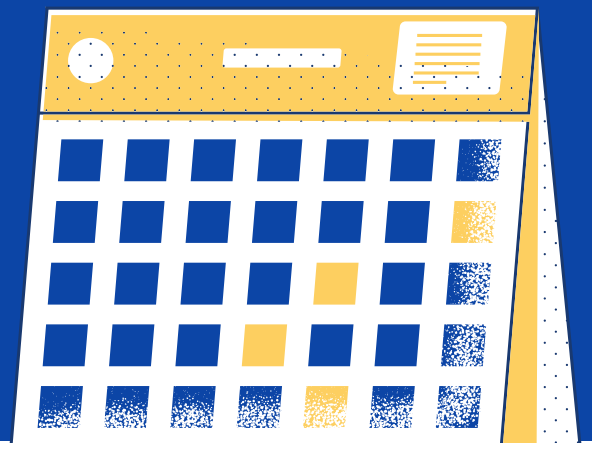
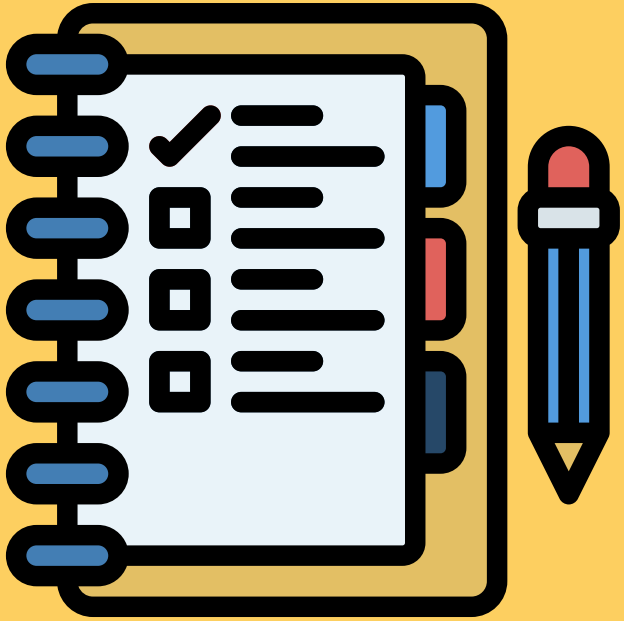
Data Fundamentals

Data Analysis Process

GOALS



- Understand the step-by-step process of Data Analysis.
- Learn how to formulate precise research questions.
- Grasp the significance of data wrangling, its techniques, and its relevance in the analysis pipeline.
- Explore Exploratory Data Analysis (EDA) in-depth.
- Discover how to effectively draw conclusions and make predictions from analyzed data.
- Master the art of effectively communicating results to stakeholders.

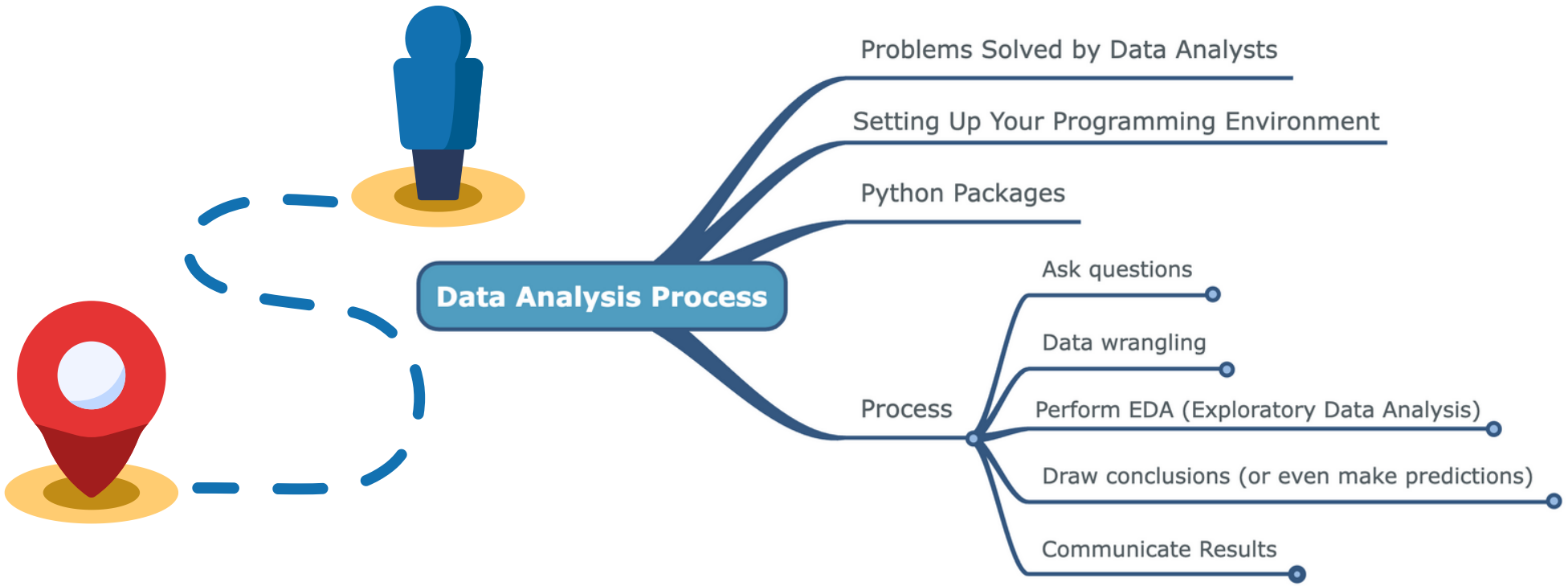


AGENDA

- Introduction to Data Analysis
- Setting Up the Environment
- Formulating Research Questions
- Data Wrangling Techniques
- Exploring Data with EDA
- Drawing Conclusions & Making Predictions
- Effective Communication & Visualization
- Q&A Session
- Feedback & Evaluation
- Resources

Behind every data point, there's a story waiting to be told.

ROADMAP


Data Analysis Process
Problems Solved by Data Analysts
Setting Up Your Programming Environment
Python Packages
Process
Ask questions
Data wrangling
Perform EDA (Exploratory Data Analysis)
Draw conclusions (or even make predictions)
Communicate Results

DATA ANALYSIS

PROBLEMS SOLVED BY DATA ANALYSTS

Company	Solution	Description
	Personalized Recommendations	Personalized movie/show recommendations using viewing habits and ratings.
	Optimal Pricing	Pricing suggestion tool for hosts based on local rental rates and demand.
	Fraud Detection	Real-time fraud detection by analyzing purchase behaviors.
	Optimizing Supply Chains	Optimized supply chain by predicting product demand using sales data.
	Predicting Outbreaks	Predicted flu outbreaks via search query data.
	Player Performance	Player performance analysis using the SportVU system.

SETTING UP YOUR PROGRAMMING ENVIRONMENT

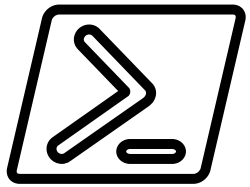
SOFTWARE SETUP

Python3.x

Libraries

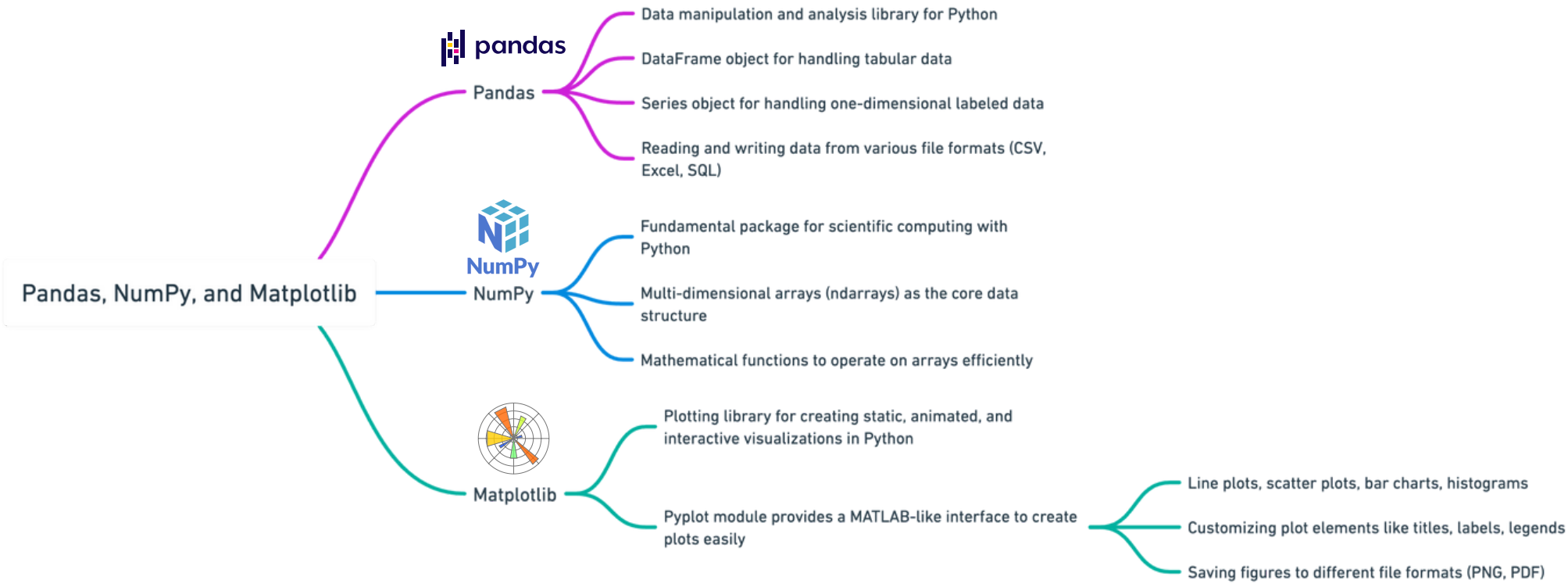
Text Editor

Terminal

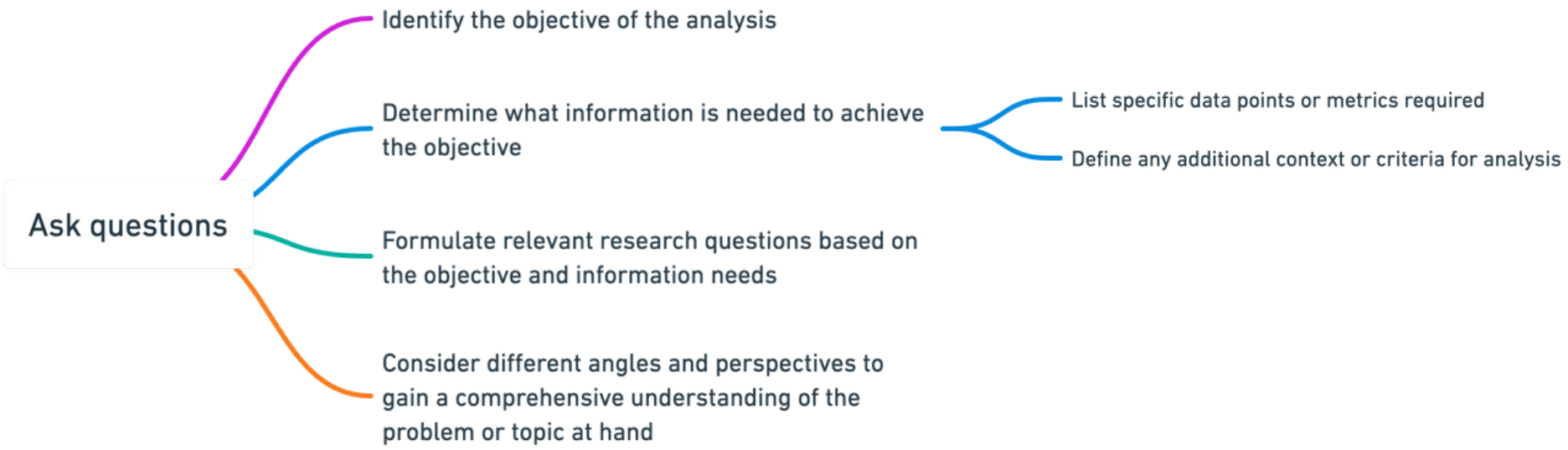


Anaconda is an open-source distribution of Python and R, tailored for data science and machine learning tasks, providing package management, Jupyter Notebooks, and popular libraries in one package.

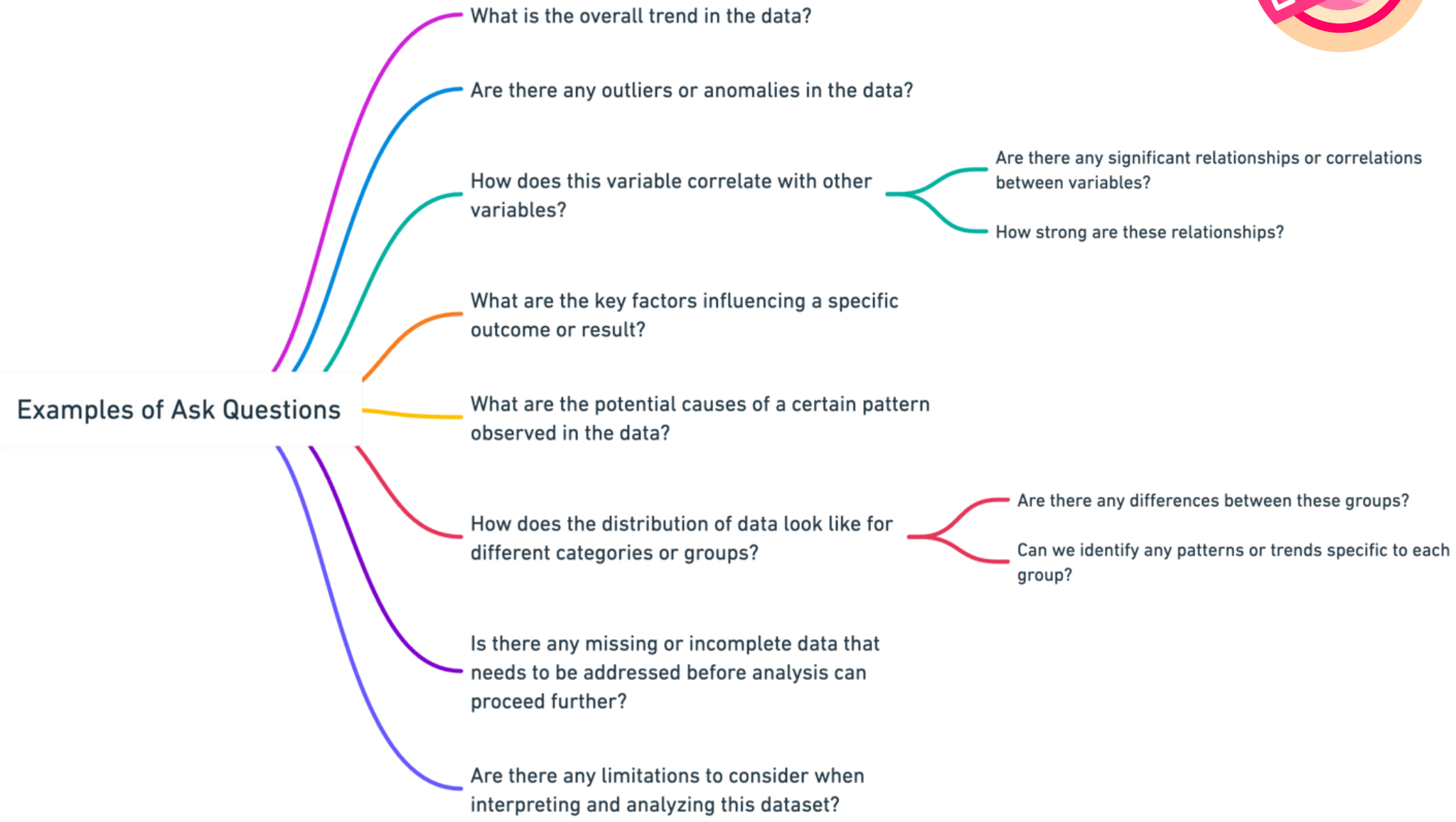
PYTHON PACKAGES



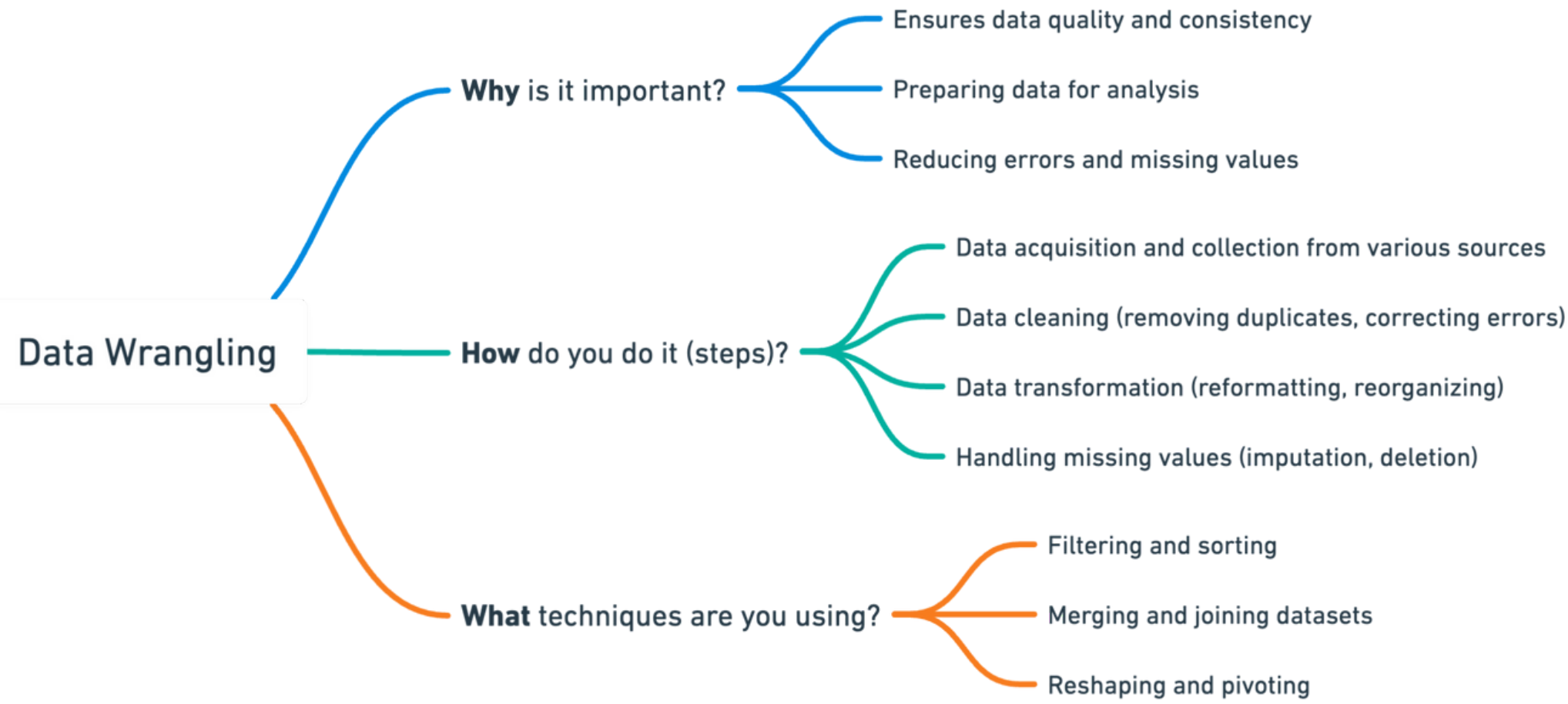
ASK QUESTIONS



ASK QUESTIONS



DATA WRANGLING



DATA WRANGLING


```python
import pandas as pd

# Data acquisition and collection from various sources
# For this example, we're creating a dummy dataset. In a real-world scenario, this could be fetched from da
data = {
    'Date': ['2023-01-01', '2023-01-02', '2023-01-02', '2023-01-03', '2023-01-04', '2023-01-04'],
    'Product': ['A', 'B', 'A', 'C', None, 'C'],
    'Sales': [100, 150, 50, 200, 300, None]
}
df = pd.DataFrame(data)

# Data cleaning
# Removing duplicates
df.drop_duplicates(inplace=True)

# Correcting errors (for this example, let's say Product 'B' was recorded incorrectly and should be 'A')
df['Product'] = df['Product'].replace('B', 'A')

# Data transformation
# Reformatting: For example, changing date from string to datetime object
df['Date'] = pd.to_datetime(df['Date'])

# Reorganizing: Aggregating sales per date and product
df = df.groupby(['Date', 'Product']).sum().reset_index()

# Handling missing values
# Imputation: Filling missing sales values with 0
df['Sales'].fillna(0, inplace=True)
```

```python
# Deletion: Removing rows where product name is missing
df = df.dropna(subset=['Product'])

# Filtering and sorting
# Filtering rows where sales are more than 100
df = df[df['Sales'] > 100]

# Sorting by date
df = df.sort_values('Date')

# Merging and joining datasets
# For this, let's consider another dataset containing product details
product_details = {
    'Product': ['A', 'C'],
    'Category': ['Electronics', 'Furniture']
}
df_product = pd.DataFrame(product_details)

# Joining the datasets on the 'Product' column
df = pd.merge(df, df_product, on='Product', how='left')

# Reshaping and pivoting
# Pivoting the table to see sales of each product per date
df_pivot = df.pivot(index='Date', columns='Product', values='Sales').reset_index()
df_pivot.fillna(0, inplace=True)

print(df_pivot)
```

EDA


EDA (Exploratory Data Analysis)
Why is EDA important?
Gain a better understanding of the data
Identify patterns and relationships
Detect outliers and anomalies
Determine suitable preprocessing techniques
How to Perform EDA?
Visualize data using plots and graphs
Histograms
Box plots
Scatter plots
Bar charts
Calculate descriptive statistics (mean, median, mode)
Assess data quality (missing values, duplicates)
What to Explore during EDA?
Univariate analysis (individual variables)
Distribution of numeric variables
Frequency counts for categorical variables
Skewness and kurtosis assessment
Variable transformations
Bivariate analysis (relationship between two variables)
Missing data analysis
Outlier detection and treatment
Distribution of variables
Summary statistics (mean, median, mode, etc.)
Correlation analysis
Data visualization (using plots and charts)
Feature engineering possibilities

CONCLUSIONS

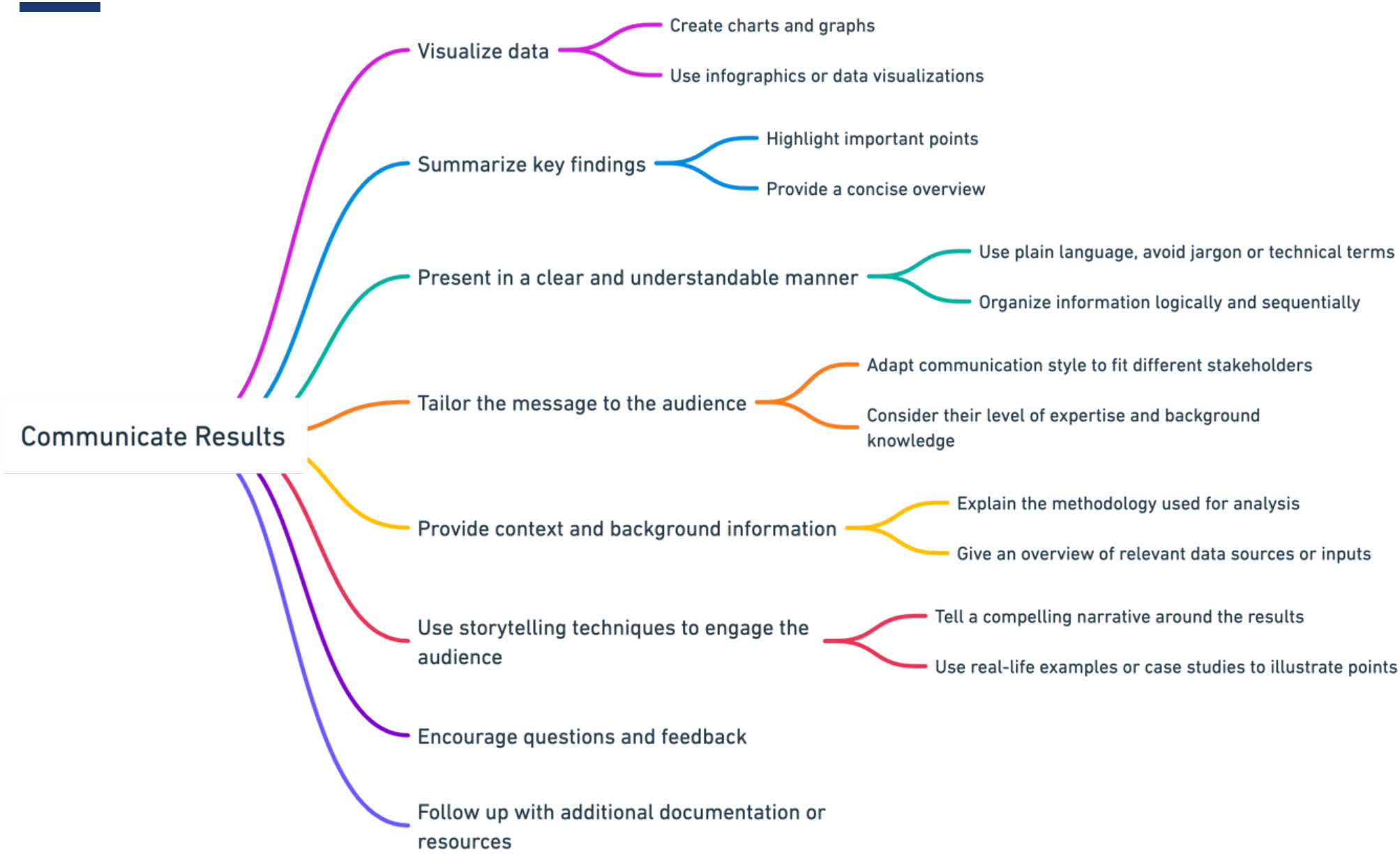
- Conclusions and Predictions
  - Analyze patterns and trends
    - Identify significant correlations
    - Determine statistical significance
  - Draw conclusions based on findings
    - Summarize main observations
    - Make inferences and generalizations
  - Impact assessment and predictions
    - Evaluate potential outcomes based on data analysis results
    - Make predictions for future scenarios

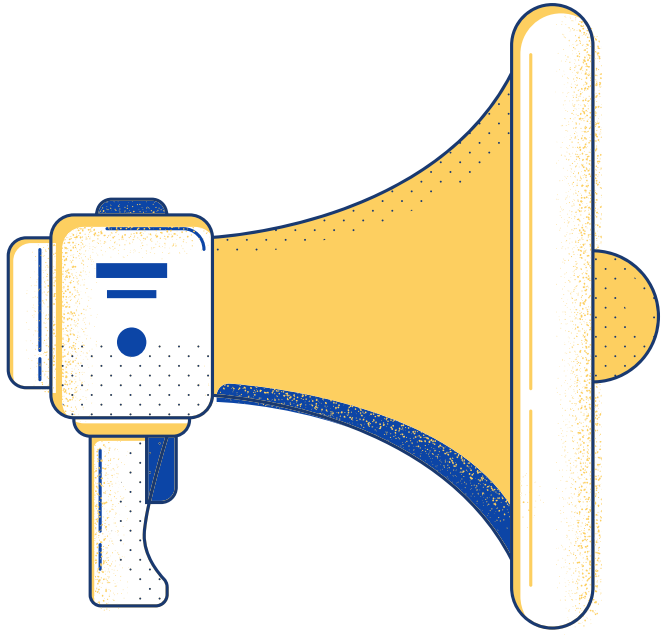
Example of a company is reviewing its sales data

Drawing Conclusions
They might find that every time they run a particular marketing campaign, there's a 20% increase in sales.

Making Predictions
Given the consistent pattern observed, they might predict that running a similar marketing campaign in the future will result in a similar increase in sales.

COMMUNICATE RESULTS





Q&A Session: Let's explore and understand together

WE NEED YOUR FEEDBACK

Help us improve our sessions! Complete the **Oman Makeen Student Satisfaction Survey**.

Rate my competency, our engagement, and your overall satisfaction.

- Confidential and used solely for session improvement.

CLICK HERE

RESOURCES

- Data Analysis Process
- Merge dataframes
- Correlation

Your presence today has added value to our shared learning journey. Thank you for joining us!