2024-17-04

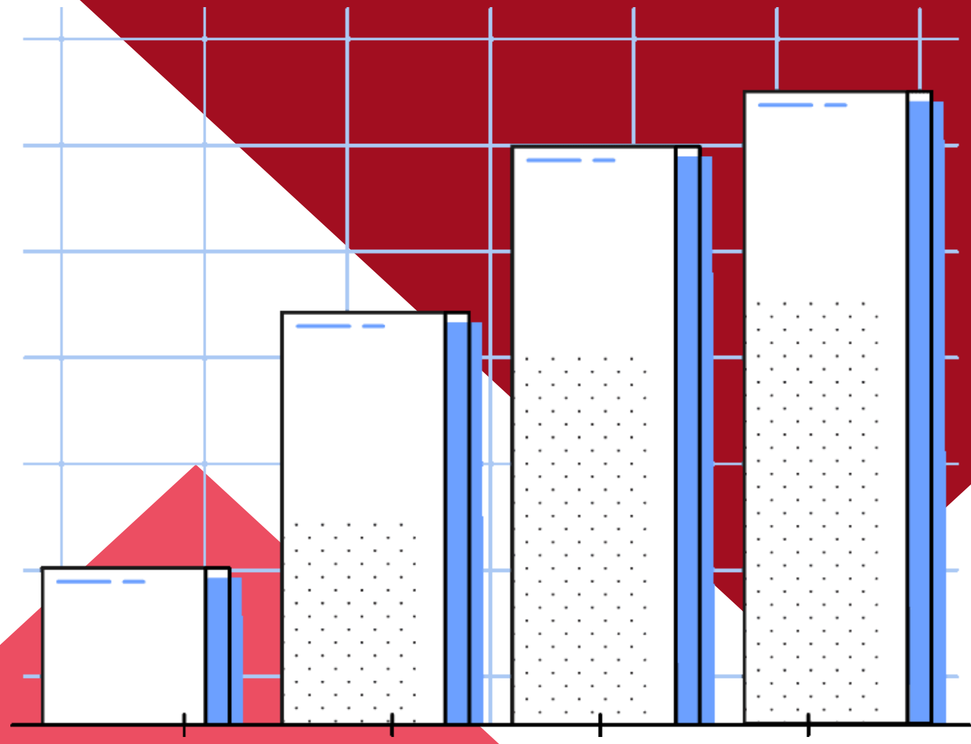
Amazon Web Scraping And Visualization

Northeastern University

PRESENTED BY

Jay Shukla

NUID: 002619454



About The Project

- Use Scrapy to mine data from Amazon
- Clean the Raw data using Pandas and analyze it
- Discover meaningful pattern and trends and report on it

Project Objectives

Amount of product on Amazon is very huge. So my focus: Laptops

Objective # 1

Scrape data of laptops for different brands

Objective # 2

Clean the scraped data for analysis

Objective # 3

Discover pattern and trends

USE CASES

- Pricing Analysis
- Product development decisions
- Sales Forecasting
- Customer Insights
- Inventory Management
- Brand Monitoring

TOOLS AND TECHNOLOGIES USED

Scrapy

RegEx

Workflow

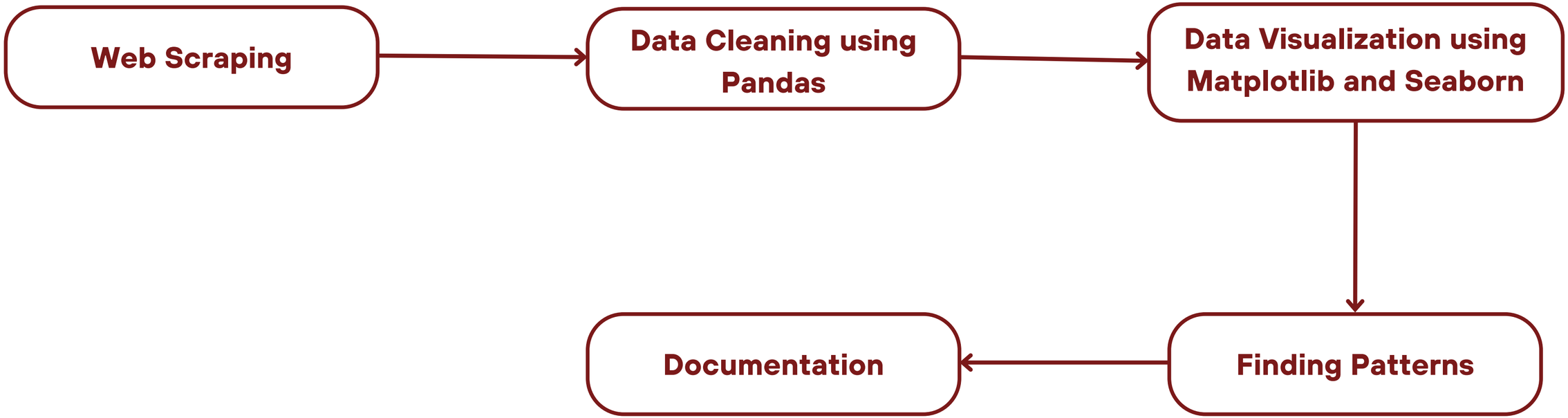


Figure 1: Flowchart showing workflow of project

Web Scraping

Roll over image to zoom in

ASUS 17 Chromebook 17.3 Inch FHD Laptop 2023 Newest, Intel Celeron N4500 Up to 2.8Ghz, 4GBRAM, 192GB Storage, USB C, Wifi6, Bluetooth, 17hours Battery Life, Chrome OS

Visit the ASUS Store

★★★☆☆ 2 ratings

Price: $299.90

Capacity:

4GB RAM | 192GB Storage(64GB eMMC+128GB ...

Brand	ASUS
Screen Size	17.3 Inches
Hard Disk Size	64 GB
CPU Model	Celeron N
Operating System	Chrome OS
Card Description	Integrated
Graphics Coprocessor	Intel UHD Graphics 630
CPU Speed	2.8 GHz
Hard Disk Description	SSD
Resolution	1080p, 4K

See less

Amazon Laptop Data

Sample data

Title	Brand	Screen_Size	CPU_Model	Ram	Operating_System	Rating	Series	Color	Resolution	Weight	USB	Battery	Graphics	Disk_size	Card_desc	Price
Microsoft	Microsoft	13.5 Inches	1.2GHz Cort	8 GB	Windows 10	4.4	Intel	Platinum	nan	nan	nan	nan	Intel HD G	256 GB	Integrated	nan
Microsoft	Microsoft	13.5 Inches	Core i5	8 GB	Windows 10	4.4	Microsoft	Matte Blac	nan	nan	nan	nan	nan	nan	nan	nan
Microsoft	Microsoft	13.5 Inches	Core i7	16 GB	Windows 10 S	3.9	Surface La	Platinum	nan	nan	nan	nan	Intel Iris Pl	512 GB	Integrated	nan
Microsoft	Microsoft	13 Inches	Core i5-103	8 GB	Windows 10 Home	4.2	Surface La	Cobalt Blu	nan	nan	nan	nan	Intel Iris Pl	nan	Integrated	nan
2022 Micr	Microsoft	12.4 Inches	Core i5	4 GB	Windows 11 Pro	3.9	Surface La	Silver	nan	nan	nan	nan	nan	256 GB	Integrated	$359.99
HP 14in Hi	HP	14 Inches	Ryzen 3 230	4 GB	Windows 10	4.3	Notebook	Silver	nan	nan	nan	nan	AMD Rade	128 GB	Integrated	nan
HP Envy 34	HP	34 Inches	Intel Core i7	32 GB	Windows 11 Home	4.4	HP ENVY 3	nan	nan	nan	nan	nan	nan	nan	nan	$2,167.00
Dell Latitu	Dell	15.6 Inches	Intel Core i5	16 GB	Windows 10 Pro	4.3	Latitude 3	Black	nan	nan	nan	nan	Intel UHD	256 GB	Integrated	$818.00
MSI Bravo	MSI	15.6 Inches	Ryzen 5	16 GB	Windows 11	4.5	Bravo 15 E	Aluminum	nan	nan	nan	nan	nan	512 GB	RX 5500M	$699.00
MSI Latest	MSI	17.3 Inches	Intel Core i9	64 GB	Windows 11 Pro	5	GE76-12U	Titanium B	nan	nan	nan	nan	nan	2 TB	Dedicated	$2,449.99

Amazon Laptop Data

SN	Column Names	Number of data
1	Title	2843
2	Brand	2810
3	Screen_Size	2756
4	CPU_Model	2631
5	Ram	2711
6	Operating_System	2631
7	Rating	2631
8	Series	2501
9	Colour	2501
10	Resolution	182
11	Weight	66
12	USB	35
13	Battery	7
14	Graphics	2063
15	Disk_size	2248
16	Card_desc	2248
17	Price	1863

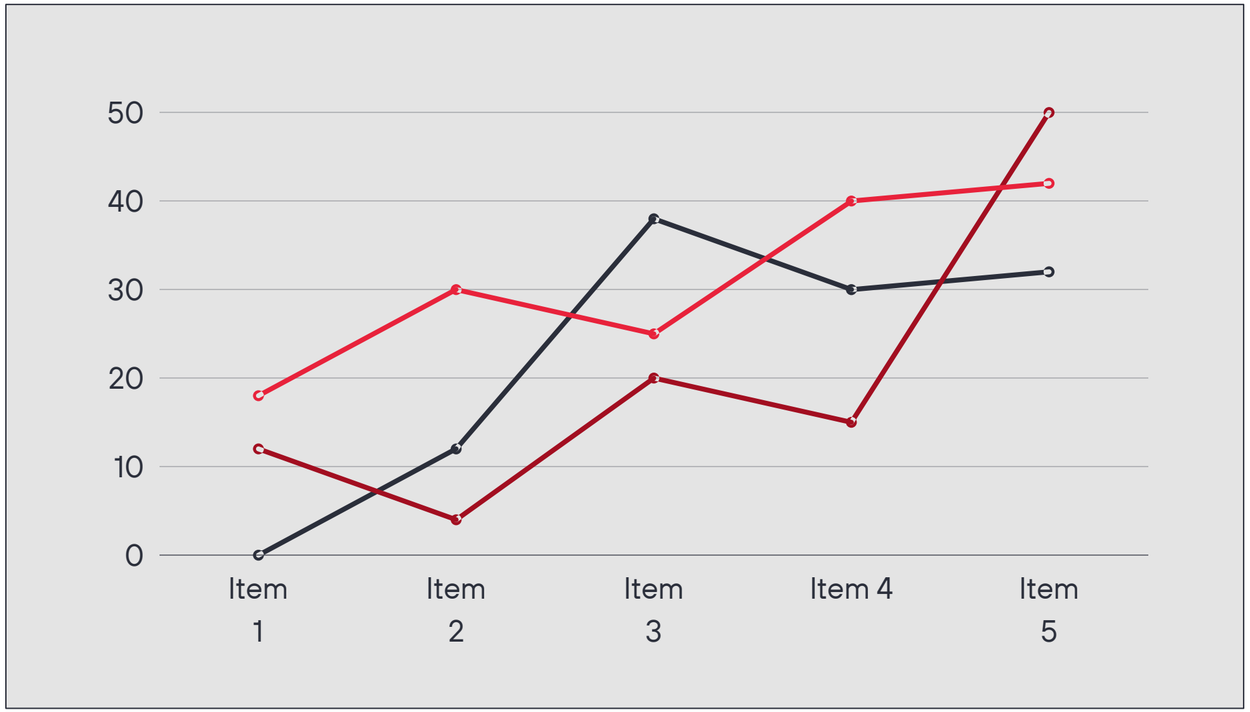
Data Cleaning

- Fill nan values in Rating column with 0
- Drop rows with nan values in other columns
- Converted data in RAM and Disk_Size column to numeric values
- Removed duplicate rows
- Removed irrelevant Columns/Features

Data Visualization

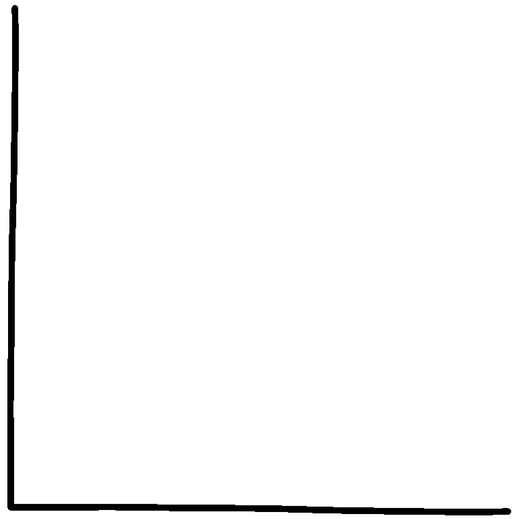
Matplotlib and Seaborn

Open-source data visualization in Python



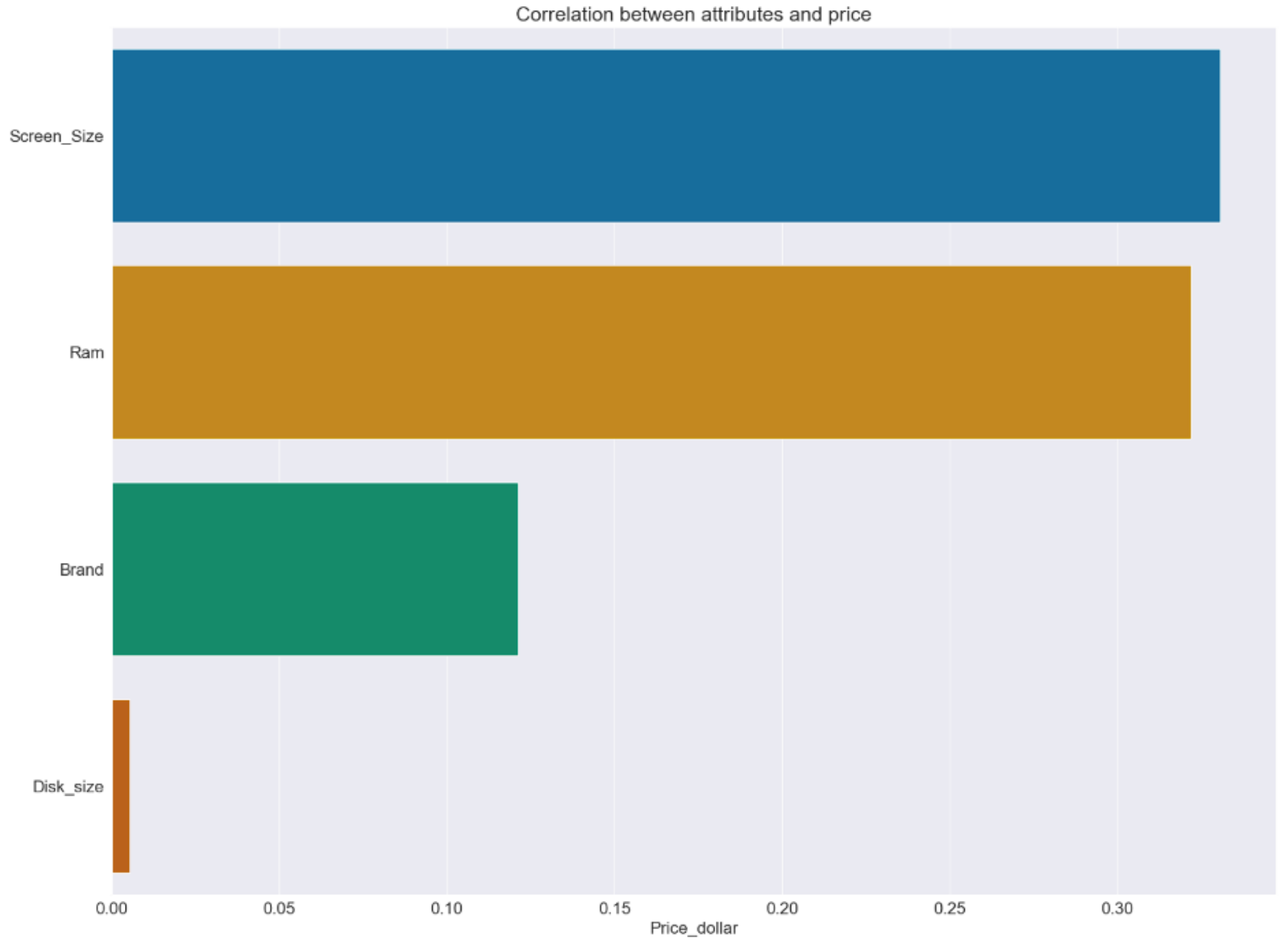
- Generates various types of plots and graphs
- Integrates with other libraries such as NumPy, Pandas.
- Offers customization and high-quality output.

DATA VISUALIZATIONS



Correlation

The best attribute appears to be screen size to determine price.



High Correlation

Screen Size, RAM

Irrelevant

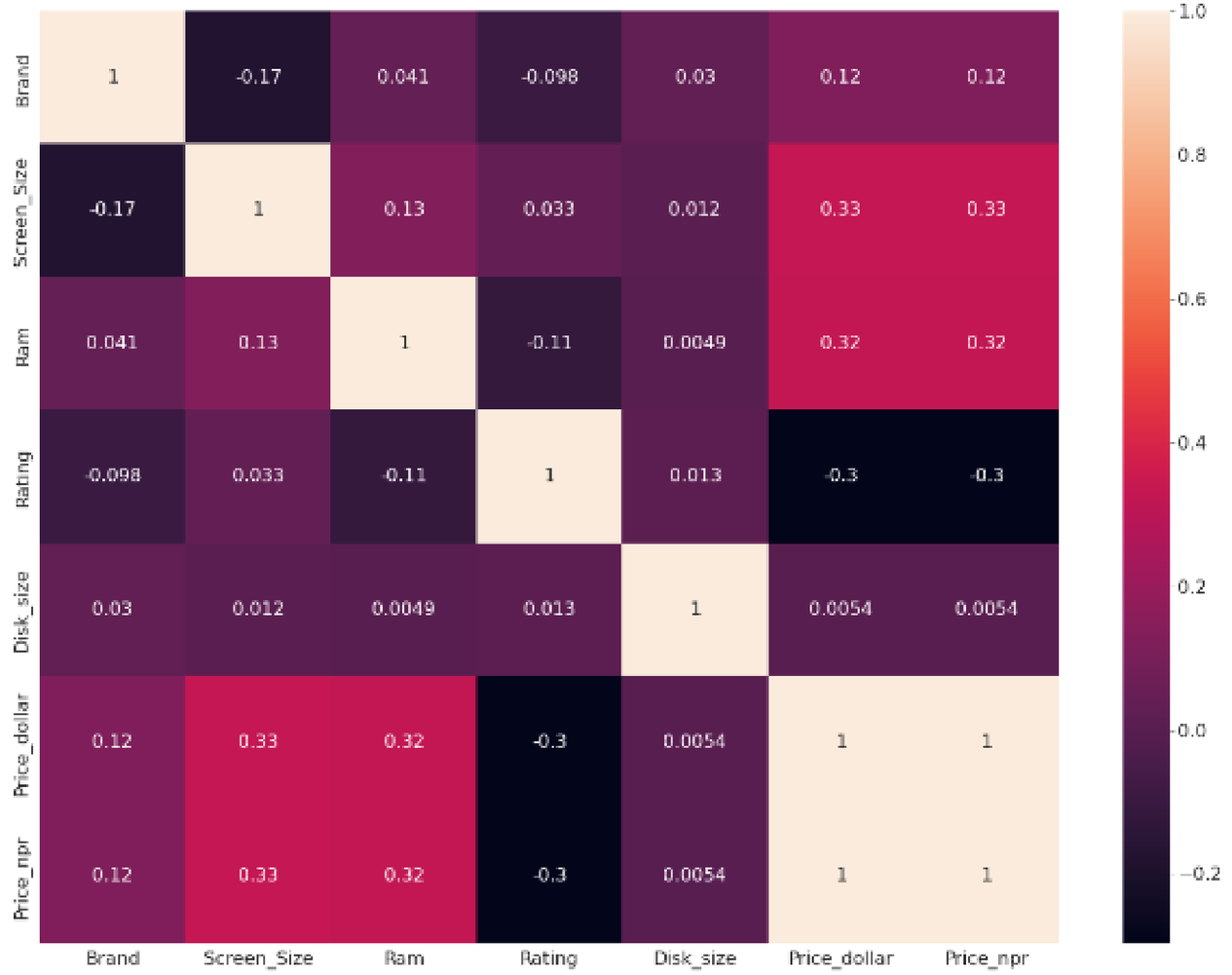
Disk Size

Medium Correlation

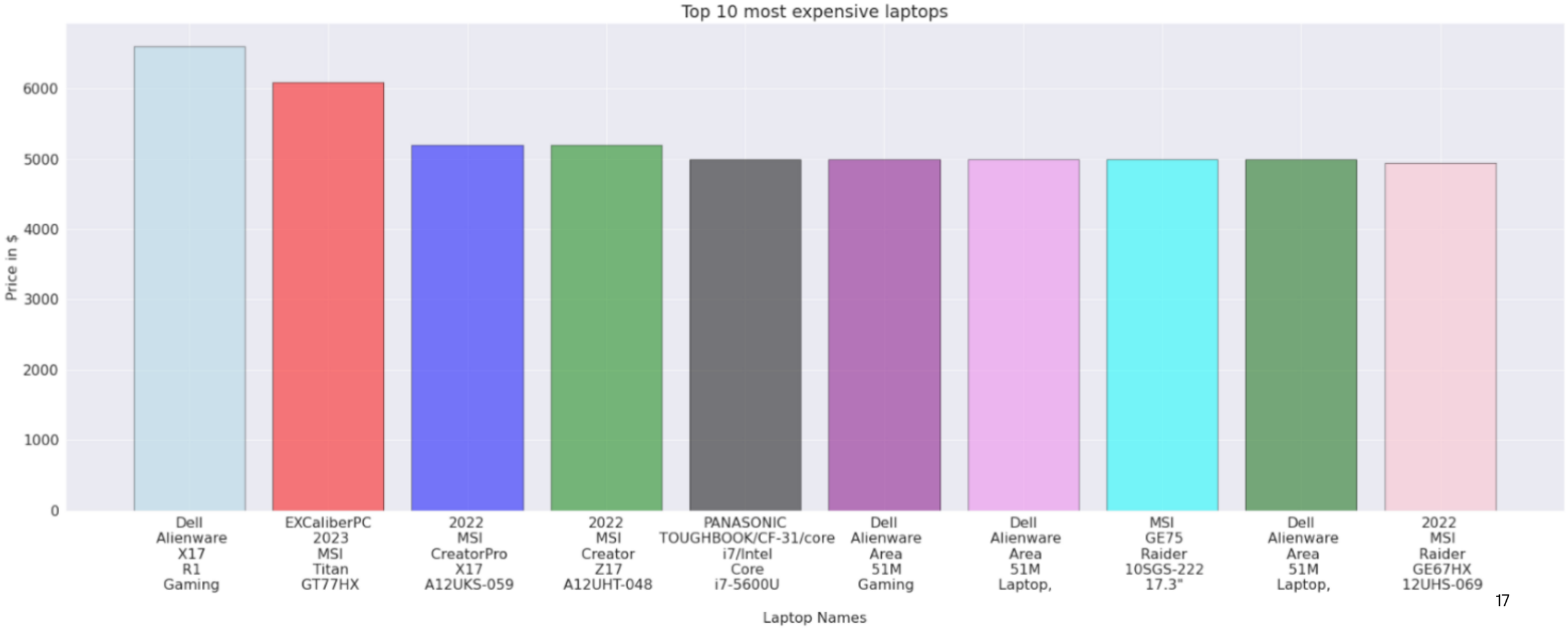
Brand

Correlation Heatmap

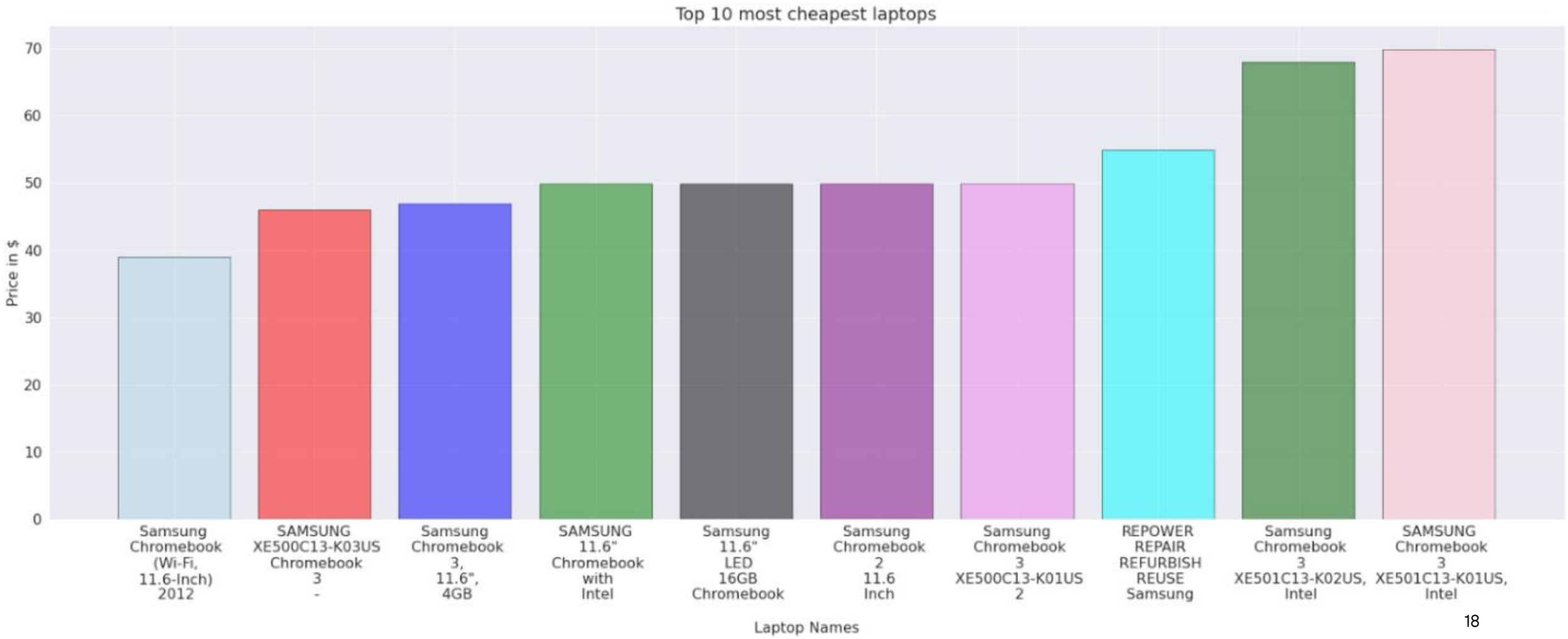
The best attribute appears to be screen size to determine price.



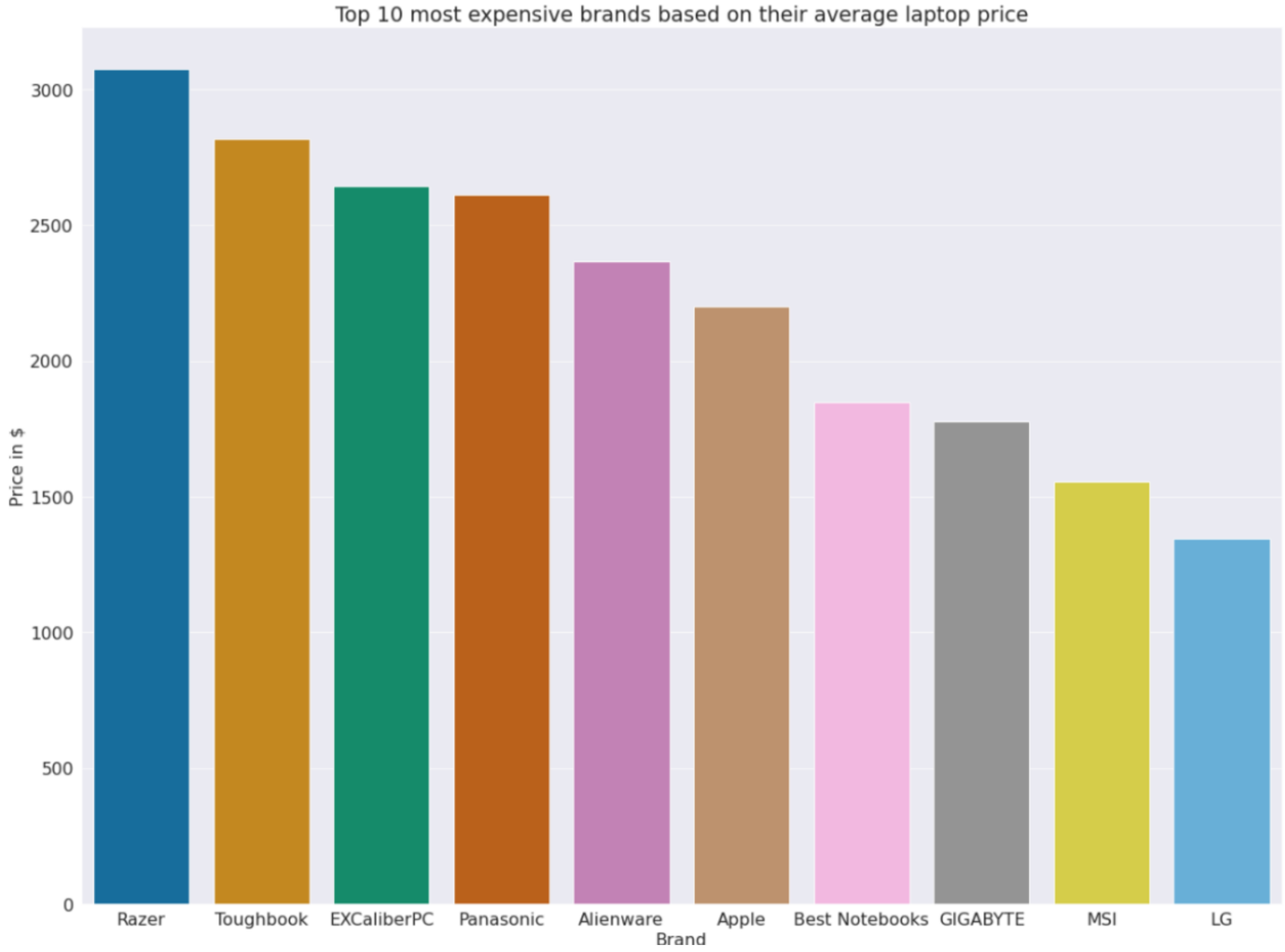
Top 10 Most Expensive Laptops



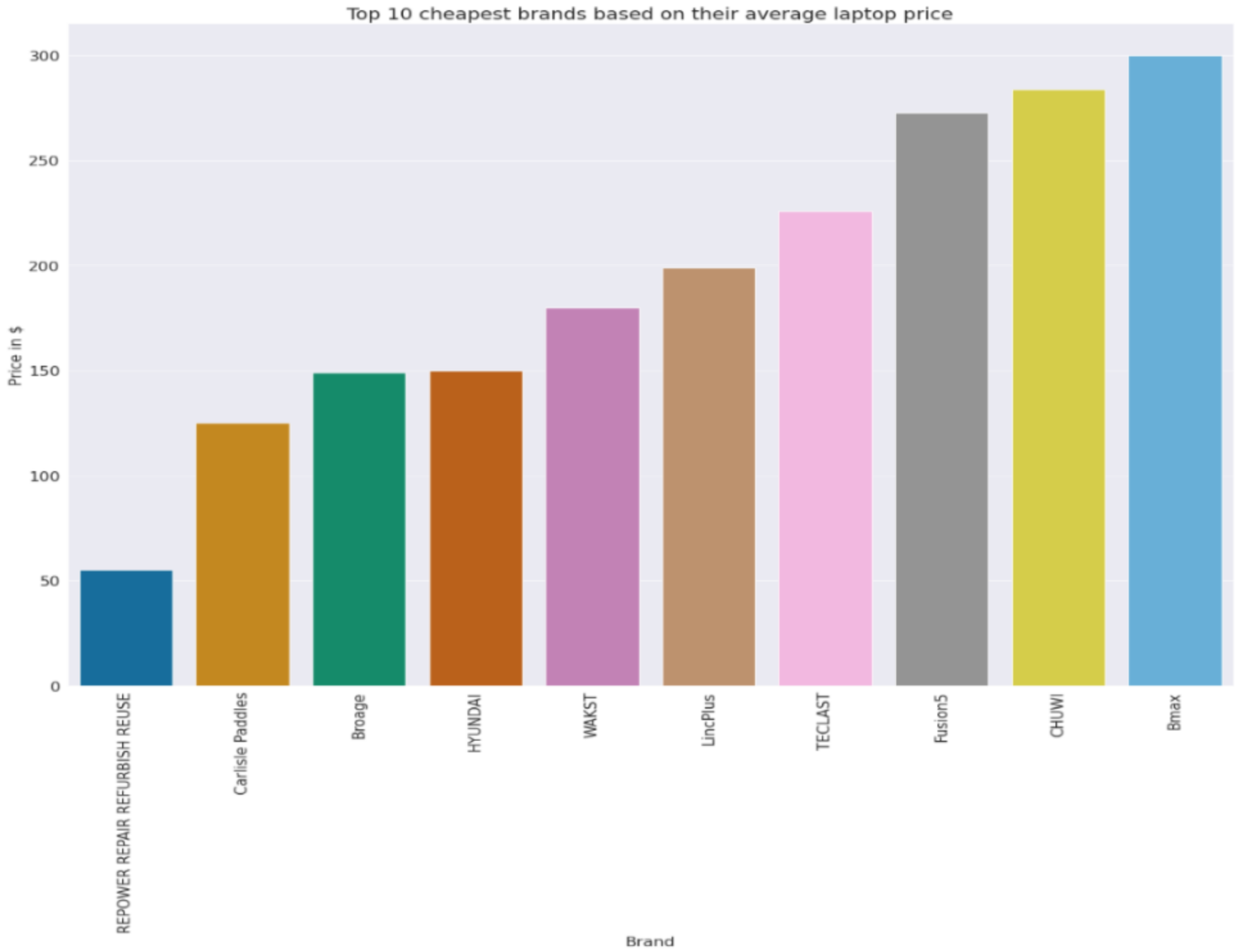
Top 10 Cheapest Laptops



Top 10 Expensive Laptops Brands



Top 10 Cheapest Laptops Brands



Average Price Of Different Brands

RAZER HAVE HIGHEST AVERAGE PRICE

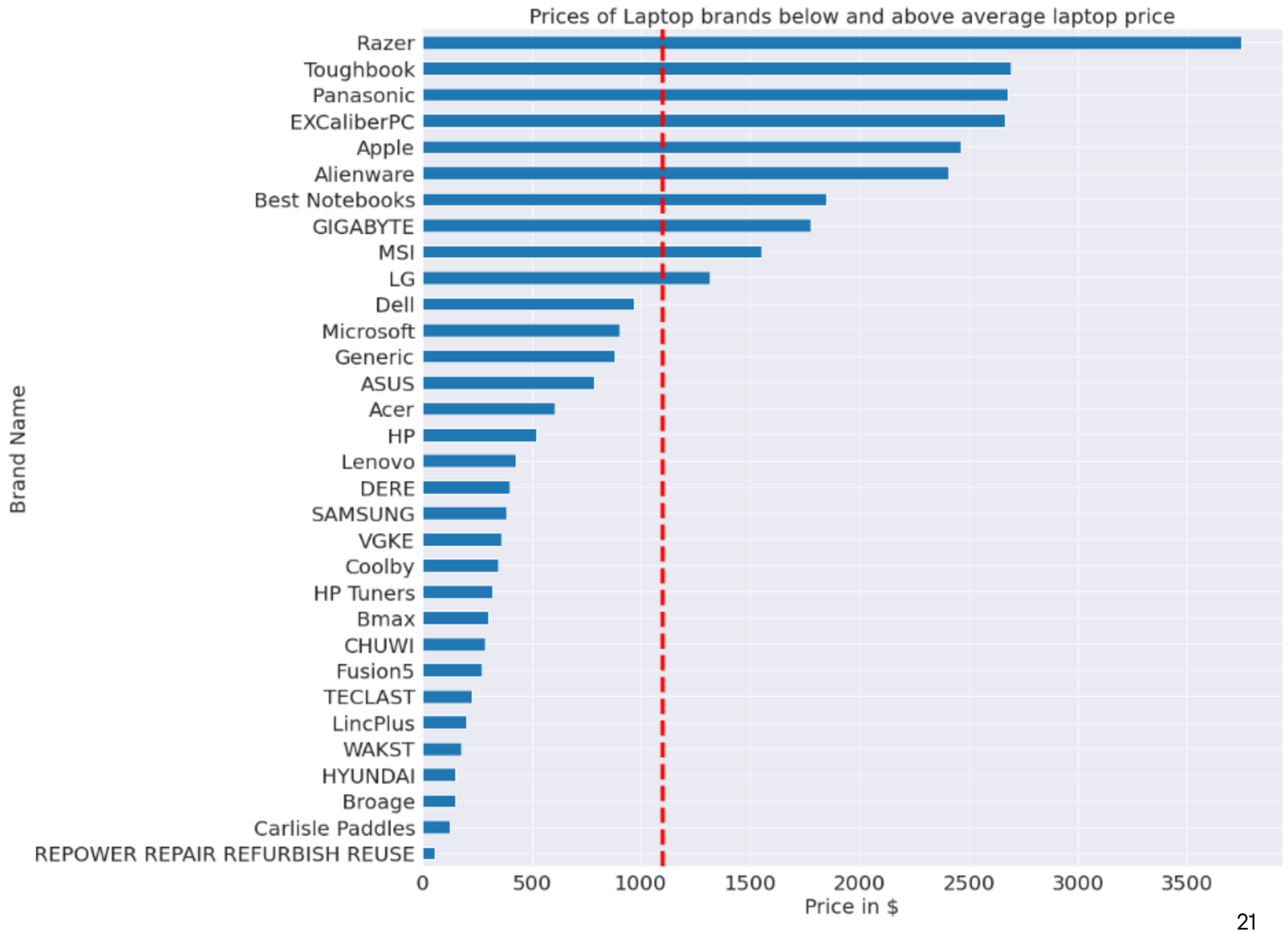
Most Razer laptops have above average price

USED HAVE LEAST AVERAGE PRICE

Similar story to previous case

AVERAGE PRICE NEAR 1000 DOLLARS

Brands like acer have varied pricing so have low average



Distribution of Price Among Laptop Brands

Scatter plot of price against brands.

HIGH END

EXCALIBER PC
ALIENWARE
RAZER

ACROSS THE BOARD

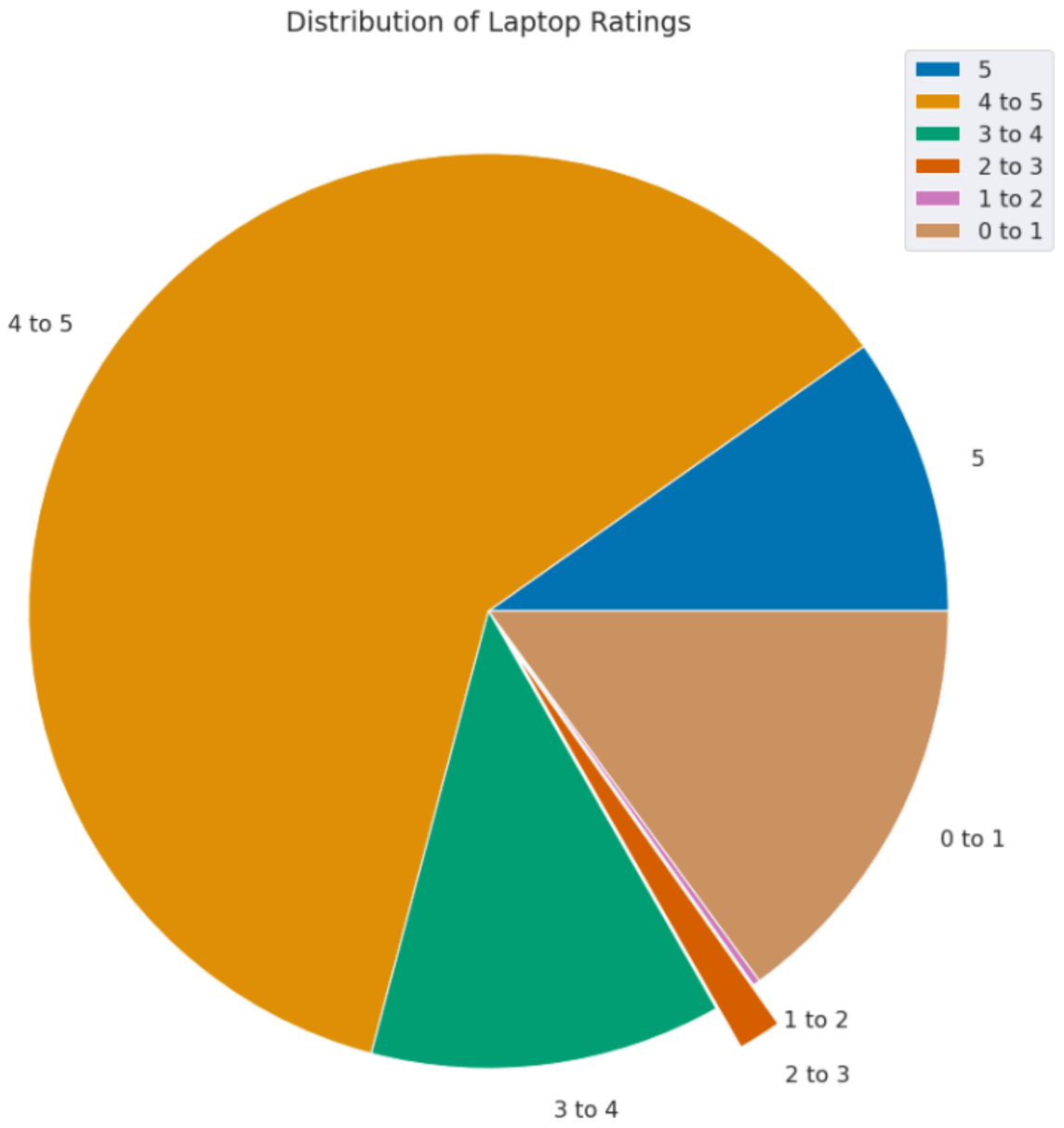
DELL
ASUS
MSI

LOW END

LENOVO
SAMSUNG

About the Ratings

- Most are rated between 4 and 5 stars
- Significant amount rated 0-1 stars



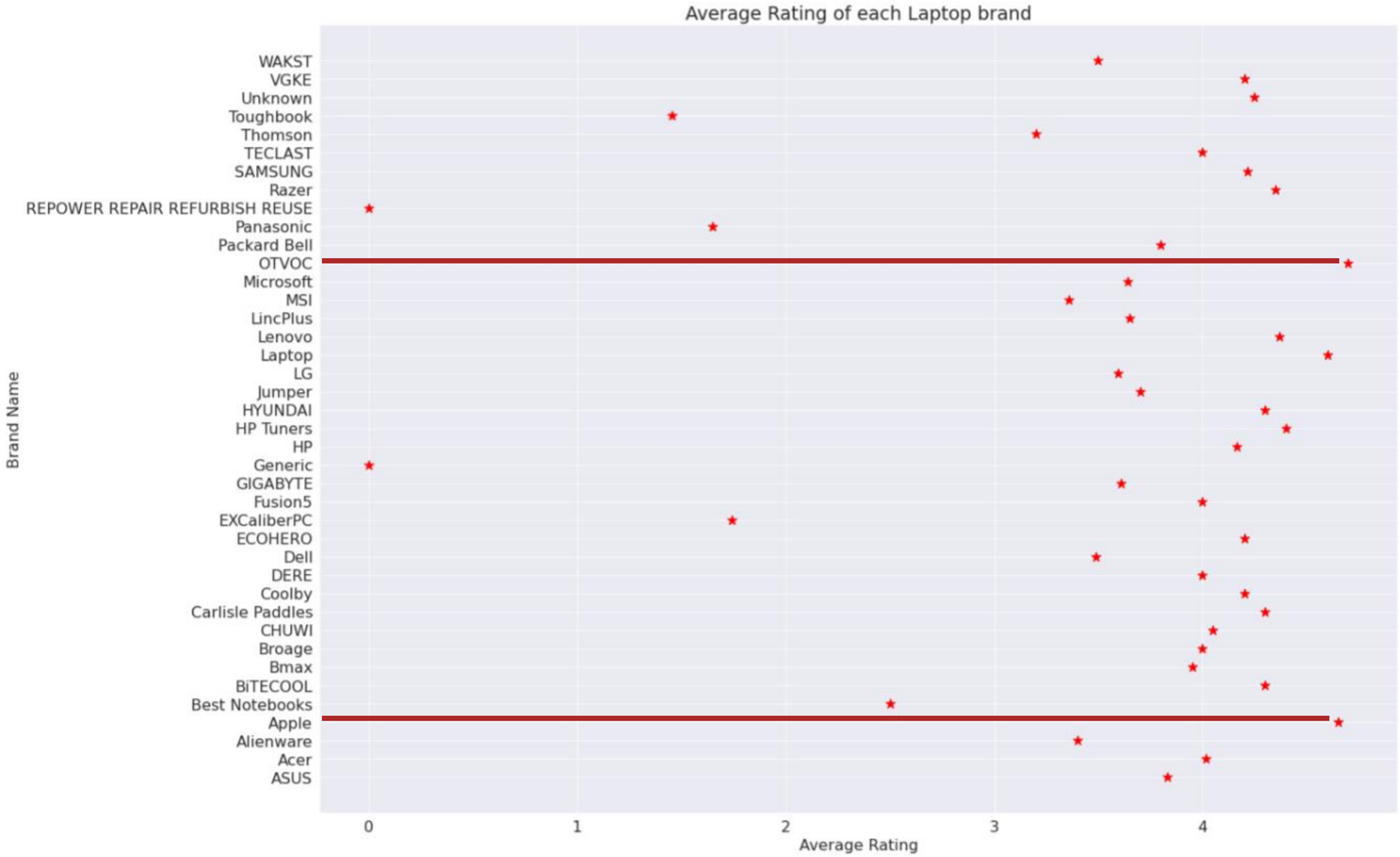
Average Rating Of Each Laptop Brand

OTVOC & APPLE HAD HIGHEST AVERAGE RATING

Over 4.0 rating

REFURBISH & GENERIC HAS LOWEST AVERAGE RATING

Not Rated

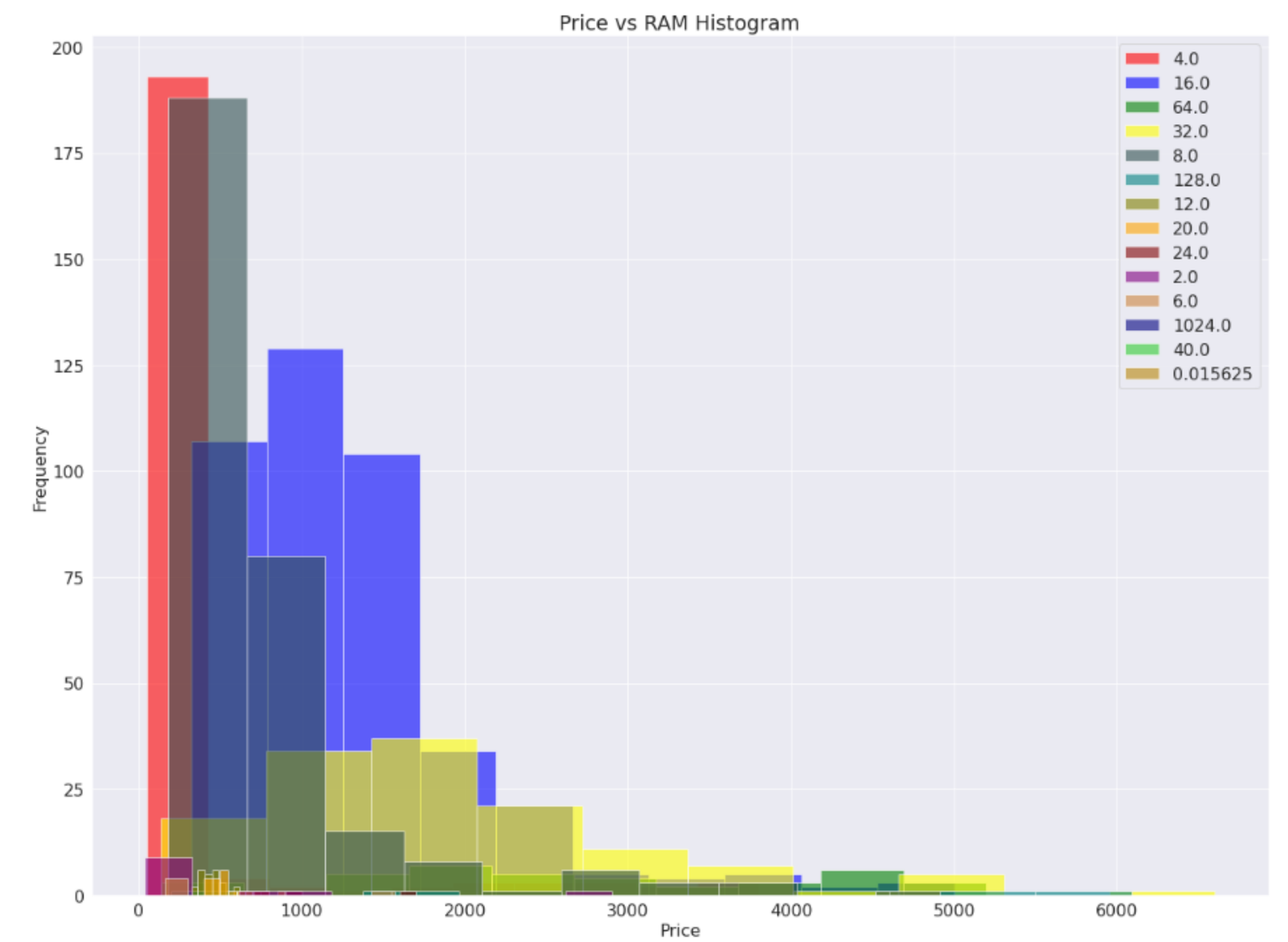


Price V/s RAM Histogram

Shows that the laptops price mostly depends on RAM size.

4 GB RAM is most common and cheapest, followed by 8 GB and 16 GB

1TB RAM laptops are most expensive



MAIN INSIGHTS

- Price changes with change in Screen size, RAM and Brand.
- Dell's Alienware and MSI's laptops take over top 10 most expensive laptops.
- Samsung Chromebook takes over top 10 most cheapest laptops.
- Razer is most expensive laptop brand followed by Toughbook and ExCalliber.
- Most laptops have 4GB RAM and is cheaper comparatively.
- Laptops with 32GB and 128GB RAM is expensive and less in number.

Challenges

- Amazon has mechanisms to prevent scraping
- Joint ventures (Sub Brands)
- Irregularities in amazon pages
- Sellers other than the authorized distributor

Conclusion

- Scraped Amazon laptop data
- Cleaned raw data
- Visualized data using Matplotlib and Seaborn
- Discovered insights and patterns

References

- Scrapy Beginners Series Part 1 - First Scrapy Spider | ScrapeOps. (n.d.). https://scrapeops.io/python-scrapy-playbook/scrapy-beginners-guide/

- Amazon.com: Laptop. (n.d.). https://www.amazon.com/s?k=laptop

THANK YOU