Data Analysis Project: A Step-by-Step Guide

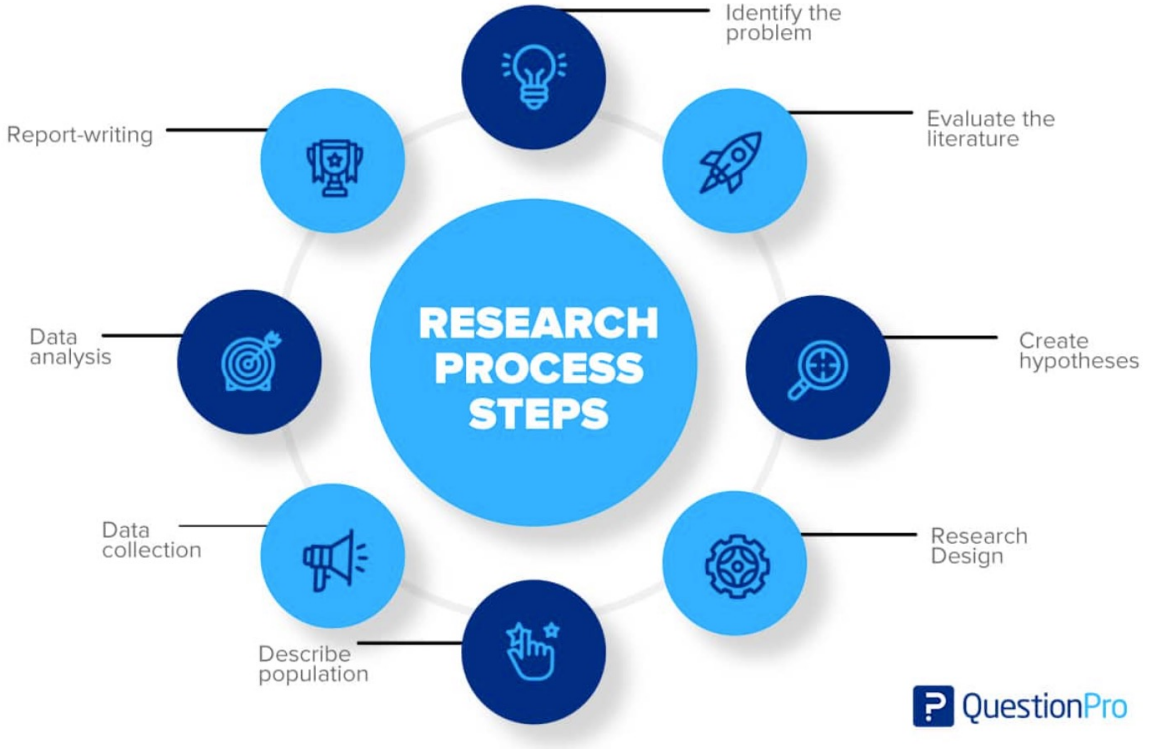
Xuejing Duan

Data Analysis Project: A Step-by-Step Guide

Xuejing Duan

What is a Research Project?

A research project is a systematic investigation aimed at answering a specific research question or testing a hypothesis. It follows rigorous scientific methods to discover new knowledge and contribute to existing understanding.



What is a Data Analysis Project?

Research Project vs. Data Analysis Project

- **Research Project:** Broad scope, focusing on exploring a hypothesis or research question.

- **Data Analysis Project:** A type of research project that specifically focuses on working with data to answer a question or support decision-making.

Key Steps in Your Data Analysis Project

- Step 1: Choose Your Topic
- Step 2: Conduct a Literature Review
- Step 3: Define Your Research Question
- Step 4: Select or Collect Data
- **Step 5: Perform EDA (Exploratory Data Analysis)**
- **Step 6: Clean** and Preprocess **the Data**
- **Step 7: Analyze & Visualize the Data**
- Step 8: Interpret and Present Results

Step 1: Choose Your Topic

- **What is a Topic?**
  - A general area of interest
  - A broad field for investigation
  - Starting point of your project
- **Tips for Topic Selection:**
  - Choose what interests you
  - Consider data availability
  - Think about course scope
  - Use familiar domains
- **Common Mistakes:**
  - Too broad: Avoid topics like "Everything about the economy."
  - No data availability: Ensure data can be accessed for your topic.

Step 2 – Conduct a Literature Review

1. Why Conduct a Literature Review?

1. To understand what research has been done in your chosen topic.
2. Identify gaps, trends, and methods used in previous studies.

2. How to Use the Literature Review:

1. Refine your research question.
2. Understand key variables, methods, and datasets used by others.

In Our Course Project:

- Brief background research
- Learning from similar projects
- Understanding analysis methods
- Getting domain knowledge

Step 3: Define Your Research Question

Research Question: Specific question you want to answer

Examples:

- **Topic**: Student Performance & Learning Patterns
- **Research Questions**:
  - "How does class attendance affect final grades in online vs in-person courses?"
  - "What is the relationship between assignment submission time and grades?"
  - "Do students who participate in study groups achieve higher test scores?”

- **Topic:** Online Food Delivery Services
- **Research Questions:**
  - “How do weather conditions affect delivery times in Westminister?"
  - "What factors influence customer ratings for food delivery?"
  - "Is there a relationship between delivery distance and customer satisfaction?"

Step 3 - Define Your Research Question

Strong Research Questions:

✅ Specific & focused

✅ Answerable with data

✅ Clear variables to analyze

✅ Manageable scope

Weak Research Questions:

❌ "How does weather affect business?" (too vague)

❌ "Why do some students get good grades? " (too broad)

❌ "What is the future of food delivery?" (not data-focused)

Question Check:

1. Can I measure/quantify this?
2. Do I have access to relevant data?
3. Can I answer this in project timeframe?
4. Can I use course tools to analyze?

Step 4 – Finding Your Data: **Data Sources**

Why is Data Important?

- Data is the foundation of your entire project.
- Without quality data, you cannot answer your research question.

Where to Find Data?

- **Public Data Sources:**
  - Kaggle (variety of datasets on many topics).
  - UCI Machine Learning Repository.
  - Government open data portals (e.g., data.gov).
- **Company/Internal Data:**
  - If you have access to internal or proprietary data (from work or other sources).

Step 4 – Finding Your Data: **Evaluating a Dataset**

- **Sample Size:** The dataset should have enough rows and columns for meaningful analysis.
- **Completeness:** Avoid datasets with too many missing values or outliers.
- **Documentation:** Well-documented datasets help you understand the context and variables.
- **Contains Needed Variables**: Ensure the dataset includes all the necessary variables to answer your research question.
- **Recent Data (if relevant):** For analyses where timing matters (e.g., market trends), the data should be recent enough to remain relevant.

Step 4 – Finding Your Data: **Common Pitfalls**

Size Problems

❌ "This dataset only has 30 rows"

❌ "My dataset has 1 million rows"

❌ “I have 100 different features”

❌ “ This dataset only has 40 rows but 70 features ”

Quality Issues

❌ " 70% of my data is missing"

❌ "The dates are all different formats"

❌ “ Poor documentation: I don't know what these columns mean”

❌ "Irrelevant features"

❌ "Too many text columns"

Step 5 –**Exploratory Data Analysis (EDA)**

What is EDA?

- EDA is the process of analyzing and visualizing data to understand its main characteristics before formal modeling.
- It helps uncover patterns, spot anomalies, test hypotheses, and check assumptions.

- **Why is EDA Important?**
  - Identify Data Issues
  - Understand Variable Relationships
  - Shape Data Cleaning Decisions

Step 5 – **(EDA) Key Steps in EDA**

- **1. Data Overview**
- Check:
  - Number of rows
  - Number of columns
  - Data types
  - Basic structure

How do study hours, attendance rate, and high school math scores affect first-year college exam scores? And does family income also have an impact?

```
[11]: df.shape

[11]: (300, 12)

[12]: df.dtypes

[12]: Student_ID                int64
      Age                       int64
      Gender                   object
      Study_Hours_Per_Week    float64
      Attendance_Rate         float64
      High_School_Math_Score  float64
      Family_Income            object
      Parent_Education         object
      Distance_to_School      float64
      Part_Time_Job            object
      Sleep_Hours             float64
      Exam_Score              float64
      dtype: object
```

```
[13]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 300 entries, 0 to 299
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Student_ID              300 non-null    int64
 1   Age                     300 non-null    int64
 2   Gender                  300 non-null    object
 3   Study_Hours_Per_Week    270 non-null    float64
 4   Attendance_Rate         300 non-null    float64
 5   High_School_Math_Score  300 non-null    float64
 6   Family_Income           300 non-null    object
 7   Parent_Education        300 non-null    object
 8   Distance_to_School      300 non-null    float64
 9   Part_Time_Job           300 non-null    object
 10  Sleep_Hours             286 non-null    float64
 11  Exam_Score              300 non-null    float64
dtypes: float64(6), int64(2), object(4)
memory usage: 28.3+ KB
```

Step 5 – **(EDA) Key Steps in EDA**

2. Descriptive Statistics

Look at:

- Mean, median, mode
- Min, max values
- Standard deviation
- Value counts
- Frequency

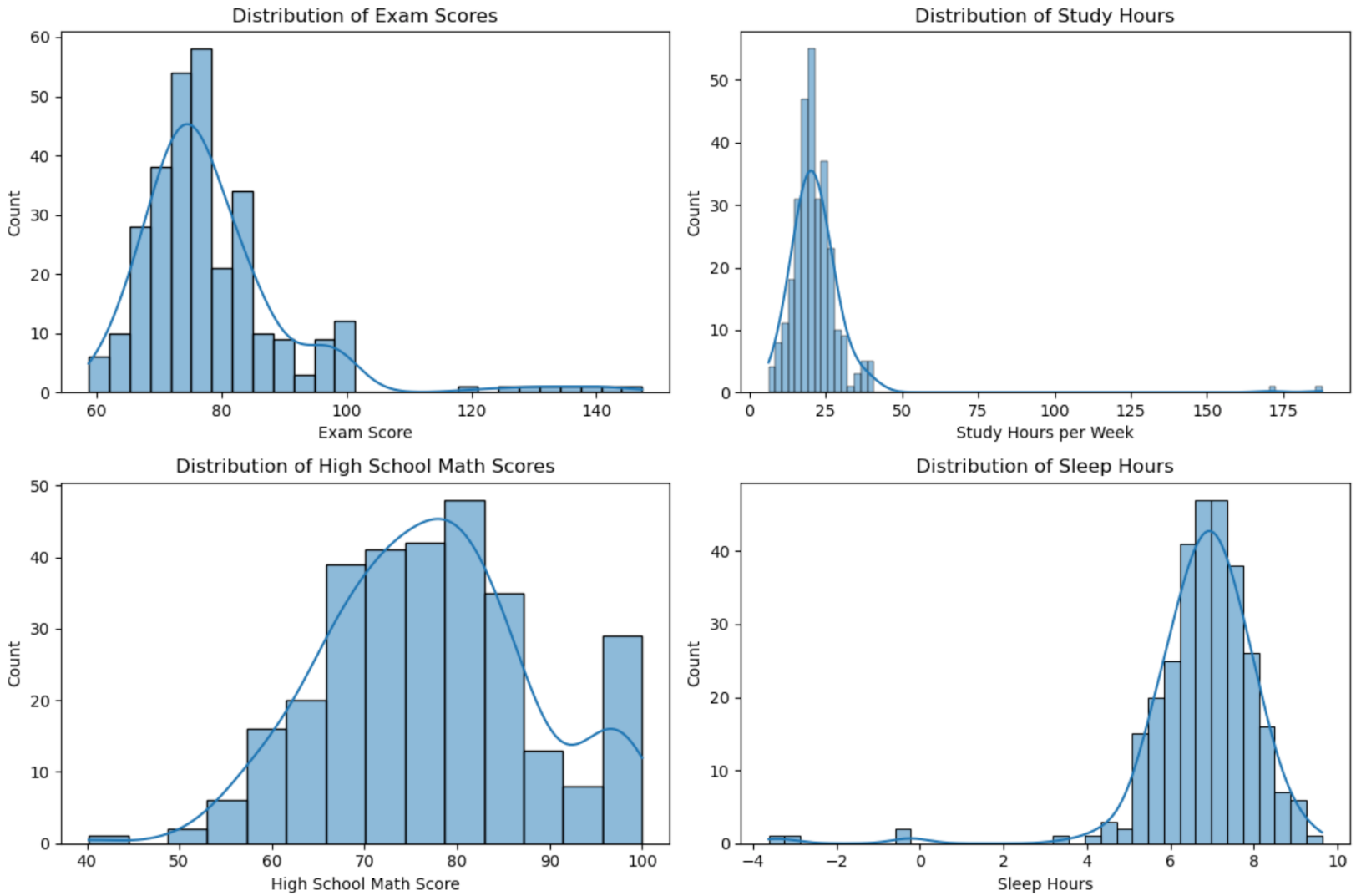
Descriptive Statistics:

	Age	Study_Hours_Per_Week	Attendance_Rate	High_School_Math_Score	Exam_Score	Sleep_Hours
count	300.00	300.00	300.00	300.00	300.00	300.00
mean	21.04	22.09	0.80	77.10	78.51	6.78
std	2.01	14.33	0.12	11.17	12.79	1.40
min	18.00	6.30	0.60	40.19	58.67	-3.63
25%	19.00	17.27	0.69	69.44	71.56	6.25
50%	21.00	20.33	0.80	77.13	75.94	6.90
75%	23.00	24.58	0.89	83.89	82.86	7.55
max	24.00	187.69	1.00	100.00	147.34	9.63

Step 5 – **(EDA) Key Steps in EDA**

3. Distribution Check

- Examine:
  - Data spread
  - Outliers
  - Patterns
  - Unusual values



Step 5 – **(EDA) Key Steps in EDA**

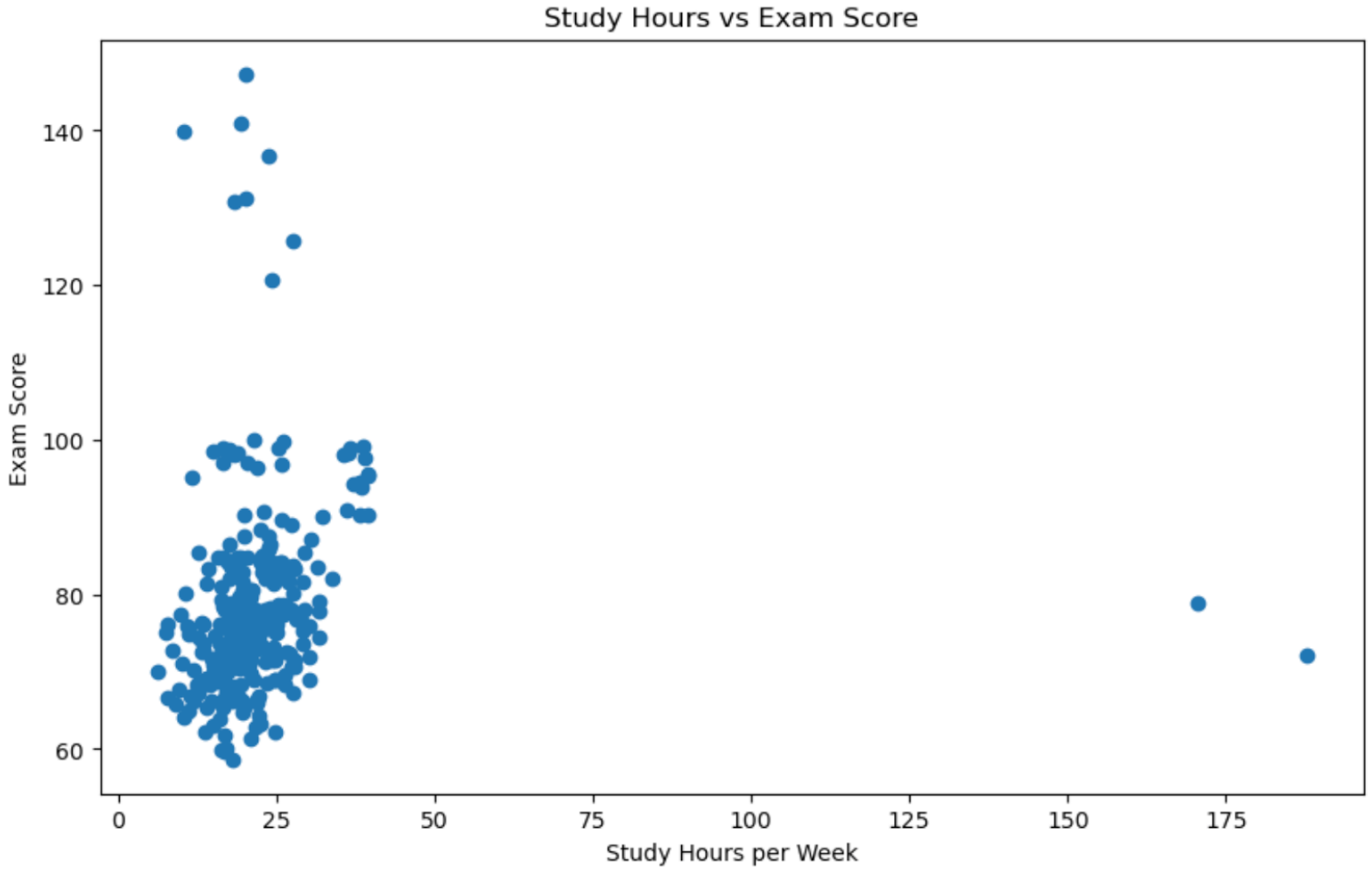
- **4. Relationships**
- Look for:
  - Correlations, Patterns, Groups, Trends

Correlation Matrix after cleaning:

	Exam_Score	Study_Hours_Per_Week	Attendance_Rate	High_School_Math_Score	Sleep_Hours
Exam_Score	1.000	0.432	0.341	0.389	-0.016
Study_Hours_Per_Week	0.432	1.000	-0.034	0.177	0.038
Attendance_Rate	0.341	-0.034	1.000	-0.080	-0.072
High_School_Math_Score	0.389	0.177	-0.080	1.000	0.049
Sleep_Hours	-0.016	0.038	-0.072	0.049	1.000

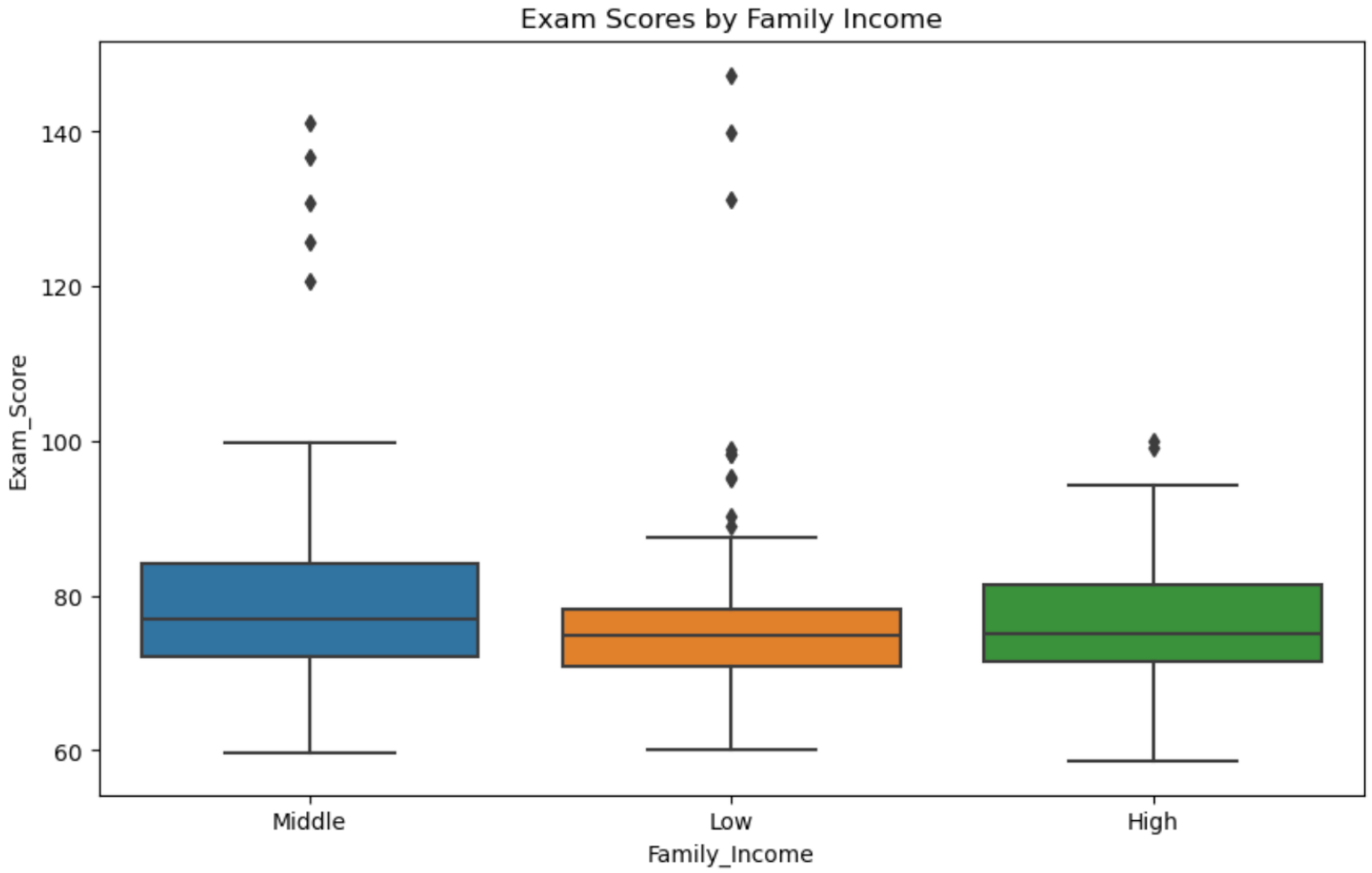
Step 5 – **(EDA) Key Steps in EDA**

- **4. Relationships**
- Look for:
  - Correlations
  - Patterns
  - Groups
  - Trends



Step 5 – **(EDA) Key Steps in EDA**

- **4. Relationships**
- Look for:
  - Correlations
  - Patterns
  - Groups
  - Trends



Step 5 – **What to Look for in EDA?**

- **Data Issues:**
  - Missing values
  - Outliers
  - Unusual patterns
  - Inconsistencies
- **Data Insights:**
  - Value ranges
  - Common values
  - Relationships
  - Distributions

Step 6 – **What is Data Cleaning?**

What is Data Cleaning?

- **Detect and correct** errors, inconsistencies, and inaccuracies in data.
- **Remove invalid** or unusable data entries.
- **Prepare the data** so it is reliable and ready for analysis or modeling.

Poor Data Leads to:

- Wrong conclusions
- Misleading results
- Complexity in processing
- Failed analysis

Step 6 – **Common Data Cleaning Tasks**

- **Handle Missing Values:**
  - Remove rows with missing data (if appropriate).
  - Impute missing values using techniques like mean, median, or mode.
- **Handle Outliers:**
  - Detect and decide whether to remove or keep outliers based on the context.
- **Remove Duplicates:**
  - Identify and eliminate duplicate records to avoid skewing results.
- **Select or Remove Irrelevant Variables:**
  - Remove features that do not contribute to the analysis or prediction.
  - Focus on the most relevant variables to reduce noise and improve model performance.

Step 6 – **Handle Missing Values**

- **Types of Missing Data:**
  - Completely Empty (NULL)
  - Placeholders ("N/A", "Unknown")
  - Spaces or Special Characters("? ")
- **Consider:**
  - How many are missing?
  - Why are they missing?
  - Is the missing data random?
- **Key Decisions:**
  - Remove missing values?
  - Fill in (impute) values?
  - Keep as missing?

Step 6 – **Handle Outliers**

- **Identify Outliers:**
  - Extreme values
  - Impossible values
  - Suspicious patterns

- **Key Questions:**
  - Is it a real value?
  - Is it an error?
  - Should we keep it?

Examples:

Obviously Incorrect:

❌ Age = 200 (impossible for humans)

❌ Income = -5000 (cannot be negative)

❌ Temperature = 1000°C (too high for most scenarios)

Requires Verification:

❓ Age = 100 (rare but possible)

❓ House price = $10 million (unusual but could be luxury home)

❓ Student study hours per week = 80 (extreme but possible during finals)

❓ Running speed = 25mph (could be elite athlete)

Step 6 – **Select Relevant Variables**

What to Remove?

1. Irrelevant Variables

- Not related to research question
- No logical connection
- Too indirect

2. Redundant Variables

- Duplicate information
- Highly correlated features
- Derived from other variables

3. Low Quality Variables

- Too many missing values
- Poor quality data
- Unreliable collection

Examples:

House Price Analysis:

❌ Remove:

- House owner's name
- ID number
- Listing agent's birthday

✅ Keep:

- Square footage
- Number of rooms
- Location
- Year built
- Recent sale prices

Step 6 – **Fix Inconsistent Values**

- **Category Inconsistencies:**
  - "Male" vs "M" vs "male"
  - "NY" vs "New York"
  - "YES" vs "Y" vs "1”
- **Number Format Issues:**
  - "1,000" vs "1000"
  - "$1000" vs "1000"

Example:

If our gender column has:

- "Male"
- "male"
- "MALE"
- "female"

- Problem: Computer treats these as different categories!
- When we count or analyze by gender:
  - We get 4 groups instead of 2!

Step 6 – **Data Preprocessing**

- **Feature Encoding:** Convert categorical data into numerical values (e.g., one-hot encoding).
- **Data Normalization:** Scale numerical features to a consistent range, especially for models sensitive to scale.
- **Text Preprocessing (if relevant):** Tokenization, removing stop words, or stemming for text-based features.
- **Feature Selection:** Select the most important variables and remove irrelevant or redundant features.

Step 7 – **Analyze & Visualize the Data**

- **Extract Insights:**
  - Identify patterns and trends
  - Summarize the data in a meaningful way
  - Uncover relationships between variables

- **Answer Research Questions:**
  - Use the analysis to directly address your research objectives
  - Provide data-driven answers
  - Support your findings with visualizations or model results

Step 7 –**Types of Data Analysis**

1. Descriptive Analysis

"What happened?"

- Summary statistics
- Patterns and trends
- Data distributions
- Basic insights

2. Statistical Analysis

"How significant?"

- Hypothesis testing
- Correlation analysis
- Regression models
- ANOVA

3. Machine Learning

"What patterns/predictions?"

- Classification
- Clustering
- Prediction
- Pattern recognition

Step 7 – **Data Visualization**

Why Visualize Data?

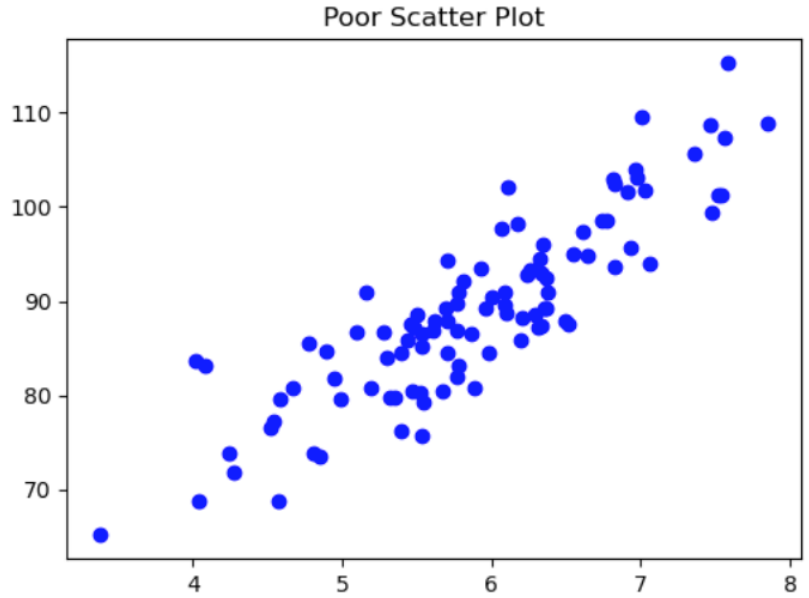
- Helps identify patterns, trends, and outliers.
- Makes data easier to understand and communicate.

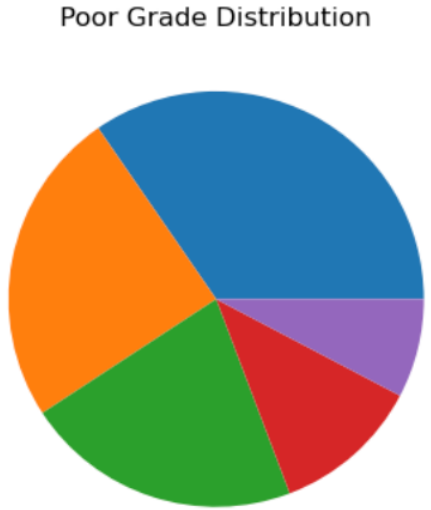
Tools for Data Visualization:

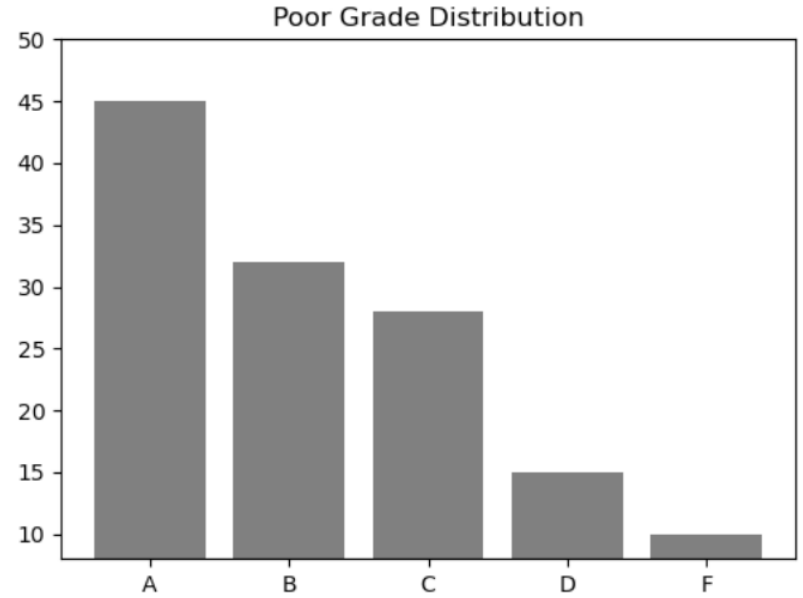
- Use **matplotlib** and **seaborn** for simple, customizable visualizations.
- Use **pandas** built-in plotting functions for quick visualizations.

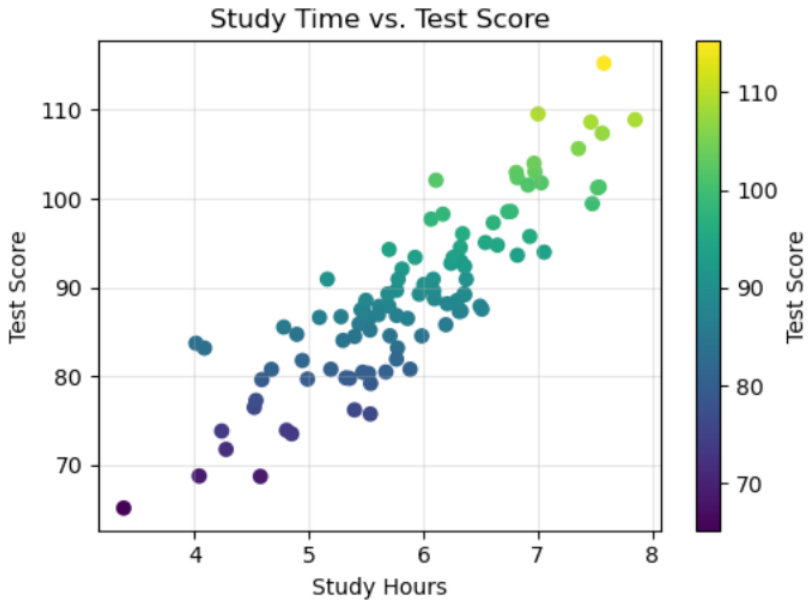
Step 7 – **Data Visualization**

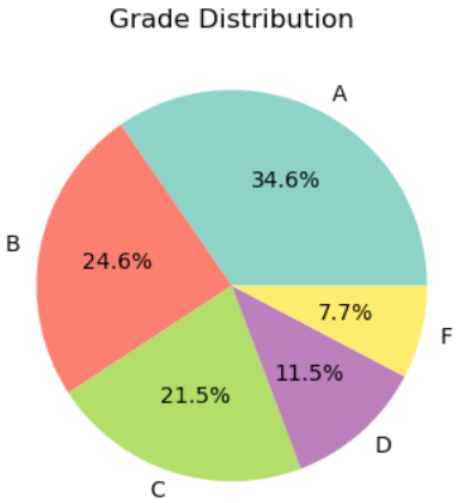
- **Common Visualization Techniques:**
  - **Histogram**: show the distribution of data
  - **Box Plot**: show data spread and identify outliers.
  - **Scatter Plot**: display the relationship between two variables
  - **Bar Chart**: compare categories
  - **Heatmap**: show the correlation between variables, usually in a correlation matrix.
  - **Line Chart**: show trends over time

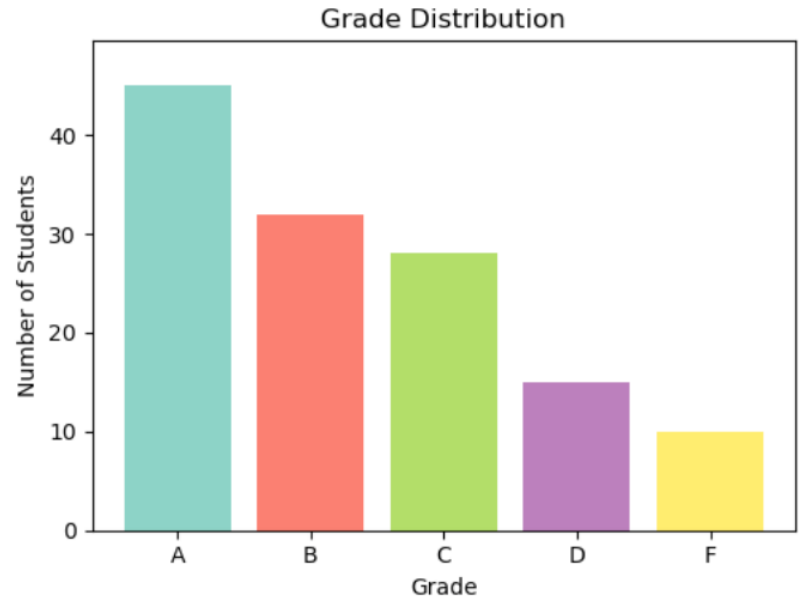

Poor Scatter Plot
110
100
90
80
70
4
5
6
7
8


Poor Grade Distribution

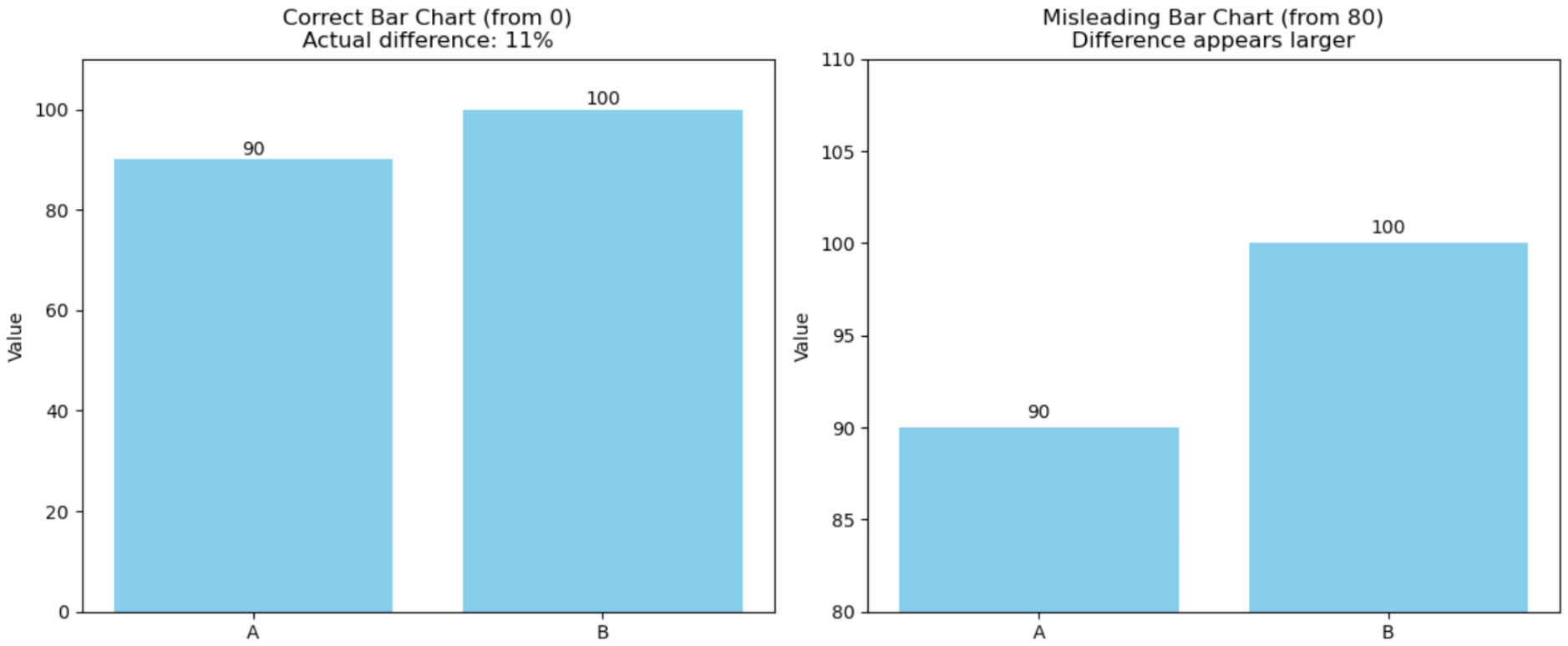

Poor Grade Distribution
50
45
40
35
30
25
20
15
10
A
B
C
D
F


Study Time vs. Test Score
Test Score
Study Hours
110
100
90
80
70
4
5
6
7
8

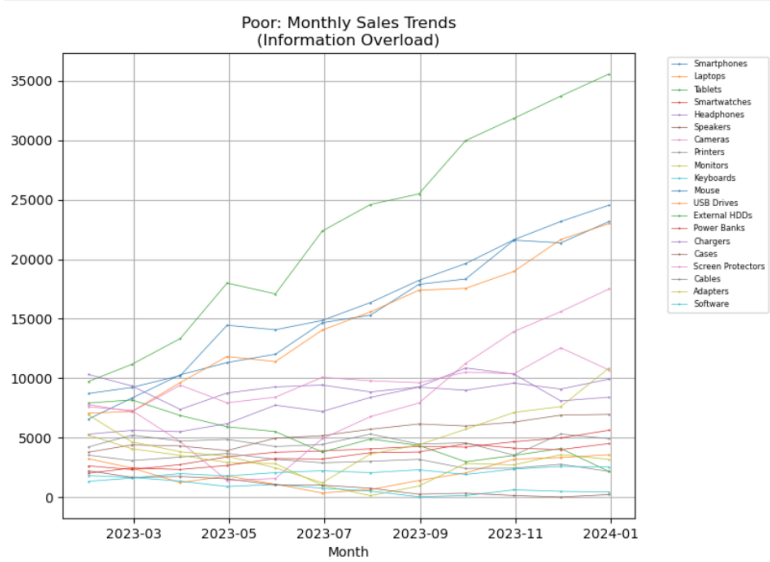

Grade Distribution
A
B
C
D
F
34.6%
24.6%
21.5%
11.5%
7.7%

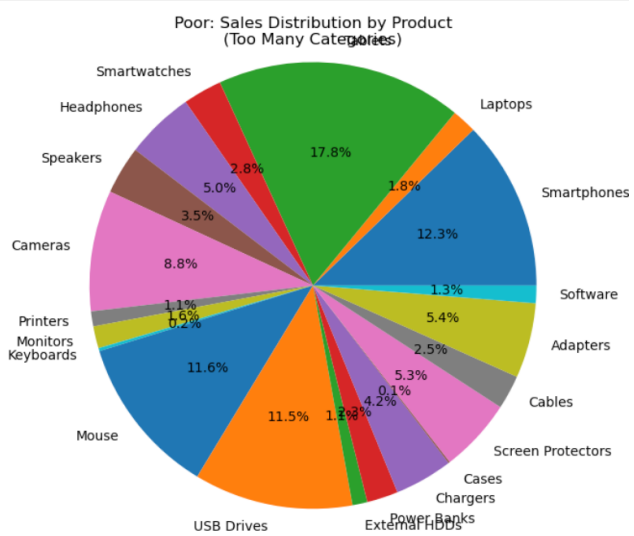

Grade Distribution
Number of Students
Grade
40
30
20
10
0
A
B
C
D
F

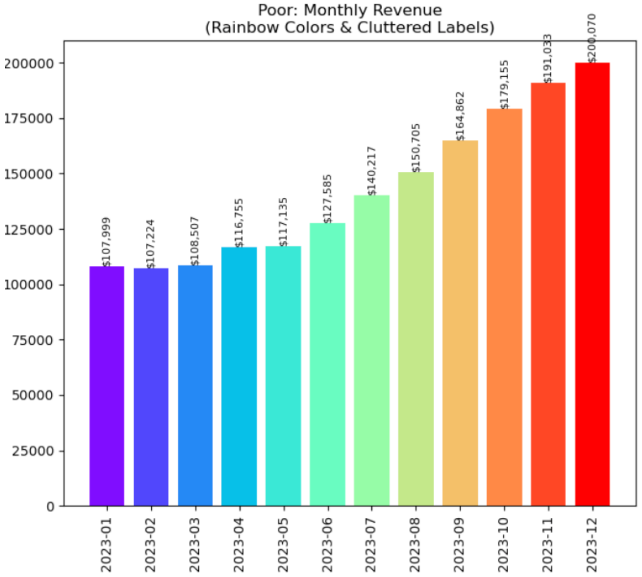
Bar Chart Zero Base Comparison

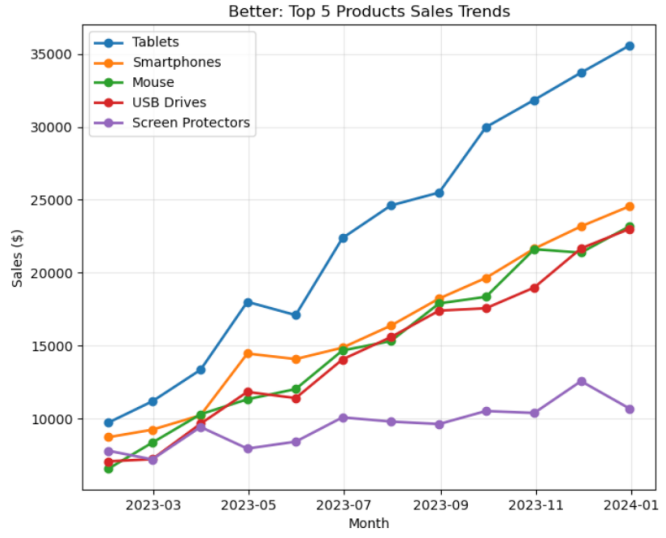


Overcrowded and Poor Data Visualization Examples









Step 8 – **Interpret and Present Results**

Understanding Your Results

- **About the Analysis**
  - Are results reasonable?
  - Statistically significant?
  - What patterns emerged?
  - Any unexpected findings?
- **About Context**
  - How does it fit with research?
  - What factors influence results?
  - What are the limitations?
  - What might you have missed?

Step 8 – **Interpret and Present Results**

Presenting Your Findings & Write the Final Report

1. **Abstract:**
   - Provide a brief summary of your entire project, including the research question, methods, key findings, and conclusion. Typically 150–250 words.
2. **Introduction:**
   - Recap your research question and explain the significance of the problem you are addressing.
   - Briefly mention the data and methods used.
3. **Methodology:**
   - Summarize the steps you took during data cleaning, preprocessing, and analysis.
   - Include any tools or models used (e.g., EDA, statistical models, machine learning algorithms).

Step 8 – **Interpret and Present Results**

Presenting Your Findings & Write the Final Report

4. **Results:**
    1. Present your key findings with supporting visualizations and statistical evidence.
    2. Ensure charts and graphs are clearly labeled and easy to interpret.

5. **Conclusion:**
    1. Provide final insights based on your analysis.
    2. Discuss any limitations of your data or methods (e.g., sample size, missing data, assumptions).
    3. Suggest directions for future research or practical applications of your findings.

Step 8 – **Interpret and Present Results**

Presenting Your Findings & Write the Final Report

6. **References:**
   - List all data sources, tools, and any academic or external references used in your analysis.
   - Ensure proper citation format (e.g., APA, MLA, etc.).

7. **Appendices (if necessary):**
   1. Include additional visualizations, data tables, or supplementary materials that support your findings but are too detailed for the main report.

Any Questions?

Time to Start Your Project!