1. Raw data clean karane
2. Raw data madhe atta paryant cha 2025 cha data add karane
3. Raw data madhe required columns add karane

* Date madhun date, month (word format), year and time separate karane
* Order date madhun ship date and delivery date find out karane
* Exchange rate is 0.6% in total purchase
* Generate customer\_id, rename transaction\_id as order\_id and product\_id

1. Raw file is divided into three separate files such as customer\_info, order\_info, product\_info in CSV format and save those files into bronze rename folder
2. Then join those files or tables and create new file or table, in the new created table or file add useful columns which are helpful in analysis
3. Then remove null values, duplicates and required data manupulation from origin separated tables and save them into new folder as silver data
4. Do visualizations based on given information and save that visual data into gold rename folder
5. Then upload those files into github
6. Then use Microsoft azure services and create pipeline to attach files or data from github to azure data storage using azure data factory
7. Then mount that data into azure databricks and perform some analysis
8. Then use those output for report and show these reports into powerBI
9. Then use azure synapse analysis

9860101303 - om dress shop

**Key Purposes of EDA:**

* **Understanding Data Structure:**

EDA helps data scientists understand the structure and characteristics of a dataset, including identifying data types, missing values, and potential issues.

* **Detecting Patterns and Anomalies:**

EDA techniques, such as visualizations and descriptive statistics, can reveal patterns, trends, and outliers within the data that might not be immediately apparent.

* **Testing Hypotheses:**

EDA can be used to test preliminary hypotheses or assumptions about the data, providing insights into the relationships between variables.

* **Checking Assumptions:**

EDA helps determine if the statistical techniques or models you are considering for data analysis are appropriate for the data's characteristics.

* **Data Cleaning and Preparation:**

EDA can identify data quality issues (e.g., missing values, outliers) that need to be addressed before further analysis or modeling.

**Common EDA Techniques:**

* **Descriptive Statistics:**

Calculating summary statistics like mean, median, standard deviation, and quartiles to understand the central tendency, variability, and distribution of the data.

* **Data Visualization:**

Creating plots and charts (histograms, scatter plots, box plots, etc.) to visualize the data and identify patterns or relationships.

* **Bivariate Analysis:**

Examining the relationship between two variables to understand their correlation or dependency.

* **Multivariate Analysis:**

Exploring the relationships between multiple variables to understand complex interactions.

* **Outlier Detection:**

Identifying data points that deviate significantly from the rest of the data using techniques like box plots, z-score analysis, and scatter plots.

* **Data Transformation:**

Applying transformations to the data (e.g., scaling, normalization) to make it more suitable for analysis or modeling.

**Why EDA is Important:**

* **Provides a Foundation for Further Analysis:**

EDA helps data scientists gain a solid understanding of the data before moving on to more complex analysis or modeling tasks.

* **Identifies Data Quality Issues:**

EDA can reveal data quality problems that need to be addressed before building accurate models.

* **Guides Model Selection:**

Understanding the data's characteristics through EDA can help data scientists choose the most appropriate models and techniques for their analysis.

* **Facilitates Communication:**

Visualizations and summaries from EDA can help data scientists communicate their findings to stakeholders in a clear and concise manner.