**INFO 7250 Engineering Big-Data Systems**

**Final Project Report on**

**‘Amazon Product Review’**

**Dataset**

SUMANTH HAGALAVADI GOPALAKRISHNA

Northeastern University

MS-IS

**NUID**: 001824723

**SUMMARY:**

This project is based on the Analysis of the ‘Amazon Product Review’ dataset. The dataset

consists of reviews of products from Amazon. The data span a period of more than 10 years,

including all ~500,000 reviews up to 2015. Reviews include product id and user information,

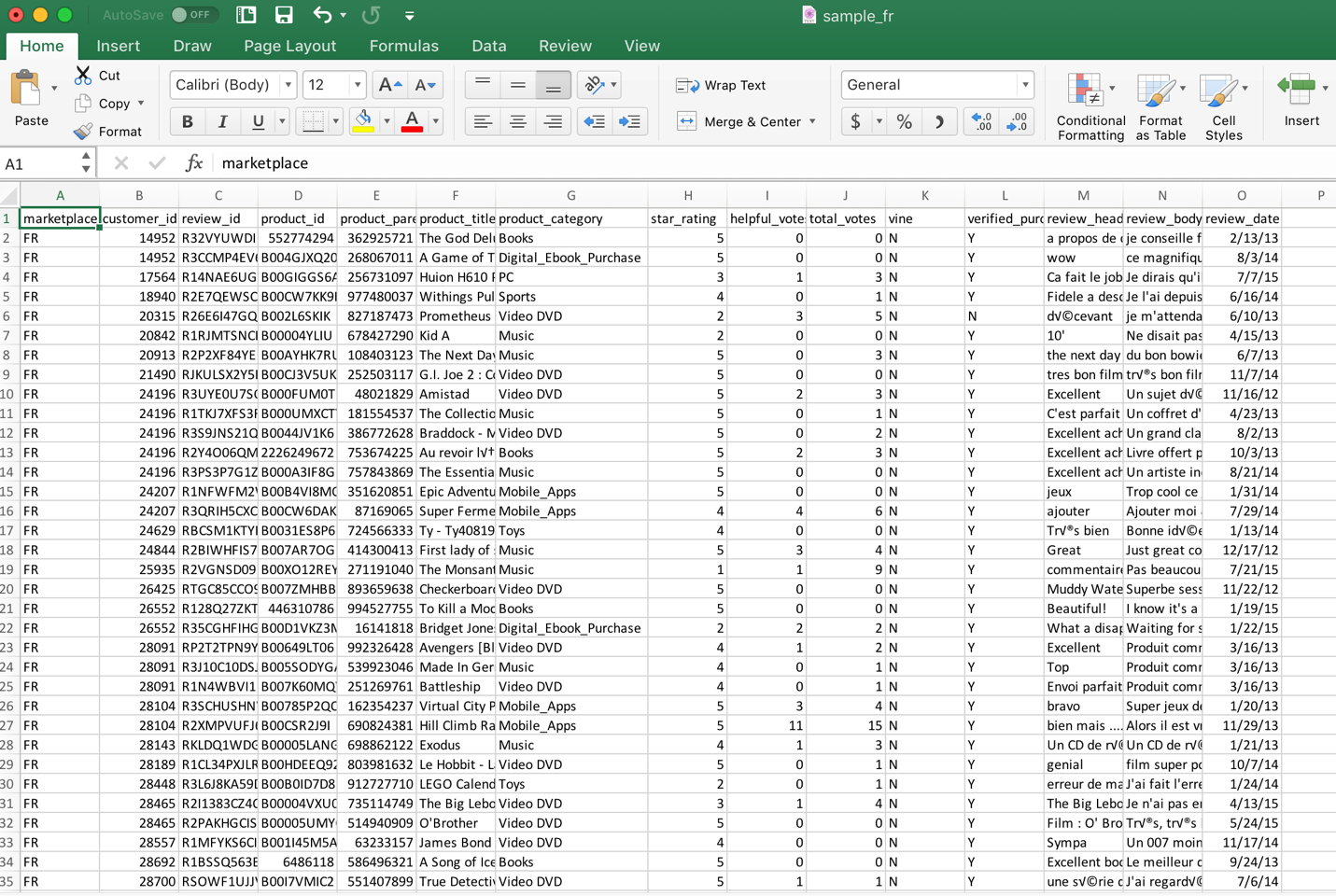
and product ratings.

Map reduce analysis are done using Hadoop, Mongo dB and Pig. All the statistics are shown using Spring MVC application.

**DATASET LINK:**

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

**SAMPLE PRODUCT REVIEW DATASET:**

****

**DATA CLEANING AND FILTERING:**

Since the combined dataset was very huge, Random Filtering is used to get random data from the dataset.

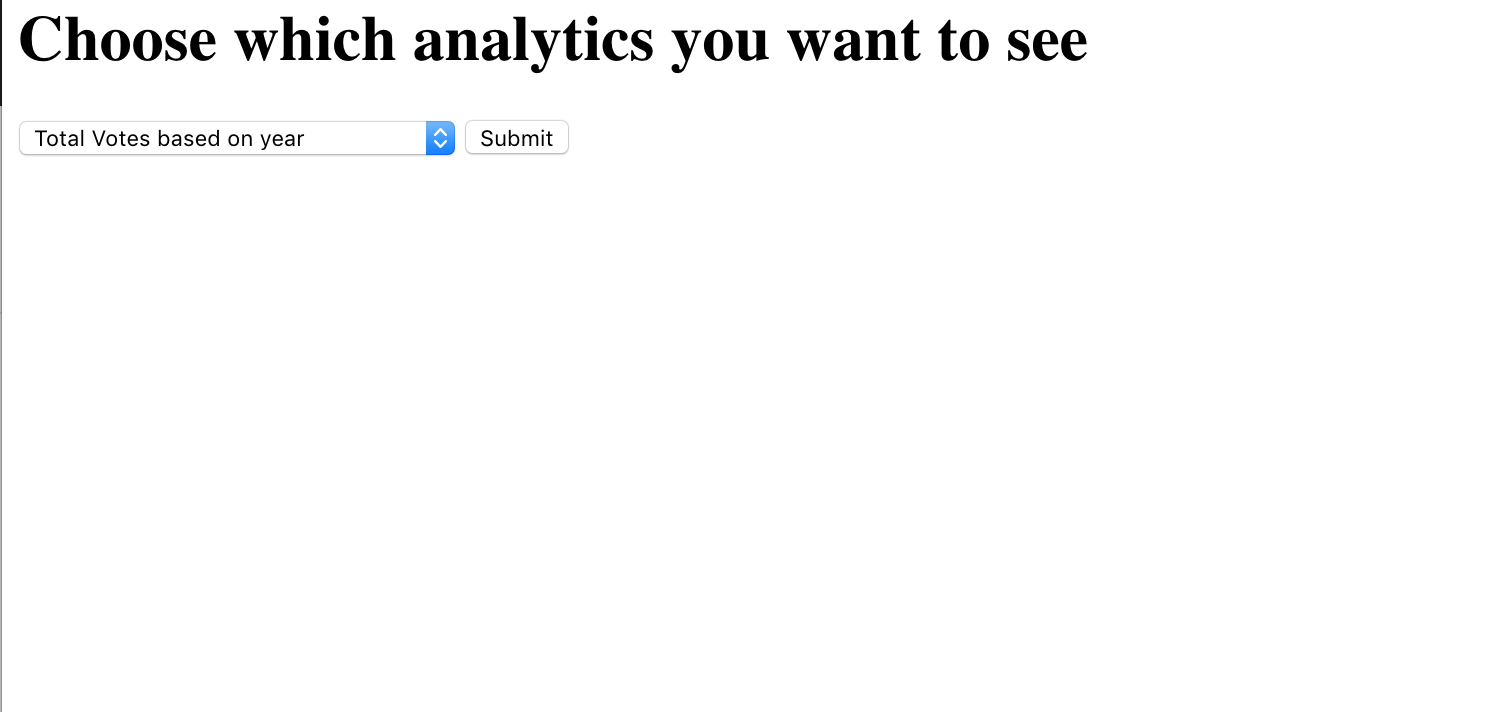
**DATA STATISTICS AND ANALYSIS:**

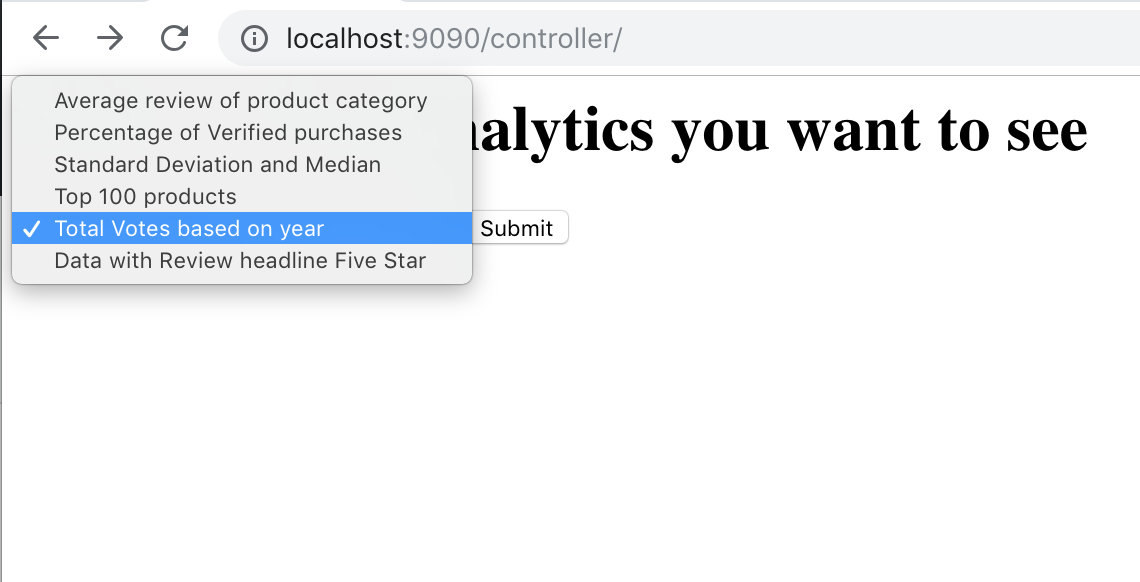
Following are the analysis done using MapReduce:

* Average review of the product based on product category
  + Mongo DB
* Percentage of verified purchases for each product
  + Hadoop
* Standard deviation and median based on Rating for the products
  + Hadoop
* Top 100 products based on Total votes
  + Hadoop
* Total votes based on year
  + Hadoop
* Filtering data for Review Headline which has “Five Stars”
  + Pig

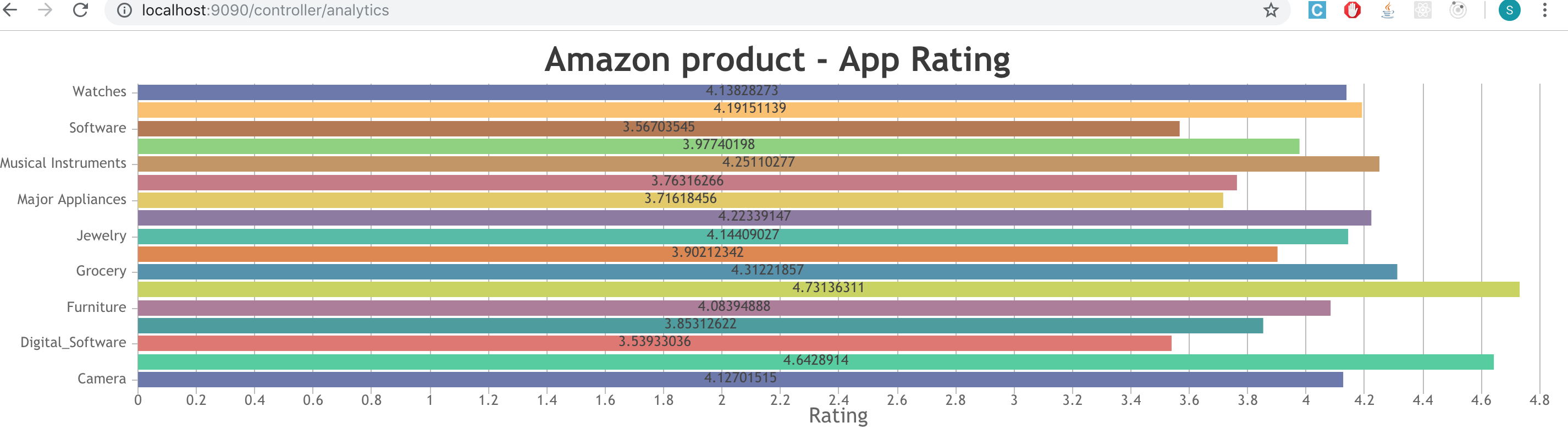
**SPRING MVC WEB APPLICATION:**

Spring MVC is used to show the statistics for each analysis which has been done using different Map Reduce jobs.

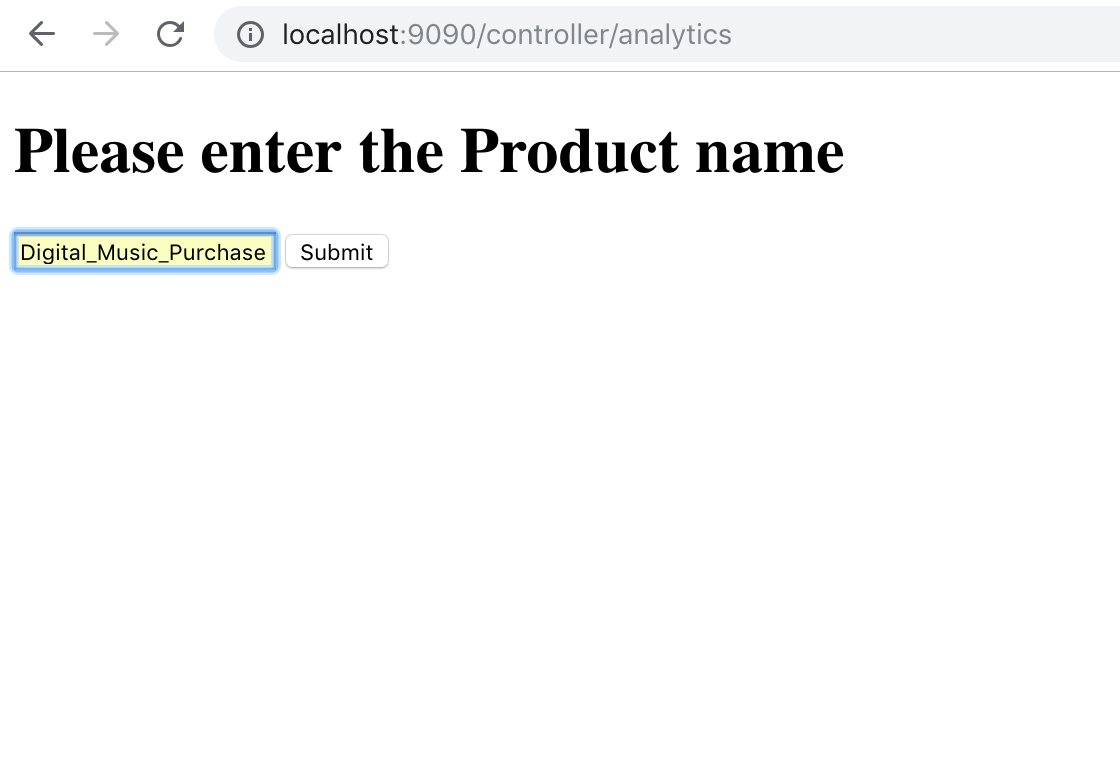


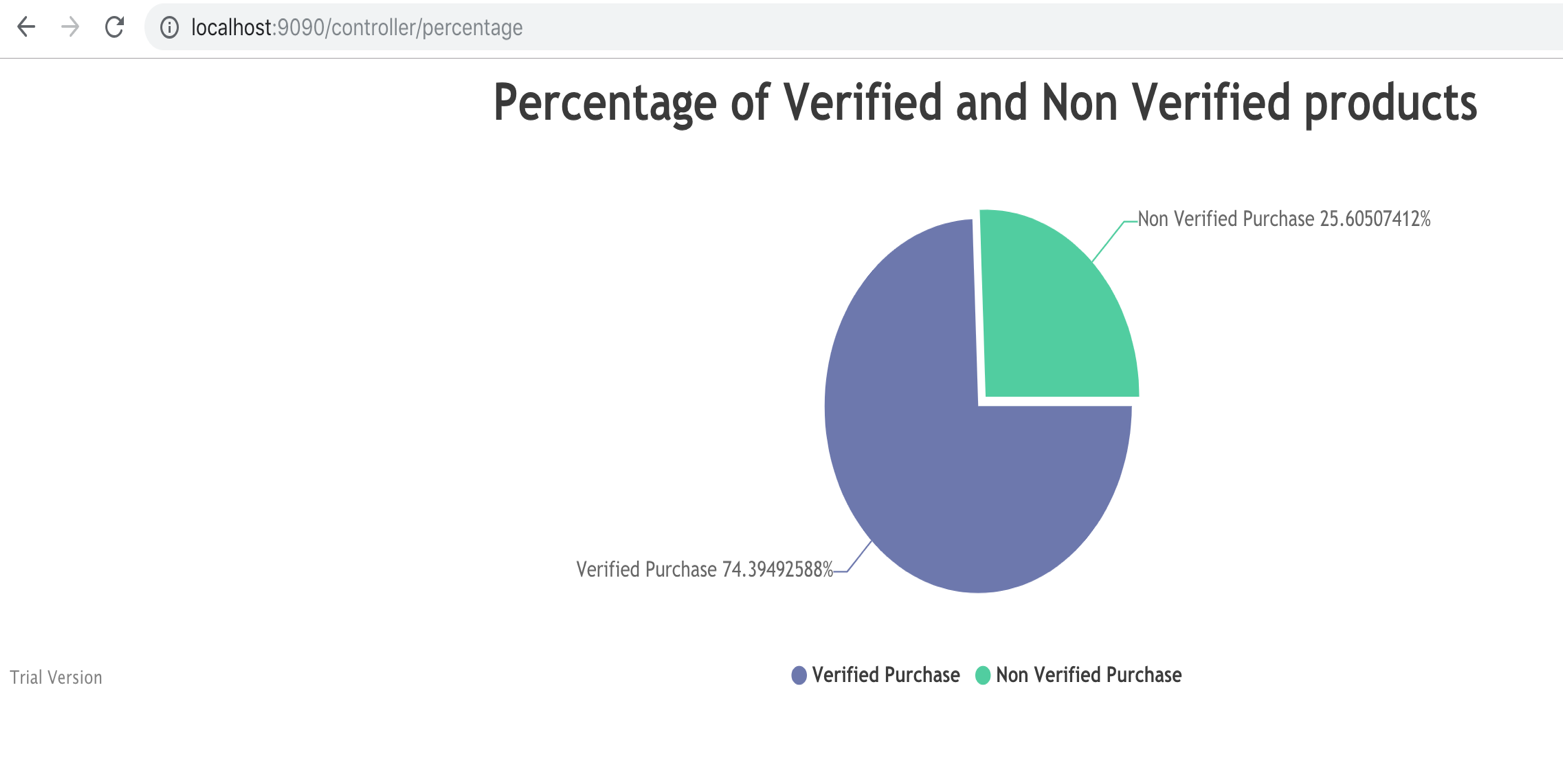


1. **AVERAGE REVIEW OF THE PRODUCT BASED ON PRODUCT CATEGORY**

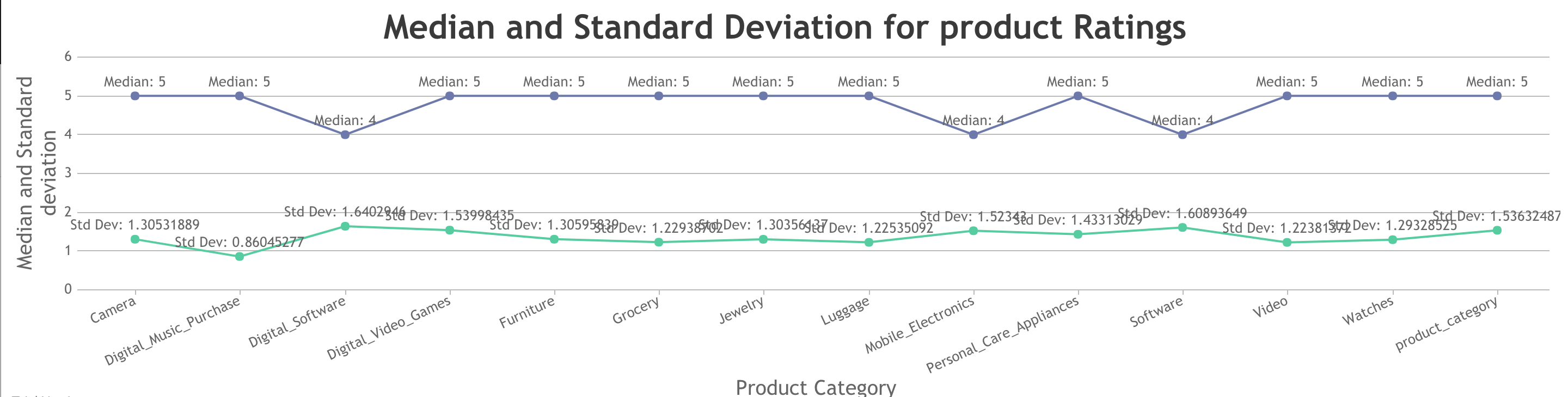
****

1. **PERCENTAGE OF VERIFIED PURCHASES FOR EACH PRODUCT**

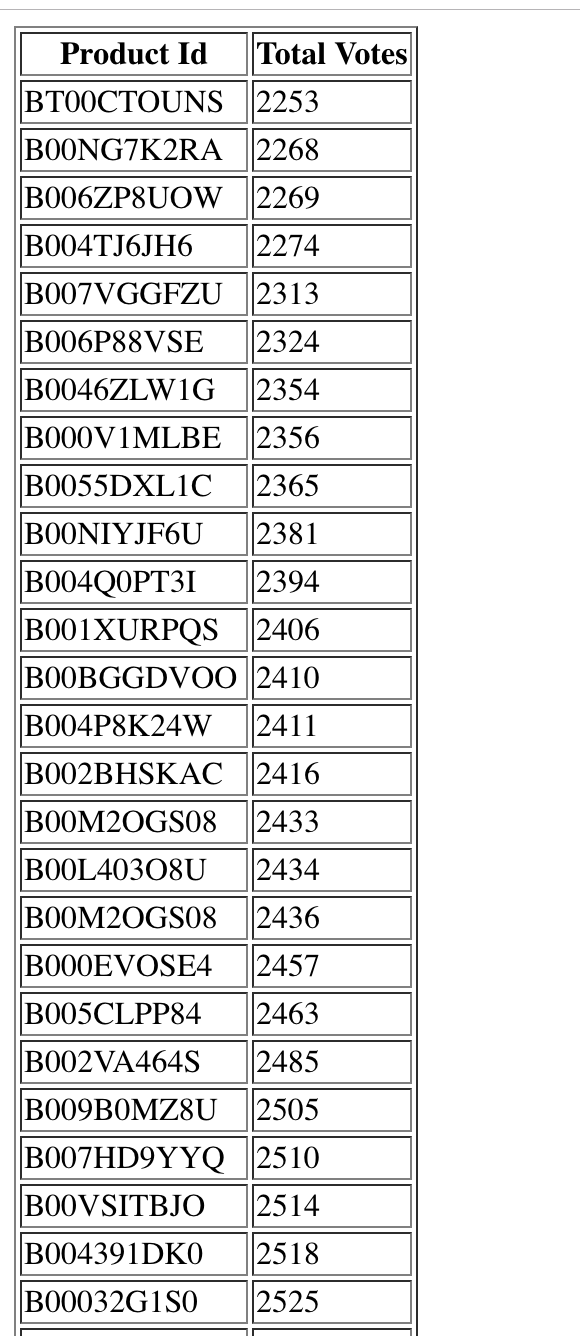
****

****

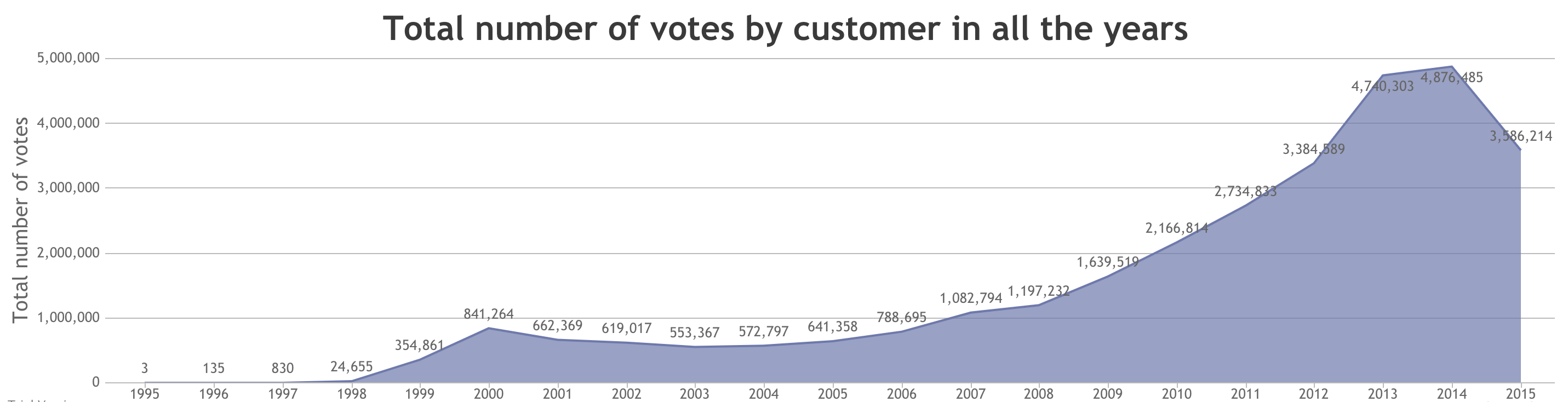
1. **STANDARD DEVIATION AND MEDIAN BASED ON RATING FOR THE PRODUCTS**

****

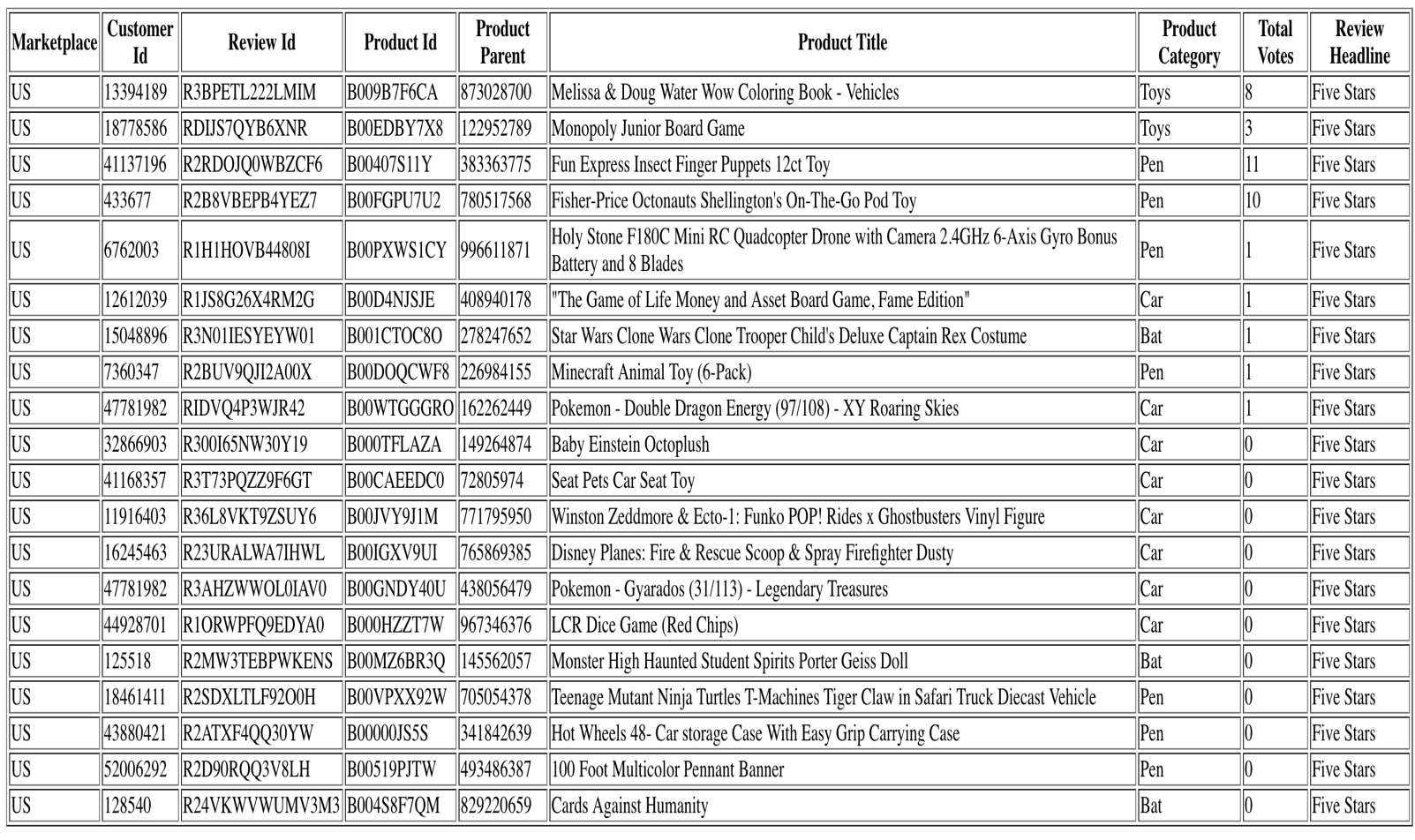
1. **TOP 100 PRODUCTS BASED ON TOTAL VOTES**

****

1. **TOTAL VOTES BASED ON YEAR**

****

1. **FILTERING DATA FOR REVIEW HEADLINE WHICH HAS “FIVE STARS”**

****

**CONCLUSION:**

The MapReduce algorithms performed in the project are java codes with implementation on Hadoop, HDFS, Mongo DB, pig. Spring MVC is used to show the statistical analysis of the data.