MIS 6346

Big Data Class Project – Part 1

Shwitaan Sreenivas Ravikumar

SXR190013

Introduction and Problem Description

In this project the amazon review dataset will analysed and business case will be developed. The analytics is only based on the reviews given by the users. Thus is not completely true because not all customers give their reviews. The most important reviews considered are given by the vine members because these reviewers are chosen by the Amazon based on the reviews given by them and the accuracy of the ratings.

In this analysis the products are catagorized based on different parameters and analysed with respect to the parameters. Initially the dataset is loaded into hive as a table and by using Hive Query Language the required records are retrieved for analysis. At first, multiple customer reviews for the same product is deleted from the dataset because these reviews are not reliable and it becomes irrelevant.

In the first part of the analysis the work is done generally. The product categories considered are wireless, automotive, music, digital music purchase, sports, toys, digital video games, video games and only the products after 2005 is considered. In the second part, head to head analysis is done between 'Music' and 'Digital music' and between 'Video games' and 'Digital video games'.

The most important parameters considered are the star ratings and the product id. Because these parameters give information about the quality of the product and the most used products. The majority of the analysis is done based on these two parameters. The business models can be built based on these two parameters.

Vine membership also plays a significant part in the analysis. Because this benefit can change the behaviour of the customer and also the type of the products purchased. Products with good ratings given by vine members are brought at a larger scale as analysed in the dataset. Thus it would be more profitable if these products are stocked at warehouse.

The reviews are also being analysed based on the marketplace because the behavior of the customers change with respect to their culture and the type of place they reside. Thus some products would be very popular in US and it not be the favorite choice for Europeans. And some products are exclusively available to particular regions. The ratings with respect to marketplace is considered as a very important parameter especially if the particular product is available globally. Thus certain products must be customised based on the regions it is being sold.

Trend analysis is also done in this project which gives insight into the change in the customer behaviours over time. This shows the type of products which gained popularity and also we can predict the type of products which will be mostly sought after in the future. Trend analysis must be refered to produce the type of products which will cut the production cost and it will also reduce the wastage of products.

Step 1:

View of a table is created to consider the records from the year 2005:

```
CREATE VIEW amazon_review.amazon_reviews AS SELECT * FROM
amazon_review.amazon_reviews_parquet WHERE year >= 2005;
```

Step 2:

View is created with all the duplicate records removed:

```
create view amazon_review.amazon_reviews_include as
select s.marketplace,t.customer_id,s.review_id,t.product_id,s.product_parent,s.product_title,
s.star_rating,s.helpful_votes,s.total_votes,s.vine,s.verified_purchase,s.review_headline,
s.review_body,s.review_date,s.year,t.product_category
from amazon_review.amazon_reviews s
join (
   SELECT
   customer_id,product_id,product_category,
   COUNT(*)
FROM
   amazon_review.amazon_reviews
GROUP BY
   customer_id,product_id,product_category
HAVING
   COUNT(*) == 1
) t on s.customer_id = t.customer_id and
s.product_id = t.product_id and
s.product_category = t.product_category
```

Step 3:

Final table is created:

```
CREATE EXTERNAL TABLE amazon_review.amazon_reviews_v2(
  `marketplace` string,
  `customer_id` string,
  `review_id` string,
  `product_id` string,
  `product_parent` string,
  `product_title` string,
  `star_rating` int,
  `helpful_votes` int,
  `total_votes` int,
  `vine` string,
  `verified_purchase` string,
  `review_headline` string,
  `review_body` string,
  `review_date` DATE,
  `year` int)
PARTITIONED BY (
```

```
  `product_category` string)
--ROW FORMAT DELIMITED
--STORED AS PARQUET
ROW FORMAT SERDE
  'org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe'
STORED AS INPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat'
LOCATION
  'hdfs:///hive/amazon-reviews-pds/parquet/'
TBLPROPERTIES (
  'transient_lastDdlTime'='1583454851');
```

Step 4:

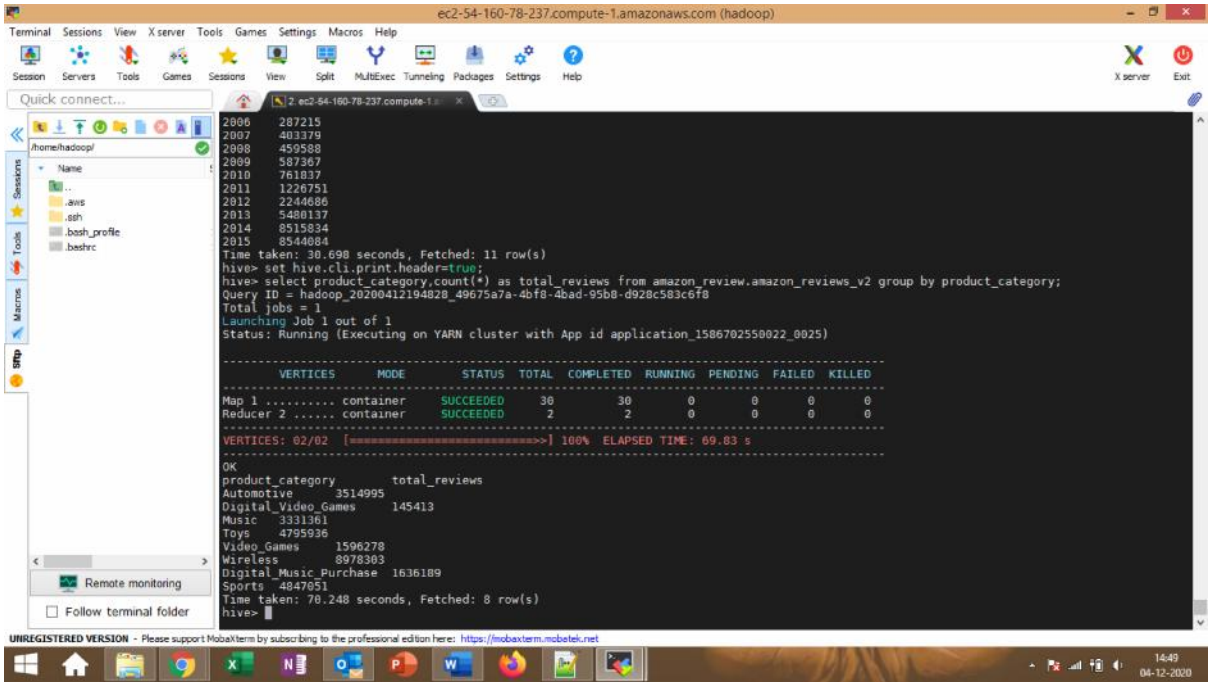
Records are inserted into the table. Example is given below

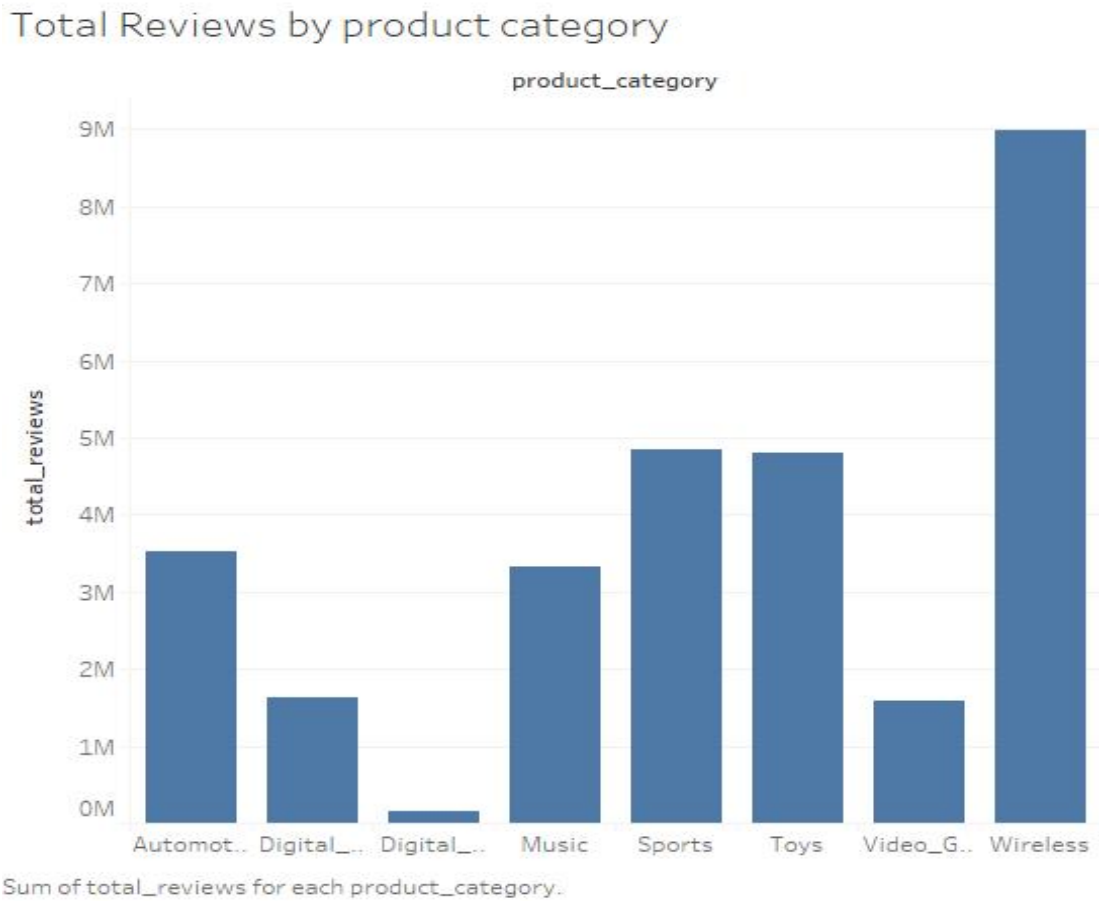
```
insert overwrite table amazon_review.amazon_reviews_v2 partition(product_category='Wireless')
select marketplace,customer_id,review_id,product_id,product_parent,product_title,star_rating,
helpful_votes,total_votes,vine,verified_purchase,review_headline,review_body,review_date,year
from amazon_review.amazon_reviews_include where product_category='Wireless';
```

Basic exploratory analysis:

No. of reviews by category

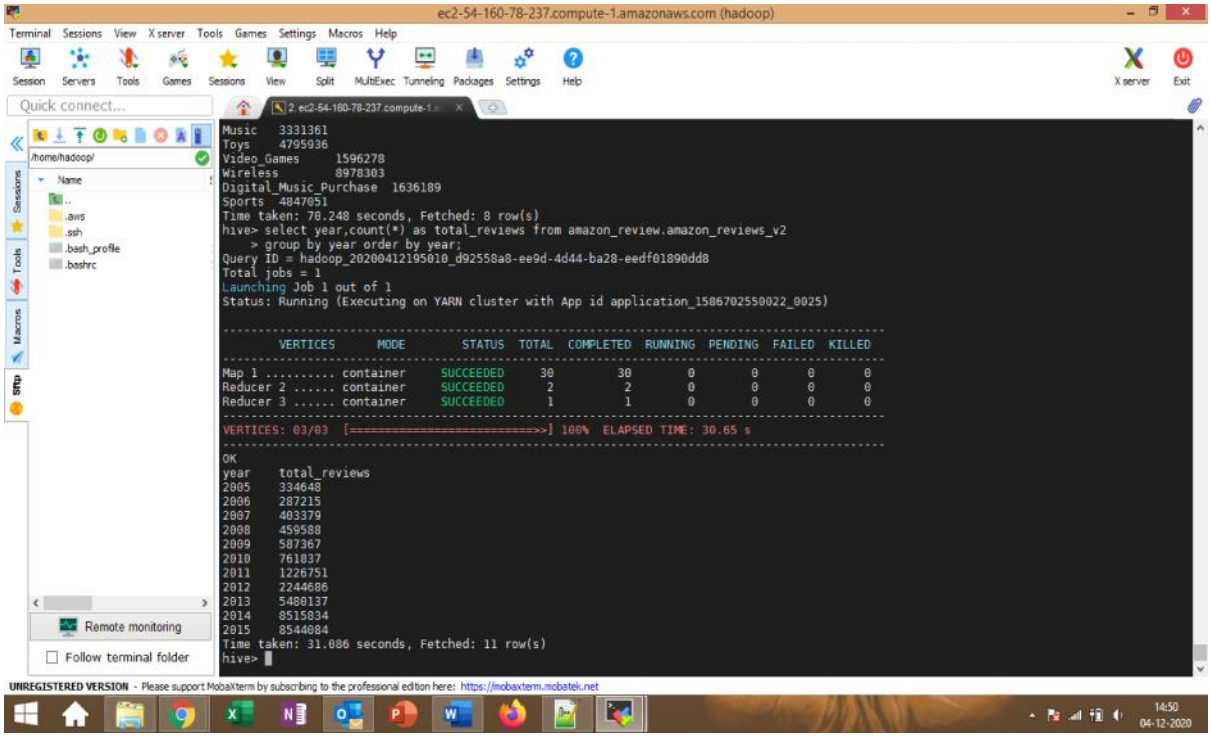
select product_category,count(*) as total_reviews from amazon_review.amazon_reviews_v2 group by product_category;

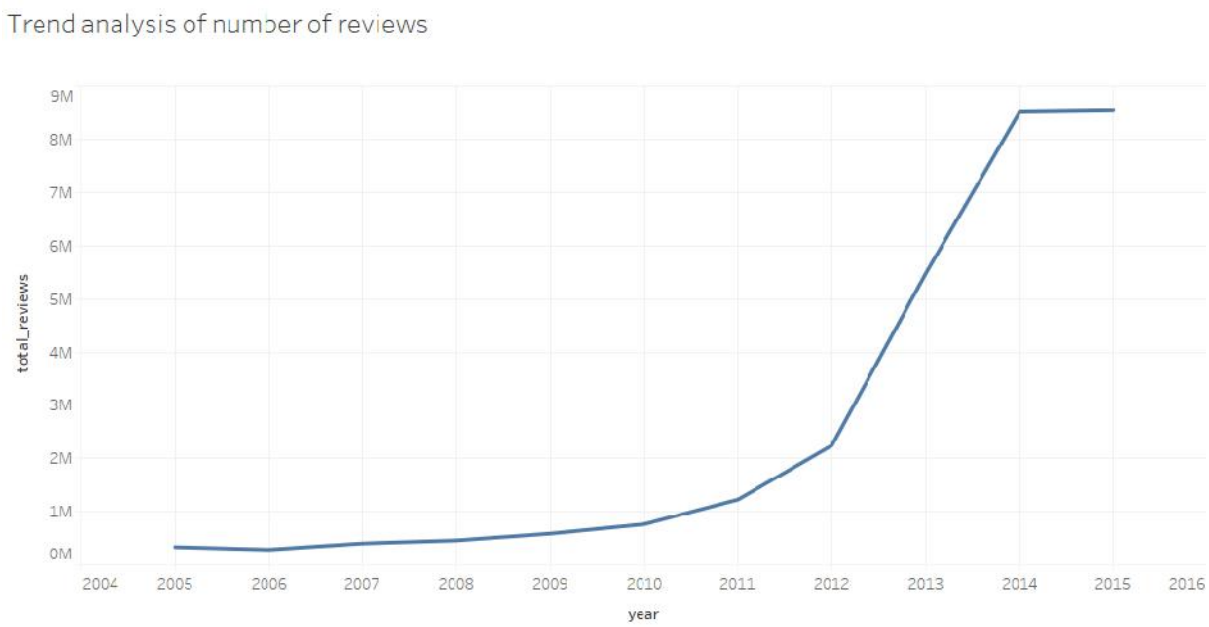




Trend analysis of no.of reviews

select year,count(*) as total_reviews from amazon_review.amazon_reviews_v2
group by year order by year;

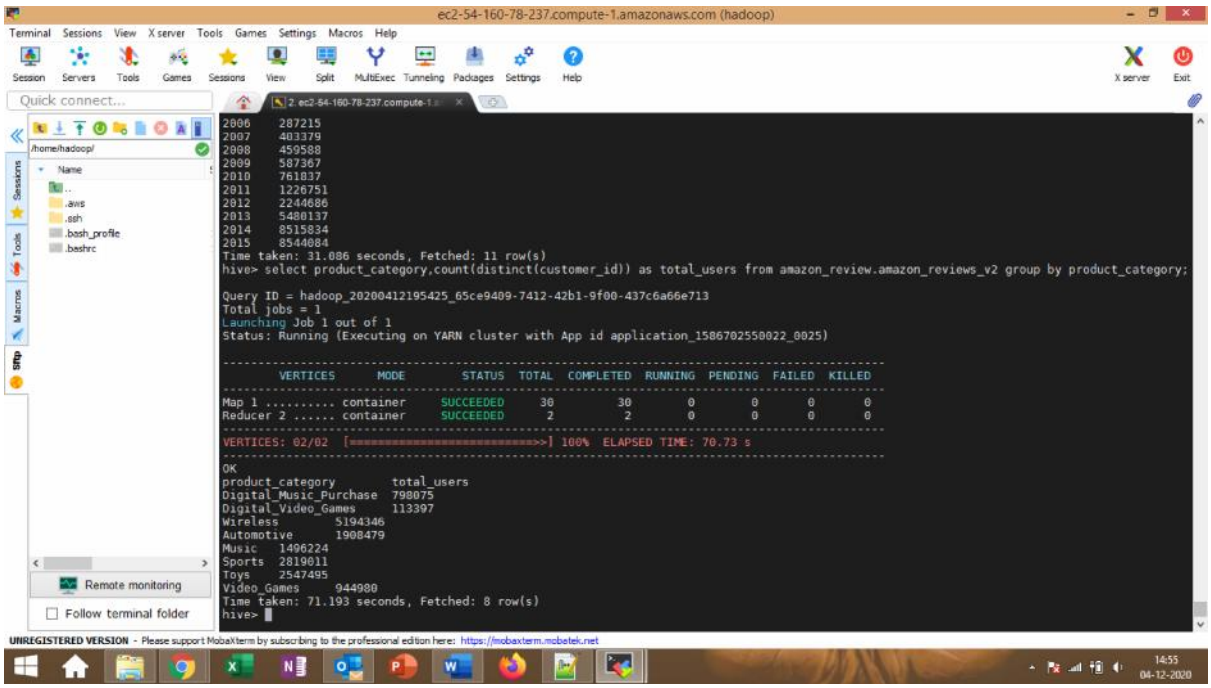


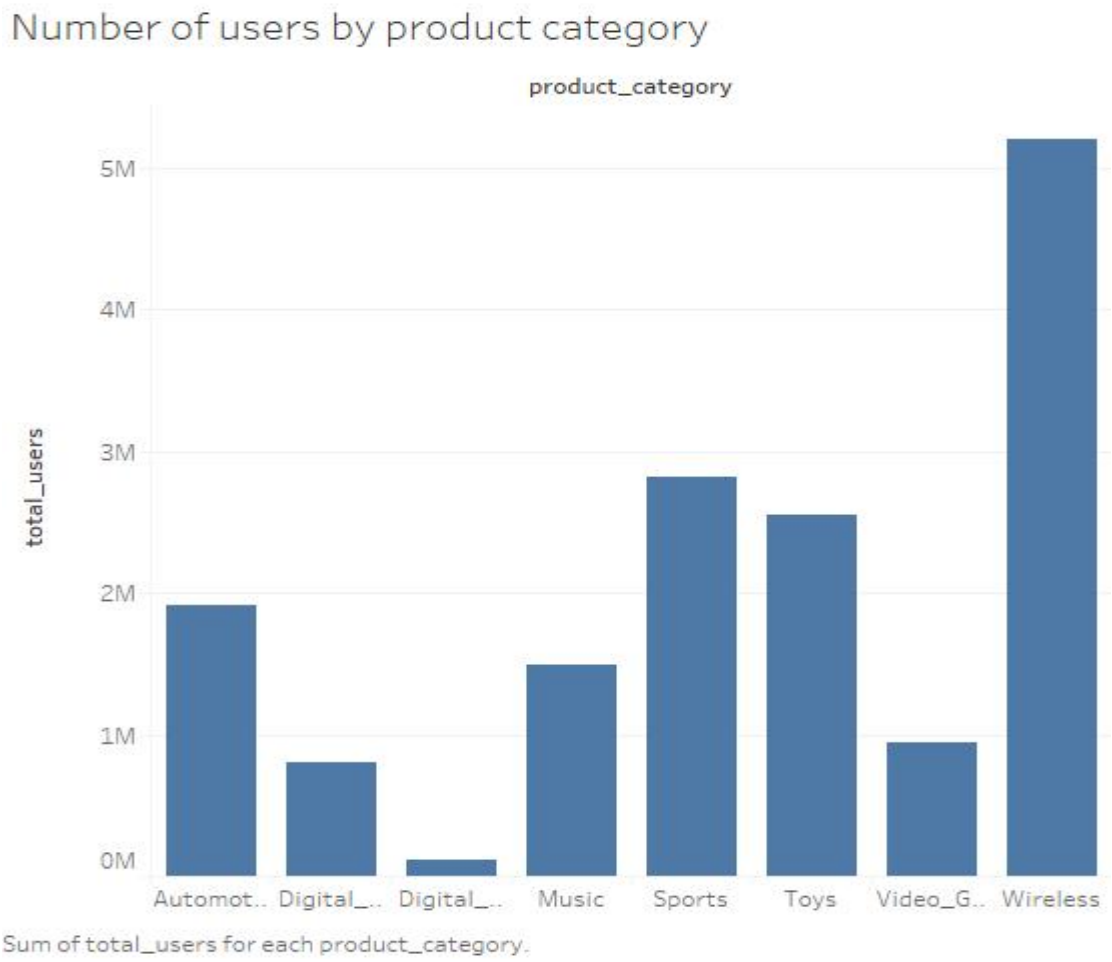


The trend of sum of total_reviews for year.

No. of users by product_category

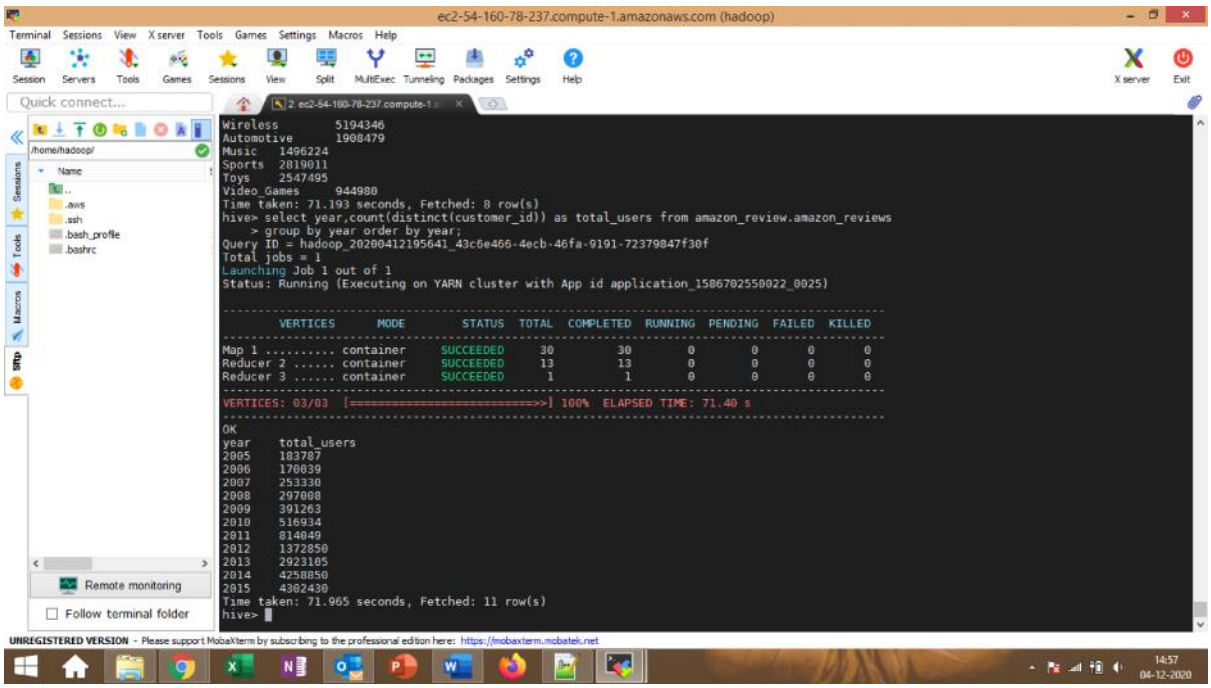
select product_category,count(distinct(customer_id)) as total_users from amazon_review.amazon_reviews_v2 group by product_category;

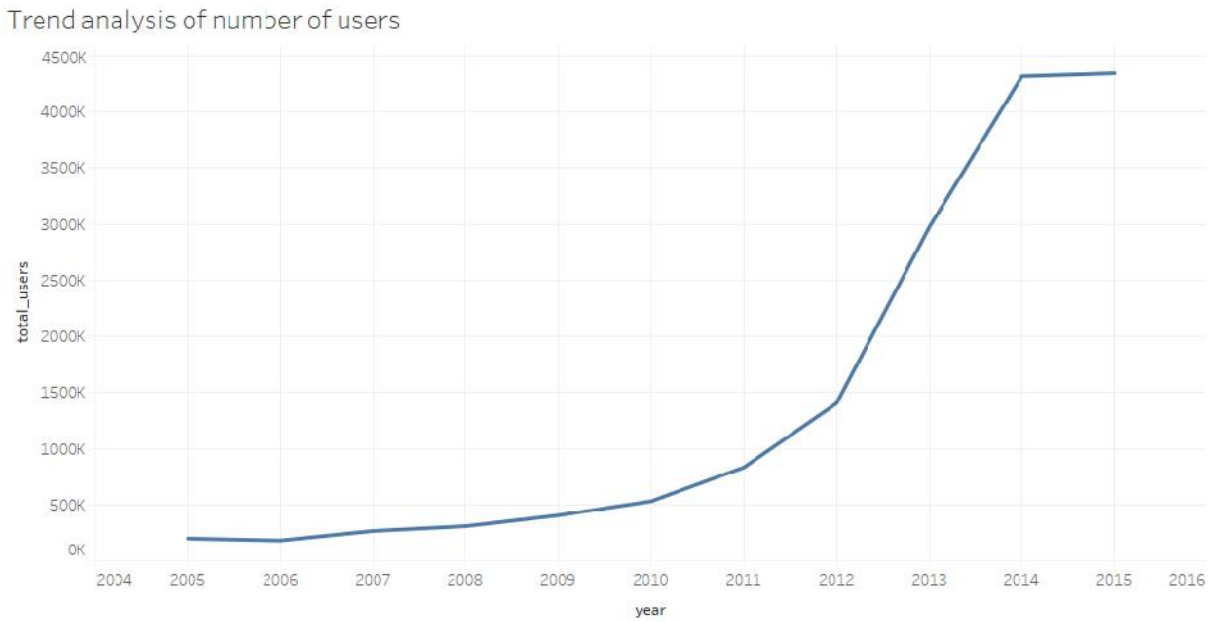




Trend analysis of no.of users

select year,count(distinct(customer_id)) as total_users from amazon_review.amazon_reviews group by year order by year;

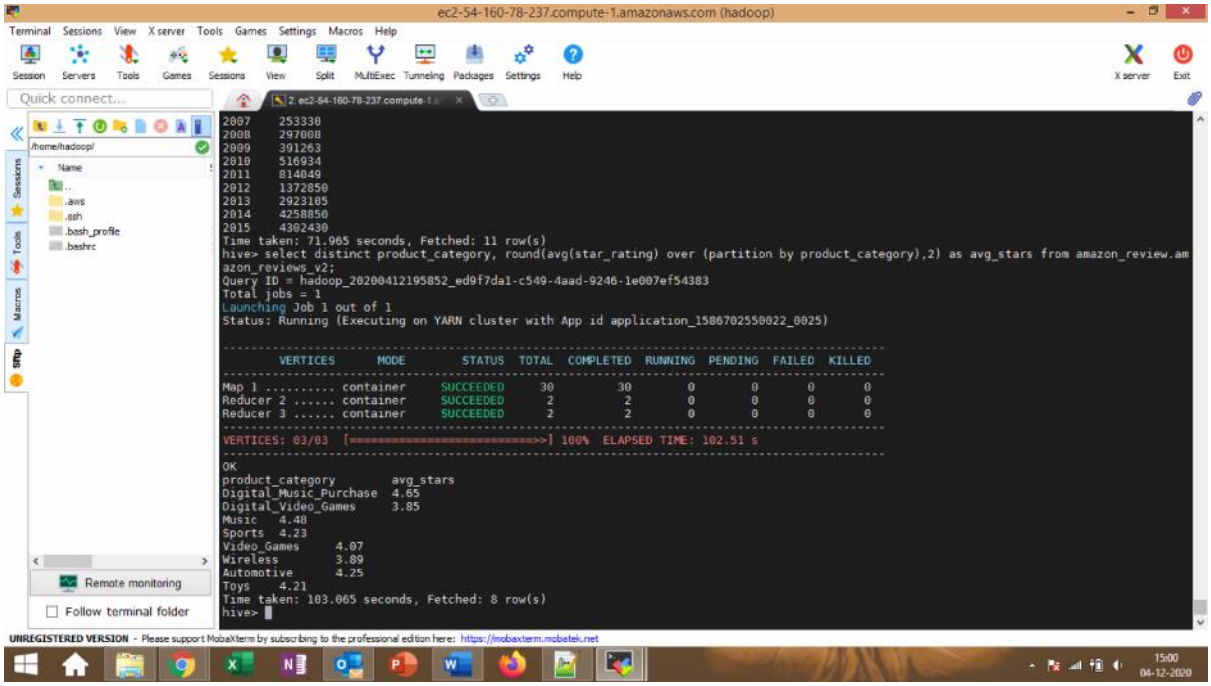


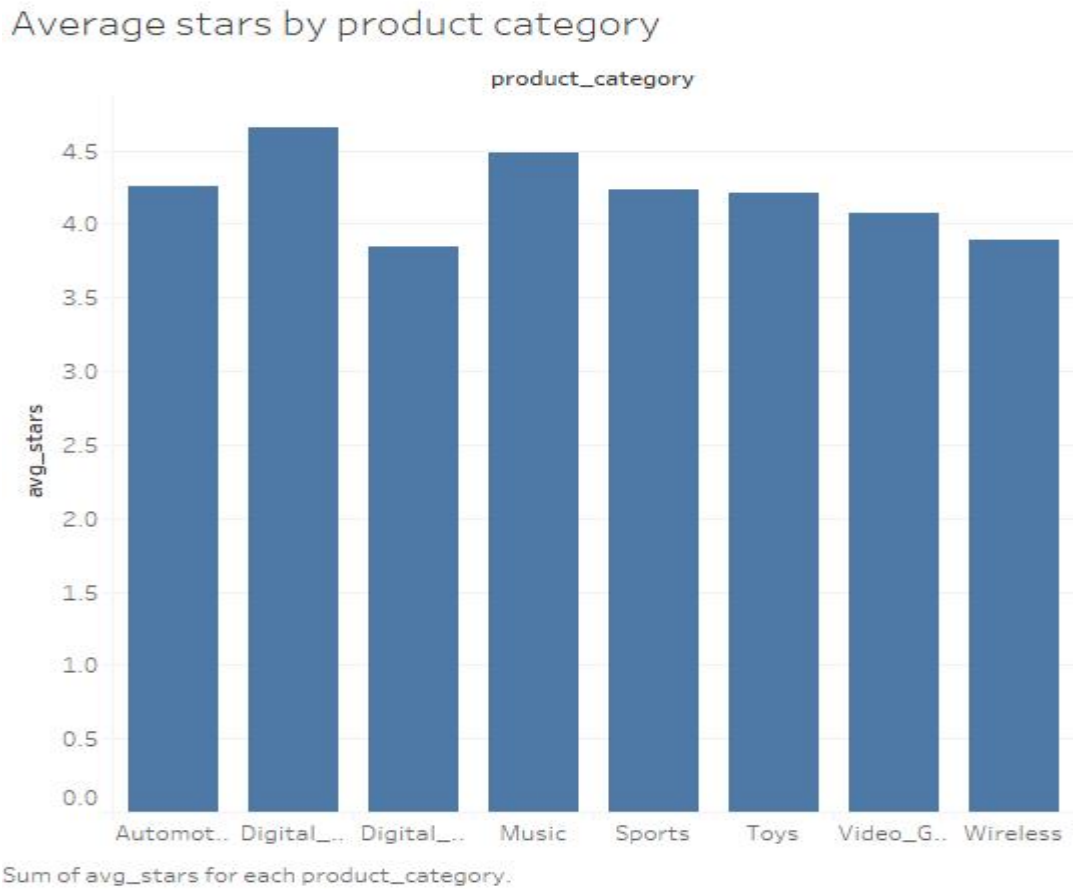


The trend of sum of total_users for year.

Average stars by product_category

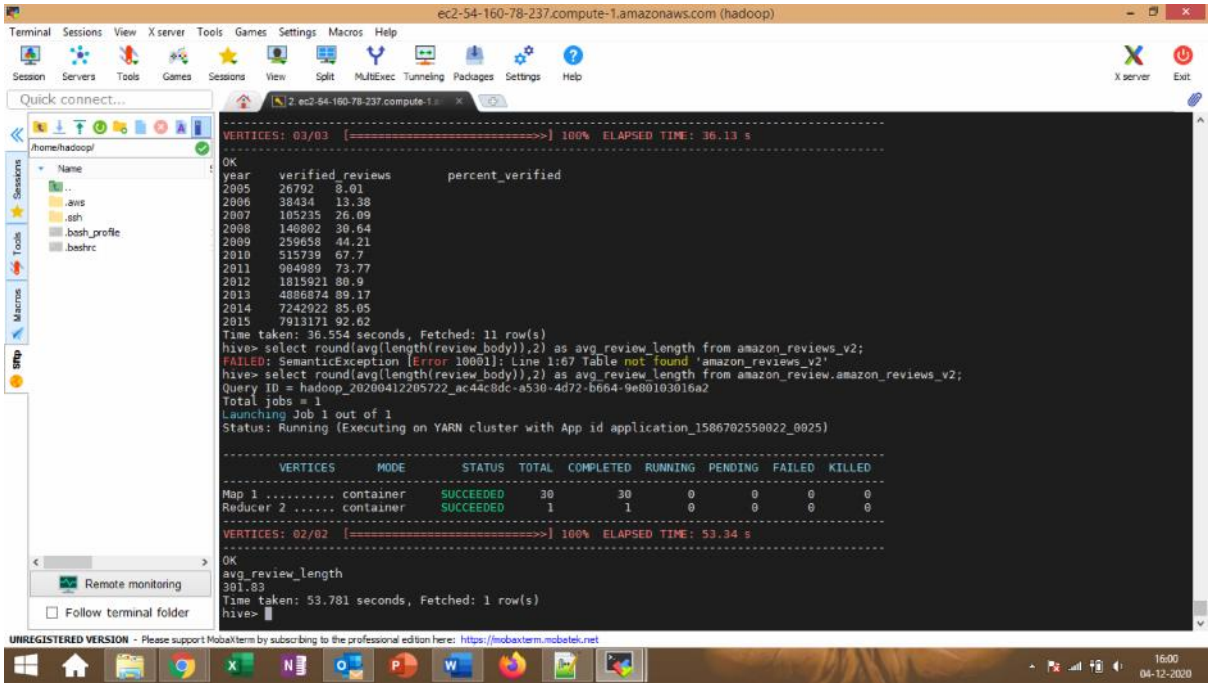
select distinct product_category, round(avg(star_rating) over (partition by product_category),2) as avg_stars from amazon_review.amazon_reviews_v2;





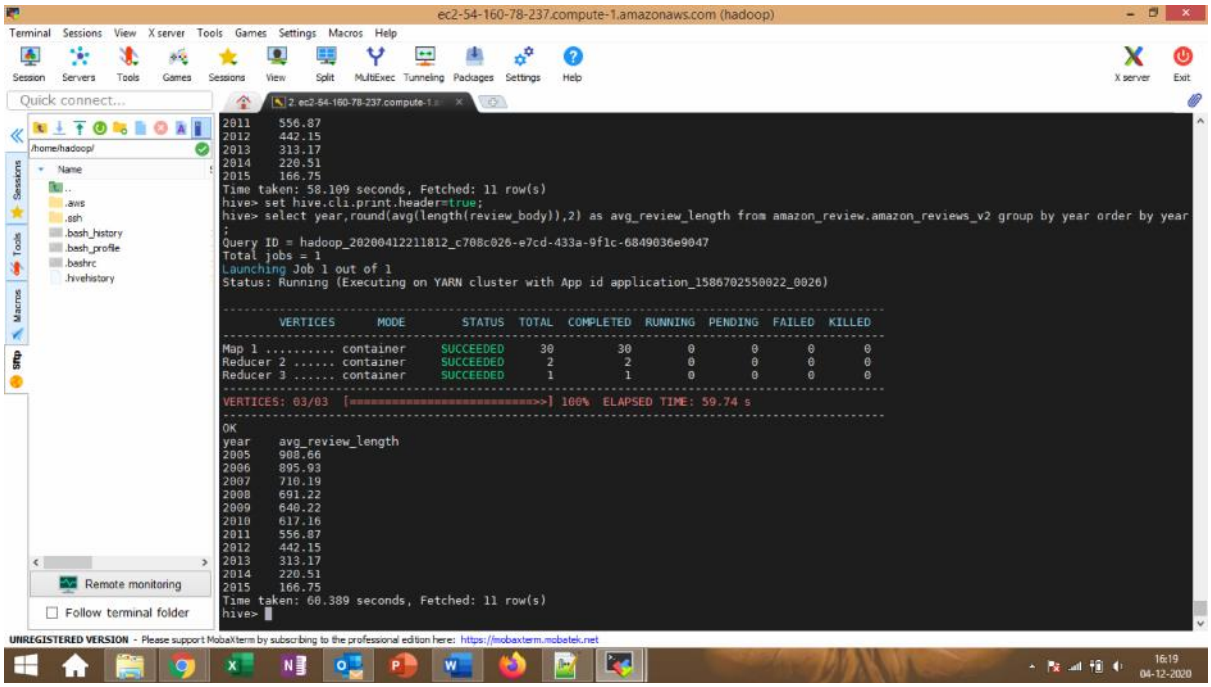
Average length of review

select round(avg(length(review_body)),2) as avg_review_length from amazon_review.amazon_reviews_v2;



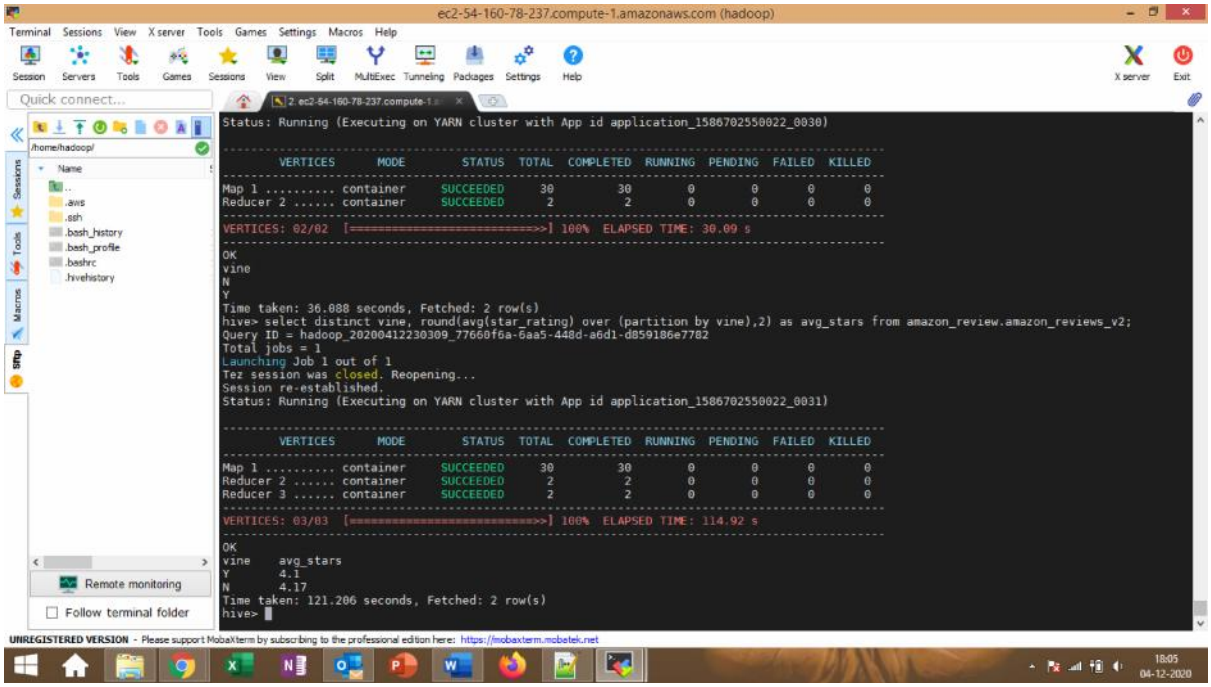
Trend analysis of average length of review

select year,round(avg(length(review_body)),2) as avg_review_length from amazon_review.amazon_reviews_v2 group by year order by year;



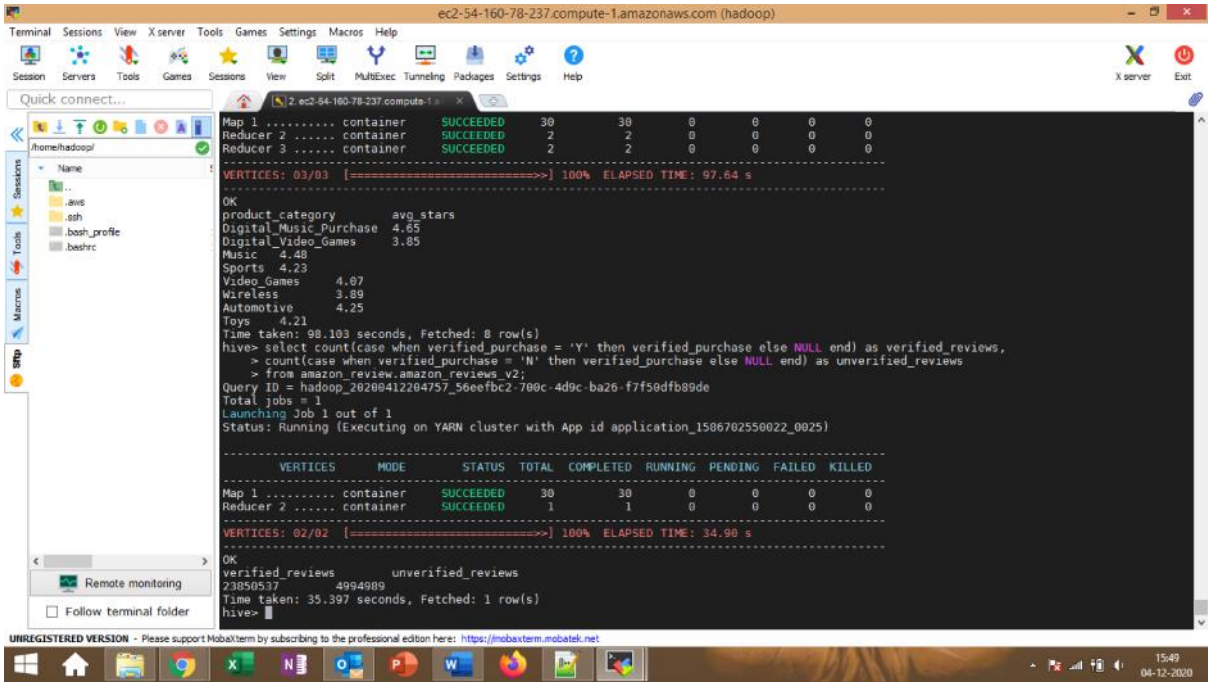
Average stars by vine membership

select distinct vine, round(avg(star_rating) over (partition by vine),2) as avg_stars from amazon_review.amazon_reviews_v2;



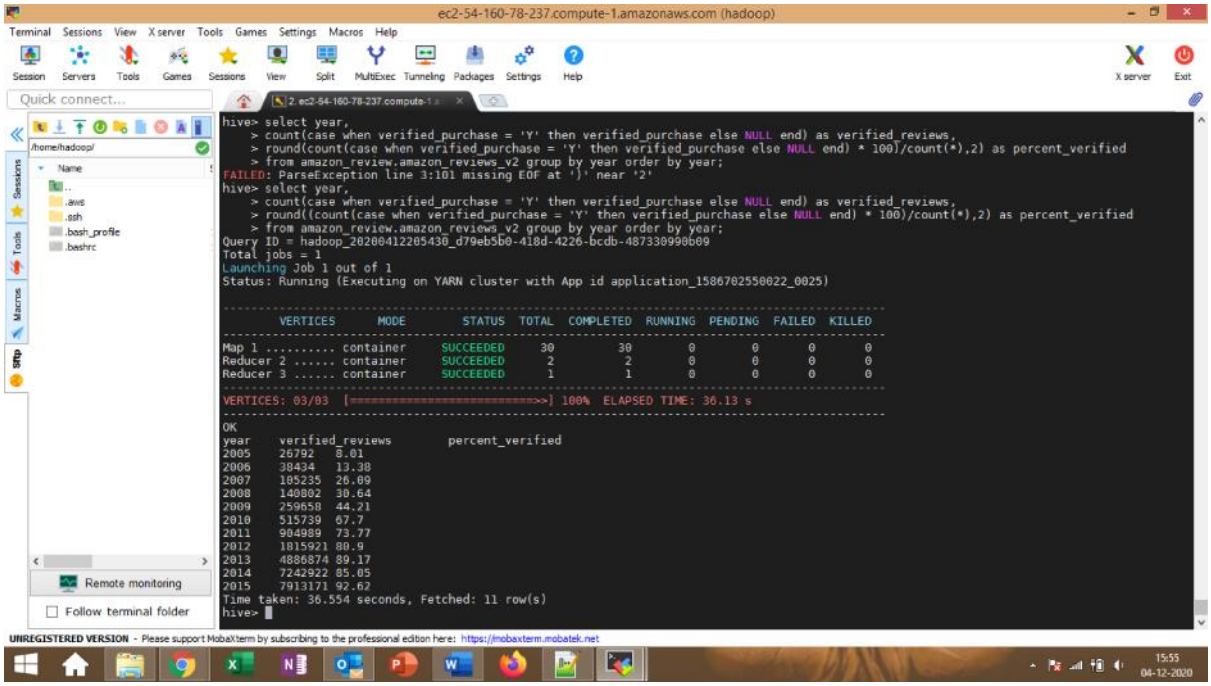
Number of verified versus unverified reviews

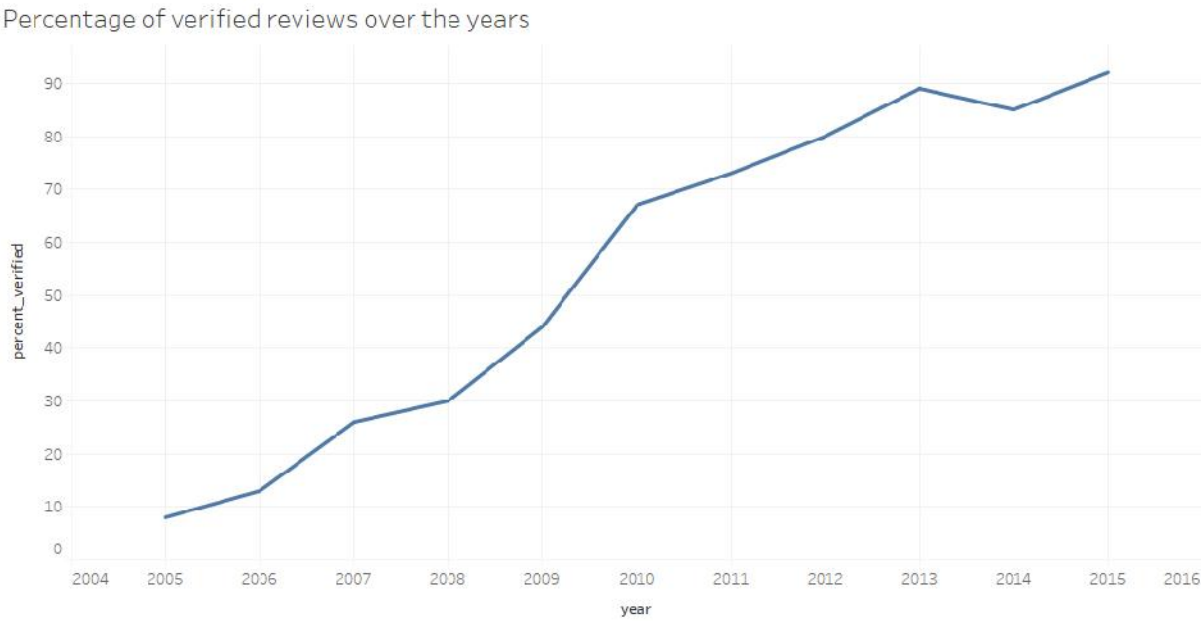
select count(case when verified_purchase = 'Y' then verified_purchase else NULL end) as verified_reviews, count(case when verified_purchase = 'N' then verified_purchase else NULL end) as unverified_reviews from amazon_review.amazon_reviews_v2;



Trend analysis of no. of verified reviews

select year, count(case when verified_purchase = 'Y' then verified_purchase else NULL end) as verified_reviews, round((count(case when verified_purchase = 'Y' then verified_purchase else NULL end) * 100)/count(*),2) as percent_verified from amazon_review.amazon_reviews_v2 group by year order by year;

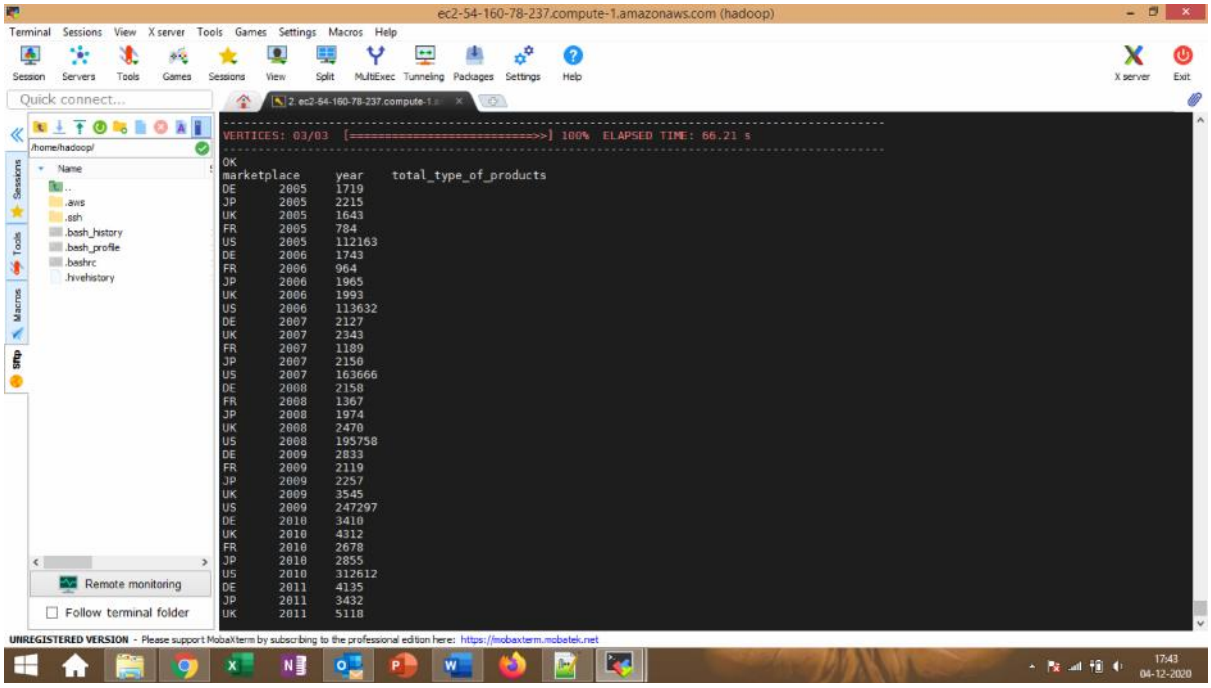


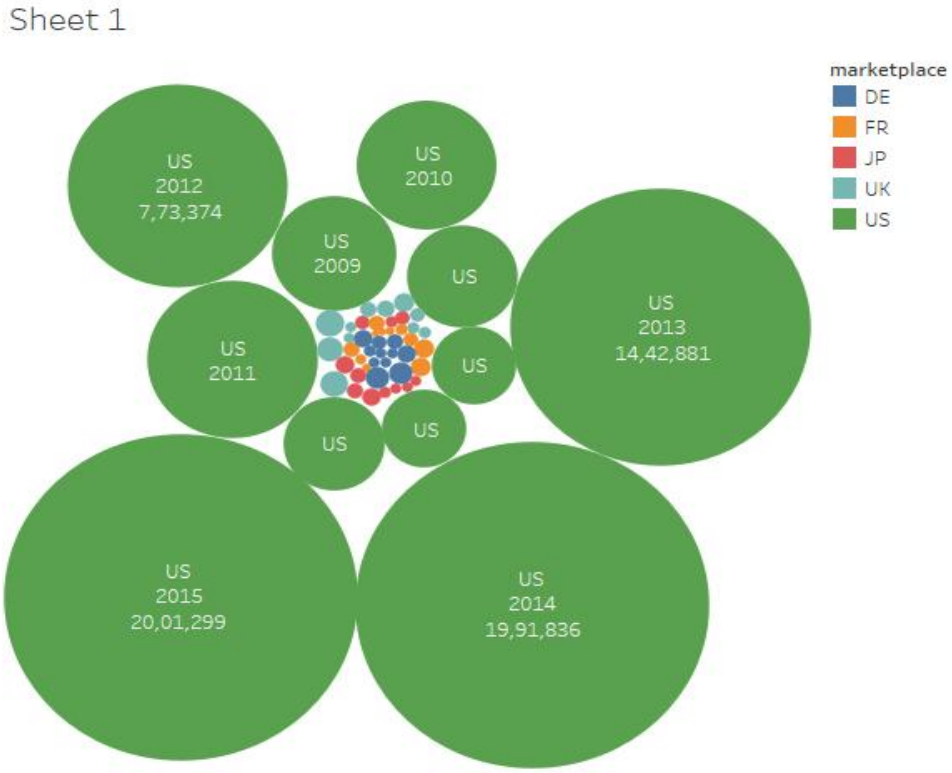


The trend of sum of percent_ver fied for year.

Trend analysis of no. of different types of products by marketplace

select marketplace,year,count(distinct(product_id)) as total_type_of_products from amazon_review.amazon_reviews_v2 group by marketplace,year order by year;

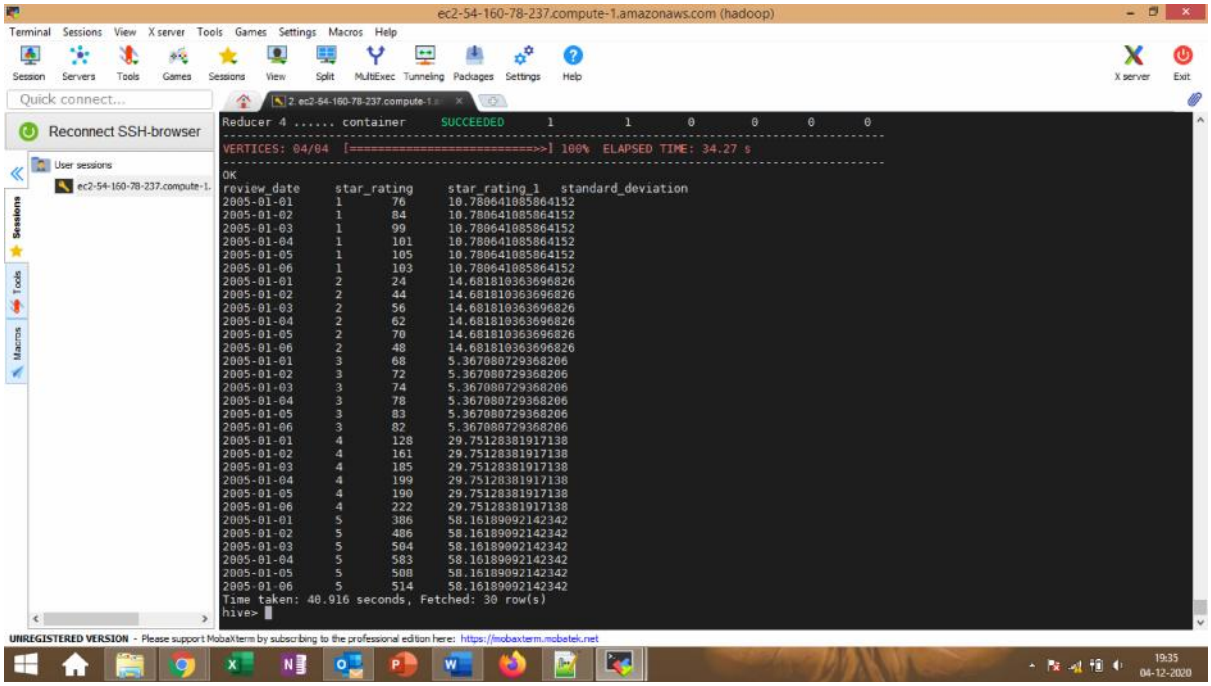




Marketplace, year and sum of total_type_of_products. Colour shows details about marketplace. Size shows sum of total_type_of_products. The marks are labelled by marketplace, year and sum of total_type_of_products.

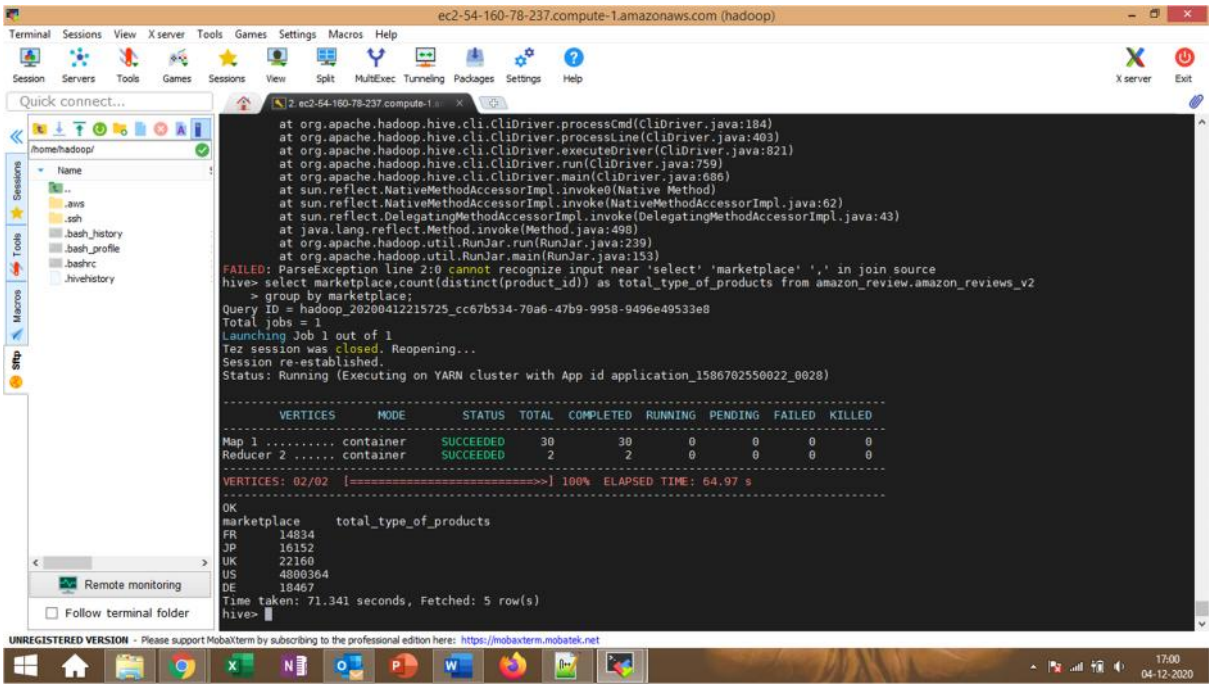
Standard deviation in star ratings

select review_date,star_rating,star_rating_1,stddev(star_rating_1) over (partition by star_rating order by review_date,star_rating asc rows between unbounded preceding and unbounded following) as standard_deviation from (select review_date,star_rating,count(star_rating) as star_rating_1 from amazon_review.amazon_reviews_v2 group by review_date,star_rating order by review_date,star_rating asc limit 30)s;



No. of different types of products by marketplace

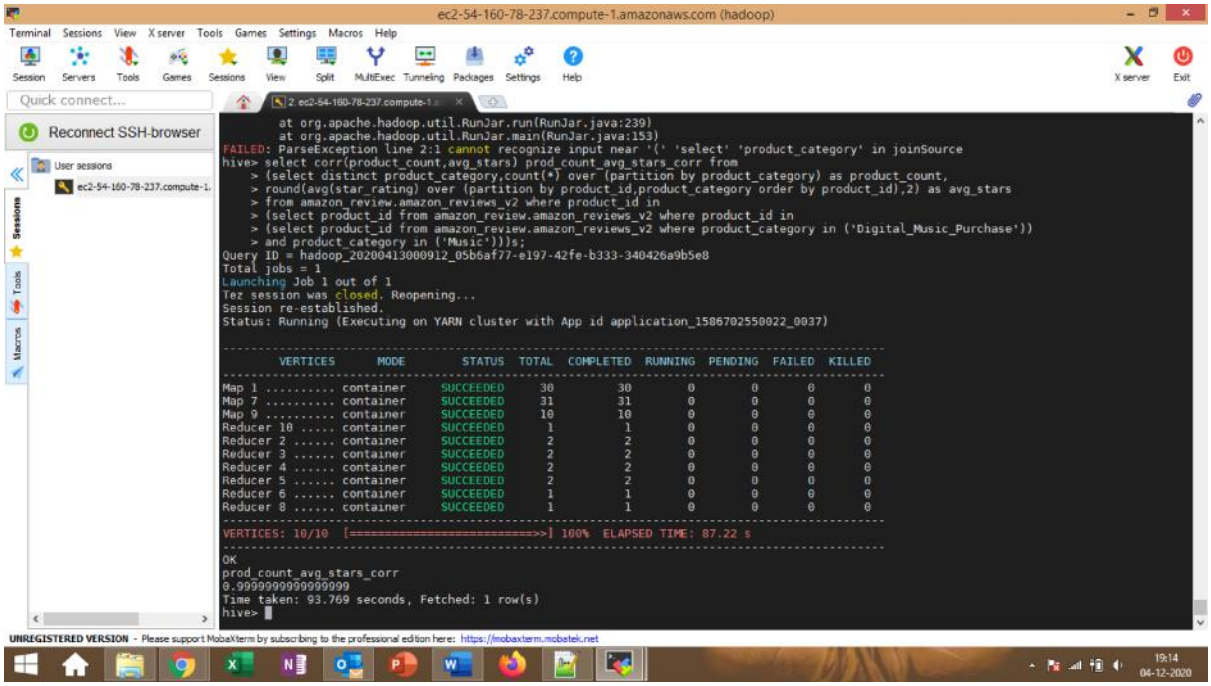
select marketplace,count(distinct(product_id)) as total_type_of_products from amazon_review.amazon_reviews_v2 group by marketplace;



Detailed analysis of Music/Digital_Music_Purchase and Digital_Video_Games/Video_Games over time.

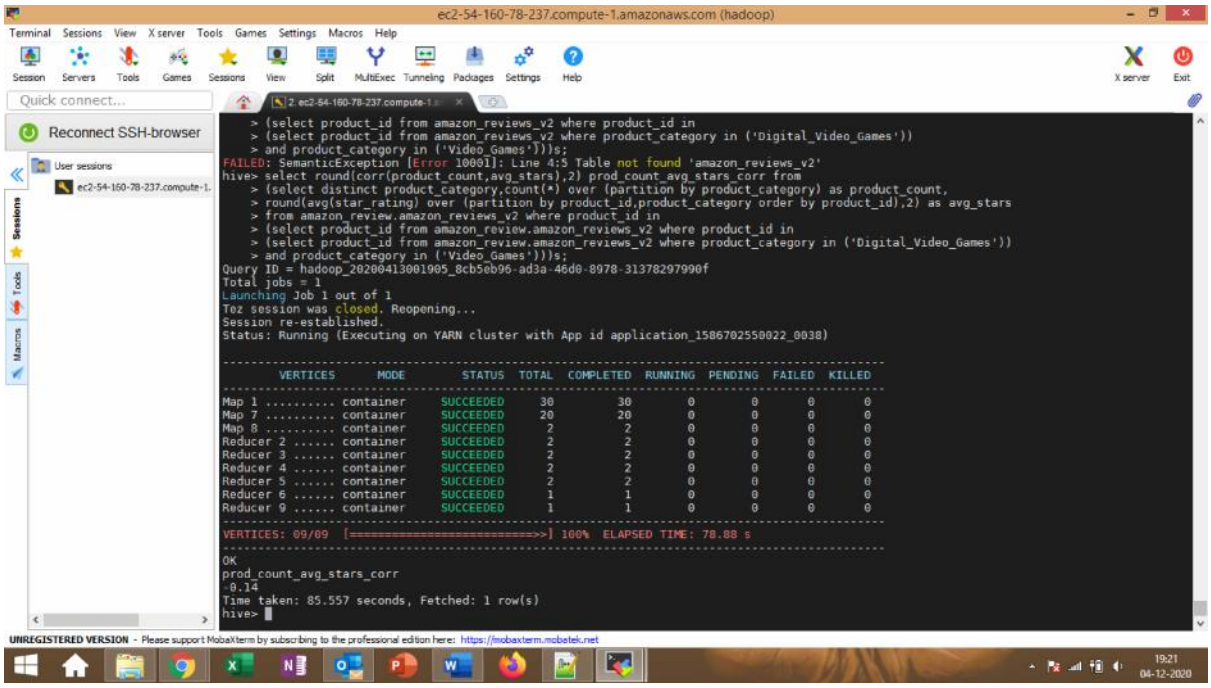
Correlation between Music and Digital_Music_Purchase

select corr(product_count,avg_stars) prod_count_avg_stars_corr from (select distinct product_category,count(*) over (partition by product_category) as product_count, round(avg(star_rating) over (partition by product_id,product_category order by product_id),2) as avg_stars from amazon_review.amazon_reviews_v2 where product_id in (select product_id from amazon_review.amazon_reviews_v2 where product_id in (select product_id from amazon_review.amazon_reviews_v2 where product_category in ('Digital_Music_Purchase')) and product_category in ('Music')))s;



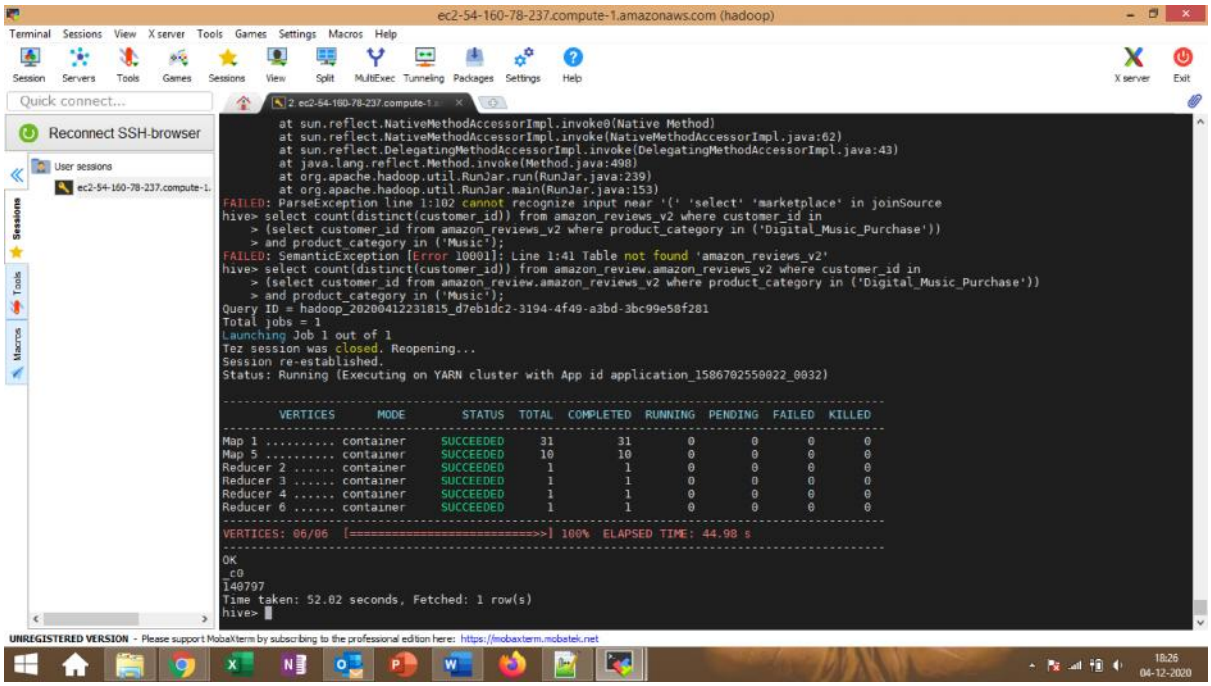
Correlation between Digital_Video_Games and Video_Games

select round(corr(product_count,avg_stars),2) prod_count_avg_stars_corr from (select distinct product_category,count(*) over (partition by product_category) as product_count, round(avg(star_rating) over (partition by product_id,product_category order by product_id),2) as avg_stars from amazon_reviews_v2 where product_id in (select product_id from amazon_reviews_v2 where product_id in (select product_id from amazon_reviews_v2 where product_category in ('Digital_Video_Games')) and product_category in ('Video_Games')))s;



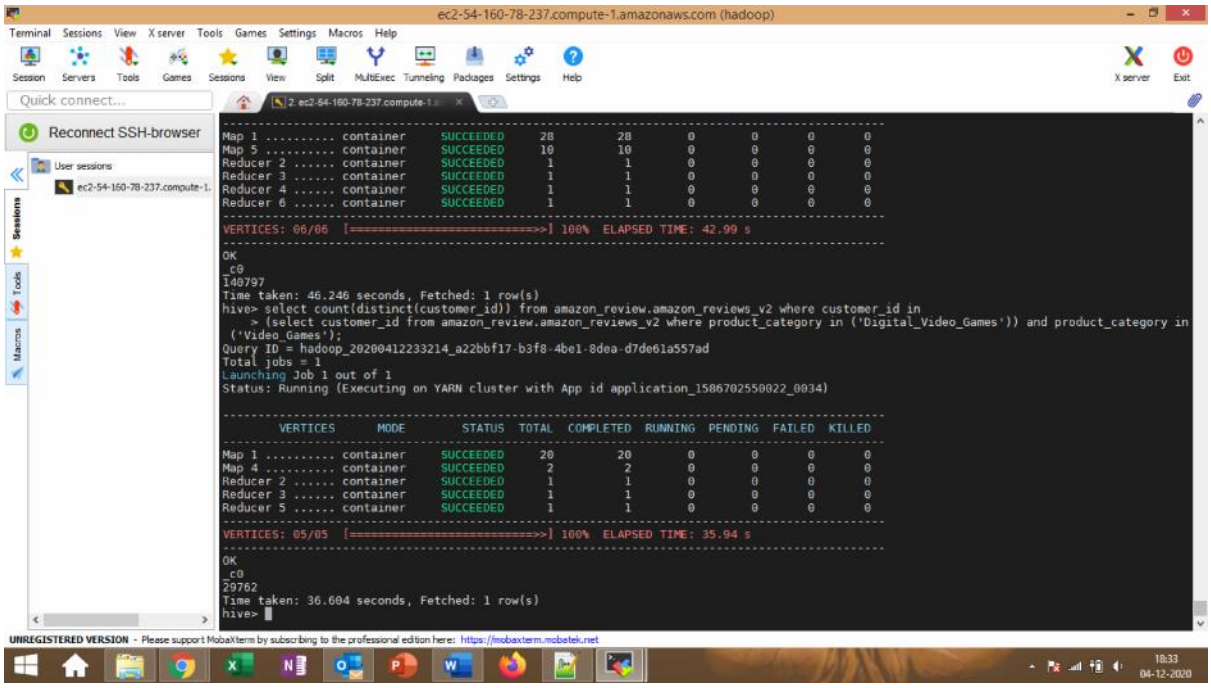
Total number of users reviewing in both Music and Digital_Music_Purchase

select count(distinct(customer_id)) from amazon_review.amazon_reviews_v2 where customer_id in (select customer_id from amazon_review.amazon_reviews_v2 where product_category in ('Digital_Music_Purchase')) and product_category in ('Music');



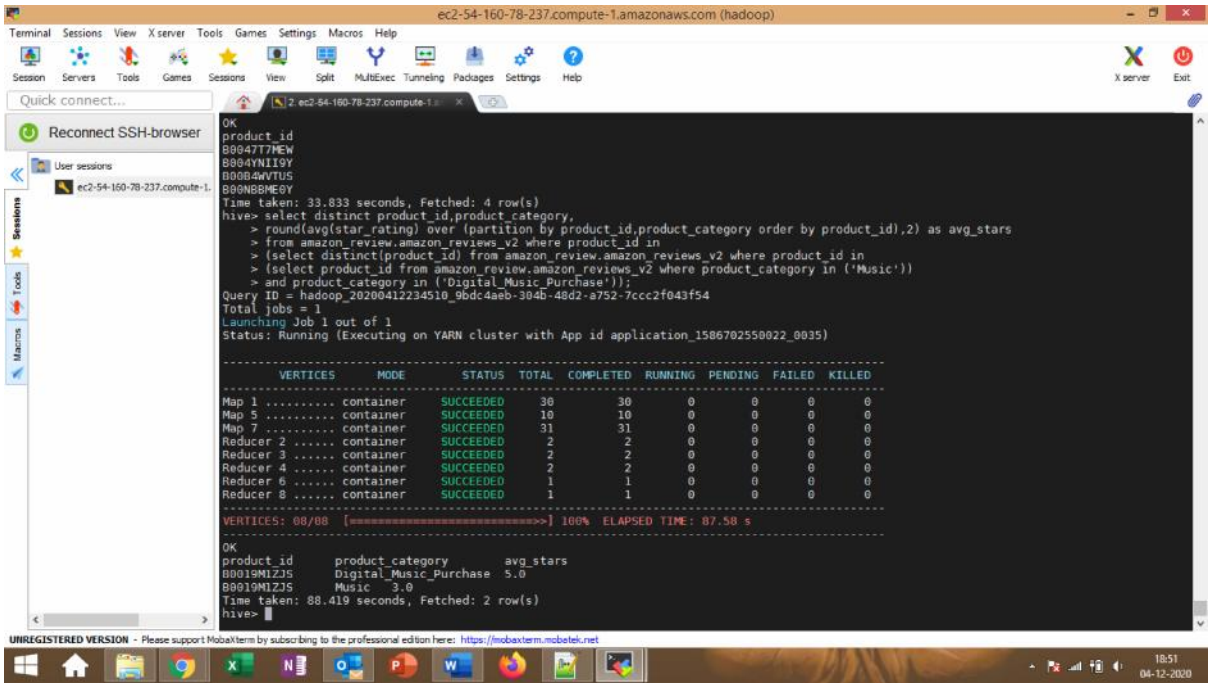
Total number of users reviewing in both Digital_Video_Games and Video_Games

select count(distinct(customer_id)) from amazon_review.amazon_reviews_v2 where customer_id in (select customer_id from amazon_review.amazon_reviews_v2 where product_category in ('Digital_Video_Games')) and product_category in ('Video_Games');



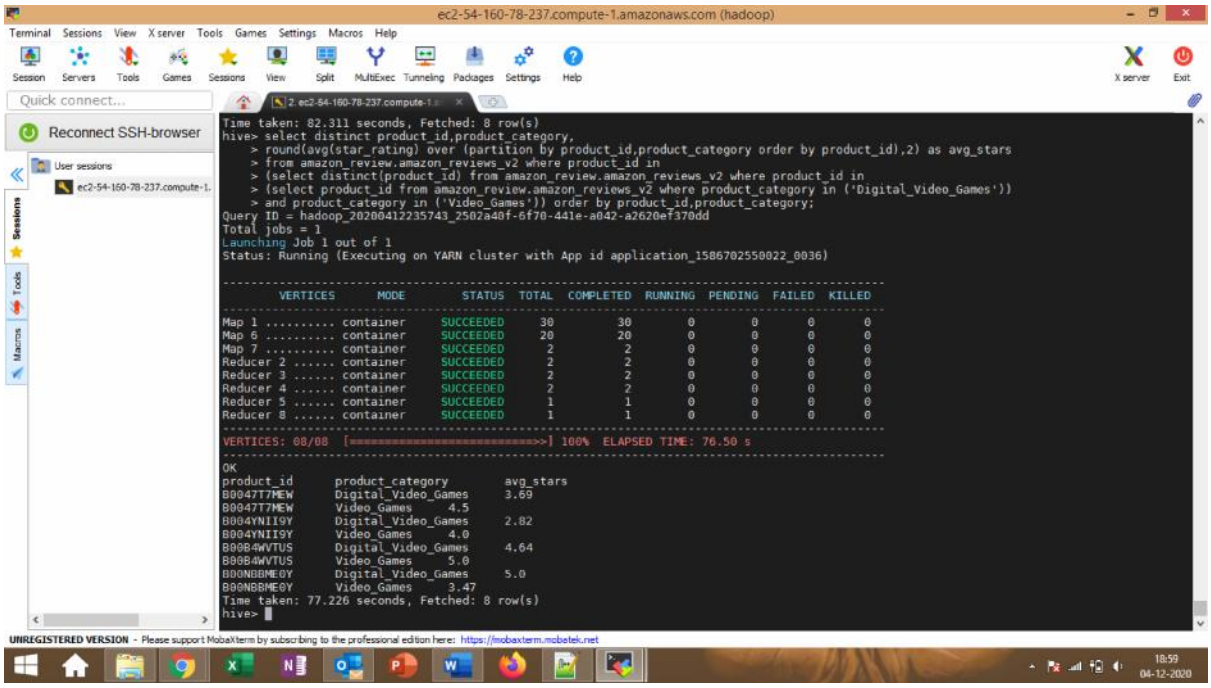
Average rating of similar products in Music and Digital_Music_Purchase

select distinct product_id,product_category, round(avg(star_rating) over (partition by product_id,product_category order by product_id),2) as avg_stars from amazon_review.amazon_reviews_v2 where product_id in (select distinct(product_id) from amazon_review.amazon_reviews_v2 where product_id in (select product_id from amazon_review.amazon_reviews_v2 where product_category in ('Music')) and product_category in ('Digital_Music_Purchase'));



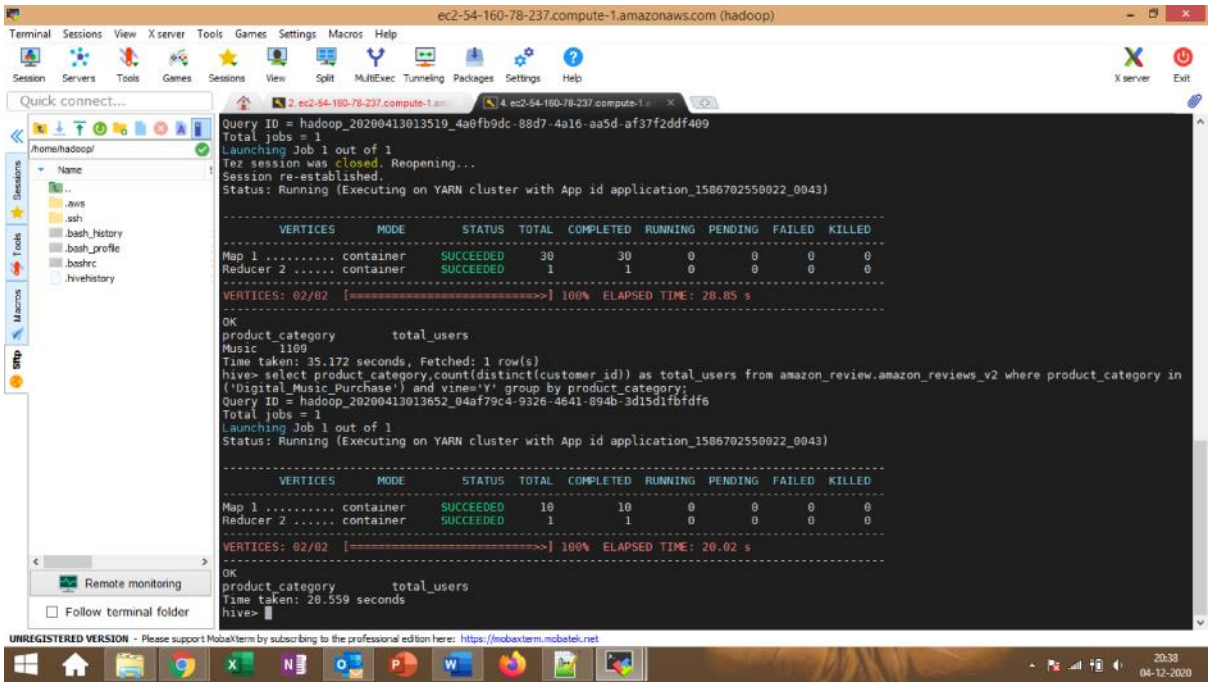
Average rating of similar products in Digital_Video_Games and Video_Games

select distinct product_id,product_category, round(avg(star_rating) over (partition by product_id,product_category order by product_id),2) as avg_stars from amazon_review.amazon_reviews_v2 where product_id in (select distinct(product_id) from amazon_review.amazon_reviews_v2 where product_id in (select product_id from amazon_review.amazon_reviews_v2 where product_category in ('Digital_Video_Games')) and product_category in ('Video_Games')) order by product_id,product_category;



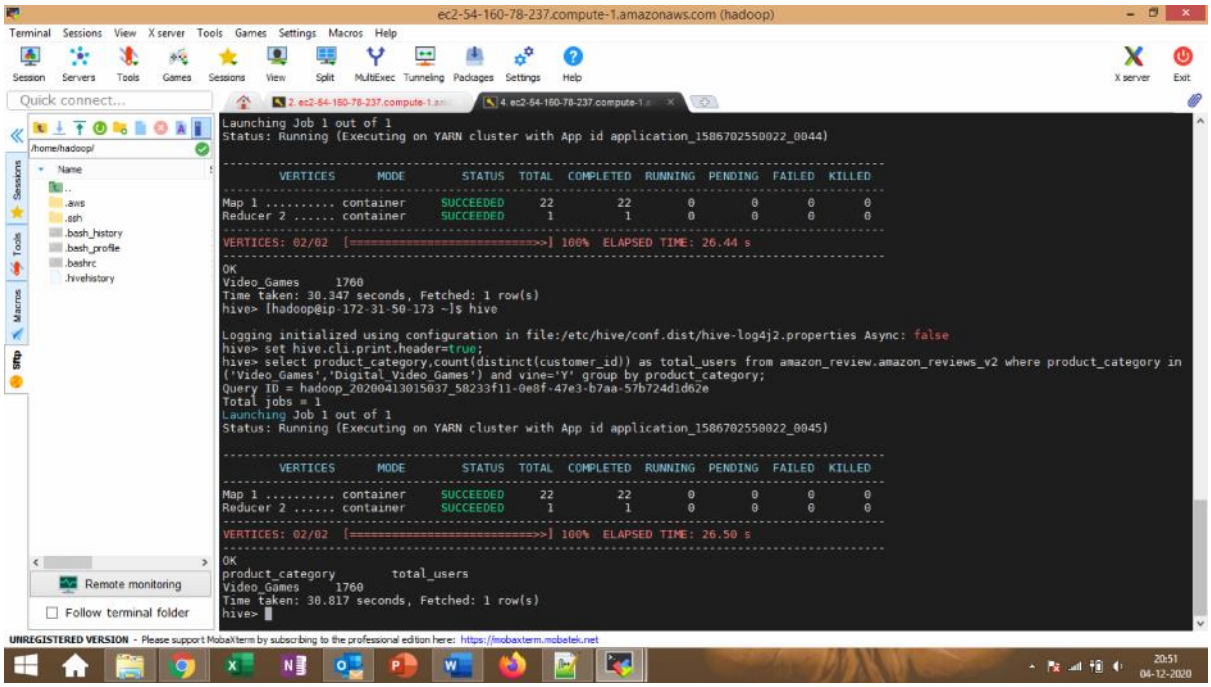
Number of vine users in Music and Digital_Music_Purchase

select product_category,count(distinct(customer_id)) as total_users from amazon_review.amazon_reviews_v2 where product_category in ('Music','Digital_Music_Purchase') and vine='Y' group by product_category;



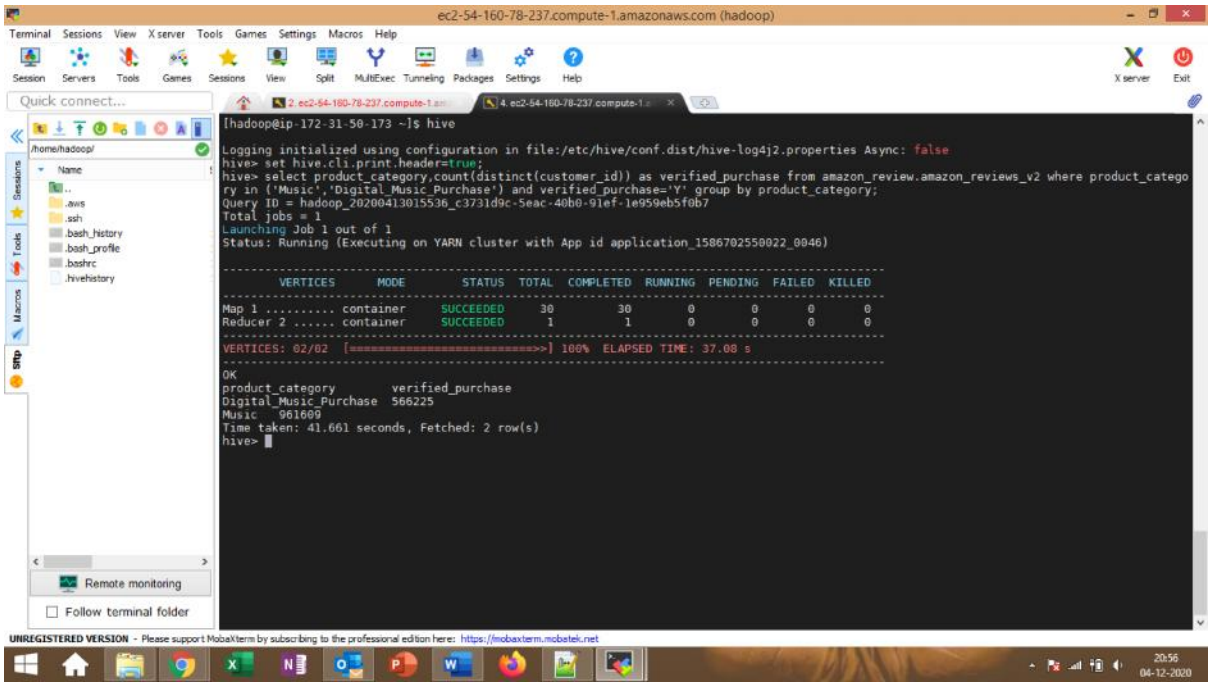
Number of vine users in Digital_Video_Games and Video_Games

select product_category,count(distinct(customer_id)) as total_users from amazon_review.amazon_reviews_v2 where product_category in ('Video_Games','Digital_Video_Games') and vine='Y' group by product_category;



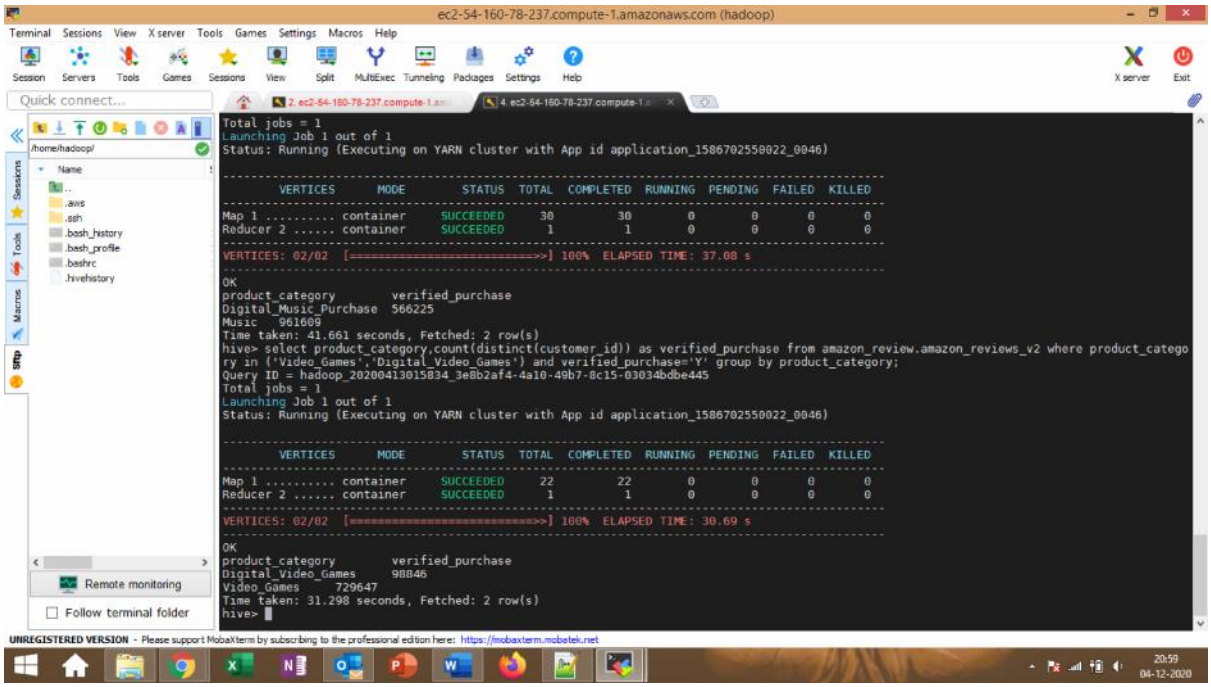
Number of verified_reviews in Music and Digital_Music_Purchase

select product_category,count(distinct(customer_id)) as verified_purchase from amazon_review.amazon_reviews_v2 where product_category in ('Music','Digital_Music_Purchase') and verified_purchase='Y' group by product_category;



Number of verified_reviews in Digital_Video_Games and Video_Games

select product_category,count(distinct(customer_id)) as verified_purchase from amazon_review.amazon_reviews_v2 where product_category in ('Video_Games','Digital_Video_Games') and verified_purchase='Y' group by product_category;



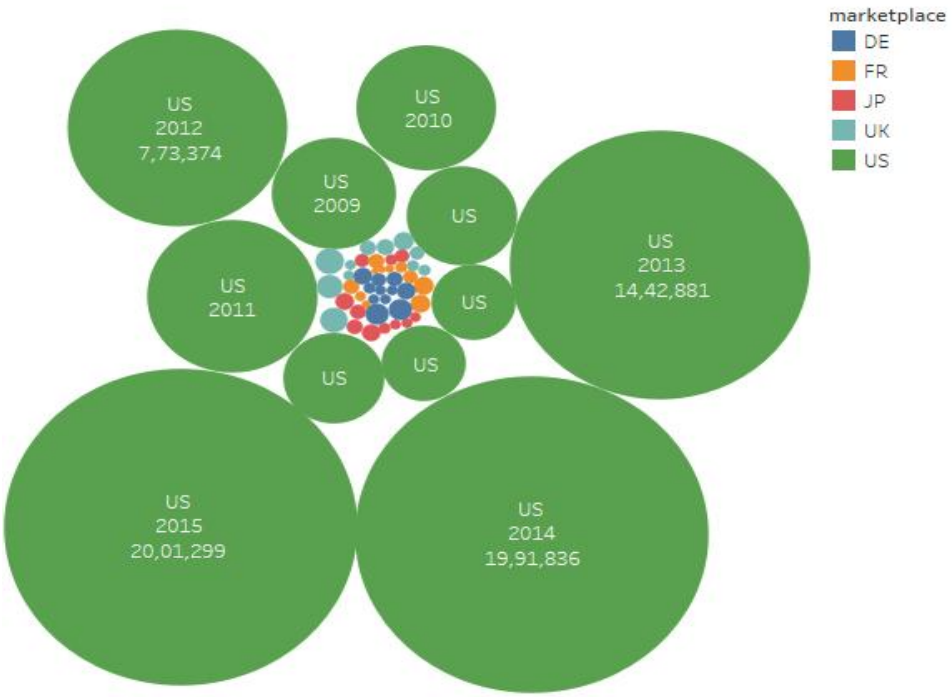
Conclusion

The finding provided about the dataset is with respect to the marketplace parameter. The marketplace gives detail about the customer behaviour based on their region. The following visualization shows the number of product types based on the marketplace. If you see the result, US has the greatest number of product types and the number of products sold in US is higher than any other region. It can be due to two factors which are the number of customers who shop online and the frequency of purchasing a product.

US is the largest consumer of consumer products in the world and it also the lowest to use recycled products. Consider European market for example, the Europeans recycle and reuse many products and thus the frequency of purchasing a product is less when compared to the US. And the number of people who shop online is larger in US when compared to other parts of the world.

Because the convenience in shopping online in US is significant when compared to Europe and Asia. This might be because of the distance of shops from the place of residence in Europe and Asia is less when compared to the US and this might be an important factor on the customer's choice of shopping. This can be significantly seen when comparing Japan and the US. As the proximity of shops in Japan is nearer when compared to the US, more people purchase in physical stores rather than online stores.

Sheet 1



Marketplace, year and sum of total_type_of_products. Colour shows details about marketplace. Size shows sum of total_type_of_products. The marks are labelled by marketplace, year and sum of total_type_of_products.

References

https://minimaxir.com/2014/06/reviewing-reviews/
https://dzone.com/articles/100-shades-of-grey