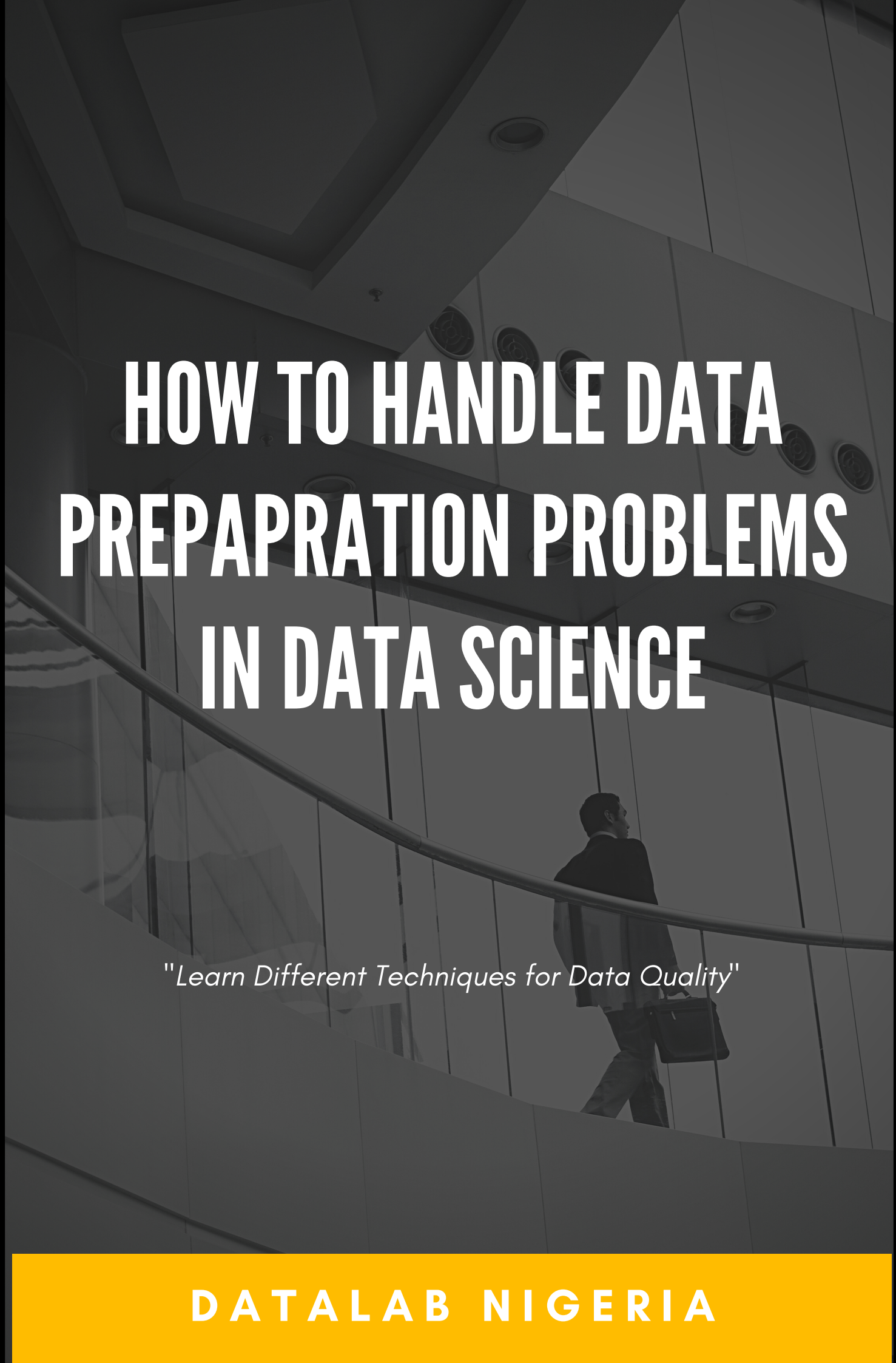


HOW TO HANDLE DATA PREPAPRATION PROBLEMS IN DATA SCIENCE

"Learn Different Techniques for Data Quality"

DATALAB NIGERIA



INTRODUCTION

Data Preparation is an important step in any data science project. It has been reported that a good Data Scientist spends up to 80% of time and resources in cleaning data and maintaining data quality in the pipeline. Data preparation comes with varying challenges which often makes the process time consuming. There is always negative impact on analysis done with an unclean data or data that has not been properly processed. Below are some popular issues in data preparation steps and how you can deal with them.

INSUFFICIENT DATA

Insufficient data can hinder the use of some machine learning algorithms that requires enough data for them to learn well. Such algorithms like Artificial Neural Networks (ANN) and other complex models in data science require enough data for deep exploration or deep learning. To overcome this challenge, you can:

- Go for simple algorithms like Logistics Regression, Naïve Bayes and Linear Regression when working with small or non-complex dataset.
- Tree Algorithms like Decision Tree and Random Forest can also work with small datasets provided you use a small number of trees to prevent overfitting on the small dataset
- Use Pre-trained Models if you must do deep learning with small dataset. Pre-trained models in ANN have been trained on a large sample of data and they can easily be adopted to learn from your small dataset using what they already know.
- You can argument your data by using data argumentation technique, like rotating an image to generate more samples of the same image.
- **You can do transfer learning by using elements of a pre-trained model in another model in solving related problem. In ANN, the last layer of the network can be taken and used as feature extractor in a different model that lacks data.**







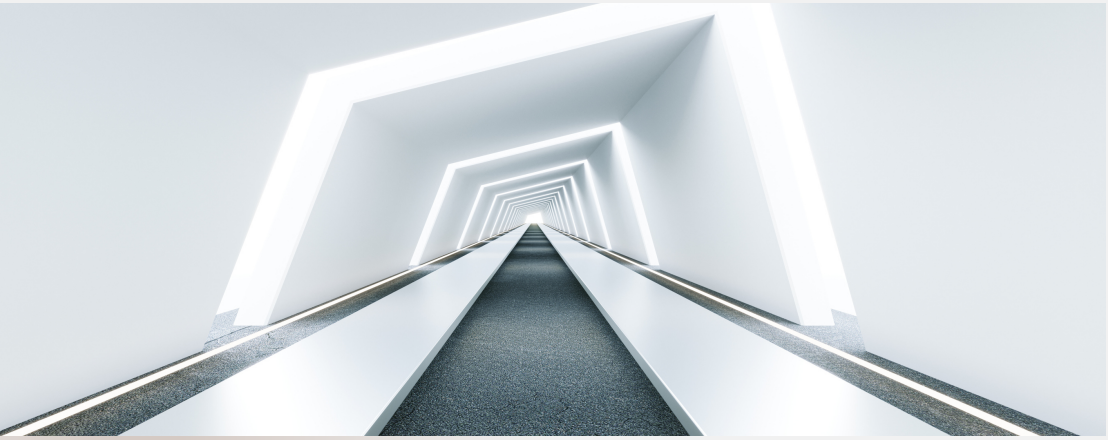
TOO MUCH DATA

It seems weird that enough data could also constituent problem in data preparation. Too much data could lead to multiple problems such as increasing computational time, difficulty in visualization, outdated historical data, redundant features, and curse of dimensionality in machine learning. To solve this problem, you can do the following:

- Feature Selection: Scale down your data to only relevant features using appropriate feature selection techniques such as variance threshold, statistical analysis such as Chi-square and Anova, embedded methods etc.

- Feature Engineering: Feature engineering allows you to combine and transform many low-level features into few useful ones. This reduces the size of your dataset. E.g., transforming latitudes and longitudes into a single feature such as Zip Code.

- Dimensionality Reduction: You can use different dimensionality reduction techniques such as Low Variance Filter, High Correlation filter, Random Forest Feature Importance, Backward Feature Elimination, Forward Feature Selection, Factor Analysis and Principal Component Analysis (PCA) to reduce the size of your dataset without you loosing much information in the dataset.



NON-REPRESENTATIVE

Another data quality issue is the problem of non-representative data. This can arise when the data you are working with is not a true representation of the real world. It can also arise when you are using outdated or static historic data that does not portray the dynamic changes in the real world. Human or algorithm bias in data gathering or generation can also lead to non-representative data. This can lead to biased models that perform poorly in practice. To overcome this problem, you can do the following:

- Mitigate non-representative data by using historic but near current data.
- Adopt mild oversampling or undersampling to reshuffle your data.
- Adopt simple or stratified random sampling from population.
- Excercise caution when using artificial or augmented data.



DUPLICATE DATA

Data duplicate can result in same set of data duplicated multiple times in your dataset. The approach in dealing with this is simple. You can just remove the duplicates or filter out rows with duplicate records.

- Interpolation from nearby values. This is useful if your values follow a certain pattern.
- Imputation by using machine learning algorithm like the K-Nearest Neighbour (KNN) Imputer to predict missing values in your data based on other column values. This is a good option.

MISSING DATA

Missing data is the most popular data quality problem in data science. There are simple to complex techniques for dealing with missing data as follows:

- Deletion is the easiest, but could drastically reduce your dataset, so not the best method. You could also loose many important information from the data.
- Use column mean, median, mode or frequently occurring values in a column. This method works if you already have a domain knowledge on the problem you are working on.
- Using mean is not always the best though because mean is highly sensitive to outliers if they exist in your dataset.

OUTLIER

Another common problem in data preparation is dealing with outliers. Outliers are data points that differs significantly from other observations. This could be due to variability of measurement and imputation error or bias. Other causes are sampling problems, unusual conditions and natural phenomena. The following techniques can be adopted in dealing with outliers:

- You can use distance from the mean or distance from the fitted line (regression) to identity potential outliers. Potential outliers can be set to the Mean.
- Data points that lie more than 3 standard deviation threshold from the mean can be termed as outliers. You can drop data above this threshold. This method is otherwise known as using Z Scores (the number of standard deviations an observation is from the column mean).

- You can use machine learning algorithm for anomaly or outlier detection and isolation like Isolation Forest and Density-based spatial clustering of applications with noise (DBSCAN).

- Inter-quartile range can also be used to detect and isolate outliers from your data.

- If your outliers are genuine or legitimate due to natural occurrences you can leave them if your model need to learn this, provided your model is not distorted.

- You can also cap your positive outliers to +3 standard deviation away from the mean and cap your negative outliers to -3 standard deviation away from the mean.

- You can use scaling techniques which are proven to be robust to outliers such as the robust standardization or robust scalar.

- You don't need to worry about the presence of outliers if you are going to use machine learning tree-based algorithms like Decision Tree, Random Forests, Gradient Boosting techniques. All these algorithms are robust to outliers and are less impacted by them.



START A CAREER IN DATA SCIENCE

Experience our Data School!



Building Brilliant Minds in Data Science and Artificial Intelligence

The Data Science internship is delivered on-site, online or hybrid. Our programs give you access to the best resources obtainable in world-class universities.

Get to Know our Programs

www.datalab.com.ng

NIG: +2348038518576

U.K: +447485644689