Data points that substantially differ from the bulk of the data in a dataset are called outliers. By distorting distributions, inflating or deflating statistics, decreasing the precision of models, and possibly pointing out inaccuracies in data processing or collecting, they can have a substantial effect on data analysis. Therefore, in order to guarantee the quality and dependability of data analysis, it is crucial to recognize and deal with outliers effectively (Smiti and A. 2020).

Smiti, A. (2020). A critical overview of outlier detection methods. *Computer Science Review*, *38*, 100306.

A critical overview of outlier detection methods

https://www.sciencedirect.com/science/article/abs/pii/S1574013720304068

This PC\TECNO SPARK 8\Internal shared storage\Android\media\com.whatsapp\WhatsApp\Media\WhatsApp Documents

A statistical method known as linear regression is used to determine the linear relationship

between one variable (the independent variable) and another (the dependent variable). Plotting the spots where minor, consistent changes in the X variable result in the Y variable would be one way to visualize this. The regression determines which line fits through those points the best. Although it can't establish a cause and effect link, it can be used to forecast future values based on trends. (James et.al 2023)

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Linear regression. In *An introduction to statistical learning: With applications in python* (pp. 69-134). Cham: Springer International Publishing.

https://link.springer.com/chapter/10.1007/978-3-031-38747-0\_3

Before using data for any kind of study, data cleaning is an essential step. If you attempted to construct a house on an unstable foundation, the outcome would not be trustworthy. In the same way, employing soiled data—that is, data that contains mistakes, inconsistencies, or missing information—can result in erroneous conclusions and, ultimately, poor decisions. By guaranteeing that the data is correct, consistent, and comprehensive, data cleaning provides you with a strong basis for reliable analysis and efficient decision-making. To put it another way, the secret to seeing the worth of your data is clean data. Chai and C (2020).  
  
 Chai, C. P. (2020). The importance of data cleaning: Three visualization examples. *Chance*, *33*(1), 4-9.

The Importance of Data Cleaning: Three Visualization Examples

https://www.tandfonline.com/doi/abs/10.1080/09332480.2020.1726112

Error plots in regression analysis are graphical tools used to diagnose how well your regression model fits the data and identify potential problems. They focus on visualizing the **residuals**, which are the difference between the actual values of your dependent variable and the values predicted by your model. Beale *et.al* (2020).

Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J., & Elston, D. A. (2020). Regression analysis of spatial data. *Ecology letters*, *13*(2), 246-264.

Regression analysis of spatial data

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1461-0248.2009.01422.x

This project explores the power of data exploration and analysis to transform raw business data into actionable insights. By leveraging tools like Tableau and Python, we can uncover hidden patterns, trends, and relationships within data sets. Rashed, T. (2022)

Rashed, T. (2022). *Exploring strategies to leverage open source data visualization platforms in developing big data visual analytics* (Doctoral dissertation, Colorado Technical University).

*Exploring Strategies to Leverage Open Source Data Visualization Platforms in Developing Big Data Visual Analytics* https://www.proquest.com/openview/d5198b7961230faba781044de7c17444/1?pq-origsite=gscholar&cbl=18750&diss=y

Forecasting involves predicting future events or trends by analyzing historical data and applying statistical methods. Its value lies in enabling informed decision-making, risk mitigation, opportunity identification, resource optimization, and effective planning (Xie, N. 2022).

Xie, N. (2022). A summary of grey forecasting models. *Grey Systems: Theory and Application*, *12*(4), 703-722.

Forecasting

https://www.emerald.com/insight/content/doi/10.1108/GS-06-2022-0066/full/html

One of the most significant statistical techniques is the study of linear regression. It looks at the linear relationship between one or more metric-scaled independent variables (also known as exogenous, explanatory, control, or p) and a metric-scaled dependent variable (also known as endogenous, explained, response, or predicted variable (Hope *et.al* 2020).

Hope, T. M. (2020). Linear regression. In *Machine learning* (pp. 67-81). Academic Press.

Machine Learning - Linear regression

https://www.sciencedirect.com/science/article/abs/pii/B9780128157398000043

In order to achieve quality data, data that has been collected from various sources may contain dirty data. Therefore, data should be cleaned before being placed into a warehouse. According to (Swapna *et.al* 2016) When the data mining query is applied to the cleaned data, exact results will be obtained

Swapna, S., Niranjan, P., Srinivas, B., & Swapna, R. (2016, March). Data cleaning for data quality. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 344-348). IEEE.

*Data cleaning for data quality*

https://ieeexplore.ieee.org/abstract/document/7724284

. According to Strauss, R. (2010), through the implem**ent**ation of targeted marketing and advertising campaigns, the organization may successfully draw in more customers, boost sales, and customize their approaches for particular branch locations. Increasing inventory levels can help the business improve its overall profit, especially in Branch B and C. Average spend and revenue can be increased by giving "Premium" clients priority and

Strauss, R. (2010). *Marketing planning by design: Systematic planning for successful marketing strategy*. John Wiley & Sons.

https://books.google.co.ke/books?hl=en&lr=&id=qNRb3cFDGWcC&oi=fnd&pg=PR9&dq=+implementation+of+targeted+marketing+and+advertising+campaigns,+the+organization+may+successfully+&ots=qdZT1RFXpo&sig=O2Y8D2dZpWkgq-8FynDuirkofEw&redir\_esc=y#v=onepage&q=implementation%20of%20targeted%20marketing%20and%20advertising%20campaigns%2C%20the%20organization%20may%20successfully&f=false

**Data Splitting:**

With our dataset well defined and complete, it is now possible to divide it wisely in regards to model creation. According to Nguyen *et.al* (2021) The accuracy of a machine learning model changes under different training and testing ratios. Basing on statistical knowledge a ratio of 70:30 for training and testing dataset is considered the best ratio for training and evaluating the models.

Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., ... & Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, *2021*, 1-15. *Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil*<https://www.hindawi.com/journals/mpe/2021/4832864/#conclusions>

One advantage of this is that randomized Search offers lesser processing time than Grid Search. It offers parameters, such as estimator, param\_distributions, scoring, n\_iter, cv. According to Elgeldawi *et.al* (2021), Random Search samples the search space and evaluates sets from a specified probability distribution. In brief, it is a technique in which random combinations of the hyperparameters are used to find the best solution for the model under consideration.

Elgeddawy, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021, December). Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. In *Informatics* (Vol. 8, No. 4, p. 79). Multidisciplinary Digital Publishing Institute. *Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis*<https://www.mdpi.com/2227-9709/8/4/79>

Missing values can skew analysis. There are two possible stages of missing values data extraction and data collecting. Improper handling of these missing information could cause a model to perform less well or become biased. Brownlee and J. (2020).

Brownlee, J. (2020). *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery.

Data Preparation for Machine Learning

https://books.google.co.ke/books?hl=en&lr=&id=uAPuDwAAQBAJ&oi=fnd&pg=PP1&dq=data+cleaning&ots=Cm0NAifTqY&sig=Y6ME5rFbziQ3uBfpIItXFUNzMw8&redir\_esc=y#v=onepage&q=data%20cleaning&f=false

In order to better comprehend the nature of the data, data analysts employ data visualization and statistical tools to explain dataset characterizations, such as size, amount, and correctness. This process is known as data exploration. This involves effectively obtaining knowledge from data, even when we are unsure of the precise information we need. As a stand-alone method of doing data analytics, data exploration enables us to swiftly identify the characteristics and patterns in the data sets you are working with (Idreos et.al 2015)

Idreos, S., Papaemmanouil, O., & Chaudhuri, S. (2015, May). Overview of data exploration techniques. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data* (pp. 277-281).

Overview of Data Exploration Techniques

https://dl.acm.org/doi/abs/10.1145/2723372.2731084

According to Boukerche et.al (2020) an excessively high or low data point in relation to the closest data point and the remaining nearby co-existing values in a data graph is called an outlier. Extreme values that significantly deviate from the general pattern of values in a dataset or graph are called outliers.

1

Boukerche, A., Zheng, L., & Alfandi, O. (2020). Outlier detection: Methods, models, and classification. *ACM Computing Surveys (CSUR)*, *53*(3), 1-37.

# Outlier Detection: Methods, Models, and Classification

https://dl.acm.org/doi/abs/10.1145/3381028

2

An excessively high or low data point in relation to the closest data point and the remaining nearby co-existing values in a data graph is called an outlier. Extreme values that significantly deviate from the general pattern of values in a dataset or graph are called outliers (Alghushairy et.al 2020)

Alghushairy, O., Alsini, R., Soule, T., & Ma, X. (2020). A review of local outlier factor algorithms for outlier detection in big data streams. *Big Data and Cognitive Computing*, *5*(1), 1.

A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams

https://www.mdpi.com/2504-2289/5/1/1

CORRELATION

In the field of data exploration, knowing how variables relate to one another is essential. To meet this problem, correlation analysis offers a statistical method to measure the direction and strength of a linear relationship between two variables (Diba *et.al* 2020).

Diba, K., Batoulis, K., Weidlich, M., & Weske, M. (2020). Extraction, correlation, and abstraction of event data for process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1346.

# Extraction, correlation, and abstraction of event data for process mining

https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1346

Data, the fuel that propels countless endeavors in today's world, holds immense potential for generating valuable insights. However, the quality of this data directly influences the reliability of those insights. Data riddled with inconsistencies, inaccuracies, and missing values can lead to skewed results, misleading trends, and ultimately, faulty conclusions Teh *et.al* (2020).

Teh, H. Y., Kempa-Liehr, A. W., & Wang, K. I. K. (2020). Sensor data quality: A systematic review. *Journal of Big Data*, *7*(1), 11.

Sensor data quality: a systematic review

https://link.springer.com/article/10.1186/s40537-020-0285-1

As mentioned by Rusdah *et.al* (2020) while imputation fairly fills in the gaps left by missing values, data validation guarantees that predetermined standards are followed. Data is prepared for models by normalization, encoding, and deduplication, which gets rid of duplicates. When combined, these methods improve data reliability to the point where it is essential for precise machine learning results.

Rusdah, D. A., & Murfi, H. (2020). XGBoost in handling missing values for life insurance risk prediction. *SN Applied Sciences*, *2*, 1-10.

*XGBoost in handling missing values for life insurance risk prediction*

https://link.springer.com/article/10.1007/s42452-020-3128-y

. According to Olson, (2003). all data governance initiatives inside an organization depend on data quality, which quantifies how well a dataset satisfies requirements for completeness, precision, validity, uniformity, distinctiveness, accuracy, and fit for use.

Olson, J. E. (2003). *Data quality: the accuracy dimension*. Elsevier.

*Data Quality*

https://books.google.co.ke/books?hl=en&lr=&id=x8ahL57VOtcC&oi=fnd&pg=PP1&dq=data+quality+issues+and+remedies&ots=p-OwkkU-BZ&sig=MHaCKoVhSgLKPwtxf8Gd1OTI2yQ&redir\_esc=y#v=onepage&q=data%20quality%20issues%20and%20remedies&f=false

One of the most significant statistical techniques is the study of linear regression. It looks at the linear relationship between one or more metric-scaled independent variables (also known as exogenous, explanatory, control, or p) and a metric-scaled dependent variable (also known as endogenous, explained, response, or predicted variable). Skiera *et.al* (2021).

Skiera, B., Reiner, J., & Albers, S. (2021). Regression analysis. In *Handbook of market research* (pp. 299-327). Cham: Springer International Publishing.

Regression analysis

https://link.springer.com/referenceworkentry/10.1007/978-3-319-57413-4\_17

The process of turning unstructured data into useful features for machine learning models is known as feature engineering. In feature engineering, pertinent data points are chosen, existing features are combined to create new, informative ones, and the data is formatted so the model can readily grasp it. In order to construct precise models, this procedure is essential. Feature engineering helps machine learning models learn from the data and improve their predictions by producing clear and meaningful features. Behnke *et.al* (2021).

Behnke, M., Briner, N., Cullen, D., Schwerdtfeger, K., Warren, J., Basnet, R., & Doleck, T. (2021). Feature engineering and machine learning model comparison for malicious activity detection in the dns-over-https protocol. *IEEE Access*, *9*, 129902-129916.

# Feature Engineering and Machine Learning Model Comparison for Malicious Activity Detection in the DNS-Over-HTTPS Protocol

https://ieeexplore.ieee.org/abstract/document/9540699/

When preparing categorical data for machine learning models, one-hot encoding is essential. The original variable's distinct categories are converted into individual binary columns. A data point is assigned a value of 1 in the corresponding new column and a value of 0 in all other new columns if it falls into that category. This strategy has a number of benefits. First of all, unlike label encoding, which allocates numerical values, it refrains from giving the categories an artificial order. Second, it makes the distribution of the data across categories more clearly represented, which helps the model comprehend the connections between these categories and other numerical variables. All things considered, one-hot encoding is an effective method that aids machine learning models in understanding categorical data and eventually leads to to more accurate and effective predictions Yu *et.al* (2022)

Yu, L., Zhou, R., Chen, R., & Lai, K. K. (2022). Missing data preprocessing in credit classification: One-hot encoding or imputation?. *Emerging Markets Finance and Trade*, *58*(2), 472-482.

<https://www.tandfonline.com/doi/abs/10.1080/1540496X.2020.1825935>

# Missing Data Preprocessing in Credit Classification: One-Hot Encoding or Imputation?

Because they offer a rapid means of understanding what is regarded as the middle value of a given data set, measures of central tendency are crucial in data science. Like a bullseye, they provide you a quick idea of where most of the numbers in your data are located in general (Whatley, 2022).

Whatley, M. (2022). Measures of Central Tendency. In *Introduction to Quantitative Analysis for International Educators* (pp. 11-22). Cham: Springer International Publishing.https://link.springer.com/chapter/10.1007/978-3-030-93831-4\_2

Tableau's interactive dashboards, in contrast to static reports, turn data analysis from a passive process into an interactive investigation Prokofieva and M. (2021). With the help of these dynamic tools, users may find hidden correlations, filter information depending on their interests, and drill down into individual data points. Imagine being able to examine a thorough analysis of each product's sales within a category by hovering over a bar chart that represents that category of products.

Prokofieva, M. (2021). Using dashboards and data visualizations in teaching accounting. *Education and Information Technologies*, *26*(5), 5667-5683. Using dashboard and data visualisation in Tablaeu https://link.springer.com/article/10.1007/s10639-021-10525-z

Businesses can find significant patterns and trends by analyzing data sets. These trends may point to untapped markets, giving businesses the chance to modify their products and services to suit changing consumer demands (Kharakhash and O, 2023).

Kharakhash, О. (2023). Data visualization: transforming complex data into actionable insights. *Automation of technological and business processes*, *15*(2), 4-12.

https://journals.ontu.edu.ua/index.php/atbp/article/view/2520

**Tablaeu facilitates the interactive examination of trends and patterns by transformingintricate data sets into visually stunning dashboards and reports (KURIAN et al., 2022). Decision-makers are empowered to discern a coherent story from the data thanks to this approachable style, which enables them to make well-informed decisions grounded in a thorough comprehension of sales performance**

KURIAN, J., Bhalla, S., & Tuteja, S. (2022). A brief study on data visualization techniques and its implementation using tableau software. *NeuroQuantology*, *20*(12), 2841.*A brief study on data visualization techniques and its implementation using tableau software*

https://www.proquest.com/openview/60aff6036ca14a33bf0b4df3db208e44/1?pq-origsite=gscholar&cbl=2035897

Tableau turns you become a strategic leader guiding your sales team toward steady success Patel, A. (2021)

Patel, A. (2021). Data Visualization Using Tableau. Data visualisation for insights in tableau

https://www.theseus.fi/bitstream/handle/10024/652129/Patel\_Ashwin.pdf?sequence=4

.

**Step 4: Interactive Visualizations by Tableau**

Visuals are processed significantly more efficiently by the human brain than is plain data. This is where data visualization enters the picture, turning intricate datasets into comprehensible, visually striking images that convey a message. Data visualization provides a deeper understanding of information by exposing trends, patterns, and correlations that could otherwise go unnoticed by using charts, graphs, and other visual features Qin *et.al* (2020).

Qin, X., Luo, Y., Tang, N., & Li, G. (2020). Making data visualization more efficient and effective: a survey. *The VLDB Journal*, *29*(1), 93-117. *Making data visualisation more efficient*

https://link.springer.com/article/10.1007/s00778-019-00588-3