AI-Powered Exploratory Data Analysis (EDA) Cheat Sheet

Unlock Data Insights Faster with ChatGPT and Copilot

Introduction:

Exploratory Data Analysis (EDA) is the first step in understanding your data. This cheat sheet provides a structured approach to EDA, using AI to accelerate the process. Always clean your data before starting EDA. Replace bracketed placeholders ([column name]) with your specific data.

I. Getting Started: Initial Data Overview

Before diving deep, understand your dataset's structure.

1. Dataset Description:

- **Task:** Get basic information (rows, columns, data types, purpose).

Prompt (ChatGPT):

- "*Describe the dataset I've uploaded. Provide the number of rows, columns, column names, and data types for each column. Briefly summarize the apparent purpose of the dataset.*"

Prompt (CoPilot/Excel):

- "*Summarize this data.*" *(Review the summary)*
- **When to Use:** Always the first step with any new dataset.

2. Missing Value Analysis:

- **Task:** Determine the extent and pattern of missing values.

Prompt (ChatGPT):

- "*Identify any columns with missing values. For each, report the number and percentage of missing entries. Is there any apparent pattern to the missingness?*"

Prompt (CoPilot/Excel):

- "*List all columns with missing values, and count the number of missing entries in each.*" (Use COUNTBLANK or a PivotTable.)

- **When to Use:** Essential for understanding data quality and planning data cleaning.

3. Descriptive Statistics (Numerical Columns):

- **Task:** Calculate key statistics (mean, median, std. dev., min, max, quartiles).

Prompt (ChatGPT):

- *"Generate descriptive statistics for these numerical columns: [list column names]. Include mean, median, standard deviation, minimum, maximum, and quartiles."*

Prompt (CoPilot/Excel):

- *"Provide descriptive statistics for [column name(s)]." (Or use Excel's Data Analysis Toolpak.)"*

- **When to Use:** To quickly understand the distribution, range, and typical values of numerical features. Useful for:
  - Initial data understanding.
  - Identifying potential data quality issues (e.g., unrealistic min/max values suggesting errors).
  - Gauging the scale and spread of variables before further analysis or modeling.

4. Unique Value Counts (Categorical/Text Columns):

- **Task:** Determine the number of unique values and their frequencies.

Prompt (ChatGPT):

- *"For each of these columns: [list column names], report the number of unique values and list the top 10 most frequent values. Also, identify any categories with very low frequency (e.g., less than 1% of total entries)."*

Prompt (CoPilot/Excel):

- Use Excel's UNIQUE and COUNTIF functions or PivotTables. Ask Copilot: *"Create a formula to count the unique values in the 'Category' column."*

- **When to Use:** When to Use: When working with categorical or text-based columns to:
  - Verify expected categories and identify unexpected values (potential errors).

- Assess the cardinality of categorical features (important for feature engineering).
- Detect data entry inconsistencies.

II. Getting Started: Initial Data Overview

Explore relationships, trends, and group differences.

5. Time Series Analysis (If Applicable)::

- **Task:** Analyze data over time (trends, seasonality, cycles).

Prompt (ChatGPT):

- *"Analyze the 'Order Date' and 'Sales' columns. Identify any trends, seasonal patterns, or cyclical fluctuations. Create a line chart. Decompose the time series into trend, seasonal, and residual components if possible."*

Prompt (CoPilot/Excel):

- *"Create a line chart showing 'Sales' over time, using 'Order Date' as the x-axis. Add a trendline."* (Visually inspect and ask follow-up questions.)

- **When to Use:** When your dataset includes a time-based component and you need to:
  - Understand trends, seasonality, and cycles in data over time (e.g., sales, website traffic, stock prices).
  - Forecast future values based on historical patterns.
  - Identify anomalies or deviations from expected seasonal behavior.

6. Correlation Analysis (Numerical Columns):

- **Task:** Determine linear relationships between numerical variables.

Prompt (ChatGPT):

- *"Calculate the correlation matrix for these numerical columns: [list column names]. Highlight any strong positive or negative correlations (above 0.7 or below -0.7). Explain the meaning of these correlations. Also, calculate the VIF for each column to assess multicollinearity."*

Prompt (CoPilot/Excel):

- *"Calculate the correlation matrix for [list of columns]."*(Use Data Analysis Toolpak or CORREL.)

- **When to Use:** When you want to explore linear relationships between numerical variables to:
  - Identify potential predictors for a target variable (for regression or feature selection).
  - Understand how variables move together (e.g., price and demand).
  - Detect multicollinearity issues in regression modeling (highly correlated predictors).
  - Form hypotheses about relationships that warrant further investigation.

7. Correlation Analysis (Numerical Columns):

- **Task:** Compare numerical values across different categories.

Prompt (ChatGPT):

- *"Compare the average 'Sales Amount' for each 'Product Category.' Are there significant differences? Present results in a table and a bar chart. Perform an ANOVA test if appropriate."*

Prompt (CoPilot/Excel):

- *"Create a pivot table showing average 'Sales Amount' by 'Product Category.' Add a bar chart."*

- **When to Use:** When you want to compare a numerical metric across different categories to:
  - Identify performance differences across groups (e.g., sales by region, customer satisfaction by product type).
  - Understand how categorical factors influence numerical outcomes.
  - Segment your data based on categories for targeted analysis or strategies.

8. Correlation Analysis (Numerical Columns):

- **Task:** Understand the shape and spread of numerical data.

Prompt (ChatGPT):

- *"Generate a histogram and box plot for the 'Customer Age' column. Describe the distribution (normal, skewed, bimodal, multimodal). Are there any significant deviations from normality? Is there evidence of zero inflation?"*

Prompt (CoPilot/Excel):

- *"Create a histogram for the 'Customer Age' column."* and *"Create a box plot for the 'Customer Age' column."*

  - **When to Use:** When you need to understand the shape and spread of numerical data to:
    - Assess normality (an important assumption for some statistical tests).
    - Identify skewness and modality (uni-modal, bi-modal, etc.)
    - Detect potential outliers by visualizing spread.

9. Segmented EDA (Conditional Analysis):

- **Task:** Task: Explore relationships within subgroups.

Prompt (ChatGPT):

- *"Examine the relationship between 'Order Quantity' and 'Discount Amount' separately for each 'Customer Segment'. Are there differences across segments? Present results for each segment."*

Prompt (CoPilot/Excel):

- *Use PivotTables/Charts, filtered by segment.*

  - **When to Use:** When you suspect relationships might not be uniform across your entire dataset.

III. Anomaly Detection

Identify data points that need further investigation.

10. Outlier Identification (Numerical):

- **Task:** Identify extreme values.

Prompt (ChatGPT):

- *"Identify any outliers in the 'Order Quantity' column using the IQR method. Explain the IQR method and list any outliers with row numbers. Suggest potential reasons for these outliers."*

Prompt (CoPilot/Excel):

- *Use Conditional Formatting or formulas (e.g., based on standard deviations or IQR).*

  - **When to Use:** When you suspect or need to identify unusual or erroneous data points in numerical columns for:
    - Data quality checks.
    - Identifying genuine rare events.
    - Improving the robustness of analyses.

11. Text Anomaly Detection (Textual):

  - **Task:** Find unusual text entries.

Prompt (ChatGPT):

- *"Examine the 'Product Description' column. Identify any entries that are significantly different in length, style, or content. Explain your reasoning. Flag any entries with less than 10 words or non-standard formatting."*

Prompt (CoPilot/Excel):

- *Not directly applicable within base Excel and requires more advanced text analysis.*

  - **When to Use:** When dealing with textual data and, you need to identify:
    - Data entry errors in text fields.
    - Spam or irrelevant content.
    - Unusual customer feedback.

IV. Hypothesis Generation and Question Refinement

Turn insights into testable ideas.

12. Open-Ended Exploration (ChatGPT):

- **Task:** Generate initial hypotheses and research questions.

Prompt (ChatGPT):

- *"Based on the dataset, what are the three most interesting or surprising insights? Explain your reasoning. Generate three specific, testable hypotheses about customer behavior."*

- **When to Use:** When you are in the early stages of EDA and want to:
  - Generate initial hypotheses about potential relationships and patterns.
  - Brainstorm potential research questions.
  - Uncover unexpected or surprising findings.

13. Drill-Down (ChatGPT):

- **Task:** Investigate specific insights in more detail.

Prompt (ChatGPT):

- *"Based on your previous response, provide more details on [Insight 1]. Include supporting data points."*

- **When to Use:** After you have identified a broad insight or pattern and, you need to:
  - Investigate a specific insight in more detail.
  - Gather supporting evidence and data points.
  - Refine understanding.

V. Preparing for Machine Learning (Optional)

14. Feature Importance

- **Task:** Identify potentially important features

Prompt (ChatGPT):

- *"If I were to build a machine learning model to predict '[Target Variable],' which features in this dataset seem most likely to be important predictors, based on your EDA? Explain your reasoning."*

- **When to Use:** If you are performing EDA in preparation for a Machine Learning task

VI. Key Tips for AI-Powered EDA:

- **Clean Data First:** Always clean your data before performing EDA.
- **Iterative Process:** EDA is exploratory; refine prompts based on AI responses.
- **Domain Knowledge:** Use your expertise to interpret findings.
- **Visualize:** Use charts and graphs to make patterns clear.
- **Correlation ≠ Causation:** AI identifies relationships, you determine causality.
- **Document Everything:** Keep a record of cleaning steps, insights, and hypotheses.
- **Large Datasets:** For millions of rows, use sampling, summarization, or specialized tools.

Disclaimer:

All content and material on the upGrad website is copyrighted material, either belonging to upGrad or its bonafide contributors and is purely for the dissemination of education. You are permitted to access, print, and download extracts from this site purely for your own education only and on the following basis:

- You can download this document from the website for self-use only.
- Any copies of this document, in part or full, saved to disc or to any other storage medium may only be used for subsequent self-viewing purposes or to print an individual extract or copy for non-commercial personal use only.
- Any further dissemination, distribution, reproduction, copying of the content of the document herein or the uploading thereof on other websites or use of the content for any other commercial/unauthorized purposes in any way which could infringe the intellectual property rights of upGrad or its contributors, is strictly prohibited.
- No graphics, images, or photographs from any accompanying text in this document will be used separately for unauthorized purposes.
- No material in this document will be modified, adapted, or altered in any way.
- No part of this document or upGrad content may be reproduced or stored on any other website or included in any public or private electronic retrieval system or service without upGrad's prior written permission.
- Any rights not expressly granted in these terms are reserved.