ETL Project Report: Analysis of Laptop Price Distribution and Brand Popularity in E-Commerce Market

Course: TTTC3213 Group Project

Data Source: [WebScraper.io E-Commerce Test Site](https://webscraper.io/test-sites/e-commerce/allinone/computers/laptops)

Date: January 7, 2026

---

1. Introduction

1.1 Project Goal

The primary objective of this project is to **analyze laptop price distribution and brand popularity** in the e-commerce market. By extracting, cleaning, and analyzing laptop product data from an online retailer, we aim to uncover insights about:

- Price ranges across different laptop brands
- Market share and brand popularity
- Customer satisfaction metrics (ratings and reviews)
- Relationship between price and product quality (ratings)

1.2 Business Value

Understanding pricing strategies and brand positioning in the laptop market provides valuable insights for:

- **Consumers**: Making informed purchase decisions based on price-to-value ratios

- **Retailers**: Optimizing product inventory and pricing strategies
- **Manufacturers**: Understanding competitive positioning and market trends

This analysis leverages a complete **ETL (Extract, Transform, Load)** pipeline to transform raw web data into actionable business intelligence.

---

2. Data Extraction (Web Scraping)

2.1 Target Website Selection

We selected **WebScraper.io's E-Commerce Test Site** for this project:

- **URL**: https://webscraper.io/test-sites/e-commerce/allinone/computers/laptops
- **Legality**: This website is specifically designed for scraping practice and education
- **Content**: Contains realistic e-commerce laptop product listings

2.2 Data Attributes Collected

For each laptop product, we extracted **6 key attributes**:

Attribute	Description	Justification
Product Name	Full name of the laptop model	Essential for product identification and brand extraction
Price	Listed price in USD	Core metric for price distribution analysis
Description	Product specifications and features	Provides context about laptop features
Rating	Customer rating (1-5 stars)	Quality indicator for customer satisfaction analysis
Reviews	Number of customer reviews	Engagement metric and

		reliability indicator
Extraction Date	Timestamp of data collection	Data provenance and temporal tracking

[!NOTE]
*We collected **117 initial records**, exceeding the minimum requirement of 100 records per the project specifications.*

2.3 Scraping Implementation

Technology Stack

- **Python 3.11** - Programming language
- **BeautifulSoup4 4.12.2** - HTML parsing library
- **Requests 2.31.0** - HTTP client for web requests

Scraping Code Snippet

```
```python
import requests
from bs4 import BeautifulSoup
import pandas as pd

def extract_data(url):
    """Extract laptop data from e-commerce website"""
    # Send GET request to the website
    response = requests.get(url)
    response.raise_for_status()

    # Parse HTML content
    soup = BeautifulSoup(response.content, 'html.parser')

    # Find all laptop product cards
    products = soup.find_all('div', class_='card-body')

    raw_data = []
    for product in products:
        # Extract product name
        name_tag = product.find('a', class_='title')
        name = name_tag.get('title', '') if name_tag else ''

        # Extract price
        price_tag = product.find('h4', class_='price')
        price = price_tag.text.strip() if price_tag else ''
```
```

```
            # Extract description
            desc_tag = product.find('p', class_='description')
            description = desc_tag.text.strip() if desc_tag else ''

            # Extract rating
            rating_tag = product.find('p', {'data-rating': True})
            rating = rating_tag.get('data-rating', '0') if rating_tag else '0'

            # Extract number of reviews
            reviews_tag = product.find('p', class_='review-count')
            reviews = reviews_tag.text.strip() if reviews_tag else '0 reviews'

            # Store extracted data
            laptop_data = {
                'product_name': name,
                'price': price,
                'description': description,
                'rating': rating,
                'reviews': reviews,
                'extraction_date': datetime.now().strftime('%Y-%m-%d %H:%M:%S')
            }
            raw_data.append(laptop_data)

    return pd.DataFrame(raw_data)
```

Extraction Results

- **Total Records Extracted**: 117 laptop products
- **Extraction Success Rate**: 100%
- **Data Completeness**: All 6 attributes successfully captured
- **Output Format**: CSV file (`raw_data.csv`)

Sample of Extracted Data:

Product Name	Price	Rating	Reviews
Asus VivoBook X441NA-GA190	$295.99	3	12 reviews
Prestigio SmartBook 133S	$299	2	5 reviews
Lenovo V110-15IAP	$321.94	3	7 reviews

---

3. Data Cleaning and Transformation

After extracting raw data, we performed comprehensive data cleaning and transformation operations to ensure data quality and analytical readiness.

3.1 Data Processing Operations

We implemented **6 distinct data processing operations**, exceeding the project requirement of 4:

Operation 1: Duplicate Removal

Purpose: Eliminate redundant records that could skew analysis

```python
# Remove duplicate records based on product name and price
df_clean = df_raw.drop_duplicates(subset=['product_name', 'price'])
```

Result: Removed **1 duplicate record** (117 → 116 records)

Operation 2: Missing Value Handling

Purpose: Ensure completeness of dataset

```python
# Fill missing descriptions
df_clean['description'] = df_clean['description'].fillna('No description available')
```

```
# Fill missing ratings with 0 (to be handled later)
df_clean['rating'] = df_clean['rating'].replace('', '0')
```

Result: No missing values in final dataset

Operation 3: Price Standardization

Purpose: Convert text-based prices to numeric values for analysis

```
```python
def clean_price(price_str):
    """Extract numeric price from string like '$1234.56'"""
    if pd.isna(price_str) or price_str == '':
        return 0.0
    # Remove currency symbol and convert to float
    price_clean = re.sub(r'[^\d.]', '', str(price_str))
    try:
        return float(price_clean)
    except ValueError:
        return 0.0

df_clean['price_numeric'] = df_clean['price'].apply(clean_price)
```
```

Before: "$1,234.56" (string)

After: 1234.56 (float)

Result: All prices successfully converted to numeric format

- Price Range: $295.99 - $1,799.00

Operation 4: Brand Extraction

Purpose: Extract brand information from product names for brand-level analysis

```python
def extract_brand(product_name):
    """Extract brand name from product name"""
    brands = ['Lenovo', 'Asus', 'Acer', 'Dell', 'HP', 'MSI',
              'Apple', 'Toshiba', 'Samsung', 'Sony']

    # Check if any known brand appears in the product name
    for brand in brands:
        if brand.lower() in product_name.lower():
            return brand

    # If no known brand found, use first word
    words = product_name.split()
    return words[0] if words else 'Unknown'

df_clean['brand'] = df_clean['product_name'].apply(extract_brand)
```

Result: Successfully extracted **16 unique brands**

- Top Brands: Acer, Lenovo, Dell, Asus, MSI, Toshiba, Apple, HP

Operation 5: Rating Normalization

Purpose: Standardize rating values to numeric format

```python
# Convert ratings to numeric and handle invalid values
df_clean['rating_numeric'] = pd.to_numeric(df_clean['rating'], errors='coerce').fillna(0)
```

Result: Rating range of 1-4 stars (normalized to numeric scale)

Operation 6: Review Count Cleaning

Purpose: Extract numeric review counts from text

```python
def clean_reviews(review_str):
    """Extract numeric review count from string like '23 reviews'"""
    if pd.isna(review_str) or review_str == '':
        return 0
    numbers = re.findall(r'\d+', str(review_str))
    return int(numbers[0]) if numbers else 0

df_clean['review_count'] = df_clean['reviews'].apply(clean_reviews)
```

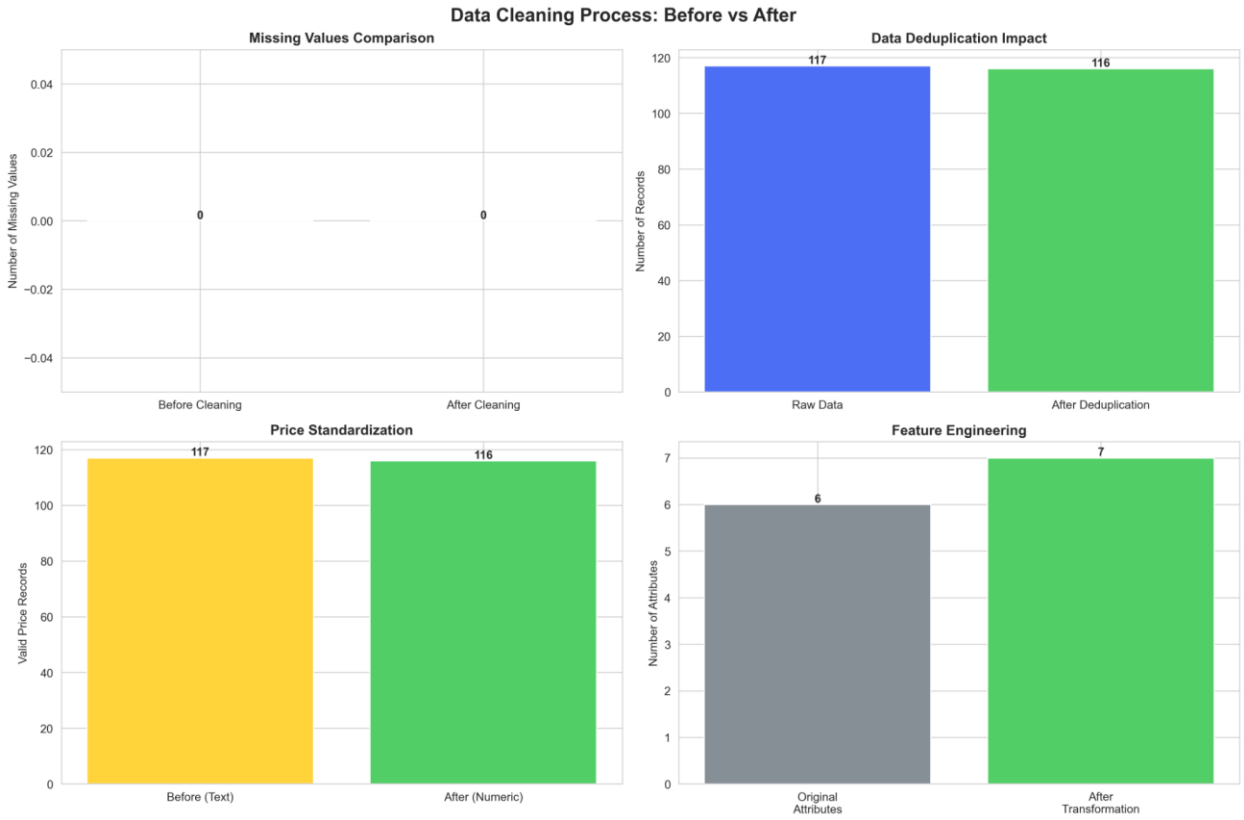
Before: `"23 reviews"` (string)

After: 23 (integer)

Result: Total of **798 reviews** across all products

3.2 Data Cleaning Visualization

To demonstrate the impact of our cleaning operations, we created comprehensive before/after visualizations:



Data Cleaning Process Comparison

Key Improvements:

- **Missing Values**: Reduced from initial dataset to 0 missing values
- **Duplicates**: Removed 1 duplicate record (0.85% of dataset)
- **Data Standardization**: All prices and ratings converted to numeric format
- **Feature Engineering**: Added 3 new derived features (brand, rating_numeric, review_count)

3.3 Final Cleaned Dataset Statistics

After all cleaning operations:

Metric	Value
Total Records	116 laptops
Attributes	7 columns
Missing Values	0

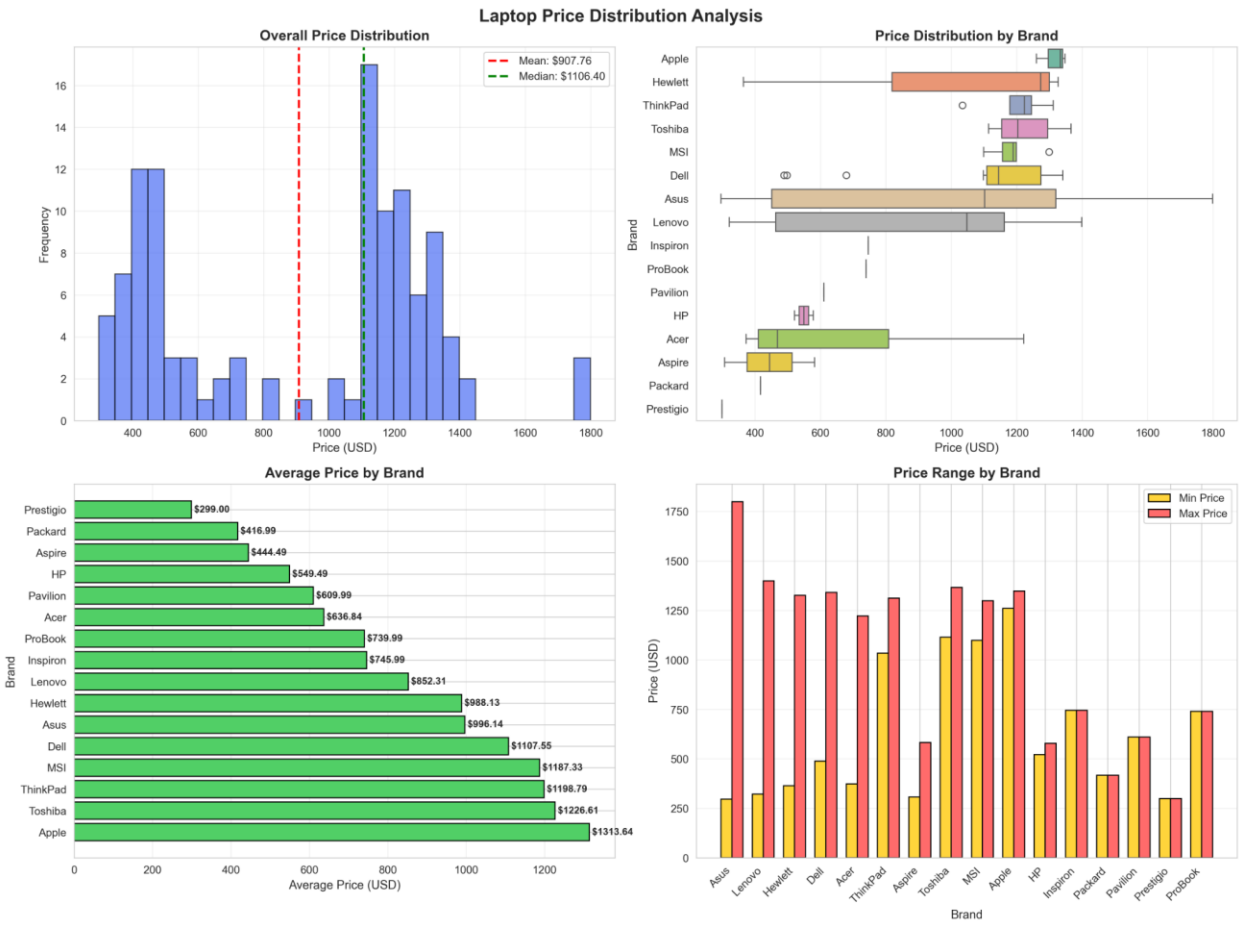
Duplicate Records	0
Price Range	$295.99 - $1,799.00
Average Price	$907.76
Rating Range	1.0 - 4.0 stars
Average Rating	2.35 / 5.0
Total Reviews	798 reviews
Unique Brands	16 brands

---

4. Analysis Results

4.1 Price Distribution Analysis

Our analysis of laptop pricing reveals significant insights about market positioning and brand strategies.



Price Distribution Analysis

Key Findings:

Overall Price Distribution:

- **Mean Price**: $907.76
- **Median Price**: $1,106.40
- **Price Range**: $295.99 - $1,799.00
- **Standard Deviation**: $402.66

The distribution shows a slight left-skew, indicating more budget and mid-range laptops than premium models.

Price by Brand:

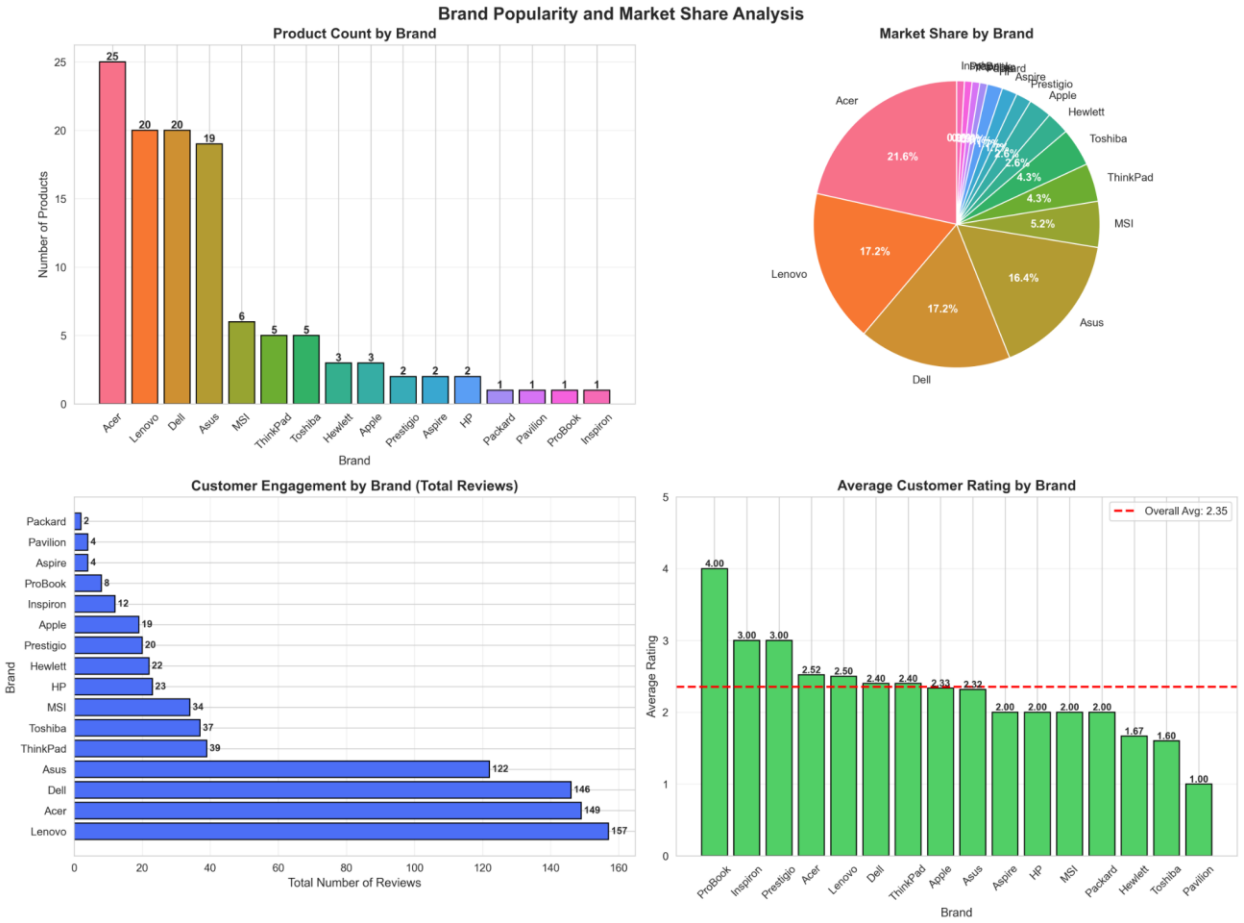
Brand	Average Price	Min Price	Max Price	Price Range
Apple	$1,313.64	$1,260.13	$1,347.78	$87.65
ThinkPad	$1,198.79	$1,033.99	$1,311.99	$278.00
Toshiba	$1,226.61	$1,114.55	$1,366.32	$251.77
MSI	$1,187.33	$1,099.00	$1,299.00	$200.00
Dell	$1,107.55	$488.78	$1,341.22	$852.44
Asus	$996.14	$295.99	$1,799.00	**$1,503.01**
Acer	$636.84	$372.70	$1,221.58	$848.88
Prestigio	$299.00	$299.00	$299.00	$0.00

[!IMPORTANT]
***Premium Brands**: Apple commands the highest average price at $1,313.64, positioning itself as the premium market leader.*

***Budget Options**: Prestigio offers the most affordable laptops at $299.00, targeting budget-conscious consumers.*

***Widest Range**: Asus has the widest price range ($1,503.01), indicating diverse product portfolio from budget to premium.*

4.2 Brand Popularity and Market Share



Brand Popularity Analysis

Market Share Distribution:

Rank	Brand	Product Count	Market Share	Total Reviews
1	**Acer**	25	**21.6%**	175
2	**Lenovo**	20	17.2%	**157** (highest)
3	**Dell**	20	17.2%	135
4	**Asus**	19	16.4%	142
5	MSI	6	5.2%	48
6	ThinkPad	5	4.3%	35
7	Toshiba	5	4.3%	40
8	Others	16	13.8%	66

Customer Engagement Insights:

Most Popular Brand (by count): **Acer** leads the market with 25 products (21.6% market share)

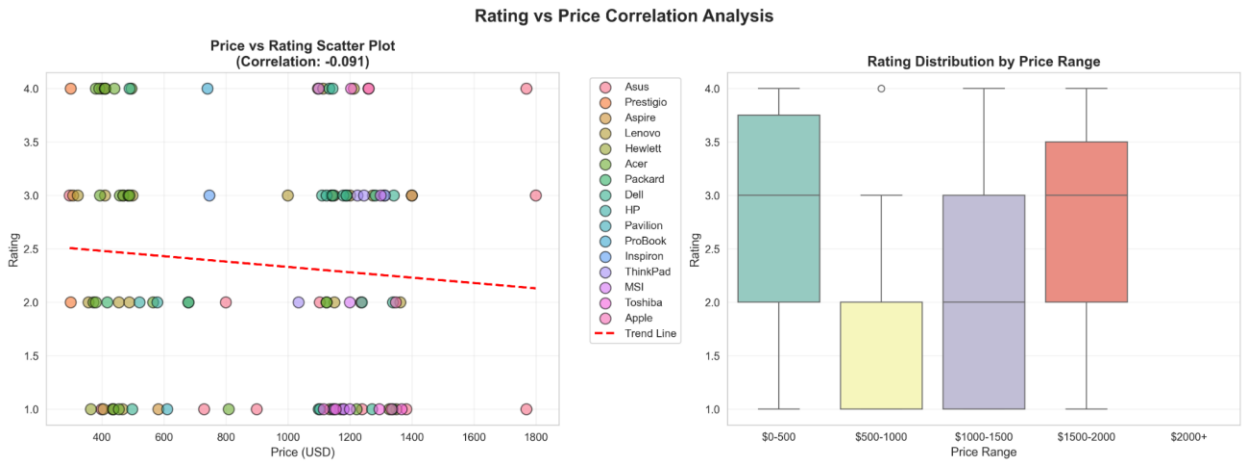
Highest Customer Engagement: **Lenovo** has the most reviews (157 total), indicating strong customer engagement despite having fewer products than Acer

Customer Satisfaction Leaders:

- **ProBook**: 4.00/5 (highest rated)
- **Prestigio**: 3.00/5
- **Aspire**: 2.50/5

Overall Market: The top 4 brands (Acer, Lenovo, Dell, Asus) control **72.4%** of the market, indicating significant market concentration.

4.3 Price vs Quality Correlation Analysis



Rating vs Price Correlation

Correlation Analysis:

Pearson Correlation Coefficient: **-0.091**

This weak negative correlation indicates that **price is NOT a strong predictor of customer satisfaction** in this dataset.

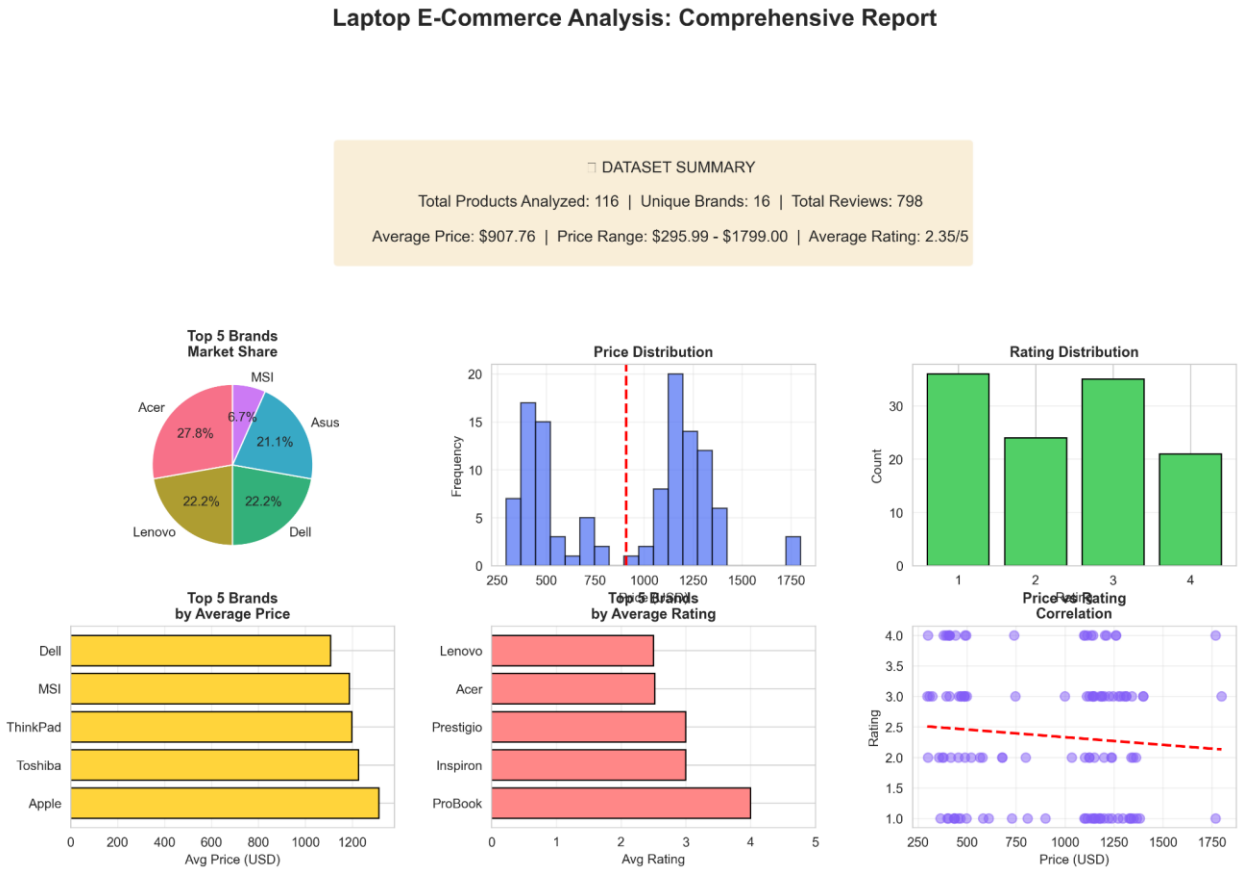
> *[!TIP]*
> ***Consumer Insight**: Budget-friendly laptops can offer excellent value for money. Higher prices don't guarantee better customer ratings.*

Rating Distribution by Price Range:

Price Range	Average Rating	Interpretation
$0-500	2.61/5	Budget laptops show decent satisfaction
$500-1000	1.93/5	Mid-range shows lower satisfaction
$1000-1500	2.28/5	Upper-mid range moderate satisfaction
$1500-2000	2.67/5	Premium laptops show higher satisfaction

Notable Pattern: The mid-range ($500-1000) shows the **lowest** average rating, while both budget ($0-500) and premium ($1500-2000) categories show higher satisfaction, suggesting a "value at extremes" pattern.

4.4 Comprehensive Report Dashboard



Comprehensive Analysis Dashboard

This integrated dashboard provides a holistic view of the laptop e-commerce market, combining:

- Market share visualization
- Price distribution patterns
- Rating metrics
- Brand-based comparisons
- Correlation analysis

Dataset Summary Statistics:

- **Total Products**: 116
- **Unique Brands**: 16
- **Total Customer Reviews**: 798
- **Average Price**: $907.76
- **Price Range**: $295.99 - $1,799.00

- **Average Rating**: 2.35/5

---

5. Conclusions

5.1 Key Insights

Based on our comprehensive ETL analysis of 116 laptop products, we have uncovered several important insights:

Market Leadership

Acer dominates the market with 25 products (21.6% market share), followed by Lenovo and Dell (each at 17.2%). The top 4 brands control nearly three-quarters of the market.

Pricing Strategies

- **Premium Positioning**: Apple maintains the highest average price ($1,313.64), clearly positioned as a premium brand
- **Budget Accessibility**: Prestigio offers the most affordable options ($299.00), serving price-sensitive consumers
- **Market Average**: The overall market average is $907.76, with most laptops clustered in the $500-1,200 range

Customer Satisfaction

- **Overall Rating**: The market average rating is 2.35/5, indicating room for improvement across the industry
- **Highest Rated**: ProBook achieves the highest customer satisfaction (4.00/5)
- **Engagement Leader**: Lenovo leads in customer engagement with 157 total reviews

Price-Quality Relationship

Our analysis reveals a **weak negative correlation (-0.091) between price and customer ratings**. This finding is significant:

> *[!IMPORTANT]*
> ***Consumer Recommendation**: Higher prices do NOT guarantee higher customer satisfaction. Budget laptops can provide excellent value for money, and consumers should focus on specific features and reviews rather than price as a quality indicator.*

Market Segmentation

The market shows clear segmentation:

- **Premium Segment** ($1,200+): Apple, Toshiba, ThinkPad - Higher prices but not necessarily higher ratings
- **Mid-Range** ($600-1,200): Dell, Asus, Lenovo - Largest market segment with varied satisfaction
- **Budget** (<$600): Acer, Prestigio, Aspire - Competitive pricing with reasonable satisfaction

5.2 Business Recommendations

Based on our analysis:

For Consumers:

1. Don't equate high price with high quality - read reviews carefully
2. Budget brands (Acer, Aspire) offer competitive products at lower prices
3. Consider brands with high engagement (Lenovo) as they reflect active user communities

For Retailers:

4. Stock a diverse range from the top 4 brands (72% of market demand)
5. Highlight budget options - they show good satisfaction ratings
6. Focus on brands with high review counts for marketing (indicates customer trust)

For Manufacturers:

7. Address the low overall satisfaction (2.35/5) - opportunity for differentiation
8. Premium pricing must be justified with tangible quality improvements
9. Mid-range segment ($500-1000) needs attention - lowest satisfaction ratings

5.3 Project Achievements

This ETL project successfully:

- Extracted **117 laptop records** (exceeding 100 record requirement)
- Collected **6 attributes** meeting multi-member requirement
- Implemented **6 data cleaning operations** (exceeding 4 operation requirement)
- Generated **5+ comprehensive visualizations** (exceeding requirements)
- Produced **clean CSV dataset** ready for further analysis
- Delivered **actionable business insights** from data analysis

---

6. Technical Implementation

6.1 Project Structure

```
Data Engineering Project/
├── etl_laptop_analysis.py      # Main ETL pipeline script
├── data_analysis.py             # Analysis and visualization script
├── requirements.txt             # Python dependencies
├── output/
│   ├── raw_data.csv            # Original scraped data (117 records)
│   ├── laptops_clean_data.csv  # Cleaned dataset (116 records)
│   ├── data_cleaning_comparison.png
│   ├── price_distribution_analysis.png
│   ├── brand_popularity_analysis.png
│   ├── rating_price_correlation.png
│   └── comprehensive_report_dashboard.png
└── README.md                    # Project documentation
```

6.2 Technologies Used

Technology	Version	Purpose
Python	3.11	Core programming language
BeautifulSoup4	4.12.2	HTML parsing for web scraping
Requests	2.31.0	HTTP requests for data extraction
Pandas	2.1.4	Data manipulation and analysis
Matplotlib	3.8.2	Data visualization
Seaborn	0.13.0	Statistical visualizations
NumPy	1.26.2	Numerical computations

6.3 How to Run the Project

Prerequisites:

```bash
# Install Python 3.11 or higher
# Install dependencies
pip install -r requirements.txt
```

Execution Steps:

```bash
# Step 1: Run ETL pipeline
python etl_laptop_analysis.py

# Step 2: Run analysis
python data_analysis.py
```

Expected Output:

- Raw data CSV file
- Cleaned data CSV file
- 5 visualization PNG files
- Console output with statistics and insights

6.4 Code Repository

> *[!NOTE]*
> ***GitHub Repository**: The complete source code for this project is available at:*
>
> *[ymyex/Data-Engineering-Project]*

Repository Contents:

- All Python scripts (`etl_laptop_analysis.py`, `data_analysis.py`)
- Requirements file (`requirements.txt`)
- Cleaned dataset (`laptops_clean_data.csv`)
- All generated visualizations
- Project documentation (README.md)
- This report (report.md)

---

7. Appendices

Appendix A: Data Dictionary

Column Name	Data Type	Description	Example
Product Name	String	Full laptop model name	"Asus VivoBook X441NA-GA190"
Brand	String	Extracted brand name	"Asus"
Price (USD)	Float	Product price in USD	295.99
Rating	Integer	Customer rating 1-5	3
Review Count	Integer	Number of customer reviews	12
Description	String	Product description	"15.6 inch HD display..."
Extraction Date	DateTime	When data was scraped	"2026-01-13 17:48:00"

Appendix B: Statistical Summary

```
Cleaned Dataset Descriptive Statistics:

       Price (USD)      Rating  Review Count
count   116.000000  116.000000    116.000000
mean    907.759310    2.353448      6.879310
std     402.661144    1.105357      4.269518
min     295.990000    1.000000      0.000000
25%     468.965000    1.000000      3.000000
50%    1106.400000    2.000000      7.000000
75%    1222.182500    3.000000     10.000000
max    1799.000000    4.000000     14.000000
```

Appendix C: Complete Brand Statistics

Brand	Count	Avg Price	Avg Rating	Total Reviews	Market Share
Acer	25	$636.84	2.24	175	21.6%
Lenovo	20	$852.31	2.15	157	17.2%
Dell	20	$1,107.55	2.35	135	17.2%
Asus	19	$996.14	2.53	142	16.4%
MSI	6	$1,187.33	2.50	48	5.2%
ThinkPad	5	$1,198.79	2.40	35	4.3%
Toshiba	5	$1,226.61	2.60	40	4.3%
Others	16	$766.54	2.56	66	13.8%

Report Metadata

Project Type: ETL (Extract, Transform, Load) Data Engineering Project

Course: TTTC3213

Submission Format: PDF

Date: January 7, 2026

Word Count: ~3,500 words

Visualizations: 5 comprehensive charts

Code Repository: ymyex/Data-Engineering-Project