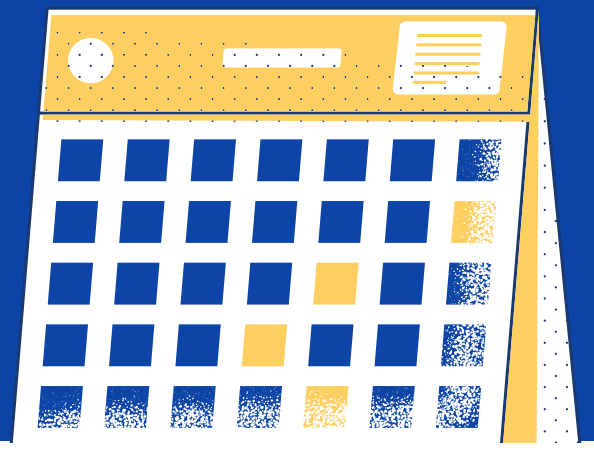
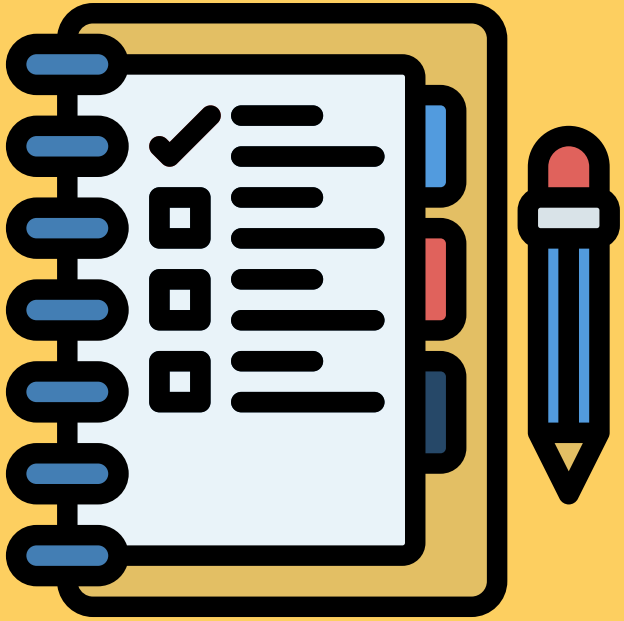
Data Fundamentals

Investigate a Dataset

GOALS



- **Understand the Project** : Analyze and share dataset findings.
- **Master Key Tools** : Numpy, Pandas, and Matplotlib.
- **Installation & Setup** : Use Anaconda for ease.
- **Detailed Analysis** : From dataset choice to result sharing.
- **Submission Standards** : Workspace vs. local machine guidance.
- **Meet Rubric Criteria** : Emphasize code quality, analysis, and communication.

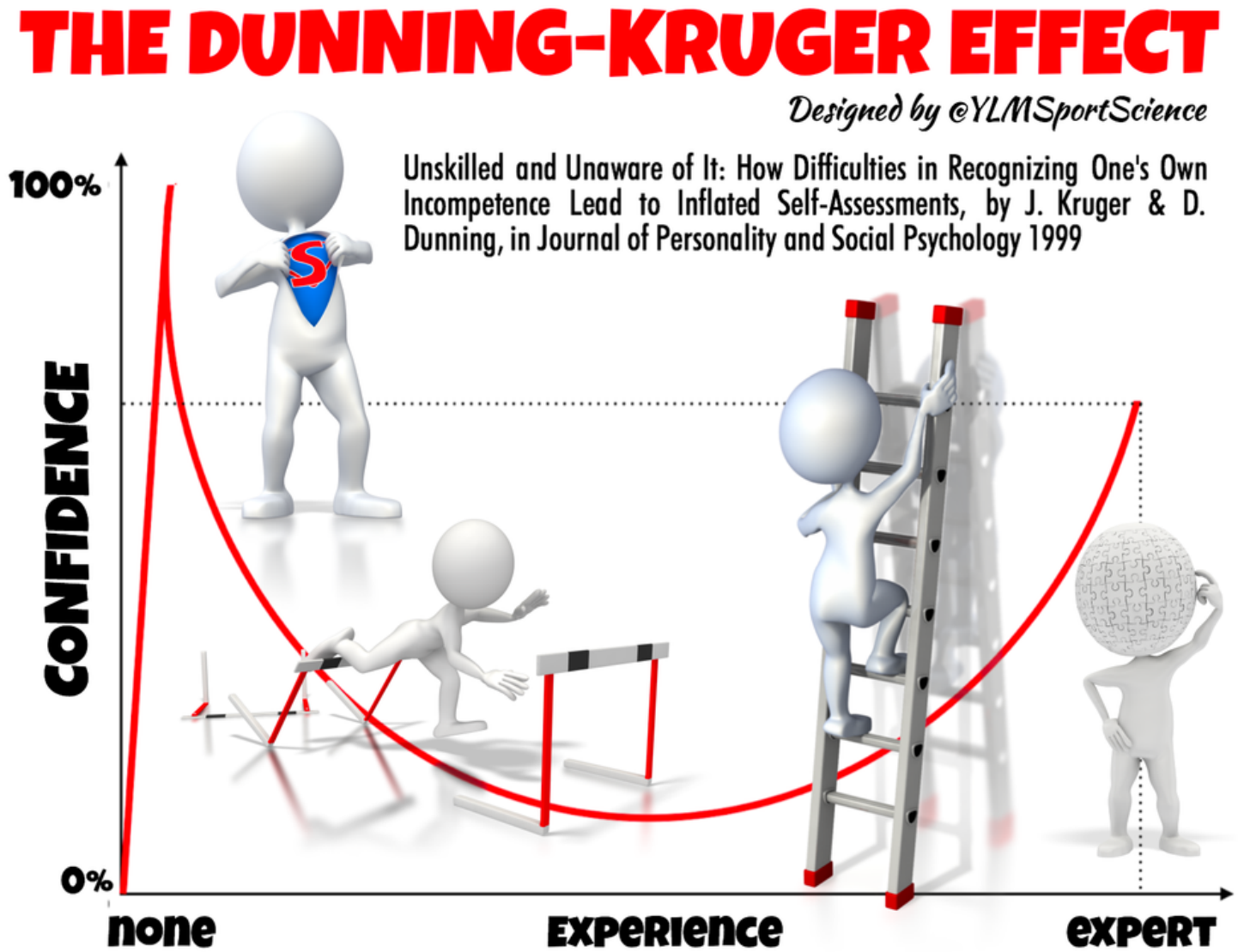


AGENDA

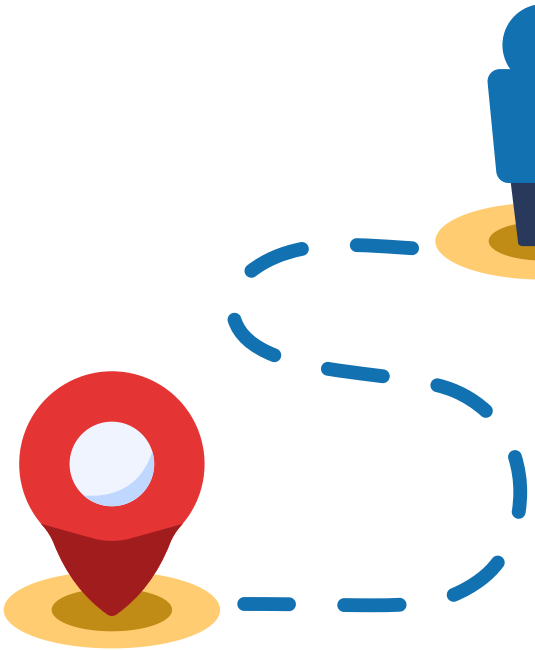
- Project Overview
- Installation Guide
- Project Details
- Completion & Submission
- Rubric Discussion

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

KRUGER EFFECT



ROADMAP



- **Investigate a Dataset**
  - Project Overview
    - Objective
    - Tools used
    - Why this Project?
    - What will I learn?
  - Installation Requirements
    - What to install
    - Recommended platform
    - Installation steps
  - Project Details
    - Introduction
    - Step One: Choose Your Data Set
    - Step Two: Get Organized
    - Step Three: Analyze Your Data
    - Step Four: Share Your Findings
    - Step Five: Review
  - How to Complete and Submit this Project
    - Option #1: Use the provided project workspace environment
    - Option #2: Work on a local machine
  - Project Rubric
    - Code Functionality
    - Quality of Analysis
    - Data Wrangling Phase
    - Exploration Phase
    - Conclusions Phase
    - Communication

PROJECT OVERVIEW

OBJECTIVE

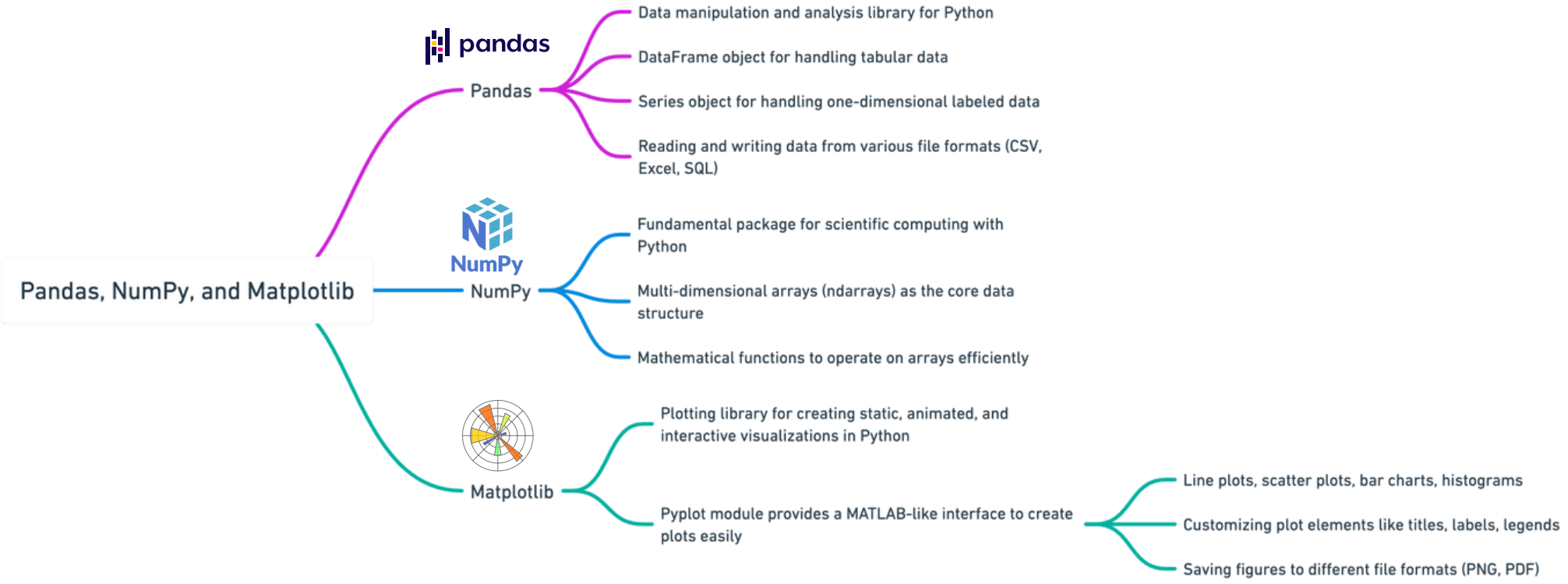
Extract meaningful insights from data.

Determine trends, patterns, and anomalies.

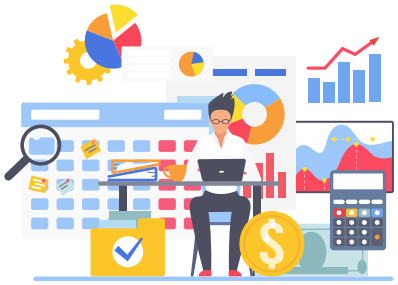
Communicate results in an understandable manner.

TOOLS USED


jupyter


Pandas, NumPy, and Matplotlib
pandas
Pandas
Data manipulation and analysis library for Python
DataFrame object for handling tabular data
Series object for handling one-dimensional labeled data
Reading and writing data from various file formats (CSV, Excel, SQL)
NumPy
NumPy
Fundamental package for scientific computing with Python
Multi-dimensional arrays (ndarrays) as the core data structure
Mathematical functions to operate on arrays efficiently
Matplotlib
Plotting library for creating static, animated, and interactive visualizations in Python
Pyplot module provides a MATLAB-like interface to create plots easily
Line plots, scatter plots, bar charts, histograms
Customizing plot elements like titles, labels, legends
Saving figures to different file formats (PNG, PDF)

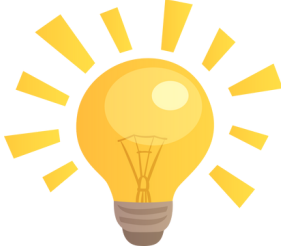
WHY THIS PROJECT?



Learn the data analysis process

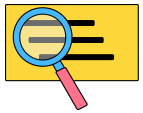
Acquire valued job skills.

WHAT WILL I LEARN?



Understand steps in data analysis

Pose and answer dataset-specific questions



Investigate and wrangle dataset problems

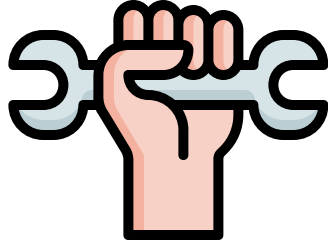
Communicate analysis results

Use vectorized operations in NumPy and pandas

Familiarize with pandas' Series and DataFrame

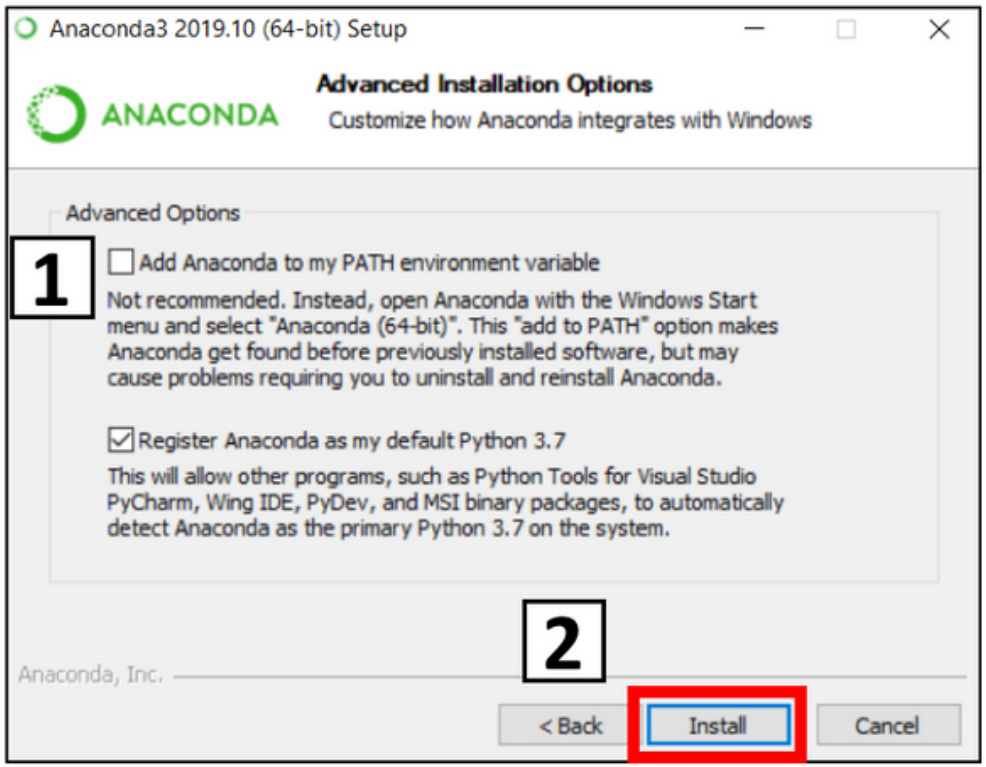
Use Matplotlib for visual findings

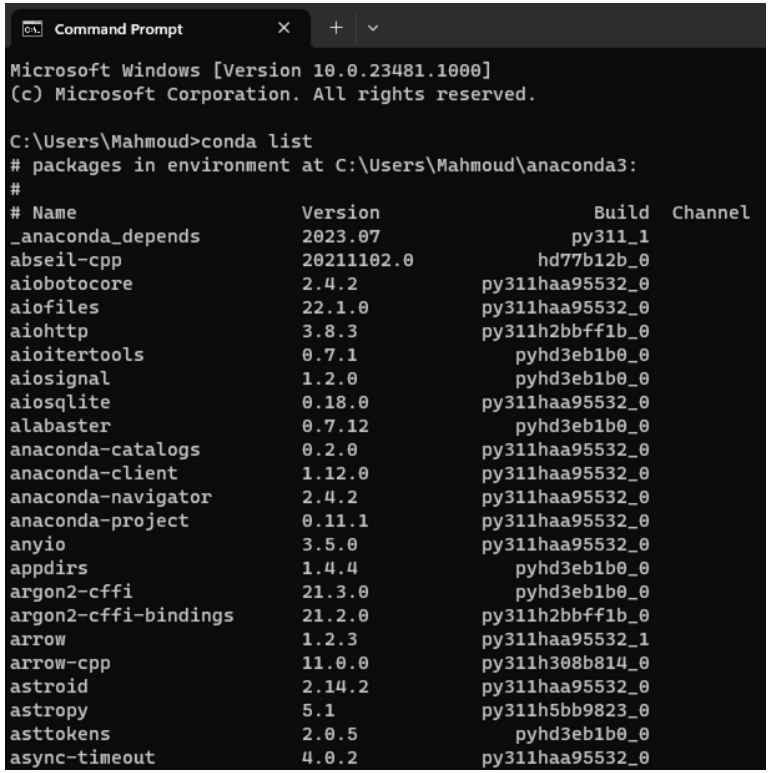
INSTALLATION



- Download from: https://www.anaconda.com/download/
- Select Python 3.7+ and appropriate system version (64/32-bit).
- Verify installation for your OS.

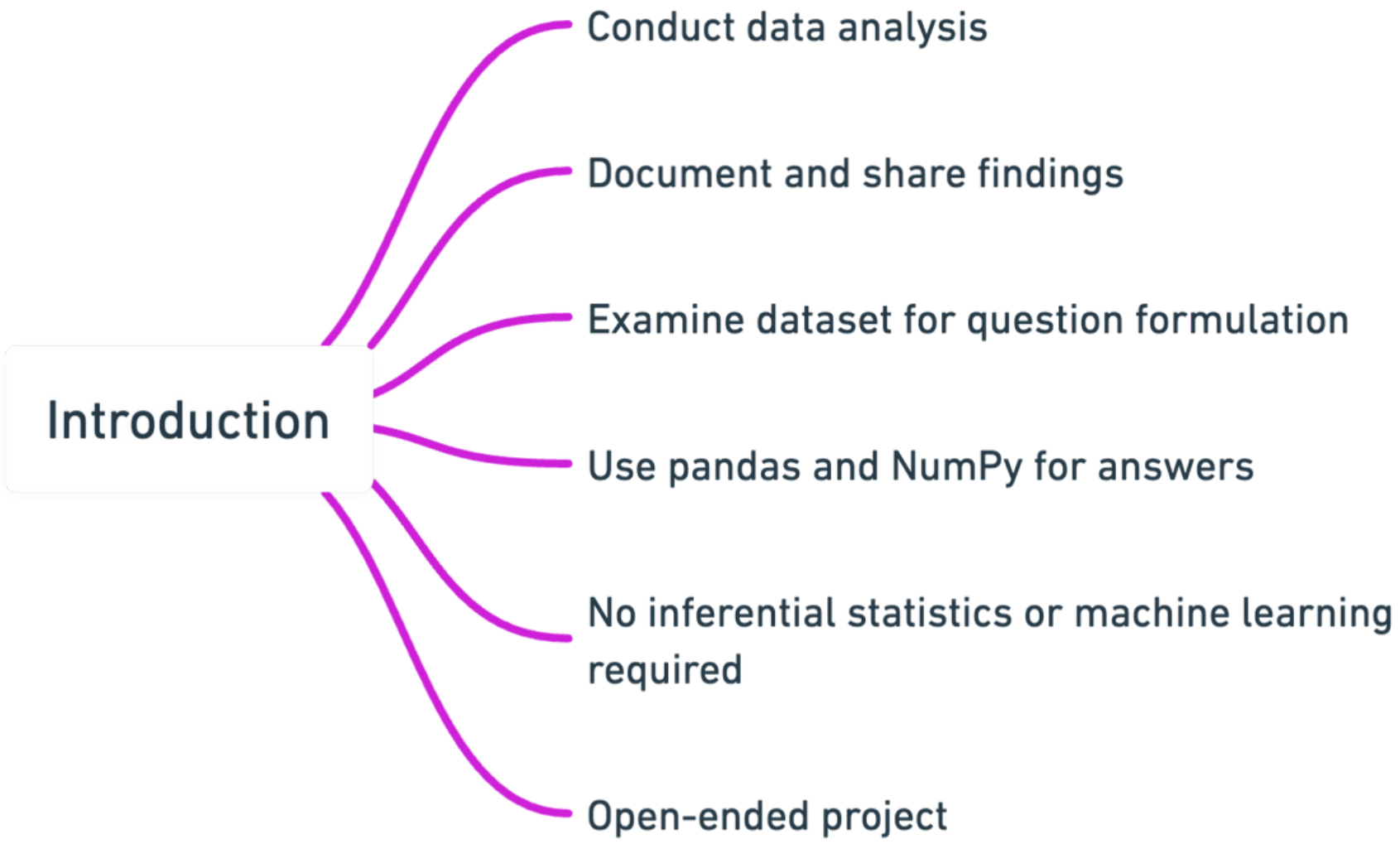
After installation, use **conda list** to view default packages.





Installation guide

PROJECT DETAILS



STEP ONE

1

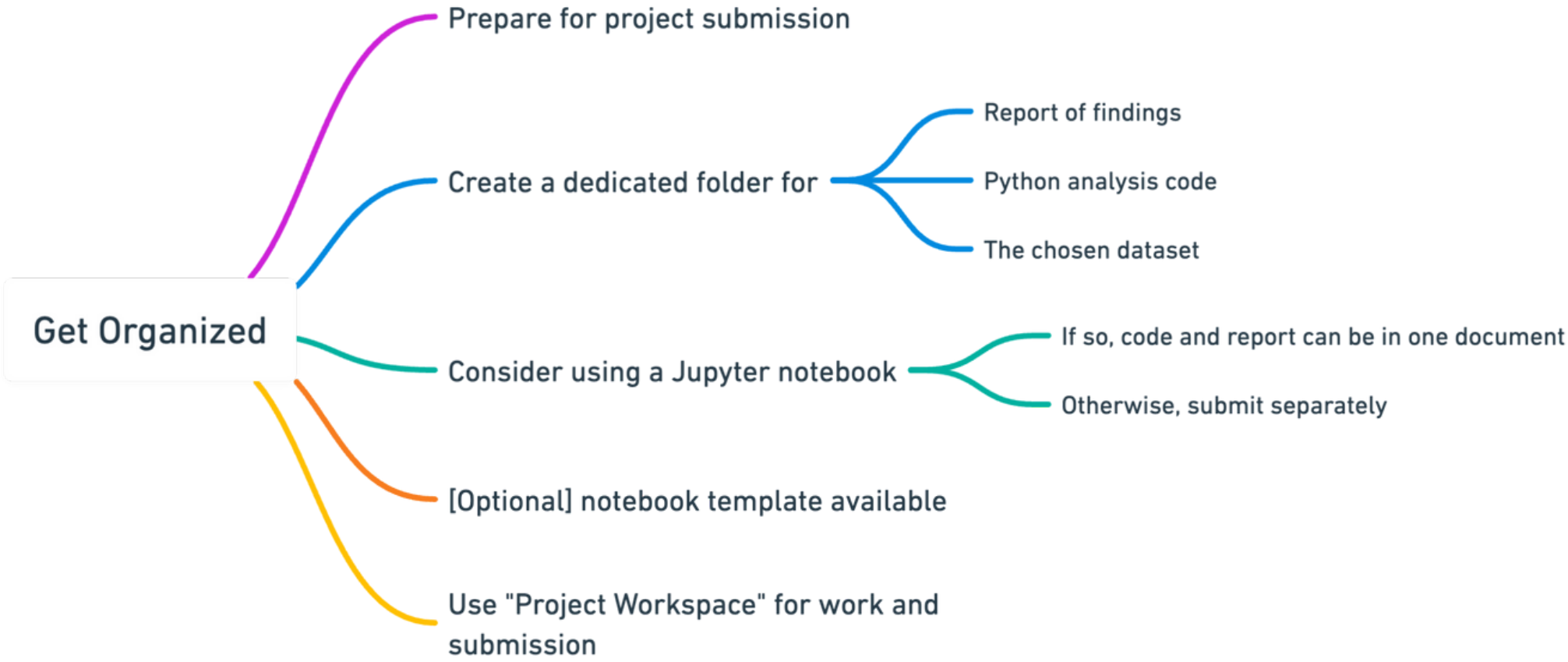
Choose One Dataset

Datasets to Investigate

- **TMDb movie data**
  - Which genres are most popular from year to year?
  - What kinds of properties are associated with movies that have high revenues?
- **No-show appointments**
  - What factors are important for us to know in order to predict if a patient will show up for their scheduled appointment?
- **Gapminder World**
  - Have certain regions of the world been growing in your selected metrics better than others?
  - Are there trends that can be observed between the selected metrics?
- **Soccer Database**
  - What teams improved the most over the time period?
  - Which players had the most penalties?
  - What team attributes lead to the most victories?
- **FBI Gun Data**
  - What census data is most associated with high gun per capita?
  - Which states have had the highest growth in gun registrations?
  - What is the overall trend of gun purchases?

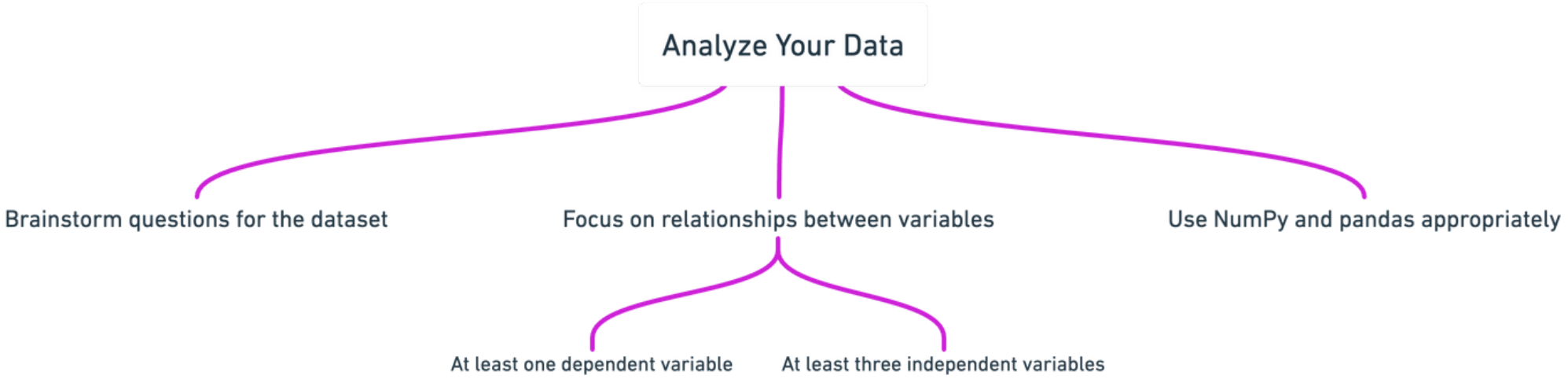
STEP TWO

2


Get Organized
Prepare for project submission
Create a dedicated folder for
Report of findings
Python analysis code
The chosen dataset
Consider using a Jupyter notebook
If so, code and report can be in one document
Otherwise, submit separately
[Optional] notebook template available
Use "Project Workspace" for work and submission

STEP THREE


3


Analyze Your Data
Brainstorm questions for the dataset
Focus on relationships between variables
Use NumPy and pandas appropriately
At least one dependent variable
At least three independent variables

STEP FOUR

4


Share Your Findings
Draft a report of significant findings
If using Jupyter:
Report should be in HTML or PDF format
Share findings and code together
Use Markdown for clarity between comments and code

STEP FIVE

5


Review
Compare against Project Rubric
Refine as needed before final submission

HOW TO COMPLETE AND SUBMIT THIS PROJECT

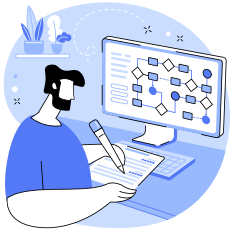
OPTION #1

- Workspace Environment
  - Workspace Details:
    - Work directly on Jupyter Notebook in workspace
    - Saves work automatically
    - Located on the "Project Workspace: Complete and Submit Project" page
  - Submission:
    - Submit directly from the workspace
    - No need to upload .ipynb files; automatically uploaded
    - Directions provided for adding HTML or PDF in the workspace
  - Workspace Features [Note: Tip: Can't download entire folders at once. Use zip command in Jupyter terminal to bundle files]
    - Reset or Restore: via Menu option on the bottom-left
    - Upload, Download, Delete files: through the Jupyter server interface
      - Accessible by clicking on the jupyter icon on the top-left

OPTION #2

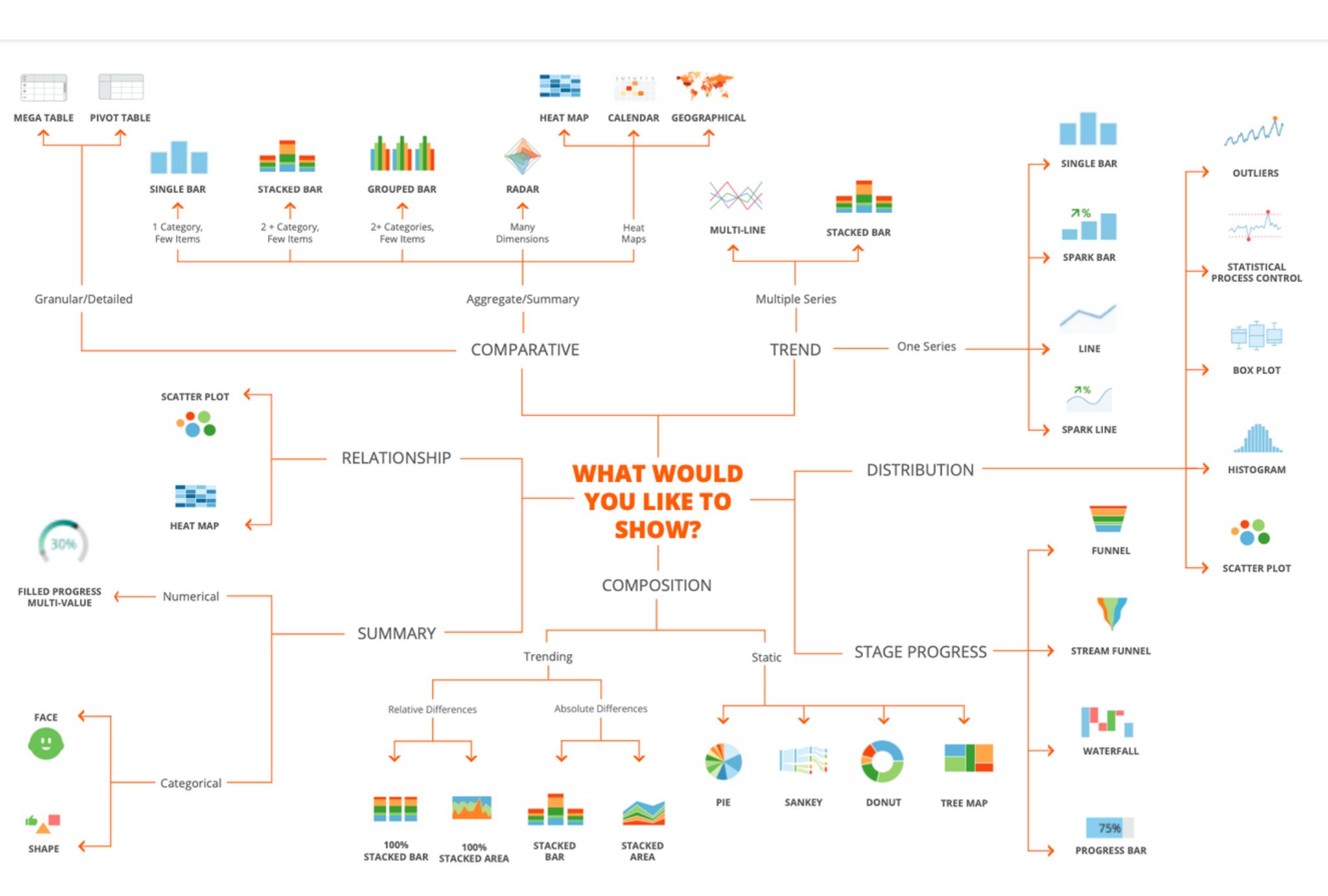
- Work on a local machine
  - Requirements
    - "Investigate_a_Dataset.ipynb" starter file (available for download)
    - Choose one of the databases (multiple options provided)
      - Database_Gapminder_World has specific download instructions
  - Submission
    - Zip project files
    - Use "SUBMIT PROJECT" button on top-right of the last lesson page
    - Follow the submission page instructions

PROJECT RUBRIC

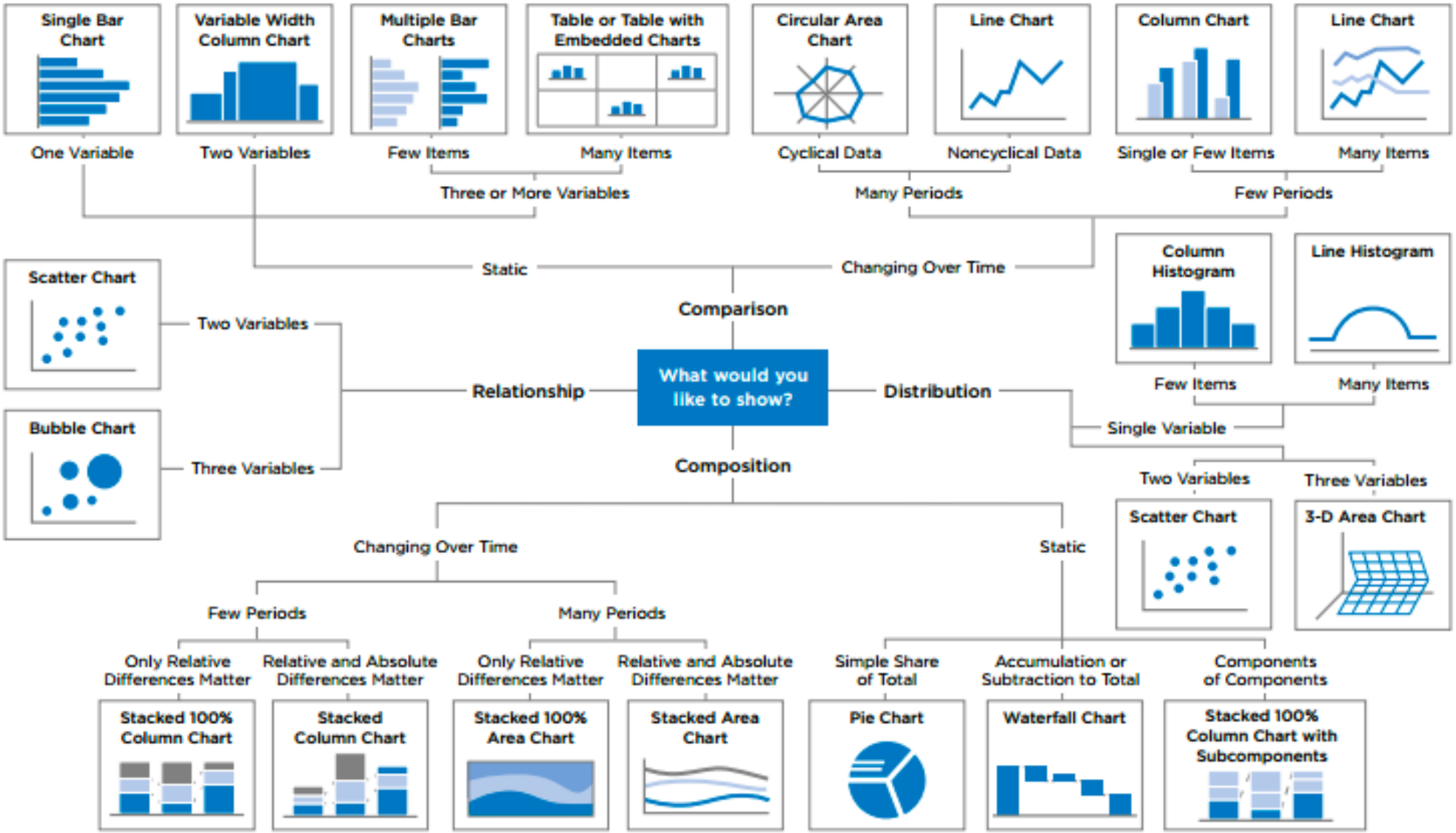


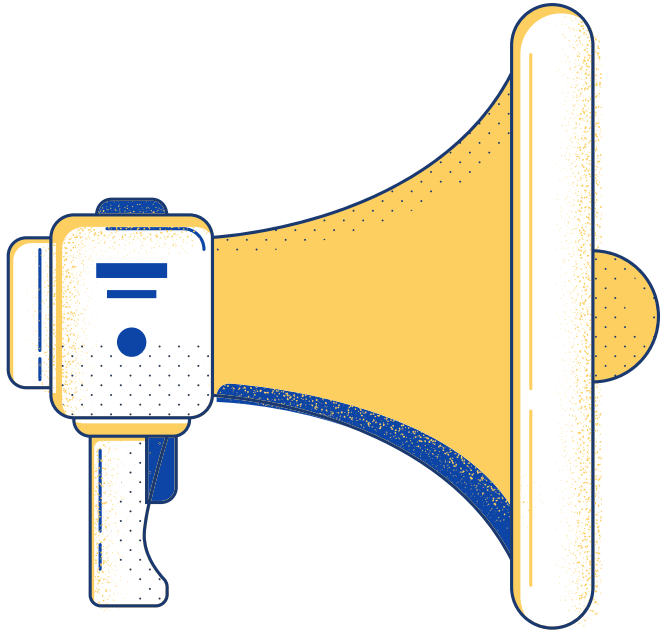
- Project Rubric
  - Code Functionality
    - Does the code work?
      - Code functional without errors
      - Reproducibility of results
    - Use of NumPy and Pandas
      - Use of NumPy arrays and Pandas structures
      - Use vectorized operations and avoid loops
    - Good coding practices
      - Use of functions for repetitive code
      - Code readability: comments and variable names
  - Quality of Analysis
    - Clearly posed question
      - Define and address questions
  - Data Wrangling Phase
    - Data cleaning documentation
      - Document steps like merging and handling missing values
  - Exploration Phase
    - Multi-angle data exploration
      - Investigate questions from multiple angles
        - Explore relationships between variables
        - 1d and 2d explorations
    - Variety in visualizations and statistical summaries
      - Varied visualizations for comparisons and trends
      - At least two types of plots

PAINTING DATA STORIES


MEGA TABLE
PIVOT TABLE
HEAT MAP
CALENDAR
GEOGRAPHICAL
SINGLE BAR
STACKED BAR
GROUPED BAR
RADAR
1 Category, Few Items
2 + Category, Few Items
2+ Categories, Few Items
Many Dimensions
Heat Maps
MULTI-LINE
STACKED BAR
SINGLE BAR
SPARK BAR
LINE
SPARK LINE
OUTLIERS
STATISTICAL PROCESS CONTROL
BOX PLOT
HISTOGRAM
SCATTER PLOT
Granular/Detailed
Aggregate/Summary
Multiple Series
COMPARATIVE
TREND
One Series
SCATTER PLOT
HEAT MAP
RELATIONSHIP
DISTRIBUTION
WHAT WOULD YOU LIKE TO SHOW?
30%
FILLED PROGRESS MULTI-VALUE
Numerical
SUMMARY
COMPOSITION
FUNNEL
STREAM FUNNEL
WATERFALL
75%
PROGRESS BAR
STAGE PROGRESS
Trending
Static
Relative Differences
Absolute Differences
FACE
SHAPE
Categorical
100% STACKED BAR
100% STACKED AREA
STACKED BAR
STACKED AREA
PIE
SANKEY
DONUT
TREE MAP

SELECTING THE APPROPRIATE CHART FOR STRATEGY PRESENTATIONS





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

- Investigate a Dataset
- Data Analysis Process
- Merge dataframes
- Correlation
- Jupyter Notebooks
- Conda cheatsheet
- Anaconda installation
- IpyWidgets
- Markdown Cheatsheet
- Jupyter Keyboard Shortcuts
- Best Practices for Choosing Chart Types

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