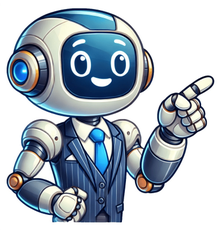
I'm not a bot



In the world of data analysis, outliers can significantly impact the insights and decisions derived from datasets. These anomalies can skew results, leading to inaccurate conclusions if not properly addressed. For analytics professionals and aspiring analysts, mastering the art of identifying and handling outliers is essential for producing reliable and robust analyses. This comprehensive guide will provide expert-level guidance on the topic, enriched with practical insights, real-world examples, and actionable advice. Imagine you're analyzing a dataset for customer purchase behavior. Most customers spend between $20 and $200 per transaction, but suddenly, you encounter a transaction of $10,000. This value is significantly different from the rest and is what we call an outlier. Ignoring such outliers can lead to skewed insights and misguided business decisions. In this article, we'll delve into what outliers are, why they occur, and how to effectively identify and handle them in data analysis. Outliers are data points that significantly differ from other observations in a dataset. They can be unusually high or low values that do not fit the general pattern of the data. Outliers can occur due to various reasons, such as measurement errors, data entry mistakes, or genuine anomalies in the data. Identifying these outliers is crucial as they can impact statistical analyses and models. Outliers can have a profound effect on statistical measures such as mean, standard deviation, and regression coefficients. They can: Skew Statistical Measures: Outliers can distort the mean, making it less representative of the dataset. Influence Analyses: In regression analysis, outliers can disproportionately affect the slope and intercept, leading to inaccurate models. Mask Patterns: Outliers can obscure underlying patterns in the data, making it challenging to detect trends and relationships. Affect Machine Learning Models: In machine learning, outliers can degrade model performance by influencing training and evaluation metrics. Scatter plots are a simple yet effective way to visually identify outliers. Plotting the data points on a graph can reveal any anomalies that stand out from the overall trend. Box plots, or whisker plots, display the distribution of data and highlight outliers as points outside the "whiskers." This method is useful for quickly identifying values that are significantly higher or lower than the rest. The Z-score measures how many standard deviations a data point is from the mean. A Z-score above 3 or below -3 is often considered an outlier. import numpy as np data = [10, 12, 12, 13, 12, 14, 14, 16, 18, 100] mean = np.mean(data) std = np.std(data) z_scores = [(x - mean) / std for x in data] outliers = [x for x, z in zip(data, z_scores) if np.abs(z) > 3] The IQR method identifies outliers as data points that fall below Q1 – 1.5 * IQR or above Q3 + 1.5 * IQR, where Q1 and Q3 are the first and third quartiles, respectively. Q1 = np.percentile(data, 25) Q3 = np.percentile(data, 75) IQR = Q3 - Q1 outliers = [x for x in data if x < Q1 - 1.5 * IQR or x > Q3 + 1.5 * IQR] Isolation Forest is an unsupervised machine learning algorithm designed to identify anomalies in data. It works by isolating observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. from sklearn.ensemble import IsolationForest data = np.array(data).reshape(-1, 1) clf = IsolationForest(contamination=0.1) outliers = clf.fit_predict(data) DBSCAN is a clustering algorithm that can identify outliers as points that do not belong to any cluster. from sklearn.cluster import DBSCAN data = np.array(data).reshape(-1, 1) clustering = DBSCAN(eps=3, min_samples=2).fit(data) labels = clustering.labels_ outliers = [data[i] for i in range(len(data)) if labels[i] == -1] Once identified, the next step is handling outliers. Here are some methods to consider: In some cases, it might be appropriate to remove outliers from the dataset. This is common when the outliers result from data entry errors or are not relevant to the analysis. Applying transformations, such as log or square root, can reduce the impact of outliers. This method is useful when the outliers are genuine but need to be scaled down to fit the analysis. Imputing outliers involves replacing them with more representative values, such as the mean or median of the data. This approach is suitable when the outliers are likely errors or when removing them would significantly reduce the dataset's size. Using robust statistical methods, such as median absolute deviation (MAD) or robust regression, can minimize the influence of outliers on the analysis. Capping involves setting a threshold to limit the maximum and minimum values in the data. This method is useful when outliers are extreme but not errors. In a retail dataset, most transactions fall between $20 and $200. However, a few transactions are significantly higher due to bulk purchases. Using the IQR method, these transactions can be identified and analyzed separately to understand customer behavior better. In a manufacturing process, sensor data might show occasional spikes due to malfunctioning equipment. Isolation Forest can be used to identify these anomalies, allowing for timely maintenance and preventing potential issues. Financial datasets often contain outliers due to fraud or errors. DBSCAN can help detect these anomalies, enabling further investigation and corrective actions. Understand the Context: Before removing or modifying outliers, understand the context and reason behind their occurrence. Document Your Process: Keep a detailed record of how outliers were identified and handled. This ensures transparency and reproducibility. Use Multiple Methods: Employ multiple techniques to identify outliers for a more comprehensive analysis. Analyze Impact: Assess how outliers affect your analysis and results before making any changes. Consult Domain Experts: Collaborate with domain experts to gain insights into the nature of outliers and the best approach to handle them. Blindly Removing Outliers: Removing outliers without understanding their context can lead to loss of valuable information. Ignoring Outliers: Failing to address outliers can skew results and lead to incorrect conclusions. Overfitting Models: Overfitting occurs when models are too complex and fit outliers too closely, reducing their generalizability. Underfitting Models: Underfitting happens when models are too simple and fail to capture important patterns in the data, including outliers. Outliers are an integral part of data analysis, and handling them effectively is crucial for accurate and reliable insights. By understanding what outliers are, why they matter, and how to identify and handle them, analytics professionals and aspiring analysts can enhance the quality of their analyses. Employing a combination of visual, statistical, and machine learning methods provides a robust approach to outlier detection and management. Remember to consider the context, document your process, and consult domain experts to make informed decisions. By following best practices and avoiding common pitfalls, you can ensure that your data analysis remains precise and trustworthy. The DBSCAN algorithm is a clustering method that can be used to detect outliers, especially in multi-dimensional data. from sklearn.cluster import DBSCAN import numpy as np # Sample 2D data X = np.array([[10, 30], [12, 32], [13, 35], [15, 36], [100, 150], [20, 40], [22, 42], [25, 45], [26, 48]]) # Fit DBSCAN model dbscan = DBSCAN(eps=3, min_samples=2) dbscan.fit(X) # Identify outliers (labeled as -1) outliers_dbscan = X[dbscan.labels_ == -1] print("Outliers using DBSCAN:", outliers_dbscan) 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. Data analysis involves scrutinizing datasets to derive meaningful insights, identify patterns, and support decision-making. Among the various concepts in data analysis, understanding outliers is crucial as they can significantly influence statistical calculations and the overall interpretation of data. This article delves into outliers, methods to describe data, ways to identify outliers, and the calculation of quartiles in datasets with odd and even numbers of observations. An outlier is an observation in a dataset that deviates markedly from the other observations. This deviation can be due to variability in the data, or it may indicate an error or a rare event. Outliers can be problematic because they can skew the results of an analysis, leading to misleading conclusions. Therefore, identifying and understanding outliers is essential for accurate data interpretation. Describing data effectively is crucial in various contexts, from scientific research to business analytics and beyond. How data is described can influence decisions, interpretations, and the overall understanding of its significance. Here are several key ways to represent data comprehensively and accurately: Contextual Background: Begin by providing a clear and concise data background. Explain where it comes from, its source, how it was collected, and any relevant details about the data generation process. This contextual information helps stakeholders understand the basis of the data and its potential limitations. Descriptive Statistics: Use descriptive statistics to summarize the main features of the dataset. This includes measures such as mean, median, mode, standard deviation, range, and percentiles. These statistics show the data's central tendency, dispersion, and distribution. Visual Representation: Present data visually using tools such as charts, graphs, and plots. Bar charts, histograms, scatter plots, and pie charts can convey patterns, trends, and relationships within the data that may not be immediately apparent from numerical descriptions alone. Data Distribution: Describe the distribution of the data points across various categories or intervals. Understanding whether the data is usually distributed, skewed, or exhibits other patterns is crucial for making informed decisions about analysis methods and interpretations. Data Quality: Assess and describe the quality of the data. This includes considerations such as completeness (whether all expected data points are present), accuracy (how closely the data reflects reality), consistency (whether data points are uniformly formatted), and relevance (how well the data aligns with the analysis objectives). Temporal Trends: If applicable, analyze and describe temporal trends in the data. Highlight changes over time, seasonal variations, or any other time-based patterns that may influence the interpretation of results. Correlations and Relationships: Explore correlations and relationships between different variables within the dataset. Use correlation coefficients, regression analysis, or other statistical methods to quantify and describe the strength and direction of relationships between variables. Outliers and Anomalies: Identify and describe any outliers or anomalies in the data. Explain their potential impact on analysis results and decision-making processes, and consider whether these outliers should be included, excluded, or investigated further. Data Interpretation: Provide interpretations and insights derived from the data analysis. Explain the implications of findings to the research question or business problem at hand. Offer recommendations or actions based on the data insights. Visualization Enhancement: Enhance data visualization with appropriate labels, titles, legends, and annotations to make the visual representation clear and meaningful. Ensure that the visual elements support rather than distract from the main message conveyed by the data. Clear Communication: Finally, communicate the described data effectively to the intended audience. Use language that is clear, concise, and accessible, avoiding jargon or technical terms that may not be familiar to all stakeholders. Identifying outliers in a dataset is an essential step in data analysis, as outliers can significantly impact the results and interpretations of statistical analyses. Outliers are data points that deviate markedly from other observations in the dataset. They may indicate variability in measurement, errors in data collection, or novel phenomena. Here's an elaborate look at how to identify outliers in a dataset: Box Plot (Box-and-Whisker Plot): A box plot displays the distribution of data based on a five-number summary: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. Outliers are typically plotted as individual points beyond the whiskers, which usually extend to 1.5 times the interquartile range (IQR) from the quartiles. Scatter Plot: Scatter plots can help identify outliers by displaying individual data points for two-dimensional data. Points that fall far away from the general cluster of data points can be considered outliers. Histogram: A histogram shows a dataset's frequency distribution. Outliers may appear as isolated bars at the extreme ends of the distribution. Z-Score: The Z-score measures how many standard deviations a data point is from the mean. Data points with a Z-score greater than 3 or less than -3 are often considered outliers. Interquartile Range (IQR): The IQR is the range between the first quartile (Q1) and third quartile (Q3). An outlier is defined as any value below Q1 - 1.5IQR or above Q3 + 1.5IQR. Modified Z-Score: For smaller datasets, the modified Z-score, which uses the median and median absolute deviation (MAD) rather than the mean and standard deviation, can be more effective. Isolation Forest: This algorithm works by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. Outliers are isolated more quickly than normal observations. DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN is a clustering method that identifies points in low-density regions as outliers. Autoencoders: In anomaly detection, autoencoders can be trained to reconstruct normal data points accurately, whereas outliers will have larger reconstruction errors. Arrange the Data: First, sort the data points in ascending order. Find the Median (Q2): The median is the middle value for a dataset with an odd number of data points. If the dataset has nnn data points, the median is the (n+1)/2(n+1)/2(n+1)/2th value. Determine the First Quartile (Q1): Q1 is the median of the lower half of the dataset, excluding the overall median. For an odd number of data points, the lower half includes all data points below the overall median. Find the median of this lower half to get Q1. Determine the Third Quartile (Q3): Q3 is the median of the upper half of the dataset, excluding the overall median. For an odd number of data points, the upper half includes all data points above the overall median. Find the median of this upper half to get Q3. Consider a dataset with 9 data points: 3,7,8,12,13,14,18,21,223, 7, 8, 12, 13, 14, 18, 21, 223,7,8,12,13,14,18,21,22 Step-by-Step Calculation Sort the Data (Already Sorted in This Example): 3,7,8,12,13,14,18,21,223, 7, 8, 12, 13, 14, 18, 21, 223,7,8,12,13,14,18,21,22 Find the Median (Q2): There are 9 data points, so n=9n = 9n=9. The median is the (9+1)/2=5(9+1)/2 = 5(9+1)/2=5th value. Median (Q2) = 13. Determine the Lower Half: The lower half includes: 3,7,8,123, 7, 8, 123,7,8,12 Number of data points in the lower half = 4. Find the First Quartile (Q1): The median of the lower half: There are 4 data points. The median of the lower half is the average of the 2nd and 3rd values. Q1 = (7 + 8) / 2 = 7.5. Determine the Upper Half: The upper half includes: 14,18,21,2214, 18, 21, 2214,18,21,22 Number of data points in the upper half = 4. Find the Third Quartile (Q3): The median of the upper half: There are 4 data points. The median of the upper half is the average of the 2nd and 3rd values. Q3 = (18 + 21) / 2 = 19.5. Summary of Quartiles for the Example Dataset First Quartile (Q1): 7.5 Median (Q2): 13 Third Quartile (Q3): 19.5 Enroll in the Post Graduate Program in Data Analytics to learn over a dozen of data analytics tools and skills, and gain access to masterclasses by Purdue faculty and IBM experts, exclusive hackathons, Ask Me Anything sessions by IBM. Arrange the Data: Sort the data points in ascending order. Find the Median (Q2): The median is the average of the two middle values for a dataset with an even number of data points. If the dataset has nnn data points, the median is the average of the n/2n/2n/2th and (n/2)+1(n/2) + 1(n/2)+1th values. Determine the First Quartile (Q1): Q1 is the median of the lower half of the dataset, including the overall median if the dataset is even. For an even number of data points, the lower half includes all data points below the median. Determine the Third Quartile (Q3): Q3 is the median of the upper half of the dataset, including the overall median if the dataset is even. For an even number of data points, the upper half includes all data points above the median. Consider a dataset with 10 data points: 2,4,5,7,10,12,14,18,21,232, 4, 5, 7, 10, 12, 14, 18, 21, 232,4,5,7,10,12,14,18,21,23 Step-by-Step Calculation Sort the Data (Already Sorted in This Example): 2,4,5,7,10,12,14,18,21,232, 4, 5, 7, 10, 12, 14, 18, 21, 232,4,5,7,10,12,14,18,21,23 Find the Median (Q2): There are 10 data points, so n=10n = 10n=10. The median is the average of the 5th and 6th values. Median (Q2) = (10 + 12) / 2 = 11. Determine the Lower Half: The lower half includes: 2,4,5,7,102, 4, 5, 7, 102,4,5,7,10 Find the First Quartile (Q1): The median of the lower half: There are 5 data points. The median of the lower half is the 3rd value. Q1 = 5. Determine the Upper Half: The upper half includes: 12,14,18,21,2312, 14, 18, 21, 2312,14,18,21,23 Find the Third Quartile (Q3): The median of the upper half: There are 5 data points. The median of the upper half is the 3rd value. Q3 = 18. Summary of Quartiles for the Example Dataset First Quartile (Q1): 5 Median (Q2): 11 Third Quartile (Q3): 18 Outliers are data points that significantly deviate from other observations in the dataset. They can result from measurement errors, data entry errors, or actual variability in the data. Example 1: Temperature Data Consider the temperature readings for a week in degrees Celsius: 22,23,21,24,30,22,23,4522, 23, 21, 24, 30, 22, 23, 4522,23,21,24,30,22,23,45 In this dataset, 45°C is an outlier because it is much higher than the other temperature readings. Example 2: Exam Scores Consider the exam scores of students out of 100: 55,60,62,65,70,75,80,85,90,92,95,3055, 60, 62, 65, 70, 75, 80, 85, 90, 92, 95, 3055,60,62,65,70,75,80,85,90,92,95,30 In this dataset, 30 is an outlier because it is significantly lower than the other scores. Example 3: Salary Data Consider the annual salaries of employees in a company (in thousands of dollars): 50,52,53,54,55,56,60,20050, 52, 53, 54, 55, 56, 60, 20050,52,53,54,55,56,60,200 In this dataset, 200 is an outlier because it is much higher than the other salaries. Understanding and calculating quartiles, whether in odd or even datasets, is essential for summarizing and analyzing data distributions. Quartiles provide a way to measure the spread and central tendency of data. Identifying outliers is crucial as they can significantly affect statistical analyses and interpretations. Various methods, including visual inspection, statistical techniques, and machine learning algorithms, can be employed to detect outliers. Properly handling outliers ensures the accuracy and reliability of data analysis, leading to more robust and meaningful conclusions. Enrolling in a Professional Certificate Program in Data Analytics and Generative AI can equip individuals with the skills needed to master these techniques and apply them effectively in real-world scenarios. Outliers can be identified in text data by analyzing unusual patterns, frequencies, or anomalies in word usage and context. Techniques such as Natural Language Processing (NLP) and text mining detect these outliers, which may indicate errors, unique events, or atypical content within the text. 2. How can outliers be handled in image processing applications? In image processing, outliers can be handled through filtering, thresholding, and anomaly detection algorithms. These methods help remove noise, enhance image quality, and identify unusual patterns or defects that may indicate errors or essential features in the image. 3. Can outliers provide valuable insights into unusual events? Yes, outliers can provide valuable insights into unusual events or rare occurrences that deviate from the norm. By analyzing these anomalies, organizations can detect fraud, identify unique opportunities, or uncover underlying issues that require attention, leading to more informed decision-making. 4. Can outliers be subjective based on the context of the analysis? Outliers can indeed be subjective based on the context of the analysis, as what is considered an outlier in one scenario may be expected in another. The definition of an outlier depends on the specific goals, data distribution, and domain-specific knowledge, making contextual understanding crucial for accurate outlier detection. 5. How do outliers affect the reliability of statistical analyses? Outliers can significantly affect the reliability of statistical analyses by skewing results, affecting measures of central tendency, and inflating variance. If not properly accounted for, they can lead to misleading conclusions, making it essential to identify and address outliers to ensure accurate and trustworthy analysis outcomes. Raw data is always murky; before it can be analyzed, it needs to be inspected, cleansed, and prepared. Data may be incomplete, inaccurate, inconsistent, or erratic. Outliers are a common anomaly in data analysis. An outlier occurs when the value of one or more data points falls unusually far from the values of most other data points. Outliers, if not corrected, can dramatically skew the outcomes of your analysis, causing you to make incorrect conclusions and/or predictions. In this post, we will explain why outliers occur; illustrate how outliers can impact your results; discuss how to detect outliers; and provide strategies commonly used to correct for outliers.Reasons Outliers OccurOutliers can occur for many reasons, with incorrect data entry being a likely culprit. Outliers can also occur because the sample we have drawn can have unique characteristics not common in other possible samples, or the population itself does not follow a normal distribution. And, sometimes, outliers just occur naturally.Let's assume a local bank wants to determine the average income of car loan applicants, so it randomly selects 10 applications, and pulls their stated incomes:Notice the last observation. All the other values range from $48,000 to $60,000, but the last observation is much greater: $154,000.It may well be that this applicant has an income of $154,000 per year; that would be a naturally occurring outlier. Or the applicant's income was just $54,000 and the data was entered erroneously as $154,000. What if the applicant included his/her income from all sources: wages, alimony, investment income, etc., and all the others just listed their wage income? That would be a measurement error. Or maybe the bank is in a remote area, and it's the only lender in the immediate vicinity, so everyone, regardless of income, applies there. Then the $154,000 income would not just be a naturally occurring outlier but would also be indicative of an outlier due to sampling and/or population distribution.How Outliers Impact ResultsIf outliers are not addressed, they can distort your analysis greatly and lead to incorrect findings and bad business decisions. Certain statistical measures are more sensitive to outliers than others. Outliers have a strong effect on the mean (average) and the standard deviation (which measures how spread out the data is), but little or none on the median (the middle value of the data set).If, for example, the last observation was supposed to be just $54,000, then the mean and median incomes of these 10 applicants would each be $54,000, and the standard deviation would be $3,916, suggesting that the applicant incomes are largely similar.However, if this $154,000 is indeed correct (or incorrect and uncorrected), the mean income would be $64,000 and the standard deviation would be $31,864, suggesting a considerable amount of spread. The median increases slightly, to $55,000. The fact that the mean income is much greater than the median income indicates the average is skewed.This skewed data can cause problems. If the bank relies solely on the mean income of the applicant in making decisions, it might design an auto loan promotion targeting high income households, which may then fall short because most of its applicants are lower income. On the flip side, when the bank must report its efforts in low income lending to regulators, the mean income alone will suggest that its borrowers have higher incomes than they actually do, and the bank could face fines and/or other repercussions for noncompliance.Outliers should rarely, if ever, be ignored.Detecting OutliersOutliers can be detected in several ways, one of which is simply calculating the descriptive statistics, which was done above. How spread out is the data? The range will indicate that: $48,000 to $154,000, in this example. But just because the range is wide doesn't mean the data has outliers. The next step is to look at the mean, median, and standard deviation. If the mean and the median are close, that's a good – but not reliable – indicator that there are no outliers. The standard deviation is a better indicator: if the standard deviation is large compared to the mean – or greater than the mean – outliers are likely present.Visual inspection of the data is another way to check for outliers. Visual inspection methods include histograms, box plots, and scatterplots. These methods are a little more difficult to explain and will be covered in a later blog post.Logical detection is another way to identify outliers. Simply put, does a particular value make logical sense? For example, if body temperatures were being recorded, a temperature of 130°F is illogical, since it would be terminal to the patient; it's more likely the patient's temperature is 103°F, and the numbers were transposed.Strategies for Handling OutliersOnce you've detected outliers, you need to decide how you will address them. You should first ask whether those data points are accurate. Check the original source of the data and correct any values that are incorrect.Your second consideration should be the data mining algorithm you intend to use. Association-based techniques, such as decision trees, are often robust when outliers are present, since they concentrate on the rank, rather than the actual value of the number. However, outliers can seriously disrupt neural networks and regression analyses.In the case of the decision tree algorithm, you likely don't need to do anything to the outliers. For the other algorithms, your third consideration will be how to adjust the outliers in some form or another; approaches include:Analyze the outlier(s) separately. It may well be that these outliers are truly exceptional. In the bank example, the applicant with the $154,000 annual income might be the wealthiest person in an area with a small population, so the bank should analyze the other 9 applicants for its main analysis and think of other ways to address its wealthier applicant.Eliminate the record.If outliers are significantly out of range, you might just exclude them from your analysis altogether, but this can bias both your sample and your results, especially if the sample is small. Although the sample in the bank example is small, eliminating the highest value introduces little bias since all the other values are clustered close together.Categorize, or bin, the values. This places values in ranges, so outliers will be classified appropriately: in the bank example above, the bank can create a binary variable, where a value of 1 can represent "More than $54,000" and a value of 0 can represent "$54,000 or less." In other examples, an analyst can break values into "low," "medium," or "high." Still, values can be binned "Less than $15,000," "$15,001-$30,000," and so on.Transform the outlying values. Transformation can take on many forms, including capping the value of the outlier to minimum or maximum value; predicting a likelier value using regression analysis or some other imputation method; substituting the average of all the other non-outlier values; or taking the natural log of all the values and using that in your analysis instead. Note, however, that transforming your variables, especially by using the natural log, can also transform your analysis.There is no perfect way to handle outliers; the process is as much an art as it is a science. The most effective strategy for handling outliers is contingent on your domain knowledge: how the data was collected and entered; your company's business rules; and the objectives of your analysis. Your domain knowledge can help you identify whether and why an outlier occurred; assess its impact; and determine the best approach to address it. Outliers are a regular occurrence in data sets and shouldn't be ignored, lest they distort your findings. Data Science Learn about Z-score and IQR methods for detecting outliers in data analysis. Understand their workings, strengths, and weaknesses. May 27, 2024 — 5 min read Outliers are data points that significantly deviate from the rest of the dataset, potentially indicating errors or unique phenomena worth investigating. Identifying outliers is crucial in data analysis as they can distort statistical analyses and lead to misleading conclusions. Two commonly used methods for detecting outliers are the Z-score method and the Interquartile Range (IQR) method. Each method has its strengths and weaknesses, and their applicability depends on the nature of the dataset.Z-ScoreThe Z-score is a measure that tells us how far away a data point is from the mean (average) of the data set, in terms of standard deviations. Here's how it works:Mean: The average of all data points.Standard Deviation: A measure of how spread out the data points are from the mean.Z-score: This is calculated for each data point and shows how many standard deviations that point is from the mean. The formula for the Z-score is:$Z=(X-)$Z=σ(X−μ)where XX is the data point, μ is the mean, and σ is the standard deviation.Outliers: In this method, any data point with a Z-score greater than 3 or less than -3 is considered an outlier. This means the point is more than 3 standard deviations away from the mean, which is quite rare.from scipy import stats import numpy as np import matplotlib.pyplot as plt # Sample data np.random.seed(42) data = np.random.normal(0, 5, 1000) # Calculating Z-scores z_scores = stats.zscore(data) outliers = np.where(np.abs(z_scores) > 3) # Plotting data with outliers highlighted plt.figure(figsize=(10, 5)) plt.plot(data, 'bo', label='Data') plt.plot(outliers[0], data[outliers], 'ro', label='Outliers') plt.title("Outliers using Z-score method") plt.legend() plt.show() print("Outliers using Z-score method:", data[outliers]) Strengths:Suitable for normally distributed data: This method is very effective for identifying outliers in data that follows a normal distribution. It considers the mean and standard deviation to identify outliers.Quantifies the extremity: The Z-score provides a standardized measure, making it clear how extreme a data point is relative to the mean.Weaknesses:Assumes normality: This method assumes that the data is normally distributed. If the data is highly skewed or has heavy tails, the Z-score method might not be appropriate.Sensitive to mean and standard deviation: Outliers can skew the mean and standard deviation, which can affect the Z-score calculation, especially in small datasets.IQR MethodThe Interquartile Range (IQR) method is a way to find outliers in a set of data. Outliers are data points that are much lower or higher than most of the other data points. Here's how the IQR method works:Quartiles:Q1 (First Quartile): This is the value below which 25% of the data points fall.Q3 (Third Quartile): This is the value below which 75% of the data points fall.IQR (Interquartile Range): This is the range between Q1 and Q3. It's calculated as:IQR=$Q3-Q1$IQR=Q3−Q1Outlier Boundaries:Lower Bound: This is calculated as $Q1-1.5\times IQR$Q1−1.5×IQR. Any data point below this value is considered an outlier.Upper Bound: This is calculated as $Q3+1.5\times IQR$Q3+1.5×IQR. Any data point above this value is considered an outlier.import numpy as np import matplotlib.pyplot as plt # Sample data np.random.seed(42) data = np.random.normal(0, 5, 1000) # Calculating Q1, Q3, and IQR Q1 = np.percentile(data, 25) Q3 = np.percentile(data, 75) IQR = Q3 - Q1 # Defining outlier bounds lower_bound = Q1 - 1.5 * IQR upper_bound = Q3 + 1.5 * IQR # Identifying outliers outliers = data[(data < lower_bound) | (data > upper_bound)] # Plotting data with outliers highlighted plt.figure(figsize=(10, 5)) plt.plot(data, 'bo', label='Data') plt.plot(np.where((data < lower_bound) | (data > upper_bound))[0], outliers, 'ro', label='Outliers') # Plotting upper and lower bounds plt.axhline(y=lower_bound, color='g', linestyle='--', label='Lower Bound') plt.axhline(y=upper_bound, color='r', linestyle='--', label='Upper Bound') plt.title("Outliers using IQR method") plt.legend() plt.show() print("Outliers using IQR method:", outliers) Strengths:Non-parametric: This method does not assume any specific distribution for the data (e.g., normal distribution). It relies on the median and quartiles, making it robust to skewed data and outliers.Simple to understand and implement: The concept of quartiles and the IQR is straightforward and easy to calculate.Effective for small sample sizes: It can be more effective with small datasets where extreme values can skew the mean and standard deviation.Weaknesses:May miss outliers in normally distributed data: It might not be as sensitive to outliers if the data is normally distributed, as it primarily focuses on the middle 50% of the data.Choosing the Right MethodData Distribution:IQR Method: Use this method if your data is not normally distributed or if you have no information about the underlying distribution.Z-score Method: Use this method if your data is approximately normally distributed.Sample Size:IQR Method: Can be more reliable with smaller sample sizes.Z-score Method: More appropriate for larger datasets where the mean and standard deviation are more stable.Sensitivity to Extremes:IQR Method: Less sensitive to extreme values and more robust to skewed data.Z-score Method: More sensitive to extreme values, providing a precise measure of how extreme a value is.Practical ExampleLet's consider practical scenarios:Skewed Data: If you are analyzing house prices in a city, where there might be a few extremely high values (luxury homes), the IQR method would be more suitable as it is not influenced by these extreme values.Normally Distributed Data: If you are analyzing heights of adult men, where the distribution is likely to be normal, the Z-score method would be effective in identifying unusually short or tall individuals.ConclusionBoth methods are useful tools in statistics for identifying outliers, and the choice of method depends on the characteristics of your data and the context of your analysis. Understanding the strengths and weaknesses of each method will help you make an informed decision on which method to use for your specific dataset. Your Step-by-Step Guide to Handling Outliers and Boosting Data AccuracyOutliers — those data points that just don't fit in — can be quite the troublemakers in your analysis. Imagine them as loud voices in a room that skew the conversation, pushing conclusions in directions that don't truly represent the data. □ If left unchecked, outliers can mess with your stats, throw off model predictions, and even hide important patterns you want to see. So, let's dive into understanding and handling these odd data points.1□ What Are Outliers? □An outlier is simply a data point that stands far apart from the others in your dataset. Picture a class average height being around 5 feet, but one student is 7 feet tall. That 7-footer? Definitely an outlier! Outliers can be the result of data entry errors or could reveal something uniquely important about your dataset.Example: Let's say you're analyzing monthly incomes in a neighborhood, and most people earn around ₹30,000 — ₹50,000. If one income shows up as ₹500,000, that's likely an outlier. Whether it's an error or a real unique case, it can throw off your calculations if you don't manage it.2□ Why Are Outliers Important? □ , the free encyclopedia that anyone can edit. 117,937 active editors 7,001,389 articles in English The English-language Wikipedia thanks its contributors for creating more than seven million articles! Learn how you can take part in the encyclopedia's continued improvement. GL Mk. II transmitter van Radar, Gun Laying, Mark I, or GL Mk. I for short, was an early World War II radar system developed by the British Army to provide information for anti-aircraft artillery. There were two upgrades, GL/EF (elevation finder) and GL Mk. II (pictured), both improving the ability to determine a target's bearing and elevation. GL refers to the radar's ability to direct the guns onto a target, known as gun laying. The first GL sets were developed in 1936 using separate transmitters and receivers mounted on gun carriages. Several were captured in 1940, leading the Germans to believe falsely that British radar was much less advanced than theirs. The GL/EF attachment provided bearing and elevation measurements accurate to about a degree: this caused the number of rounds needed to destroy an aircraft to fall to 4,100, a tenfold improvement over early-war results. The Mk. II, which was able to directly guide the guns, lowered the rounds-per-kill to 2,750. About 410 Mk. Is and 1,679 Mk. IIs were produced. (Full article...) Recently featured: Andrea Navagero Nosy Komba McDonnell Douglas Phantom in UK service Archive By email More featured articles About Lieke Klaver ahead in the women's 400 metres final ... that a 400-metre race in 2025 (pictured) was won by Lieke Klaver, who pretended that an absent competitor was running in front of her? ... that the land snail Drymaeus poecilus is notable for the striking variety of colors and patterns on its shell? ... that a forensic investigation of Signalgate has determined how a journalist was included in a group chat about Operation Rough Rider? ... that two of the players involved in the 2005 Vietnamese football match-fixing scandal did not accept payment because they felt ashamed? ... that a rebellion against a peace treaty with the Yuan dynasty operated out of the Historic Site of Anti-Mongolian Struggle on Jeju Island? ... that Nathan Frink fled the United States with enslaved children to settle in Canada, where he was elected as a Member of the Legislative Assembly and caught in a smuggling conspiracy? ... that Seattle's women's ice hockey team has an expected rival, despite not even having played their first game? ... that Cave Johnson Couts was separately acquitted for shooting his foreman, firing on funeral mourners, and whipping a native laborer to death? ... that characters' scars in an episode of The Last of Us were made with a paste-based appliance and a food mixer? Archive Start a new article Nominate an article Ngũgĩ wa Thiong'o Kenyan writer and activist Ngũgĩ wa Thiong'o (pictured) dies at the age of 87. In sumo, Ōnosato Daiki is promoted to yokozuna. In association football, Liverpool win the Premier League title. In motor racing, Álex Palou wins the Indianapolis 500. In basketball, the EuroLeague concludes with Fenerbahçe winning the Final Four Playoff. Ongoing: Gaza war M23 campaign Russian invasion of Ukraine timeline Sudanese civil war timeline Recent deaths: Phil Robertson Mary K. Gaillard Peter David Alan Yentob Gerry Connolly Sebastião Salgado Nominate an article May 30: Statehood Day in Croatia (1990) Johann Sebastian Bach 1431 – Hundred Years' War: After being convicted of heresy, Joan of Arc was burned at the stake in Rouen, France. 1723 – Johann Sebastian Bach (pictured) assumed the office of Thomaskantor in Leipzig, presenting the cantata Die Elenden sollen essen in St. Nicholas Church. 1922 – The Lincoln Memorial in Washington, D.C., featuring a sculpture of the sixteenth U.S. president Abraham Lincoln by Daniel Chester French, opened. 1963 – Buddhist crisis: A protest against pro-Catholic discrimination was held outside the National Assembly of South Vietnam in Saigon, the first open demonstration against President Ngô Đình Diệm. 2008 – The Convention on Cluster Munitions, prohibiting the use, transfer, and stockpiling of cluster bombs, was adopted. Ma Xifan (d. 947)Colin Blythe (b. 1879)Norris Bradbury (b. 1909)Wynonna Judd (b. 1964) More anniversaries: May 29 May 30 May 31 Archive By email List of days of the year About Seventeen performing "Oh My!" in 2018 South Korean boy band Seventeen made their debut on May 26, 2015, when they performed a showcase for their debut EP 17 Carat in front of a crowd of 1,000 people. Since then, the group have held 9 concert tours, 13 fan meetings, and have performed at a number of music festivals and awards shows. Their concert tours include the Right Here World Tour, which sold over one million tickets, and the Follow Tour, which was noted by Billboard as being the top grossing K-pop tour of 2023. In 2024, Seventeen made their first appearances at festivals in Europe, when they were the first South Korean act to perform at Glastonbury Festival's Pyramid Stage and as headliners for Lollapalooza Berlin. Seventeen's live performances are well regarded by fans and critics alike, and garnered them the award for Top K-pop Touring Artist at the 2024 Billboard Music Awards. (Full list...) Recently featured: Accolades received by Top Gun: Maverick National preserve 76th Primetime Emmy Awards Archive More featured lists Ignace Tonené (1840 or 1841 – 15 March 1916), also known as Nias or by his Ojibwe name Maiagizis ('right/correct sun'), was a Teme-Augama Anishnabai chief, fur trader, and gold prospector in Upper Canada. He was a prominent employee of the Hudson's Bay Company. Tonené was the elected deputy chief before being the lead chief and later the life chief of his community. In his role as deputy, he negotiated with the Canadian federal government and the Ontario provincial government, advocating for his community to receive annual financial support from both. His attempts to secure land reserves for his community were thwarted by the Ontario premier Oliver Mowat. Tonené's prospecting triggered a 1906 gold rush and the creation of Kerr Addison Mines Ltd., although one of his claims was stolen from him by white Canadian prospectors. This photograph shows Tonené in 1909. Photograph credit: William John Winter; restored by Adam Cuerden Recently featured: Australian white ibis Hell Gate Bridge Anemonoides blanda Archive More featured pictures Community portal – The central hub for editors, with resources, links, tasks, and announcements. Village pump – Forum for discussions about Wikipedia itself, including policies and technical issues. Site news – Sources of news about Wikipedia and the broader Wikimedia movement. Teahouse – Ask basic questions about using or editing Wikipedia. Help desk – Ask questions about using or editing Wikipedia. Reference desk – Ask research questions about encyclopedic topics. Content portals – A unique way to navigate the encyclopedia. Wikipedia is written by volunteer editors and hosted by the Wikimedia Foundation, a non-profit organization that also hosts a range of other volunteer projects: CommonsFree media repository MediaWikiWiki software development Meta-WikiWikimedia project coordination WikibooksFree textbooks and manuals WikidataFree knowledge base WikinewsFree-content news WikiquoteCollection of quotations WikisourceFree-content library WikispeciesDirectory of species WikiversityFree learning tools WikivoyageFree travel guide WiktionaryDictionary and thesaurus This Wikipedia is written in English. Many other Wikipedias are available; some of the largest are listed below. 1,000,000+ articles العربية Deutsch Español فارسی Français Italiano Nederlands 日本語 Polski Português Русский Svenska Українська Tiếng Việt 中文 250,000+ articles Bahasa Indonesia Bahasa Melayu Bân-lâm-gú Български Català Čeština Dansk Eesti Ελληνικά Esperanto Euskara עברית Հայերեն Magyar Norsk bokmål Română Simple English Slovenčina Srpski Srpskohrvatski Suomi Türkçe Oʻzbekcha 50,000+ articles Asturianu Azərbaycanca □□□□□ Bosanski کوردی Frysk Gaeilge Galego Hrvatski ქართული Kurdî Latviešu Lietuvių □□□□□□ Македонски □□□□□□□□□□ Norsk nynorsk □□□□□□ Shqip Slovenščina □□□ □□□□□□ اردو Retrieved from " 2EP by Seventeen 17 CaratEP by SeventeenReleasedMay 29, 2015 (2015-05-29)GenreK-popdance-pophip hopLength16:48LanguageKoreanLabelPledis EntertainmentLOEN EntertainmentSeventeen chronology 17 Carat(2015) Boys Be(2015) Singles from 17 Carat "Adore U"Released: May 29, 2015 17 Carat is the debut extended play (EP) by South Korean boy group Seventeen. It was released on May 29, 2015, by Pledis Entertainment and distributed by LOEN Entertainment. "Adore U" serves as the lead single for the EP. 17 Carat features five tracks written, co-written, and co-produced by Seventeen's group members. "Adore U" was chosen as the lead single for the EP and was performed on multiple music shows by the group. "Shining Diamond" was used as a pre-single on the group's reality debut show. The group stated that the tracklist was chosen to reflect Seventeen's core concept of "boys' passion".[1] The album has two physical versions: one with a "black" themed photo card set, and the other with a "white" themed photo card set. All copies include a CD containing the songs and a fold-up poster/lyric sheet. "Adore U" is the lead single of the extended play. It was written by Woozi, S.Coups, and Yeon Dong-geon.[2] The Korea Herald states "'Adore U' is a funky pop song about a teenage boy trying to navigate through puppy love."[3] It marks the beginning of the group's trilogy composed of the singles Adore U, Mansae, and Pretty U about a boy meeting, falling in love and asking out a girl. The track was composed and arranged by Woozi, Bumzu, and Yeon Dong-geon. The music video for the single was released on May 29, 2015, and was directed by Dee Shin. The dance choreography accompaniment to the song was choreographed by Hoshi and focuses on "storytelling, and on highlighting each member's strengths onstage".[4] The single has sold more than 38,000 digital copies and peaked at number 13 on the Billboard US World Chart. The EP has sold over 82,972 copies in South Korea.[5] It peaked at number 4 on the Korean Gaon Album Chart[6] and number 8 on the US World Billboard Chart.[7] Year-end lists Critic/publication List Rank Ref. Billboard The 10 Best K-pop Album of 2015 Placed [8] Hoshi participated in the choreography of "Adore U" and "Shining Diamond", Dino choreographed "Jam Jam".[9] Official track list[10]No.TitleLyricsMusicArrangementsLength1."Shining Diamond"WooziVernonS.CoupsKim Min-jeongWooziMasterKeyRishiMasterKeyRishi3:242.""Adore U"" (; Akkinda)WooziVernonS.CoupsBumzuWooziBumzuYeon Dong-geonWooziBumzuYeon Dong-geon3:073."Ah Yeah" (Hip-Hop unit)S. CoupsVernonWonwooMingyuCream DoughnutRishiCream DoughnutRishi3:294."Jam Jam" (Performance unit + Vernon)WooziHoshiDinoVernonWooziCream DoughnutCream Doughnut3:255."20" (Vocal unit)WooziWoozi Won Yeong-heonWon Yeong-heonDong Ne-hyeong3:23 Weekly chart performance for 17 Carat Chart (2015-2023) Peakposition Japanese Albums (Oricon)[11] 46 South Korean Albums (Gaon)[12] 4 US World Albums (Billboard)[13] 8 Year-end chart performance for 17 Carat Chart (2015) Peakposition South Korean Albums (Gaon)[14] 47 ^ "Seventeen hopes to shine like diamonds with '17 Carat'". The Korea Herald. 26 May 2015. Retrieved 30 November 2016. ^ "Adore U". Color Coded Lyrics. 29 May 2015. Retrieved 29 November 2016. ^ "Seventeen hopes to shine like diamonds with '17 Carat'". The Korea Herald. 26 May 2015. Retrieved 30 November 2016. ^ "Seventeen hopes to shine like diamonds with '17 Carat'". The Korea Herald. 26 May 2015. Retrieved 30 November 2016. ^ Cumulative sales of 17 Carat: "2015 Album Chart". "2016 12 Album Chart". "2017 11 Album Chart". ^ "2015 Albums". Gaon Music Chart. Korea Music Content Industry Association. Archived from the original on September 10, 2016. Retrieved November 29, 2016. ^ "June 27, 2015". Billboard. Retrieved 29 November 2016. ^ Benjamin, Jeff; Oak, Jessica (December 12, 2015). "The 10 Best K-Pop Albums of 2015". Billboard. Archived from the original on September 18, 2021. Retrieved October 31, 2021. ^ , (18 June 2015). "[My Name] (3) - , , , | ".   (in Korean). The Korea Economic Daily. . Retrieved 18 July 2021. ^ "SEVENTEEN 1st Mini Album '17 CARAT'". ^ "週間 アルバムランキング 2023年07月10日付" [Weekly album ranking as of July 10, 2023]. Oricon News (in Japanese). Archived from the original on July 5, 2023. Retrieved February 18, 2024. ^ "2015 27 Album Chart". Gaon Chart (in Korean). Archived from the original on August 7, 2017. Retrieved February 18, 2024. ^ "Seventeen Chart History (World Albums)". Billboard. Retrieved February 17, 2024. ^ "2015 Album Chart". Gaon Chart (in Korean). Archived from the original on May 7, 2017. Retrieved February 17, 2024. Retrieved from " 3 The following pages link to 17 Carat External tools (link count transclusion count sorted list) · See help page for transcluding these entries Showing 50 items. View (previous 50 | next 50) (20 | 50 | 100 | 250 | 500)Main Page (links | edit) Pledis Entertainment discography (links | edit) List of 2015 albums (links | edit) 2015 in South Korean music (links | edit) Seventeen (South Korean band) (links | edit) S.Coups (links | edit) Vernon (rapper) (links | edit) Wonwoo (links | edit) List of awards and nominations received by Seventeen (links | edit) Seventeen discography (links | edit) Love & Letter (links | edit) Joshua (singer) (links | edit) Seventeen TV (links | edit) 17 carat (redirect page) (links | edit) Going Seventeen (links | edit) List of Seventeen live performances (links | edit) Teen, Age (links | edit) All (links | edit) Bumzu (links | edit) Boys Be (EP) (links | edit) You Make My Day (links | edit) You Made My Dawn (links | edit) Jun (Chinese entertainer) (links | edit) List of Stray Kids live performances (links | edit) The8 (links | edit) An Ode (links | edit) Seungkwan (links | edit) Fallin' Flower (links | edit) Heng:garæ (links | edit) Semicolon (EP) (links | edit) Your Choice (links | edit) Going Seventeen (web series) (links | edit) Not Alone (Seventeen song) (links | edit) Attacca (EP) (links | edit) Rock with You (Seventeen song) (links | edit) Woozi (links | edit) Hoshi (South Korean singer) (links | edit) Don't Wanna Cry (Seventeen song) (links | edit) Darl+ing (links | edit) Face the Sun (links | edit) Left & Right (Seventeen song) (links | edit) 24H (EP) (links | edit) We Make You (links | edit) Hot (Seventeen song) (links | edit) Dream (Seventeen EP) (links | edit) BSS (band) (links | edit) FML (EP) (links | edit) Super (Seventeen song) (links | edit) Always Yours (album) (links | edit) Seventeenth Heaven (links | edit) View (previous 50 | next 50) (20 | 50 | 100 | 250 | 500) Retrieved from " WhatLinksHere/17_Carat"