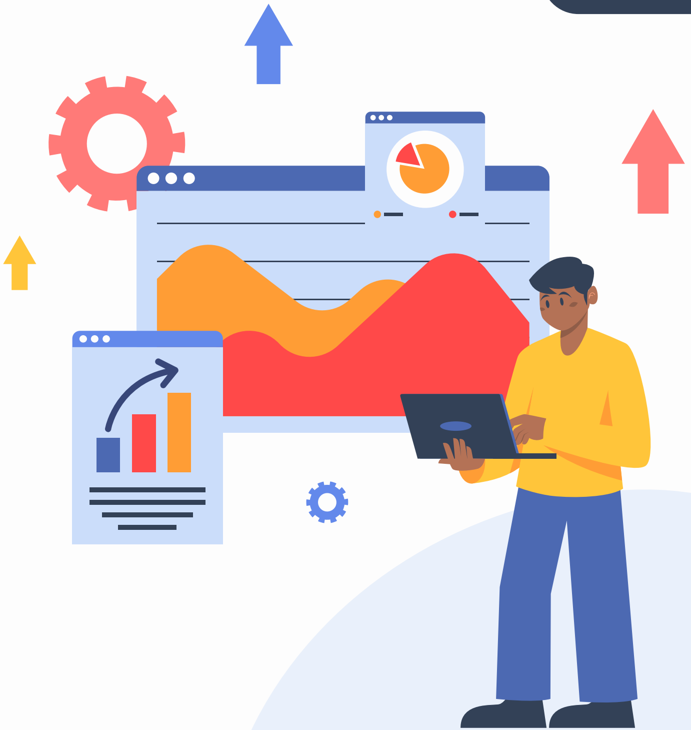
06 – LoveData

Brand Fashion Market Trend Analysis

Progressing Check-In



Contents of project

Current Status of the Project	A summary of entire current progress of the project, from its inception to the current stage
Current Issues	Focus on identifying and analyzing problems being encountered in the project, to clarify the challenges that hinder the progress and quality of the project
Proposed Solutions to Current Issues	Focus on how to solve the challenges listed in the "Current Issues" section, to demonstrate project initiative and problem management.
Plan for Remaining Time	Detailed plan to complete the remaining work in the project
Appendices	Provides additional information that helps clarify the report and aids in checking or referencing details

Table of contents

01 Current Status of the Project

Critical thinking skills and mathematical reasoning

02 Current Issues

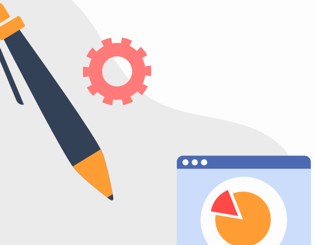
Analyzing and interpreting mathematical models

03 Proposed Solutions to Current Issues

Identifying mathematical concepts in real-world scenarios

04 Plan for Remaining Time

Integration of technology tools in mathematical exploration



01

Current Status of the Project

Data Collection

Sources

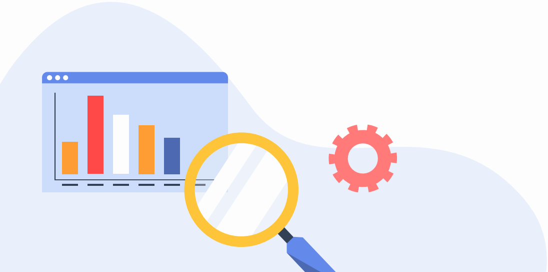
- **Tiki:** Product data (prices, brands, sales, reviews).
- **Uniqlo:** Data on inventory structure and product categorization.
- **Shein:** Scraped dynamic content using **Selenium**.

Tools used

- **BeautifulSoup** and **Selenium** for web scraping.
- **Requests** for static data extraction.
- **APIs** for direct data extraction.

Process

- Automated data extraction with validation and cleaning.



Data Preprocessing

Handling Missing Values

- **Tiki dataset:** Brand (Thương hiệu) had null values.
- Products with 0 reviews and sales were flagged for removal or replacement.

Removing Duplicates

Ensured no duplicated products (e.g., same product listed under slightly different names).

Removing Non-Relevant Features

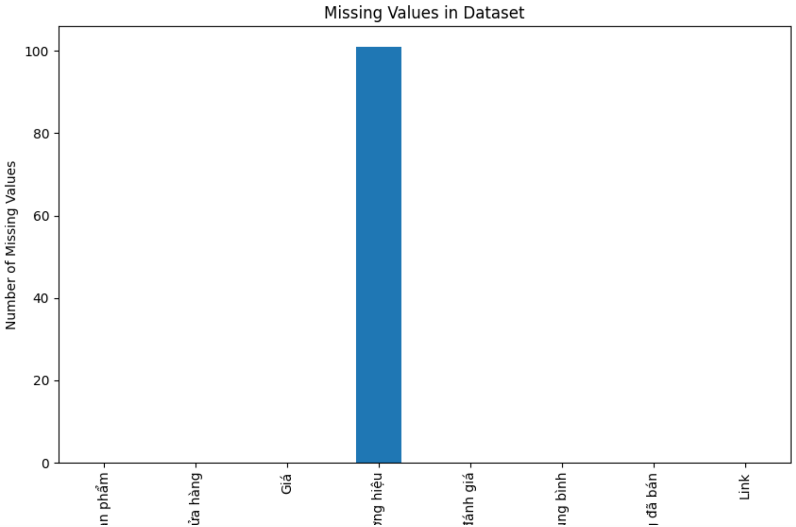
Irrelevant columns (e.g., URLs) were retained for validation and categorization only.

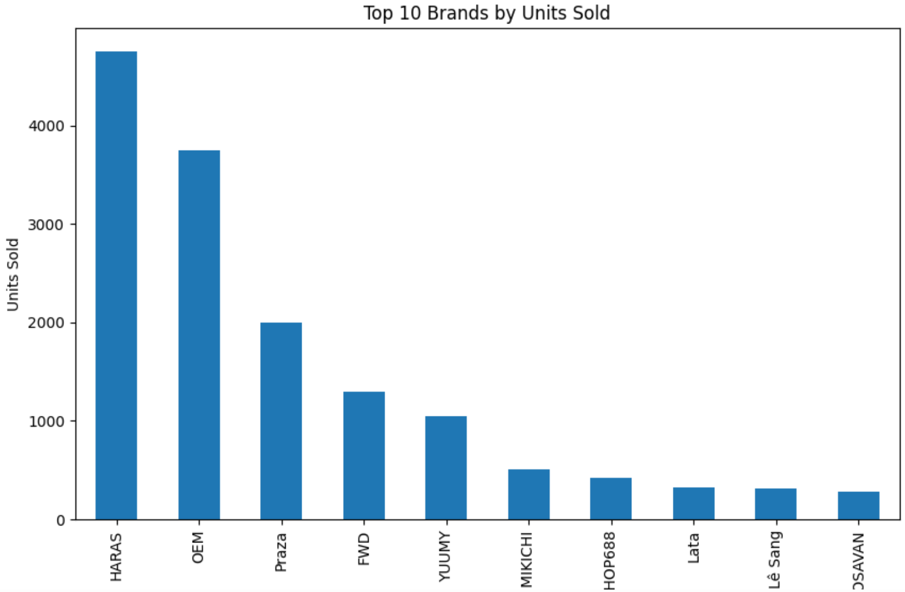
Statistical Insights

Price Analysis: Median product price is **$155,000 VND**, with a range between **$50,000VND and $2,000,000 VND**.

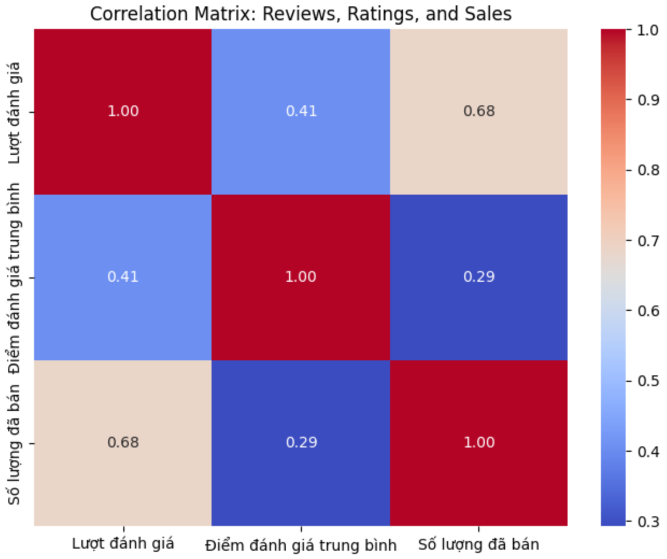
Sales Performance: Approximately 60% of products have fewer than 5 sales recorded

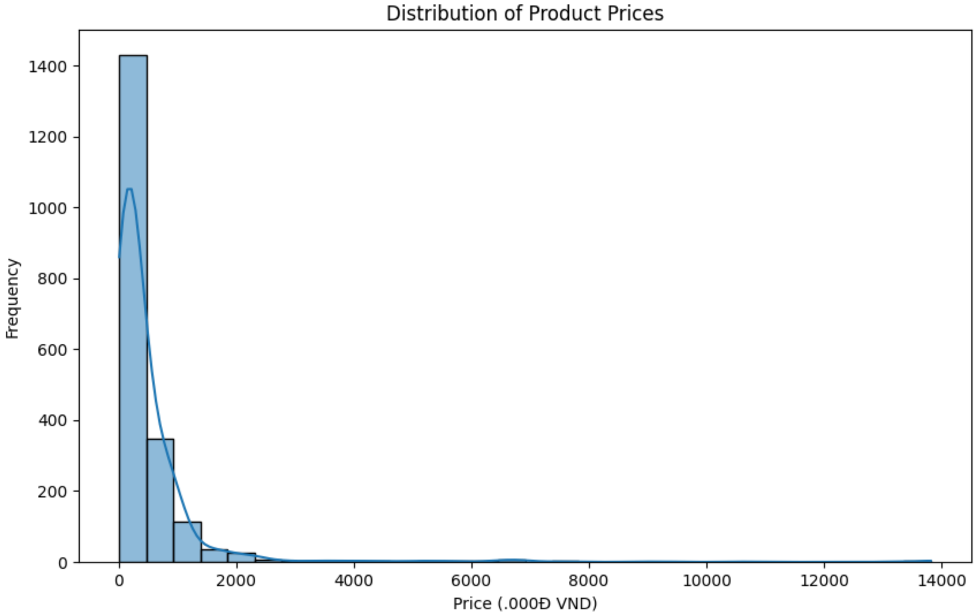
Visualization





Visualization





Modeling (Theory)

Regression

- **Objective:** Predict pricing trends based on sales, reviews, and brand.
- **Models:**
- **Linear Regression:** Simple, interpretable.
- **Random Forest:** Handles non-linearity and feature importance.

Clustering

- **Objective:** Group products by price, sales, and engagement.
- **Models:**
- **K-Means:** Identifies product categories.
- **DBSCAN:** Handles varying densities and outliers.

Performance Metrics

- **Regression:** R^2, MAE.
- **Clustering:** Silhouette Score, Davies-Bouldin Index.

02

Current Issues

2.1 Data Quality Issues

- **Missing Values:** Brand names (Thương hiệu) have missing entries.

- **Inactive Data:** Many products show **0 reviews** and **0 units sold**, indicating potentially invalid or irrelevant entries.

- **Inconsistent Brand Names:** Non-standardized formats for brand names lead to inconsistencies in the data.

2.2 Technical Challenges

Dynamic Content

Platforms like **Shein** use JavaScript for rendering content, requiring tools like **Selenium**, which increases resource usage and scraping time.

API Limitations

Tiki: Pagination and rate limits required adding delays between API requests to avoid blocks.

Shein: Limited access to APIs, making scraping necessary but time-consuming.

Crawl Prevention: Some platforms, like **Shopee**, restrict access to reviews and important data, which can only be fully extracted via their APIs.

03

Proposed Solutions to Current Issues

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 forprices).

Improving Model Performance

Feature engineering, such as **creating composite** features (e.g., price-to-rating ratio).

Hyperparameter tuning for optimal model performance.

Technical Adjustments

Use **Scrapy** for faster scraping and parallel processing.

Employ proxies **to avoid IP bans** during large-scale data collection.

Time Management

Prioritize preprocessing and modeling over secondary tasks like extended visualizations.

04

Plan for Remaining Time

Immediate Next Steps

- Handle **missing values** and remove **low-quality data**
    - Perform exploratory data analysis (EDA) with **visualizations**.
- Begin testing regression and clustering models

Timeline

40%

Week 1

Complete data cleaning and feature engineering

70%

Week 2

Develop and test initial models

100%

Week 3

Finalize analysis and compile documentation

Milestones

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

Final Deliverables

Comprehensive final report with actionable **insights**.

Model performance summary and **visualization** dashboard.

05

Appendices

Code Snippets Example

API URL Syntax:

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

Filter Out Unsold Products:

```
data['Thương hiệu'] = data['Thương
hiệu'].fillna(data.groupby('Category')['Thương hiệu'].transform('mode'))
data = data[data['Số lượng đã bán'] > 0]
```

Data Sample

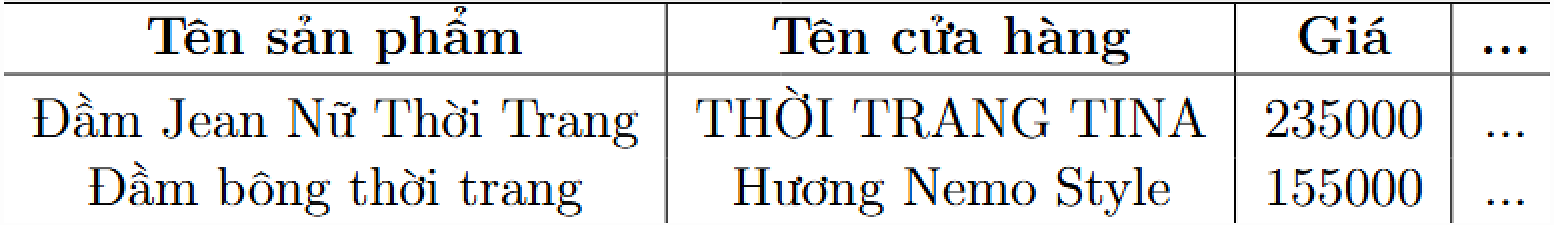
tiki.csv

- **Total Records:** 2000 (per category)
- **Columns:**
  - Product Name
  - Store Name
  - Price
  - Brand (Some missing values)
  - Number of Reviews
  - Average Rating
  - Units Sold
  - Product URL

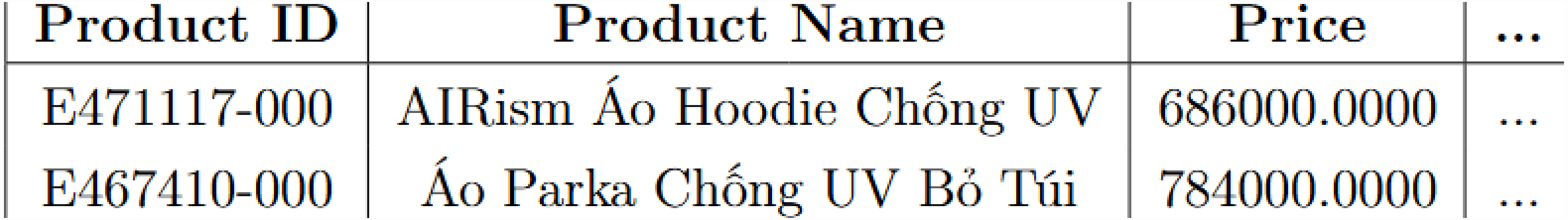
uniqlo.csv

- **Total Records:** 955 (Official Uniqlo products)
- **Columns:**
  - Product ID
  - Product Name
  - Price Currency
  - URL
  - Rating
  - Total Ratings
  - Fit
  - Rating Count

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	...



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

References and Links

- Tiki API Documentation
- Selenium Documentation
- BeautifulSoup Documentation

Thanks!

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**