EDA of Amazon Reviews

Project Final Report – CSE 578

MUJAHED SYED, PRANAY REDDY VANCHA, SAKTHI LAKSHMI, GAURAV SHARMA
GROUP 7, TEAM NAME: DATA DETECTIVES

ABSTRACT

With the tremendous growth of the e-commerce industry, the data associated with those companies are also growing exponentially. To be the leader in the industry and to compete successfully with several other competitors, it is critical to put those data into good use. Exploratory Data Analysis, which is the process of obtaining more insight into the available data by discovering trends and revealing the hidden patterns.

Amazon is one of the largest e-commerce platforms having a huge user base. The customers have the option to leave reviews on the products purchased. The project focuses on performing Exploratory Data Analysis (EDA) on Amazon reviews. The 5-core dataset for each category is used for data analysis and visualization. The objective of the project is to develop a data dashboard which will be beneficial to both the customers and the sellers to understand more about the products.

To achieve this data dashboard, exploratory data analysis is conducted on the product reviews. Tools like Python, Pandas, Numpy, NLTK and Tableau are used in the project. The final dashboard contains several visualizations to depict the trend of the average rating of a particular product, the most frequently occurring words in the reviews, helpfulness of a review and the polarity of the review. Sentiment Analysis is conducted to determine the polarity of the review. The visualizations developed are sure to aid customers in product selection and help sellers to improve their product quality based on the customer feedback and hence increasing sales.

Table of Contents

1 INTRODUCTION

Background - Online shopping has gained popularity in recent years as customers could shop for items sitting on their couch. Also, customers can easily browse through several online shopping websites, compare and order desired products. One critical thing the customers lack in online shopping is that they do not get a chance to find the quality of the product. They solely rely on the reviews of fellow shoppers. However, one product could have several reviews and it is difficult for a customer to read every review.

Purpose - It is time consuming and a lot of work for a person to analyses reviews especially with the tremendous amount of data available. The purpose of this project is to perform Exploratory Data Analysis on such product reviews. Exploratory Data Analysis is the process of analyzing datasets to explore the main characteristics and present as visuals [1]. The project focuses on analyzing the customer reviews for the products on Amazon.

Significance - This EDA on the Amazon reviews would be beneficial to two categories of end users - the customers looking to buy a product as well as the sellers of the product. The Customers could easily get a picture of the quality of the product and make buying decisions. Meanwhile the sellers can also understand the negatives of their products easily and can aim towards Customer satisfaction. Our project will significantly improve customers' shopping experience and provide great customer satisfaction.

2 TEAM GOALS AND BUSINESS OBJECTIVES

Today one of the most prized possessions with large businesses is the tons of user data available to them, especially for a company like Amazon which has millions of daily active users (DAU). Thus, the Business objective of the project is to leverage the power of Visual Analytics to improve the understanding of the customer review data[3] within the company and utilize it to improve customer experience and help the primary stakeholders make business decisions that will ultimately result in increasing the revenue of the company.

The goal of the team is to extract valuable information from the data using various analytical techniques and then represent them visually, making it easier to understand the information. Precisely, we aim to create a dashboard for the stakeholders, where they will be able to see a few key performance indicators (KPIs) of a given product or product category, analyzed visually and represented using charts and graphs such as line chart, Donut chart, etc.

3 SCOPE AND LIMITATIONS

The scope of the project is to perform extensive analysis on the dataset available and is limited to creating visuals based on the reviews. Other features like product suggestions based on the reviews will not be done in this phase of the project and can be taken up for future enhancements.

4 ASSUMPTIONS

Business Assumptions:

- Agile methodology to be followed to deliver the project.
- The Project Team will work remotely and on a flexible schedule.
- All discussions pertaining to the project shall be done on a remote platform.
- The project team will document all the project deliverables and details by following the Agile documentation methodology.
- The license(s) of tools and software are required to be obtained at any stage of the project, it will be the stakeholder's sole responsibility.

Technical Assumptions:

- The 5-core dataset for each product category will be used for the project.
- The 5-core dataset for each category is the subset of the actual customer review data for each category.
- In the 5-core dataset, all users and items (each unique reviewer and product id) have at least 5 reviews.
- Python 3.6 and required packages will be used for the preprocessing and natural language processing tasks.
- Tableau Desktop 2020.2 will be used for performing visual analytics.

5 USER STORIES

We have two target audiences for this project, namely users buying the product and Amazon itself. So, we'll define the user stories for these two personas:

1.) As a customer, I want to find the polarity of product reviews before buying the product, so that I can make more informed decisions while buying products.
2.) As an e-commerce company (Amazon), we want to extract value from a huge amount of customer review data so that we can understand and further improve upon customer experience.
    a. Perform helpfulness analysis of the review and show the most helpful reviews to the customer.
    b. Perform review text sentiment analysis to understand the user's experience with the product and improve on suggestions.
3.) I would like to be able to select one or more product ID(s) and see only the information pertaining to those product ID(s) on the dashboard.
4.) I would like all the visualizations easily interpretable with proper labels and color schemes.

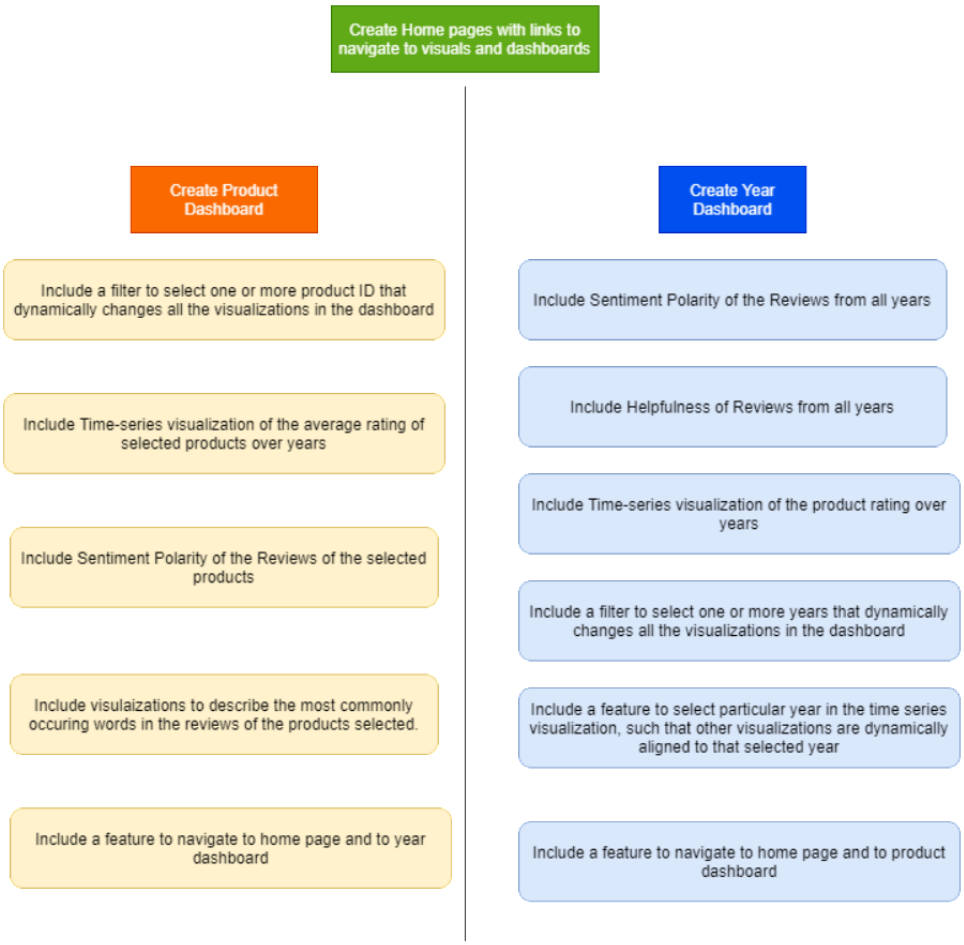

Figure 1: User Stories

6 INDIVIDUAL CONTRIBUTION

Role	**Name**	**Individual Contribution**
1. Team Lead	Mujahed Syed	Preprocessed the Data and Calculated Helpfulness rating of reviews using Python. Used the preprocessed data and created all year specific visualizations, year level dashboard, project start page and project contents (Home) page in Tableau. Developed the complete UI for the project and Published the project to Tableau Public to make the project accessible to everyone.

Team Member	Pranay Reddy	Pre-processed the text data and calculated the sentiment score of the reviews using Python. Used the pre-processed text data and normalised the text to use it in the Word Cloud. Pre-processed the data for Product level, and created Product specific visualizations and dashboard, which was later integrated into the project.
Team Member	Sakthi Lakshmi	Handled the entire documentation of the project from Milestone 1 till the final project report. Apart from documentation, she is involved in identifying project goals, creating user stories blocks, in project discussions and aided in visuals and dashboard creation.
Team Member	Gaurav Sharma	Gaurav worked on determining team goals and business objectives and using that information for creating user stories. He also worked on Data Acquisition, Data Preprocessing and helped with creating the Dashboard.

7 PROCESS FLOW

Our project was divided into the four main phases in order to divide responsibilities among the team members and to ease the overall execution of the project. The four phases are:

- Data Acquisition
- Data Preprocessing
- Visual Analytics
- Documentation

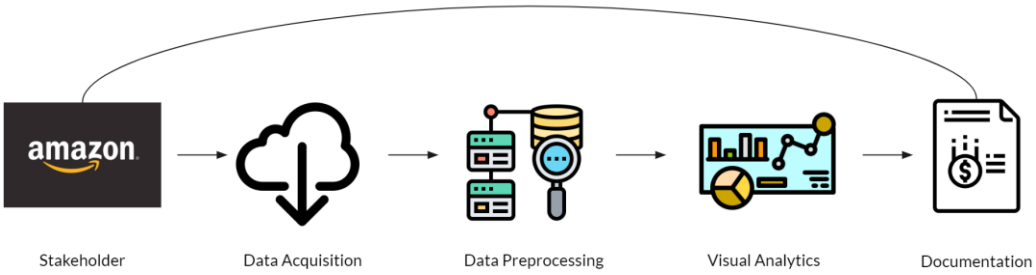

Figure 2: Project Phases/Process Flow.

Data Acquisition:

In the very first step of our project, we acquire the required data from the data sources provided. We obtained the required amazon customer review data made public by Prof. Julian McAuley's [3] research group. After acquiring the data, we examined the data to get an idea of what fields in the data could be used to extract useful insights. The data we acquired was in JSON format comprised of the following fields:

- reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
- asin - ID of the product, e.g. 0000013714
- reviewerName - name of the reviewer
- vote - helpful votes of the review
- style - a dictionary of the product metadata, e.g., "Format" is "Hardcover"
- reviewText - text of the review
- overall - rating of the product
- summary - summary of the review
- unixReviewTime - time of the review (unix time)
- reviewTime - time of the review (raw)
- image - images that users post after they have received the product.

Data Preprocessing:

After acquisition and examination of the data we realized that not all the fields present in the dataset would be necessary to extract useful insights and perform visual analytics. Therefore, we removed the unwanted fields like 'image', 'unixReviewTime', 'style', 'reviewerName', etc. from our dataset and also dropped the records containing null values (NaNs) from our dataset. Originally our dataset had 1,828,971 records but after dropping the null values, the dataset contained around 6,200 records. In addition to the aforementioned tasks, we also performed a Natural Language Processing (NLP) task called Sentiment Analysis to calculate the sentiment polarity of the review summary ('summary' column).

The preprocessing was done using Python and the major python packages used are gzip, pandas, numpy and sentiment analysis was done using the NLTK package available in Python.

Visual Analytics:

In this phase our task was to visually analyze the preprocessed data and to make sense out of it. In order to perform the visual analytics tasks, we made use of one of the leading data visualization and business intelligence tools available in the market today, Tableau. The main reason for choosing Tableau over any other native programming tool is the high quality of graphics Tableau can generate and the ease of creating visualizations. The User Interfaces for the project were also designed in this phase using Tableau's inbuilt features. We created the following charts and dashboards, in addition to enhancing the overall project UI in this phase:

- A Line Chart showing the Average Rating of products over time.
- A Word Cloud showing the Popular words associated with the reviews of products selected.
- A Donut Chart to show the sentiment polarities of reviews.
- A Donut Chart to show the helpfulness of reviews.
- A Year Level Dashboard.
- A Product Level Dashboard.

Documentation:

Documentation is critical for any project. Every step of the project is meticulously documented. Right from Goal of the project, all the assumptions, scope and limitations, user stories, the process involved in the project, steps to create data visualizations, instructions to work on the dynamic visualizations and dashboards, issues faced till the future enhancements, the project details are completely documented both in Systems documentation as well as the Final Project Report.

8 ACCESSING THE PROJECT

This project has been published to Tableau Public's Viz Gallery and is available for anyone to view and interact with at the following link : **Project Start Page - Tableau Public .**

1. On clicking the link to the project, the project start page will open up on Tableau Public and the following screen is displayed:
2. Next, click on the icon inside the highlighted red rectangle in Figure 3 to enter the Full-Screen view. **Note:** if your browser zoom is at 100%, zoom out to 90% or less to ensure the visualizations get rendered properly.
3. After entering the full screen view, click on the "Start Here" (Shown in Figure 3 inside the blue rectangle) button to navigate to the project homepage or the contents page (See Figure 4).

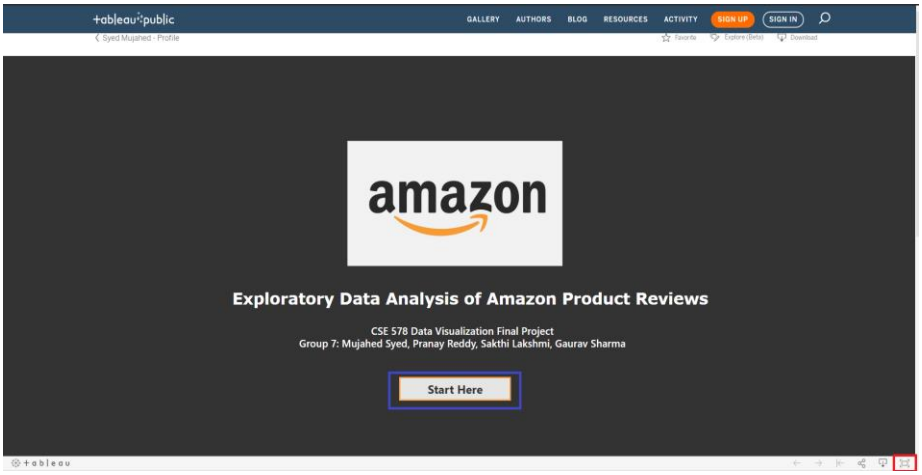

Figure 3: Project Start Page on Tableau Public

4. By clicking on the buttons on the homepage (shown in Figure 4), the users can navigate to their desired visualizations or dashboards.
5. By Clicking on the homepage title ("EDA of Amazon Reviews") the users can navigate back to the Project Start Page.
6. To exit the full-screen mode press "Esc" key on the keyboard.

Figure 4: Project Homepage/Contents Page

9 VISUALIZATIONS AND DASHBOARDS

As shown in Figure 4, all the charts and dashboards can be accessed by clicking on the respective buttons. By clicking on any of the buttons corresponding to individual charts the following views are displayed as shown in Figure 5 and the users can interact with the visualizations with the help of drop down filters in order to filter the visualization to view the data of their interest. To return to the homepage from any of the visualizations, the users can click the "Home" button at the bottom of the page and they will be redirected to the Project Homepage.

Figure 5: Visualizations (Line Chart, Word Cloud, Sentiment and Helpfulness Donut Charts)

In addition to the individual visualizations, the users can access the year level and product level dashboards to view all the charts on a single screen. As shown in Figure 6, the year-level dashboard uses the year as the filter for all the charts and the product-level dashboard uses the "Product ID" of the products to filter the charts on the dashboard. The users can toggle between the project homepage, the year level and the product level dashboard while accessing the dashboard view.

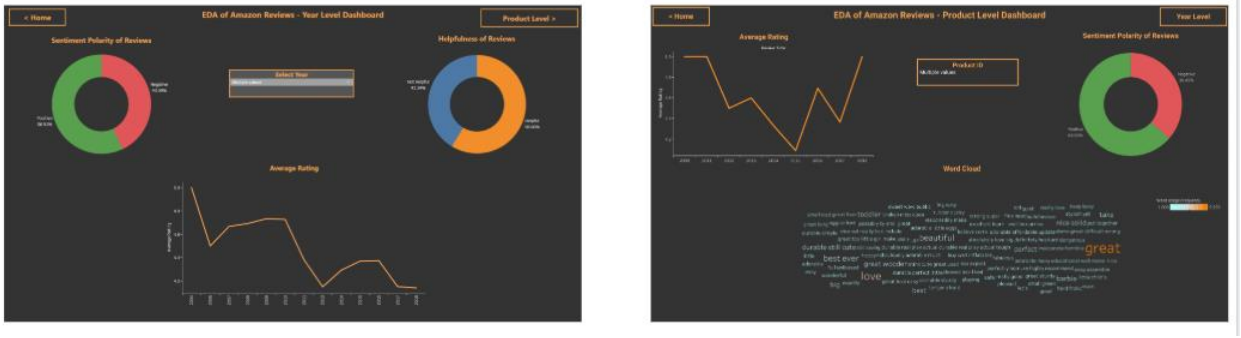

Figure 6: Year and Product Level Dashboards

10 CHALLENGES FACED

1. **Cluttered (On-Screen Clutter) Word Cloud:** Initially, the text in the word cloud was too cluttered and had unnecessary prepositions and articles which contributed to screen clutter. To resolve this issue, we made further modifications to the pre-processed data, such as including only Adjectives, Verbs and Adverbs in the "summary" field. Doing this not only reduced the clutter but also gave much more clear insights on the data.
2. **Choosing Products for Product Filter:** For the Product-level Dashboard, we wanted to include only products having more than 4 reviews in the dataset. In order to do this, we made use of a quick fix in Tableau itself and created a calculated filter that filters our product IDs having less than 4 occurrences in the dataset and includes the rest of the product IDs in the filter.
3. **Choosing a metric to Calculate Helpfulness:** In order to calculate the helpfulness of the reviews, we devised two approaches as discussed in the systems documentation report earlier. However, our first approach didn't work well as it was giving very low helpfulness ratings even for reviews having more helpfulness votes. Therefore, we devised a much simpler second approach and used the average of these ratings in the donut chart.

2 Second Approach

$$H_i\% = \frac{V_i}{\sum_{j=1}^{n} V_j} * 100 \qquad (3)$$

Figure 7: Calculation of Helpfulness Ratings

Where, H_i = Helpfulness rating of reviews corresponding to review "i". V_i = Number of helpful votes for the product's review "i" and ΣV_j = sum of helpful votes of all reviews corresponding to product ID "j".

11 CONCLUSION AND LEARNING

Conclusion: To conclude, the exploratory data analysis EDA on the customer reviews done in the project has resulted in two dashboards, one for product level and year level. The dashboards are sure to provide new customer experience as well as beneficial to other stakeholders[2] like Amazon. The project can be extended to include several other retailers and to add enhancements to suggest a product to the customer based on the data analysis done in this phase.

Learning: This project has been a big learning experience for the team**.** We learnt and applied data visualization principles in making our visualization simple and appealing. Each stage was a learning experience for the team. We got to learn several new techniques to process data and create visualizations in Tableau that are learnt during the course of the project. And moreover, we also came to realize the limitations of visualization that came with the type of data we were using. One such example is the cluttered word cloud. We overcame the challenge by modifying the data and making it useful for our word cloud. Thus, this project taught us how we can take raw data and present it in a way, such that it helps unravel many interesting patterns, such as analyzing the sentiments of the review.

REFERENCES

[1].https://en.wikipedia.org/wiki/Exploratory_data_analysis#:~:text=In%20statistics%2C%20exploratory%20data%20analysis,modeling%20or%20hypothesis%20testing%20tas

[2]. https://www.supinfo.com/articles/single/3023-stakeholders-in-agile-projects

[3]. Jianmo Ni, Jiacheng Li, Julian McAuley, "Justifying recommendations using distantly-labeled reviews and fined-grained aspects", *Empirical Methods in Natural Language Processing (EMNLP)*, 2019

[4]**https://towardsdatascience.com/sentiment-analysis-a-how-to-guide-with-movie-reviews-9ae335e6bcb2**