Effective Data Cleaning Techniques for Data Analysts.

Author name: B.Harika

1. ABSTRACT

Data cleaning can be very helpful for identifying errors, and correcting them, data can contain inconsistencies and inaccuracies while cleaning. Data cleaning plays a very important and crucial role in examining data. This process ensures that the data has been correctly and trustworthy analyzed, resulting in precise, accurate results and decision-making. This method becomes a challenging task while working on larger datasets, this becomes a difficult task because data cleaning takes time.

2. INTRODUCTION

Data cleaning is a very crucial step for data analysts, it is also known as data cleansing and data scrubbing. The possibility of having duplicates is high when data is combined with multiple datasets. When data is cleaned using the correct techniques and processes, it can result in high-quality data that can produce remarkable outcomes. This article aims to find the most effective data-cleaning steps and techniques.

3. Problem statement

The main issue arises when the data is inaccurate and contains several inaccuracies, missing values, and so on; these improper analytical skills and conclusions, which can lead to serious consequences. Hence, this demands for having a proper Data Cleaning process for ensuring accurate and reliable results, which will further help in decision-making.

4. Methods:

This article is collected using information from blogs, industry publications, and informational blogs. This article is mainly focused on effective methods and techniques used for data cleaning. Different methods have been used to clean the data like checking for duplicates, Null values, Handling missing values, and outliers detection.

Ⅰ. Checking for Duplicates:

Checking for duplicates is a very important step in order to remove all the unwanted observations from the dataset and duplicated observations. Duplicates in the data occur when there is a merge or join of multiple datasets, multiple duplicate entries, or scrape data. Duplicate data can lead to distractions and confuse the results, hence the removal of duplicates makes analysis more efficient.

ⅠⅠ. Detecting the Null Values:

Null values are the missing values in the datasets, null values occur due to different reasons such as incomplete data, missing data entry, or errors while collecting the data. Checking for the null values is very important in order to avoid problems during analysis. Handling the null values can be tricky, hence Effective handling of null values can improve the quality and reliability of data analysis and modeling.

ⅠⅠⅠ. Handling the missing values:

Missing values or null values depend on your analysis goal and what you want to achieve. Based on that, there are two major solutions for handling the missing values: Removing or deleting the null values which automatically removes the complete observations / the entire row of missing values. The second method effective method is Imputing the missing values using mean, median, and mode methods. The other methods are filling with zeros, and using forward and backward fills.

Ⅳ. Outliers Detection:

The importance for a data analyst is to find outliers and remove them from the dataset, the presence of outliers in the dataset can lead to lower predictive modeling performance. There are different techniques to find the outliers Z-Score, Percentile, IQR, and Visualization methods. Two ways to treat the outliers: Trimming and Capping, this technique is to cap the outliers of the dataset and limits by a particular value high or low then all values are considered outliers. Treating the outliers as missing values by using the imputing technique of missing values.

5. Results:

The article provides an overview of effective data-cleaning techniques, the methods include checking for duplicates, and removing the duplicated data, detecting the null values, handling the missing values, detecting, and treatment of the outliers. The importance of checking for duplicate values and removing them to impact the better understanding of data. Usage of different techniques to check for duplicates.

The importance of detecting the Null values and different methods to resolve the null values error and the need for solving the null values to improve the quality and reliability of data, 4 different functions are used for detecting the null values to option different forms of output and the solution led to the same outcome. Handling missing values has two different solutions: deleting the null values, imputing with mean, median, and mode methods, filling with zeros, and using backward and forward fill.

The highlight of detecting the outliers and treatment with different techniques such as Z-score, Percentile, IQR, and Visualization methods. Two main treatment methods are trimming, cap, and missing values by using imputing methods. Overall these insights into effective methods and techniques for data cleaning can help improve the quality and reliability of data analysis and modeling.

6. Conclusion:

A crucial step in ensuring the precision and dependability of data analysis and modeling is data cleansing. Data analysts can collect high-quality data that produces exact and accurate results by following efficient approaches like checking for duplicates, recognizing null values, addressing missing numbers, and identifying and treating outliers. Despite the difficulties involved in cleaning huge datasets, it is necessary that we dedicate the required time and resources to this process in order to make judgments based on reliable data.

7. References

[1] Inés Roldós ’’Effective Data Cleaning Techniques for Better Data” from MonkeyLearn. [online]
https://monkeylearn.com/blog/data-cleaning-techniques/ [Accessed: 03-04-2023]

[2] Tableau. “Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data” [online]
https://www.tableau.com/learn/articles/what-is-data-cleaning
[Accessed: 10-04-2023]

[3] Chirag Goyal “Outlier Detection & Removal | How to Detect & Remove Outliers” From Analytics Vidhya. [online]
https://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-python-code/ [Accessed: 20-04-2023]