Comprehensive Assignment Guide

Exploratory Data Analysis, Statistical Exploration & Visualization Using Python

1 Intent and Scope of This Assignment

This assignment is intended to simulate how data is explored, questioned, validated, and communicated in real industry environments. Students are expected to behave as analysts who have been given an unfamiliar dataset and asked to extract trustworthy insights.

This guide goes beyond basic exploratory data analysis. It integrates statistical reasoning, dimensionality reduction, clustering, and visual interpretation into a single coherent exploratory workflow.

Students may use any structured dataset of their choice. The evaluation focuses on analytical rigor rather than dataset complexity.

2 Analyst Mindset and Expectations

Throughout this assignment, students are expected to:

- Ask questions continuously, not only at the beginning
- Treat every plot as an analytical argument
- Validate visual impressions using statistics
- Recognize uncertainty and analytical limitations
- Separate observation from interpretation

Every section explicitly requires formulation of analytical questions.

3 Required Technical Stack

The following Python libraries are mandatory:

- pandas
- numpy
- matplotlib
- seaborn
- plotly
- scikit-learn
- scipy and/or statsmodels

4 Universal Exploratory Workflow (Foundational Framework)

All activities must follow this iterative workflow:

1. Observe the data structure
2. Ask analytical questions
3. Visualize appropriately
4. Quantify patterns statistically
5. Validate or reject assumptions
6. Refine questions and repeat

Skipping steps or treating this as a linear checklist is discouraged.

5 Section A: Structural and Contextual Exploration

5.1 Data Context and Generation

Students must answer the following:

- What real-world process produced this data?
- Who or what is represented by each row?
- What does one observation truly mean?
- What factors might influence data quality?

All assumptions regarding data origin and context must be explicitly stated.

5.2 Data Structure and Integrity

Required analyses include:

- Dataset dimensions and schema
- Data type validation and correction
- Missing value patterns across variables
- Duplicate and inconsistent record detection

Required analytical questions include:

- Are missing values random or systematic?
- Which variables are reliable enough for analysis?
- Which variables may distort results due to scale or sparsity?

6 Section B: Univariate Analysis and Question Expansion

6.1 Numerical Variable Exploration

For each numerical variable, students must ask:

- What is the typical value?
- How dispersed is the data?
- Is the distribution symmetric or skewed?
- Are extreme values meaningful or erroneous?

Required analyses include:

- Histogram and boxplot
- Mean, median, and standard deviation
- Quantile analysis and tail behavior

6.2 Categorical Variable Exploration

For each categorical variable, students must ask:

- Which categories dominate the data?
- Are there rare categories worth investigating?
- Does category imbalance affect interpretation?

Required analyses include:

- Frequency plots
- Proportion tables
- Conceptual discussion of category diversity

7 Section C: Bivariate Relationships and Statistical Validation

7.1 Numerical–Numerical Relationships

Required analytical questions include:

- Is there an apparent relationship?
- Is the relationship linear or non-linear?
- Is the relationship strong or weak?

Required analyses include:

- Scatter plots with trend indication
- Correlation coefficients
- Confidence-aware interpretation of results

7.2 Numerical–Categorical Relationships

Required analytical questions include:

- Does category membership shift distributions?

- Are observed differences meaningful or random?

Required analyses include:

- Boxplots or violin plots
- Group-wise descriptive statistics
- Appropriate statistical comparison tests

7.3 Categorical–Categorical Relationships

Required analytical questions include:

- Are the variables independent?
- Does one category dominate outcomes?

Required analyses include:

- Stacked bar charts
- Contingency tables
- Association measures where applicable

8 Section D: Multivariate Analysis and Feature Interactions

8.1 Conditional Exploration

Students must ask:

- Does a relationship persist under conditioning?
- Does introducing a third variable change conclusions?

Required analyses include faceted plots and color-encoded scatter plots.

8.2 Statistical Interaction Reasoning

Students must discuss:

- Possible interaction effects
- Limitations of pairwise analysis
- Risks of overinterpretation

9 Section E: Advanced Statistical Exploration

9.1 Distribution Comparison

Students must ask:

- Are two distributions statistically distinguishable?

- Does visual difference imply statistical difference?

Required analyses include:

- Distribution overlays
- Appropriate statistical tests
- Effect size interpretation

9.2 Hypothesis Formulation and Testing

Students must:

- Formulate at least three exploratory hypotheses
- Select appropriate statistical tests
- Clearly state assumptions
- Interpret results in practical terms

Hypothesis testing must be treated as exploratory rather than confirmatory.

9.3 Correlation Versus Causation Reflection

Students must identify:

- At least one misleading correlation
- Possible confounding variables
- Additional data required for causal inference

10 Section F: Dimensionality Reduction and PCA

10.1 Motivation and Readiness Check

Students must ask:

- Is dimensionality reduction appropriate?
- What information may be lost?

10.2 PCA Execution and Interpretation

Required analyses include:

- Feature scaling justification
- Explained variance analysis
- Interpretation of principal components

10.3 PCA-Based Questioning

Students must ask:

- Do natural groupings appear?
- Are patterns clarified or obscured?
- Which variables dominate variance?

11 Section G: Clustering for Exploratory Insight

11.1 Clustering Motivation

Students must ask:

- Why might natural groupings exist?
- What domain meaning could clusters have?

11.2 Clustering Execution

Required analyses include:

- Application of at least two clustering methods
- Visualization in original and reduced spaces

11.3 Cluster Validation and Skepticism

Students must ask:

- Are clusters stable?
- Are they artifacts of scaling or method choice?
- How would domain knowledge validate them?

12 Section H: Visualization Design, Ethics, and Perception

12.1 Aesthetic Optimization

Students must refine visualizations considering:

- Color semantics
- Scale integrity
- Label clarity

12.2 Ethical Risk Assessment

Students must ask:

- Could this visualization mislead?
- What assumptions does it impose on viewers?

13 Section I: Reading, Critiquing, and Stress-Testing Visuals

13.1 Self-Critique

For each major visualization, students must state:

- Core message
- Hidden information
- Potential misinterpretations

13.2 External Visualization Critique

Students must analyze one published visualization and identify:

- Strengths
- Weaknesses
- Ethical concerns

14 Section J: Interactive Visualization and Exploratory Tools

14.1 Interactive Question Expansion

Students must ask:

- What questions cannot be answered statically?
- How does interaction change exploration?

Required features include hover-based detail, dynamic filtering, and zoom or drill-down capability.

15 Section K: Tool Comparison and Industry Reflection

15.1 Same Insight, Different Tools

Students must recreate one insight using multiple visualization libraries and compare:

- Development effort
- Analytical clarity
- Suitability for stakeholders

16 Final Submission Expectations

Students must submit:

- Well-organized Python notebooks
- Integrated explanations alongside code
- Clear distinction between observation and inference
- Reproducible analysis workflow

17 Evaluation Philosophy

Evaluation will emphasize:

- Quality of questions asked
- Depth of statistical reasoning
- Strength of visual arguments
- Awareness of limitations
- Professional analytical communication

End of Assignment Guide