Laptop Dataset — Data Cleaning & Feature Engineering Documentation

Project Scope

This document is about the work I did on the Laptop Dataset. I used MySQL to clean up the data and make it look nice and neat. The Laptop Dataset had a lot of information about laptop hardware. I wanted to make it easy to understand and use. So I changed the update into a simple and organised format. Now the Laptop Dataset is ready for people to look at and analyse. They can use it to make reports and graphs with tools, like EDA and BI dashboards. The Laptop Dataset is also good to use for modelling.

Initial Raw Columns

Company, TypeName, Inches, ScreenResolution, Cpu, Ram, Memory, Gpu, OpSys, Weight, Price, MyUnknownColumn

Column Audit & Profiling

Performed null checks, blank checks, value distributions, regex validation, and outlier scans across all columns. Result: data mostly consistent with targeted normalization required for hardware spec fields.

Dropped Column

I got rid of the MyUnknownColumn because it did not have any information in it most of the time and it was not useful, for analysis. The MyUnknownColumn was mostly empty. Did not help us understand anything.

Price Normalization

The price is now set to a fixed number of places, which is two. So the final type of this price is DECIMAL(10,2). This means the price will always have two places, like dollars and cents.

RAM Transformation

We took the values like 8GB and 16GB and converted them to a numeric column called ram_gb which is an integer.

Weight Transformation

We take values such as 2.3kg. We convert them into a numeric weight in kilograms. This is done by removing the unit text, which's kg" in this case and we are left with just the number. The result is a weight in kilograms that is stored as a DOUBLE, which's a type of numeric value. We do this for values, like 2.3kg so that we can work with the weight in kilograms easily.

Screen Resolution Decomposition

I found out the width and the height of the screen. I also got the resolution details, in a form. The type of panel used in the screen is noted down. It is also clear if the screen is a touchscreen or not. The resolution class of the screen is identified. The details of the resolution are very important. Include the width and the height and the type of resolution class. The screens resolution class and panel type and touchscreen flag are all part of the details that were found out..

CPU Feature Engineering

Extracted cpu_brand, cpu_family, cpu_model, cpu_generation (when derivable), and clock_speed_ghz..

GPU Feature Engineering

Extracted gpu_brand, gpu_series, and gpu_model with normalized naming.

Storage (Memory) Decomposition

We need to split the storage into two parts: storage and secondary storage. For each part we have to include the size in gigabytes and the type of storage it is, such, as Solid State Drive, Hard Disk Drive, Hybrid or Flash. We also have to make sure that all the storage sizes that are given in terabytes are changed to gigabytes. This way the primary storage and secondary storage will each have their size in gigabytes and their type, like Solid State Drive or Hard Disk Drive.

Operating System Normalization

Split into os_family and os_version with normalized families (Windows, macOS, Linux, Chrome OS, Android, No OS).

Screen Size & Category

We are going to keep the size of the screen in inches, as a number so it is called screen_size_inch.

The type of laptop is also kept this is called laptop_type. It is something that we categorize.

.

Validation Methods

Used regex checks, numeric sanity ranges, vendor consistency scans, and format validation before conversion.

Final Engineered Schema

Company, laptop_type, screen_size_inch, display_resolution_width, display_resolution_height, display_resolution, panel_type, is_touchscreen, resolution_class, cpu_brand, cpu_family, cpu_model, cpu_generation, clock_speed_ghz, ram_gb, primary_storage_gb, primary_storage_type, secondary_storage_gb, secondary_storage_type, gpu_brand, gpu_series, gpu_model, os_family, os_version, weight_kg, price.