I'm not a robot

Box Plot is a graphical method to visualize data distribution for gaining insights and making informed decisions. Box plot is a type of chart that depicts a group of numerical data through their quartiles. In this article, we are going to discuss components of a box plot, how to create a box plot, uses of a Box Plot, and how to compare box plots. What is a Box Plot?The idea of box plot was presented by John Tukey in 1970. He wrote about it in his book "Exploratory Data Analysis" in 1977. Box plot is also known as a whisker plot, box-and-whisker plot, or simply a box-and whisker diagram. Box plot is a graphical representation of the distribution of a dataset. It displays key summary statistics such as the median, quartiles, and potential outliers in a concise and visual manner. By using Box plot you can provide a summary of the distribution, identify potential and compare different datasets in a compact and visual manner. Elements of Box Plot A box plot gives a five-number summary of a set of data which is- Minimum - It is the minimum value in the dataset excluding the outliers.First Quartile (Q1) - 25% of the data lies below the First (lower) Quartile.Median (Q2) - It is the mid-point of the dataset. Half of the values lie below it and half above.Third Quartile (Q3) - 75% of the data lies below the Third (Upper) Quartile.Maximum - It is the maximum value in the dataset excluding the outliers.Note: The box plot shown in the above diagram is a perfect plot with no skewness. The plots can have skewness and the median might not be at the center of the box. The area inside the box (50% of the data) is known as the Inter Quartile Range. The IQR is calculated as - IQR = Q3-Q1Outlies are the data points below and above the lower and upper limit. The lower and upper limit is calculated as - Lower Limit = Q1 - 1.5*IQRUpper Limit = Q3 + 1.5*IQRThe values below and above these limits are considered outliers and the minimum and maximum values are calculated from the points which lie under the lower and upper limit. How to create a box plots?Let us take a sample data to understand how to create a box plot. Here are the runs scored by a cricket team in a league of 12 matches - 100, 120, 110, 150, 110, 140, 130, 170, 120, 220, 140, 110. To draw a box plot for the given data first we need to arrange the data in ascending order and then find the minimum, first quartile, median, third quartile and the maximum. Ascending Order 100, 110, 110, 110, 120, 120, 130, 140, 140, 150, 170, 220 Median (Q2) = (120+130)/2 = 125; Since there were even values To find the First Quartile we take the first six values and find their median. Q1 = (110+110)/2 = 110For the Third Quartile, we take the next six and find their median. Q3 = (140+150)/2 = 145Note: If the total number of values is odd then we exclude the Median while calculating Q1 and Q3. Here since there were two central values we included them. Now, we need to calculate the Inter Quartile Range. IQR = Q3-Q1 = 145-110 = 35We can now calculate the Upper and Lower Limits to find the minimum and maximum values and also the outliers if any. Lower Limit = Q1-1.5*IQR = 110-1.5*35 = 57.5Upper Limit = Q3+1.5*IQR = 145+1.5*35 = 197.5 So, the minimum and maximum between the range [57.5,197.5] for our given data are - Minimum = 100Maximum = 170 The outliers which are outside this range are - Outliers = 220Now we have all the information, so we can draw the box plot which is as below- We can see from the diagram that the Median is not exactly at the center of the box and one whisker is longer than the other. We also have one Outlier. Use-Cases of Box PlotBox plots provide a visual summary of the data with which we can quickly identify the average value of the data, how dispersed the data is, whether the data is skewed or not (skewness).The Median gives you the average value of the data.Box Plots shows Skewness of the data-a) If the Median is at the center of the Box and the whiskers are almost the same on both the ends then the data is Normally Distributed.b) If the Median lies closer to the First Quartile and if the whisker at the lower end is shorter (as in the above example) then it has a Positive Skew (Right Skew).c) If the Median lies closer to the Third Quartile and if the whisker at the upper end is shorter than it has a Negative Skew (Left Skew). The dispersion or spread of data can be visualized by the minimum and maximum values which are found at the end of the whiskers.The Box plot gives us the idea of about the Outliers which are the points which are numerically distant from the rest of the data.How to compare box plots?As we have discussed at the beginning of the article that box plots make comparing characteristics of data between categories very easy. Let us have a look at how we can compare different box plots and derive statistical conclusions from them. Let us take the below two plots as an example: - Compare the Medians If the median line of a box plot lies outside the box of the other box plot with which it is being compared, then we can say that there is likely to be a difference between the two groups. Here the Median line of the plot B lies outside the box of Plot A.Compare the Dispersion or Spread of data The Inter Quartile range (length of the box) gives us an idea about how dispersed the data is. Here Plot A has a longer length than Plot B which means that the dispersion of data is more in plot A as compared to plot B. The length of whiskers also gives an idea of the overall spread of data. The extreme values (minimum &maximum) give the range of data distribution. Larger the range more scattered the data. Here Plot A has a larger range than Plot B.Comparing Outliers The outliers give the idea of unusual data values which are distant from the rest of the data. More number of Outliers means the prediction will be more uncertain. We can be more confident while predicting the values for a plot which has less or no outliers.Compare Skewness Skewness gives us the direction and the magnitude of the lack of symmetry. We have discussed above how to identify skewness. Here Plot A is Positive or Right Skewed and Plot B is Negative or Left Skewed.This is all for Box Plots. Now you might have got the idea of Box Plots how to make them and how to derive information from them. For any queries do leave a comment down below. In this post, I want to share how Python can be used to automate the documentation of machine-learning (ML) experiments using AsciiDoc.The search for the best-performing ML model is an empirical process, which involves fitting models with differing parameters and evaluating their predictive performance. Only after a multitude (e.g. hundreds or thousands) of models have been evaluated, is it possible confidently proclaim that a suitable model has been identified. The major challenge of running vast numbers of experiments is that they are time- and compute-intensive because results usually have to be delivered within a certain time frame (e. Radar visualizations for technological choices have been pioneered by ThoughtWorks. In the meantime, many organizations have created their own tech radars to map out which technologies should be considered for use by members of the organization.The German online fashion retailer Zalando has even made the source code of their tech radar publicly available. Since technological decisions for data science and AI projects are distinct from conventional applications, I decided to adapt Zalandos tech radar. Protocol buffers (Protobuf) are a language-agnostic data serialization format developed by Google. Protobuf is great for the following reasons: Low data volume: Protobuf makes use of a binary format, which is more compact than other formats such as JSON. Persistence: Protobuf serialization is backward-compatible. This means that you can always restore previous data, even if the interfaces have changed in the meantime. Design by contract: Protobuf requires the specification of messages using explicit identifiers and types. Are you a researcher in data science? Are you in desparate need for GPU ressources for your next project? Then you should know that a GPU server may be just around the corner.HOSTKEY is currently hosting a competition where you can win a grant for free GPU ressources. The competition is open to all researchers in the data science sphere.Application Criteria for the Grant Program If you want to apply, you have to send the following information: Companies usually have firewalls in place, which ensure that the internal network is protected. To access the outside world, all traffic must be routed through a proxy. When you are using the standard operating system (typically Windows), you are automatically authenticated with this proxy.However, when you are using a non-standard operating system (e.g. through a virtual machine running Linux), you are not automatically authenticated with the companys proxy. The sad result: you wont be able to access the internet out of the box. Flask is a lightweight Python web development framework that is becoming more and more popular, as you can see from this comparisonagainst Django. AWS (Amazon Web Services) certifications are among the most lucrative certifications in the IT sector. This is due to the growing demand for professionals with cloud expertise, as more and more companies are adopting cloud technology. Furthermore, AWS upholds high quality standards when it comes to certification. So, while certification can be challenging, there is a lot to learn along the way.I only recently had my first exposure to cloud computing when I took on a DevOps role in industry in 2019. The Cambridge Dictionary defines plagiarism as the process or practice of using another persons ideas or work and pretending that it is your own. In the last years, there have been several famous Germans who lost their PhD titles due to plagiarizing their doctoral theses. In Germany, VroniPlag is the largest open community that analyzes scientific work with respect to plagiarism. Most notably, in 2011, Guttenplag (a specific group of plagiarism hunters) published a detailed analysis of the doctoral thesis by Karl-Theodor zu Guttenberg, the German defense minister at that time. Lets say you are currently adding new arguments to an installation script for your software. After some work, your commit history may look different than youwould like. When I started working in the IT sector, I was impressed by the large number of different roles that exist and it took me quite a bit of time to understand their individual responsibilities. That is why I thought it would be nice to share my understanding of the most common roles you will encounter in IT projects.You should definitely read this post if you are thinking about applying for position in the information technology sector but are unsure which one is the right fit for you or if youre already working in IT and want to improve your understanding of other roles. A boxplot, also known as a box plot or box-and-whisker plot, is a standardized way of displaying the distribution of a data set based on its five-number summary of data points: the minimum, first quartile [Q1], median, third quartile [Q3] and maximum. Heres an example.Different parts of a boxplot | Image: Michael Galarnyk Boxplots can tell you about your outliers and what their values are. They can also tell you if your data is symmetrical, how tightly your data is grouped and if and how your data is skewed. A boxplot is a standardized way of displaying the distribution of data based on its five-number summary (minimum, first quartile [Q1], median, third quartile [Q3] and maximum). Boxplots can tell you about your outliers and their values, if your data is symmetrical, how tightly your data is grouped and if and how your data is skewed. In this tutorial Ill answer the following questions: What is a boxplot? How can I understand the anatomy of a boxplot by comparing a boxplot against the probability density function for a normal distribution? How do you make and interpret boxplots using Python? As always, the code used to make the graphs is available on my GitHub. With that, lets get started. More Statistics From Built In ExpertsWhat Is Descriptive Statistics? Box and Whisker Plots Explained. | Video: Math with Mr. J What Is a Boxplot? A boxplot is a graph that gives a visual indication of how a data sets 25th percentile, 50th percentile, 75th percentile, minimum, maximum and outlier values are spread out and compare to each other. A boxplot is drawn as a box with a line inside of it, and has extended lines attached to each of its sides (known as whiskers). The box is used to represent the interquartile range (IQR) or the 50 percent of data points lying above the first quartile and below the third quartile in the given data set. The whiskers are used to represent the variability of the minimum, maximum and any outlier data points in comparison to the IQR (the longer the whisker, the wider the variability of the attached data points to the IQR). The boxs left edge or bottom end represents the first/lower quartile (Q1; the 25th percentile) of the data. The line inside the box represents the median (Q2; the 50th percentile) of the data. The boxs right edge or top end represents the third/upper quartile (Q3; the 75th percentile) of the data. If a dot, cross or diamond symbol is present inside the box, this represents the mean of the data. As for whiskers of the boxplot, the left whisker shows the minimum data value and its variability in comparison to the IQR. The right whisker shows the maximum data value and its variability in comparison to the IQR. Whiskers also help present outlier values in comparison to the rest of the data, as outliers sit on the outside of whisker lines.Different parts of a boxplot | Image: Author Median (Q2/50th percentile): The middle value of the data set. First Quartile (Q1/25th percentile): The middle number between the smallest number (not the minimum) and the median of the data set. Third Quartile (Q3/75th percentile): The middle value between the median and the highest value (not the maximum) of the dataset. Interquartile Range (IQR): 25th to the 75th percentile. Whiskers (shown in blue) Outliers (shown as green circles) Minimum: Q1 - 1.5*IQR Maximum: Q3 + 1.5*IQR When to Use a Boxplot A boxplot may help when you need more information from a data set/distribution than just the measures of central tendency (mean, median and mode). Boxplots can illustrate the variability or dispersion of all data points present within a set, giving a good indication of outliers and how symmetrical the data is. Although boxplots 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 data sets. What defines an outlier, minimum or maximum may not be clear yet. The next section will try to clear that up for you. Related Reading From Built InHow to Find Outliers With IQR Using Python Boxplot on a Normal DistributionComparison of a boxplot of a nearly normal distribution and a probability density function (PDF) for a normal distribution | Image: Author The image above is a comparison of a box-and-whisker plot of a nearly normal distribution and the probability density function (PDF) for a normal distribution. The reason why I am showing you this image is that looking at a statistical distribution is more commonplace than looking at a boxplot. In other words, it might help you understand a boxplot. This section will cover: How outliers are (for a normal distribution) 0.7 percent of the data. What a minimum and a maximum are. Probability Density Function and Boxplots This part of the post is very similar to my 689599.7 rule article (normal distribution), but adapted for a boxplot. To be able to understand where the percentages come from, its important to know about the probability density function (PDF). A PDF is used to specify the probability of the random variable falling within a particular range of values, as opposed to taking on any one value. This probability is given by the integral of this variables PDF over that range that is, it is given by the area under the density function but above the horizontal axis and between the lowest and greatest values of the range. This definition might not make much sense so lets clear it up by graphing the probability density function for a normal distribution. The equation below is the probability density function for a normal distribution:PDF for a normal distribution. | Image: Author Lets simplify it by assuming we have a mean () of 0 and a standard deviation () of 1.PDF for a normal distribution. | Image: Author You can graph this using anything, but I choose to graph it using Python. # Import all libraries for this portion of the blog postfrom scipy.integrate import quadimport numpy as npimport matplotlib.pyplot as plt%matplotlib inlinex = np.linspace(-4, 4, num = 100)constant = 1.0 / np.sqrt(2*np.pi)pdf_normal_distribution = constant * np.exp((-x**2) / 2.0)fig, ax = plt.subplots(figsize=(10, 5));ax.plot(x, pdf_normal_distribution);ax.set_ylim(0);ax.set_title('Normal Distribution', size = 20);ax.set_ylabel('Probability Density', size = 20);Normal distribution graphed | Image: Author The graph above does not show you the probability of events but their probability density. To get the probability of an event within a given range we will need to integrate. Suppose we are interested in finding the probability of a random data point landing within the interquartile range .6745 standard deviation of the mean, we need to integrate from -.6745 to .6745. You can do this with SciPy. # Make PDF for the normal distribution a functiondef normalProbabilityDensity(x): constant = 1.0 / np.sqrt(2*np.pi) return(constant * np.exp((-x**2) / 2.0))# Integrate PDF from -.6745 to .6745result_50p, _ = quad(normalProbabilityDensity, -.6745, .6745, limit = 1000)print(result_50p)Image: Author You can do the same for minimum and maximum. # Make a PDF for the normal distribution a functiondef normalProbabilityDensity(x): constant = 1.0 / np.sqrt(2*np.pi) return(constant * np.exp((-x**2) / 2.0))# Integrate PDF from -2.698 to 2.698result_99_3p, _ = quad(normalProbabilityDensity, -2.698, 2.698, limit = 1000)print(result_99_3p)Image: Author As mentioned earlier, outliers are the remaining 0.7 percent of the data. It is important to note that for any PDF, the area under the curve must be one (the probability of drawing any number from the functions range is always one).Image: Author More on Data ScienceHow to Use the Z-Table How to Graph and Interpret a Boxplot This section is largely based on a free preview video from my Python for Data Visualization course. In the last section, we went over a boxplot on a normal distribution, but as you obviously wont always have an underlying normal distribution, lets go over how to utilize a boxplot on a real data set. To do this, we will utilize Python and the Breast Cancer Wisconsin (Diagnostic) Data Set. If you dont have a Kaggle account, you can download the data set from my GitHub. Boxplots using Matplotlib, Pandas, and Seaborn Libraries (Python) | Video: Michael Galarnyk Read in the Data Before graphing, lets read in the data in Python. The code below reads the data into a pandas DataFrame. import pandas as pdimport seaborn as snsimport matplotlib.pyplot as plt# Put dataset on my github repo df = pd.read_csv(') More From Our ExpertsThe Poisson Process and Poisson Distribution, Explained (With Meteors!) How to Graph a Boxplot We use a boxplot below to analyze the relationship between a categorical feature (malignant or benign tumor) and a continuous feature (area_mean). There are a couple ways to graph a boxplot through Python. You can graph a boxplot through Seaborn, Matplotlib or pandas. Graphing a Boxplot With Seaborn The code below passes the pandas DataFrame df into Seaborns boxplot. sns.boxplot(x='diagnosis', y='area_mean', data=df)Image: Author Graphing a Boxplot With Matplotlib I made the boxplots you see in this post through Matplotlib. This approach can be far more tedious, but can give you a greater level of control. malignant = df[df['diagnosis']=='M']['area_mean']benign = df[df['diagnosis']=='B']['area_mean']fig = plt.figure()ax = fig.add_subplot(111)ax.boxplot([malignant,benign], labels=['M', 'B'])You can make this a lot prettier with a little bit of work. | Image: Author Notched Boxplot in Matplotlib The notched boxplot allows you to evaluate confidence intervals (by default 95 percent confidence interval) for the medians of each boxplot. malignant = df[df['diagnosis']=='M']['area_mean']benign = df[df['diagnosis']=='B']['area_mean']fig = plt.figure()ax = fig.add_subplot(111)ax.boxplot([malignant,benign], notch = True, labels=['M', 'B']);Its not the prettiest it can be yet. | Image: Author Graphing a Boxplot With Pandas You can plot a boxplot by invoking .boxplot() on your DataFrame. The code below makes a boxplot of the area_mean column with respect to different diagnosis. df.boxplot(column = 'area_mean', by = 'diagnosis');plt.title('')Image: Author More on Distributions4 Probability Distributions Every Data Scientist Needs to Know How to Interpret a Boxplot Data science is about communicating results so keep in mind you can always make your boxplots a bit prettier with a little bit of work (see the code here).Image: Author Using the graph, we can compare the range and distribution of the area_mean for malignant and benign diagnoses. We observe that there is a greater variability for malignant tumor area_mean as well as larger outliers. Also, since the notches in the boxplots do not overlap, you can conclude that with 95 percent confidence, the true medians do differ. Here are a few other things to keep in mind about boxplots: You can always pull out the data from the boxplot in case you want to know what the numerical values are for the different parts of a boxplot. Matplotlib does not estimate a normal distribution first and instead calculates the quartiles from the estimated distribution parameters. The median and the quartiles are calculated directly from the data. In other words, your boxplot may look different depending on the distribution of your data and the size of the sample (e.g. asymmetric and with more or fewer outliers). Hopefully this wasnt too much information on boxplots. My next tutorial goes over How to Use and Create a Z Table (Standard Normal Table). If you have any questions or thoughts on the tutorial, feel free to reach out through YouTube or X. A boxplot shows the distribution of values in a data set based on its five-number summary. The five-number summary is the minimum, first quartile, median, third quartile and maximum in a data set. To draw a boxplot, do the following: Determine the data set's five-number summary (minimum, first quartile, median, third quartile and maximum values). Draw a number scale, and number it so it can contain the minimum and maximum values. Mark where the five-number summary values fall on the scale. Draw a box where the edges connect at the first quartile and third quartile. Draw a line in the box at the median. Draw lines (whiskers) from the edges of the box that reach to the minimum and maximum values on each side. In a boxplot graph, the box represents the datas interquartile range (IQR), which is the 50 percent of data points above the first quartile and below the third quartile. Each whisker (line) on the side of a boxplot represents the top and bottom 25 percent of data points, where the line at the start of the box goes to the minimum value and the line at the end of the box goes to the maximum value. The longer the whiskers, the larger the variability may be in the data set. Any circles or points outside of the whiskers represent outliers in the data. Boxplots are best used to summarize a data set and show a high-level distribution of data points, especially in comparison to multiple groups or other data sets. A boxplot doesnt show the exact shape of the data distribution (like detailed peaks or data skews in the distribution), individual data points or the total number of data points in a data set. Boxplots also dont always show the mean and mode of a data set. In descriptive statistics, a box plot or boxplot (also known as a box and whisker plot) is a type of chart often used in explanatory data analysis. Box plots visually show the distribution of numerical data and skewness by displaying the data quartiles (or percentiles) and averages. Box plots show the five-number summary of a set of data: including the minimum score, first (lower) quartile, median, third (upper) quartile, and maximum score. A boxplot is a visual representation of the five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. Interquartile Range (IQR): The box plot shows the middle 50% of scores (i.e., the range between the 25th and 75th percentile). Whiskers: The upper and lower whiskers represent scores outside the middle 50% (i.e., the lower 25% of scores and the upper 25% of scores). Minimum Score: The lowest score, excluding outliers (shown at the end of the left whisker). Lower Quartile: Twenty-five percent of scores fall below the lower quartile value (also known as the first quartile). Median: The median marks the mid-point of the data and is shown by the line that divides the box into two parts (sometimes known as the second quartile). Half the scores are greater than or equal to this value, and half are less. Upper Quartile: Seventy-five percent of the scores fall below the upper quartile value (also known as the third quartile). Thus, 25% of data are above this value. Maximum Score: The highest score, excluding outliers (shown at the end of the right whisker). Outliers: Points beyond the whiskers, indicating potential extreme values. Why are box plots useful?Quick summary of key statisticsBy displaying the median, interquartile range, and outliers at a glance, boxplots provide a concise snapshot of a datasets center and spread.Box plots divide the data into sections containing approximately 25% of the data in that set.Box plots are useful as they provide a visual summary of the data enabling researchers to quickly identify mean values, the dispersion of the data set, and signs of skewness.Note that the image above represents data that has a perfect normal distribution, and most box plots will not conform to this symmetry (where each quartile is the same length).The median is the average value from a set of data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value, and half are less.The box plot shape will show if a statistical data set is normally distributed or skewed.When the median is in the middle of the box, and the whiskers are about the same on both sides of the box, then the distribution is symmetric.When the median is closer to the bottom of the box, and if the whisker is shorter on the lower end of the box, then the distribution is positively skewed (skewed right).When the median is closer to the top of the box, and if the whisker is shorter on the upper end of the box, then the distribution is negatively skewed (skewed left).In statistics, dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed.The smallest and largest values are found at the end of the whiskers and are useful for providing a visual indicator regarding the spread of scores (e.g., the range).The interquartile range (IQR) is the box plot showing the middle 50% of scores and can be calculated by subtracting the lower quartile from the upper quartile (e.g., Q3Q1).An outlier is an observation that is numerically distant from the rest of the data.When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot.Outliers can be genuine rare events (e.g., a patient who recovers extraordinarily quickly) or the result of data-entry mistakes, faulty instruments, or other errors. Source: example, outside 1.5 times the interquartile range above the upper quartile and below the lower quartile (Q1 1.5 * IQR or Q3 + 1.5 * IQR). Boxplots make it easy to compare multiple data groups (e.g., different populations or experimental conditions) by aligning the boxes next to each other, highlighting differences in medians, overall spread, or outlier prevalence. How to compare box plotsBox plots are a useful way to visualize differences among different samples or groups. They manage to provide a lot of statistical information, including medians, ranges, and outliers.Note although box plots have been presented horizontally in this article, it is more common to view them vertically in research papers.Step 1: Compare the medians of box plotsCompare the respective medians of each box plot. If the median line of a box plot lies outside of the box of a comparison box plot, then there is likely to be a difference between the two groups. Source: 2: Compare the interquartile ranges and whiskers of box plotsCompare the interquartile ranges (that is, the box lengths) to examine how the data is dispersed between each sample. The longer the box, the more dispersed the data. The smaller, the less dispersed the data.Next, look at the overall spread as shown by the extreme values at the end of two whiskers.This shows the range of scores (another type of dispersion). Larger ranges indicate wider distribution, that is, more scattered data.Step 3: Look for potential outliers (see the above image)When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot.Because the mean is sensitive to extreme values, a single outlier can substantially shift the average, potentially giving a misleading picture of the dataset.Recognizing this effect is essential for accurate data interpretation.Step 4: Look for signs of skewnessIf the data do not appear to be symmetric, does each sample show the same kind of asymmetry?Start by listing all the numbers in your dataset.Example: 3, 5, 7, 10, 12, 14, 18Now, put those numbers in order from smallest to largest. This makes it much easier to find the important values.The minimum is the smallest number in your list.The maximum is the largest number.In our example (3, 5, 7, 10, 12, 14, 18):Minimum = 3Maximum = 18The median is the middle number when your data is sorted.If you have an odd number of values, its the one right in the center.If you have an even number, you average the two middle numbers.In our example (3, 5, 7, 10, 12, 14, 18):There are 7 numbers, so the middle one is the 4th one.Median = 10Q1 is the median of the lower half of your data (the numbers below the overall median).Q3 is the median of the upper half of your data (the numbers above the overall median).Important: when finding Q1 and Q3, if there is an even amount of numbers in the upper or lower half, you must average the two center numbers.From our example:Lower half: 3, 5, 7 (below the median 10)Q1 is the middle of (3, 5, 7). Thats 5. So Q1 = 5.Upper half: 12, 14, 18 (above the median 10)Q3 is the middle of (12, 14, 18). Thats 14. So Q3 = 14.The IQR tells you how spread out the middle 50% of your data is.IQR = Q3 Q1For our example:Q3 = 14Q1 = 5IQR = 14 5 = 9Outliers are numbers that are very different from the rest.Lower Outlier Cutoff = Q1 (1.5 IQR)Upper Outlier Cutoff = Q3 + (1.5 IQR)Any numbers below the Lower Cutoff or above the Upper Cutoff are outliers.Using our example:Lower Cutoff = 5 (1.5 9) = 5 13.5 = 8.5Upper Cutoff = 14 + (1.5 9) = 14 + 13.5 = 27.5Since our minimum is 3 and our maximum is 18, we have no outliers.Draw a number line that covers the range of your data (from minimum to maximum).Mark Q1, the median, and Q3 above the number line to create the box.Draw the whiskers.The left whisker goes from Q1 to the smallest value within the lower outlier cutoff.The right whisker goes from Q3 to the largest value within the upper outlier cutoff.If you have outliers, plot them as individual points beyond the whiskers.Remember to use visual aids along with these steps. Always sort your data before finding quartiles. Use (Q3 Q1) for the IQR, then multiply by 1.5 to find the outlier cut-off. If an outlier exists, plot it as a separate point outside the whiskers. Label your axes carefully and include a title for clarity A researcher measures short-term memory test scores for 15 participants. Each participant completes a recall test scored out of 30 points. The data are: 23, 16, 19, 27, 14, 29, 22, 18, 12, 30, 25, 20, 15, 28, 17 Sort the data. Find the minimum, Q1, median, Q3, and maximum. Calculate the interquartile range (IQR). Determine if there are any outliers using the 1.5 IQR rule. Draw a boxplot of the data. Interpret your findings. Are the scores clustered, spread out, or skewed? A psychologist uses a stress scale ranging from 0 (no stress) to 40 (extreme stress). Fifteen patients reported the following scores: 6, 9, 15, 15, 18, 22, 5, 27, 32, 33, 10, 14, 16, 8, 40 Order the scores and calculate the median. Find Q1 and Q3, then compute the IQR. Determine lower and upper outlier boundaries (Q1 1.5 IQR, Q3 + 1.5 IQR). Draw a boxplot, labeling any outliers as individual points. Discuss any potential reasons for outliers on a stress scale (e.g., unique personal situations). A clinical psychologist measures anxiety in a group of clients before and after a short therapy program. Anxiety is scored from 0 to 50. The data are: Before: 10, 12, 15, 15, 24, 20, 28, 31, 35, 15 After: 7, 10, 10, 17, 20, 15, 21, 30, 28, 10 Construct a boxplot for the Before scores. Construct a separate boxplot for the After scores (on the same scale). Compare the distributions:Which group has a higher median? Which group has a wider spread (IQR)? Are there any outliers in either group? Interpret the difference in anxiety levels. Based on the boxplots, does therapy appear to have an effect?Lack of Detail About Distribution Shape: Although boxplots can show skewness to some extent (through asymmetry in the box or whiskers), they do not reveal whether a distribution is unimodal, bimodal, or has other distinct peaks. Limited Information on Data Density: A boxplot will not show you how many data points lie at different values within each quartile. Two datasets with the same boxplot could still have very different internal distributions. Context for Outliers: While outliers are marked, boxplots alone do not explain why those points are far from the bulk of the datafurther context or other visualizations might be necessary to understand anomalies.A boxplot visually summarizes the central tendency and spread of a dataset, much like a distribution curve. The center line of the box represents the median, while the box edges mark the first and third quartiles. The whiskers capture the extremes within 1.5 times the interquartile range, and any points beyond are outliers. This complements the smooth shape of a distribution curve, which shows a continuous view of how the datas values are spread.Quick Summary of Key Statistics: By displaying the median, interquartile range, and outliers at a glance, boxplots provide a concise snapshot of a datasets center and spread. Simple, Side-by-Side Comparisons: Boxplots make it easy to compare multiple data groups (e.g., different populations or experimental conditions) by aligning the boxes next to each other, highlighting differences in medians, overall spread, or outlier prevalence. Outlier Detection: Because boxplots explicitly plot values beyond 1.5 IQR (or another chosen threshold) as separate points, unusual or extreme values are immediately visible. Olivia Guy-Evans, MSc BSc (Hons) Psychology, MSc Psychology of Education Associate Editor for Simply Psychology Olivia Guy-Evans is a writer and associate editor for Simply Psychology. She has previously worked in healthcare and educational sectors. Saul McLeod, PhD Editor-in-Chief for Simply Psychology BSc (Hons) Psychology, MRes, PhD, University of Manchester Saul McLeod, PhD., is a qualified psychology teacher with over 18 years of experience in further and higher education. He has been published in peer-reviewed journals, including the Journal of Clinical Psychology. Share copy and redistribute the material in any medium or format for any purpose, even commercially. Adapt remix, transform, and build upon the material for any purpose, even commercially. The licensor cannot revoke these freedoms as long as you follow the license terms. Attribution You must give appropriate credit , provide a link to the license, and indicate if changes were made . You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. ShareAlike If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original. No additional restrictions You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits. You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation . No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material. A box plot is a type of plotthat displays the five number summary of a dataset, which includes:The minimum valueThe first quartile (the 25th percentile)The median valueThe third quartile (the 75th percentile)The maximum valueWe use three simple steps to create a box plot for any dataset:1. Draw a box from the first to the third quartile2. Draw a vertical line at the median3. Draw whiskers from the quartiles to the minimum and maximum valueWe typically create box plots in one of three scenarios:Scenario 1: To visualize the distribution of values in a dataset.A box plot allows us to quickly visualize the distribution of values in a dataset and see where the five number summary values are located.Scenario 2: To compare two or more distributions.Side-by-side box plots allow us to visualize the differences between two or more distributions and compare the median values and the spread of values between distributions.Scenario 3: To identify outliers.In box plots, outliers are typically represented by tiny circles that extend beyond either whisker. An observation is defined to be an outlier if it meets one of the following criteria:An observation is less than Q1 1.5*(Interquartile range)An observation is greater than Q3 + 1.5*(Interquartile range)By creating a box plot, we can quickly see whether or not a distribution has any outliers.The following examples show how we would use a box plot in each scenario.Scenario 1: Visualize the Distribution of Values in a DatasetSuppose a basketball coach wants to visualize the distribution of points scored by players on his team so he creates the following box plot:Based on this box plot, he can quickly see the following values:Minimum: 5Q1 (First Quartile): About 8Median: About 13Q3 (Third Quartile): About 18Maximum: 25This allows the coach to quickly see that the points scored by players ranges from 5 to 25, the median points scored is about 13, and 50% of his players score between about 8 and 18 points per game.Scenario 2: Compare Two or More DistributionsSuppose a sports analyst wants to compare the distribution of points scored by basketball players on three different teams so he creates the following box plots:Using these plots, he can quickly see that Team C has the highest median points scored and Team A has the lowest median points scored.He can also quickly see that Team B has the highest spread of points scored since the box plot for Team B has the longest box.Scenario 3: Identify OutliersSuppose a basketball coach wants to know if any of his players are outliers in terms of points scored. He decides to create the following box plot to visualize the distribution of points scored by his players:Using this plot, the coach can see that the tiny dot at the top of the plot indicates an outlier.Specifically, one of the players scored about 50 points which is considered an outlier compared to all of the other points scored.Additional ResourcesThe following tutorials offer in-depth explanations of how to use box plots in practice:How to Find the Interquartile Range (IQR) of a Box PlotHow to Identify Skewness in Box PlotsHow to Compare Box PlotsThe following tutorials explain how to create box plots in different statistical software:How to Make a Box Plot in Google SheetsHow to Create Box Plots in SPSSHow to Create Side-by-Side Box Plots in ExcelHow to Create Side-by-Side Box Plots in R The 5 values to be identified areThe minimum value the smallest number in the data setThe maximum value the largest number in the data setThe median - the middle value. 50% of the data is larger than this value and 50% of the data is smaller.The lower quartile is the value halfway between the minimum value and the median.75% of the data is larger than this value and 25% is lower than this value .The upper quartile is the value half-way between the median and the maximum value.25% of the data is larger than this value and 75% is lower than this value.The box has the lower quartile at one end and the upper quartile at the other.The median is always somewhere inside the box. The height of the box is not important.A box plot can be drawn from a data set using 5 key values.Minimum, Maximum, Median, Lower Quartile and Upper Quartile. A box plot (aka box and whisker plot) uses boxes and lines to depict the distributions of one or more groups of numeric data. Box limits indicate the range of the central 50% of the data, with a central line marking the median value. Lines extend from each box to capture the range of the remaining data, with dots placed past the line edges to indicate outliers. The example box plot above shows daily downloads for a fictional digital app, grouped together by month. From this plot, we can see that downloads increased gradually from about 75 per day in January to about 95 per day in August. There also appears to be a slight decrease in median downloads in November and December. Points show days with outlier download counts: there were two days in June and one day in October with low downloads compared to other days in the month. The box and whiskers plot provides a cleaner representation of the general trend of the data, compared to the equivalentline chart. Box plots are used to show distributions of numeric data values, especially when you want to compare them between multiple groups. They are built to provide high-level information at a glance, offering general information about a group of datas symmetry, skew, variance, and outliers. It is easy to see where the main bulk of the data is, and make that comparison between different groups. On the downside, a box plots simplicity also sets limitations on the density of data that it can show. With a box plot, we miss out on the ability to observe the detailed shape of distribution, such as if there are oddities in a distributionsmodality(number of humps or peaks) and skew. The datasets behind bothhistogramsgenerate the same box plot in the center panel. Construction of a box plot is based around a datasetsquartiles, or the values that divide the dataset into equal fourths. The first quartile (Q1) is greater than 25% of the data and less than the other 75%. The second quartile (Q2) sits in the middle, dividing the data in half. Q2 is also known as the median. The third quartile (Q3) is larger than 75% of the data, and smaller than the remaining 25%. In a box and whiskers plot, the ends of the box and its center line mark the locations of these three quartiles. The distance between Q3 and Q1 is known as the interquartile range (IQR) and plays a major part in how long the whiskers extending from the box are. Each whisker extends to the furthest data point in each wing that is within 1.5 times the IQR. Any data point further than that distance is considered an outlier, and is marked with a dot. There areother ways of defining the whisker lengths, which are discussed below. When a data distribution is symmetric, you can expect the median to be in the exact center of the box: the distance between Q1 and Q2 should be the same as between Q2 and Q3. Outliers should be evenly present on either side of the box. If a distribution is skewed, then the median will not be in the middle of the box, and instead off to the side. You may also find an imbalance in the whisker lengths, where one side is short with no outliers, and the other has a long tail with many more outliers. Visualization tools are usually capable of generating box plots from a column of raw, unaggregated data as an input; statistics for the box ends, whiskers, and outliers are automatically computed as part of the chart-creation process. When a box plot needs to be drawn for multiple groups, groups are usually indicated by a second column, such as in the table above. Box plots are at their best when a comparison in distributions needs to be performed between groups. They are compact in their summarization of data, and it is easy to compare groups through the box and whisker markings positions. It is less easy to justify a box plot when you only have one groups distribution to plot. Box plots offer only a high-level summary of the data and lack the ability to show the details of a data distributions shape. With only one group, we have the freedom to choose a more detailed chart type like ahistogramor a density curve. If the groups plotted in a box plot do not have an inherent order, then you should consider arranging them in an order that highlights patterns and insights. One common ordering for groups is to sort them by median value. As observed through this article, it is possible to align a box plot such that the boxes are placed vertically (with groups on the horizontal axis) or horizontally (with groups aligned vertically). The horizontal orientation can be a useful format when there are a lot of groups to plot, or if those group names are long. It also allows for the rendering of long category names without rotation or truncation. On the other hand, a vertical orientation can be a more natural format when the grouping variable is based on units of time. Certain visualization tools include options to encode additional statistical information into box plots. This is useful when the collected data represents sampled observations from a larger population. Notches are used to show the most likely values expected for the median when the data represents a sample. When a comparison is made between groups, you can tell if the difference between medians are statistically significant based on if their ranges overlap. If any of the notch areas overlap, then we cant say that the medians are statistically different; if they do not have overlap, then we can have good confidence that the true medians differ. This plot suggests that Process B creates components with better (higher) failure times, but the overlapping notches indicate the difference in medians is not statistically significant. Box width can be used as an indicator of how many data points fall into each group. Box width is often scaled to the square root of the number of data points, since the square root is proportional to the uncertainty (i.e. standard error) we have about true values. Since interpreting box width is not always intuitive, another alternative is to add an annotation with each group name to note how many points are in each group. There are multiple ways of defining the maximum length of the whiskers extending from the ends of the boxes in a box plot. As noted above, the traditional way of extending the whiskers is to the furthest data point within 1.5 times the IQR from each box end. Alternatively, you might place whisker markings at other percentiles of data, like how the box components sit at the 25th, 50th, and 75th percentiles. Common alternative whisker positions include the 9th and 91st percentiles, or the 2nd and 98th percentiles. These are based on theproperties of the normal distribution, relative to the three central quartiles. Under the normal distribution, the distance between the 9th and 25th (or 91st and 75th) percentiles should be about the same size as the distance between the 25th and 50th (or 50th and 75th) percentiles, while the distance between the 2nd and 25th (or 98th and 75th) percentiles should be about the same as the distance between the 25th and 75th percentiles. This can help aid the at-a-glance aspect of the box plot, to tell if data is symmetric or skewed. When one of these alternative whisker specifications is used, it is a good idea to note this on or near the plot to avoid confusion with the traditional whisker length formula. As developed byHofmann, Kafadar, and Wickham, letter-value plots are an extension of the standard box plot. Letter-value plots use multiple boxes to enclose increasingly-larger proportions of the dataset. The first box still covers the central 50%, and the second box extends from the first to cover half of the remaining area (75% overall, 12.5% left over on each end). The third box covers another half of the remaining area (87.5% overall, 6.25% left on each end), and so on until the procedure ends and the leftover points are marked as outliers. The letter-value plot is motivated by the fact that when more data is collected, more stable estimates of the tails can be made. In addition, more data points mean that more of them will be labeled as outliers, whether legitimately or not. While the letter-value plot is still somewhat lacking in showing some distributional details like modality, it can be a more thorough way of making comparisons between groups when a lot of data is available. As noted above, when you want to only plot the distribution of a single group, it is recommended that you use ahistogramrather than a box plot. While a histogram does not include direct indications of quartiles like a box plot, the additional information about distributional shape is often a worthy tradeoff. With two or more groups, multiple histograms can be stacked in a column like with a horizontal box plot. Note, however, that as more groups need to be plotted, it will become increasingly noisy and difficult to make out the shape of each groups histogram. In addition, the lack of statistical markings can make a comparison between groups trickier to perform. For these reasons, the box plots summarizations can be preferable for the purpose of drawing comparisons between groups. One alternative to the box plot is theviolin plot. In a violin plot, each groups distribution is indicated by a density curve. In a density curve, each data point does not fall into a single bin like in a histogram, but instead contributes a small volume of area to the total distribution. Violin plots are a compact way of comparing distributions between groups. Often, additional markings are added to the violin plot to also provide the standard box plot information, but this can make the resulting plot noisier to read. Depending on the visualization package you are using, the box plot may not be a basic chart type option available. Even when box plots can be created, advanced options like adding notches or changing whisker definitions are not always possible. However, even the simplest of box plots can still be a good way of quickly paring down to the essential elements to swiftly understand your data. The box plot is one of many different chart types that can be used for visualizing data. Learn more from our articles onessential chart types,how to choose a type of data visualization, or by browsing the full collection ofarticles in the charts category.

What are box plots used for in statistics. What is a box plot best used for. Box plot explained. What is a box plot. What is a box plot example. What are box plots used for.