📚 Book Data Analysis – Final Project Report

**1. Executive Summary**

**This project analyzes book data scraped from "Books to Scrape" (**[**http://books.toscrape.com/**](http://books.toscrape.com/)**) to uncover market trends. We developed a Streamlit dashboard with interactive visualizations and conducted statistical analyses on 1,000+ books, examining relationships between price, ratings, and availability.**

**2. Project Overview**

**🧠 Objectives**

* **Perform exploratory data analysis (EDA) on book attributes**
* **Identify correlations between price and ratings**
* **Test statistical hypotheses about market trends**
* **Build an interactive dashboard for data exploration**

**📈 Key Insights**

* **Average book price: £35.38 ± £11.51 (SD)**
* **Rating distribution skews positive (Mean: 3.9/5)**
* **No significant price difference between available/out-of-stock books (p=0.42)**

**3. Methodology**

**📥 Data Collection & Cleaning**

|  |  |
| --- | --- |
| Component | Details |
| Source | **Scraped 50 pages (1,000+ books) using BeautifulSoup** |
| Cleaning | **Removed special characters, standardized ratings (1-5 scale)** |
| Enhancements | **Added price categories (Cheap/Medium/Expensive)** |
| Storage | **MongoDB Atlas → CSV for analysis** |

**📊 Analytical Approach**

1. **Descriptive Statistics**
   * **Calculated central tendency and dispersion metrics**
   * **Generated variance-covariance matrices**
2. **Inferential Statistics**
   * **Shapiro-Wilk normality tests**
   * **T-tests for price comparisons**
   * **Chi-square tests for categorical relationships**
3. **Visual Analytics**
   * **Interactive Plotly charts**
   * **Correlation heatmaps**
   * **Word frequency analysis**

**4. Key Findings**

**🔍 Statistical Results**

|  |  |  |  |
| --- | --- | --- | --- |
| Test Type | Price | Rating | Conclusion |
| Normality (Shapiro) | **p < 0.001\*** | **p = 0.018\*** | **Non-normal distribution** |
| Price vs. Rating | **r = 0.13** | **-** | **Weak positive correlation** |
| Availability | **χ² p = 0.36** | **-** | **No inventory-rating relationship** |

**📉 Visual Discoveries**

1. **Price Distribution**
   * **68% of books priced £20-£50**
   * **Right-skewed with luxury outliers (>£100)**
2. **Rating Analysis**
   * **41% of books have 4-star ratings**
   * **Only 6% received 1-star reviews**
3. **Text Mining**
   * **Top title keywords: "Guide" (12%), "Novel" (9%), "Life" (7%)**

**5. Dashboard Features**

**🚀 Interactive Tools**

* **Dynamic Filtering: By price range, rating, availability**
* **Statistical Testing: On-demand hypothesis tests**
* **Export Capability: Download filtered datasets**

**📊 Visualization Suite**

1. **Histograms with availability overlay**
2. **Scatter plots with trendlines**
3. **Box plots by rating category**
4. **Word clouds from book titles**

**7. Conclusions & Recommendations**

**🎯 Key Takeaways**

1. **Price and rating show minimal correlation (r=0.13)**
2. **Inventory status doesn't affect ratings (p>0.05)**
3. **Strong clustering in £20-50 price range**

**💡 Business Implications**

* **For Sellers: Focus on mid-range priced books (highest demand)**
* **For Buyers: Higher prices don't guarantee better ratings**
* **For Developers: API integration potential for real-time updates**

**8. Team & Contributions**

|  |  |  |
| --- | --- | --- |
| Member | Role | Key Contribution |
| Ahmed Hassan Ali | **Project Lead** | **Scraper development, Statistical analysis** |
| Fares Ahmed | **Data Engineer** | **Database design, Cleaning pipeline** |
| Mohamed Ashraf | **Visualization Specialist** | **Dashboard UI, Interactive charts** |
| Farouq Ashraf | **QA Analyst** | **Validation, Hypothesis testing** |
| Nour Mamduh | **Technical Writer** | **Documentation, Final report** |