EXPLORATORY DATA ANALYSIS USING PYTHON - LAPTOP-DATASET

By Arunachalam N
DA &DS
MAY 2025

INTRODUCTION

This project analyses a laptop dataset containing detailed specifications such as company, type, screen size, RAM, storage, GPU, OS, and price. The primary objective is to clean and analyse this dataset to extract meaningful insights and understand key drivers of laptop pricing.

KEY ATTRIBUTES VARIABLES

Company: The brand or manufacturer of the laptop (e.g., Dell, HP, Apple).
TypeName: The category or type of laptop (e.g., Ultrabook, Gaming, Notebook).
Inches: The size of the laptop screen, measured diagonally in inches.
Ram : The amount of RAM (system memory), typically in gigabytes (e.g., 4GB, 8GB, 16GB).
Memory: The storage capacity and type (e.g., 1TB HDD, 512GB SSD, or hybrid options).
OpSys: The operating system installed on the laptop (e.g., Windows 10, macOS, Linux).
Weight_kg: The physical weight of the laptop, usually given in kilograms.
Price: The selling price of the laptop, which is the target variable for analysis.

DATA CLEANING

Data cleaning is a crucial step to ensure accuracy and reliability before analysis. For a laptop dataset, this process targets correcting inconsistencies, handling missing or wrong values, and reformatting data into usable forms. Here's a breakdown of essential steps and practical techniques applied when cleaning such datasets.

Removing Duplicates	Handling Missing Values	Standardizing Formats
Identify and eliminate duplicate laptop entries to avoid double-counting and skewed results.	Drop rows with extensive missing or nonsensical data. For fields with occasional missing values: Fill in with statistically appropriate values (e.g., mean, median, or mode).	• Ensure consistency in text fields: • Convert all brand or type entries to uniform casing • Convert numerical columns stored as text (like "8GB" RAM or "2.5 kg" weight) into proper numeric fields for easier analysis.

EDA

Exploratory Data Analysis (EDA) is the initial step in the data analysis process. It involves examining and visualizing datasets to uncover key characteristics, relationships, and trends. EDA helps identify patterns, detect anomalies, check assumptions, and set the foundation for further modeling or in-depth study.

Types of Analysis in EDA

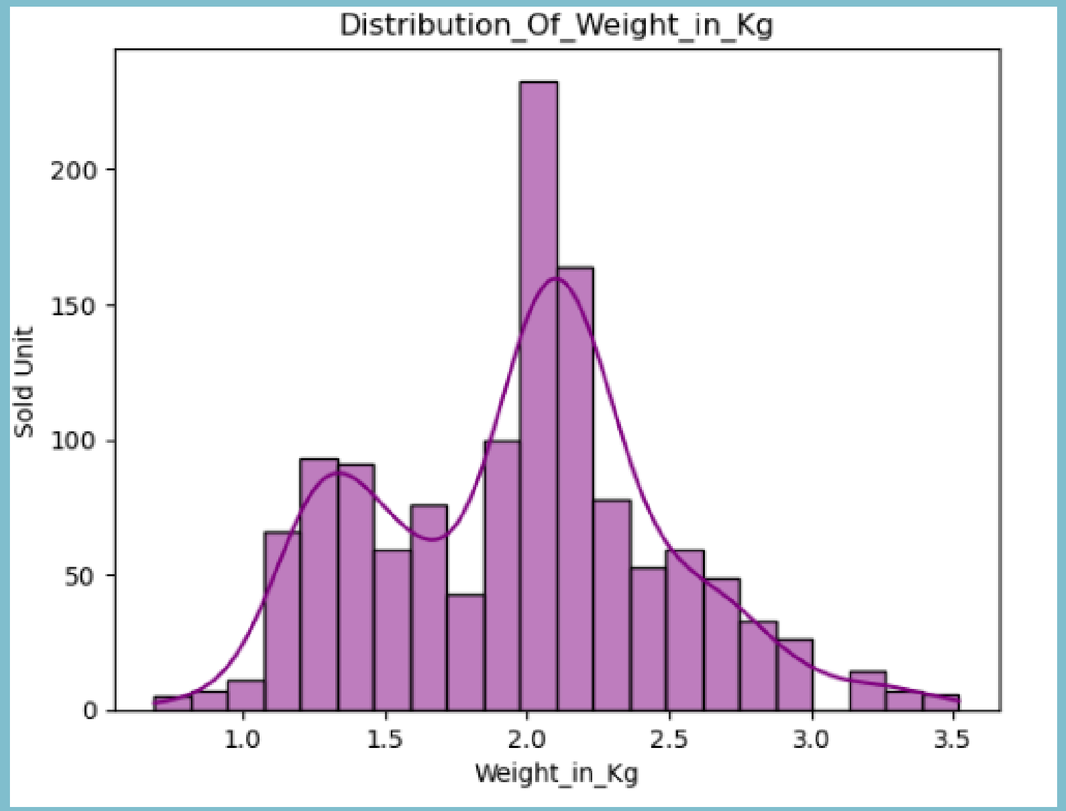
1. Univariate Analysis

2. Bivariate Analysis

3. Multivariate Analysis

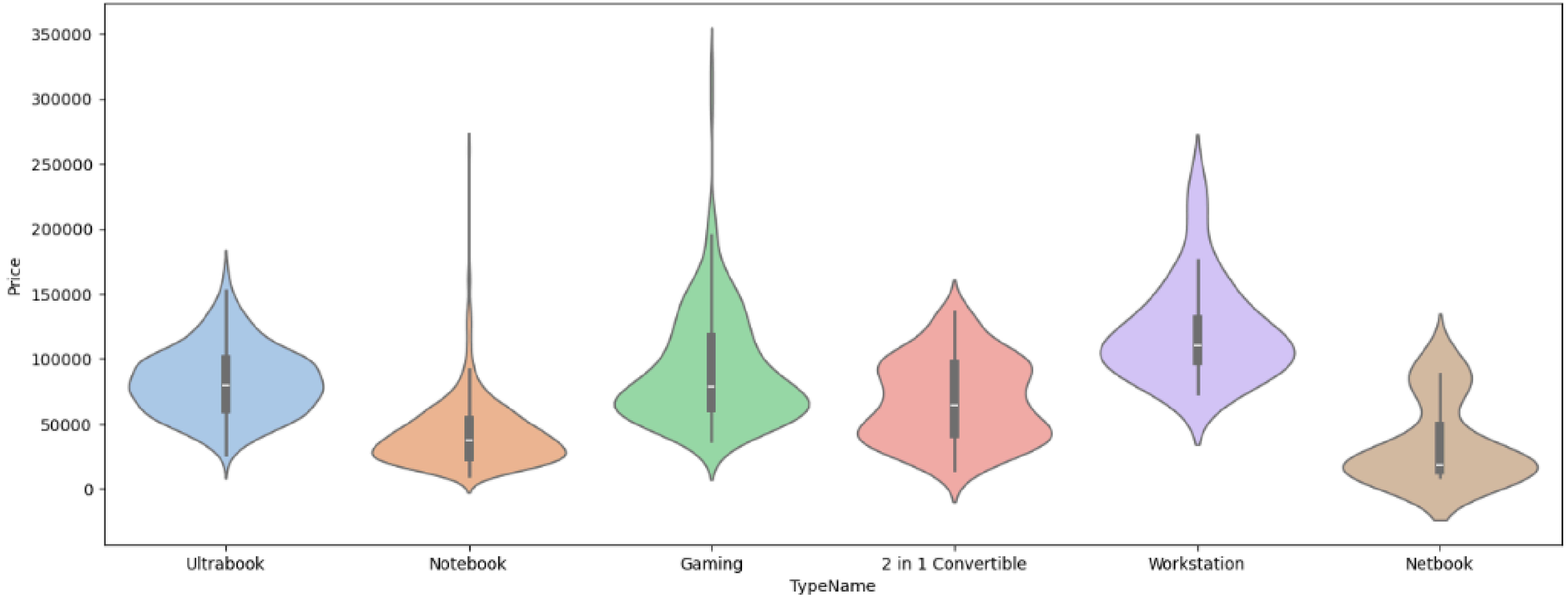
UNIVARIATE ANALYSIS

Univariate Analysis is the simplest form of data analysis focused on examining one variable at a time. It describes the distribution, central tendency (mean, median, mode), and variability (range, variance, standard deviation) of that single variable without exploring relationships with others.



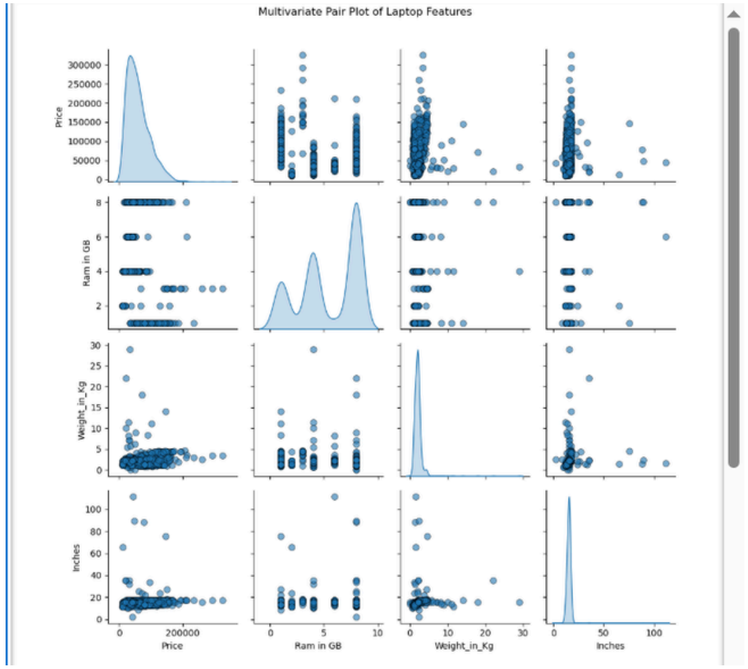
BIVARIATE ANALYSIS

Bivariate analysis is a statistical technique used to examine the relationship between two variables, often denoted as X and Y.

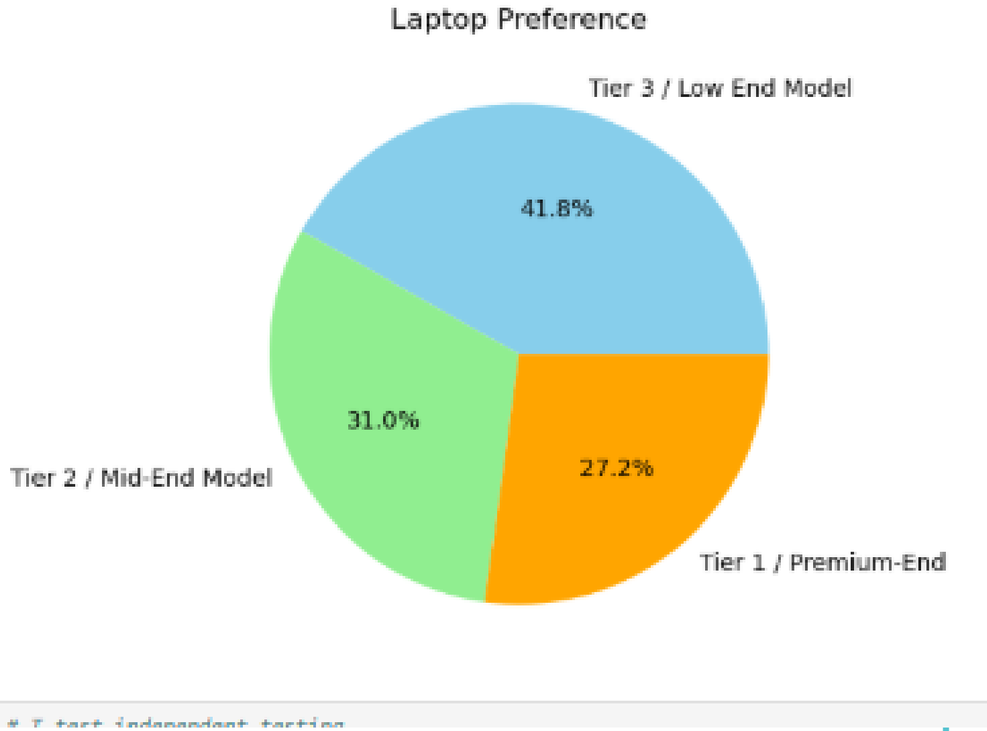


MULTIVARIATE ANALYSIS

Multivariate analysis (MVA) is a set of statistical methods used to analyze data sets with multiple variables simultaneously



FEATURES ANALYSIS



HYPOTHESIS

Descriptive Statistics
Computed mean, median, mode, and standard deviation
Summarized features like price, Ram, inches.
Hypothesis Testing
Two-sample t-test: Compared mean prices across different view categories
One-way ANOVA: Tested price differences among multiple view groups
Chi-square Test: Checked dependency between view and condition

```
# T test independent testing
from scipy.stats import ttest_ind
dell_prices = lap_data[lap_data['Company'] == 'Dell']['Price']
hp_prices = lap_data[lap_data['Company'] == 'HP']['Price']

stat, p_value = ttest_ind(dell_prices, hp_prices)
print("T-statistic:", stat)
print("P-value:", p_value)
if p_value < 0.05:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")
```

```
T-statistic: 2.0366358749657705
P-value: 0.04216408728364471
Reject the null hypothesis
```

```
# ANNOVA One way Annova
from scipy.stats import f_oneway

ram_sizes = lap_data['Ram in GB'].unique()
grouped_prices = [lap_data[lap_data['Ram in GB'] == r]['Price'] for r in ram_sizes]

# Perform the ANOVA test
stat, p = f_oneway(*grouped_prices)
print("F-statistic:", stat)
print("P-value:", p)

if p < 0.05:
    print("Reject the null hypothesis: At least one RAM group has a different mean price.")
else:
    print("Fail to reject the null hypothesis: No significant mean price difference between RAM groups.")
```

```
F-statistic: 287.67428985106627
P-value: 9.181674559738192e-206
Reject the null hypothesis: At least one RAM group has a different mean price.
```

CONCLUSION

- Laptops with 8GB & 16GB RAM are the most sold among all configurations.
- Laptops with 2 GB RAM have the lowest sales
- Top-Selling Company:
- The company with the highest number of laptop sales is e.g., HP / Dell / Lenovo].

THANK YOU