How to Conduct an Effective Exploratory Data Analysis (EDA)

The Ultimate Guide

Hello there, back with a hot topic in the data science world — Exploratory Data Analysis (EDA). Think of EDA as your preliminary detective work; it’s the part that can make or break your subsequent data modeling. I will try to explain EDA through my understanding in this article. So, let’s get started!

Why EDA?

Imagine you’re an archaeologist. Would you dig randomly? — No, you’d survey the land first, right? Similarly, EDA helps you know where to dig for insights. It’s about preventing flawed models and wrong conclusions by understanding your data’s existing structure and variables. (features)

## A Structured Approach: The APP Framework

Let’s employ my beloved APP framework for this:

1. **Attention:** Understand your dataset.
2. **Purpose:** Establish your goals.
3. **Process:** Do the actual EDA.
4. **Pay-off:** Extract and implement insights.

## Step 1: Attention — Know Your Data

### Data Types

Run df.dtypes in Python to understand the variable types you have.

Knowing whether you're dealing with numerical or categorical data, for example, will inform your subsequent steps.

### Data Size and Shape

To see the size and shape of your data, run df.shape.

If you're dealing with Big Data, you might have to employ specialized tools for your EDA.

### Missing Values

Execute df.isnull().sum() to find the missing values.

This can indicate whether you can rely on a column for analysis or if you need to take steps for imputation.

## Step 2: Purpose — Goals & Objectives

Are you trying to understand user behavior? Perhaps predict future sales? By understanding your purpose, you can target your EDA effectively.

## Step 3: Process — The Heart of EDA

Here’s where the action happens. We’ll dive into four essential components:

### Data Summarization

Utilize df.describe() to summarize your data. This gives you the mean, median, standard deviation, etc.

**For categorical data**, use df['column\_name'].value\_counts() to see the frequency of each category.

### Data Cleaning

**1. Handling Missing Values:** If a column has too many missing values, consider dropping it. If only a few are missing, think about imputation methods. There are several imputation methods you can consider:

* **Mean/Median/Mode Imputation:** Replace missing values with the mean (for normally distributed data), median (for skewed data), or mode (for categorical data).