Project report on Amazon Customer Reviews
MSDA, San Jose State University
Group 9

Yogita Suryavanshi *SJSU ID : 015274077*
Cheuk Ip Hong *SJSU ID : 015265588*
Hrushikesh Pokala *SJSU ID : 015349321*
Saroj Saran *SJSU ID : 015276014*
Preeti Khatri *SJSU ID : 015261259*

***Abstract*—The world we live in is becoming increasingly digitized. In this digitized environment, e-commerce is gaining traction by bringing goods closer to consumers without requiring them to leave their homes. Since so many people nowadays depend on online items, the value of a review is increasing. A consumer must read thousands of reviews to understand a product before making a purchase. However, in this day and age of machine learning, sorting through thousands of feedback and learning from them would be much easier if a model was used to polarize and learn from them. On a large-scale Amazon dataset, we used supervised learning methods to polarize it and achieve adequate accuracy.**

1. Introduction

Motivation

The motivation behind the project was based on our significant paper presentation and industry case study where both are amazon related studies. Our significant paper presentation is based on Amazon.com recommendations where amazon uses one such recommendation algorithm to personalize the online store for each customer. The store radically changes based on customer interests, Thus highlighting the products which might interest them based on their search pattern like showing programming titles to a software engineer and baby toys or related products to a new mother. Since this paper was highly dependent on the reviews and ratings of customers, this led us to choose the amazon customer review dataset for our project as well.

Our industry case study is based on a checkout-free and convenience store called Amazon Go where it takes the idea of convenience to a new level by Quick access to everyday products, especially groceries and convenience goods. The Consumer scans the grocery store with Amazon App on smartphones. Consumers go around the store, pick up items, add to bags, shop like normal, and consumer exits. This eventually led us to

1.1. Literature Survey(based on significant paper :

Introduction

This literature review is essential to obtain a clear picture of the studies that have

been conducted in these areas to date, and where it is possible to identify gaps in the existing research. Our literature survey is dependent on our significant paper presentation "Amazon.com Recommendations". In this paper, amazon uses one such recommendation algorithm to personalize the online store for each customer. The store radically changes based on customer interests, Thus highlighting the products which might interest them based on their search pattern like showing programming titles to a software engineer and baby toys or related products to a new mother.

Several challenges to implementing the recommendation algorithms include: millions of customers contributing to huge data, real-time analysis for a high-quality recommendation, new customers have limited history and recommendation suggestions based on the limited available data, old customers have huge data history and browsing through all the data to better recommend as per the current interest.

The review of literature is divided into categories based on our significant paper presentation, designing and analyzing the project, and using data visualization. The common approaches that play an important role to solve the recommendation problems are collaborative filtering, cluster models, and search-based methods. The main approach used by amazon is Item to Item collaborative filtering which is efficient enough to handle the number of customers and number of items in the product catalog. Thus producing recommendations in real-time, scales to massive data sets, and generates high-quality recommendations.

The variables that play an important role in analyzing and designing the project are mean, standard deviation, regression, and hypothesis testing.

Collaborative Filtering: The algorithm then uses the filtering criteria for recommendations to list down the items vector component values to recommend the current customer, hence it uses the similarity between the current customer and other customer's details from the database for a recommendation.

Cluster Models: Cluster Model follows the algorithm of finding the customer with the most similarity with the current customer and assigning it to the segment of users. It then uses the purchases and ratings of the customers in the segment to generate recommendations.

Search-Based Methods: Search-based algorithms mostly search for the similar item as searched, i.e the algorithm constructs a searching look based on the user's purchased and rated items, algorithm construct query to find the other similar items based on the keyword search. Like books by similar authors, or similar products by different brands.

Item to Item Collaborative Filtering: Existing recommendation algorithms may not be efficient and suitable for tens of millions of customers at amazon, so at Amazon, they developed item to item

collaborative filtering, which scales to their massive datasets and produces high-quality recommendations. With the Item to Item filtering method rather than matching the user to a similar customer, it matches each of the customers purchased and rated item to similar items and combines those similar items into a recommendation list. This algorithm builds a similar item table by finding items that customers tend to purchase together.

Mean: The mean, also referred to as the average, is the most common statistic used to measure the center of a numerical data set. If you have a data set with a wide range of numbers, knowing the mean can give you a general sense of how these numbers could essentially be put together into a single representative value.

Standard Deviation: Standard deviation measures the spread of data distribution. The more spread out a data distribution is, the greater its standard deviation. Interestingly, the standard deviation cannot be negative. A standard deviation close to 0 indicates that the data points tend to be close to the mean.

Regression: Regression Analysis, a statistical technique, is used to evaluate the relationship between two or more variables. Regression analysis helps an organization to understand what their data points represent and use them accordingly with the help of business analytical techniques in order to do better decision-making.

Hypothesis Testing: Hypothesis testing is the process used to evaluate the strength of evidence from the sample and provides a framework for making determinations related to the population, ie, it provides a method for understanding how reliably one can extrapolate observed findings in a sample understudy to the larger population from which the sample is drawn.

Data Visualization Technology: Data visualization is scientific and technological research on the visual representation of data. Among them, the visual representation of this kind of data is defined as a kind of information extracted in a certain summary form, including various attributes and variables of corresponding information units. Data visualization includes basic concepts such as data space, data development, data analysis, and visualization.

Charts: Data visualization charts can be classified into the following categories according to the function and functions of data: comparison, distribution, process, map, proportion, interval, association, time, and trend. Each type of chart can contain different data visualization graphics such as bar chart, pie chart, bubble chart, thermal chart, trend chart, histogram, radar chart, color block diagram, funnel chart, chord chart, dashboard, area chart, broken line chart, K-line chart, ring chart, word cloud, etc.

Pandas: According to P.Lemenkova(2019) Pandas is a tool based

on Numpy, which is created to solve data analysis tasks. Pandas include a large number of libraries and some standard data models, providing the tools needed to operate large data efficiently. Functions and methods of Pandas can enable us to process data quickly and conveniently.

1.2. Methodology:

We have used MYSQL and MYSQL Workbench to store and query the data in order to create a well-structured relation schema between all tables and have a better understanding of the relationship from the data by modeling ER Diagrams. We have used the Talend ETL tool and pandas for data cleaning and loading. We have made use of Python sklearn for data analysis. Finally have used Tableau for the data visualization to create a graphical representation of detailed information such as charts, graphs, and maps to understand trends and outliers.

2. Project Walk-through:

Data sourcing and modelling :

Our Project dataset was taken from Kaggel which Focuses on customer reviews data for one of the online retailer business "Amazon.com". Describing below is the ER diagram with details on the entities and entity relation model. Entities involved are :

1) Customer : Holding details about customers. With identifier key as usernameID
2) Products : includes details about the amazon products, with identifier key as productID
3) Review : holding the details about the customer review,
4) Categories: holding the information about the product categories, with CategoriesID as the identifier key.
5) Brand: Brand of the products, brandID
6) ReviewDate: holding the dates at which reviews were added,ReviewDateID is the identifier key.
7) ProductDate: Holding the products added dates, ProductDateID is the identifier key

With snowflake schema, we have a factless fact table as 'Associate' holding the foreign key for the associated entities.

2.1. ER Diagram:

For the ER model, we decided to use a snowflake schema with a factless fact table for a couple of reasons. First, we wanted to eliminate a data redundancy to make a better data quality. Second, we want to utilize as little disk storage as possible. Third, our datasets are selected between 2015 to 2018, so we use a factless fact table for this model because the measurements are not continuous. Finally, denormalized data is hard to understand the relationship between all the data in the dataset to analyze. To achieve our goal for snowflake schema, we used a normalization to get rid of the duplication and dimensional

hierarchy for attributes that we did not use often. Refer fig 1.

Relational Model : Relation model is also provided below. With details on the relational attributes and the association between each relational table. Refer fig 2.

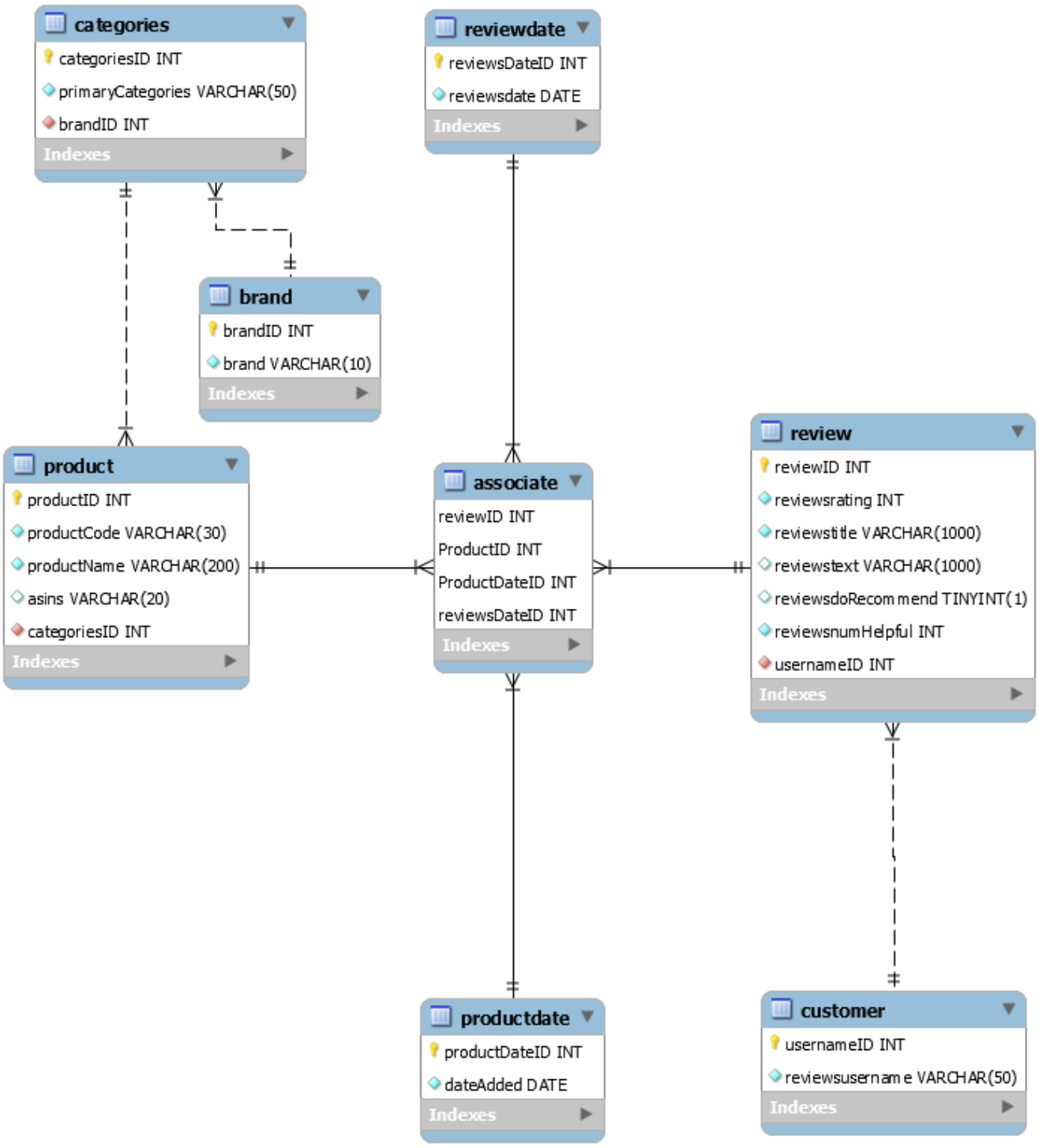


Figure 1. ER Diagram

2.2. Data cleaning and loading :

We have used python Pandas for data cleaning and Talend ETL for data loading.

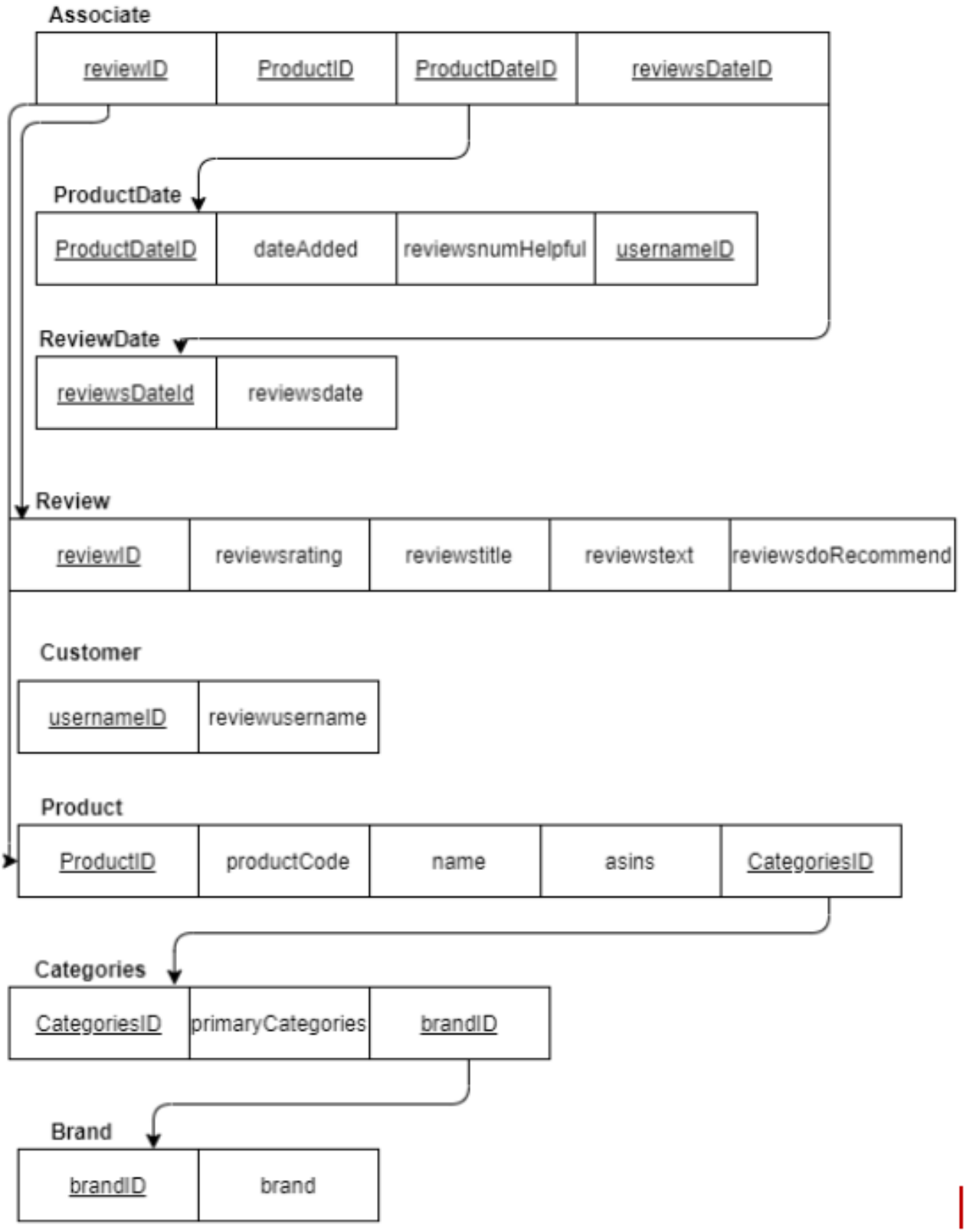


Figure 2. Relational Model

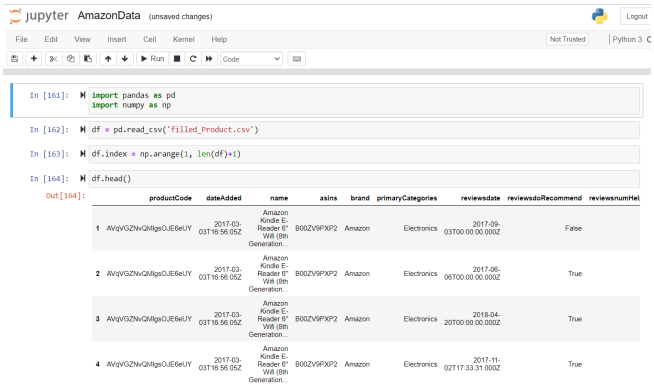


Figure 3. Python Pandasl

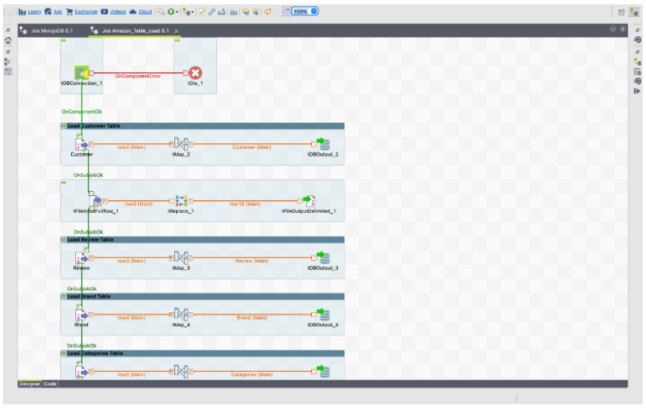


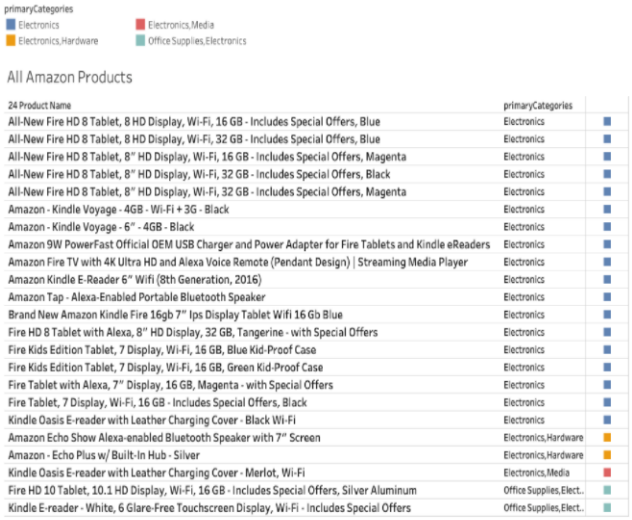
Figure 4. Talend ETL

Code for python data cleaning and ETLjob screens are placed at github repository: https://github.com/preetikhatrisjsu/DATA225_Project

2.3. Data Analysis and Visualization:

Showing all Amazon products in dataset :
Our goal is to conduct quantity and quality reviews based on the Amazon Products below: fig 5

Sentiment Analysis :



24 Product Name	primaryCategories
All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi, 16 GB - Includes Special Offers, Blue	Electronics
All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi, 32 GB - Includes Special Offers, Blue	Electronics
All-New Fire HD 8 Tablet, 8" HD Display, Wi-Fi, 16 GB - Includes Special Offers, Magenta	Electronics
All-New Fire HD 8 Tablet, 8" HD Display, Wi-Fi, 32 GB - Includes Special Offers, Black	Electronics
All-New Fire HD 8 Tablet, 8" HD Display, Wi-Fi, 32 GB - Includes Special Offers, Magenta	Electronics
Amazon - Kindle Voyage - 4GB - Wi-Fi + 3G - Black	Electronics
Amazon - Kindle Voyage - 6" - 4GB - Black	Electronics
Amazon 9W PowerFast Official OEM USB Charger and Power Adapter for Fire Tablets and Kindle eReaders	Electronics
Amazon Fire TV with 4K Ultra HD and Alexa Voice Remote (Pendant Design) | Streaming Media Player	Electronics
Amazon Kindle E-Reader 6" Wifi (8th Generation, 2016)	Electronics
Amazon Tap - Alexa-Enabled Portable Bluetooth Speaker	Electronics
Brand New Amazon Kindle Fire 16gb 7" Ips Display Tablet Wifi 16 Gb Blue	Electronics
Fire HD 8 Tablet with Alexa, 8" HD Display, 32 GB, Tangerine - with Special Offers	Electronics
Fire Kids Edition Tablet, 7 Display, Wi-Fi, 16 GB, Blue Kid-Proof Case	Electronics
Fire Kids Edition Tablet, 7 Display, Wi-Fi, 16 GB, Green Kid-Proof Case	Electronics
Fire Tablet with Alexa, 7" Display, 16 GB, Magenta - with Special Offers	Electronics
Fire Tablet, 7 Display, Wi-Fi, 16 GB - Includes Special Offers, Black	Electronics
Kindle Oasis E-reader with Leather Charging Cover - Black Wi-Fi	Electronics
Amazon Echo Show Alexa-enabled Bluetooth Speaker with 7" Screen	Electronics,Hardware
Amazon - Echo Plus w/ Built-In Hub - Silver	Electronics,Hardware
Kindle Oasis E-reader with Leather Charging Cover - Merlot, Wi-Fi	Electronics,Media
Fire HD 10 Tablet, 10.1 HD Display, Wi-Fi, 16 GB - Includes Special Offers, Silver Aluminum	Office Supplies,Elect..
Kindle E-reader - White, 6 Glare-Free Touchscreen Display, Wi-Fi - Includes Special Offers	Office Supplies,Elect..

Figure 5. Visual 1 for all amazon products in dataset

For an overall amazon product rating analysis, we used a Sentiment Analysis method to get an overview of all ratings from all the amazon products to see how people think of all Amazon products. In the Sentiment Analysis method, we assigned sentiment of 1 for rating is greater than 4 as a good review and sentiment of 0 for rating is less than 2 as a bad review, and we got rid of rating of 3 which is neutral. For example,refer fig 678

After we assigned sentiment, we used a python library and function called sklearn and train_test_split to split our datasets into the 75% for training set and the 25% for testing set.

We used Logistic Regression for our model since our training and testing models are binary 0 and 1.

reviewsrating	sentiment
5	1
4	1
1	0

Figure 6. sentiment analysis

```
#75% Train and 25%test
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0)
```

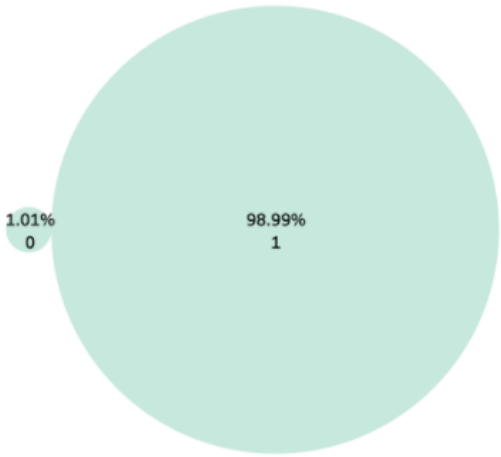
Figure 7. Python code : Sentiment analysis

```
from sklearn.feature_extraction.text import CountVectorizer
workcount = CountVectorizer()
numTFTR = workcount.fit_transform(X_train)
numTFtest = workcount.transform(X_test)
```

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
model = LogisticRegression()
model.fit(numTFTR, y_train)
```

Figure 8. Python code : Sentiment analysis

Here is the result after all the computation. The big circle represents the percentage of all the good reviews and the small one represents the percentage of all the bad reviews in a prediction model. Based on the result of accuracy_score, the predicted model and the testing set have over 98% accuracy, so we can say that over 98% of customers recommend their products.Refer fig 9




```
accuracy_score(y_test, predict)
```

```
0.9816819317235637
```

Figure 9. visual : Sentiment analysis

Querying and Visualization :

For our querying and visualization section, we are connecting Tableau to MYSQL to create some useful queries and graphs representing the query.

1) What is the average rating on each product?
The highest average rating is Amazon Fire TV with 5 star average rating. refer fig 10

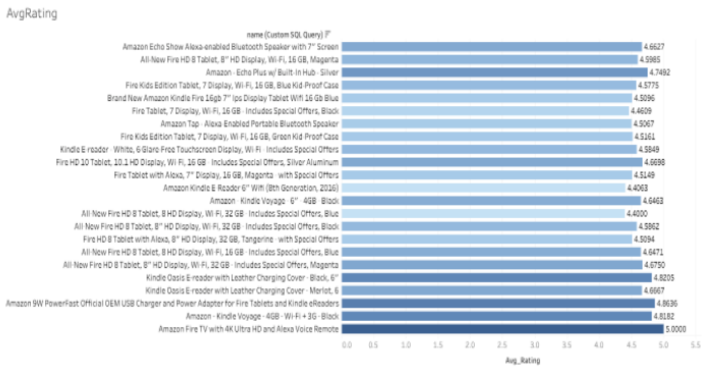


Figure 10. visual : Average rating on products

2) What is the total amount of reviews for each product from the most to the less?
We can see that Amazon Echo Show Alexa has the most reviews and we can also see that Amazon Fire TV has the least reviews on the bottom of the graph. refer fig 11

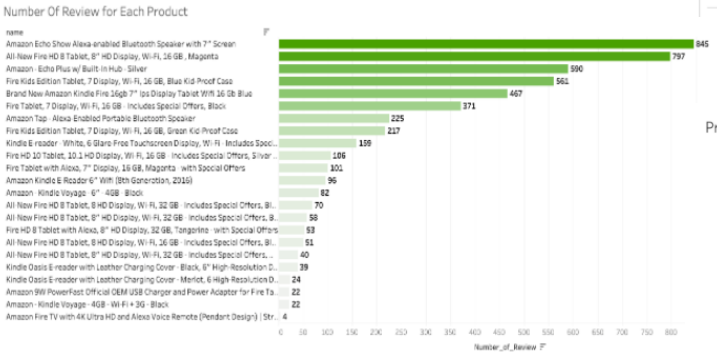


Figure 11. visual : total amount of reviews per product

3) Year Wise listing the count of products added.
We can notice which Amazon product is the most popular and the most reviewed product each year.We are listing all the new products and number of products which are added every year to the item catalog.Refer fig 12

4) Categorizing the products based on category,
Database holds Electronics products, hence gives a distinct of the product in each category.Refer fig 13

5) Year-wise listing the products added :
We are listing all the new products which

Figure 12. visual : year-wise count of products

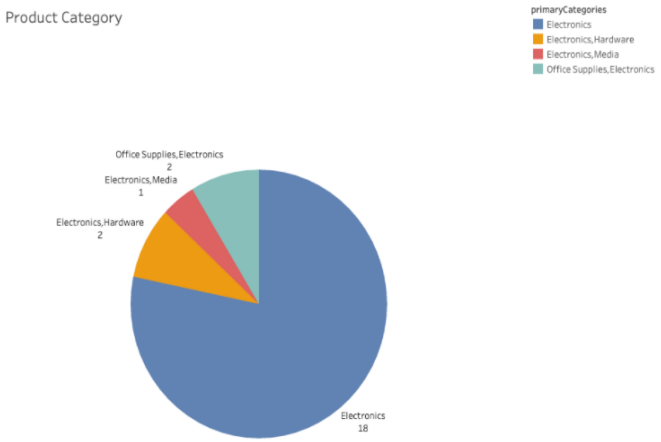


Figure 13. visual : categories of products

are added every year to the item catalog, Database hold data from 2015 to 2018, we can see form the visual that there we re 15 new products added to catalog in the year of 2017, i.e max of the rest years.Refer fig 14

6) Finding from the reviewer ratings to understand how likely the reviewers are to recommend the product in percentage, for all the products, and for all five ratings, what is the percentage likely of a customer reviewed a product to recommend it to other customers.

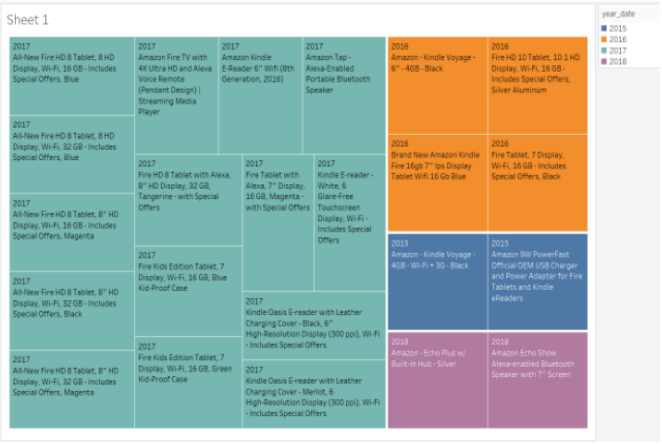


Figure 14. visual : categories of products

Figure 15, shows the Tabular representation of all the products review ratings and in percentage representing how likely of unlikely it is that the customer will recommend the product, we extracted this information from the field, reviewsdoRecommend which hold true of false in the form of 1 or 0,representing the customers who rated the product will recommend the product to other customer or not.

Second image of figure 15, shows the visuals for all the amazon customers who have rated products from 1-5 and for each ratings what is the percentage prediction of how likely the customer is going to recommend the product, Blue dots represents that the customer are highly likely to recommend products.and the one with red shows that the customers would not wish to recommend the product.

Figure 16, shows the visuals for all the amazon customers who have rated products from 1-5 and for each ratings what is the percentage prediction of how unlikely the customer is going to recommend the product, Blue dots represents that the customer are highly unlikely to recommend products.and the one with red shows that the customers would wish to recommend the product.

7) Top 10 Customers who have been actively rating and reviewing the products, based on their review and ratings records.Refer fig 17

Blue bar represent the customer who has rated the maxi mun so far in our databases, these kind of in formations would be useful for retailer company like Amazon reward the customer's who have been giving inputs to manage their business better.

8) Top three most 5 star rated products :

These are products which are most 5 rated product, which would help the amazons recommendation system with its algorithm.Refer fig 18

productname	reviewsrating	Likely_To_Recommend(%)	Not_Likely_To_Recommend(%)
All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi, 16 GB - Includes Special Offers, Blue	4	93.75	6.25
All-New Fire HD 8 Tablet, 8" HD Display, Wi-Fi, 16 GB - Includes Special Offers, Magenta	5	99.6241	0.3759
All-New Fire HD 8 Tablet, 8" HD Display, Wi-Fi, 16 GB - Includes Special Offers, Magenta	4	98.6842	1.3158
All-New Fire HD 8 Tablet, 8" HD Display, Wi-Fi, 16 GB - Includes Special Offers, Magenta	3	40	60
All-New Fire HD 8 Tablet, 8" HD Display, Wi-Fi, 16 GB - Includes Special Offers, Magenta	2	16.6667	83.3333
All-New Fire HD 8 Tablet, 8" HD Display, Wi-Fi, 32 GB - Includes Special Offers, Black	3	33.3333	66.6667
Amazon - Echo Plus w/ Built-In Hub - Silver	4	97.8947	2.1053
Amazon - Echo Plus w/ Built-In Hub - Silver	3	64.2857	35.7143
Amazon - Echo Plus w/ Built-In Hub - Silver	2	14.2857	85.7143
Amazon - Kindle Voyage - 6" - 4GB - Black	5	98.3333	1.6667
Amazon - Kindle Voyage - 6" - 4GB - Black	4	88.2353	11.7647
Amazon Echo Show Alexa-enabled Bluetooth Speaker with 7" Screen	5	99.8405	0.1595
Amazon Echo Show Alexa-enabled Bluetooth Speaker with 7" Screen	4	99.4186	0.5814
Amazon Echo Show Alexa-enabled Bluetooth Speaker with 7" Screen	3	32.2581	67.7419
Amazon Echo Show Alexa-enabled Bluetooth Speaker with 7" Screen	2	11.1111	88.8889
Amazon Echo Show Alexa-enabled Bluetooth Speaker with 7" Screen	1	16.6667	83.3333
Amazon Kindle E-Reader 6" Wifi (8th Generation, 2016)	5	98.2143	1.7857
Amazon Kindle E-Reader 6" Wifi (8th Generation, 2016)	4	93.1034	6.8966
Amazon Kindle E-Reader 6" Wifi (8th Generation, 2016)	3	57.1429	42.8571
Amazon Tap - Alexa-Enabled Portable Bluetooth Speaker	4	94.2308	5.7692
Amazon Tap - Alexa-Enabled Portable Bluetooth Speaker	3	53.8462	46.1538
Amazon Tap - Alexa-Enabled Portable Bluetooth Speaker	2	28.5714	71.4286
Brand New Amazon Kindle Fire 16gb 7" Ips Display Tablet Wifi 16 Gb Blue	5	99.6764	0.3236
Brand New Amazon Kindle Fire 16gb 7" Ips Display Tablet Wifi 16 Gb Blue	4	95.7265	4.2735
Brand New Amazon Kindle Fire 16gb 7" Ips Display Tablet Wifi 16 Gb Blue	3	75	25
Brand New Amazon Kindle Fire 16gb 7" Ips Display Tablet Wifi 16 Gb Blue	1	7.6923	92.3077
Fire HD 10 Tablet, 10.1 HD Display, Wi-Fi, 16 GB - Includes Special Offers, Silver Aluminum	3	33.3333	66.6667
Fire HD 8 Tablet with Alexa, 8" HD Display, 32 GB, Tangerine - with Special Offers	3	50	50
Fire Kids Edition Tablet, 7 Display, Wi-Fi, 16 GB, Blue Kid-Proof Case	4	98.5612	1.4388
Fire Kids Edition Tablet, 7 Display, Wi-Fi, 16 GB, Blue Kid-Proof Case	3	57.6923	42.3077
Fire Kids Edition Tablet, 7 Display, Wi-Fi, 16 GB, Green Kid-Proof Case	3	63.6364	36.3636
Fire Tablet, 7 Display, Wi-Fi, 16 GB - Includes Special Offers, Black	5	99.1228	0.8772
Fire Tablet, 7 Display, Wi-Fi, 16 GB - Includes Special Offers, Black	4	97.2727	2.7273
Fire Tablet, 7 Display, Wi-Fi, 16 GB - Includes Special Offers, Black	3	78.9474	21.0526
Fire Tablet, 7 Display, Wi-Fi, 16 GB - Includes Special Offers, Black	2	50	50
Kindle E-reader - White, 6 Glare-Free Touchscreen Display, Wi-Fi - Includes Special Offers	5	97.1698	2.8302
Kindle E-reader - White, 6 Glare-Free Touchscreen Display, Wi-Fi - Includes Special Offers	4	97.619	2.381
Kindle E-reader - White, 6 Glare-Free Touchscreen Display, Wi-Fi - Includes Special Offers	3	22.2222	77.7778
Kindle Oasis E-reader with Leather Charging Cover - Black, 6" High-Resolution Display (300 ppi), Wi-Fi - Includes Special Offers	5	97.1429	2.8571

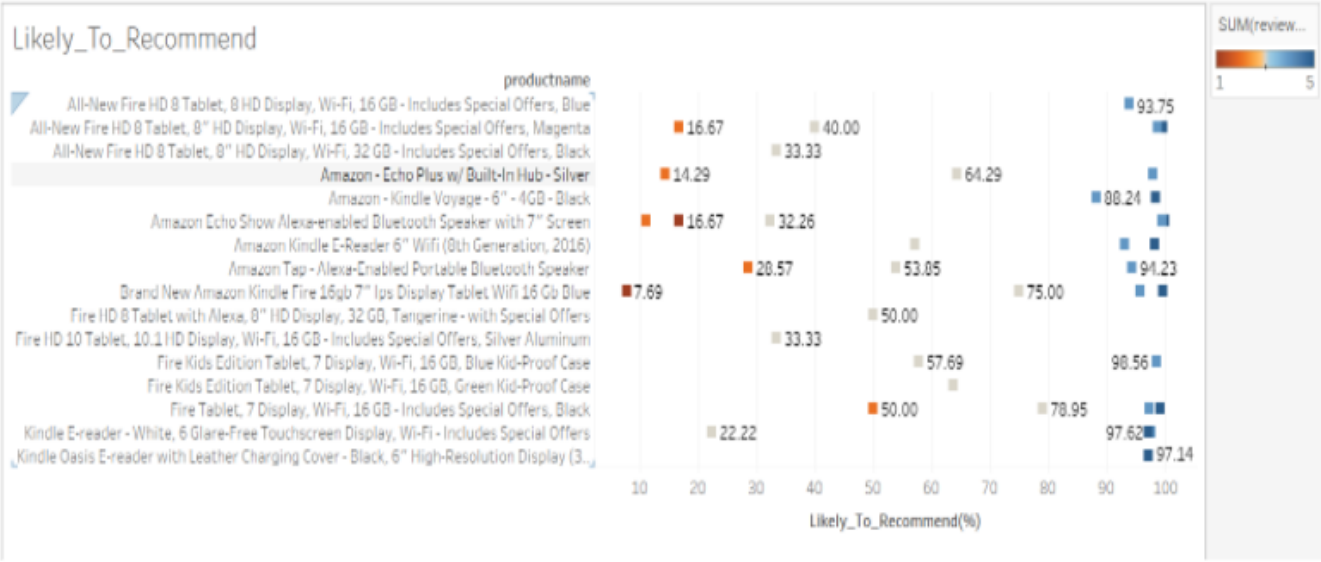


Figure 15. visual : categories of products

9) Search for the top 30 words that show up in Amazon Echo Show Alexa review.

Based on the result of showing the number of reviews on each product, we choose the product that has the most reviews to see how customers think of this

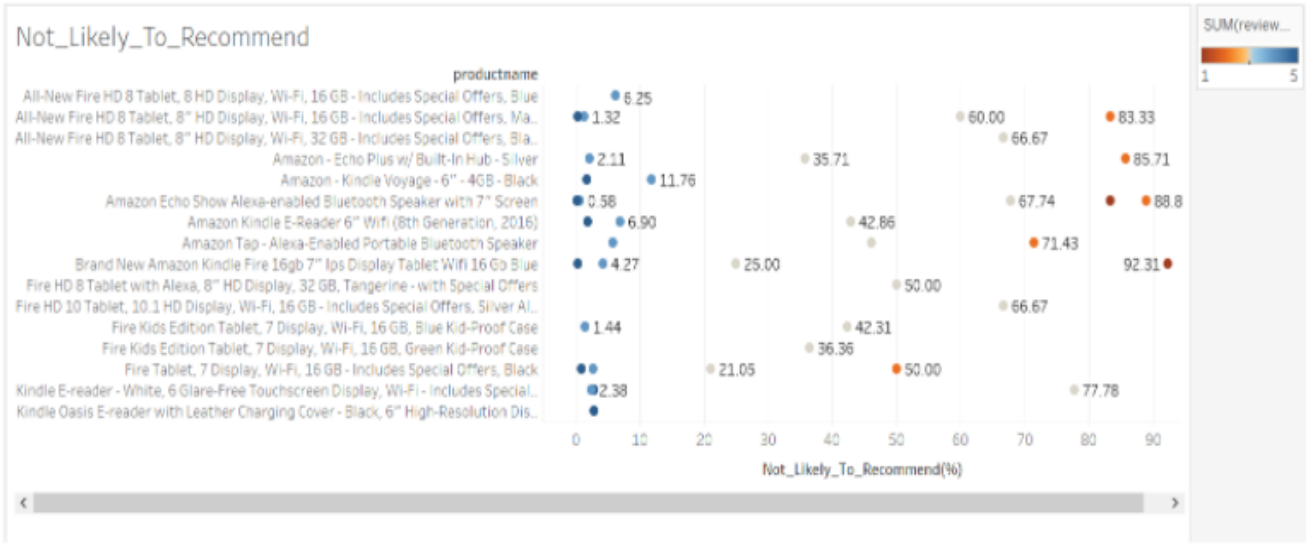


Figure 16. visual : categories of products

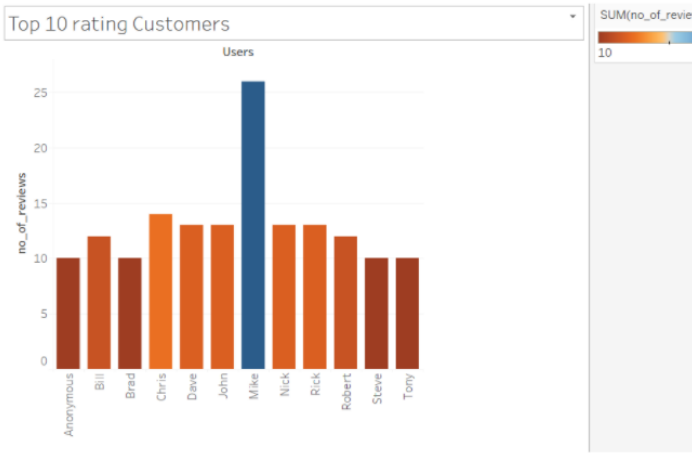


Figure 17. visual : Top 10 Rating customers

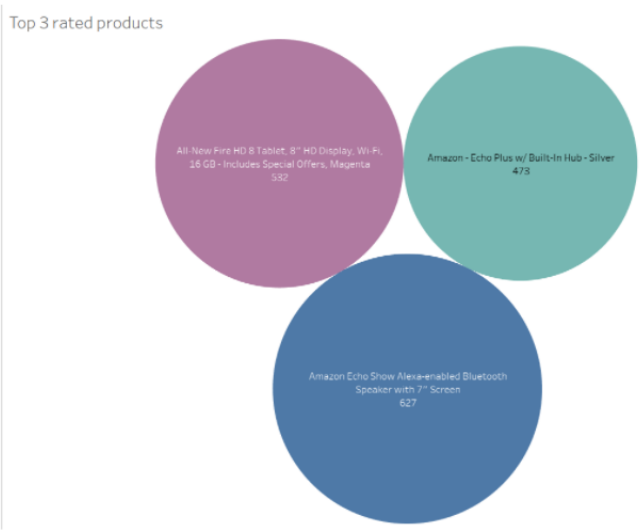


Figure 18. visual : Top 3 most 5 rated products

product on their review texts. Because we cannot read all the review texts one by one, we decide to do a word count method by using a function called CountVectorizer in the sklearn library in python to find the top 30 words that show up the most in the review text. Here are some examples of coding and the result.Refer fig 19

Based on the top 30 words out of 2190 words that show up in review, we can re-construct a sentence from all those words to see or guess what people write about the product review on Echo Show Alexa.Refer fig 20

```
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer
```

```
vectorizer = CountVectorizer()
matrix = vectorizer.fit_transform(df3['reviewstext'])
matrix
```

```
counts = pd.DataFrame(matrix.toarray(),
                      columns=vectorizer.get_feature_names())
s = counts.sum(axis = 0)
```

```
s = s.sort_values(ascending=False)
s.count()
```

```
2190
```

```
wordscount = s.head(30)
wordscount.to_csv('wordcount.csv',index = True)
```

Figure 19. visual : Top 30 word mostly used in reviews

- Amazon echo show alexa is great.
- We love the screen.
- The music and video is great on Amazon echo show alexa.

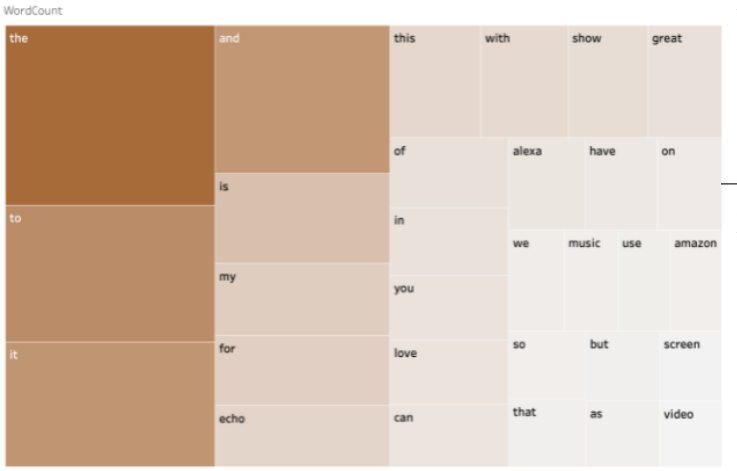


Figure 20. visual : Top 30 word mostly used in reviews

3. Queries :

Query Questions	Sql Queries
1)What is the total amount of reviews for each product from the most to the less?	Select productName , Count(productName) as Number_of_Product from Product Natural Join Associate Group by productName order by Number_of_Product Desc;
2)What is the average rating on each product?	Select productName , Avg(reviewsrating) as Avg_Rating from Product Natural Join Associate Natural Join Review Group by productName order by Count(name) Desc;
3) Year Wise listing the count of products added.	select productName, count(p.ProductName) product_count, year_date from product p, productdate pd, associate a ,(select distinct Product-DateID,SUBSTRING (dateAdded, 1,4) AS year_date from productdate) temp where a.ProductDateID= pd.ProductDateID and a.ProductID= p.ProductID AND temp.ProductDateID= a.ProductDateID group by temp.year_date, p.productName
4)Categorizing the products based on category,database holds Electronics products.	select C.categoriesID, primaryCategories, productID,count(*) as product_count from categories c, product p where c.categoriesID= p.categoriesID group by p.productName ,categoriesID
5)Year wise listing the products added	select p.productName, p.ProductID,year_date from product p ,productdate pd, associate a ,(select distinct ProductDateID, SUBSTRING(dateAdded, 1,4) AS year_date from productdate) temp where a.ProductDateID= pd.ProductDateID and a.ProductID= p.ProductID AND temp.ProductDateID= a.ProductDateID group by p.productName, temp.year_date;

6) Finding from the reviewers ratings to understand how likely the reviewers are to recommend the product in percentage	Select recom.productname, recom.reviewsrating, norecom.recommend_count /(norecom.recommend_count + recom.recommend_count) *100 as ” Not_Likely_To_Recommend (%)” ,recom.recommend_count /(norecom.recommend_count +recom.recommend_count)*100 as ”Likely_To_Recommend(%)” from (select p.productName as productname, reviewsrating, reviewsdoRecommend, count(*) as recommend_count from product p,review r,associate a where p.productID=a.ProductID and a.reviewID=r.reviewID and reviewsdoRecommend=0 group by p.productName, reviewsrating, reviewsdoRecommend order by p.productName, reviewsrating, reviewsdoRecommend desc) norecom, (select p.productName as productname, reviewsrating, reviewsdoRecommend, count(*) as recommend_count from product p,review r,associate a where p.productID= a.ProductID and a.reviewID= r.reviewID and reviewsdoRecommend=1 group by p.productName, reviewsrating,reviewsdoRecommend order by p.productName,reviewsrating, reviewsdoRecommend desc) recom where norecom.productname= recom.productname and norecom.reviewsrating= recom.reviewsrating order by productname, reviewsrating desc;
7)Top 10 Customers, based on their purchases and ratings records	select temp.usernameID, c.reviewsusername as Users, temp.no_of_reviews from customer c,(select usernameID,count(*) as no_of_reviews from review group by usernameID having no_of_reviews¿=10 order by no_of_reviews desc) temp where c.usernameID= temp.usernameID;
8)Top three rated products	SELECT distinct p.productName,count(*) as count,p.productID from review r,product p,associate a where a.ProductID=p.ProductID and a.reviewID=r.reviewID and r.reviewsrating=5 group by p.productName order by count desc limit 1;

4. Tools and Technologies:

Python pandas: We have made use of python pandas for data cleaning. We found that the coding below are useful for extracting data values with a format of regular expression.
In order to assign the same integer as ID number to a duplicated data such as ProductID and Product, we used the coding below. Ref fig 21

```
df['dateAdded'] = df['dateAdded'].str.extract(r'^(\s*\d+\s*\W+\s*\d+\s*\W+\s*\d+)\S*')

df.insert(loc=13,column='ProductDateID',value=df['dateAdded'].factorize()[0]+1)
df.insert(loc=15,column='reviewsDateID',value=df['reviewsdate'].factorize()[0]+1)
```

ProductID	productCode	name	asins
1	AVqVGZNvQMlgsOJE6eUY	Amazon Kindle E-Reader 6" Wifi (8th Generation...	B00ZV9PXP2
1	AVqVGZNvQMlgsOJE6eUY	Amazon Kindle E-Reader 6" Wifi (8th Generation...	B00ZV9PXP2

Figure 21. Python Tools used

Python sklearn: we have made use of python sklearn for data analysis such as LogisticRegression and CountVectorizer. Ref fig 22

Mysql workbench: We have used MYSQL workbench for database creation and

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
model = LogisticRegression()
model.fit(numTFTR, y_train)

LogisticRegression()

from sklearn.feature_extraction.text import CountVectorizer
workcount = CountVectorizer()
numTFTR = workcount.fit_transform(X_train)
numTFtest = workcount.transform(X_test)
```

Figure 22. Python Tools used

querying the tables, list of queries involving the table creation and data loading is attached in the git hub, link shared at the end of the report. This includes the DDL queries for database and table creation and DML queries to insert and query the tables. Ref fig 23

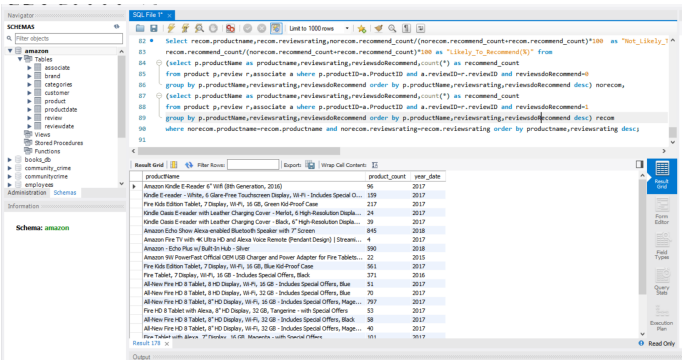


Figure 23. MYSQL workbench

Tableau for visualization: We used Tableau to analyze and visualize the data from Amazon database. All the above represented visual where extracted from Tableau for different kind of Custom SQL. Ref fig 24

Talend Data Integration (ETL Tool) : Talend has been used for data extraction, Transforming, and loading to the MySQL database. Some of the data cleaning has also been done with talend jobs. Job has been created in such a way that it takes all

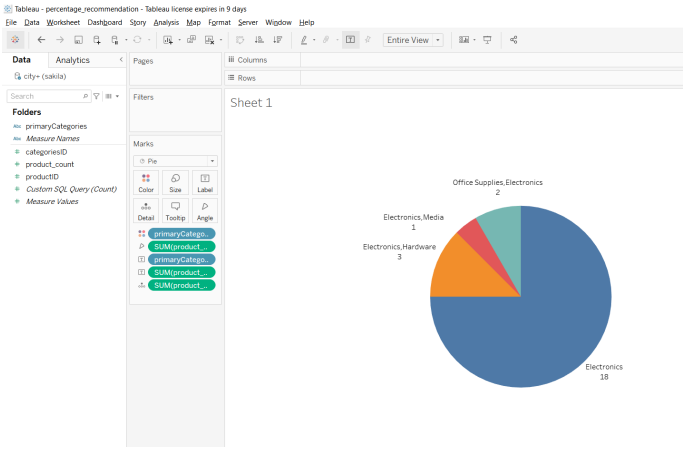


Figure 24. Tableau for visualization

the data csv file as input and transforms it according to requirement and loads it in the MySQL database. Talend job has been loaded in the repository. Ref fig 25

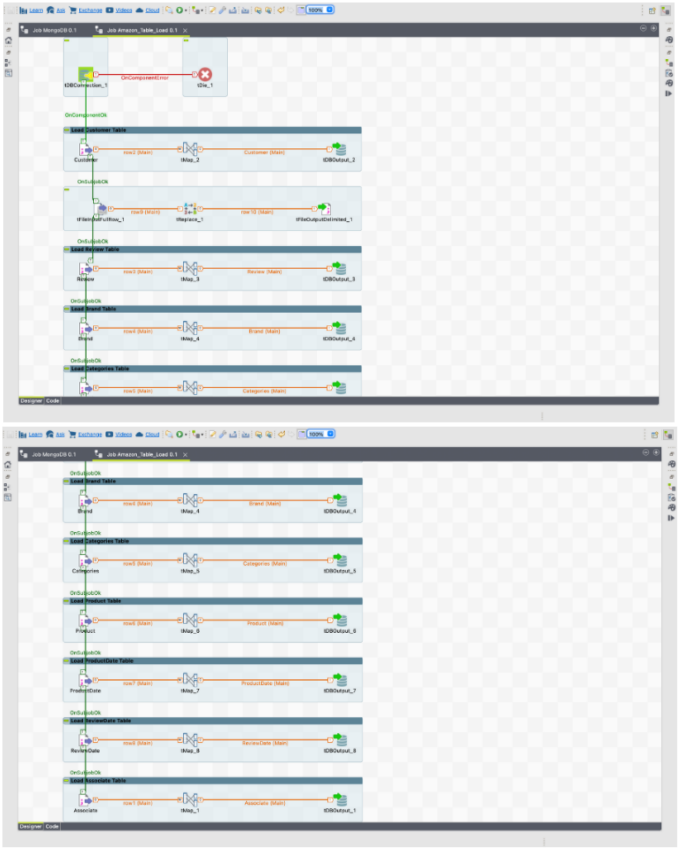


Figure 25. Talend Data Integration (ETL Tool)

MongoDB : Apart from MySQL, we

have loaded our data in no sql database as well which is MongoDB using JAVA. We have created one cluster on mongoDB cloud and loaded our amazon data in it using MongoDB compass. Ref fig 26 https://cloud.mongodb.com/

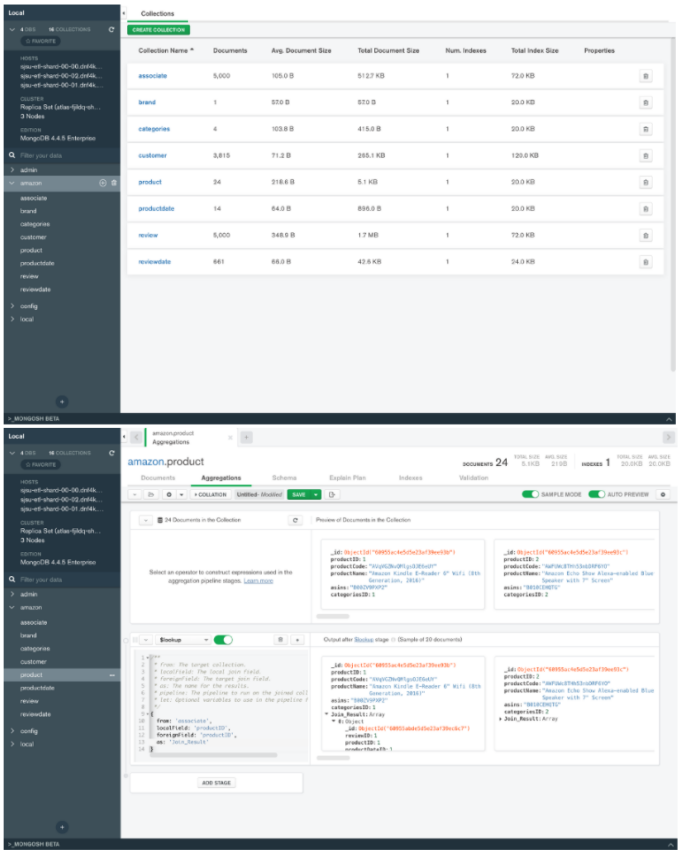


Figure 26. MongoDB

Queries:

1) Total amount of reviews for each product from the most to the less.Refer fig 27

[{$lookup: {from: 'associate',localField: 'productID',foreignField: 'productID',as: 'associate'}}, { $group: {_id: " $productName " ,numberOfProduct: {$sum:1}}}, { $sort: {numberOfProduct:-1}}, { $project: {"_id": 0,"productName": "$_id","numberOfProduct": 1}}]

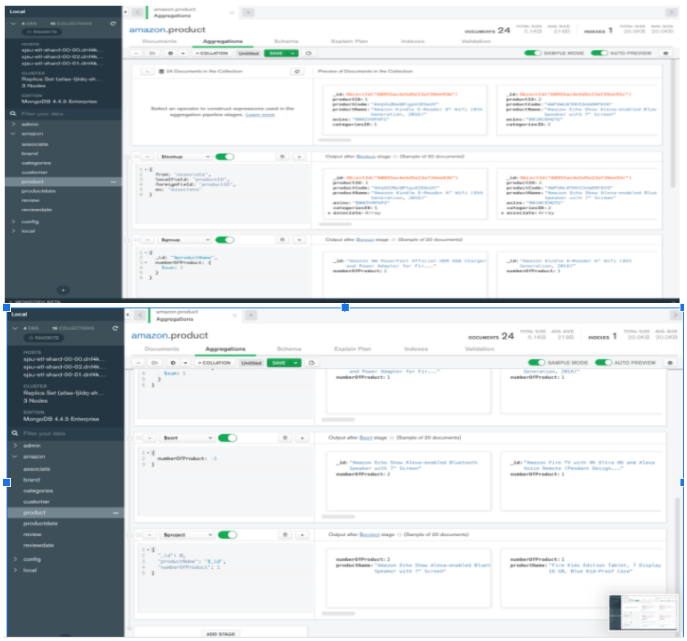


Figure 27. MongoDB Query1 Result

2) Average rating on each product ,Refer fig 28

[{$group: {_id: "$productName", numberOfProduct: {$sum: 1} }}, {$project: { "_id": 0,"productName": "$_id","numberOfProduct": 1}}, {$lookup: { from: 'product',localField: 'productName',foreignField: 'productName',as: 'products' }}, {$unwind: {path: "$products",preserveNullAndEmptyArrays: false}}, {$project: { "productID": "$products.productID","productCode": "$products.productCode", "productName": 1,"asins": "$products.asins","categoriesID": "$products.categoriesID", "numberOfProduct": 1}}, {$lookup: { from: 'associate',localField: 'productID',foreignField: 'productID',as: 'associates'}}, {$unwind: {path: "$associates",preserveNullAndEmptyArrays: false}}, {$lookup: { from: 'review',localField: 'associates.reviewID',foreignField: 'reviewID',as: 'reviews' }}, {$unwind: {path: "$reviews",preserveNullAndEmptyArrays: false}}, {$project: { "_id": 0, "productName": 1, "numberOfProduct": 1, "reviewsrating": "$reviews.reviewsrating"}}, {$group: {_id: {productName: "$productName", numberOfProduct: "$numberOfProduct"}, avgRating: { $avg: "$reviewsrating"}}}, {$project: {"_id": 0,"productName": "$_id.productName",

”numberOfProduct”: ”$_id.numberOfProduct”, ”avgRating”: 1}},

{$sort: {numberOfProduct: -1}}]

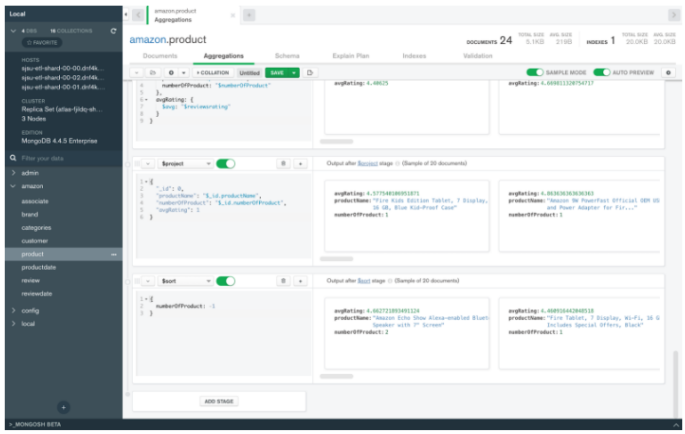


Figure 28. Latex

Latex : We used Latex for reporting and documentation, following the required IEEE format, Refer fig 29

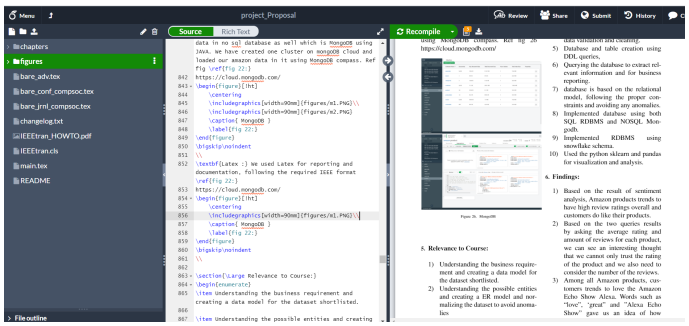


Figure 29. Latex

5. Relevance to Course:

1) Understanding the business requirement and creating a data model for the dataset shortlisted.
2) Understanding the possible entities and creating a ER model and normalizing the dataset to avoid anomalies
3) Cleaning of the data using ETL tool, and loading the data to the database using ETL tool.
4) Making use of python pandas for data validation and cleaning.
5) Database and table creation using DDL queries,
6) Querying the database to extract relevant information and for business reporting.
7) database is based on the relational model, following the proper constraints and avoiding any anomalies.
8) Implemented database using both SQL RDBMS and NOSQL Mongodb.
9) Implemented RDBMS using snowflake schema.
10) Used the python sklearn and pandas for visualization and analysis.

6. Findings:

1) Based on the result of sentiment analysis, Amazon products trends to have high review ratings overall and customers do like their products.
2) Based on the two queries results by asking the average rating and amount of reviews for each product, we can see an interesting thought that we cannot only trust the rating of the product and we also need to consider the number of the reviews.
3) Among all Amazon products, customers trends to love the Amazon Echo Show Alexa. Words such as “love”, ‘great” and ”Alexa Echo Show” gave us an idea of how

much the customers like this product. Based on the result, we can conclude that Echo Show Alexa is the best-selling product.

4) Based on the customers ratings and reviews analysis, It could be used in the recommendation system to better recommend a good quality products to other customers.
5) We are able to keep a track of all the products being added to the products catalog.
6) We are able to predict the sale of products based on the users ratings, the products with most rates are highly expected to to sell faster than the products which are rated low by other customers.

7. Technical difficulties and their solutions :

1) Data cleaning: Dataset included several, blank columns and special characters, hence transforming data was a challenge, to efficiently clean the data without data loss we used python pandas and ETL tools for cleaning and loading.
2) Identifier key: Few of the files were not having a unique identifier, we had to generate identifier keys for them.
3) Dataset: Initially we picked a different dataset, and the data wasn't accurate for analysis and reporting so we had to change the dataset to this Amazon customer review.

8. Members and their roles :

Members	Roles
Cheuk Ip Hong	Data cleaning and visualization through python
Yogita Suryavanshi	Data analysis, reporting and visualization through Tableau
Preeti Khatri	Data cleaning and loading through ETL and mongodb
Hrushikesh Pokala	ER modelling and reporting
Saroj Saran	data modelling and Analysis

9. Conclusion

Main objective of our project was to understand the back-end operations of a business involving large databases, hence picked one of the leading online retailer Amazon.com to understand the customer reviews processing and to understand how this data can be used further for Business Intelligence. As our project was a motivation from our Industry Case study "Amazon Go" and Significant paper presentation "Amazon recommendations", focusing on a single business helped us understand the different aspects of a business and how it is managed.

10. Acknowledgments

We would like to thank our professor, Dr Vishnu Pendyala for his guidance and supervision throughout the course and giving us this opportunity to understand the

project aspects and its use in real world, we would also like to extend our appreciation to Teaching Assistant Varun Valla for his support.

11. Sources

https://www.kaggle.com/datafiniti/
consumer-reviews-of-amazon-products

https://ieeexplore.ieee.org/abstract/document/
1167344?casa_token=nbY3WL2heqoAAAAA:
LmQ7G5TPffD0toVB0BLxP5c1Nlu6pFSl7Ixh9ypaR7CYHiZV_
BEPZ5z9wOIoT24nOc-_iXXd3RQ

https://towardsdatascience.com/
a-complete-sentiment-analysis-algorithm-in-python-with-amazon-product-review-data-step-by-step-2680d2e2c23b

12. Appendix

Criteria	Comments
Version Control (Use of git hub), including code and reports generated as part of this project	https://github.com/preetikhatrisjsu/DATA225_Project
Lessons learned	Learnings Listed
Technical Difficulties	Technical difficulties and their solutions Listed
Practiced agile / scrum	Yes, we used to have a recurring zoom call at 3PM everyday to get the teams update on task completed and what task to be taken next
Tools for language	Yes, we used Grammerly for the documentation
Used unique tools	yes, Used latex for IEEE report format
Includes DB Connectivity	Yes, used Python for DB connectivity, click for details DB Conectivity
Team Work	Team Members and their Roles Listed
NOSQL database	Used Mongo DB
Analytics Components	Data Visualization and Analytics
Code Walkthrough	Project Walk-through