Exploratory Data Analysis (EDA) Guide

Exploratory Data Analysis (EDA) is a crucial part of data science. It helps you gain insights into your data, recognize relationships, detect anomalies, and identify important trends. However, it can be challenging to know what questions to ask when exploring your data. Here are 15 questions to guide your investigation and provide the information you need to make informed, data-driven decisions.

Questions to Guide Your EDA

1. **What are the basic characteristics of my dataset?**
    - Determine the type of data, the number of observations, and the number of variables. This initial understanding will help guide further exploration.
2. **What is the overall structure of my dataset?**
    - Examine variable types (categorical vs continuous) and their distributions. This overview helps understand how all components fit together and guides further exploration into specific areas of interest.
3. **What patterns exist in the data?**
    - Identify trends or relationships between variables. Understanding these patterns can reveal interactions and insights into what drives them.
4. **Are there any outliers present?**
    - Outliers may indicate anomalies or errors in the dataset or something significant. Analyzing outliers helps determine their causes and how to address them.
5. **What are the missing values in the dataset?**
    - Identify missing values, which may indicate problems with data collection or entry. Understanding why values are missing allows for appropriate corrective actions.

6. **What is the correctness of the data?**
    - Assess data quality, including its source authenticity and the presence of duplicate values. The quality of analysis depends on the quality of the data.
7. **Is there any correlation between variables?**
    - Examine correlations to reveal hidden relationships between variables, leading to new insights and further exploration.
8. **How does this data compare to past performance?**
    - Compare current performance metrics with previous periods to identify behavioral changes and forecast future outcomes based on past trends.
9. **Is there any seasonality present?**
    - Look for recurring patterns over time, such as seasonal fluctuations in sales, to understand external factors affecting the data.
10. **How much variability exists within each variable?**
    - Measure variability to understand the spread of data points within a variable, helping decide the appropriate analysis techniques (e.g., clustering algorithms vs linear regression).
11. **Are there any discrepancies between observed and expected values?**
    - Investigate mismatches between observations and expectations to determine if adjustments are needed, such as cleaning bad records or changing model parameters.
12. **What are some potential explanations for unexpected results?**
    - Unexpected results may reveal interesting insights or highlight issues needing attention, such as improving accuracy by targeting specific features.
13. **How do different subsets of my dataset behave differently?**
    - Break down datasets into smaller samples based on criteria (e.g., geographic location) to gain additional insights and tailor strategies for maximum effectiveness.

14. **Do I need to transform any variables before analysis?**
    - Identify necessary transformations (e.g., scaling) before detailed analysis to ensure accurate results.
15. **Are there any gaps in our understanding that need filling before deeper analysis?**
    - Ensure you have all relevant information before conducting detailed analyses to avoid unreliable results due to a lack of contextual understanding.

Additional Considerations

- How many features and observations do you have?
- What is the data type of each feature?
- Do the data types make sense? Do any need changing?
- Are there null values?
- How much memory does the dataset use? Could this pose a problem later on?
- What is the distribution of each variable?
- Are there any outliers?
- Are the max/min values reasonable? Are there any errors?
- What is the mean for each variable? What do the means reveal about the dataset?

This guide provides a structured approach to EDA, helping you uncover valuable insights and make informed decisions based on your data.