

**CS4818 DATA SCIENCE**

**Project  
based on**

**AMAZON FINE FOOD REVIEW ANALYSIS**

**By:**

**Abdul Wasay 1812254**

**Submitted To: Dr. Imran Amin**

**Date: 11th January, 2022**

**What Our Dataset Is About?**

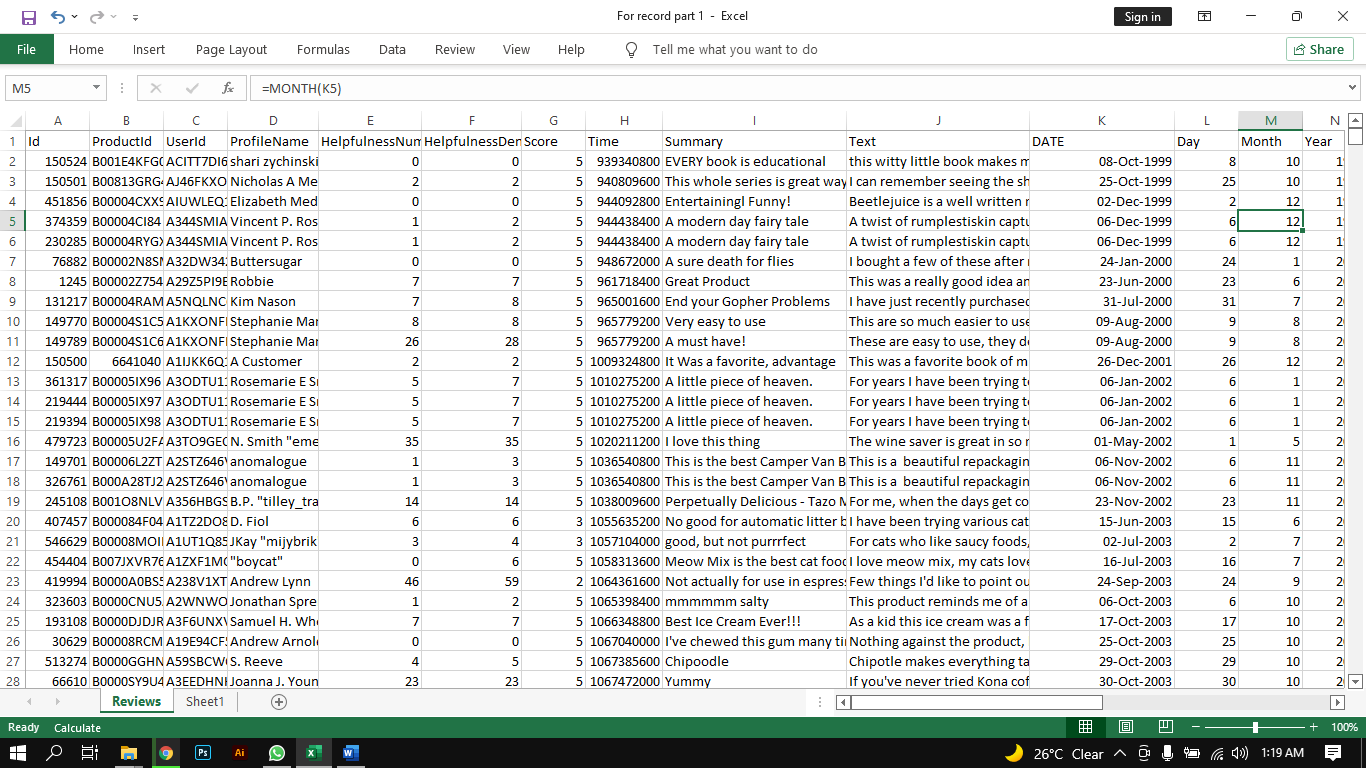
Our dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all 568,454 reviews up to October 2012. Reviews include product and user information, ratings, and a plain text review. It also includes reviews from all other Amazon categories.

**Contents**

* Reviews.csv: Pulled from the corresponding SQLite table named Reviews in database.sqlite
* database.sqlite: Contains the table 'Reviews'

Data includes:

* Reviews from Oct 1999 - Oct 2012
* 568,454 reviews
* 256,059 users
* 74,258 products
* 260 users with > 50 reviews

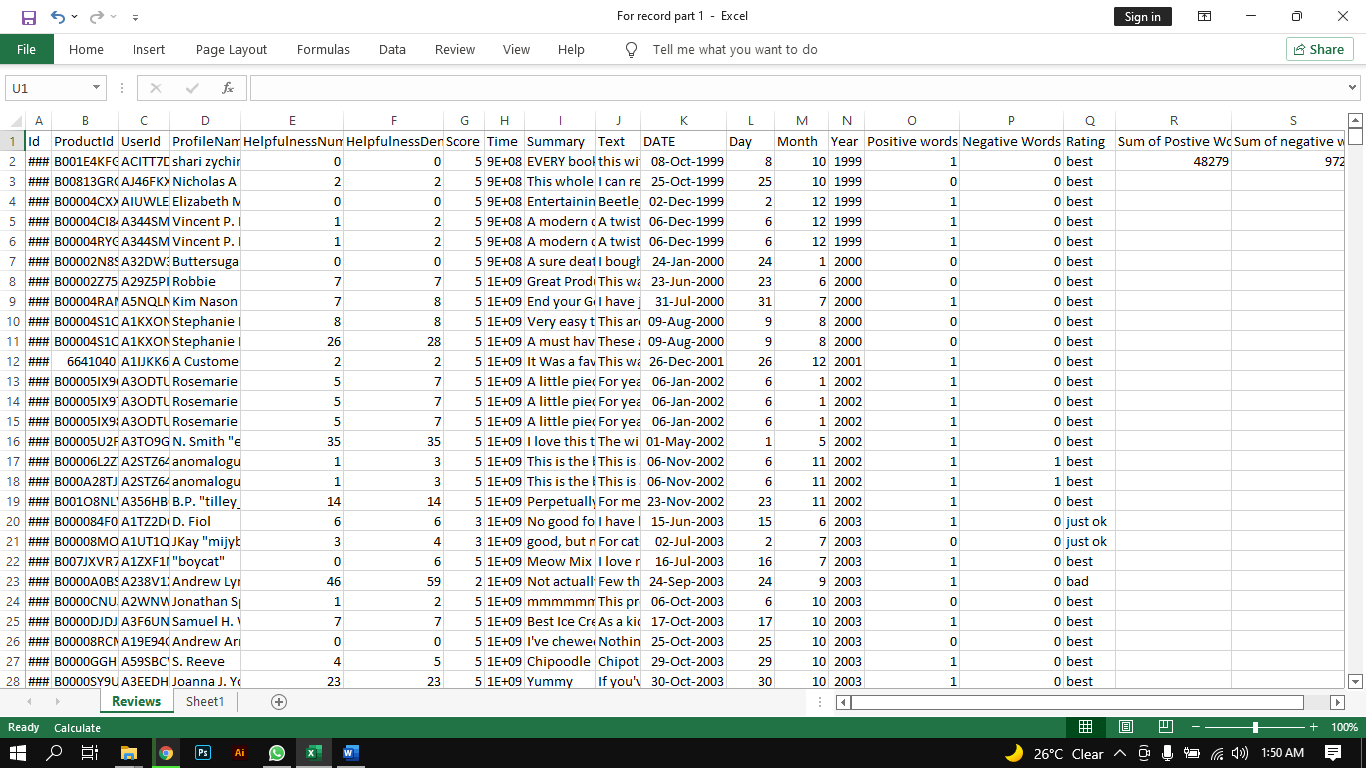


**Glossary**

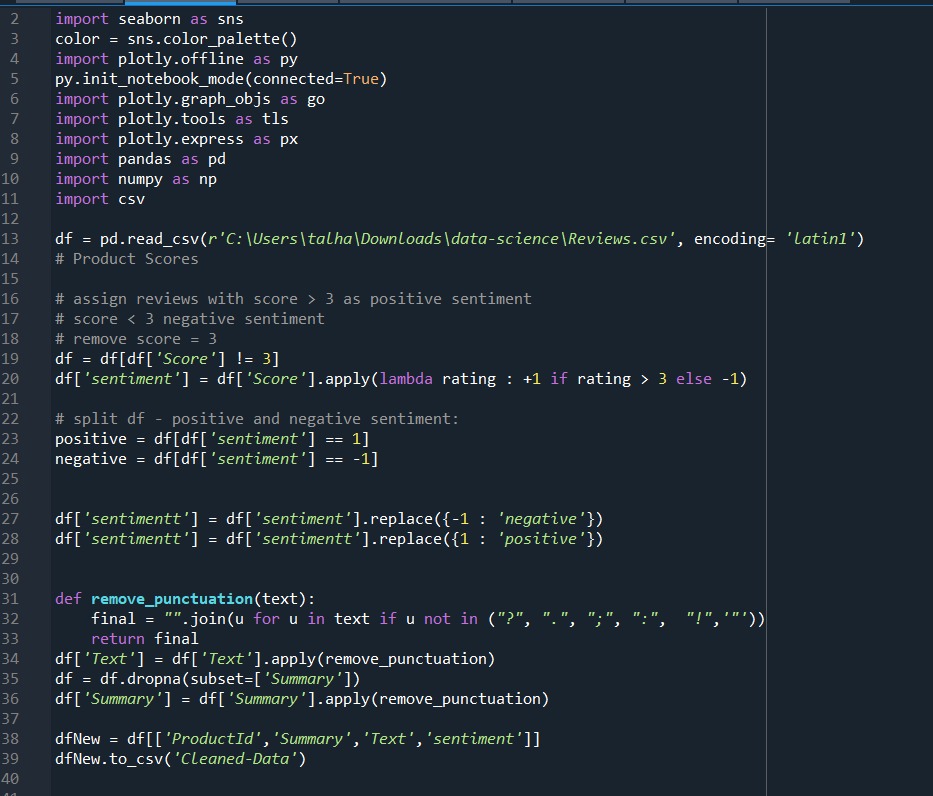
**ProductId:** Product ID **UserId:** User ID **ProfileName:** User Profile Name **HelpfulnessNumerator:** number of users who found the review helpful **HelpfulnessDenominator:** number of users who indicated whether they found the review helpful or not **Score:** The product rating provided by the customer between 1 and 5 **Time:** Time of the Review **Summary:** This is a summary of the entire review **Text:** This variable contains the complete product review information

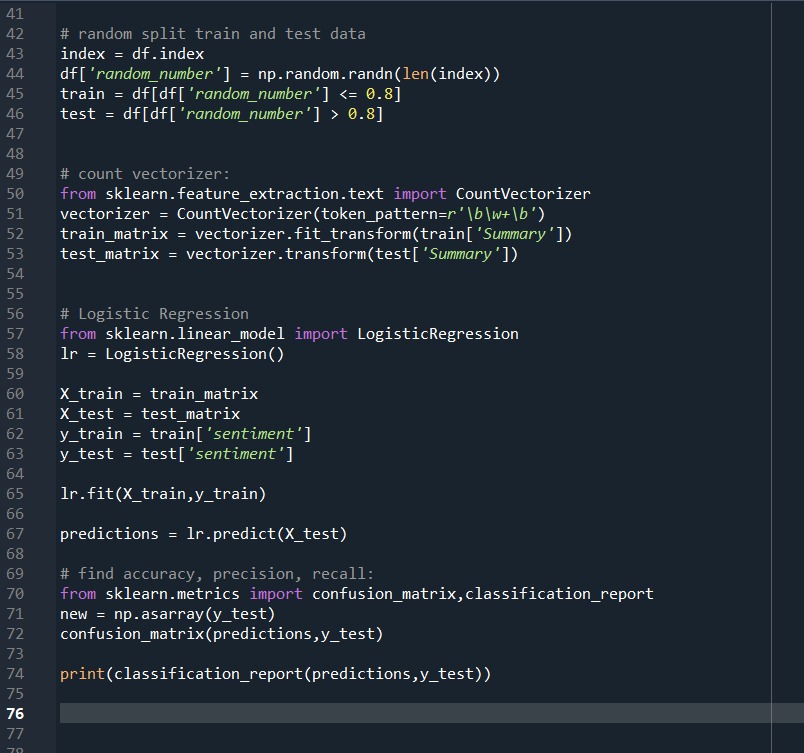
**What We Did With The Dataset?**

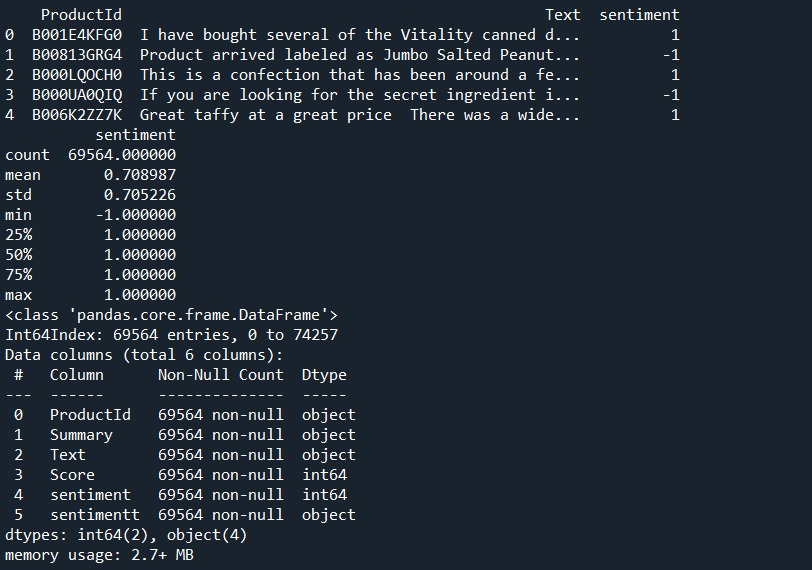
Firstly, we checked if there are any missing values in the dataset so that we can handle them accordingly.  
Then we cleaned the data, we removed punctuations which were in between the words for our own future help according to project scope and needs. Time was given in UNIX format, we applied formula **“=(((H3/60)/60)/24)+DATE(1970,1,1)”** in excel to calculate equivalent user human readable time.  
Then we categorized positive and negative reviews using ‘text’. Formula used for this purpose is **“IF(OR(ISNUMBER(SEARCH("WORD",COLUMN))”**.

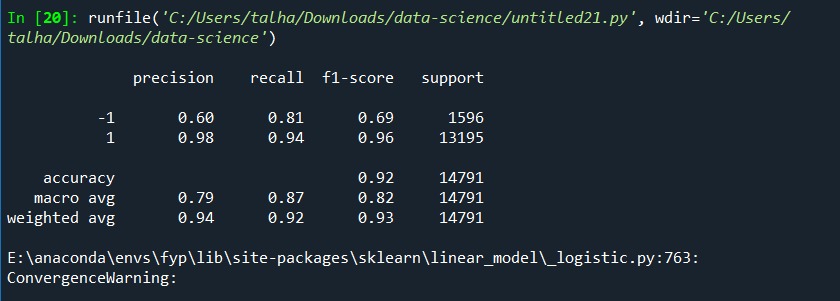
  
*Fig 1.1: Screenshot of Dataset after processing*

Sentiment analysis using ‘score’ from dataset. Scores are given out of 5, so we made 3 as neutral, less than 3 as negative sentiment and greater than 3 as positive ones. We used python language to split into positive and negative sentiments which was represented by +1 and -1. Then we passed our ‘text’ to count vectorization for training and testing. After that we applied logistic regression on the determined sentiments (+1 & -1) for prediction purpose. We calculated precession and accuracy.

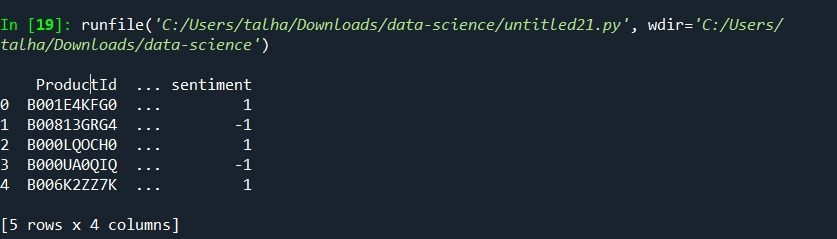
  
*Fig 1.2: Screenshot of sentiment determination in python*

  
*Fig 1.3: Training and Classification in Python*

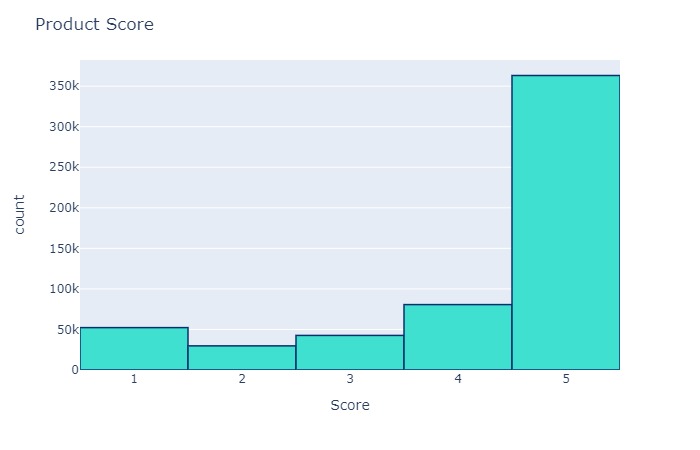
  
*Fig 1.4: Dataset details after cleaning and sentiment determination*



*Fig 1.5: Final accuracy test of code*

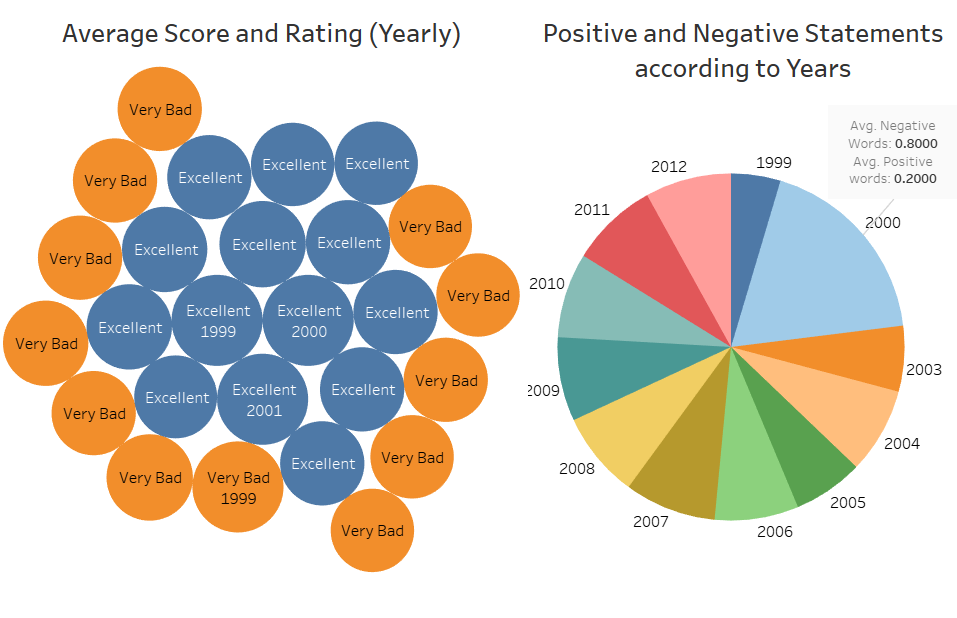
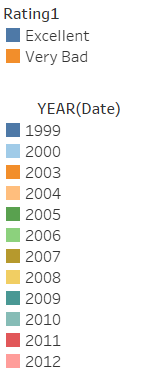


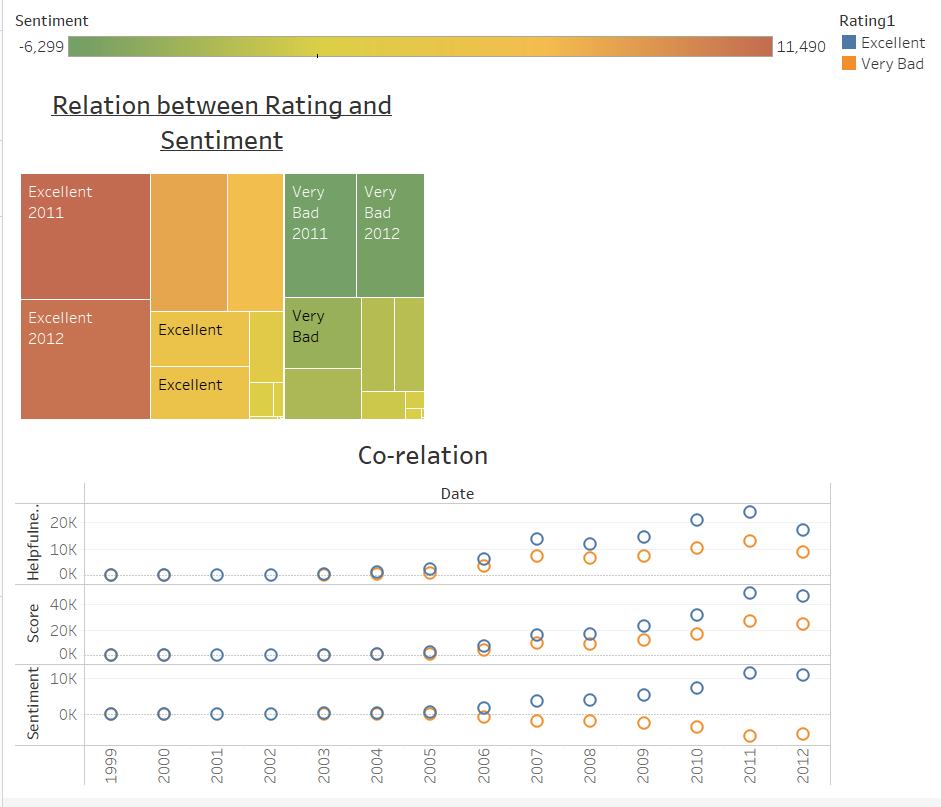
*Fig 1.6: Sentiment according to product as determined by algorithm in Python*

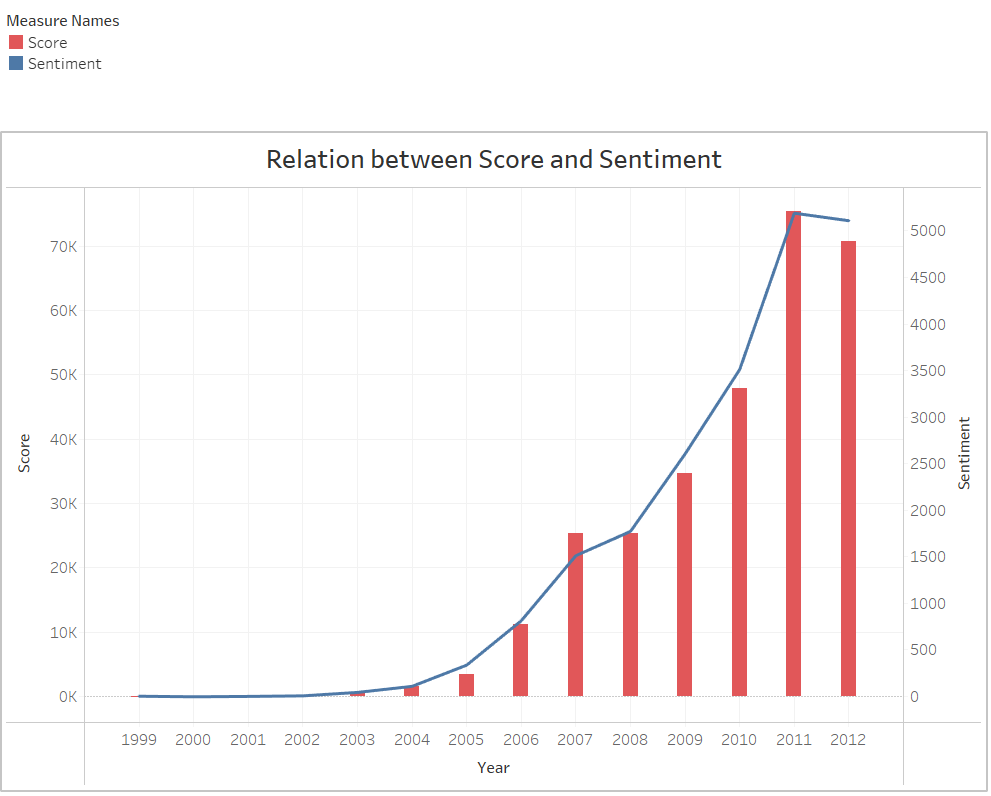


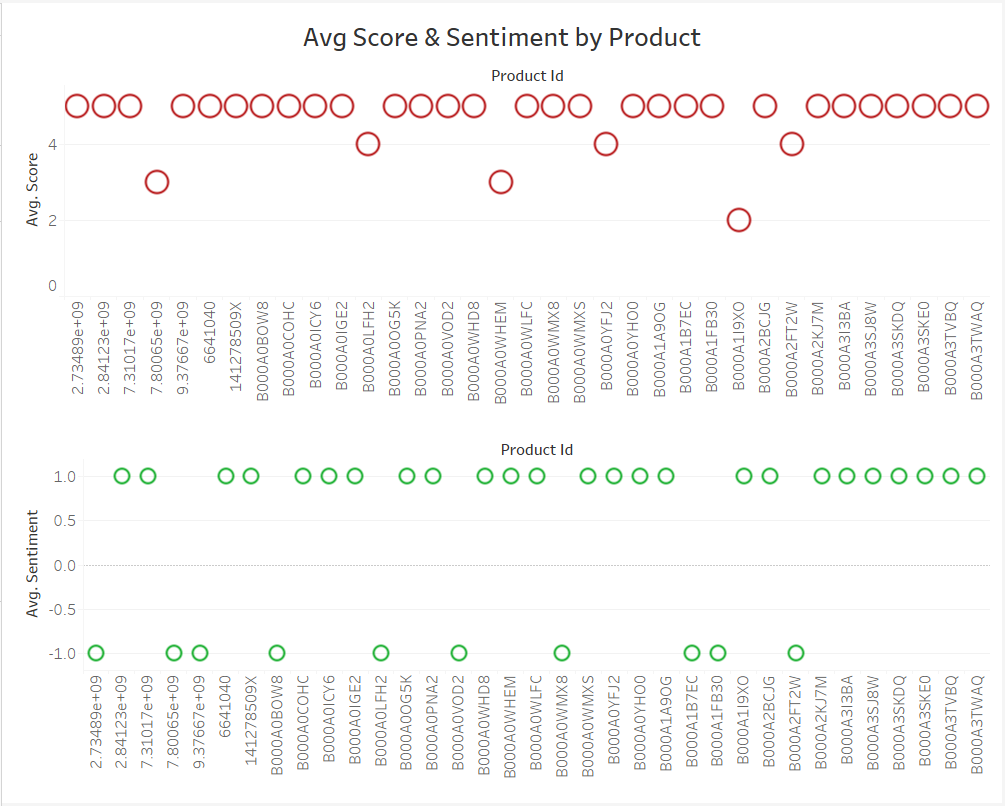
*Fig 1.7: Bar chart of Score against Count*

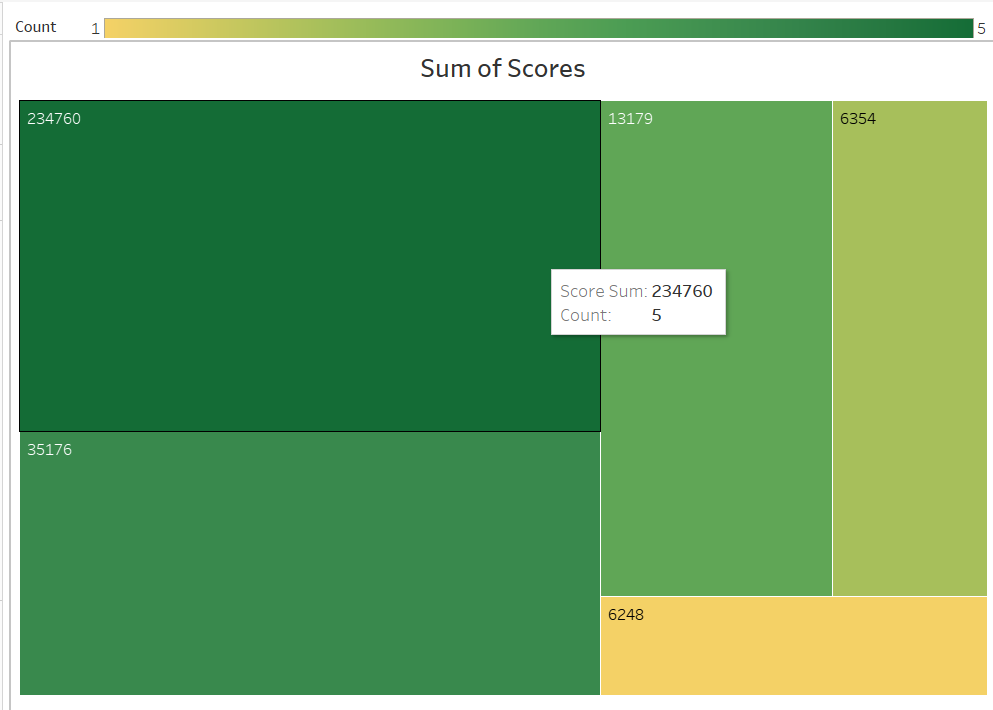
**Dashboard Visualization**

**** ***Fig 2.1: Average Score Rating (yearly) and Positive & Negative Statements according to Years*

*****Fig 2.2: This graph correlation between rating & sentiment and sentiment, score and helpfulness according to year.*

*****Fig 2.3: Graph shows correlation between score and sentiment with time.*

*****Fig 2.4: This shows average score and sentiment of each product determined by ProductId.*

*****Fig 2.5: Sum of the count of all the scores from 1 to 5 according to the rating.*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_