Vietnam National University Ho Chi Minh City

University of Science

Faculty of Information Technology

PROJECT PROGRESS REPORT
Brand Fashion Market Trend Analysis

Course	Introduction to Data Science
Class	22CLC
Students	**22127225 – Trần Thị Thiên Kim** **22127357 – Phạm Trần Yến Quyên** **22127374 – Lê Thanh Tâm** **22127449 – Mai Đức Vân**
Github:	Brand Fashion Market Seasonal Trend Analysis

HCMC, 2024

To have a good experience, please use a PDF reader that supports hyperlinks

Mục lục

1 Current Status of the Project

1.1 Data Collection

- We've collected data from reputable sources that include branded fashion shops, including:
  - Tiki: Collected product data including prices, brands, sales numbers, and review statistics. This platform provides a wide range of fashion items that cater to diverse customer needs.
  - Uniqlo: Data was scraped to understand how a global brand structures its inventory and categorizes products for customers.
  - Shein: Using Selenium for dynamically rendered pages.
- **Tools Used**:
  - `BeautifulSoup` and `Selenium` for scraping structured and dynamic web content.
  - `Requests` library for making HTTP requests to extract static data.
  - APIs where available for direct data extraction, reducing scraping overhead.
- **Process:**
  - Data was extracted using automation scripts.
  - Data cleaning and basic validation were applied during extraction to minimize errors.

1.2 Data Preprocessing

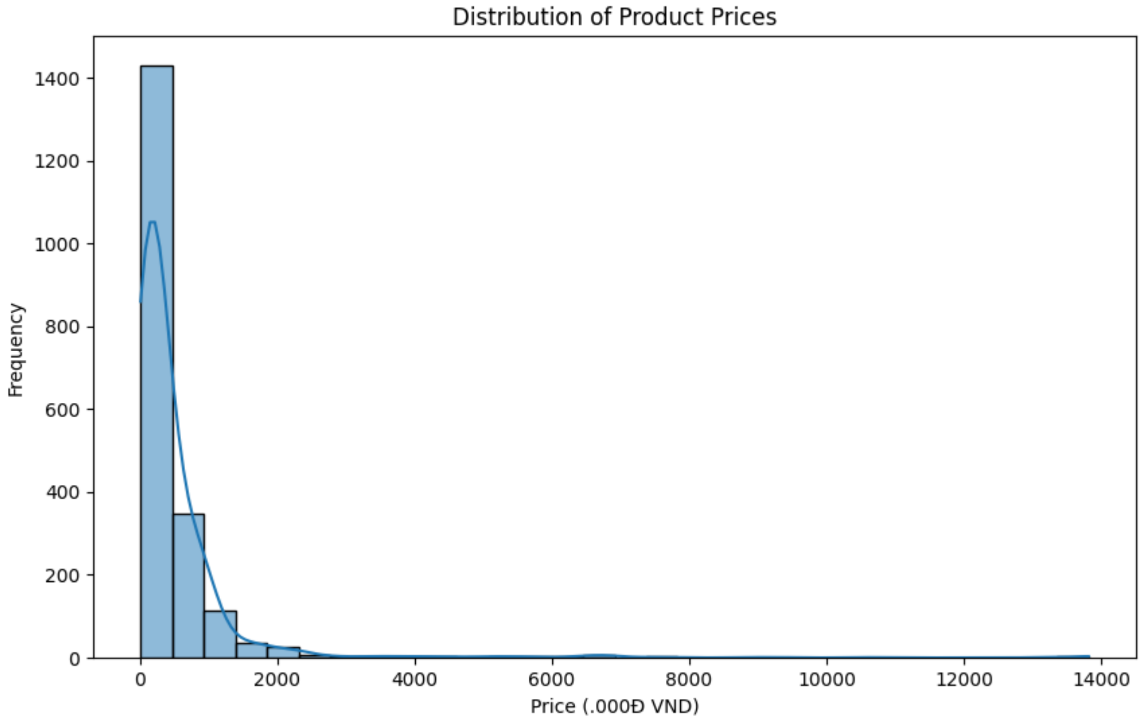
1. Handling missing values:
   (a) For the Tiki dataset, Thương hiệu (Brand) contains null values.
   (b) Products with 0 reviews and sales were flagged for removal or replacement.
2. Removing duplicates: Ensured no duplicated products in the datasets (Same product from the same shop but under a slighty different name, Ex: multiple listing in one product page, etc.)
3. Removing non-relevant features: Irrelevant columns (e.g., URLs for modeling) were retained only for validation and categorization purposes.

1.3 Initial Analysis

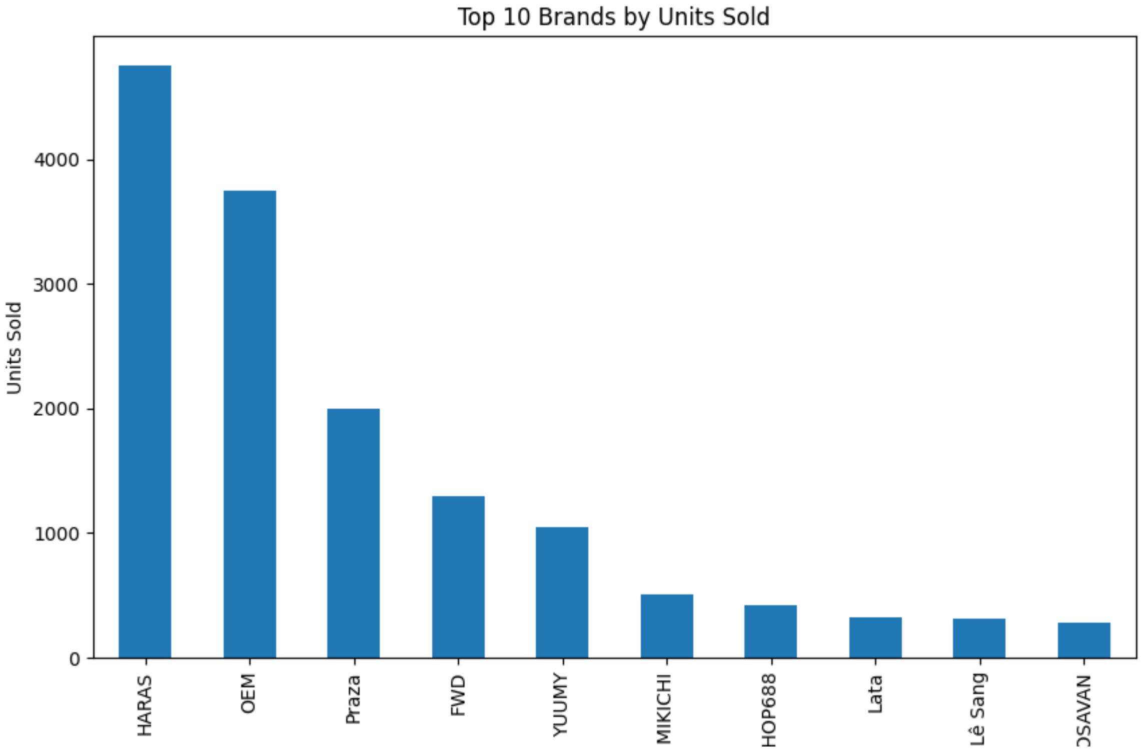
1.3.1 Visualization

Visualization tools were used to analyze patterns. In `tiki.csv`:

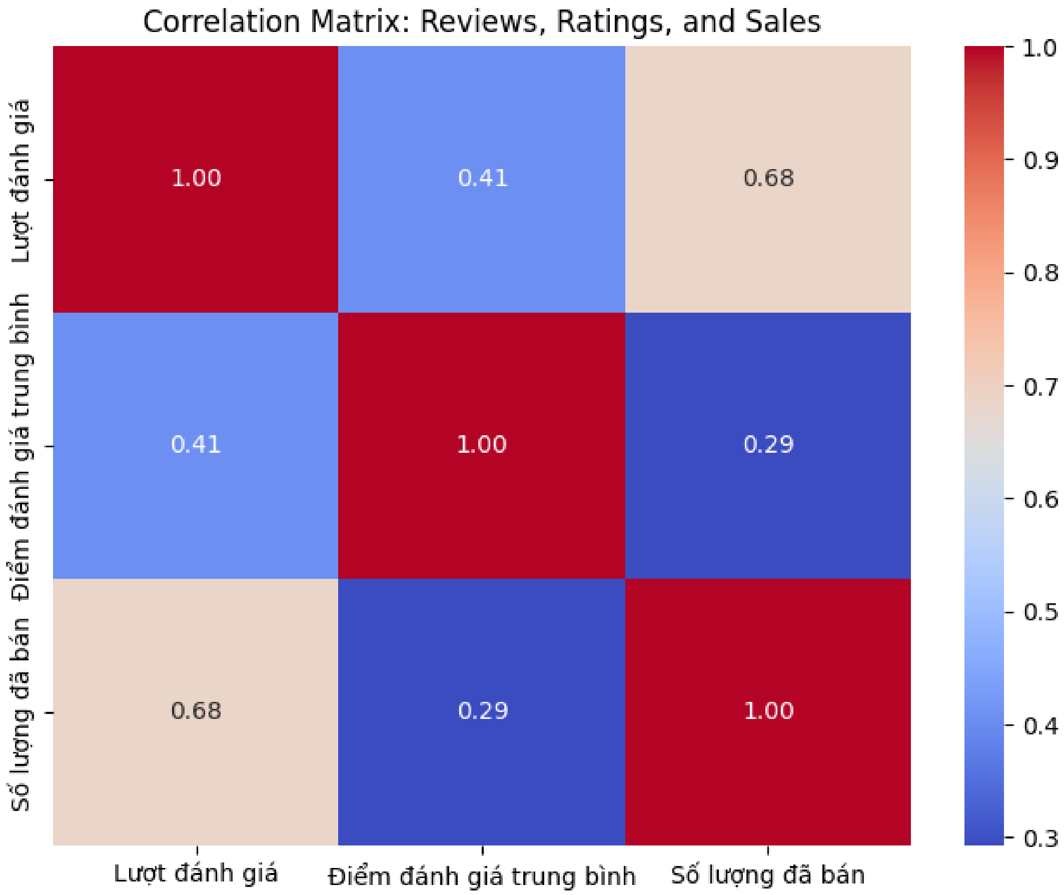
- Distribution of product prices.



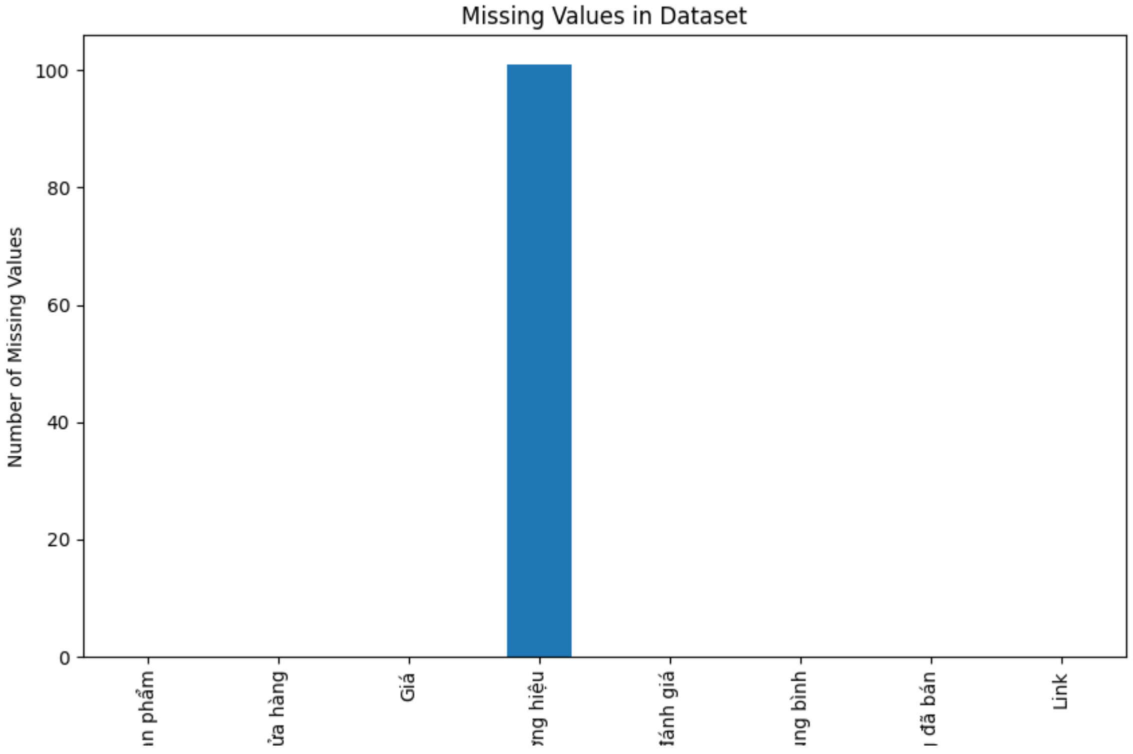
- Sales trends for different brands.



- Reviews and ratings correlation with sales.



- Missing values in the dataset.



1.3.2 Statistical Insights

- **Price Analysis:** Median product price is $155,000 VND, with a range between $50,000 VND and $2,000,000 VND.
- **Sales Performance:** Approximately 60% of products have fewer than 5 sales recorded.

1.4 Modeling (Theory only)

1. **Regression**:
   - Objective: Predict pricing trends based on features like sales, reviews, and brand.
   - Reason: Helps identify optimal pricing strategies and competitive positioning.
   - Proposed Models:
     - Linear Regression: Quick to implement, interpretable results.
     - Random Forest Regressor: Handles non-linear relationships and feature importance analysis.
2. **Clustering**:
   - Objective: Group products into clusters based on price, sales, and engagement.
   - Reason: Useful for segmentation and targeted marketing campaigns.
   - Proposed Models:
     - K-Means: Efficient for identifying distinct product categories.
     - DBSCAN: Handles clusters with varying densities and outliers.

3. **Performance Metrics**:
   - Regression: R^2 score for goodness-of-fit, Mean Absolute Error (MAE) for interpretability.
   - Clustering: Silhouette Score and Davies-Bouldin Index for cluster quality.

2 Current Issues

2.1 Data Quality Issues

1. Some of the current quality issues are:
   - Missing values in `Thương hiệu`.
   - Many products have `Lượt đánh giá` (reviews) and *Số lượng đã bán* (units sold) as 0, indicating potential invalid or inactive data.
   - Non-standardized format for brand names, leading to data inconsistencies.

2.2 Technical Challenges

1. **Dynamic Content**: Shein's use of JavaScript requires Selenium, which increases resource consumption and scraping time.
2. **API Limitations**:
   - Tiki API pagination and rate limits required delays between requests.
   - Websites like Shein doesn't provide easy to access APIs for crawling data.
   - **Crawl-prevention**: Some website like Shopee make it so that reviews and some other important data can only be crawl fully through their APIs.

3 Proposed Solutions to Current Issues

3.1 Data Quality Solutions

- **Data Enrichment:** Incorporate additional data sources to improve data coverage and variety.
- **Imputation:** Use statistical methods to estimate missing values (e.g., median imputation for prices).

3.2 Improving Model Performance

- Feature engineering, such as creating composite features (e.g., price-to-rating ratio).
- Hyperparameter tuning for optimal model performance.

3.3 Technical Adjustments

- Use `Scrapy` for faster scraping and parallel processing.
- Employ proxies to avoid IP bans during large-scale data collection.

3.4 Time Management

- Prioritize preprocessing and modeling over secondary tasks like extended visualizations.

4 Plan for Remaining Time

4.1 Immediate Next Steps

– Finalize preprocessing steps:

  (a) Handle missing values and remove low-quality data.

  (b) Perform exploratory data analysis (EDA) with visualizations.

  (c) Begin testing regression and clustering models.

4.2 Timeline

Break down remaining weeks with tasks:

- **Week 1:** Complete data cleaning and feature engineering.
- **Week 2:** Develop and test initial models.
- **Week 3:** Finalize analysis and compile documentation.

4.3 Milestones

Specify milestones such as completing data preprocessing, achieving model performance goals, and generating insights.

4.4 Final Deliverables

- Comprehensive final report with actionable insights.
- Model performance summary and visualization dashboard.

5 Appendices

5.1 Code Snippets Example:

1. Syntax API urls:

```
api_url = f"{self.base_url}?limit={self.limit_per_page}&q={self.query}&page={page}"
```

2. Filter out unsold products:

```
data[’Thương hiệu’] = data[’Thương hiệu’].fillna(data.groupby(’Category’ [’Thương hiệ
data = data[data[’Số lượng đã bán’] > 0]
```

5.2 Data Samples

1. `tiki.csv`:
   (a) Dataset Information:
   - Total Records: 2000 (per category)
   - Columns: 8
     - Tên sản phẩm (Product Name)
     - Tên cửa hàng (Store Name)
     - Giá (Price)
     - Thương hiệu (Brand) - Some null values
     - Lượt đánh giá (Number of Reviews)
     - Điểm đánh giá trung bình (Average Rating)
     - Số lượng đã bán (Units Sold)
     - Link (URL to Product)

   (b) Some Data Overview:

Tên sản phẩm	**Tên cửa hàng**	**Giá**	**...**
Đầm Jean Nữ Thời Trang	THỜI TRANG TINA	235000	...
Đầm bông thời trang	Hương Nemo Style	155000	...

2. `uniqlo.csv`:
   (a) Dataset Information:
   - Total Records: 955 (Uniqlo's official product)
   - Columns: 8
     - Product ID
     - Product Name
     - Price Currency
     - URL
     - Rating
     - Total Ratings
     - Fit
     - Rating Count

   (b) Some Data Overview:

Product ID	**Product Name**	**Price**	**...**
E471117-000	AIRism Áo Hoodie Chống UV	686000.0000	...
E467410-000	Áo Parka Chống UV Bỏ Túi	784000.0000	...

5.3 References and Links

- Tiki API Documentation
- Selenium documentation.
- BeautifulSoup Documentation