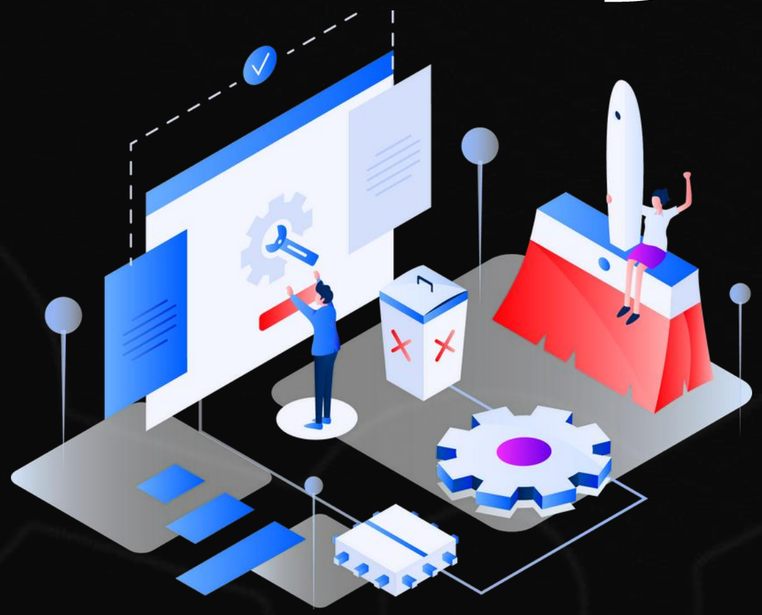
DATA CLEANING WITH

James Ehiabhi

James Ehiabhi

Why Data Cleaning??



Data Cleaning is the process of identifying and correcting errors, inconsistencies, and inaccuracies in datasets. The goal of data cleaning is to improve the quality of data by addressing issues such as missing values, duplicate entries, outliers, and formatting errors.

Accuracy of Analysis

Clean data ensures accurate and reliable analysis. If the dataset contains errors, inconsistencies, or missing values, any conclusions drawn from the analysis may be flawed or misleading.

Decision-Making

Businesses and organizations rely on data to make informed decisions. If the data used for decision-making is not clean, it can lead to poor choices, misguided strategies, and financial losses.

Modeling and Machine Learning

Accurate and reliable models depend on high-quality training data. Data cleaning is a crucial step before building models to ensure that the model is not influenced by noise, outliers, or inconsistencies in the data.

James Ehiabhi

Real Life Consequences Of Poor Data Cleaning Practices

Incident: Data Entry Errors at the Hawaii State Department of Taxation, The Hawaii State Department of Taxation encountered significant issues due to data entry errors.

Mistake: The department erroneously input tax credit data from taxpayers, leading to overpayments of tax refunds.

Cause: The errors originated from insufficient data cleaning procedures.

Failure: Data cleaning procedures couldn't identify and correct inaccuracies in tax data prior to refund processing.

Financial Impact: The errors resulted in millions of dollars in costs.

Real Life Consequences Of Poor Data Cleaning Practices

Incident: In 2012, the United Kingdom's National Health Service (NHS) faced a significant data handling error during a massive data transfer between the NHS and the Primary Care Trusts (PCTs) as part of NHS restructuring.

Mistake: Due to poor data cleaning practices, critical data fields were missing or improperly mapped, leading to errors in patient records, including misidentification of patients, missing medical histories, and inaccurate treatment information.

Cause: Insufficient data cleaning procedures were responsible for the errors.

Failure: Data cleaning procedures failed to detect and rectify inaccuracies in patient data before transfer.

Impact: The errors resulted in substantial financial and performance losses for the NHS, including costs for rectifying data inaccuracies, implementing data cleaning protocols, and addressing operational inefficiencies, amounting to millions of pounds.

PROCESSES IN DATA CLEANING

1

Handling Inaccuracies and Inconsistencies

Identify and correct inaccuracies in data entries, such as typos, incorrect values, formats, and inconsistent formatting.

2

Handling Missing Data

Identifying and addressing missing values in the dataset, either by imputing values or removing incomplete records.

3

Removing duplicates

Identifying and eliminating duplicate entries by comparing records. Remove duplicate records, keeping only one instance of each unique data point.

4

Handling outliers

Detect outliers or anomalous data points that deviate significantly from most of the data.

James Ehiabhi

As a Junior Data Professional at your organization, you are required by your Team Lead to deploy data cleaning techniques in handling inconsistencies in client’s project’s data. You are also required to provide a report on their respective enquiries.

Vector Inc.

- Clean the data
- Separate the customer's name into first name and last name
- Calculate the overall revenue, cost, and profit in $.
- Calculate the discount for each sales across the following discount levels 5%, 7%, and 10%

James Ehiabhi

As a Junior Data Professional at your organization, you are required by your Team Lead to deploy data cleaning techniques in handling inconsistencies in client's project's data. You are also required to provide a report on their respective enquiries.

Solutions-Savvy

- Clean the data and make it usable for further analysis
- Find the average age of founders in the dataset
- Separate the founder's name into First Name and Last Name
- Determine the gross sales value for each company and calculate the net sales value considering a VAT of 7.5% and a discount of 10%.
- What is the proportion of each company's revenue to the total revenue of the entire companies in the dataset, expressed as a percentage.

- VAT = Net Amount * %VAT
- Gross Sales = Net Amount + VAT
- Discount = Gross Sales * Discount
- Net Sales = Gross Sales – Discount
- Revenue Share = Gross Sales/Total Gross

James Ehiabhi

Basic Functions for Data Cleaning

- Text to column
- Trim function
- Substitute function
- Remove duplicates
- Flash fill
- Proper, Upper and Lower function
- Find and replace function
- Paste special