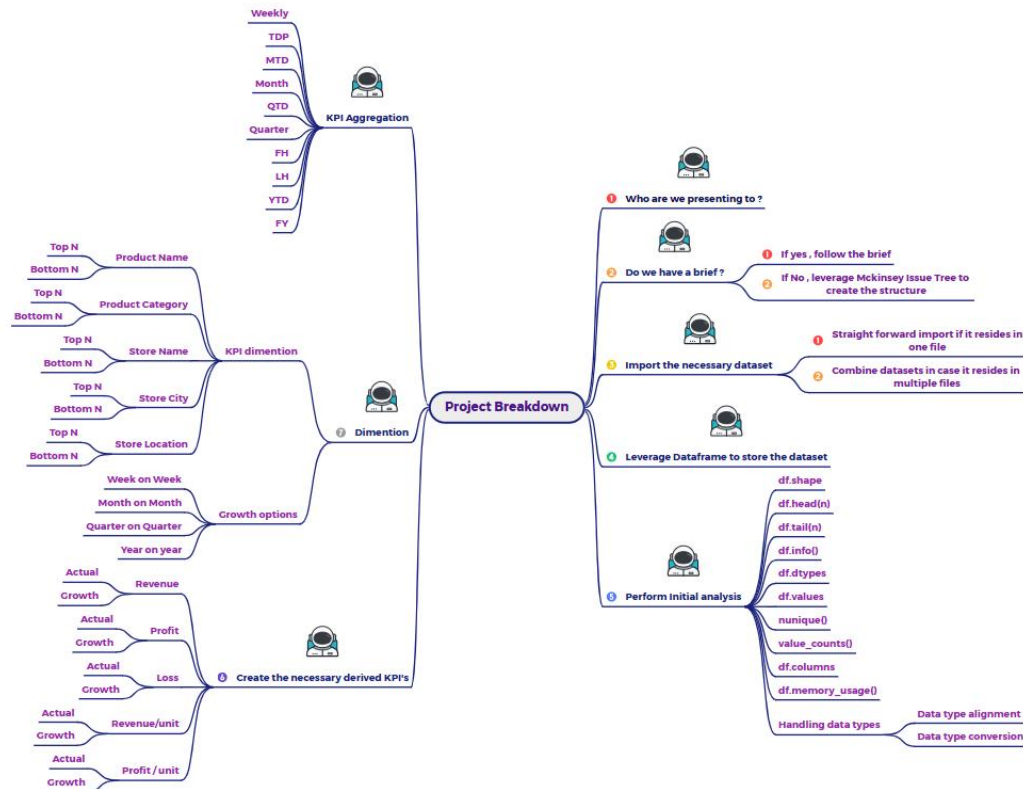# How to structure the EDA project?

Overall Framework: Project Breakdown for Data Analysis/EDA

The mind map centers around "Project Breakdown" and branches into five main phases:

1. **Audience and Briefing (Who are we presenting to? & Do we have a brief?):**

   o **Purpose:** This initial step emphasizes understanding the project's context and audience.

   o **Key Actions:**

     ▪ If a brief exists, follow it to structure your analysis.

- If no brief, leverage a "McKinsey Issue Tree" (a problem-solving framework) to structure the analysis, implying a need to define the core questions and hypotheses.

- **Why for EDA**: Crucial for focusing your EDA efforts on questions that matter to your stakeholders, ensuring your exploration is purposeful.

2. **Import the Necessary Dataset:**

- **Purpose**: Getting your data into the analytical environment.

- **Key Actions:**

  - Straightforward import if data resides in a single file.

  - Combine datasets if they reside in multiple files (implying merging, concatenating, or joining data).

- **Why for EDA**: The absolute first technical step. You can't analyze what you haven't loaded.

3. **Leverage DataFrame to store the dataset:**

- **Purpose**: Once imported, the data should be in a structured format suitable for manipulation (like a Pandas DataFrame).

- **Why for EDA**: DataFrames are the standard for tabular data in Python, providing powerful tools for subsequent steps.

4. **Perform Initial Analysis:**

- **Purpose**: A quick overview of the dataset's structure, content, and basic properties. This is the very beginning of the "exploration" phase of EDA.

- **Key Actions (DataFrame methods):**

  - df.shape: Get the number of rows and columns.

  - df.head(n): View the first 'n' rows.

  - df.tail(n): View the last 'n' rows.

  - df.info(): Get a summary of the DataFrame, including data types and non-null values.

- df.dtypes: Check data types of each column.

- df.values: Access the data as a NumPy array.

- df.nunique(): Count unique values in each column.

- df.value_counts(): Get frequency counts for unique values in a Series.

- df.columns: List column names.

- df.memory_usage(): Check memory consumption.

- **Handling Data Types:**

  - Data type alignment.

  - Data type conversion.

- **Why for EDA:** Essential for understanding the raw state of your data, identifying potential issues (missing values, incorrect types), and planning further cleaning and transformation.

5. **Create the necessary Derived KPIs (Key Performance Indicators):**

   - **Purpose:** Generating new, more insightful metrics from existing data. This is a crucial part of feature engineering in EDA.

   - **Key Actions:**

     - **KPI Aggregation:** Summarizing data over different time periods (e.g., Weekly, TDP (Today), MTD (Month-to-Date), Month, QTD (Quarter-to-Date), Quarter, FH (First Half), LH (Last Half), YTD (Year-to-Date), FY (Fiscal Year)).

     - **KPI Dimension:** Breaking down KPIs by different categorical dimensions (e.g., Product Name, Product Category, Store Name, Store City, Store Location) and identifying Top N and Bottom N performers within these dimensions.

     - **Growth Options:** Calculating growth metrics (e.g., Week on Week, Month on Month, Quarter on Quarter, Year on Year).

     - **Specific KPIs:** Calculating Actual vs. Growth for:

       - Revenue

- Profit

- Loss

- Revenue/unit

- Profit/unit

- o **Why for EDA:** Derived KPIs transform raw data into actionable business metrics, allowing you to explore performance, trends, and identify areas of strength or weakness, which are core goals of EDA.

How to use this framework for your EDA project:

You can follow these steps sequentially:

1. **Start with the "Why":** Before touching code, define what questions you want to answer with your EDA. Use the "Audience and Briefing" section to guide this.

2. **Get Your Data Ready:** Use the "Import the Necessary Dataset" and "Leverage DataFrame" steps to load your data into Pandas DataFrames.

3. **Initial Data Health Check:** Immediately apply the "Perform Initial Analysis" steps. This is where you'll discover missing values, incorrect data types, and get a feel for your data's scale and distribution.

4. **Enrich Your Data:** Based on your initial questions and data understanding, move to "Create the necessary Derived KPIs." This is where you'll calculate new columns that are more relevant to your analysis (e.g., profit margin, weekly sales growth).

5. **Deep Dive Visualization (Your Seaborn/Matplotlib skills come in!):** While not explicitly a separate branch in this mind map, this is where you'd extensively use your knowledge of Seaborn (relational, distribution, categorical, regression, multi-plot grids) and Matplotlib (customization, text/annotations) to visualize the raw data and your newly created KPIs. This iterative process of deriving and visualizing is the heart of EDA.

6. **Summarize and Present:** Once you've gained insights, use the initial "Audience and Briefing" context to structure your findings and present them effectively.