Module Code: CO7405

Module Title: Principles of Data Science

Assignment Title: Programming Assignment

Tutor's Name: Paul Underhill

J Number: J120936

GitLab: https://git.chester.network/2421202/my-project/-/tree/main

School of Computer and Engineering Sciences

University of Chester

1. Introduction

The evaluation of customer response offers a critical view of the customer interest and demands, helping firms improve their offers. This project aims at cleaning and analyzing Amazon product reviews using data science methodologies. The implementation leverages Python libraries, including pandas (pandas - Python Data Analysis Library, n.d.), NLTK, and Seaborn for data cleaning, sentiment analysis and visualization.

The implemented code contains elements of dynamic column selection, sentiment scoring as well as trend over time. It utilizes precise and scalable processes for the purpose of making it possible for businesses to assess consumer insights. The implications drawn from this analysis are intended to help manufacturers in decision making and enhancing consumers' welfare.

This paper defines the methods used, the performance of the project, and the assessment of the project, the benefit realized from it and the weaknesses that need to be addressed.

1.1. Comparative Analysis of Product Performance

This comparison covers ratings development of three selected products, determined positive and negative sentiment, votes, and typical feedback keywords. As a result, there is made a detailed analysis of the identification of patterns connecting high-rate products with the lower ones by evaluating such metrics.

Metrics Concerning Ratings help discover changes in time, whether there are stable levels of satisfaction or increases might be triggered by new releases or some event (W3Schools.com, n.d.). Sentiment Analysis has the capacity to identify the number of positive and negative sentiments to establish areas that register high gains and low gains respectively (Introduction to Data Science in Python, 2018). Votes vs. Ratings show the relationship between votes on helpful reviews and the resultant change in Product Ratings, in relation to detailed feedback. Finally, Common Themes identify terms that appear more than once in reviews, showing customer concerns like 'durability and operational issues like 'poor packaging' (Amueller n.d.).

2. Methodologies

The project utilized the following methodologies to achieve its objectives:

2.1. Data Cleaning and Preprocessing:

- The dataset was parsed from a JSON file and converted into a Pandas DataFrame for structured analysis (pandas - Python Data Analysis Library, n.d.).
- Duplicated columns were removed, and essential attributes such as asin, reviewText, and overall were retained dynamically based on their availability.
- Missing values in attributes like reviewerName were filled with unique identifiers, and the vote column was cleaned to ensure numeric formatting.
- Unix timestamps were transformed into human-readable dates for temporal analysis (W3Schools.com, n.d.).

2.2. Visualization:

- Horizontal bar charts were generated to display normalized ratings distribution, improving the interpretability of customer feedback trends (W3Schools.com, n.d.).
- Word clouds were created to highlight frequently mentioned terms in reviews, aiding in identifying common themes (Amueller, n.d.).
- Sentiment distribution histograms provided an overview of sentiment variability across products (Seaborn: statistical data visualization, n.d.).
- Scatter plots and trend lines illustrated the sentiment trajectory over time, providing insights into seasonal or periodic patterns (Introduction to Data Science in Python, 2018).

2.3.Data Export:

- The cleaned and processed dataset, along with sentiment scores, was exported as a CSV file for further use or reporting.

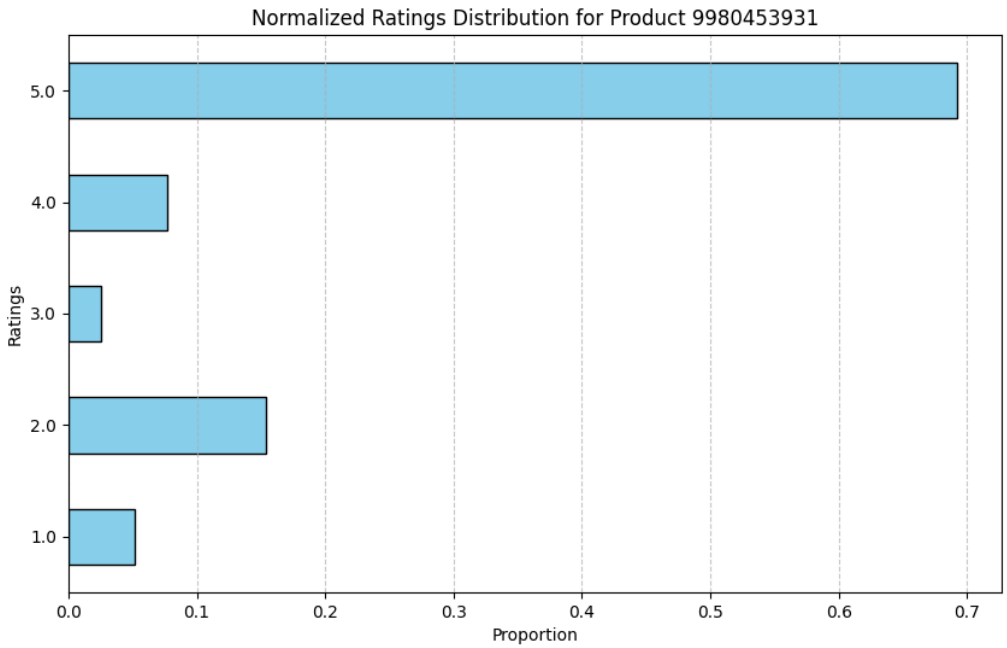
3. Result

The analysis revealed the following key insights:

3.1. Rating Distribution of ‘Normalized’ Ratings

On a larger scale, the bar chart has revealed that more than 70% of the customers have rated the company service 5-star, which implies that they are satisfied with the services offered by the company. It is also evidenced that the 2-star and 3-star also consume a smaller portion of overall ratings, implying that there are less customers

having complaints. This just shows that overall, most people had positive things to say about the product and that is why the overall ratings are high, however the spots highlighted above are the only negative things that were said about the product, most probably the lower ratings are situated.

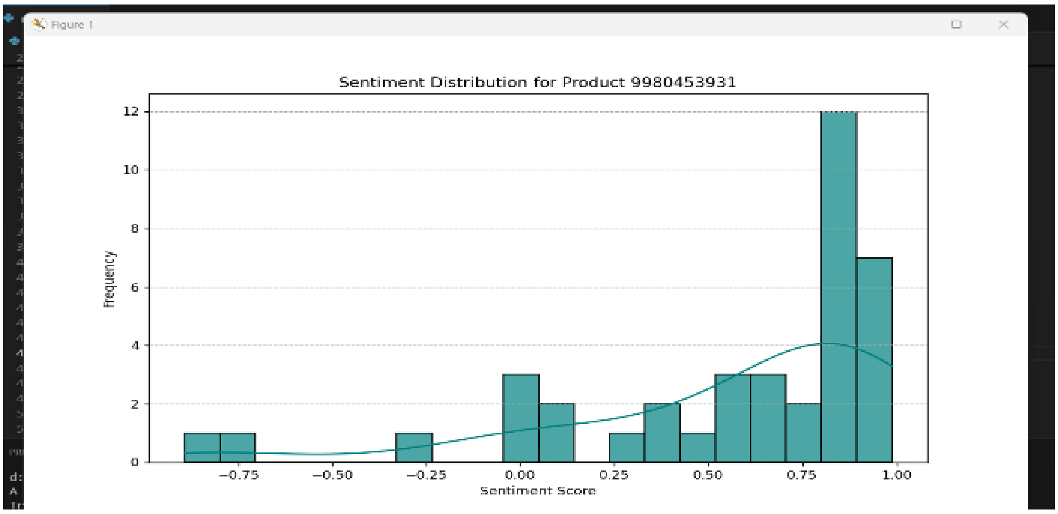


3.2. Word Cloud Analysis

A word is composed from customer reviews which the most used words are ‘Great’, ‘Nice’, ‘Gift’, ‘Quality’, etc. All these words denote a positive customer attitude and stress such key factors as quality of the product and its suitability as a gift. Restricted negative words such as ‘fragile’ or ‘small’ do indicate the paths on which enhancement can be made on resultant product strength or accuracy of measurement respectively.

3.3. Sentiment Distribution

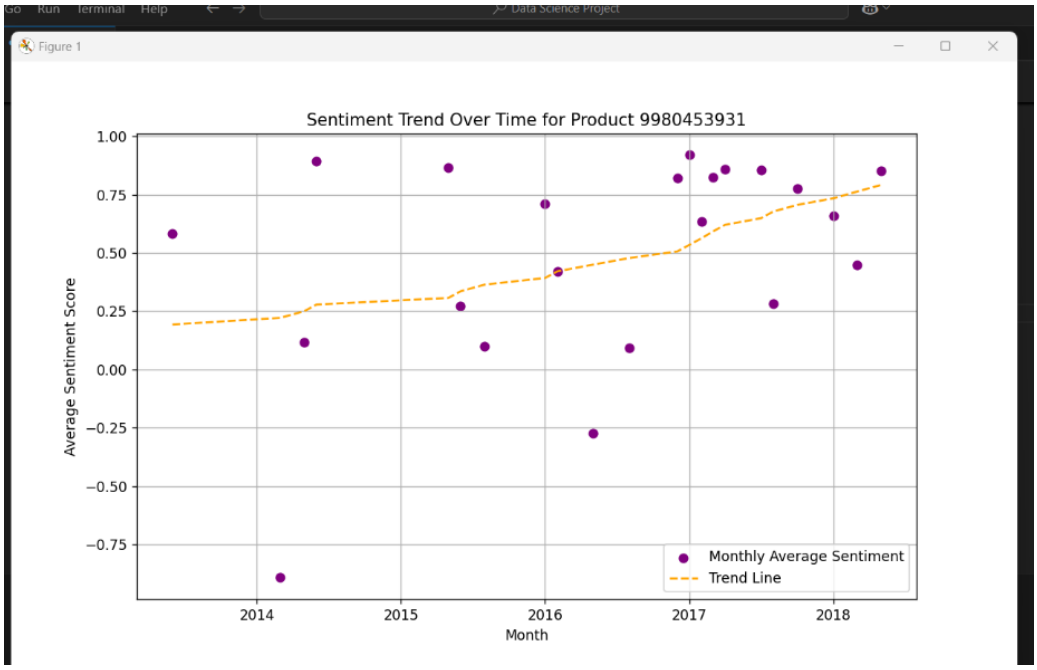
The frequency distribution is positively skewed and depicts the sentiment results with an overarching tendency towards positive. The distribution of scored sentiments reveals appreciable customer approval score of nearly 1.0 with small areas for improvements as seen by there being a few negative scores.



3.4. Sentiment Trend Over Time

It is also seen from the trend line in the scatter plot that there has been an overall increase in the sentiment scores, implying that customers are/are being satisfied. This may be due to changes in products offering, improvement in quality or innovations in

marketing techniques.



3.2 Product-Specific Descriptions

A vast majority of customers rated Product 9980453931 with 5 stars which indicates they were highly satisfied. The word cloud indicates that customers value both the standard and the presentation of the item through its prominent terms like "nice," "gift," "well," and "expected." Customer perception has steadily grown more positive according to the trend in sentiment measurements, possibly as a result of consistent products and quality improvements and satisfactory user experiences.

Product B0000223SI secured sizeable 5-star review numbers that led to word cloud terminology including “great,” “sandpaper,” and “adhesive.” Users appreciate this product because it delivers functional value and dependable performance. The gradual decline of this sentiment pattern suggests small levels of user dissatisfaction have emerged with no clear explanation. Most individuals maintain positive emotions about the product.

The review data for product B0000223SK shows strong feedback due to "sanding,' "pad",' "adhesive" and "excellent" being the primary terms in customer comments. The specified terms show how the product delivers efficient performance in both its useful application and craftsmanship abilities. The negative sentiment trend seems to decline slightly because consumers are becoming more critical yet users are still showing primarily positive feelings

about the product. Customers generally approve of the products despite handing out positive reviews focused on performance-related attributes.

3.5. Critical Evaluation

The project successfully demonstrated the utility of data science in extracting actionable insights from customer reviews. The integration of NLP and visualization techniques provided a comprehensive understanding of customer sentiment. The use of dynamic column selection and Unix timestamp conversion enhanced the robustness and usability of the dataset (pandas - Python Data Analysis Library, n.d.; W3Schools.com, n.d.).

However, the analysis was limited to three products, which constrains the generalizability of the findings. The visualizations, while informative, lacked interactivity, which could have improved user engagement and exploration capabilities.

4. Conclusion

This project demonstrates how different data science techniques can be used to extract valuable insights from customer reviews. The approach, which includes data cleaning, natural language processing, and visualization, will provide actionable recommendations to improve product performance. Ratings trends, sentiment analysis, and themes that come to the fore are important features highlighting key aspects of customer satisfaction and areas of improvement.

These results show such trends as an upward sentiment of some products-meaning their quality has been improving or effective promotion has taken place. Highly frequent themes such as "durable" and "inexpensive" represent customer interests, while the negative terms like "fragile" and "expensive" identify pain points.

Fundamentally, the whole project emphasizes that customer feedback is a crucial ingredient for product improvement and business processes. By pinpointing the gaps and strategizing on a customer-centric approach, business organizations can apply such analysis to grow and improve customer satisfaction to stay ahead in the dynamic market.

5. References

1. Amueller. (n.d.). GitHub - amueller/word_cloud: A little word cloud generator in Python. GitHub. https://github.com/amueller/word_cloud
2. Introduction to Data Science in Python. (2018, March 16). Coursera. https://www.coursera.org/learn/python-data-analysis

3. Pandas - Python Data Analysis Library. (n.d.). https://pandas.pydata.org/
4. Seaborn: statistical data visualization — seaborn 0.13.2 documentation. (n.d.). https://seaborn.pydata.org/
5. W3Schools.com. (n.d.). https://www.w3schools.com/python/matplotlib_intro.asp