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.

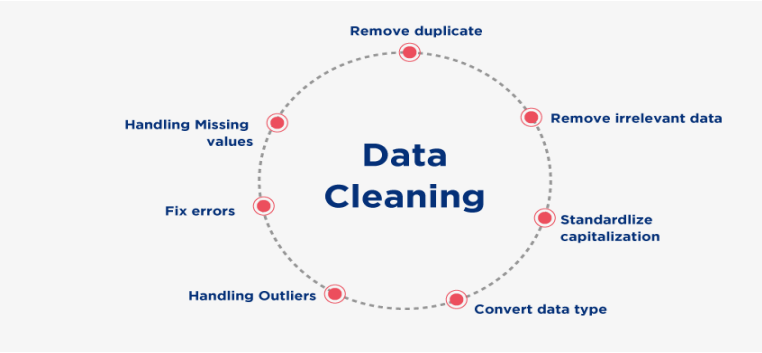

Fig Data cleaning

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?

Assignment 2

Problem Statement:

Interacting with Web APIs
Analyzing Weather Data from Open WeatherMap

API Dataset:

Weather data retrieved from Open WeatherMap

API Description:

The goal is to interact with the OpenWeatherMap API to retrieve weather data for a specific location and perform data modeling and visualization to analyze weather patterns over time.

Tasks to Perform:

1. Register and obtain API key from OpenWeatherMap.
2. Interact with the OpenWeatherMap API using the API key to retrieve weather data for a specific location.
3. Extract relevant weather attributes such as temperature, humidity, wind speed, and precipitation from the API response.
4. Clean and preprocess the retrieved data, handling missing values or inconsistent formats.
5. Perform data modeling to analyze weather patterns, such as calculating average temperature, maximum/minimum values, or trends over time.
6. Visualize the weather data using appropriate plots, such as line charts, bar plots, or scatter plots, to represent temperature changes, precipitation levels, or wind speed variations.
7. Apply data aggregation techniques to summarize weather statistics by specific time periods (e.g., daily, monthly, seasonal).
8. Incorporate geographical information, if available, to create maps or geospatial visualizations representing weather patterns across different locations.
9. Explore and visualize relationships between weather attributes, such as temperature and humidity, using correlation plots or heatmaps.

Objective:

1. The objective of analyzing weather data from OpenWeatherMap (or any other weather data source) can vary depending on the specific goals and requirements of the analysis. However, some common objectives of analyzing weather data include:
2. Weather Forecasting: One of the primary objectives is to use historical weather data to develop forecasting models. Analyzing past weather patterns and conditions can help improve the accuracy of short-term and long-term weather predictions, which is essential for various sectors such as agriculture, transportation, tourism, and emergency management.
3. Climate Studies: Weather data analysis can contribute to climate studies by examining long-term trends, identifying climate change patterns, and understanding variations in weather phenomena over extended periods. This information can be valuable for climate scientists, policymakers, and researchers concerned with environmental changes and their impacts.

4. Disaster Preparedness and Management: Analyzing weather data can aid in assessing the likelihood and severity of natural disasters like hurricanes, tornadoes, floods, and droughts. This information is crucial for disaster preparedness and management efforts, enabling communities to take proactive measures to reduce potential damages and protect lives.
5. Resource Planning: Various industries, such as energy production, agriculture, and water management, heavily rely on weather data for resource planning. By analyzing weather patterns, they can optimize resource allocation and make informed decisions to enhance efficiency and minimize risks.
6. Weather Pattern Analysis: Understanding weather patterns helps in identifying recurring climatic phenomena, which can have far-reaching effects on global weather systems. Analyzing these patterns can provide insights into their impacts on regional climates and ecosystems.
7. Health and Safety: Weather data analysis can play a role in public health and safety. Extreme weather conditions can have adverse effects on human health, such as heatwaves, cold snaps, or air pollution. By analyzing weather data, health authorities can issue timely advisories and take preventive measures.
8. Academic and Research Purposes: Weather data analysis serves as a valuable resource for academic and scientific research in meteorology, climatology, environmental science, and related fields. It contributes to a better understanding of the Earth's atmosphere and its complex interactions.
9. Urban Planning: Weather data analysis is essential for urban planning, as it helps identify areas prone to specific weather-related risks, such as flooding, urban heat islands, or air quality issues. This information can influence decisions on infrastructure development and city design.
10. Consumer Applications: Various consumer-oriented applications, such as weather apps and websites, rely on weather data analysis to provide accurate and up-to-date weather information to the public. It helps individuals make informed decisions about daily activities, travel plans, and outdoor events.

Theory:

What is OpenWeatherMap API?

OpenWeatherMap is a popular online service that provides weather data through its API (Application Programming Interface). The API allows developers to access and retrieve weather-related information programmatically, making it easier to integrate weather data into their applications, websites, and services.

Here are some key features and endpoints commonly available in the OpenWeatherMap API:

Current Weather Data: The API provides real-time weather data for a specific location, including information such as temperature, humidity, wind speed, weather conditions, and more.
Weather Forecast: OpenWeatherMap API offers weather forecasts for upcoming days, usually up to several days in advance. It provides forecasts for various time intervals, such as hourly or daily forecasts.

Historical Weather Data: Some plans or APIs might include access to historical weather data, allowing developers to analyze past weather conditions for specific locations.

Weather Maps: The API provides various weather map layers, such as temperature, precipitation, cloud cover, and more. These maps can be integrated into applications to visualize weather patterns.
UV Index: OpenWeatherMap provides information about the UV index, which indicates the strength of ultraviolet radiation from the sun.

Air Pollution Data: Some plans include air pollution data, such as the concentration of pollutants like PM2.5 and PM10, as well as pollutant indexes like AQI (Air Quality Index).
To use the OpenWeatherMap API, developers typically need to sign up for an API key, which is a unique identifier used to authenticate API requests and track usage limits. The API key is usually included in API requests as a parameter.

It's important to note that while OpenWeatherMap offers a free tier for limited usage, more extensive access to the API and additional features often require a paid subscription. Pricing and available features may vary, so developers should review the OpenWeatherMap website for the most up-to-date information on plans and usage terms.

How we can Interacting with Web APIs?

Interacting with Web APIs (Application Programming Interfaces) allows developers to access and exchange data between different software applications over the internet. Web APIs provide a standardized way for different systems to communicate and share information, enabling developers to integrate third-party services or access data from remote servers. Here's a general overview of how to interact with Web APIs:

1. Understanding API Documentation: Before using a Web API, it's essential to read its documentation thoroughly. API documentation provides information on the available endpoints, request methods (e.g., GET, POST, PUT, DELETE), required parameters, response formats, and authentication methods. The documentation is typically provided by the API provider and serves as a guide on how to use the API effectively.

2. Obtaining an API Key (if required): Some APIs require an API key for authentication. To get an API key, developers usually need to sign up on the API provider's website and generate a unique key tied to their account. The API key is often included in API requests as a parameter or in the request headers.
Choosing an HTTP Client Library: To interact with a Web API, developers can use various programming languages and HTTP client libraries.
Popular choices include:
Python: requests library
JavaScript: Fetch API or Axios (for Node.js)
Java: HttpURLConnection or OkHttp
PHP: cURL or Guzzle
Ruby: Net::HTTP or HTTParty
Select a library that suits your programming language and provides the necessary functionalities for making HTTP requests.

3. Making API Requests: Once you have selected an HTTP client library, you can start making API requests to the desired API endpoints. API requests are typically performed using HTTP methods like GET, POST, PUT, or DELETE. Depending on the API, you may need to include query parameters, request headers, and the API key (if required) in the request.
Handling API Responses: After sending an API request, the server will respond with the requested data or status information. API responses are usually in formats like JSON or XML. You'll need to parse the response data to extract the relevant information and handle any potential errors or status codes.

4.Rate Limiting and Best Practices: Many Web APIs have rate limits to prevent abuse and ensure fair usage. Review the API documentation for rate-limiting details and any other best practices recommended by the API provider to avoid potential issues.

5. Error Handling and Debugging: Implement robust error handling in your code to deal with potential issues such as network failures or incorrect API responses. Proper error handling ensures that your application can gracefully handle unexpected situations.

6.Testing and Debugging: During development, use tools like Postman, cURL, or browser-based tools to test API requests and responses. Debugging tools can help you troubleshoot and identify issues with your API interactions.

Conclusion:

By this way, we can Analyzing Weather Data from Open WeatherMap Successfully.

Oral Questions:

1. What is Open WeatherMap?
2. What is Web API?
3. How we can Interacting with Web APIs?

Assignment 3

Problem Statement:

Data Cleaning and Preparation

Problem Statement: Analyzing Customer Churn in a Telecommunications Company

Dataset: "Telecom_Customer_Churn.csv"

Description: The dataset contains information about customers of a telecommunications company and whether they have churned (i.e., discontinued their services). The dataset includes various attributes of the customers, such as their demographics, usage patterns, and account information. The goal is to perform data cleaning and preparation to gain insights into the factors that contribute to customer churn.

Tasks to Perform:

1. Import the "Telecom_Customer_Churn.csv" dataset.
2. Explore the dataset to understand its structure and content.
3. Handle missing values in the dataset, deciding on an appropriate strategy.
4. Remove any duplicate records from the dataset.
5. Check for inconsistent data, such as inconsistent formatting or spelling variations, and standardize it.
6. Convert columns to the correct data types as needed.
7. Identify and handle outliers in the data.
8. Perform feature engineering, creating new features that may be relevant to predicting customer churn. 9. Normalize or scale the data if necessary.
10. Split the dataset into training and testing sets for further analysis. 11. Export the cleaned dataset for future analysis or modeling.

Objective:

The objective of analyzing customer churn in a telecommunications company is to understand and reduce the rate at which customers are leaving or canceling their services. Customer churn, also known as customer attrition, refers to the number or percentage of customers who stop using a company's products or services during a specific period.

The primary objectives of analyzing customer churn in a telecommunications company are as follows: Identify Churn Patterns: Analyzing historical customer data helps identify patterns and trends that lead to churn. By understanding these patterns, the company can proactively address issues and implement strategies to retain customers.

1. **Predict Churn Probability**: Developing predictive models enables the company to forecast which customers are at a higher risk of churning. This helps in targeting retention efforts more effectively and allocating resources efficiently.

2. **Customer Segmentation:** Segmenting customers based on their behavior and characteristics can reveal insights into which groups are more likely to churn. This knowledge allows the company to tailor retention strategies to specific customer segments.

3. **Improve Customer Experience:** Analyzing churn reasons provides valuable feedback on the company's weaknesses and pain points experienced by customers. By addressing these issues, the company can improve overall customer experience and satisfaction, leading to higher retention rates.

4. **Optimize Marketing and Promotions:** Analyzing customer churn data can reveal which marketing strategies and promotions are most effective in retaining customers. It helps in allocating marketing budgets wisely and avoiding spending on ineffective campaigns.

5. **Leverage Customer Feedback:** Customer feedback and complaints data can be analyzed to identify recurring problems that drive customers away. Addressing these concerns can lead to a reduction in churn and improved customer loyalty.

6. **Calculate Customer Lifetime Value (CLV):** Understanding the CLV of different customer segments can help prioritize retention efforts on high-value customers who contribute more to the company's revenue.

7. **Competitor Analysis**: Analyzing churn rates in comparison to competitors can provide insights into the company's market position and competitiveness. Understanding why customers choose competitors' services can help in developing strategies to counteract these factors.

Overall, the goal of analyzing customer churn in a telecommunications company is to enhance customer retention, increase loyalty, and improve the company's bottom line by reducing the loss of valuable customers.

Theory:

What is Data preparation?

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. Key steps include collecting, cleaning, and labeling raw data into a form suitable for machine learning (ML) algorithms and then exploring and visualizing the data.

How do you prepare your data?

Data preparation follows a series of steps that starts with collecting the right data, followed by cleaning, labeling, and then validation and visualization. Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data, and combining datasets to enrich

data.
Data preparation is often a lengthy undertaking for data engineers or business users, but it is essential as a prerequisite to put data in context in order to turn it into insights and eliminate bias resulting from poor data quality.
For example, the data preparation process usually includes standardizing data formats, enriching source data, and/or removing outliers.

1.Collect data

Collecting data is the process of assembling all the data you need for ML. Data collection can be tedious because data resides in many data sources, including on laptops, in data warehouses, in the cloud, inside applications, and on devices. Finding ways to connect to different data sources can be challenging. Data volumes are also increasing exponentially, so there is a lot of data to search through. Additionally, data has vastly different formats and types depending on the source. For example, video data and tabular data are not easy to use together.

2.Clean data

Cleaning data corrects errors and fills in missing data as a step to ensure data quality. After you have clean data, you will need to transform it into a consistent, readable format. This process can include changing field formats like dates and currency, modifying naming conventions, and correcting values and units of measure so they are consistent.

3.Label data

Data labeling is the process of identifying raw data (images, text files, videos, and so on) and adding one or more meaningful and informative labels to provide context so an ML model can learn from it. For example, labels might indicate if a photo contains a bird or car, which words were mentioned in an audio recording, or if an X-ray discovered an irregularity. Data labeling is required for various use cases, including computer vision, natural language processing, and speech recognition.

4.Validate and visualize

After data is cleaned and labeled, ML teams often explore the data to make sure it is correct and ready for ML. Visualizations like histograms, scatter plots, box and whisker plots, line plots, and bar charts are all useful tools to confirm data is correct. Additionally, visualizations also help data science teams complete exploratory data analysis. This process uses visualizations to discover patterns, spot anomalies, test a hypothesis, or check assumptions. Exploratory data analysis does not require formal modeling; instead, data science teams can use visualizations to decipher the data.

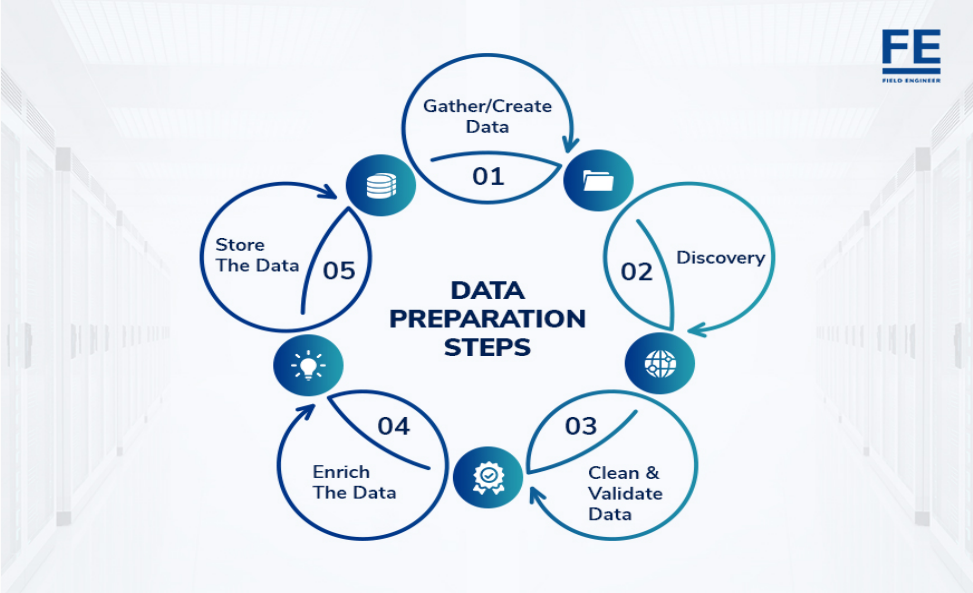

Fig. Data preparation steps

Algorithm:

Analyzing customer churn in a telecommunications company involves a systematic examination of customer data and relevant factors to understand why customers are leaving and to develop strategies for reducing churn.
Here's a step-by-step guide on how to approach customer churn analysis in a telecommunications company:

1) **Data Collection:** Gather relevant data about your customers, their interactions with the company, and their usage patterns. This data may include customer demographics, usage history, customer service interactions, billing information, contract details, and any other relevant data points.

2) **Define Churn:** Determine how churn is defined for your telecommunications company. Churn can be measured differently based on specific business needs and goals. For example, churn could be defined as customers who cancel their services, customers who have been inactive for a certain period, or customers who have downgraded their plans.

3) **Data Cleaning and Preprocessing:** Ensure that the collected data is clean, consistent, and ready for analysis. Handle missing values, outliers, and data inconsistencies appropriately. Preprocess the data to make it suitable for analysis.

4) **Exploratory Data Analysis (EDA):** Conduct exploratory data analysis to gain insights into the data. Explore churn rates over time, identify patterns, and look for correlations between churn and various customer attributes or behaviors. This step helps in understanding the current state of churn in the company.

5) **Feature Selection:** Identify relevant features or variables that may impact churn. These features could include customer demographics, usage patterns, customer service metrics, customer satisfaction scores, contract details, etc.

6) **Model Development:** Build predictive models to forecast churn probability for individual customers. Commonly used models include logistic regression, decision trees, random forests, and machine learning algorithms like XGBoost or support vector machines.

7) **Model Validation:** Validate the predictive models using appropriate techniques like cross-validation to ensure they are reliable and generalizable to new data.

8) **Interpretation of Results:** Analyze the model results to understand the factors that contribute most significantly to churn. Identify the key drivers of churn, such as poor customer service, high prices, competitive pressures, or service quality issues.

9) **Customer Segmentation:** Segment your customer base based on churn probability and other relevant characteristics. This segmentation helps tailor retention strategies to different customer groups effectively.

10) **Retention Strategies:** Develop targeted retention strategies based on the insights gained from the analysis. These strategies might include personalized offers, improved customer service, loyalty programs, and proactive outreach to at-risk customers.

11) **Implement and Monitor:** Put the retention strategies into action and closely monitor their effectiveness. Continuously track churn rates and customer feedback to assess the impact of the implemented strategies.

12) **Iterative Approach:** Customer churn analysis is an ongoing process. Regularly update the data and reevaluate the models and strategies to adapt to changing customer behavior and market conditions.
13) By following these steps, a telecommunications company can gain valuable insights into customer churn and take proactive steps to reduce churn rates, improve customer satisfaction, and foster long-term customer loyalty.

Conclusion:

In this way, we Analyzing Customer Churn in a Telecommunications Company successfully.

Oral Questions:

1) What is Data Preparation and cleaning?
2) How to perform analysis on dataset?

Assignment 4

Problem Statement:

Data Wrangling

Data Wrangling on Real Estate Market

Dataset: "RealEstate_Prices.csv"

Description: The dataset contains information about housing prices in a specific real estate market. It includes various attributes such as property characteristics, location, sale prices, and other relevant features. The goal is to perform data wrangling to gain insights into the factors influencing housing prices and prepare the dataset for further analysis or modeling.

Tasks to Perform:

1. Import the "RealEstate_Prices.csv" dataset. Clean column names by removing spaces, special characters, or renaming them for clarity.
2. Handle missing values in the dataset, deciding on an appropriate strategy (e.g., imputation or removal).
3. Perform data merging if additional datasets with relevant information are available (e.g., neighborhood demographics or nearby amenities).
4. Filter and subset the data based on specific criteria, such as a particular time period, property type, or location.
5. Handle categorical variables by encoding them appropriately (e.g., one-hot encoding or label encoding) for further analysis.
6. Aggregate the data to calculate summary statistics or derived metrics such as average sale prices by neighborhood or property type.
7. Identify and handle outliers or extreme values in the data that may affect the analysis or modeling process.

Objective:

The objective of data wrangling in the real estate market is to collect, clean, transform, and prepare raw real estate data to make it suitable for analysis and decision-making. Data wrangling, also known as data munging or data preprocessing, is a crucial step in the data analysis process. By performing data wrangling, analysts and stakeholders can gain valuable insights and make informed decisions in the real estate industry.

Here are some specific objectives of data wrangling in the real estate market:

1) **Data Collection:** Gather raw data from various sources, such as real estate listings, property databases, government agencies, real estate agents, and online platforms.
2) Data Cleaning: Identify and handle data quality issues, such as missing values, outliers, and

inconsistencies, to ensure the data is accurate and reliable.

3) **Data Transformation**: Convert data into a consistent format and structure to facilitate analysis. This may involve converting data types, standardizing units of measurement, and handling categorical variables.

4) **Data Integration:** Combine data from different sources and merge relevant information to create a comprehensive dataset for analysis.

5) **Feature Engineering:** Create new variables or features from existing data that might provide valuable insights. For example, calculating price per square foot, creating location-based features, or deriving property age from the construction year.

6) **Data Enrichment:** Augment the dataset with additional relevant information, such as demographic data, economic indicators, or market trends, to provide a broader context for analysis.

7) **Data Reduction:** If the dataset is too large or contains redundant information, data wrangling can involve reducing the data size while retaining its essential characteristics.

8) **Handling Missing Data:** Develop strategies to handle missing data points, such as imputation techniques or excluding records with missing values, based on the impact on the analysis.

9) **Data Visualization:** Generate visual representations of the data during the wrangling process to explore patterns, identify anomalies, and verify the effectiveness of cleaning and transformation steps.

10) **Data Documentation:** Maintain detailed documentation of the data wrangling process to ensure transparency, reproducibility, and collaboration among analysts and stakeholders.

11) By accomplishing these objectives, data wrangling empowers real estate professionals, investors, and policymakers to make better-informed decisions, identify market trends, understand property valuation, assess risk, and discover opportunities in the real estate market.

Theory:

What is Data Wrangling?

Data Wrangling is the process of gathering, collecting, and transforming Raw data into another format for better understanding, decision-making, accessing, and analysis in less time. Data wrangling is the practice of converting and then plotting data from one “raw” form into another.

Data Wrangling is also known as Data Munging, data cleansing, data scrubbing, data cleaning, or data remediation.

Data Wrangling in Python

Data Wrangling is a crucial topic for Data Science and Data Analysis. Pandas Framework of Python is used for Data Wrangling. Pandas is an open-source library in Python specifically developed for Data Analysis and Data Science. It is used for processes like data sorting or filtration, Data grouping, etc. Data wrangling in Python deals with the below functionalities:

1. **Data exploration:** In this process, the data is studied, analyzed, and understood by visualizing representations of data.
2. **Dealing with missing values:** Most of the datasets having a vast amount of data contain missing values of *NaN, they are needed to be taken* care of by replacing them with mean, mode, the most frequent value of the column, or simply by dropping the row having a *NaN* value.
3. **Reshaping data:** In this process, data is manipulated according to the requirements, where new data can be added or pre-existing data can be modified.
4. **Filtering data:** Some times datasets are comprised of unwanted rows or columns which are required to be removed or filtered

Fig: steps in data wrangling

There are 6 steps as follows:

1. **Data Discovery**: This is an all-encompassing term that describes understanding what your data is all about. In this first step, you get familiar with your data

2. **Data Structuring**: When you **collect** raw data, it initially is in all shapes and sizes, and has no definite structure. Such data needs to be restructured to suit the analytical model that your enterprise plans to deploy

3. **Data Cleaning**: Raw data comes with some errors that need to be fixed before data is passed on to the next stage. **Cleaning** involves the tackling of outliers, making corrections, or deleting bad data completely
4. **Data Enriching**: By this stage, you have kind of become familiar with the data in hand. Now is the time to ask yourself this question – do you need to embellish the raw data? Do you want to augment it with other data?

5. **Data Validating**: This activity surfaces data quality issues, and they have to be addressed with the necessary transformations. The rules of **validation** rules require repetitive programming steps to check the authenticity and the quality of your data

6. **Data Publishing**: Once all the above steps are completed, the final output of your data **wrangling efforts** is pushed downstream for your analytics needs

Algorithm:

Data wrangling is the process of gathering, cleaning, transforming, and organizing raw data into a format suitable for analysis. When it comes to the real estate market, data wrangling plays a crucial role in making the data usable and extracting valuable insights. Here's a step-by-step guide to data wrangling for the real estate market:

1) **Data Collection:**
   Identify the sources of real estate data you want to analyze. These could include public databases, real estate websites, APIs, or data provided by real estate agencies. Decide on the specific variables you need, such as property prices, location, property size, number of bedrooms, etc.
2) **Data Cleaning**:
   Remove any duplicate records from the dataset to ensure data accuracy. Handle missing values. You can either remove rows with missing values or use imputation techniques (e.g., mean, median, regression) to fill in missing data where appropriate. Check for and correct data entry errors or inconsistencies. Standardize data formats (e.g., converting dates to a uniform format) to facilitate analysis.
3) **Data Transformation:**
   Convert categorical variables to numerical representations using techniques like one-hot encoding or label encoding. Extract relevant features from the existing data. For example, you might extract the year from a date variable to create a separate "Year" feature. If you have unstructured data (e.g., property descriptions), consider using natural language processing (NLP) techniques to extract meaningful information.
4) **Data Integration:**
   If you have data from multiple sources, integrate them into a single dataset. Ensure that the

data formats are consistent. Merge datasets using common identifiers (e.g., property IDs) if you need to combine information from different sources.

5) **Data Exploration:**
   Perform exploratory data analysis (EDA) to understand the distribution and relationships of variables. Visualize the data through plots, histograms, scatter plots, and other relevant charts to identify patterns and outliers.
6) **Data Analysis:**
   Apply statistical methods and machine learning algorithms to extract insights from the data.
   Perform regression analysis to understand the relationship between property features and prices.
   Use clustering techniques to group similar properties together.
7) **Data Presentation:**
   Summarize your findings in a clear and understandable manner.Create visualizations and reports to present the results effectively.

Conclusion

In this way, we perform Data Wrangling on Real Estate Market .

Oral Questions:

1) What is Data Wrangling?
2) What are the different steps in data wrangling?
3) What is data cleaning?
4) How to perform Data Wrangling on Real Estate Market
5) What is data visualization?

Assignment 5

Problem Statement:

Data Visualization using matplotlib
Analyzing Air Quality Index (AQI) Trends in a City

Dataset: "City_Air_Quality.csv"

Description:
The dataset contains information about air quality measurements in a specific city over a period of time. It includes attributes such as date, time, pollutant levels (e.g., PM2.5, PM10, CO), and the Air Quality Index (AQI) values. The goal is to use the matplotlib library to create visualizations that effectively represent the AQI trends and patterns for different pollutants in the city.

Tasks to Perform:
1. Import the "City_Air_Quality.csv" dataset.
2. Explore the dataset to understand its structure and content.
3. Identify the relevant variables for visualizing AQI trends, such as date, pollutant levels, and AQI values.
4. Create line plots or time series plots to visualize the overall AQI trend over time.
5. Plot individual pollutant levels (e.g., PM2.5, PM10, CO) on separate line plots to visualize their trends over time.
6. Use bar plots or stacked bar plots to compare the AQI values across different dates or time periods.
7. Create box plots or violin plots to analyze the distribution of AQI values for different pollutant categories.
8. Use scatter plots or bubble charts to explore the relationship between AQI values and pollutant levels. 9. Customize the visualizations by adding labels, titles, legends, and appropriate color schemes.

Objective:

The objective of analyzing Air Quality Index (AQI) trends in a city is to gain valuable insights into the quality of the air over a specific period and to understand how it is changing over time. This analysis serves several important purposes:

1) **Assessing Public Health Impacts**: Air pollution can have significant effects on public health, leading to respiratory and cardiovascular diseases, among other issues. By analyzing AQI trends, health authorities can identify potential health risks and take appropriate measures to protect the well-being of the population.

2) **Environmental Monitoring**: Monitoring AQI trends helps in evaluating the effectiveness of pollution control measures and environmental regulations. It allows policymakers to determine if the implemented strategies are making a positive impact on air quality or if further actions are required.

3) **Identifying Pollution Hotspots**: AQI trends can highlight specific areas within the city that

consistently experience poor air quality. This information is crucial for targeting resources and implementing localized solutions to reduce pollution levels in those areas.

4) **Climate Change Considerations**: Air quality is interconnected with climate change, as certain air pollutants also contribute to greenhouse gas emissions. Analyzing AQI trends can help understand the linkages between air pollution and climate change, providing data for comprehensive environmental planning.

5) **Forecasting and Early Warnings**: By analyzing past AQI trends, it becomes easier to develop accurate air quality forecasting models. This information can be used to issue early warnings to citizens, particularly those sensitive to poor air quality, so they can take precautions and avoid exposure during periods of high pollution.

6) **Public Awareness and Engagement**: Sharing AQI trend data with the public raises awareness about air quality issues and encourages individuals to adopt more environmentally friendly practices. It empowers citizens to take actions that collectively contribute to improving air quality in their city.
7) **Comparative Studies**: Comparing AQI trends between cities can provide valuable insights into the effectiveness of air pollution control strategies implemented in different regions. This facilitates the exchange of best practices and encourages collaborative efforts in addressing air quality challenges on a global scale.

THEORY:

Data Visualization

Data visualization is the graphical representation of information and data to help users understand the patterns, trends, and insights hidden within the data. It is an essential tool in data analysis and communication as it allows complex data sets to be presented in a visual format, making it easier for both technical and non-technical audiences to interpret the information.

- **Understanding Data:** Visualizations provide a clear and intuitive way to grasp the underlying patterns and relationships in the data, enabling better understanding and insights.

- **Discovering Patterns and Trends:** Data visualization helps in identifying trends, outliers, and patterns that might not be apparent in raw data.

- **Effective Communication:** Visuals are more engaging and memorable than tables of numbers or textual descriptions, making it easier to communicate findings to stakeholders.

- **Decision-Making:** Visualizations facilitate data-driven decision-making, enabling organizations to make informed choices based on data insights.

- **Data Visualization Libraries**

There are various libraries available to create data visualizations in Python. One of the most popular and widely used libraries is **matplotlib**.

Introduction to Matplotlib

Matplotlib is a comprehensive data visualization library in Python that allows users to create static, interactive, and animated plots. It provides a high-level interface for drawing attractive and informative statistical graphics. Matplotlib can create various types of plots, including line plots, scatter plots, bar plots, histograms, pie charts, and more.
Key Concepts in Matplotlib

- **Figure and Axes:** A Figure is the top-level container that holds all the elements of a plot. Within a Figure, we have one or more Axes objects, which represent the actual plotting area where data is plotted. Figures and Axes can be thought of as a canvas and subplots on that canvas, respectively.

- **Plot Types:** Matplotlib supports various types of plots, including line plots, scatter plots, bar plots, histograms, box plots, etc. Each plot type can be customized with different attributes to suit the data being visualized.

- **Customization:** Matplotlib allows extensive customization of plots by providing access to a wide range of parameters for controlling colors, markers, labels, titles, legends, axis limits, and more.

- **Subplots:** Matplotlib can create multiple plots in a single Figure using subplots. Subplots help in comparing different data sets or different aspects of the same data in a structured manner.

- **Color Maps:** Color maps are used to represent data values with different colors. Matplotlib provides a variety of built-in color maps for effective representation of data.

Algorithm:

Task 1: Import the "City_Air_Quality.csv" dataset

```
import pandas as pd
import matplotlib.pyplot as plt
# Load the dataset
data = pd.read_csv("City_Air_Quality.csv")
```

Task 2: Explore the dataset

To understand the structure and content of the dataset, we will perform some exploratory data analysis.

```
# Display the first few rows of the dataset
print(data.head())

# Check the basic statistics of the dataset
print(data.describe())
```

```
# Check the data types and missing values
print(data.info())
```

Task 3: Identify relevant variables for visualizing AQI trends
Based on the dataset's structure, we need to identify the relevant variables for our visualizations. These variables include date, pollutant levels, and AQI values.

Task 4: Create line plots for overall AQI trend

Task 5: Plot individual pollutant levels over time

Task 6: Use bar plots to compare AQI values across different dates
Now, we will use bar plots or stacked bar plots to compare the AQI values across different dates or time periods.

Task 7: Create box plots to analyze the distribution of AQI values
We will create box plots or violin plots to analyze the distribution of AQI values for different pollutant categories

Task 8: Use scatter plots to explore the relationship between AQI values and pollutant levels
Lastly, we will use scatter plots or bubble charts to explore the relationship between AQI values and pollutant levels.

Task 9: Customize the visualizations
To make our visualizations more informative and presentable, we will customize them by adding labels, titles, legends, and appropriate color schemes.
Feel free to experiment with the visualizations and customize them further according to your preferences.

Conclusion:

In this way, we have successfully visualized and analyzed the Air Quality Index (AQI) trends in a city using the matplotlib library in Python.

Oral Questions:

1) What do you mean by data visualization?
2) What are the different libraries in python?
3) What is Matplotlib?
4) What are the different maps in Matplotlib?

Assignment 6

Problem Statement:

Assignment 6

Data Aggregation
Analyzing Sales Performance by Region in a Retail Company

Dataset: "Retail_Sales_Data.csv"

Description: The dataset contains information about sales transactions in a retail company. It includes attributes such as transaction date, product category, quantity sold, and sales amount. The goal is to perform data aggregation to analyze the sales performance by region and identify the top-performing regions.

Tasks to Perform:

1. Import the "Retail_Sales_Data.csv" dataset.
2. Explore the dataset to understand its structure and content.
3. Identify the relevant variables for aggregating sales data, such as region, sales amount, and product category.
4. Group the sales data by region and calculate the total sales amount for each region.
5. Create bar plots or pie charts to visualize the sales distribution by region.
6. Identify the top-performing regions based on the highest sales amount.
7. Group the sales data by region and product category to calculate the total sales amount for each combination.
8. Create stacked bar plots or grouped bar plots to compare the sales amounts across different regions and product categories.

Objective:

The objective of analyzing sales performance by region in a retail company is to gain a comprehensive understanding of how sales are performing in different geographic areas. This analysis is conducted to achieve the following key objectives:

1) **Identifying High-Performing and Underperforming Regions**: The primary goal of analyzing sales performance by region is to identify which areas are generating strong sales and which regions are lagging behind. This information helps the retail company to allocate resources effectively and focus on regions with growth potential.

2) **Understanding Customer Behavior and Preferences**: By examining sales data by region, the company can gain insights into customer behavior, preferences, and buying patterns specific to each location. This understanding helps in tailoring products, marketing strategies, and promotions to better meet the needs of customers in different regions.

3) **Evaluating the Impact of Marketing and Sales Strategies**: Analyzing sales performance by region allows the company to evaluate the effectiveness of various marketing and sales initiatives in different areas. This assessment helps in refining strategies that work well in certain regions and adapting or replacing those that do not yield desired results in others.

4) **Optimizing Inventory Management**: Understanding sales performance by region helps the retail company optimize inventory management. It enables them to adjust stock levels based on demand in each region, reducing carrying costs and minimizing stockouts or excess inventory.

5) **Spotting Growth Opportunities**: Sales performance analysis may reveal untapped markets or regions with growth potential. Identifying these opportunities allows the company to prioritize expansion efforts and invest strategically to capture new market share.

6) **Supporting Territory and Sales Force Management**: By analyzing sales performance by region, the company can assess the performance of its sales teams in different areas. This evaluation can aid in recognizing and rewarding high-performing sales representatives and providing additional support or training to those in regions facing challenges.

7) **Benchmarking and Goal Setting**: Analyzing sales performance by region provides benchmarks for setting realistic and achievable sales targets for each area. It enables the company to set performance goals based on historical data and market potential.

8) **Informing Business Expansion Strategies**: The analysis of sales performance by region provides valuable data for making informed decisions about business expansion. It helps in identifying which regions are most promising for growth and expansion, guiding the company's strategic planning.

9) **Improving Overall Sales and Profitability**: Ultimately, the objective of analyzing sales performance by region is to improve overall sales and profitability for the retail company. The insights gained from this analysis can lead to more effective sales and marketing strategies, better customer engagement, and increased revenue in specific regions.

Theory

What is Data Aggregation?

- Data aggregation is the process of collecting data to present it in summary form. This information is then used to conduct statistical analysis and can also help company executives make more informed decisions about marketing strategies, price settings, and structuring operations, among other things.
- Data aggregation is typically performed on a large scale via software programs known as data aggregators.
- Data Aggregation is the way in which data is gathered from multiple sources, compiled, and presented in a summarized manner.
- It is very necessary to collect quality content in huge amounts so that they can create relevant outcomes.
- Data aggregation plays a vital role in finance, product, operations, and marketing strategies in any business organization. Aggregated data is present in the data warehouse that can enable one to solve various issues, which helps solve queries from data sets.

Fig: Example of Data Aggregation

There are mainly 2 types of Data Aggregation:

- **Manual**
- **Automated**

1.Manual Data Aggregation:

In a **Manual Data Aggregation approach**, the data is aggregated manually by employees. A Data Aggregation Tool is used to export the data from multiple sources and then all the data is sorted through an Excel sheet manually. Employees have to manually format all the data into a common format and then they have to create charts to compare the performance of the aggregated data based on the metrics considered.

All of these tasks can become very cumbersome and there is a high chance of error. In order to prevent these errors, the whole process is automated.

2. Automated Data Aggregation:

In the **Automated Data Aggregation process**, a 3rd party device, also called **Middleware**, is used to gather data automatically from the marketing, product, SaaS, and numerous other platforms. When the process gets automated, the region of interest for the data gets expanded and this frees up time to focus on other parts of the analytical process.

Analyzing sales performance by region in a retail company is essential for identifying trends, patterns, and opportunities for improvement. Data aggregation is a crucial technique that allows you to summarize and analyze data from different regions effectively. Here's a step-by-step guide on how to perform this analysis:

- **Gather Data:** Collect all relevant sales data, including region-specific information, such as sales revenue, quantity sold, profit, customer demographics, etc. Make sure the data is organized and available in a structured format, such as a spreadsheet or a database.
- **Define Regions:** Clearly define the regions you want to analyze. Regions can be countries, states, cities, or any other geographical divisions based on your retail company's structure.

- **Data Aggregation:** Aggregate the sales data by the defined regions. Common aggregation functions include summing the sales revenue, calculating the average profit, finding the maximum or minimum sales, etc. For example, you can calculate the total sales revenue, total profit, and average quantity sold for each region.

- **Visualize the Data:** Create visualizations, such as bar charts, pie charts, or maps, to present the aggregated data in a more intuitive and understandable way. Visualizations help you identify trends and disparities between different regions more easily.

- **Compare Regions:** Compare the sales performance across different regions. Look for regions with high growth rates, consistently strong sales, or areas where there might be untapped potential. Identify regions that are underperforming and may require further investigation to understand the reasons behind their lower sales figures.

- **Analyze Trends:** Analyze historical data to identify any recurring trends or seasonal patterns in sales performance by region. Understanding these trends can help you plan better for future sales strategies and promotions.

- **Customer Segmentation:** Segment customers based on their demographics and buying behavior within each region. This can provide insights into which products are more popular in specific regions and help tailor marketing efforts accordingly.

Conclusion:

In this way, we have Analyzing Sales Performance by Region in a Retail Company successfully.