Assignment 1

Problem Statement:

Data Loading, Storage and File Formats
Analyzing Sales Data from Multiple File Formats

Dataset:

Sales data in multiple file formats (e.g., CSV, Excel, JSON)

Description:

The goal is to load and analyze sales data from different file formats, including CSV, Excel, and JSON, and perform data cleaning, transformation, and analysis on the dataset**.**

Tasks to Perform:

Obtain sales data files in various formats, such as CSV, Excel, and JSON.

1. Load the sales data from each file format into the appropriate data structures or dataframes.
2. Explore the structure and content of the loaded data, identifying any inconsistencies, missing values, or data quality issues.
3. Perform data cleaning operations, such as handling missing values, removing duplicates, or correcting inconsistencies.
4. Convert the data into a unified format, such as a common dataframe or data structure, to enable seamless analysis.
5. Perform data transformation tasks, such as merging multiple datasets, splitting columns, or deriving new variables.
6. Analyze the sales data by performing descriptive statistics, aggregating data by specific variables, or calculating metrics such as total sales, average order value, or product category distribution.
7. Create visualizations, such as bar plots, pie charts, or box plots, to represent the sales data and gain insights into sales trends, customer behavior, or product performance

Objective:

1. Consolidate sales data from various file formats (CSV, Excel, JSON) into a unified database.
2. Identify key sales metrics and trends across different products and regions.
3. Perform data cleansing to ensure accuracy and consistency of the sales data.
4. Generate visualizations and reports for better insights into sales performance.
5. Conduct comparative analysis of sales figures between different time periods and products.
6. Implement predictive modeling to forecast future sales based on historical data.
7. Explore correlations between sales and external factors (e.g., marketing campaigns, economic indicators).
8. Extract actionable insights to optimize sales strategies and improve revenue generation

Outcome:

1. Sales data from multiple file formats analyzed successfully for insights and trends.
2. Integration of diverse file formats streamlined, leading to comprehensive sales analysis.
3. Effortless extraction and processing of sales data from various formats achieved.
4. In-depth sales analysis completed by harmonizing data from multiple file types.
5. Cross-format data analysis facilitated clear understanding of sales performance.

Theory:

What is Data Cleaning in Data Visualization?

Data cleaning is the process of identifying and fixing incorrect data. It can be in incorrect format, duplicates, corrupt, inaccurate, incomplete, or irrelevant. Various fixes can be made to the data values representing incorrectness in the data.

Following eight common steps in the data cleaning process

1. Removing duplicates
2. Remove irrelevant data
3. Standardize capitalization
4. Convert data type
5. Handling outliers
6. Fix errors
7. Language Translation
8. Handle missing values

For example:
Consider data where we have the gender column. If the data is being filled manually, then there is a chance that the data column can contain records of 'male' 'female', 'M', 'F', 'Male', 'Female', 'MALE', 'FEMALE', etc. In such cases, while we perform analysis on the columns, all these values will be considered distinct. But in reality, 'Male', 'M', 'male', and 'MALE' refer to the same information. The data cleaning step will identify such incorrect formats and fix them.

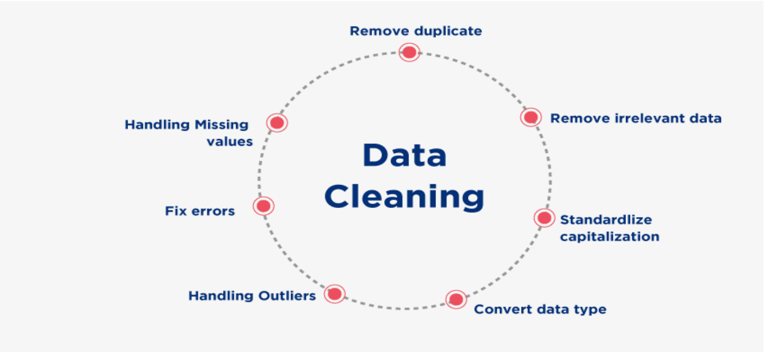

Fig Data cleaning

What is Data Transformation in Data Visualization?

Data transformation is the process of converting raw data into a format or structure that would be more suitable for model building and also data discovery in general. It is an imperative step in feature engineering that facilitates discovering insights.

When implementing supervised algorithms, training data and testing data need to be transformed in the same way. This is usually achieved by feeding the training dataset to building the data transformation algorithm and then apply that algorithm to the test set.

Data transformation is the process of converting, cleansing, and structuring data into a usable format that can be analyzed to support decision making processes, and to propel the growth of an organization. Data transformation is used when data needs to be converted to match that of the destination system. This can occur at two places of the data pipeline. First, organizations with on-site data storage use an extract, transform, load, with the data transformation taking place during the middle 'transform' step.

What is Data analysis?

It is the process of cleaning, changing, and processing raw data and extracting actionable, relevant information that helps businesses make informed decisions. The procedure helps reduce the risks inherent in decision-making by providing useful insights and statistics, often presented in charts, images, tables, and graphs.

A simple example of data analysis can be seen whenever we make a decision in our daily lives by evaluating what has happened in the past or what will happen if we make that decision. Basically, this is the process of analyzing the past or future and making a decision based on that analysis.

Algorithm:

1. Obtain sales data files in various formats, such as CSV, Excel, and JSON:
   - This task involves collecting sales data from different sources or systems, where the data might be stored in various formats like CSV, Excel spreadsheets, or JSON files.
2. Load the sales data from each file format into the appropriate data structures or dataframes:
   - After obtaining the data files, the next step is to load each file format into appropriate data structures or dataframes in a programming environment like Python or R. For instance, using libraries like pandas in Python to read CSV and Excel files, and json library for JSON files.
3. Explore the structure and content of the loaded data, identifying any inconsistencies, missing values, or data quality issues:

- Once the data is loaded, it's essential to examine the structure and content of the data. This involves checking the data types, column names, and detecting any inconsistencies, missing values, or data quality issues that might hinder the analysis.

4. Perform data cleaning operations, such as handling missing values, removing duplicates, or correcting inconsistencies:

   - Data cleaning is crucial to ensure data accuracy and reliability. This step involves handling missing values by imputing them or removing rows with missing values, identifying and removing duplicate entries, and correcting any inconsistencies or errors in the data.

5. Convert the data into a unified format, such as a common dataframe or data structure, to enable seamless analysis:

   - Since the data might have been loaded from various formats, converting it into a unified format (e.g., a single dataframe) makes it easier to perform analysis tasks without worrying about the original format.

6. Perform data transformation tasks, such as merging multiple datasets, splitting columns, or deriving new variables:

   - Data transformation involves reshaping and manipulating the data to extract valuable insights. Tasks like merging multiple datasets together, splitting columns, or creating new variables based on existing ones are performed in this step.

7. Analyze the sales data by performing descriptive statistics, aggregating data by specific variables, or calculating metrics such as total sales, average order value, or product category distribution:

   - The heart of the analysis lies in this step. Descriptive statistics such as mean, median, and standard deviation are calculated to understand the central tendencies and variabilities in the data. Data can also be aggregated based on specific variables like time periods, regions, or product categories to gain a deeper understanding of sales performance. Metrics like total sales, average order value, or product category distribution are computed to evaluate performance.

8. Create visualizations, such as bar plots, pie charts, or box plots, to represent the sales data and gain insights into sales trends, customer behavior, or product performance:

   - Data visualizations are powerful tools to communicate findings effectively. Creating visualizations like bar plots to compare sales across different categories, pie charts to represent the proportion of sales for each product category, or box plots to identify sales outliers helps to gain insights into sales trends, customer behavior, or product performance.

By following these tasks systematically, businesses can gain valuable insights from their sales data, make informed decisions, and optimize their strategies accordingly.

Conclusion:

In this way, we can analyze Sales Data from Multiple File Formats successfully.

Oral Questions:

1. What is Data Visualization?
2. What are different methods are used in Visualization?
3. What is Data Analysis?
4. What is Data Cleaning and Data transformation?