 3.9, 4.0, 4.1, 4.2, 4.3, 4.3, 4.4, 4.6, 4.8, 5.0, 5.1, 5.2, 5.4, 5.4, and 5.5. We display the data in stem and leaf plot as follows: Here, 3.5, 3.8, 3.8, 3.9, 4.0, 4.1, 4.2, 4.3, 4.3, 4.4, 4.6, 4.8, 5.0, 5.1, 5.2, 5.4, 5.4, 5.5. The data are arranged in ascending order with minimum value 3.5 and maximum value 5.5. Now we display the above information in Stem and Leaf plot as:

Stem	Leaf
3	5 8 8 9
4	0 1 2 3 3 4 6 8
5	1 2 4 4 4 5 (decimal)

Advantages of Stem and Leaf Display

- Individual identity of the observation in not lost due to grouping process as in case of frequency table.
- Stem and leaf is useful for listing data in array.

- It can easily be converted to frequency distribution but the converse cannot be done.
- It helps in finding the range as minimum and maximum values are known.

Five Number Summaries

A five-number summary is especially useful in descriptive analyses or during the preliminary investigation of a large data set. A summary consists of five values: the most extreme values in the data set (the maximum and minimum values), the lower and upper quartiles, and the median. These values are presented together and ordered from lowest to highest: minimum value, lower quartile (Q_1), median value (Q_2), upper quartile (Q_3), maximum value. These values have been selected to give a summary of a data set because each value describes a specific part of a data set: the median identifies the centre of a data set; the upper and lower quartiles span the middle half of a data set; and the highest and lowest observations provide additional information about the actual dispersion of the data. This makes the five-number summary a useful measure of spread. A five-number summary simply consists of the smallest data value, the first quartile, the median, the third quartile, and the largest data value. The five number summaries is very useful to determine the shape of distribution. It consists of the smallest value (X smallest), the first quartile or lower quartile Q_1, median (Md or Q_2), the third quartile or upper quartile (Q_3) and the largest value (X largest).

Discussion on shape of distribution on the basis of five number summaries

The distribution of the data set is said to be left skewed if the distance from the X smallest to the median is greater than the distance from the median to X largest or the distance from X smallest to Q_1 is greater than the distance from Q_3 to X largest or the distance from Q_1 to the median is greater than the distance from the median to Q_3. It is right skewed if the distance from X smallest to the median is less than the distance from the median to X largest or the distance from X smallest to Q_1 is less than the distance from Q_3 to X largest or distance from Q_1 to the median is less than the distance from the median to Q_3. It is symmetric if both distances are the same. The hypothetical examples have been given below for the explanation.

For example, the five-summary number of data set for the sample of 11 observations are as follows: minimum value = 1, $Q_1 = 5$, Q_2 or Md. = 9, $Q_3 = 18$ and maximum value = 27. We discuss on the shape of distribution as**:** Here, X smallest = 1, $Q_1 = 5$, $Q_2 = 9$, $Q_3 = 18$, X largest = 27.

We find the distance from X smallest to the median is 1 – 9 = -7

Distance from the median to X largest = 9 – 27 = -18

Distance from X smallest to Q_1 = 1 -5 = -4

Distance from the Q_3 to X largest = 18 - 27= -9

Distance from Q_1 to median = 5 – 9 = -4

Distance from Q_2 to Q_3 = 9 – 18 = -9

Now as distance from X smallest to the median is -7 which is greater than distance from x smallest to median which is -18, distance from X smallest to Q_1 is -4 which is greater than distance from Q_3 to X largest -9 and distance from Q_1 to median is -4 which is greater than distance from Q_2 to Q_3 is -9, the distribution is left skewed. Some more examples are 426, 342, 317, 545, 264, 541, 1049, 631, 512, 266, 492, 562, 298 are the sample of batteries from a day's production and used them continuously until they were drained. We list the five-number summary and comment on shape of the distribution as: Arranging the given data in ascending order 264, 266, 298, 317, 342, 426, 492, 512, 541, 545, 562, 631, and 1049.

Now Smallest value, X smallest = 264, Largest value, X largest = 1049.

Position of Q_1 = $\left(\frac{N+1}{4}\right)^{th}$ term = 13+1/4 =14/4 = 3.5th term = value of 3rd term + 0.75 (4th term -3rth term) = 298 + 0.75 (317 – 298) = 312.25.

Position of Q_2 = $\left(\frac{N+1}{2}\right)^{th}$ term = 13+1/2 =14/2 = 7th term = 492.

Position of Q_3 = $3\left(\frac{N+1}{4}\right)^{th}$ term = 3(13+1)/4 =42/4 = 10.5th term = value of 10th term + 0.25 (11th term -10th term) = 545 + 0.25 (562 – 545) = 549.25.

Therefore, the five number summary is 264, 312.25, 492, 549.25, 1049.

Again, distance from X smallest to the Md. = 264 – 492 = -228

Distance from the Md. To the X largest = 492 – 1049 = -557

Distance from X smallest to Q1 = 264 – 312.25 = -48.25

Distance from Q3 to X largest = 549.25 – 1049 = -499.75

As the distance from the X smallest to the median is -228 which is greater than distance from median to X largest -557 and distance from X smallest to Q1 is -48.25 which is greater than the distance from Q3 to X largest -499.75, the distribution is left skewed.

Requirement for five number Summary

Sometimes, it's impossible to find a five-number summary. In order for the five numbers to exist, our data set must meet these two requirements:

- Our data must be univariate. In other words, the data must be a single variable. For example, this list of weights is one variable: 120, 100, 130, 145. If we have a list of ages and we want to compare the ages to weights, it becomes bivariate data (two variables). For example, age 1 (25 pounds), 5 (60 pounds), 15 (129 pounds). The matching pairs make it impossible to find a five-number summary.

- Our data must be ordinal, interval, or ratio.

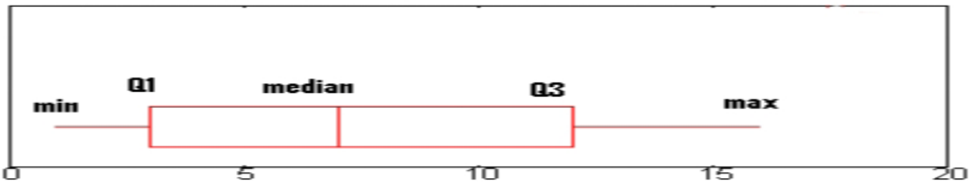
The Box and Whisker plot

A Box and Whisker Plot (or Box Plot) is a convenient way of visually displaying the data distribution through their quartile or using five-number summary. The lines extending parallel from the boxes are known as the "whiskers", which are used to indicate variability outside the upper and lower quartile. Outliers are sometimes plotted as individual dots that are in-line with whiskers. Box Plots can be drawn either vertically or horizontally. Although Box Plots may seem primitive in comparison to a Histogram or Density Plot, they have the advantage of taking up less space, which is useful when comparing distributions between many groups or datasets. A box plot is a graphical device based on a five-number summary. A rectangle (i.e., the box) is drawn with the ends of the rectangle located at the first and third quartile. The rectangle represents the middle 50 per cent of the data. A vertical line is drawn in the rectangle to locate the median. Finally, lines, called whiskers, extend from one end of the rectangle to the smallest data value and from the other end of the rectangle to the largest data value. If outliers are present, the whiskers generally extend only to the smallest and largest data values that are not outliers. Dots, or asterisks, are then placed outside the whiskers to denote the presence of outliers.

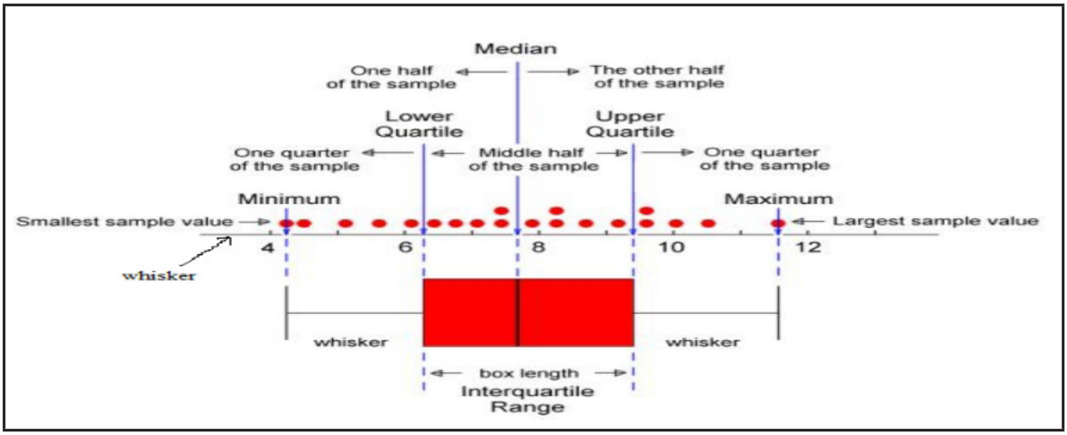
The box and whiskers chart shows how our data is spread out. Five pieces of information (the "five-number summary") are generally included in the chart:

- The minimum (the smallest number in the data set). The minimum is shown at the far left of the chart, at the end of the left "whisker."
- First quartile, Q_1, is the far left of the box (or the far right of the left whisker).
- The median is shown as a line in the center of the box.
- Third quartile, Q_3, shown at the far right of the box (at the far left of the right whisker).

- The maximum (the largest number in the data set), shown at the far right of the box.



A box-plot is a way to show a five-number summary in a chart. The main part of the chart (the "box") shows where the middle portion of the data is: the inter-quartile range. At the end of the box, we find the first quartile (the 25% mark) and the third quartile (the 75% mark). The far left of the chart (at the end of the left "whisker") is the minimum (the smallest number in the set) and the far right is the maximum (the largest number in the set). Finally, the median is represented by a vertical bar in the center of the box. Box plots aren't used that much in real life. However, they can be a useful tool for getting a quick summary of data. The complete figure is shown below:



Advantages of Box and whisker plot

- It displays the 5 –number summary and outliers
- Easy to compare two or more data sets
- Handles extremely large data sets easily.

Disadvantages of Box and whisker plot

- Exact values are not retained
- Not as visually appealing as other graphs

- It does not give information on Kurtosis
- Mean and mode cannot be found out using the box plot.

Some examples of making a box and whisker plot for the following data: 7, 3, 14, 9, 7, 8, 12 are First arranging the data in ascending order 3, 7, 7, 8, 9, 12, 14 as follows:

Now calculating five-number summary as:

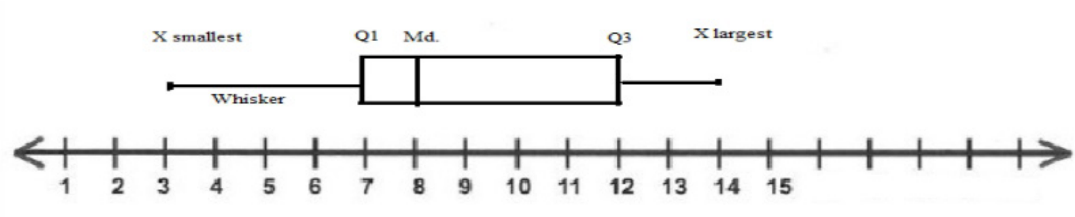
Smallest value, X smallest = 3, Largest value, X largest = 14

Position of Q_1 = $\left(\frac{N+1}{4}\right)^{th}$ term = 7+1/4 = 2nd term = 7

Position of Q_2 =2 $\left(\frac{N+1}{.}\right)^{th}$ term = 2 x 8/4 = 4th term =8

Position of Q_3 = 3 ($\frac{N+1}{4}\big)^{th}$ term = 3(7+1)/4 = 6th term = 12

Now we make box and whisker plot as the following way:



Other hypothetical example is given below as: the time taken by the students to arrive the school by bus and by car are given below as:

By bus (mins.)	14	18	16	22	25	12	32	16	15	18
By car (mins.)	12	10	13	14	9	17	11	10	8	11

We construct the box and whisker plot of the both sets of data on the same number line. We find the similarities and differences the distributions of both sets of data as here, calculating the five summary numbers for the first set of data i.e. by bus

Arranging the data in ascending order 12, 14, 15, 16, 16, 18, 18, 22, 25, 32

Smallest value, X smallest = 12,

Largest value, x largest = 32,

Q_1 = Position of (N+1)4th term = (10+1)4th term = 2.75th term = 2nd term + 0.75 (3rd term – 2nd term) = 14 + 0.75 (15 -14) = 14.75

Q_2 = Position of 2(N+1)4th term = 2(10+1)4th term =2 x 11/4 = 5.5th term = $\frac{16+16}{2}$ = 16

Q_3 = Position of 3(N+1)4th term = 3(10+1)/4th term = 3 x11/4 = 8.25th term = 8th term + 0.25 (9th term – 8th term) = 22 + 0.25 (25 -22) = 22.75

Similarly, the distance from the X smallest to the median 12 – 16 = -4

Distance from the Md. to X largest = 16 – 32 = -16

Distance from the X smallest to Q1 = 12 – 14.75 = -2.75

Distance from the Q_3 to X largest = 22.75 – 32 = - 9.25

Distance from Q_1 to Md. = 14.75 – 16 = -1.25

Distance from Md. To Q_3 = 16 – 22.75 = -6.75.

Again, calculating five-number summary for second set of data i.e. by car:

Arranging the given data in ascending order

8, 9, 10, 10, 11, 11, 12, 13, 14, 17

Now smallest value, X smallest = 8

Largest value, X largest = 17

Q_1 = Position of (n+1)/4th term = (10+1)/4 = 2.75th term = 9 + 0.75(10 – 9) = 9.75

Q_2 = position of 2(n+1)/4th term = 2 x 2.75 = (11 + 11)/2 = 11

Q_3 = Position of 3(n+1)/4th term = 3 x 2.75 = 8.25th term = 13 + 0.25 (14 -13) = 13.25

Similarly, the distance from the X smallest to the median= 8-11 = -3

Distance from the Md. To X largest = 11- 17 = -6

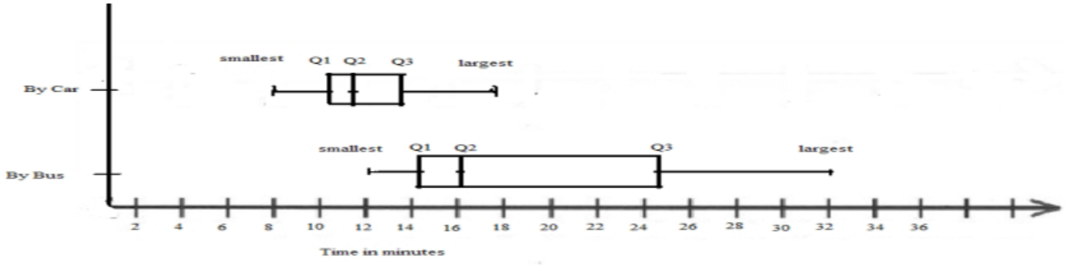
Distance from the X smallest to Q1 = 8 – 9.75 = -1.75

Distance from the Q_3 to X largest = 13.25 – 17 = -3.75

Distance from Q_1 to Md. = 9.75 – 11 = -1.25

Distance from Md. To Q_3 = 11 – 13.25 = -2.25

The box and whisker plot is shown below:



Here five-number summary using by bus is 12, 32, 14.25, 16, 24.25 and by car is 8, 17, 9.25, 11, 13.75 respectively. Similarly, as per the distances calculated for the first set of data using by bus, it showed that the distribution of data is left skewed. Also, in the second set of data using by car, it showed that the distribution of data is left skewed. The similarity is that both are left skewed distribution however second set of data (by car) is slightly less skewed than first set of data (by bus).

Another example of box and whisker plot of the grouped data is shown below and we construct box and whisker plot for the following uniformly distributed data:

Scores	$0 \le X < 10$	$10 \le X < 20$	$20 \le x < 30$	$30 \le x < 40$	$40 \le x < 50$
frequency	**3**	**6**	**8**	**4**	**2**

We calculate of Q_1, Q_2 and Q_3 as:

Scores	Frequency	Cumulative frequency (less than)
0-10	3	3
10-20	6	9
20-30	8	17
30-40	4	21
40-50	2	23
	N = 23	

Now, Q_1 = Position of Q_1 = $(N/4)^{th}$ term = 23/4 = 5.75^{th} term. The c.f greater than 5.75 is 9. Q_1 lies in the class interval 10-20. So, L = 10, f = 6, c.f = 3, h = 10.

$Q_1 = L + \frac{h}{f}(\frac{N}{4}$ - c.f) = 10 + 10/6 (5.75 – 3) = 14.58

Q_2 = Position of Q_2 = 2 x $(N/4)^{th}$ term =2 x 23/4 = 11.5^{th} term. The c.f greater than 11.5 is 17. Q_2 lies in the class interval 20-30. So, L = 20, f = 8, c.f = 9, h = 10.

$Q_2 = L + \frac{h}{f}(\frac{2N}{4}$ - c.f) = 20 + 10/8 (11.5 – 9) = 23.13.

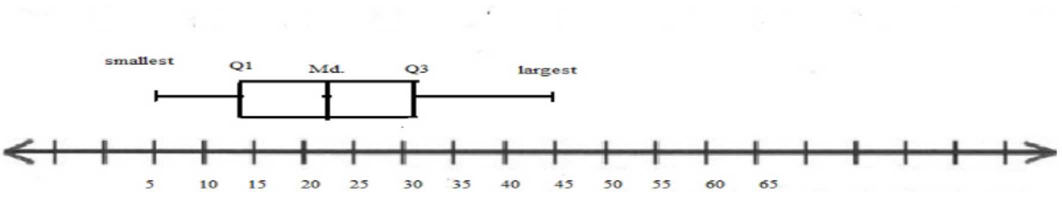
Q_3 = Position of Q_3 = 3 x$(N/4)^{th}$ term = 3 x23/4 = 17.25^{th} term. The c.f greater than

17.25 is 21. Q_3 lies in the class interval 30-40. So, L = 30, f = 4, c.f = 17, h = 10.

$Q_3 = L + \frac{h}{f} (\frac{3N}{4} - c.f) = 30 + 10/4\ (17.25 – 17) = 30.63$

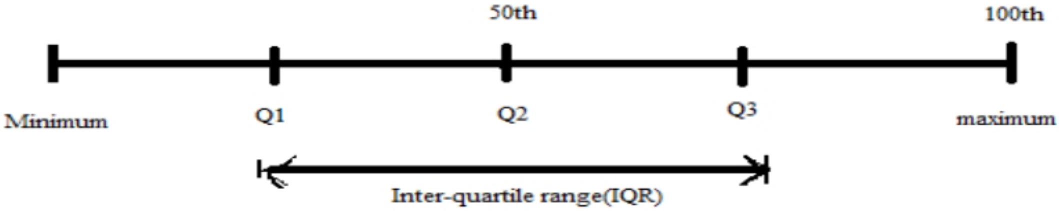
Again, smallest value, X smallest = 5 (in the class interval 0-10, any values from 0 to 10 can be the smallest value in the uniformly distributed data, so, we take the middle value between 0 to 10 so as to make no bias on smallest value between 0 to 10). And largest value is 45.

Now constructing box and whisker plot for the above information:



Outliers

If the values lie beyond the whiskers, then those values are called outliers. The values are considered outliers in the lower end or left end of the data set if it has values less than (Q1 – 1.5 times inter-quartile range) and the values are also considered outliers in the upper end or right end of the data set if its values are greater than Q_3 + 1.5 times inter-quartile range. And inter-quartile is defined as the difference between upper quartile and first quartile mathematically given as $Q_3 – Q_1$.



Let us look above diagram, at the left side is the minimum value and at the right side is the maximum value of the data and the median of the data is the second quartile. This is the 50th percentile of the data. The minimum will be zero percentiles and maximum will be 100th percentile. Now Q_1 is the median of the lower half of the data and Q3 is the median of the upper half of the data. The inter-quartile range (IQR) represents the middle 50% of the data. IQR is the difference between third quartile and the first quartile. Q_1 represents 25th percentile and Q_3 represents the 75th percentile. Sometimes we may have the data point that is either all the way to the right that is very high or to the left that is very low. The outliers exist outside of the (Q_1 – 1.5 times inter-quartile range) or (Q_3 + 1.5 time's inter-quartile range).

For example, let us take the data as 5, 8, 26, 10, 18, 3, 12, 6, 14, 11. Find out whether this data set has outliers or not.

Solution: Arranging data in ascending order 3 5 6 8 10 11 12 14 15 18 26. Here, minimum value is 3 and maximum value is 26.

Q_1 = position of (N+1)/4th term = (11+1)/4 = 3rd term = 6

Q_2 = position of 2(N+1)/4th term =2 (11+1)/4 = 6th term = 11

Q_3 = position of 3(N+1)/4th term = 3(11+1)/4 = 9th term = 15.

Inter-quartile range (IQR) = Q3 – Q1 = 15 – 6 = 9

Again, interval [Q_1 – 1.5 x (IQR), Q_3 + 1.5 x (IQR)] = [6 – 1.5 x (9), 15 + 1.5 x (9)] = [-7.5, 28.5]. Now, outlier will be any number in this list of number that is not in this range [-7.5, 28.5]. So, there are no outliers in the above given data as no list of numbers are outside this range.

Performing Exploratory Data Analysis using popular scripting languages

Data specialists perform exploratory data analysis using popular scripting languages for statistics, such as Python and R. For effective EDA, data professionals also use a variety of BI (Business Intelligence) tools, including Qlik Sense, IBM Cognos, and Tableau. Python and R programming languages enable analysts to analyze data better and manipulate it using libraries and packages such as Plotly, Seaborn, or Matplotlib. BI tools, incorporating interactive dashboards, robust security, and advanced visualization features, provide data processors with a comprehensive view of data that helps them develop Machine Learning (ML) models.

The exploratory data analysis steps that analysts have in mind when performing EDA include:

- Asking the right questions related to the purpose of data analysis
- Obtaining in-depth knowledge about problem domains
- Setting clear objectives that are aligned with the desired outcomes.

The objectives of EDA are:

- It enables unexpected discoveries in the data.
- It suggests hypotheses about the causes of observed phenomena.

- It assesses assumptions on which statistical inference will be based.
- It supports the selection of appropriate statistical tools and techniques.
- Provide a basis for further data collection through surveys or experiments

Given that EDA is not simply a set of techniques but an attitude toward the data (Tukey, 1977), are researchers conducting EDA when they compute exploratory factor analysis or other exploratory statistics? The answer depends on how the analysis is conducted. A researcher may conduct an exploratory factor analysis without examining the data for possible rogue values, outliers, or anomalies; fail to plot the multivariate data to ensure the data avoid pathological patterns; and leave all decision making up to the default computer settings. Such activity would not be considered EDA because the researcher may be easily misled by many aspects of the data or the computer package. Any description that would come from the factor analysis itself would rest on too many unassessed assumptions to leave the exploratory data analyst comfortable. Henderson and Velleman (1981) demonstrated how an interactive (EDA based) approach to stepwise regression can lead to markedly different results than would be obtained by automated variable selection. This occurs because the researcher plots the data and residuals at each stage and thereby considers numerous patterns in the data while the computer program is blind to all aspects of the data except the R 2 (Behren, 1997).

Conclusion

Exploratory Data Analysis is evidently one of the most important steps during the entire process of extracting insights out of data, even before the actual analysis or modeling begins. Therefore, for any researchers that want to truly strap up the power of data, putting their strengths and focus on the EDA phase could help them set up a solid foundation for their overall analysis efforts. So, this study attempts to analyze the basic concept and application of the explorative data analysis by using different method like visual method and so on using leaf and stem display, five-number summary, and box and whisker plot. Hence from above study, we can gain certain information of explorative data analysis to be used in the study.

References

Chatfield*, C.* (1995). *Problem solving:* A Statistician's Guide (2nd ed.). Chapman and Hall. ISBN 978-0412606304.

Good, I. J. (1983). The philosophy of exploratory data analysis*. Philosophy of science*, 50, 238- 295.

Morgenthaler, S., & Fernholz, L. T. (2000). "Conversation with John W. Tukey and Elizabeth Tukey, Luisa T. Fernholz and Stephan Morgenthaler". *Statistical*

Science. 15 (1): 79–94. doi:10.1214/ss/1009212675.

Mulaik, S. A. (1984). Empiricism and exploratory statistics. *Philosophy of science*, 52, 410-430.

Sailem, H. Z., Sero, J, E., Bakal, C. (2015). "Visualizing cellular imaging data using pheno plot". *Nature communications. 6 (1): 5825.*

Tukey, J. W. (1980). "We need both exploratory and confirmatory*". The American Statistician. 34 (1): 23–25. doi:10.1080/00031305.1980.10482706.*

Tukey, J. (1961). *The future of data analysis*. July 1961.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Pearson. *ISBN 978-0201076165.*

Nabriya, P. (1999). *Exploratory data analysis and visualization techniques in data science*. Analyticvidhya.com

Behren, J. T. (1997). Principles and procedures of explorative data analysis. *Psychological methods.*Vol.2, No. 2131-160.