

Report Title (COMP3125 Individual Project)

Alexander Vaz

Department of Computer Science, Wentworth Institute of Technology

Abstract This project investigates the factors influencing the prices of laptops using various machine learning and statistical techniques. By exploring multiple datasets, the analysis aims to uncover patterns and relationships between hardware specifications and pricing. Techniques such as regression, clustering, and data visualization were applied using Python libraries including Scikit-learn and Pandas. The insights can be useful for both consumers and manufacturers in understanding laptop market trends.

Keywords—laptop pricing, regression, clustering, data science, hardware analysis

I. INTRODUCTION Laptops are an essential part of modern life, with a wide range of models offering varying performance levels, features, and price points. However, it is not always obvious what justifies the price difference between laptops with similar functions. This project seeks to uncover the underlying factors that influence the price of laptops and how these factors interact. By analyzing the relationships between key specifications such as RAM, CPU, GPU, screen resolution, and storage types, we aim to understand what truly drives price variation.

Using regression, clustering, and visualization methods, this project not only examines which specifications are most predictive of price but also categorizes laptops into tiers based on their features. This contributes to a clearer understanding for both manufacturers setting prices and consumers making informed choices. Existing research often focuses on consumer behavior or individual hardware components, but this study combines multiple aspects to provide a holistic analysis.

II. DATASETS

A. Source of dataset The datasets used in this project were sourced from Kaggle and included detailed listings of laptops and their specifications, including price in euros. Some minor data cleaning was required, and additional columns were created for derived features such as screen resolution in pixels and GPU brand.

B. Character of the datasets Three datasets (`df1`, `df2`, `df3`) were utilized:

- **df1:** Included company, type name, screen size, CPU and GPU brands, RAM, storage, OS, weight, and price.
- **df2:** Focused on visual components like touchscreen, IPS display, screen resolution, CPU, GPU, and SSD/HDD size.

- **df3:** Provided enriched CPU and GPU brand data, along with RAM, screen features, and combined storage values.

All datasets were combined and cleaned for missing values and standardized unit formats. New columns were created for CPU brand, GPU brand, and total pixel count to help with feature engineering.

III. METHODOLOGY

A. Regression Analysis Linear regression and multiple regression were applied to understand the impact of individual variables such as RAM and SSD size on price. The Scikit-learn `LinearRegression()` model was used, with data preprocessed using standard scaling.

B. Clustering KMeans clustering was used to group laptops into price tiers. The optimal number of clusters was determined using the elbow method. Data was normalized before fitting into the model using `KMeans(n_clusters=3)`.

C. Data Visualization Seaborn and Matplotlib were used to generate boxplots, pairplots, and heatmaps. These visualizations highlighted the relationships between categorical and numerical variables and the price.

IV. RESULTS

A. Regression Results

- RAM and SSD size showed the strongest linear relationship with laptop price.
- Apple laptops consistently had higher intercept values compared to other brands.

B. Clustering Results

- Three major clusters were identified: budget, mid-range, and premium.
- The high-end cluster was dominated by Apple, MSI, and Razer laptops, while budget models were from Lenovo, Acer, and Asus.

C. Visual Analysis

- Boxplots showed a clear increase in price with RAM and SSD upgrades.
 - Heatmaps indicated strong correlations between price and storage/RAM but weaker associations with weight or screen size.
-

V. DISCUSSION While the regression models provided decent predictive accuracy, nonlinear relationships and feature interactions likely limit the full interpretability. GPU brand, while significant visually, was harder to capture in a purely linear model. Cluster analysis supported natural groupings, but boundary laptops (e.g., low-spec Apple models) were sometimes misclassified. Future work should explore classification models to predict price ranges and incorporate real-time pricing data from e-commerce platforms.

VI. CONCLUSION This project provided a comprehensive analysis of factors influencing laptop prices. RAM, storage type, CPU brand, and screen features were all shown to significantly affect pricing. The results support the development of smarter pricing strategies and purchasing decisions. In the future, including user ratings and real-world performance benchmarks could deepen the insights.

ACKNOWLEDGMENT Thanks to the COMP3125 teaching team for guidance and support throughout this project.

REFERENCES

<https://www.kaggle.com/datasets/muhammادkhairulramadhan/laptop-price-dataset>

<https://www.kaggle.com/datasets/abdocan/laptop-prices>

<https://www.kaggle.com/datasets/abdelrahmanemad594/laptop-prices>

<https://scikit-learn.org> [3] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.