A Mini- Project Report

on

"AMAZON DATASET ANALYSIS (WATCHES)"

Submitted to the

Pune Institute of Computer Technology, Pune

In partial fulfillment for the award of the Degree of

Bachelor of Engineering

in

Information Technology

by

AJINKYAKULKARNI	33126
YASH KULKARNI	33129
SHUBHAM LOYA	33133
SAHIL NAPHADE	33140

Under the guidance of

Prof. D. D. Londhe

Department of Information Technology

Pune Institute of Computer Technology College of Engineering

Sr. No 27, Pune-Satara Road, Dhankawadi, Pune - 411 043.

CERTIFICATE

This is to certify that the project report titled **Amazon Dataset Analysis,** Submitted by Ajinkya Kulkarni (33126), Yash Kulkarni (33129), Shubham Loya (33133) and Sahil Naphade (33140) is a bonafide work carried out by them under the supervision of Prof. D. D. Londhe and it is approved for the partial fulfillment of the requirement of Software Laboratory Course-2015 for the award of the Degree of Bachelor of Engineering (Information Technology) .

Prof. D. D. Londhe
Internal Guide

Dr. A. M. Bagade
Head of Department,
Department of IT

Prof. D. D. Londhe
Internal Guide

Date:

Place:
Date:

ACKNOWLEDGMENT

We thank everyone who have helped and provided valuable suggestions for successfully creating a wonderful project.

We are very grateful to our guide, Prof. D. D. Londhe, Head of Department Dr. A. M. Bagade and our principal Dr. P. T. Kulkarni. They have been very supportive and have ensured that all facilities remained available for smooth progress of the project.

We would like to thank our professor and Prof. D. D. Londhe for providing very valuable and timely suggestions and help. We would also like the entire project staff team for providing valuable reviews and suggestions from time to time.

We would like to thank our entire department and college staff for the very valuable help and co-ordination throughout the duration of the project.

We would also like to thank our families and all our friends for the valuable support they provided throughout the duration of the project.

Ajinkya Kulkarni

Yash Kulkarni

Shubham Loya

Sahil Naphade

ABSTRACT

Amazon dataset is one of the most exciting dataset that can be used for analysis purpose. It has huge amount of data to work upon and make some important predictions and analysis for future sales. Analysis of this online purchasing dataset can help customers as well as service providers to a very good extent.

For this project in particular we are using Amazon feedback dataset (data of watches in particular) for our analysis purpose. This dataset consists of feedback given for various watches purchased from year 2002 to 2015. Dataset consists of various parameters like rating, reviews, verified purchases, etc. which helps in making predictions and analysis.

In this project we have tried to visualize various graphs based on various fields given in the dataset. To mentions about the graphs we have a bar-graph representing about the number of products which has got ratings varying from 1 to 5 , we have a bar-plot representing the number of products which are verified and how many are not verified , year-wise sale , month-wise sale , and many more visualizations are made.

Here we have performed market basket analysis and predictive algorithm on our dataset . So overall future prediction is made based on the customer purchases for the improved business strategy. This algorithm uses previous data in order to find the probability of customer buying the related products thereby increasing the overall profit margin.

1. Introduction

1.1 Purpose

Reviews have a great impact over the future in the sense that it helps in improved business. Here we are considering Amazon product reviews (watches in particular) and analyzing it on the basis of ratings, feedback, etc. This could help customers as well as service providers in providing the service more efficiently.

1.2 Problem Statement

Amazon is one of the most popular platform for purchasing the products online. However, every customer must get an insight of the product before purchasing it. Reviews on any particular product helps customer in deciding ‘what to purchase. Analysis of reviews helps in improved business as well as satisfaction of the customer. This project aims to produce learning models, trained on identified features, relevant to the outcome of the reviews. Such prediction can be utilized with multiple aspects, and can also be used by the company to identify improvement areas and successfully improve on the identified areas.

1.3 Scope

This project consists of multiple phases of data processing like, extraction – transformation – loading, preprocessing, cleaning, data model planning, feature identification, data model building, visualization. The project utilizes the following learning algorithms: Logistic Regression, K Nearest Neighbors, Naïve Bayes, Support Vector Machines, Neural Network.

1.4 Objective

The objective of this project is to create a data processing, analytics, visualization and prediction for processing Amazon dataset. This project includes building and training of various learning models for prediction of outcome Amazon reviews dataset for watches.

2. Literature Survey

2.1 Introduction

Multiple approaches have been used for data analytics on amazon dataset. Amazon is one of the leading websites in the world for e-commerce. Reviews on any particular product helps customer in deciding ‘what to purchase’. Analysis of reviews helps in improved business as well as satisfaction of the customer

2.2 Detail Literature Survey

The lifeblood of retail businesses has always been sales. A retailer can never assume that his customers know all of his offerings. But rather, he must make the effort to present all applicable options in way which increases customer engagement and increase sales.

Association Rule Mining

Association Rule Mining is used when you want to find an association between different objects in a set, find frequent patterns in a transaction database, relational databases or any other information repository. The applications of Association Rule Mining are found in Marketing, Basket Data Analysis (or Market Basket Analysis) in retailing, clustering and classification.

The most common approach to find these patterns is Market Basket Analysis, which is a key technique used by large retailers like Amazon, Flipkart, etc to analyze customer buying habits by finding associations between the different items that customers place in their “shopping baskets”. The discovery of these associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers The strategies may include:

- Changing the store layout according to trends
- Customer behavior analysis
- Catalog design
- Cross marketing on online stores
- What are the trending items customers buy
- Customized emails with add-on sales etc..

Online retailers and publishers can use this type of analysis to:

- Inform the placement of content items on their media sites, or products in their catalog
- Deliver targeted marketing (e.g. emailing customers who bought products specific products with other products and offers on those products that are likely to be interesting to them.)

Difference between Association and Recommendation

Association rules do not extract an individual's preference, rather find relationships between sets of elements of every distinct transaction. This is what makes them different than Collaborative filtering which is used in recommendation systems.

To understand it better take a look at below snapshot from amazon.com and you notice 2 headings "Frequently Bought Together" and the "Customers who bought this item also bought" on each product's info page.

"Frequently Bought Together" → Association

"Customers who bought this item also bought" → Recommendation

3. System Architecture

Training Pipeline

`

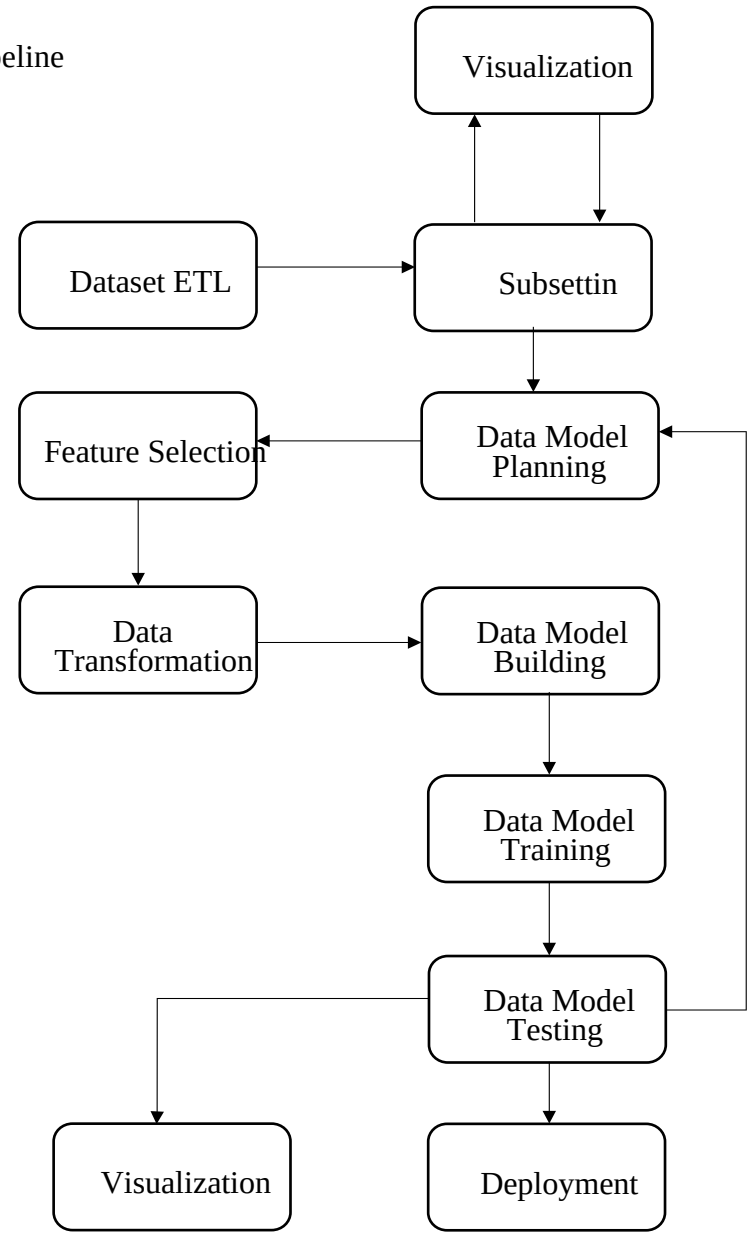


Figure 1: Architecture (Training Pipeline)

Utilization Pipeline

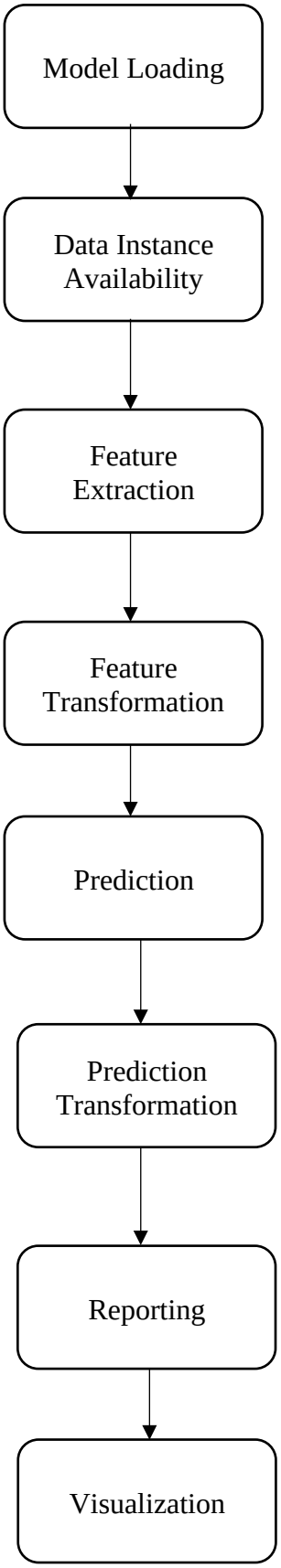


Figure 2: Architecture (Utilization Pipeline)

3.2 Dataset Analysis

The number of records in the selected dataset (watches): 960203. Other datasets are ten-folds increase in the same. The dataset contains records from 2002 to 2015. The data is already cleaned at the source, hence there is no issue of unclean or issues in quality of the data. The dataset is collected from one marketplace only, USA. Other multilingual datasets are also available.

Links for Datasets

The datasets are available at:
https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt
The various products available are:
1. Watches
2. Video games
3. Software
4. Tools (Power tools)
5. Electronics and many more

3.3 Phases

Data ETL

This phase consists of the first interaction with data. ETL stands for Extract Transform and Load. Data may be present in formats, different than what is required for carrying out different tasks. Thus, data first needs to extracted from source, in a form which can be processed by the computer. Thus, data then needs to be converted to formats, convenient for further tasks. Thus, data needs to transformed. Finally, data must actually be loaded onto the working computer memory, so that further operations can be applied on it. This is the loading phase.

Preprocessing

Preprocessing consists of the actions to be performed on loaded data, to make it usable and more meaningful, to further learning or analytical pipelines. Data may consist of errors generated to incorrect observations, incorrect storage, or even degradation of storage equipment over time. Data can also contain missing values. Inconsistent data or data with missing values may simply be excluded to prevent it from impacting learning pipelines. However, in case of scarcity of data, evert data instance is essential and strategies to mitigate the inconsistencies, such as outlier replacement are required.

Data Model Planning

Data model planning is the phase of choosing the right kind of data models which can well represent the features of the data set at hand. This requires statistical and mathematical expertise. Choosing the right kind of the data model ensures maximum efficiency and performance. Data model planning also involves interpreting the right data representation, suitable to the chosen model.

Feature Selection

Choosing the right features for data model is essential. The right features help represent maximum relevant information from the data set. It helps the data model achieve maximum efficiency and performance. Various techniques like covariance and correlation exist to identify the right features and their representation.

Data Model Building

Data model building of actually defining the appropriate data model, with specific structure. The defined model can then learn from the training data set and can be tested on the testing data set. This model can then be deployed for further use.

Visualization

Categorize the products into 5 sub-categories based on star-rating between 1 to 5. Visualize the count of verified products and unverified products based on given column. Visualize the word-cloud based on the feedback given to the product

4. Results

Most Sold Product

Year 2008-2015 - Watch Repair Kit

Year 2002 - Kenneth Cole Women's KC4232 Reaction Silver
.
Year 2003 - Hamilton H91514733 (Men's Watch)

Year 2004 - Citizen Skyhawk JR3060-59F

Year 2005 - Seiko Men's SKX779 "Black Monster"

Year 2006 - Casio Men's VA430SGA-9A1V Waveceptor Solar Atomic Watch

Year 2007 - Double Automatic Vertical Watch Winder Boxana

Prediction of Market Basket Analysis

The confidence of Buying a 21-piece watch repair kit, is 1, suggesting that people are likely to buy is again. Also, the same is the case for “Disney’s Frozen Elsa & Anna Singing watch” has confidence 1, which also suggest that the product, on its own is very likely to be sold (for children).

Observations-

Thus, it is found from the visualization that almost 90% of the total products are verified. Overall quality of the products is good based on star-rating. Subset of all the word-clouds tells that most of the feedback are positive. Top 20 and bottom 20 products are sorted for future reference. The sales are at its peak in December, owing to the Christmas vacations and trend of gifting the loved ones. The second most sold product, “Bling Jewellary plated Ladies watch”, confirms the same (Product ID: B004YM2FV2). Also, the watch “Disney’s Frozen Elsa & Anna Singing watch” is very likely to be sold, as part of gift for children.

Screenshots of observation:

Top 20 and Bottom 20 products based on no.of purchases:

```
Top 20 reviewed products are:

product_id
B000T9VK56    4390
B004YM2FV2    3050
B005JVP0LE    2047
B002SSUQFG    1945
B008D902Q2    1884
B000AR7S3A    1516
B000JQJS6M    1465
B00791QYMQ    1421
B000EQS1JW    1410
B000GAWSHM    1309
B000GAYQLI    1308
B000GAYQKY    1254
B000LTAY1U    1238
B000GAWSDG    1208
B000JQFX1G    1161
B00791R1MI    1101
B0006AAS7E    1089
B004VR9HP2    1058
B003DZDYMJ    1044
B004VR9GCQ    1022
Name: star_rating, dtype: int64

Most Reviewed Product, B000T9VK56 - has 4390 re
```

```
Bottom 20 sorted products
product_id
B004GL6QMA    1
B00DYNPVUO    1
B00DYO4SQG    1
B004GIA1BK    1
B004GI4E2C    1
B00DYXHBLQ    1
B004GHY39M    1
B004GI09LM    1
B00DYX87LO    1
B004GI32HK    1
B00DYVH0AK    1
B004GI4J2M    1
B00DYQ1ZJC    1
B00DYQ5548    1
B00DYQ50JI    1
B004GI5ZFC    1
B004GI60NS    1
B00DYQ508Y    1
B004GIA0H0    1
1380137136    1
Name: star_rating, dtype: int64
```

Most sold product of the year(2015):

```
product_id
B000T9VK56    1072
B004YM2FV2     752
B008D902Q2     595
B005JVP0LE     568
B00791QYMQ     509
              ...
B00B3EON8C       1
B00B3DIXR0       1
B00B3DIWU8       1
B00B3DIWFI       1
B008PATDA0       1
Name: product_id, Length: 59853, dtype: int64
The most selling product is product_id
B000T9VK56    1072
Name: product_id, dtype: int64
The most grossing product of year 2015 selled 1072 times
```

6. Conclusion

Conclusion-

Thus we analyzed the Amazon watches dataset, and identified the best selling products per year, most reviewed product ever, general review of the sales. Also we saw that the sales are exponential in growth and expected to continue in the future, along with its visualization. This study shows that the online market is reaching its potential to serve the serve the customers and keeping them happy. The study was conducted on a relatively smaller portion of the overall categories, showing that the online market of Amazon in US is really huge.

Future Scope-

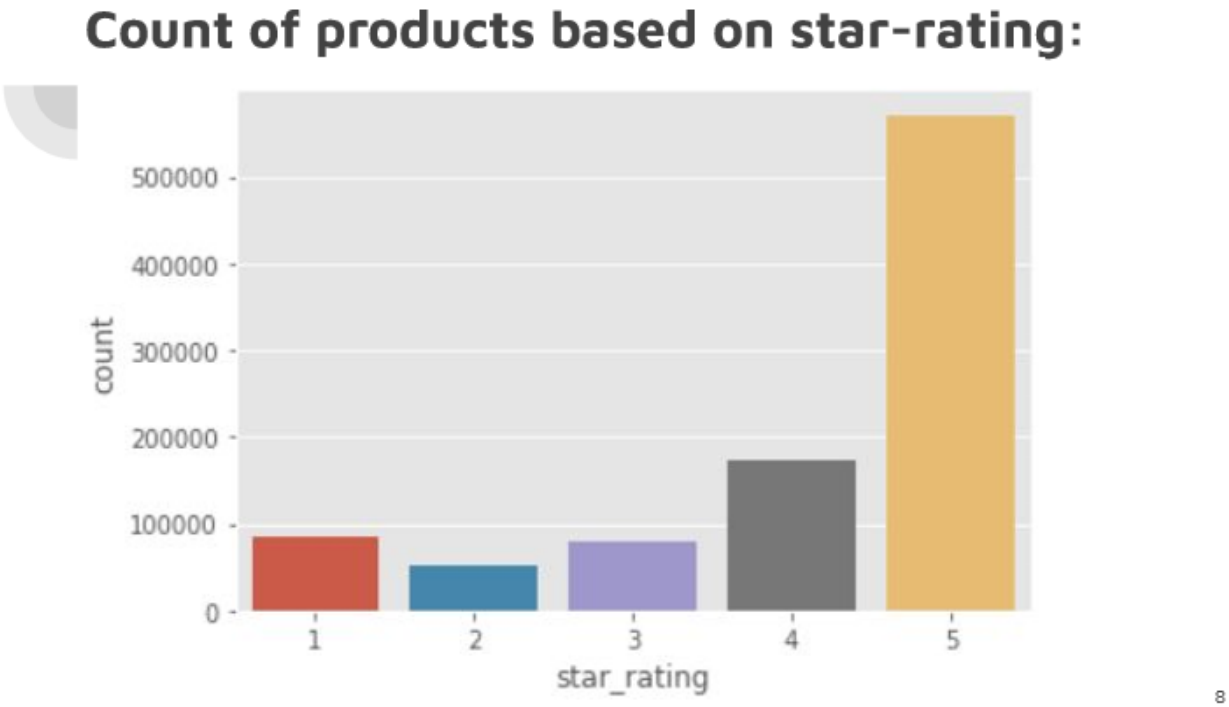
Sentiment analysis of reviews will help to learn about different products and people feedback towards it. Prediction of future sales in upcoming year, for e.g. which month to launch a particular watch to maximize product, or which product is most sold and should be made in greater quantity. Find out minimum cost of particular watch and from which portal to buy it from depending on the customer feedback. Different data sources can be integrated to predict the best portal for buying the watch.

7. References

1) https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt
2) https://towardsdatascience.com/market-basket-analysis-978ac064d8c6
3) https://www.kaggle.com/learn/intro-to-machine-learning

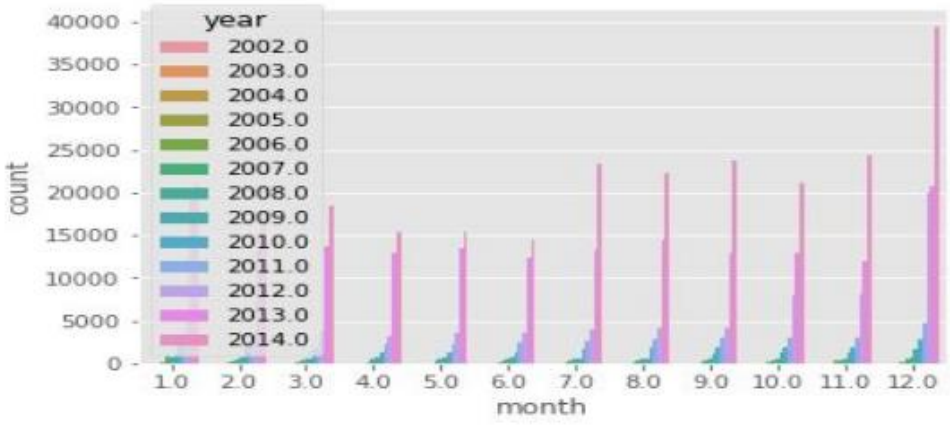
8. Appendix

Screenshots of Project

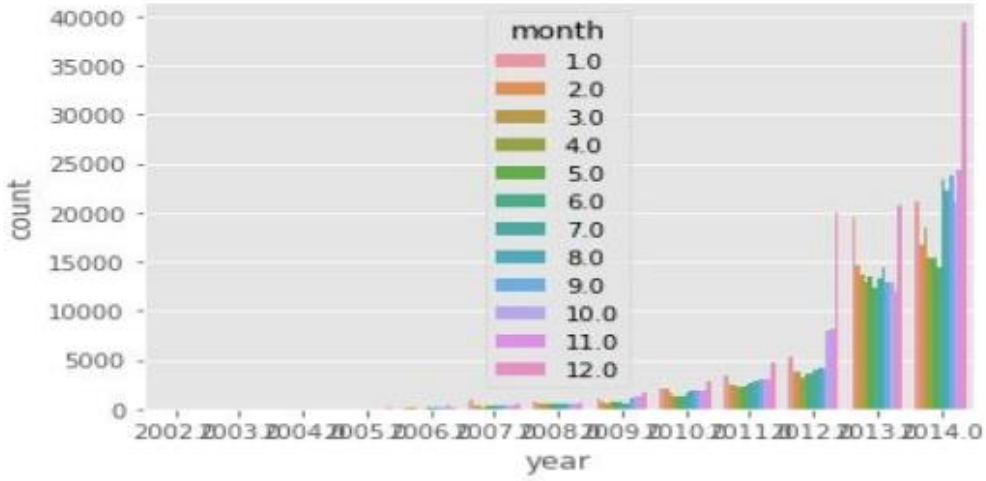


8

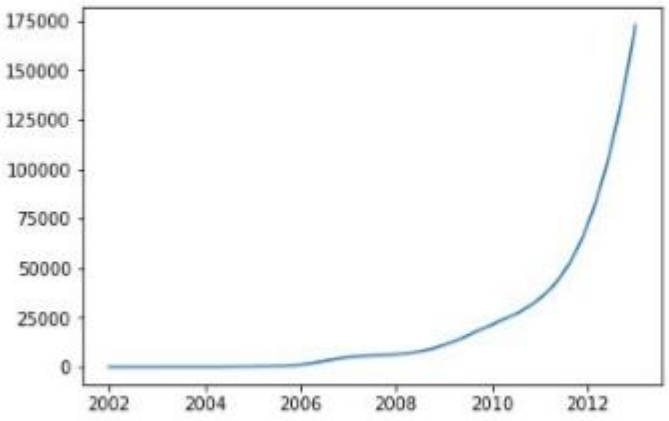
Monthly Sales -Year wise



Yearly Sales-Month Wise

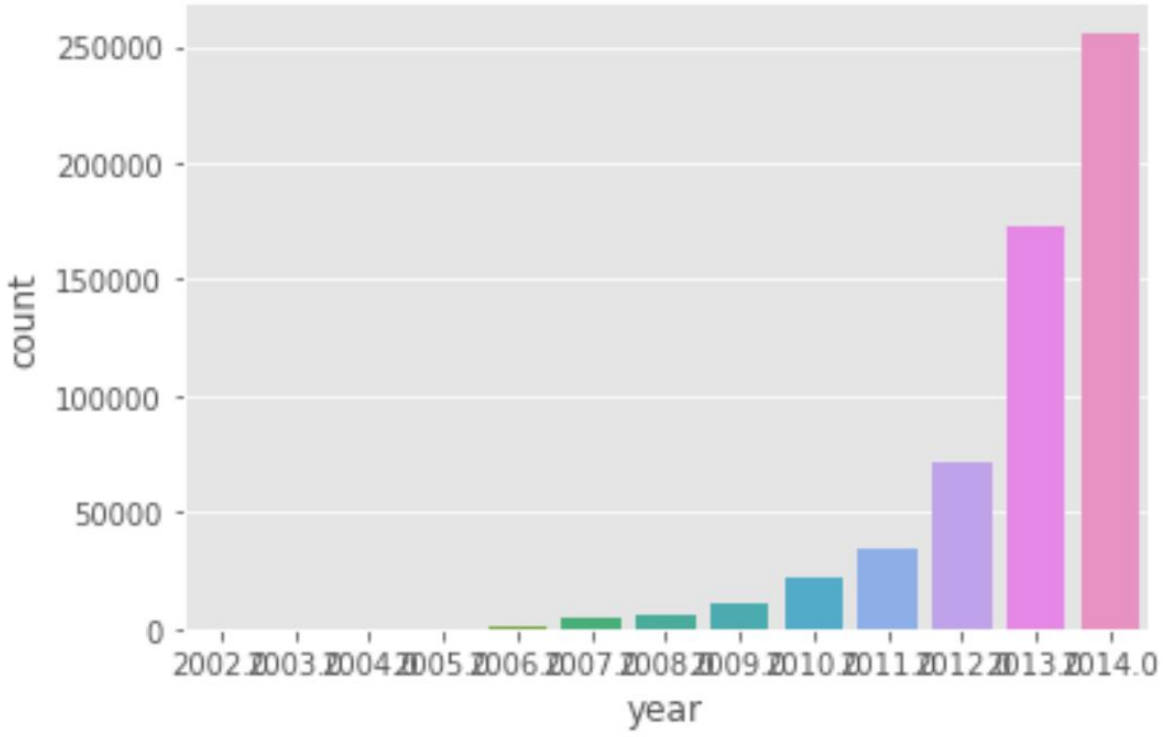


Plotting the sales vs Year

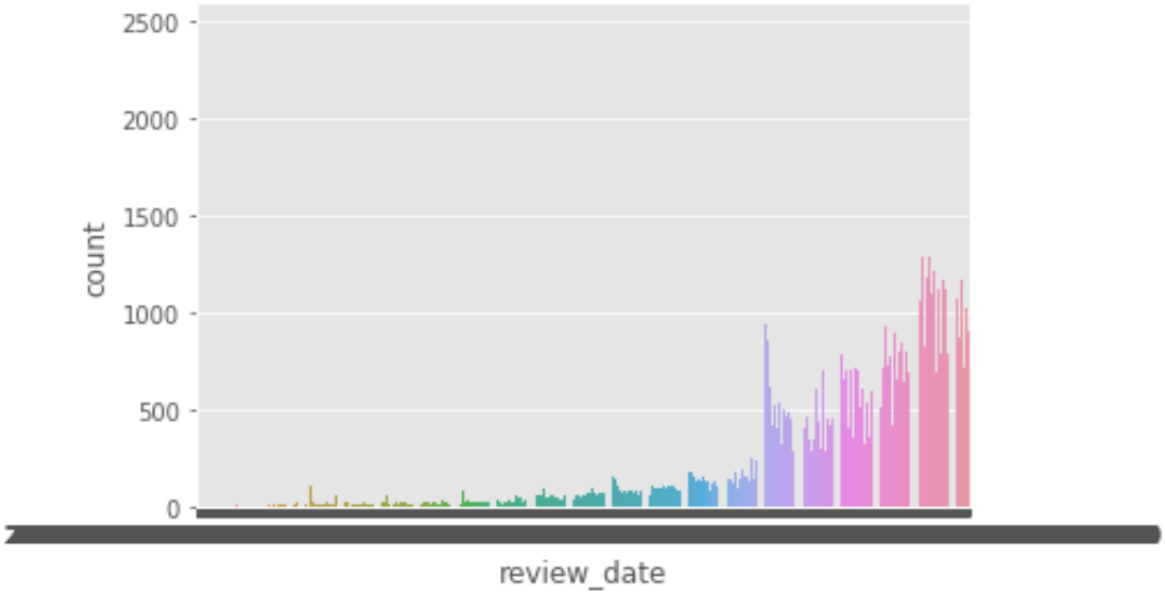


WORD-CLOUD OF FEEDBACK

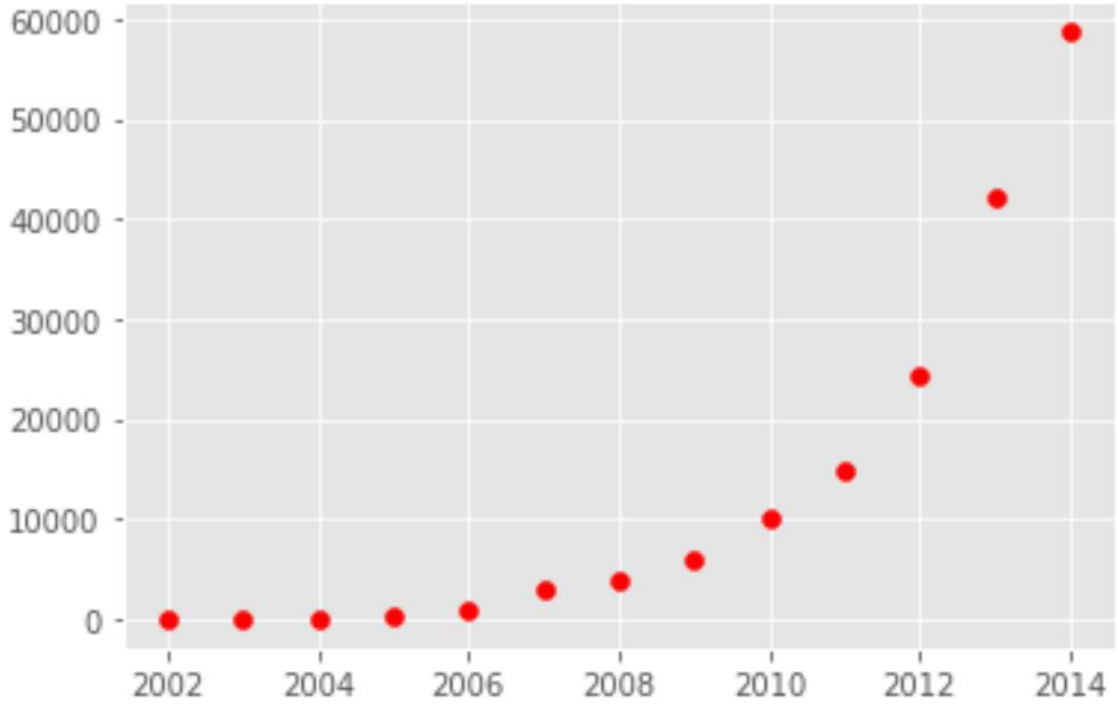
1



Sales count from 2002 to 2014



Monthly reviews as per date



Sales graph as per year (Scatter plot)