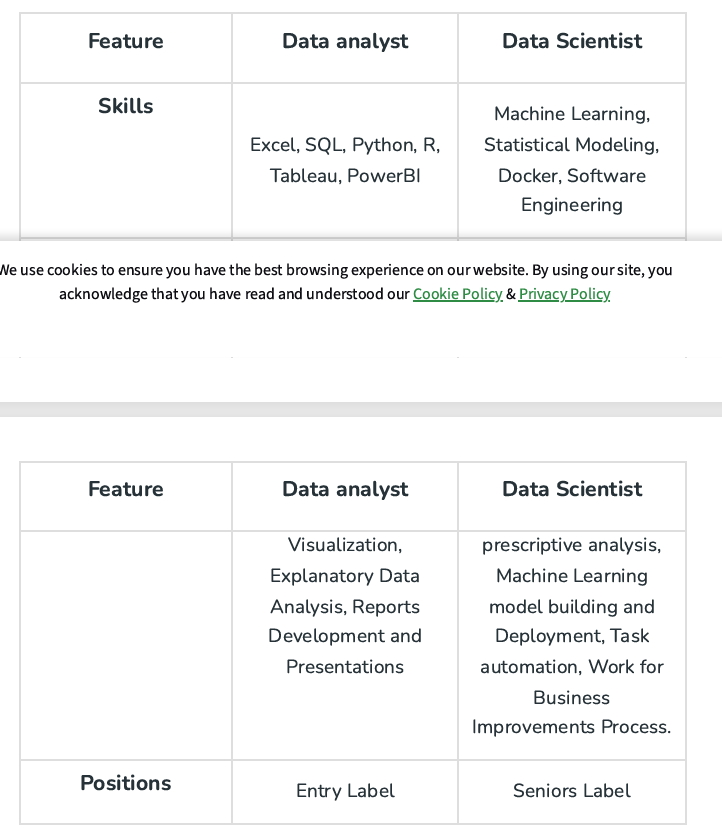
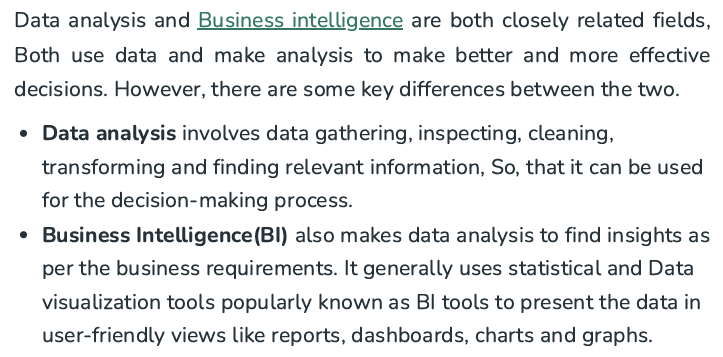
1. **How do data analysts differ from data scientists?**



1. **How Data analysis is similar to Business Intelligence?**

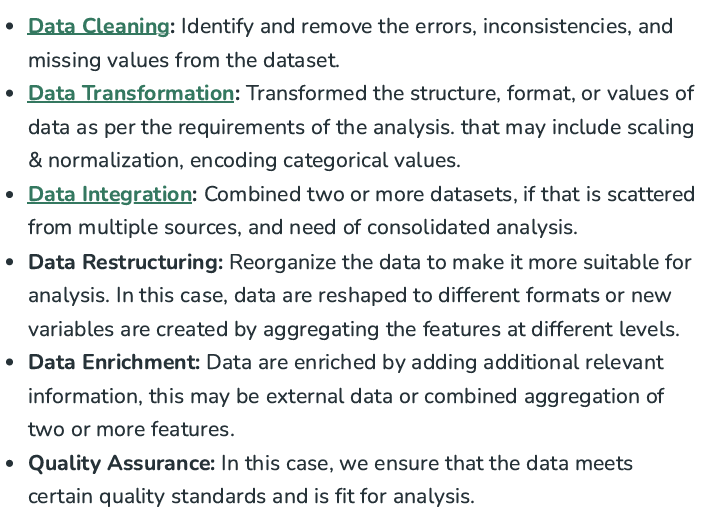
****

1. **What are the different tools mainly used for data analysis?**

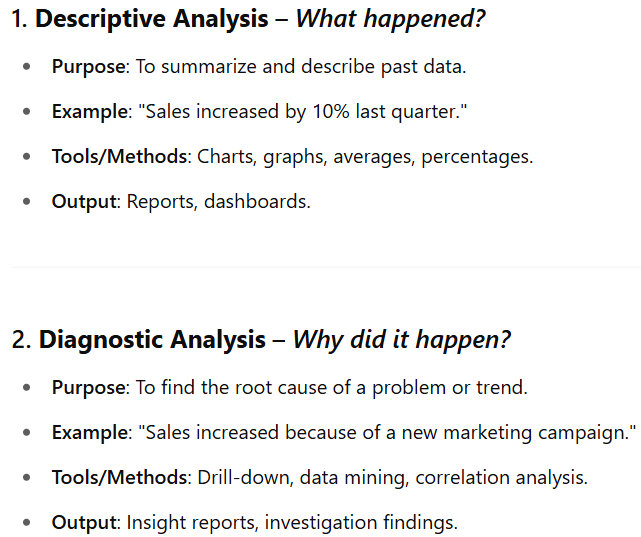
Excel, SQL, Programming lanugaues

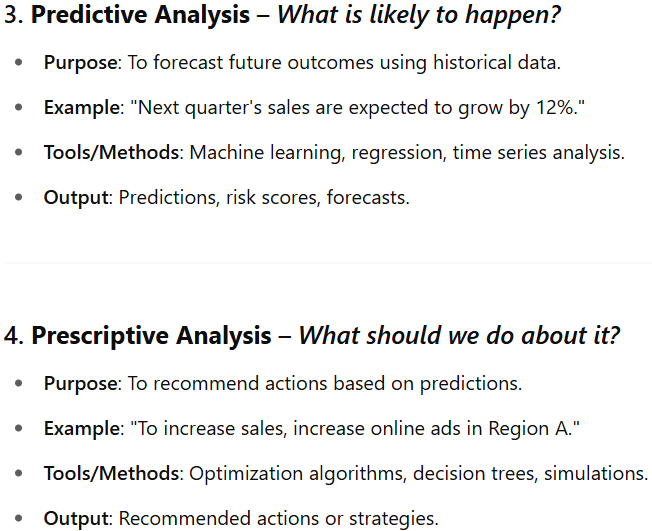
1. **What is Data Wrangling?**

Another name of data preprocessing.

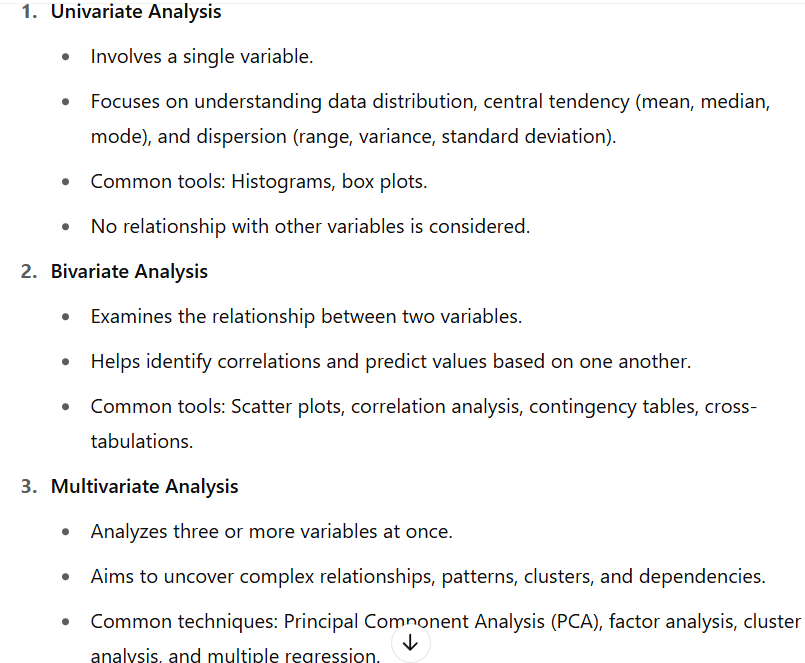


1. **Difference between descriptive, dignosis, predictive, prescriptive analysis**





1. **Difference between univariate, bivariate, multivariate analysis**

****

1. **Name some of the most popular data analysis and visualization tools used for data analysis**

Tableau, Power Bi, SAS, Google data studio free web based application

1. **What are the steps you would take to analyze the dataset**

 Problem Definition / Objective  
Clearly define the problem or question. Understand the goals to guide the entire analysis.

 Data Collection  
Gather relevant data from sources like surveys, databases, web scraping, etc. Ensure accuracy and representativeness.

 Data Preprocessing / Cleaning  
Clean the data by renaming columns, standardizing formats, and handling missing or inconsistent values.

 Exploratory Data Analysis (EDA)  
Use statistical and graphical methods to explore data, identify patterns, outliers, and gain initial insights.

 Data Visualization  
Create visual representations (charts, graphs) to simplify complex data, reveal trends, and communicate findings effectively to stakeholders.

1. **What is data cleaning?**

DC is a process of removing misleading info, it is the next process of data collection and loading, in data clearing we will do

1. Cleaning in inconsistent formatted data
2. Remove duplicate entries
3. Missing values
4. Outliers
5. **What is the importance of exploratory data analysis (EDA) in data analysis?**

**Exploratory Data Analysis (EDA)** is a **non-parametric** method that doesn't assume anything about the data. It plays a crucial role in the data analysis process by:

1. Providing deep insights into data patterns, distributions, and relationships between variables.
2. Assessing data quality through univariate analysis (e.g., mean, median, distribution plots).
3. Exploring relationships via bivariate and multivariate analyses (e.g., regression, correlation, scatter plots).
4. Identifying influential features using statistical measures like correlation and covariance.
5. Detecting and addressing outliers using box plots and statistical methods.

Overall, EDA forms the foundation for effective data processing, modeling, and insight generation.

1. **What is Time series analysis?**

Time Series Analysis is a statistical method used to examine and interpret data points collected over time at consistent intervals. Its primary goal is to understand the underlying patterns and predict future values.

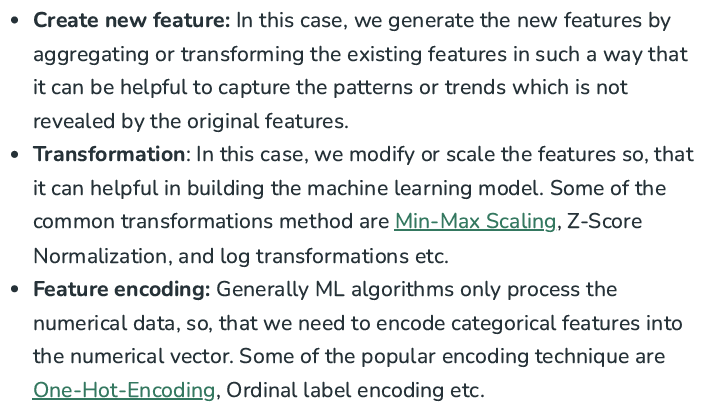
**Key Components:**

* **Trend:** Long-term direction of the data (upward, downward, or stable).
* **Seasonality:** Repeating patterns at regular intervals (daily, monthly, yearly).
* **Cyclical Patterns:** Irregular, long-term fluctuations often linked to economic or business cycles.
* **Irregular Fluctuations:** Random and unpredictable variations not explained by other components.
* **Auto-correlations:** The relationship between current data points and past values.

1. **What is Feature Engineering**

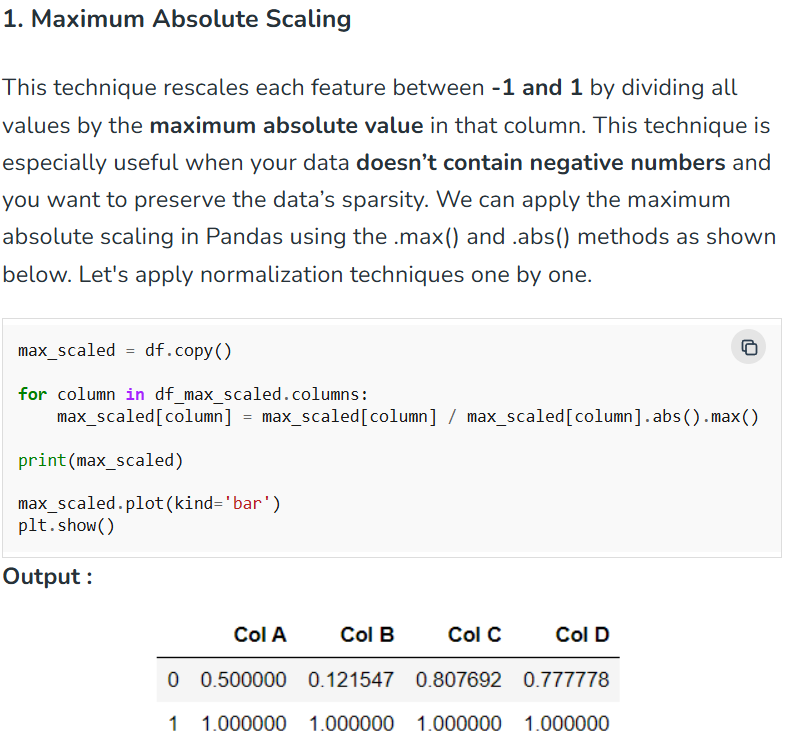
The primary goal of feature engineering is to identify the most relevant features or create the relevant features by combining two or more features using some mathematical operations from the raw data so that it can be effectively utilized for getting predictive analysis by machine learning model.

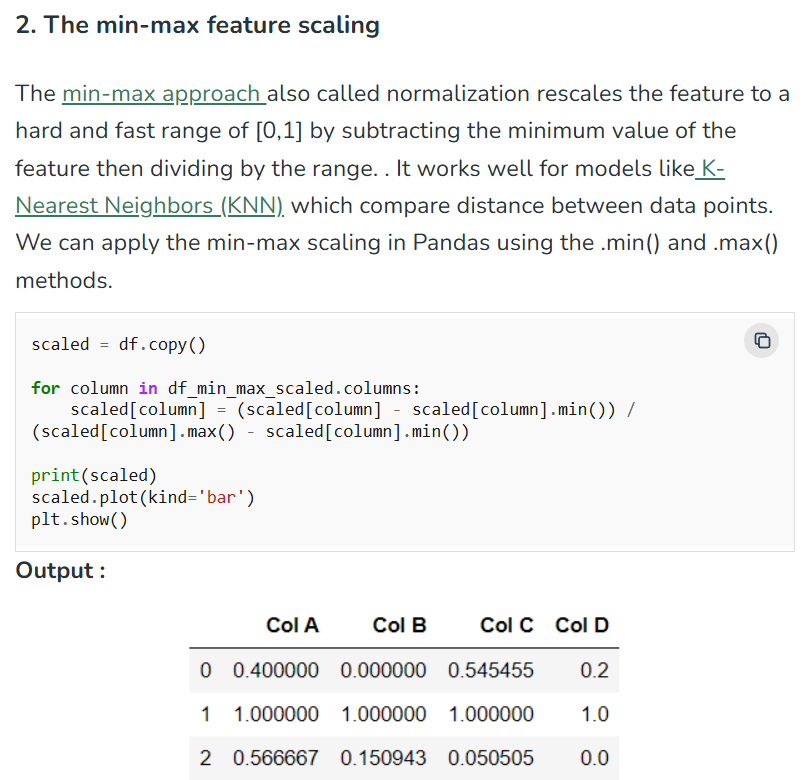
The following are the key elements of feature engineering:

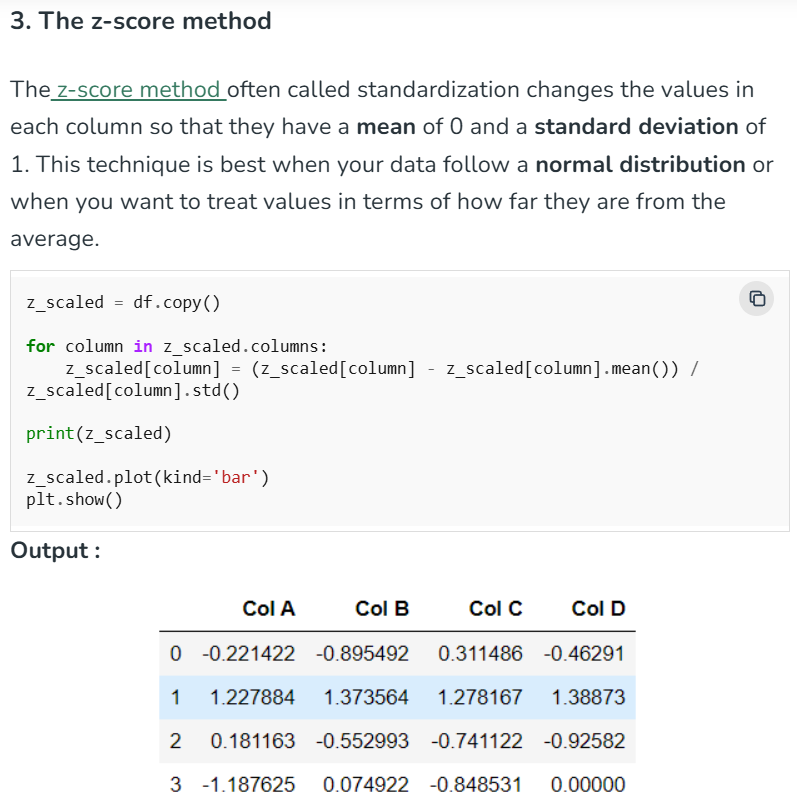


1. **What is data normalization, and why is it important?**

Data normalization is the process of transforming numerical data into standardised range. The objective of data normalization is scale the different features (variables) of a dataset onto a common scale, which make it easier to compare, analyze, and model the data. This is particularly important when features have different units, scales, or ranges because if we doesn't normalize then each feature has different impact which can affect the performance of various machine learning algorithms and statistical analyses. Common normalization techniques are as follows:

****

****

****

1. **What are the main libraries you would use for data analysis in Python?**

For data analysis in Python, many great libraries are used due to their versatility, functionality, and ease of use. Some of the most common libraries are as follows:

**NumPy:** A core Python library for numerical computations. It supports arrays, matrices, and a variety of mathematical functions, making it a building block for many other data analysis libraries.

**Pandas:** A well-known data manipulation and analysis library. It provides data structures (like as DataFrames) that make to easily manipulate, filter, aggregate, and transform data. Pandas is required when working with structured data.

**SciPy:** SciPy is a scientific computing library. It offers a wide range of statistical, mathematical, and scientific computing functions.

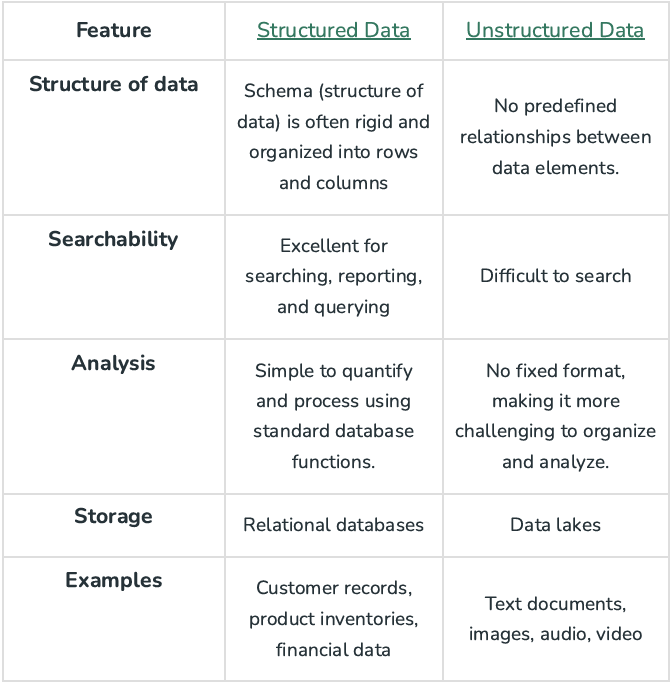
**Matplotlib:** Matplotlib is a library for plotting and visualization. It provides a wide range of plotting functions, making it easy to create beautiful and informative visualizations.

**Seaborn:** Seaborn is a library for statistical data visualization. It builds on top of Matplotlib and provides a more user-friendly interface for creating statistical plots.

**Scikit-learn:** A powerful machine learning library. It includes classification, regression, clustering, dimensionality reduction, and model evaluation tools. Scikit-learn is well-known for its consistent API and simplicity of use.

**Statsmodels:** A statistical model estimation and interpretation library. It covers a wide range of statistical models, such as linear models and time series analysis.

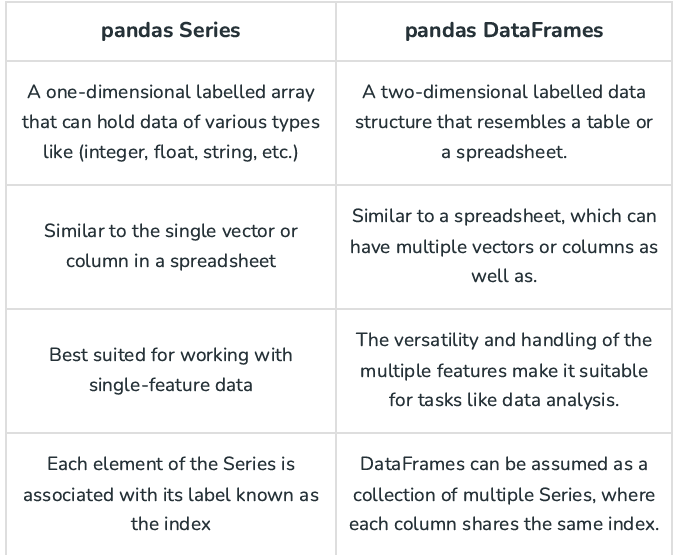
1. **What's the difference between structured and unstructured data**

****

1. **How can pandas be used for data analysis?**

It is a open source library that provides a lot of options to manipulate data.

1. Easily Loads various platform data – sql, excel, csv
2. Provides data cleaning function – fillna, drop\_duplicates, as\_type
3. Provides data transformation functions – apply, map
4. Data aggregation – groupby, sum, mean
5. Time series – resample, shift
6. Merging, Joining
7. **What is the difference between pandas Series and pandas Dataframes?**

****

1. **What is One-Hot-Encoding?**

One-hot encoding is a technique used for converting categorical data into a format that machine learning algorithms can understand. Categorical data is data that is categorized into different groups, such as colors, nations, or zip codes. Because machine learning algorithms often require numerical input, categorical data is represented as a sequence of the category variable is "color" and the potential values are "red," "green," and "blue," then three additional binary variables are created: "color\_red," "color\_green," and "color\_blue." Each of these binary variables would have a value of 1 if the matching category value was present and 0 if it was not