

**Group Alpha**

Parth Girishkumar Lakdawala

Kumud Chauhan

Priyanka Balaji

Raj Soni

ALY6110 – Data Management & Big Data

CRN: 71205

Final Project Report

**Project Name: Insights on Amazon Products Ratings and Reviews**

Contact:

[lakdawala.p@husky.neu.edu](mailto:lakdawala.p@husky.neu.edu)

[chauhan.ku@husky.neu.edu](mailto:chauhan.ku@husky.neu.edu)

[balaji.p@husky.neu.edu](mailto:balaji.p@husky.neu.edu)

[soni.raj@husky.neu.edu](mailto:soni.raj@husky.neu.edu)

Professor: Valeriy Shevchenko

Table of Contents

[Summary 1](#_Toc22762589)

[Introduction 1](#_Toc22762590)

[Understanding the rating system 1](#_Toc22762591)

[Our Approach and Methodology 1](#_Toc22762592)

[Understanding the data 2](#_Toc22762593)

[Source of Data 2](#_Toc22762594)

[Data Structure 2](#_Toc22762595)

[Data Format 2](#_Toc22762596)

[Data Dictionary 3](#_Toc22762597)

[Business Questions 4](#_Toc22762598)

[Big Data Technology Implementation 5](#_Toc22762599)

[Azure Databrick Configurations 6](#_Toc22762600)

[Analysis 7](#_Toc22762601)

[Basic setup 7](#_Toc22762602)

[Pre-elementary check of data 8](#_Toc22762603)

[Data size 8](#_Toc22762604)

[Null Value 8](#_Toc22762605)

[Exploratory data analysis 8](#_Toc22762606)

[**Metric 1: Ratings** 9](#_Toc22762607)

[**Metric 2: Reviews** 11](#_Toc22762608)

[**Metric 3: Helpfulness votes** 14](#_Toc22762609)

[Conclusion 16](#_Toc22762610)

[References 17](#_Toc22762611)

# Summary

## Introduction

In this project, we have taken Amazon customer reviews and ratings dataset which can provide useful insights about product and the seller. In our view, it is one of the most under analyzed data used in customer behavior analysis. The data consisting of reviews and ratings given by the customers about different products. These reviews and ratings provide useful information about customers psychology such as are user inclined more towards higher rating products, does the length of the review, words used in review headline influences user decision, does low rating and bad review restricts customer to purchase the product etc.

## Understanding the rating system

Amazon provides user a range of 1-5 (lowest-highest) star which can be given to a product. Amazon also insist user to give a brief review on product. Providing review is optional and hence if the user provides a review, then the importance of the latter is high to other users. Each review has a thumbs up and thumbs down icon which is helpfulness rating. This was introduced to tackle fake review provided by paid user in order to boost the sale of product. If a user gives thumbs down, it indicates the review is not useful and if the user gives thumbs up, then it indicates that the review is in fact true and useful. According to the response collected, an average helpfulness rating is calculated for each product. Helpfulness rating to a certain extent has help Amazon tackle the fake review issue to a certain extent. This has given Amazon an edge in the market as they were the first one to introduce this concept in e-commerce market.

## Our Approach and Methodology

In start we have asked few questions pertaining to the concept we were building the project upon. Asking questions is considered to be the fundamental steps while developing any analytics project, it sets the base for hypothesis and modelling. After, the initial brainstorming sessions we defined KPI’s in align with the question developed. KPI metrics provide structure to the analysis and bridges the communication gap between business and technology team. Data cleaning phase was initiated thereafter, where we cleaned the data and involved process such as removing null values etc. More emphasis was given on analyzing data using Big data technologies.

We have used Microsoft Azure Databricks and created our workspace on Azure portal. We used cluster of configuration **6.0 Conda Beta** with **2 nodes** which includes **Apache Spark 2.4.3, scala 2.11**.We used **python** programming language for our analysis and used python API (**pyspark**) and performed analysis using several sql functions. For making our analysis easier we choose to work on dataframe by using **pandas**. For visualization we used python libraries **Matplotlib.pyplot** and **seaborn**. Azure filled the gap in the Infrastructure, Spark made computation of large data easy and feasible while python helped us to interact with spark and analyzing big data and to create visualization of metrics. We have explored the basics and have performed Exploratory Data Analysis given the familiarity with the technology for a shorter timeframe.

# Understanding the data

## Source of Data

The dataset has been taken from the amazon’s official website by web scraping. We chose **multilingual reviews dataset for US market** place only. We have used multilingual reviews dataset which contains the reviews of US marketplace only. The data contains reviews of 20 years from year 1995 to 2015.

Link: <https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt>

## Data Structure

In this project, we have taken dataset which consist of approximately 7 million rows with 15 datapoints or attributes. We used **requests** library to scrap data from the website using API and loaded on our cluster directly.

## Data Format

Tab ('\t') separated text file, without quote or escape characters. First line in each file is header; 1 line corresponds to 1 record. The datafile is a tar file and we extract it on the cluster using bash command. Then we loaded this data into dbfs (databricks file system) to perform our analysis.



Fig1. Rating of a product in Amazon

## 

## Data Dictionary

|  |  |
| --- | --- |
| Name | Description |
| marketplace | 2 letter country code of the marketplace where the review was written |
| customer\_id | random identifier that can be used to aggregate reviews written by a single author |
| review\_id | The unique ID of the review. |
| product\_id | - The unique Product ID the review pertains to |
| product\_parent | random identifier that can be used to aggregate reviews for the same product |
| product\_title | Title of the product |
| product\_category | Broad product category that can be used to group reviews |
| star\_rating | The 1-5 star rating of the review. |
| helpful\_votes | Number of helpful votes |
| total\_votes | Number of total votes the review received. |
| verified\_purchase | The review is on a verified purchase. |
| review\_headline | The title of the review (summary in the image) |
| review\_body | The review text written by customer. |
| review\_date | The date the review was written. |
| vine | Review was written as part of the Vine program. |

Fig 2. Table of data dictionary

# Business Questions

In this section, business team will understand the work for data scientist team and will help them to evaluate the work. It will provide insights to business problem in the organization. These questions were later grouped together to create KPI metrics which serves as a standard across the organization to check the progress on the problem of tackling fake reviews.

* Distribution of rating system: This helps us to understand various type of ratings provided by the customer. If product is not good the rating will be low and if it is good rating will be high on the scale of 1 to 5.
* Total number of reviews by product category: This helps us to understand total number of review provided by user for each category. If the total review of a category is less, it means that the quality provided is high or consumer purchases less product of that category on Amazon.
* Product with maximum review in a category: We have displayed top 10 product with maximum review. This helps us to understand the trend on Amazon US.
* Average rating for each category: This helps us to measure one category with another. This helps us to evaluate the performance of each category compared to other
* Average Helpfulness out of total votes: This helps us to understand the proportion of helpful vote from the total votes of each category.
* Number of reviews over 20 years: We can understand the trend of reviews over 20 years in the platform. Are the reviews increasing or decreasing? In which year we received maximum reviews?
* Most frequent/top words used by the customer in review headline: We want to understand what kind of words customers used to write in reviews.
* Distribution of review length: It helps us to understand that what is length on average written by consumer for a review. How many words customer prefer to write in a review? It answers question like do consumer like short reviews or long reviews etc.

# Big Data Technology Implementation

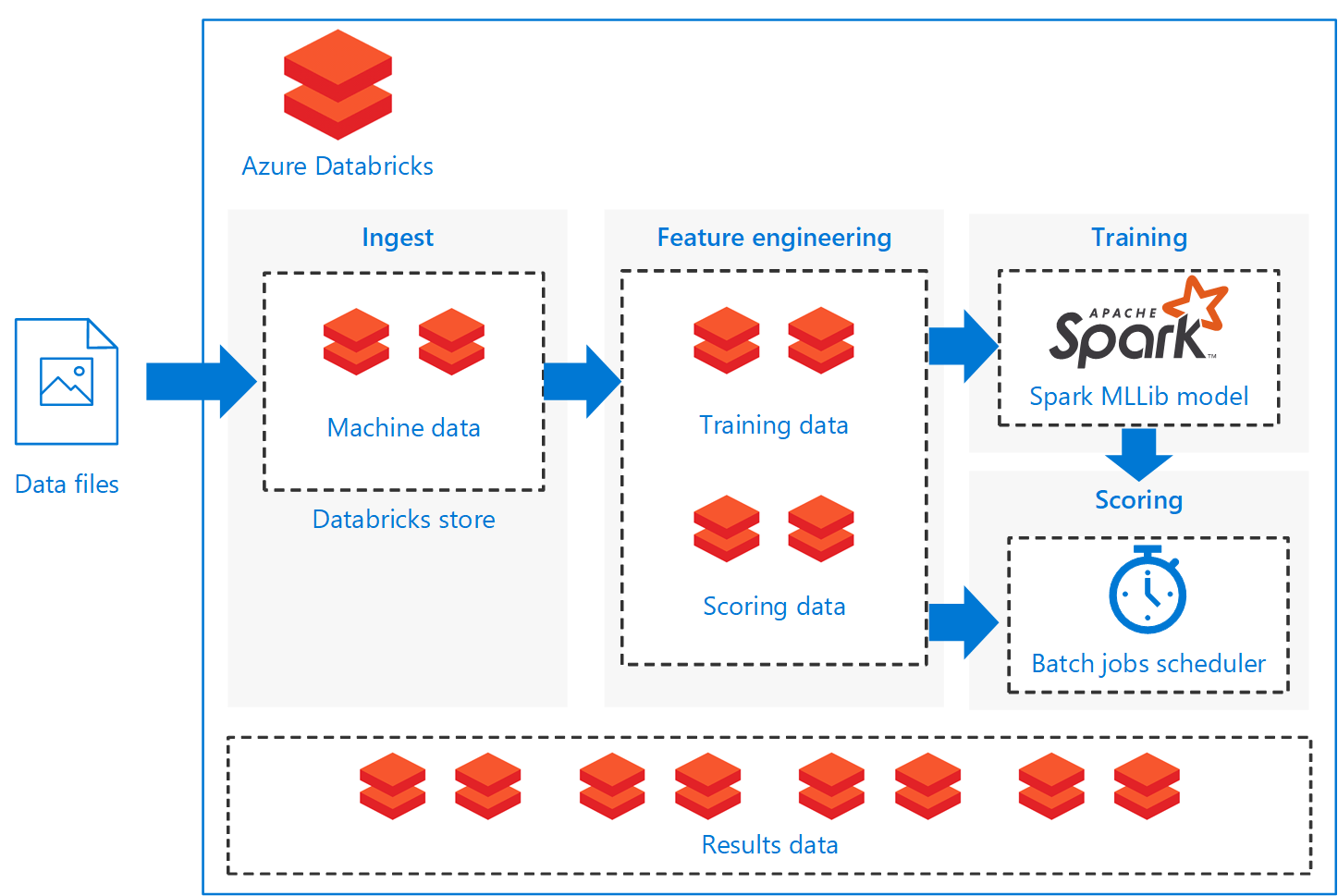


Fig 3. Azure Databricks Spark Architecture

In this project, we have utilized the Azure Databricks which provides seamless integration of Apache Spark, Database and Python Spark API. The deployment is easy and within minutes. As azure is cloud technology, it is ready 24x7. Clusters are auto created and are configurable with our needs.

In Fig3. The architecture of Azure data brick with spark is shown. We upload or download the data file in the Azure data brick. First stage is the ingest stage where the data is ingested to the database and a data frame is created with fields and its corresponding type. The second stage is feature extraction stage where we clean the data and split the data in two type training data and scoring data (or test data). Spark is used in feature engineering and is recommended to use training data. Here, exploratory data analytics and data cleaning take place. Scoring data communicates with Batch job scheduler which is a function of Spark to use its computation advantages. At the end results data is produced. We have used only training data and have communicated with spark to perform our exploratory data analytics.

# Azure Databrick Configurations

We first create an account in Microsoft Azure data brick (<https://azure.microsoft.com/en-us/services/databricks/>) in order to use the service. We have received $200 worth of free credits which can be utilized to use its service for this project.

1. The first step is to create resource. We go to create resource > Analytics > Databricks
2. We provide details such as workspace name, resource name etc to complete the setup

A screenshot of a computer

Description automatically generated

Fig 4. Azure Databrick Configuration setup

1. After creating resource, we can databrick portal launches and we can start spark cluster setup in which our analysis will run. We provide the name of cluster, spark version and python version

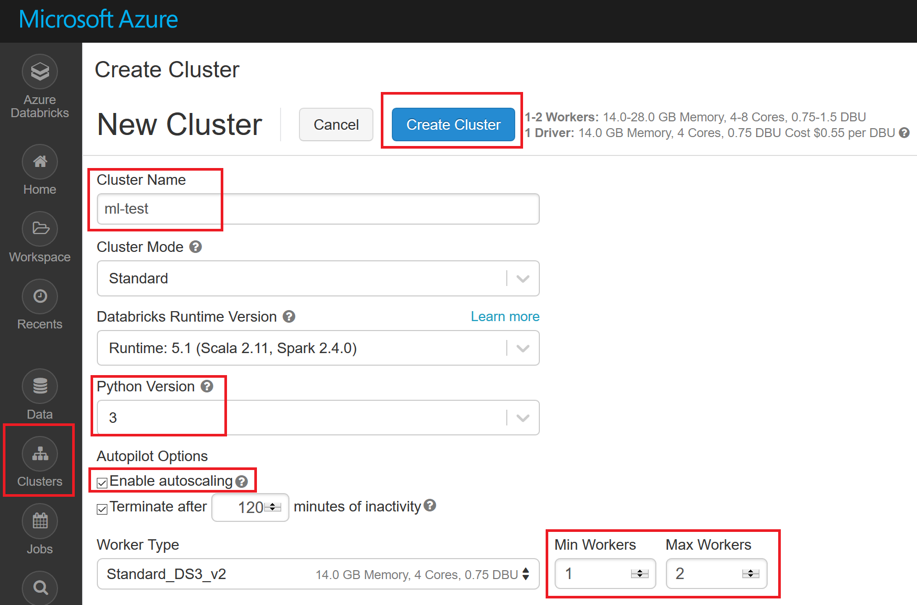


Fig 5. Creating cluster in Azure databricks

1. We can now create a notebook in workspace and start our analysis

# Analysis

## Basic setup

First, we need to import all the necessary libraries of python and spark in order to start with order analysis. The code snippet is attached below.

A screenshot of a social media post

Description automatically generated

Fig 6. Importing Python and Spark Libraries

After importing the libraries, we import our data to data brick and transfer all the data to local database file system of data brick.

A screenshot of a social media post

Description automatically generated

Fig 7. Importing data to Databrick database file system

Once the data is available in the local file system of databrick, we can load the data into spark cluster easily and use the cluster for further analysis.

A screenshot of a cell phone

Description automatically generated

Fig 8. Transferring data from local database file system to spark

## Pre-elementary check of data

### Data size

We check the size of the data by using spark function called count. This will give us total number of records. There are 6.93 million rows.

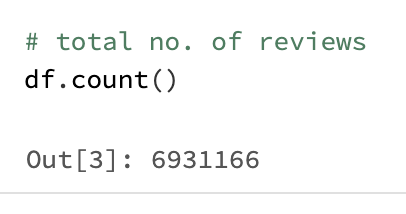


Fig 9. Total number of records using Spark count function

### Null Value

This provide us insights on how many Null values are there in our datasets. There are 16 Null values in our data sets. Spark function isNull, count and when used.

A screenshot of a cell phone

Description automatically generated

Fig 10. Displaying Null values of each columns

### Exploratory data analysis

In this section we will explore different attributes of dataset and analyze them.

* Total number of different products : There are 52380 types of products in our data.



Fig 11

* Total product categories: There are 38 product categories in our data.

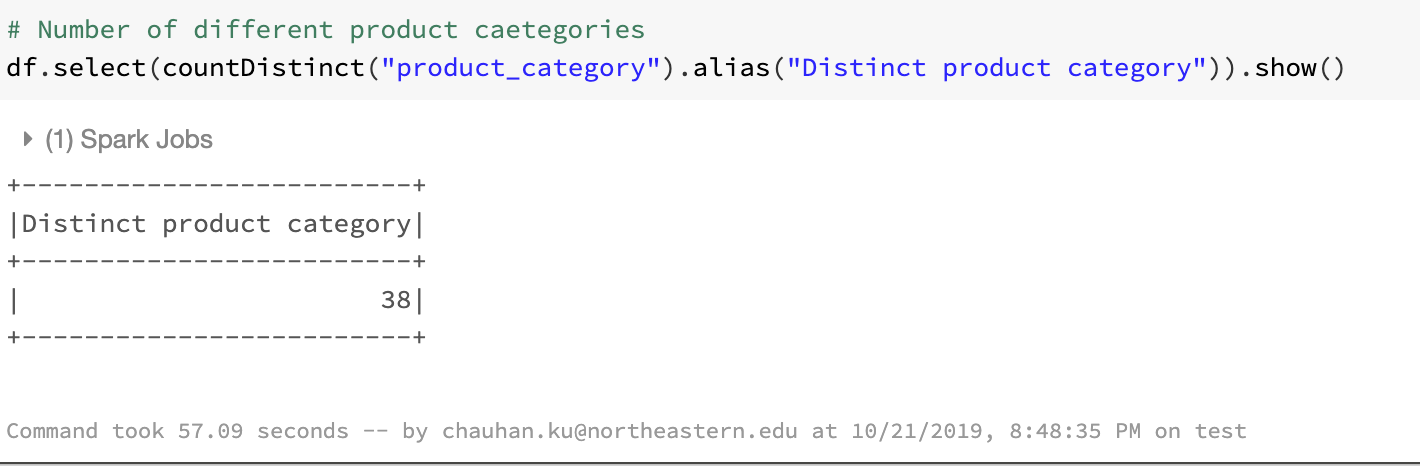


Fig 12

* Total number of customers : 4112395

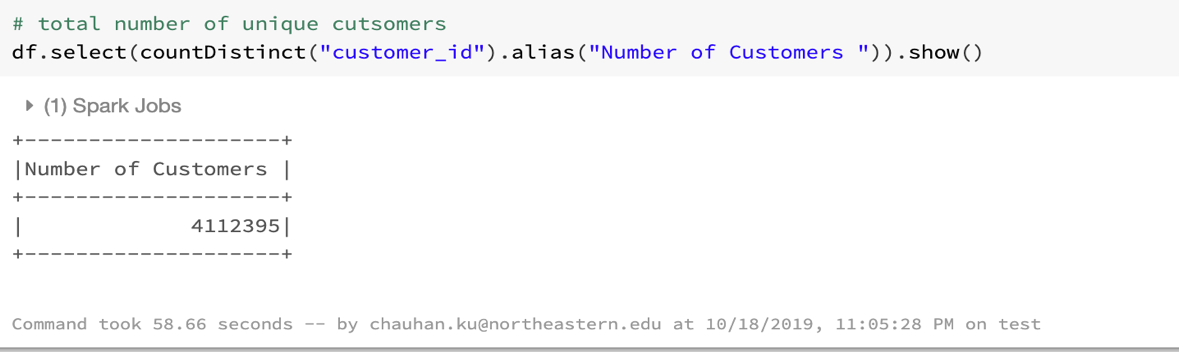


Fig 13

## **Metric 1: Ratings**

In this metric we have tried to explore the rating section of product through basic analysis explained below.

1. **Distribution of product ratings:** It has helped us to understand the product quality and the customer satisfaction. The maximum proportion of 5 star rating is 64% with a count of 4.44 million, followed by 4 star rating which means customers are highly satisfied by the products.

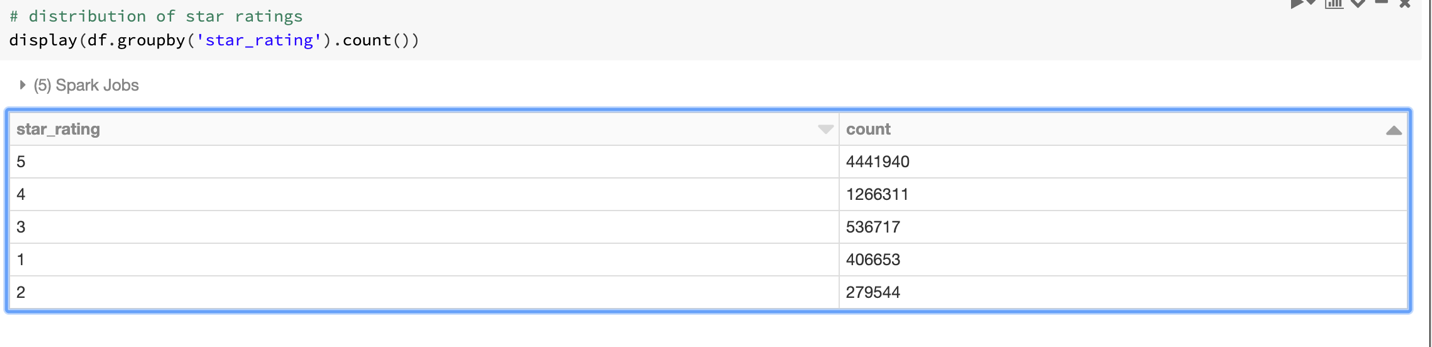


Fig 14

A picture containing screenshot

Description automatically generated

Fig 15. Distribution of rating in Pie chart.

We have used display function of python and count as well order by in spark to the desired result.

1. **Average star rating for category:** We computed the average star rating for all the product categories and sorted them high to low. We observed that “Automotive” products has the highest average rating which is 4.59/5. A close up of a logo

   Description automatically generated

Fig 16. Average star rating of each category in amazon

We also sort the average rating in ascending order and found that Health & personal care rating is the minimum 3.83/5.

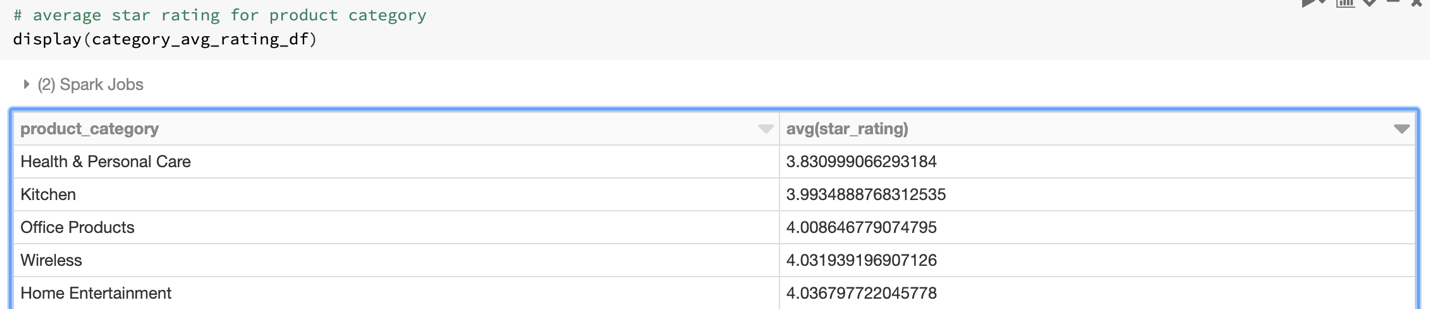


Fig 17

## **Metric 2: Reviews**

This metric helps us to understand total reviews of each product, reviews for each category and most used words in reviews. We have dissected each part of review to analyze its effect on consumer behavior.

1. **Total review in each category:** This provides us insights on how many reviews over past 20 years has been written in each category, which product category is most popular among customers and which is less. We grouped the products category wise and count the reviews. Finally, we sorted them in ascending and descending order.

**Mobile apps** got maximum reviews with a count of over 1 million and **Pet products** got only 5 reviews which is least among all product categories.

These results could be of some highly popular product got more reviews which leads to increased reviews for product category.

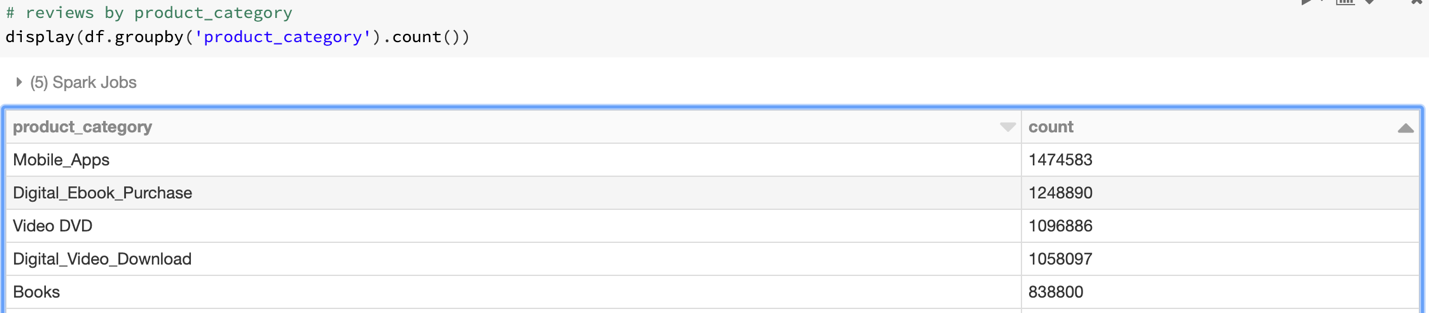


Fig 18. Top 5 most reviewed product category

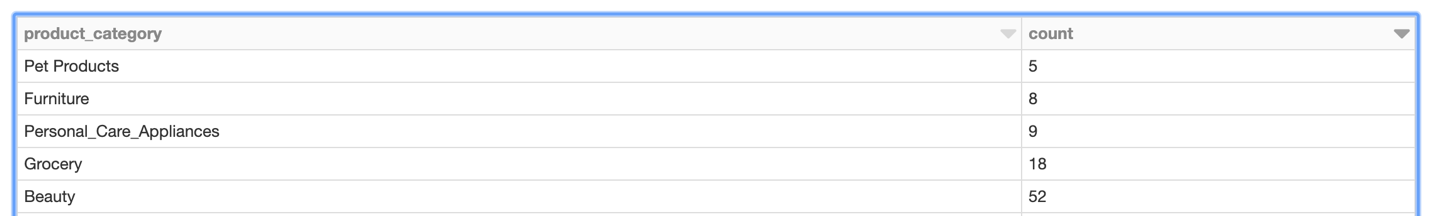


Fig 19. Least reviewed product category

1. **Product with maximum review within a category:** We are interested to know which single product got maximum reviews and to which category does it belong. So we grouped our products with category and product title and sorted top 10 products which received maximum reviews in their category.

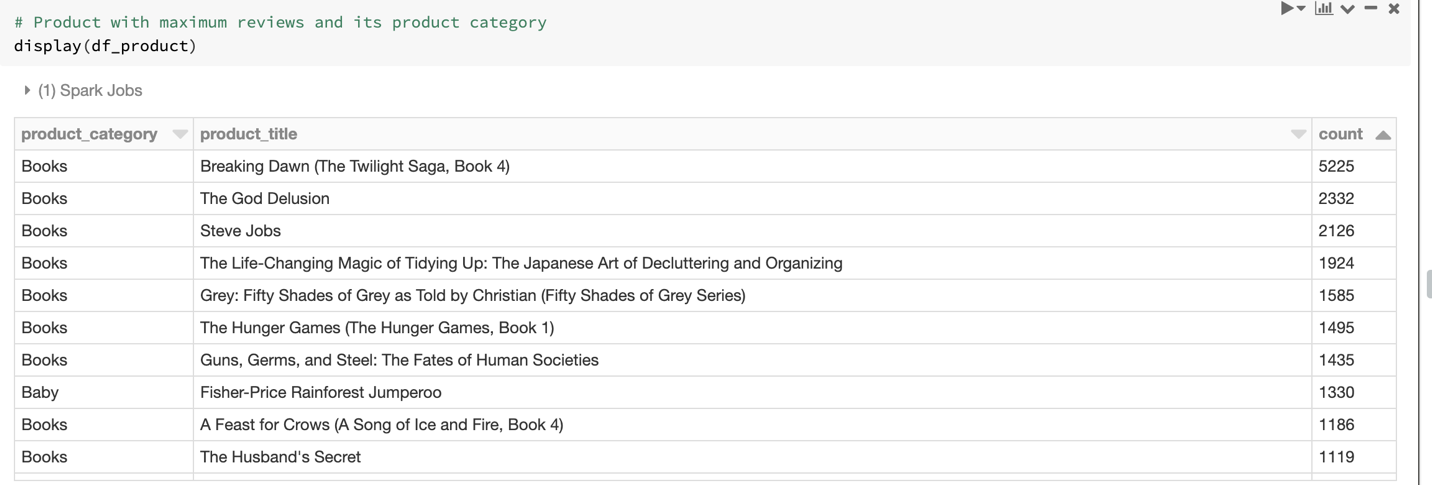


Fig 20. Product with maximum reviews

We observed that **“Breaking Dawn”** book got maximum reviews with a count of 5225, which is highest for a single product. Also, the top 10 most reviewed products belongs to Books category.

It is expected because Amazon started its business with books and in our data we have some new products which may have less reviews than books category.

1. **Growth in Number of reviews over 20 years:** This shows the **exponential trend** on increase in reviews. It also helps us to understand the growing importance of reviews.A screenshot of a cell phone

   Description automatically generated

Fig 21: Total number of reviews in each year

We observed that the number of reviews in starting years were very few compared to succeeding years as the bars are missing for year 1995 to 1997 and very short until 2010. In this bar plot, we observed that for the last 5 years the growth in reviews is exponential. Also, amazon got maximum reviews in year 2014.

Also, the number of products are increasing on amazon’s website over the years and customers are writing reviews for different products which is causing growth of reviews.

1. **Most frequent word used in writing review headline:** This has helped us to understand what words customer used in headline while writing a review. To know the most frequent words to appreciate or criticize the product in its review. To analyze text we perform text mining using NLP techniques and functions like Tokenizer, stopwordsRemover.

First, we tokenize the words in headline which are separated by space and punctuation. Then we remove the stop words which are common in all reviews otherwise words like “a, the, and” etc. would be the most frequent words which comes in almost all reviews. Then we created a vector in which the word and its frequency is stored and we sort these by the count high to low. We observed that the **stars** is the **most frequent** word with a count of approximately 1 million followed by **five.** This proves that customers are satisfied by the products and writing good reviews that’s why the proportion of 5 star rating is high and even the words used in reviews are also five star.

A screenshot of a social media post

Description automatically generated

Fig 22. Most used words in review

We have used tokenizer function of spark to create a unique list of all words and then created a frequency table to plot the above graph.

1. **Distribution of length of reviews:** This helps us to understand customer likeliness of long review or short review or medium size review. It provides insights on customer preferences and how many words customer prefer while writing a review. We tokenized the words from the review text body and created a histogram of bin size 10. We plotted the histogram which shows that most of the reviews contains 20 - 30 words. Approximately 2 million reviews were written by customers in 20-30 words.

A screenshot of a social media post

Description automatically generated

Fig 23. Length of reviews and its frequency

## **Metric 3: Helpfulness votes**

This metric will help us to understand the intensity and impact of each review provided by the user across each product category. This metric is crucial in differentiating between genuine review vs fake review. On the basis of this metric we can blacklist user from the platform.

1. **Average votes for each category:** We have calculated the average of total votes and average helpful vote for each category to understand the which product category got maximum average total votes and maximum average helpful votes.

A screenshot of a cell phone

Description automatically generated

Fig 24. Top 10 helpfulness rating of category in Amazon

1. **Proportion of votes vs helpfulness vote:** This helps us to understand the total proportion of votes given to each category product compared to total helpfulness vote. This helps us to know how many reviews were actually helpful in each category in the platform. If the helpfulness vote is more than the category has product with more genuine reviews then the when compared to other. We chose only top 10 categories to plot the graph.

We observed that the Mobile Electronics category got maximum proportion of helful votes while in our previous analysis (fig 18) the number of count of mobile apps got the maximum number of reviews.

A screenshot of a social media post

Description automatically generated

Fig 25. Proportion of helpful votes (green) vs total vote(blue) for top 10 category

# Conclusion

In this project we had dealt with both the business as well as the technology problems in our approach. We took Amazon review and ratings data to understand its effect. We found out that user having negative experience with a product writes longer review while user having positive experience provides shorter review of the product. The length of review provides us with overview of user experience. The initial stage of project involved understanding the data, identifying problem statement, asking question around problem statement and creating KPI metrics on the question asked which provided us deep insights on the data which might not have been discovered during routine analysis. Reviews and ratings have gained sufficient traction in customer support segments to improve customer support experience in the organization. We can also perform sentiment analysis on the text written by customer in reviews and can build a classifier that classifies the sentiments into positive, negative and neutral categories.

In the technology aspect, we understood the underlying framework of Azure and its services such as databricks which leverages Spark and python to its full potential. Working on such a massive dataset was not an easy task but with Apache spark we got our results in seconds. We could also use a cluster of high configuration and more CPUs to get quicker results. The easiness in deployment with variety of tracking tool to understand the consumption of resources is useful to any organization who plans to develop strong architecture for the company across all the business unit. Apache Spark made our computation and aggregation easy as it took less time to execute and provided us result by using minimum resources. We explored the batch operation used by spark to run parallel operations in Azure while Python acted as intermediate software providing us with capabilities such as visualization and manipulating of dataset in spark.

In this project, we identified three main metrics which were Ratings, Reviews and Helpfulness votes which gave us insights on customer thinking, quality of products in Amazon and impact of reviews in influencing decision making. The current project explores the various exploratory analysis which leverages the qualitative features of data and exploits various aspect of it to identify new arenas for analysis. The future scope will be do create a model which will predict the behavior pattern of a customer and will show product recommendation accordingly. If the consumer decision is not affected by reviews, then the website can show products with good rating which has no or less reviews. The analysis can be seen as a compliment data for recommendation machine learning algorithms and improve its efficiency to a certain extent.

# References

1. Woolf, M. (2017, January 2). Playing with 80 Million Amazon Product Review Ratings Using Apache Spark. Retrieved from <https://minimaxir.com/2017/01/amazon-spark/>.
2. (n.d.). Retrieved from <https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt>.
3. Azure Databricks Quickstart. (n.d.). Retrieved from <https://docs.azuredatabricks.net/getting-started/quick-start.html>.
4. Quick Start. (n.d.). Retrieved from https://spark.apache.org/docs/latest/quick-start.html.