

Group1 |Big Project1 | September 5, 2016

Amazon Fine Food Reviews

Group1- Ajay, augustine, joydeb, saurabh, vaibhav and vinu

# Project Introduction

The Amazon Fine Food Reviews dataset consists of 568,454 food reviews Amazon users left up to October 2012. This data was originally published on SNAP and the team has sourced the dataset from kaggle.com. The dataset contains user reviews covering a host of categories which include beverages, confectionaries, gourmet food, pet food etc.

Our team has picked up this dataset as part of the project requirements of Big Data course from Upx academy. We intend to extend the project to different phases to apply the learnings from Big Data technologies covered during various stages of specialization of the course. In phase1 the team will be exploring ways to apply learnings from Python, HDFS, MapReduce API- MRJob, Pig and Hive, covered as part of the Big Data foundation course.

# About the dataset

This dataset consists of a single CSV file- ***Reviews.csv***, and a corresponding SQLite table named ***database. SQLite***. The columns in the table are:

* Id
* Product ID - unique identifier for the product
* Use rid - unique identifier for the user
* Profile Name
* Helpfulness Numerator - number of users who found the review helpful
* Helpfulness Denominator - number of users who indicated whether they found the review helpful
* Score - rating between 1 and 5
* Time - timestamp for the review
* Summary - brief summary of the review
* Text - text of the review

|  |  |  |
| --- | --- | --- |
|  |  |  |
| File Name | Format | Size |
| ***Reviews.csv*** | CSV | 286MB |
| ***database. SQLite*** | SQLITE | 355MB |

# Business Questions Identified

The following questions were identified by the team after many rounds of inspection and discussions: **Take a look again TreeMap concept - Use SortComparable class & ALSO write numbers of hours require ( Length of code and logic development & Testing )**

**Output should come in the CSV Format**

1. Which are the top 10 Product IDs that get reviewed the most?
2. Which are the top 10 most favorably reviewed Product IDs? Average product ratings
3. How may product reviews are generated on a daily, monthly and yearly basis?
4. Product sales- festival mapping. By festival mapping we are trying to analyze the top 3Product IDs that get reviewed the most for each of the major festivals in the USA. --- MapSide Join for flexibility --- festival as a Key – Look up table to be created for festival table and month
5. Which are the most top 3 products reviewed by month. --- Need Improvement

🡪 Number of counts for Average RANK of each month.

6. By seasonal mapping we are trying to analyze top products reviewed during seasons like summer, autumn, spring and winter in the host country. --- Try to Make it by the Date

7 - i Customer profiling by identifying top contributors to product reviews measured using metrics like frequency of visits per month as well as attempting to gauge customer sentiments using metrics like #of positive and #of negative reviews by the top reviewers.

7- ii Customer profiling based on the monthly reviews.

* what does the product-reviewer graph look like?

Sentimental Analysis --- Take total positive and take total negative.

Formula - ( positive – negative)/ (positive + negative)\*100

**Average + Sentimental**

8. Find out the unique number of reviewers.

8:ii – Find out the number of reviews by each user

9. List out the product id reviews by the customer and total count of the product.

10: data visualization using Excel programming. – 10 to 15 hrs

11: Find the product which are getting ratings 1 or 2 and display the graphics.( Worst product) --- PIE Chart for 1 rating ,2 rating,3 rating , 4 ratings & 5 ratings

11: - (!!) Categorized product based on the ratings and display in the PIE chart.

12: Total Number of product ---- then create report which fall under different ratings – PIE Chart

Out of 100. 20 pro got 5 ratings , 40 got 4 ratings etc.

13: Each Product reviewed by 10 people gave rating five.

Same product reviewed by 8 people gave rating 2. You can derive superhit/hit/average product in the market.

--- Average rating of each product

--🡪 14: 1 -- TOP 50 reviews which helps maximum number of people to buying the product.

* Ii – Top 50 reviews who indicated whether they found the review helpful

15: Create word cloud of use sentiment.

Take maximum 200 words to making word cloud.

15: ii – Find out the sentiments on Dog’s Food product.

16: List out the worst product using parameter rating =1, HelpfulnessNumerator, HelpfulnessDenominator where zero. Output should contain prod id and reviews.

17: Find out the product id and their reviews whose rating is 5 still its HelpfulnessNumerator & HelpfulnessDenominator are null.

**16 )** [**What is the most popular food product**](https://www.kaggle.com/jenskitchen/d/snap/amazon-fine-food-reviews/what-is-the-most-popular-food-product)**.**

**17:** [Helpfulness of Frequent Reviewers](https://www.kaggle.com/jasontam/d/snap/amazon-fine-food-reviews/helpful-reviewers)

## 18: Distribution of ratings

19: What makes the Food delicious.

20:

<https://github.com/tonyandrys/cs-4980-big-data-technologies-andrys>

Use Case 1.1: Find out the unique number of reviewers.

Use Case 1.2 Find out the number of reviews given by each user.

Use Case 3: Which are the top 10 Product IDs that get reviewed the most?

Use Case 4: Which are the top 10 most favorably reviewed Product IDs?

Use Case 5: How may product reviews are generated on a daily, monthly and yearly basis?

Use Case 7: Which are the most top 3 products reviewed by month.

Use Case 8: By seasonal mapping we are trying to analyze top products reviewed during seasons like summer, autumn, spring and winter in the host country.

Use Case 9: Customer profiling by identifying top contributors to product reviews measured using metrics like customer sentiments using metrics like #of positive and #of negative reviews by the top reviewers.