

Box Plots: An Introduction to the Visual EDA Wonder

Vimal Rajaseharan¹, Sabarieswar V.^{2,*}

Abstract

A box plot, also known as a box-and-whisker plot, is a widely used statistical graphic that provides a concise summary of the distribution of numerical data. By displaying key descriptive measures such as the median, quartiles, interquartile range (IQR), minimum and maximum values, and potential outliers, box plots allow researchers to quickly understand the spread and central tendency of a dataset. They are especially useful when comparing distributions across multiple groups, as differences in location, variability, and skewness can be identified easily. This paper discusses the construction and interpretation of box plots, emphasizing the significance of their main components. The median indicates the central value of the data, while the quartiles divide the dataset into four equal parts. The interquartile range measures the spread of the middle 50% of observations and helps detect unusual values. Whiskers extend from the box to represent the range of non-outlying observations, whereas outliers are displayed separately to highlight extreme data points. In addition to the traditional box plot, several enhanced versions have been developed to provide further insights into data distributions. Notched box plots incorporate confidence intervals around the median, enabling visual comparison between groups, while violin plots combine box plot features with density estimation to reveal the underlying shape of the distribution. The paper also demonstrates how box plots can be created and customized using the ggplot2 package in R. Through its flexibility and effectiveness, ggplot2 facilitates clear and informative data visualization. Overall, box plots remain an essential tool for exploratory data analysis in statistics, economics, business, and many scientific disciplines.

Keywords: Box plot, quartiles, interquartile range (IQR), outliers, data visualization, ggplot2, r programming, notched box plot, violin plot, exploratory data analysis (EDA)

INTRODUCTION

Box plots are excellent visualization tools. These are graphs drawn to scale. A box plot summarizes the distribution of quantitative data and is particularly useful for comparing variations and trends among multiple groups. It is a graphical EDA (Exploratory Data Analysis) technique used to highlight the center and extremes of a dataset. As an EDA technique, box plots are more data-centric and display differences between populations without making many prior assumptions, unlike Classic DA techniques. Box plots make it easy to compare distributions, as they clearly show the data's center, spread, and overall range at a glance.

*Author for Correspondence

Sabarieswar V.

E-mail: sabarieswar@tax53.net

¹Statistician, Department of Statistics, Psycho Pedagogist and Aviation Expert, Tiruchirappalli, Tamil Nadu, India

²Business Analyst, Department of Statistics, Pedagogist and Aviation Expert, Tiruchirappalli, Tamil Nadu, India

Received Date: May 30, 2026

Accepted Date: June 03, 2026

Published Date: June 15, 2026

Citation: Vimal Rajaseharan, Sabarieswar V. Box Plots: An Introduction to the Visual EDA Wonder. Research & Reviews: Journal of Statistics. 2026; 15(2): 20–30p.

This type of graphical representation is useful in knowing

- Distribution shape
- Central value
- Variability

A box plot represents the data by visualizing a five-number summary as a graph. Those are:

- i. the smallest observation which is the sample minimum

- ii. Q_1 , the lower quartile, which represents the median of the lower half of the data
- iii. The second quartile (Q_2), also known as the median, represents the middle value of the dataset and serves as a measure of the central tendency.
- iv. Q_3 , the upper quartile representing the median of the upper half of the data
- v. The largest observation which is the sample maximum

A box plot can be constructed both vertically and horizontally. The vertical box plot is a useful construction when comparing datasets, and the horizontal box plot is aesthetic when studying a single data set.

Vertical box plots are formed with random variables on the vertical axis and the factor of interest on the horizontal axis when constructed vertically, and vice versa for the horizontal box plots [1].

Construction of a Box Plot

1. Sort the data values from the smallest to the largest.
2. The first (Q_1), second (Q_2), and third (Q_3) quartiles were calculated.
3. The lower edge of the box represents the first quartile (Q_1), also known as the lower hinge.
4. The upper edge of the box represents the third quartile (Q_3), also known as the upper hinge of the box plot.
5. The line drawn inside the box indicates the second quartile (Q_2) or median.
6. A line extending from the point corresponding to the smallest observation to the first quartile bottom hinge is drawn to meet its mid-point. This line is known as the lower whisker.
7. A line extending from the point corresponding to the largest observation to the third quartile top hinge is drawn, meeting its mid-point, known as the upper whisker. Hence, the box plot is also called a box-and-whisker plot [2, 3].

Box Plot - an Illustration

Let us examine the construction of the box plot through an illustration. Let us consider the following data of 37 computers whose speed was examined in seconds through a benchmarking game. The observations were compared for 16 Intel and 21 AMD processors. The observations are presented in Tables 1 and 2.

Components of a Box Plot

Now, let us look at the various components and terminologies associated with a box plot [4].

The Upper and Lower Hinges

The lower and upper hinges correspond to the first (Q_1) and third quartiles (Q_3), respectively.

For the illustration discussed above, the values of the upper and lower hinges are 21 and 17 for the AMD processors, whereas for the Intel processors, they are 25 and 19, respectively, as shown in Figure 1.

When we plot the values calculated above with the observations (time in seconds) on the y-axis and the factor of interest on the x-axis, we arrive at the box plot in Table 3 [5].

Table 1. Construction of box plot through an illustration

Intel Processors	Time (in seconds)
Intel Processors	18, 19, 20, 22, 24, 25, 26, 16, 17, 19, 25, 27, 28, 23, 23, 31
AMD Processors	15, 17, 18, 19, 20, 21, 23, 14, 17, 18, 19, 20, 21, 24, 19, 16, 17, 18, 20, 22, 28

Table 2. Constructing the box plots, let us arrange the given data in ascending order.

Processors	Time (in seconds)
Intel	16, 17, 18, 19, 19, 20, 22, 22, 23, 23, 24, 25, 25, 27, 28, 31
AMD	14, 15, 16, 17, 17, 17, 18, 18, 18, 19, 19, 19, 20, 20, 20, 21, 21, 22, 23, 24, 28

Table 3. Steps for construction:

General Steps:	Intel Calculation	AMD Calculation
The lowest or smallest observation, X_s	16	14
$Q_1 = \frac{1}{4}(n+1)^{\text{th}}$ item, where n is the number of observations	$= \frac{1}{4}(16+1) = 4.25^{\text{th}}$ item $= 4^{\text{th}}$ item + $0.25(5^{\text{th}}$ item – 4^{th} item) $= 19 + 0.25(19-19) = 19$	$= \frac{1}{4}(21+1) = 5.5^{\text{th}}$ item $= 5^{\text{th}}$ item + $0.5(6^{\text{th}}$ item – 5^{th} item) $= 17 + 0.5(17-17) = 17$
$Q_2 = \frac{1}{2}(n+1)^{\text{th}}$ item, where n is the number of observations	$= \frac{1}{2}(16+1) = 8.5^{\text{th}}$ item $= \text{mean of } 8^{\text{th}} \text{ and } 9^{\text{th}} \text{ item}$ $= \frac{1}{2}(22+23) = 22.5$	$= \frac{1}{2}(21+1) = 11^{\text{th}}$ item $= 19$
$Q_3 = \frac{3}{4}(n+1)^{\text{th}}$ item, where n is the number of observations	$= \frac{3}{4}(16+1) = 12.75^{\text{th}}$ item $= 12^{\text{th}}$ item + $0.75(13^{\text{th}}$ item – 12^{th} item) $= 25 + 0.75(25-25) = 25$	$= \frac{3}{4}(21+1) = 16.5^{\text{th}}$ item $= \text{mean of } 16^{\text{th}} \text{ and } 17^{\text{th}} \text{ item}$ $= \frac{1}{2}(21+21) = 21$
The largest observation, X_l	31	28

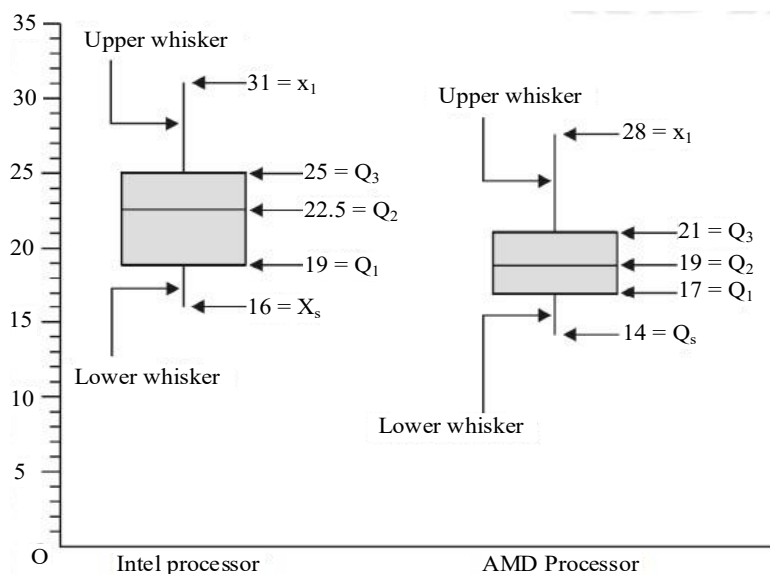


Figure 1. Box plots of Intel and AMD processor performance.

H-Spread

This value was determined by subtracting the lower hinge from the upper hinge. The H-Spread shows the spread of the elements of the data between the first and third quartiles, showing the horizontal spread of the data. This is also known as the Inter Quartile range.

For the data related to Intel in the example discussed above, the H-spread is $25 - 19 = 6$, whereas for AMD, it is $21 - 17 = 4$ [6].

Whiskers

The lines extending above and below the box are called whiskers of the box plot. The lower whisker extends from the point corresponding to the smallest observation to Q_1 and the upper whisker extends from Q_3 to the largest observation.

The maximum length of whiskers can be extended until the upper and lower fences [7].

Step

The step is calculated by multiplying the H-spread by 1.5.

For the data in the illustration above, the value of one step is $1.5 \times 6 = 9$ for Intel processors and $1.5 \times 4 = 6$ for AMD processors.

Upper and Lower Inner Fences

The upper and lower inner fences are calculated by adding one step to the upper hinge and subtracting one step from the lower hinge, respectively. In other words, the upper inner fence is equal to one step ahead of the upper hinge, and the lower inner fence is equal to one step before the lower hinge.

For the data in the illustration, for Intel processors, the upper inner fence is $25 + 9 = 34$ and the lower inner fence is $19 - 9 = 10$ [8].

Upper and Lower Outer Fences

Like the inner fences but calculated at an interval of two steps.

In the illustration, for the Intel processors, it is $25 + 2(9) = 43$ for the upper outer fence and $19 - 2(9) = 1$ for the lower outer fence.

Upper and Lower Adjacent

The upper and lower adjacent are used to represent the largest and smallest observations of the data, respectively.

For the data of Intel processors in the example discussed above, the upper and lower adjacent are 31 and 16, respectively [9].

Outside Value

A data point that lies outside the inner fence but remains within the outer fence is referred to as an outlier. These values indicate observations that are somewhat distant from the main data body. In a box plot, the outside values are typically represented using circles [10].

Far-out Value

The far-out value is beyond the upper or lower outer fence. Asterisks are used to represent these values.

Box-plots with Outliers and + Sign***Outliers, Small 'o'***

Outside value(s) and far-out values (s) are known as extreme observations. If extreme observations are present in the data, they may be represented in the box plot with an individual mark. The extreme observations are described as outliers when represented in box plots in Figure 2. [11, 12]

The outside value is a value that is beyond an inner fence but not beyond an outer fence and is denoted by a small 'o' whereas the far-out value is a value that is beyond the lower and upper outer fences denoted by an asterisk or small 'o.' The individual marks for extreme values can be plotted above and below the whiskers in the box plot. In particular, the outer values are indicated by a small 'o.'

In the data of AMD processors in the above example, 28 is the only far-out value.

Mean, + Sign

Another component that can be included in box plots is the value of some important parameters, such as the mean and mode. In a box plot, the mean value of a dataset is often represented by a '+' symbol. This provides more information about the central tendency of the data, which enables us to understand the symmetry of the distribution in Figure 3.

In the graph above, with respect to the data of AMD processors, the mean is shown by a + sign [13, 14].

Box Plot Distribution vs. Frequency Distribution

The distribution of a box plot can explain how tightly the data are grouped, how the data are skewed, and the symmetry of the data.

In a normal distribution (Bell Curve), the inter-quartile range (IQR), which is defined as $Q_3 - Q_1$, is equal to $SD \times 1.35$ and mean = median. Shown in Figure 4.

Therefore

$$Q_1 = \text{mean} - (0.5 \times \text{IQR})$$

$$Q_3 = \text{mean} + (0.5 \times \text{IQR})$$

For simply:

$$Q_1 = \text{mean} - (0.675 \times SD)$$

$$Q_3 = \text{mean} + (0.675 \times SD)$$

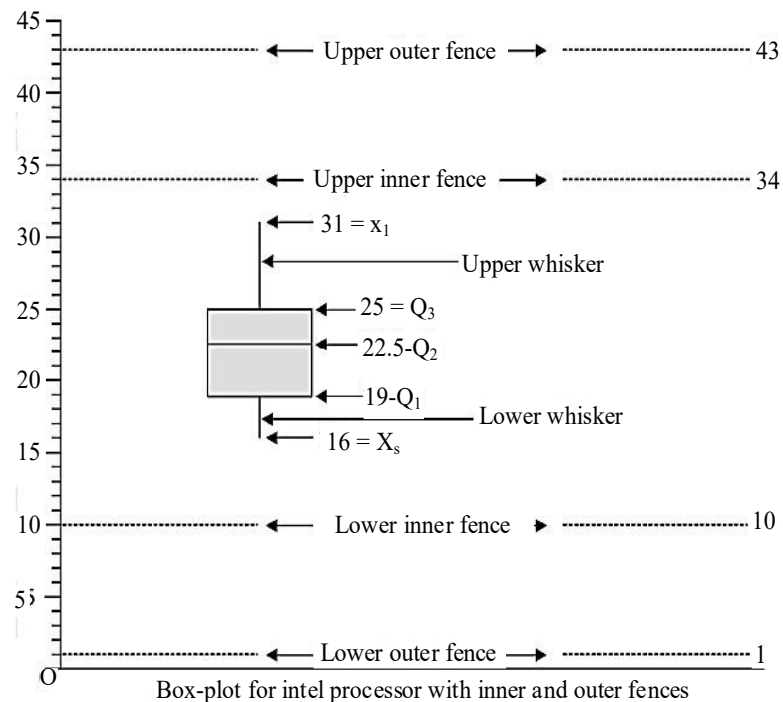


Figure 2. Box plot of Intel processor benchmarking times showing the inner and outer fences.

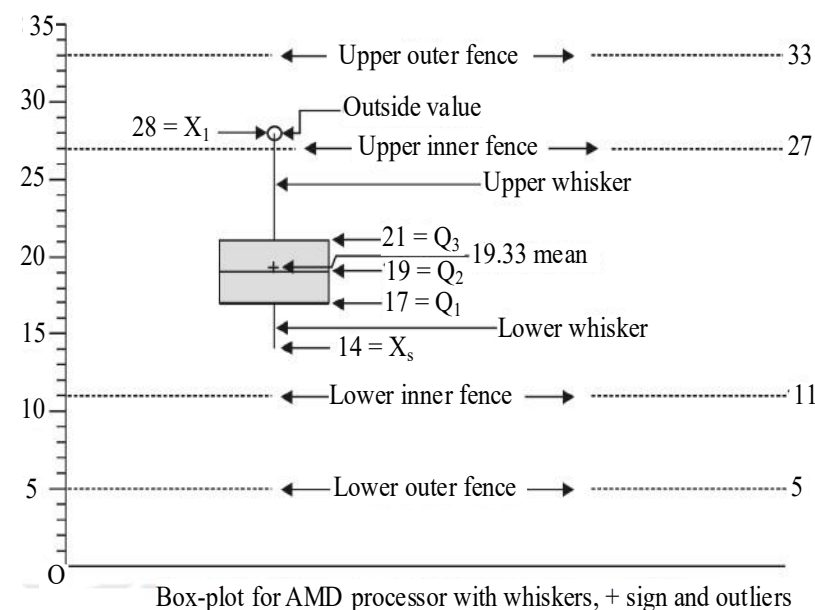


Figure 3. Box plot of AMD processor benchmarking times showing whiskers, mean (+), and an outside value.

- If the median is positioned near the center of the box and the whiskers extend approximately the same distance on both sides, the distribution is considered symmetric, as shown in Figure 5.
- If the median lies closer to the first quartile (Q1) and the lower whisker is shorter than the upper whisker, the distribution is positively skewed (right-skewed).
- If the median is located closer to the third quartile (Q3) and the upper whisker is shorter than the lower whisker, the distribution is negatively skewed (left-skewed) [15].

Width of the Box plot

While the length of the box plot represents the IQR representing 50% of the distribution of the data, the width of the box plot can also be used in some cases.

skewed data.

For example, when comparing two groups of datasets that are the outcome of different strengths of the factor of interest, as in the case of different doses of the same medicine used for the treatment of the same disease in different groups, the same fertilizer of different ratios used on the same crop, etc. [16].

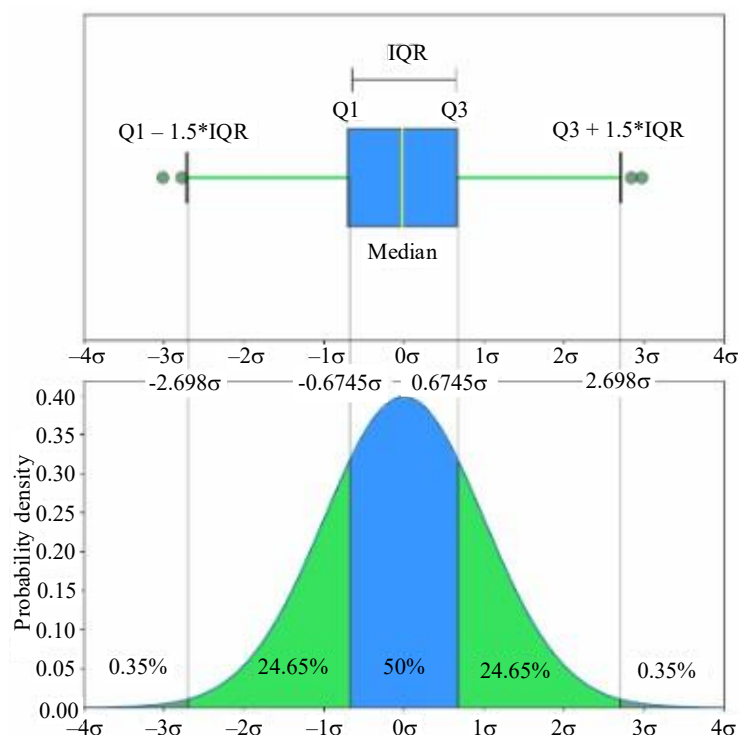


Figure 4. Relationship between the box plot and the normal distribution.

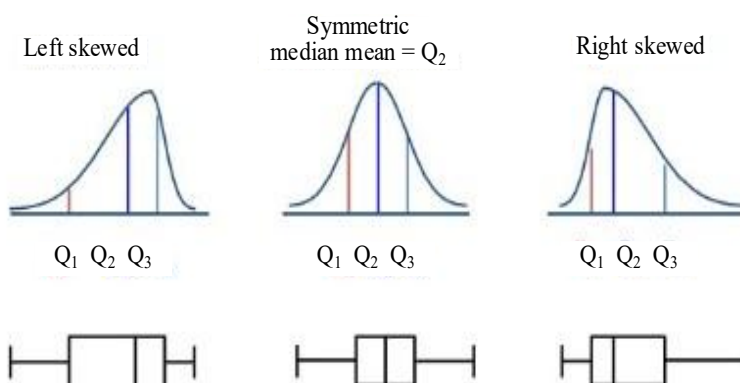


Figure 5. Box plots and corresponding distribution shapes for left-skewed, symmetric, and right-

Why Box plot?

Box plots clearly highlight important landmarks of the data.

The box plot can provide answers to the following questions:

- i. Is a factor significant?
- ii. Does the location differ between subgroups?
- iii. Does the variation differ between subgroups?
- iv. Are there any outliers?

The box plot is also an effective tool summarizing large quantities of information visually like

- Median as a measure of central tendency
- Range and quartile deviation as measures of dispersion
- Symmetry
- Skewness
- Examining outliers, useful in
 - Identifying strong skew in the distribution
 - Identifying possible data collection or data entry errors
 - Providing insight into interesting properties of the data
- Tight grouping of data
- Quick comparison between groups of data

Box plots in R

R Markdown

This RMD document uses “**ggplot2**” library and the “**diamonds**” dataset.

We can call the library ‘ggplot2’ using the **library ()** function; **data ()** function will load the ‘diamonds’ data set, whereas we can inspect the diamonds data set by using the **head ()** function.

```
library(ggplot2)
data("diamonds")
head(diamonds)

## # A tibble: 6 x 10
##   carat cut    color clarity depth table price    x    y    z
##   <dbl> <ord>  <ord> <ord>  <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1 0.23 Ideal  E    SI2    61.5  55  326  3.95  3.98  2.43
## 2 0.21 Premium E    SI1    59.8  61  326  3.89  3.84  2.31
## 3 0.23 Good   E    VS1    56.9  65  327  4.05  4.07  2.31
## 4 0.29 Premium I    VS2    62.4  58  334  4.2   4.23  2.63
## 5 0.31 Good   J    SI2    63.3  58  335  4.34  4.35  2.75
## 6 0.24 Very Good J    VVS2    62.8  57  336  3.94  3.96  2.48
```

Including Box Plots

Now, let us construct a simple box plot where we will use the factor ‘**carat**’ to understand how the distribution of diamonds is based on the attribute ‘carat.’ This is done using the **boxplot ()** in R, as shown in Figure 6.

Using ggplot

We can now use the ggplot2 library. It has a versatile **ggplot()** function that can be used to create several graphs. We can create box plots using the **ggplot()** function. Let us use the ‘cut’ attribute of the diamonds to group the diamonds and compare it based on ‘carat.’ x-axis - ‘cut’ - factor of interest y-axis - ‘carat’ - frequency of observations in Figure 7.

In the code below, **ggplot()** calls the dataset and the aesthetics of the graph; **geom_boxplot()** generates the type of graph.

```
ggplot(diamonds, aes(y=carat, x=cut)) + geom_boxplot()
```

Variants of Box plots

The original restriction of the box plot was that it was designed and computed to be drawn by hand since its introduction by John Tukey in 1970. The introduction of computers in the field of statistics has allowed for substantially more complex variations of box plots by providing a richer distributional summary of the histogram and density plot. These plots help us overcome the limitations of box plots, such as their failure to display multi-modality.

The first variation of the box plot is the **vase-plot** (Benjamin, 1988), where the box is substituted by a symmetrical display of estimated density for the data in the IQR alone, and when the display of density is substituted for all the points, the box plot becomes the **violin-plot** (Hintze and Nelson, 1998) shown in Figure 8.

A Comparison of Box Plot and Violin-Plot in R Using the ‘Diamonds’ Data Set

The bean plot is an advanced variation of the box plot, introduced by Kampstra in 2008. Its name comes from the visual similarity of the plot to a bean pod shape. Each bean is created by mirroring the density curve, resulting in a polygonal display. Individual observations are represented by short lines forming a rug plot within the bean. The outer shape illustrates the data distribution, and the internal lines resemble seeds inside a bean pod. This distinctive appearance inspired the term bean plot. Additionally, a horizontal line is typically included to indicate the overall mean (average) of the data shown in Figure 9 and 10.

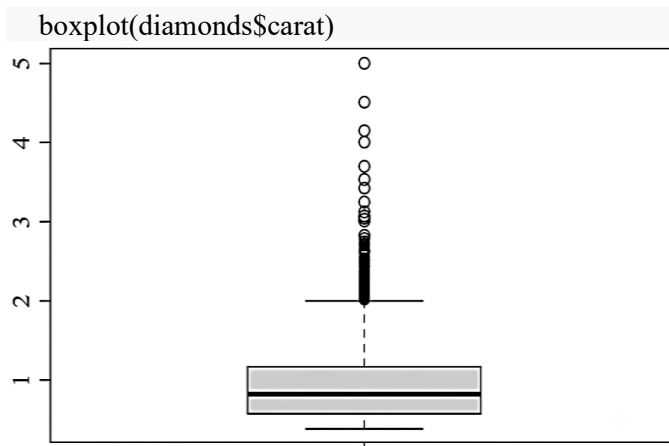


Figure 6. Box plot of diamond carat values showing the distribution and outliers.

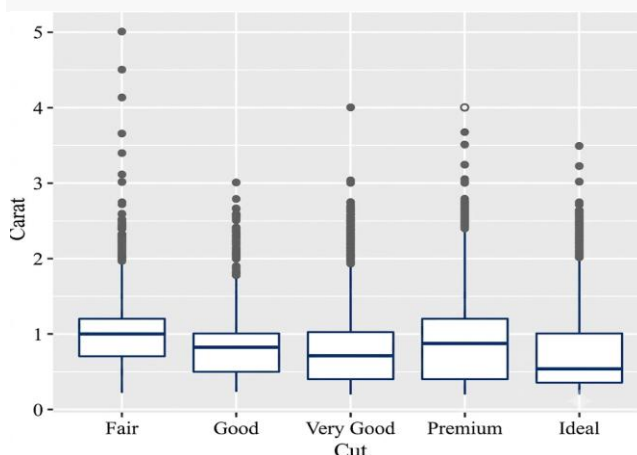
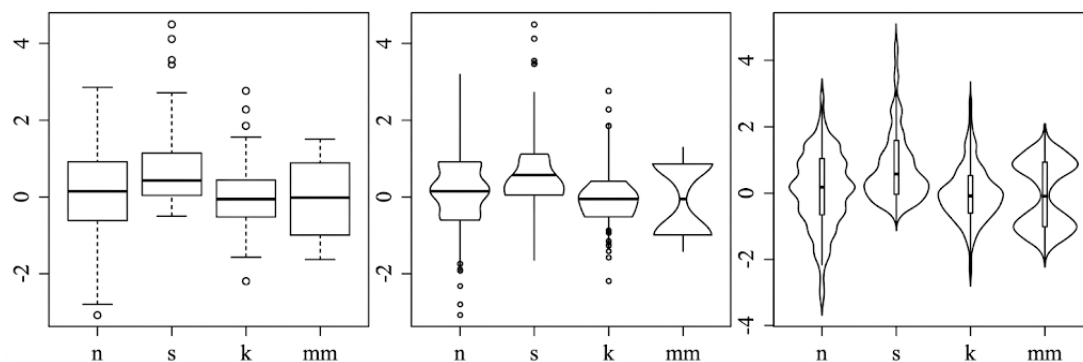


Figure 7. Box plots of diamond carat values across different cut categories using ggplot2.

A Comparison of Box Plot and bean-plot in R using the ‘penguins’ data set

(This illustration uses the ‘penguins’ dataset from the palmer penguins library to take advantage of a larger dataset)

A **letter value box plot** (Hoffmann et al 2006) is an enhancement of the box plot designed to overcome the shortcomings of the box plot for large data. The boxplot displays many outliers and does not take advantage of the more reliable estimates of tail behavior. The letter-value boxplot extends the boxplot with additional letter-values apart from the median (M) and fourths (F): eighths (E), sixteenths (D), ..., until the estimation error becomes excessively large. Each additional letter value is displayed in a slightly smaller box. This display only adds extra letter values and suffers from the same problems as the original boxplot; multimodality is almost impossible to spot, as shown in Figure 10 and 12.



Box-plot vs vase-plot vs violin-plot - A comparison.

(Hadley Wickham and Lisa Stryjewski)

Figure 8. Comparison of box plot, vase plot, and violin plot for data distribution visualization.

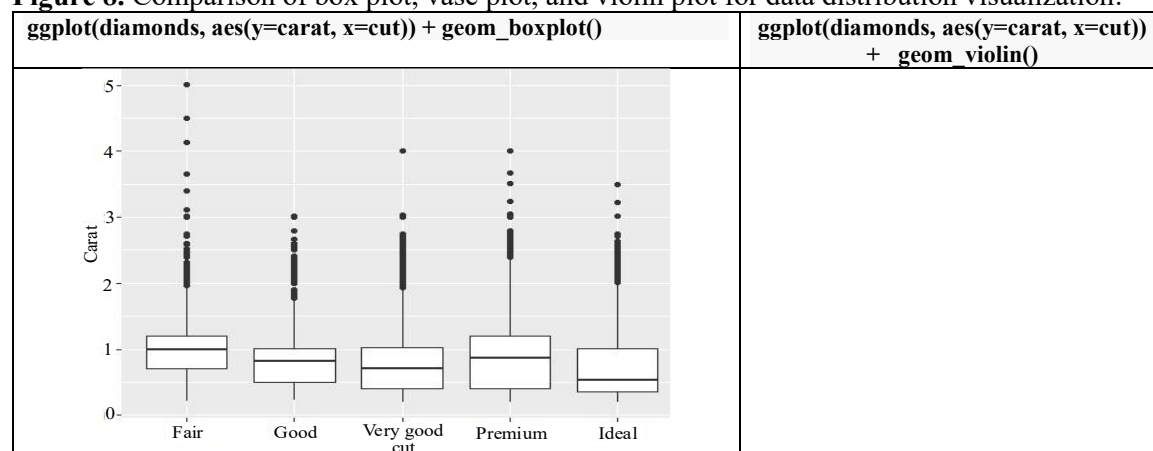


Figure 9. Comparative box plots of diamond carat values across different cut categories.

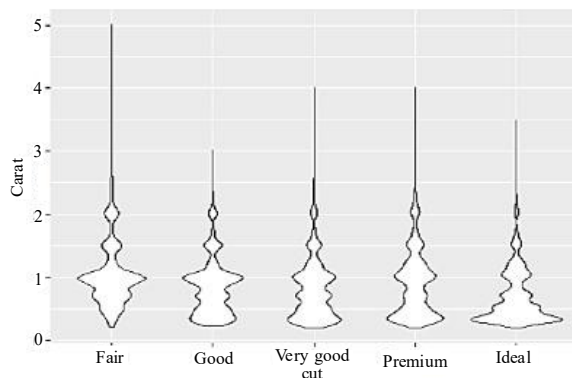


Figure 10. Violin plots of diamond carat values across different cut categories.

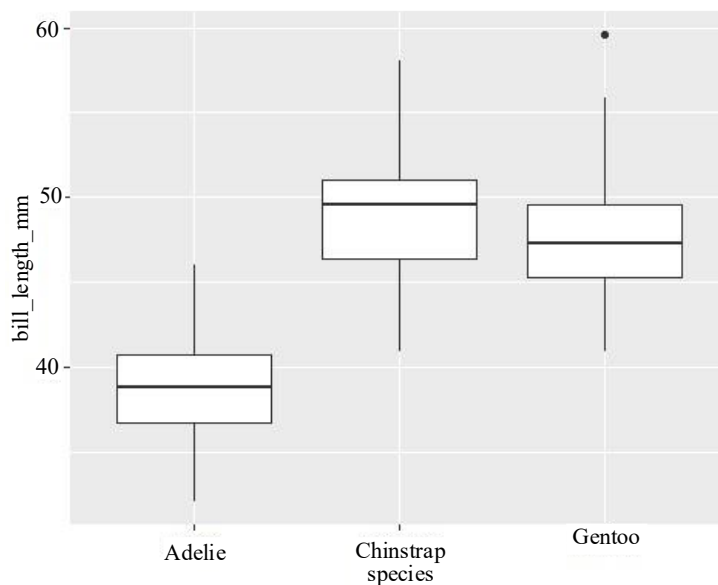


Figure 11. Box plots of penguin bill length across different species.

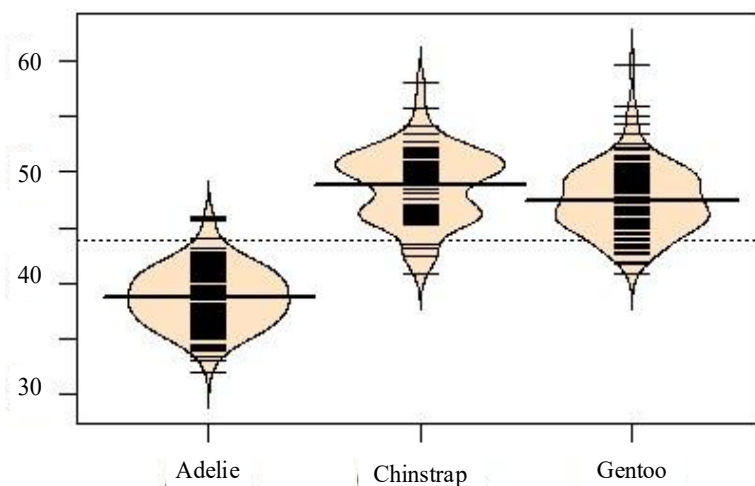


Figure 12. Bean plots of penguin bill length across different species.

CONCLUSION

Box plots are versatile and powerful graphical tools for Exploratory Data Analysis (EDA), providing a concise summary of the distribution, central tendency, variability, skewness, and potential outliers in a dataset. Their simplicity and effectiveness make them valuable for comparing multiple groups and identifying patterns that may not be apparent from numerical summaries. Advances in statistical computing have led to several enhanced variants, such as violin, vase, bean, and letter-value box plots, which overcome some limitations of the traditional box plot by providing additional information about the underlying data distribution. These visualization techniques enable researchers to better understand complex datasets and support informed decision making. With the availability of statistical software such as **R** and the **ggplot2** package, box plots and their variants can be generated efficiently, making them indispensable tools in modern data analysis across statistics, business, economics, medicine, engineering and many other scientific disciplines.

REFERENCES

1. Wickham H, Stryjewski L. 40 years of boxplots. In: Chen CH, Härdle W, Unwin A, editors. Handbook of data visualization. Berlin: Springer; 2008. p. 275-289.
2. Kampstra P. Beanplot: a boxplot alternative for visual comparison of distributions. J Stat Softw.

- 2008;28(1):1-9.
3. Lander JP. R for everyone. 2nd ed. Boston (MA): Pearson Education; 2017.
 4. McGill R, Tukey JW, Larsen WA. Variations of box plots. Am Stat. 1978;32(1):12-16.
 5. Tukey JW. Exploratory data analysis. Reading (MA): Addison-Wesley Publishing Company; 1977.
 6. Chambers JM, Cleveland WS, Kleiner B, Tukey PA. Graphical methods for data analysis. Boston (MA): Duxbury Press; 1983.
 7. Cleveland WS. Visualizing data. Summit (NJ): Hobart Press; 1993.
 8. Wickham H. ggplot2: elegant graphics for data analysis. 2nd ed. New York (NY): Springer; 2016.
 9. Hintze JL, Nelson RD. Violin plots: a box plot-density trace synergism. Am Stat. 1998;52(2):181-184.
 10. Wilke CO. Fundamentals of data visualization: a primer on making informative and compelling figures. Sebastopol (CA): O'Reilly Media; 2019.
 11. Chang W. R graphics cookbook. 2nd ed. Sebastopol (CA): O'Reilly Media; 2018.
 12. Additional Strong Journal References on Boxplots and Visualization
 13. Frigge M, Hoaglin DC, Iglewicz B. Some implementations of the boxplot. Am Stat. 1989;43(1):50-54.
 14. Hubert M, Vandervieren E. An adjusted boxplot for skewed distributions. Comput Stat Data Anal. 2008;52(12):5186-5201.
 15. Spear ME. Charting statistics: box plots. J Stat Educ. 2004;12(2):1-9.
 16. Marmolejo-Ramos F, Tian TS. The shifting boxplot: a boxplot based on essential summary statistics around the mean. Int J Psychol Res. 2010;3(1):37-45.