Project Part 1

Chandra Sekhar Padhy **CSP180002**

Introduction and Problem Description:

From the dataset provided we could compare products and develop insights as to which proudtcs are better with respect to sales and response rate. This was possible because of various parameters like Number of Reviews, Number of Users, Rating, Customer ID, Product ID wherein Customer ID is unique and could not be reused more than once. Another parameter would be customer ratings for the products. Customer ratings would give a fair insight to the potential buyer and would help them choose the right product for them.

Some products are bound to be have better ratings than others. My aim is to analyze which product is performing well and which product is not through the queries that are mentioned below. The products which have not so good reviews could be brought to the notice of the business and they could then focus on how to improve them.

As for the products with good customer reviews and ratings could be used by the business to analyze what they did right, how to improve on them and introduce the same kind of plan for future produts for them to be a success.

The approach that I have followed to complete my analysis include :

Data Cleaning, Views, Parition the dataset, Visualization for better understanding for the business, Analytic queries to achieve desired interpretations.

Technical Scripts and SQL Queries

The dataset has too many records with multiple reviews by the same user for the same product.

Use of such data could lead to misrepresentation of the product review.

I have thus decided to exclude records that have multiple reviews of the same product by the same user.

For this I have created a filter View and taken data into consideration only after 2005 as the previous data is not of much importance for our analysis.

I have also included suitable Visualizations for the queries and their interpretaion.

The data has various product categories and they have been analyzed

Create View with excluded data

```
CREATE view filter_excluded AS
```

SELECT *

FROM amazon_review.amazon_reviews_parquet

WHERE review_id IN

(SELECT x.review_id FROM

(SELECT customer_id, product_id, review_id,

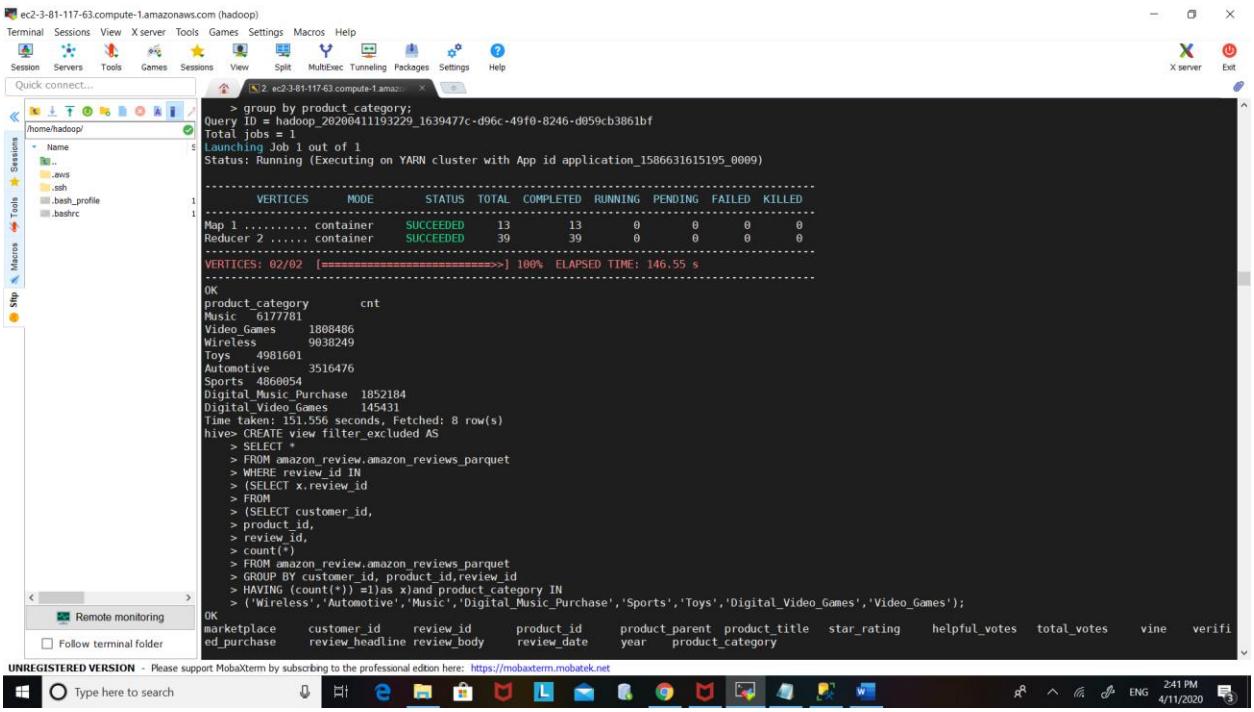
count(*)

FROM amazon_review.amazon_reviews_parquet

GROUP BY customer_id, product_id,review_id

HAVING (count(*)) =1)as x)and product_category IN

('Wireless','Automotive','Music','Digital_Music_Purchase','Sports','Toys','Digital_Video_Games','Video_Games');



Create table with Excluded data

CREATE TABLE amazon_review.amazon_review_filtered_data

AS

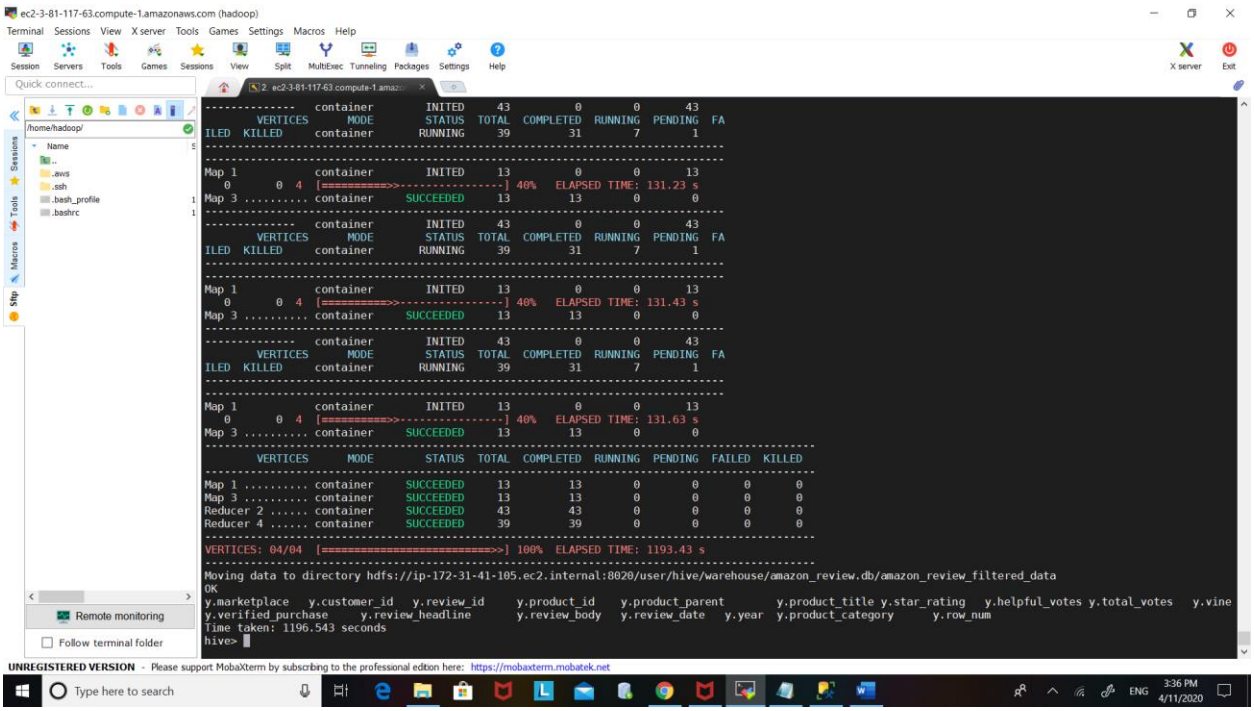
SELECT y.* from

(SELECT *,

row_number()

OVER (partition by customer_id, product_id) AS row_num FROM filter_excluded)y

WHERE row_num=1;



Q1) Explore the dataset and perform basic exploratory analysis

SELECT year, count(review_id) AS No_of_Reviews,

count(distinct(customer_id)) AS No_of_Users,

round(avg(star_rating), 2) AS stars_avg_review,

round(avg(length(review_body)),2) AS length_avg_review,

sum(case

WHEN verified_purchase ='Y'then 1

ELSE 0 end) AS verified_users, sum(case

WHEN verified_purchase='N' THEN

1

ELSE 0 end) AS unverified_users, sum(case

WHEN helpful_votes= 1 THEN

1

ELSE 0 end) AS Number_of_Helpful_Votes, count(distinct(product_id)) AS

Total_Number_of_Products

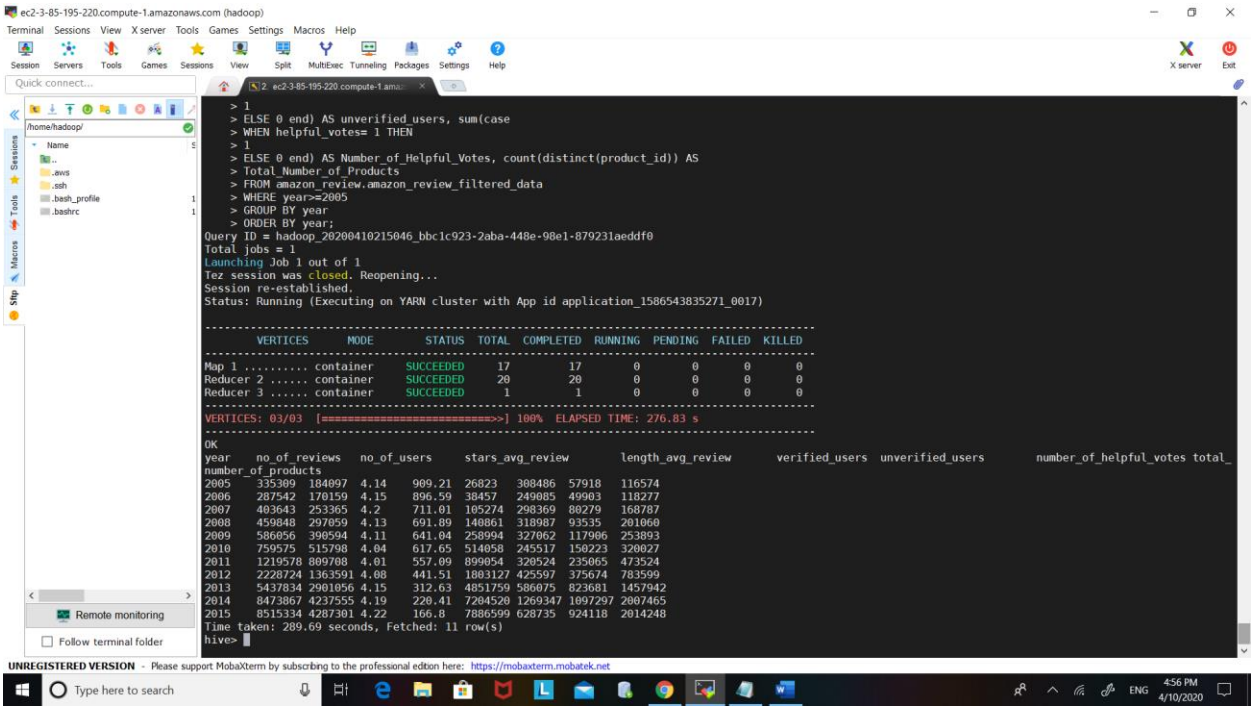
FROM amazon_review.amazon_review_filtered_data

WHERE year>=2005

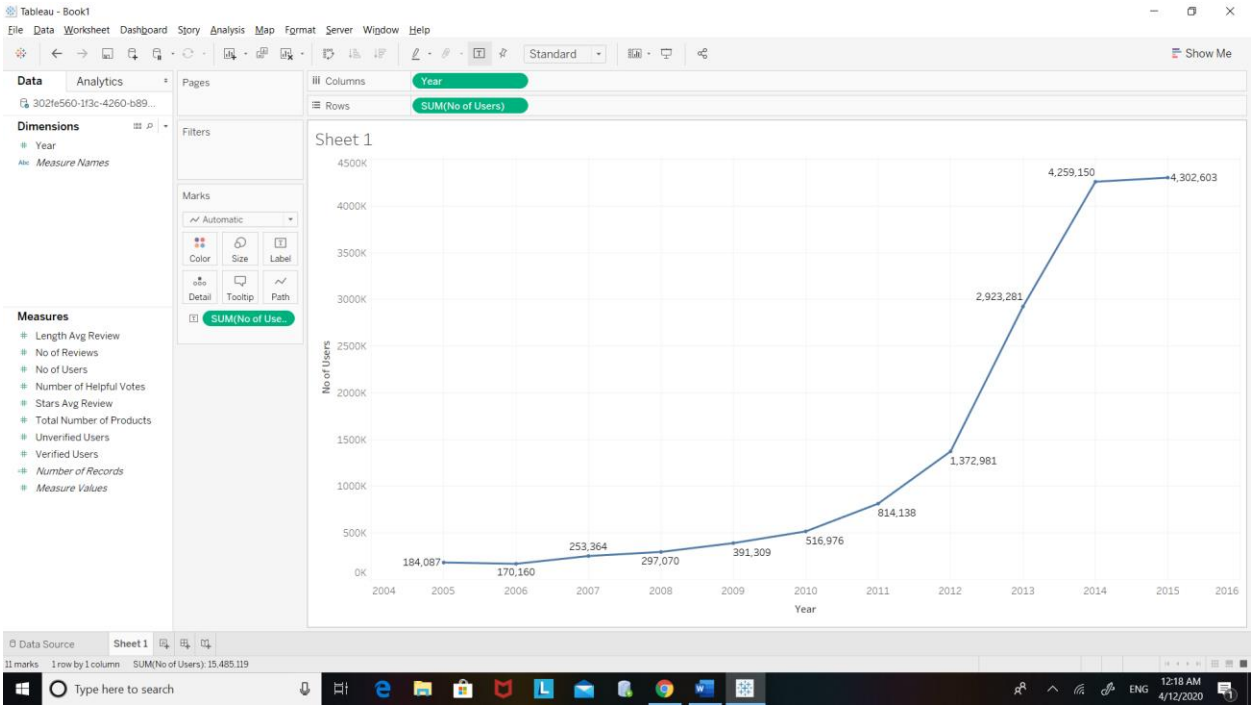
GROUP BY year

ORDER BY year;

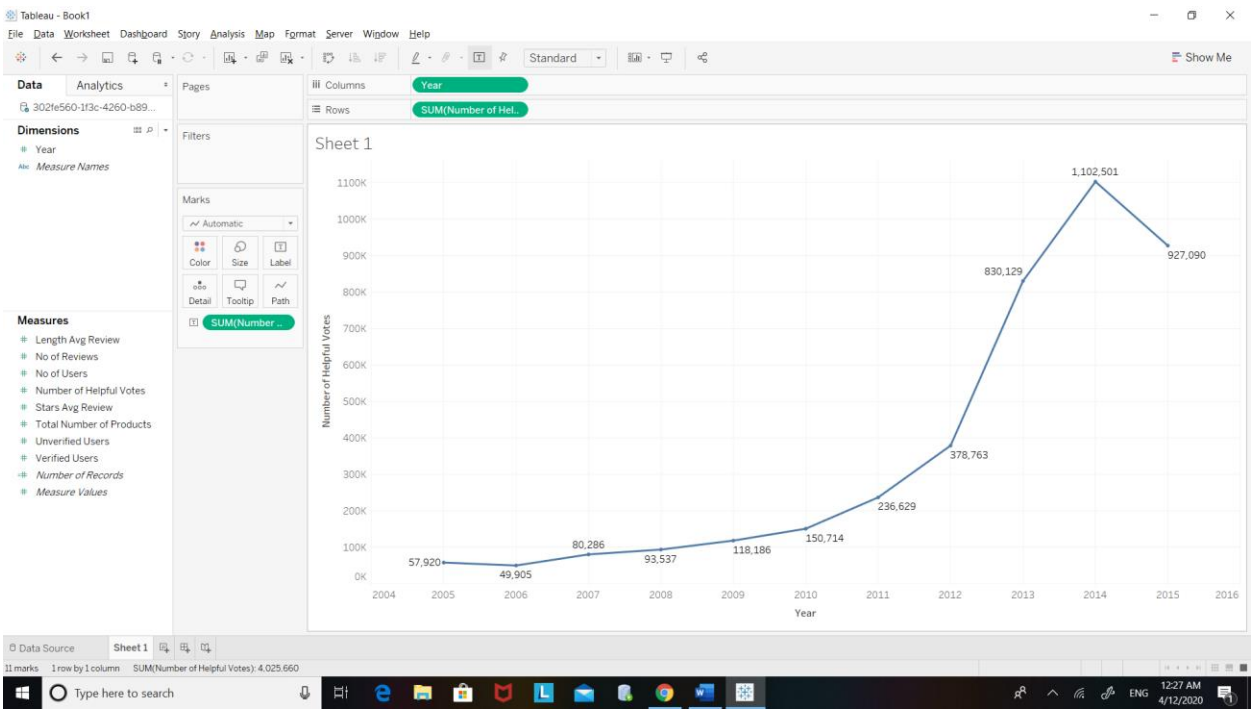
Output:



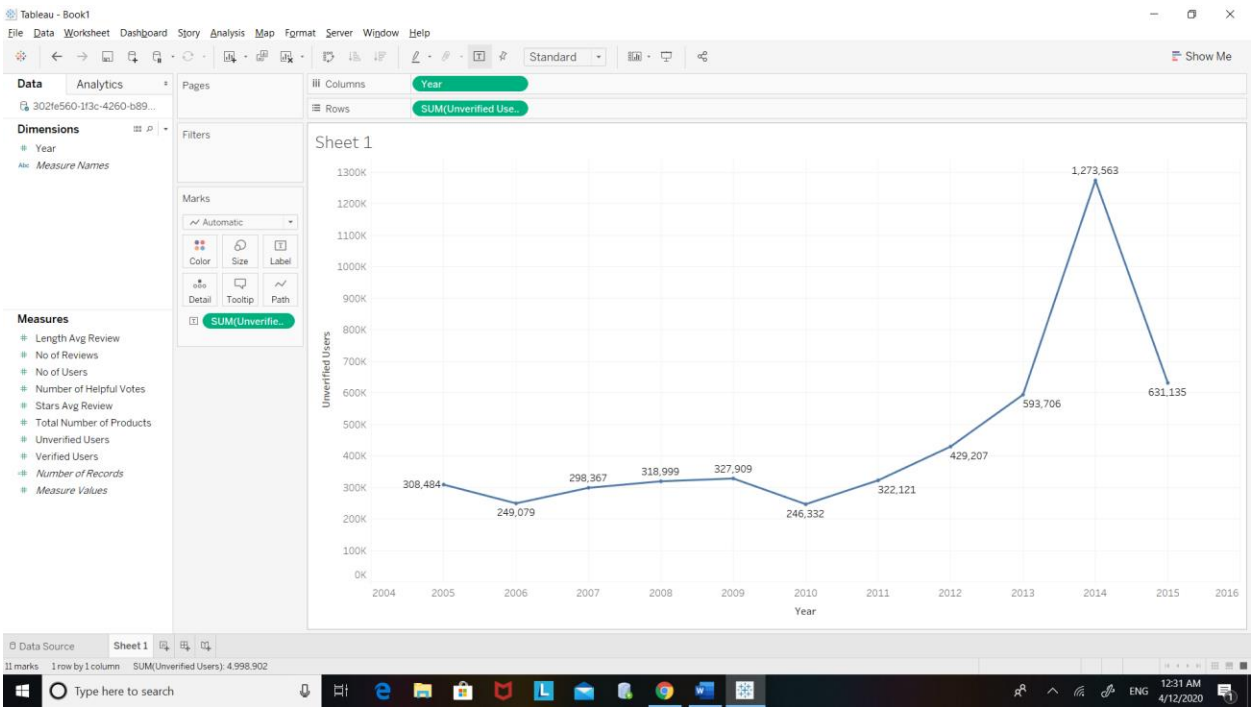
Visualizations:



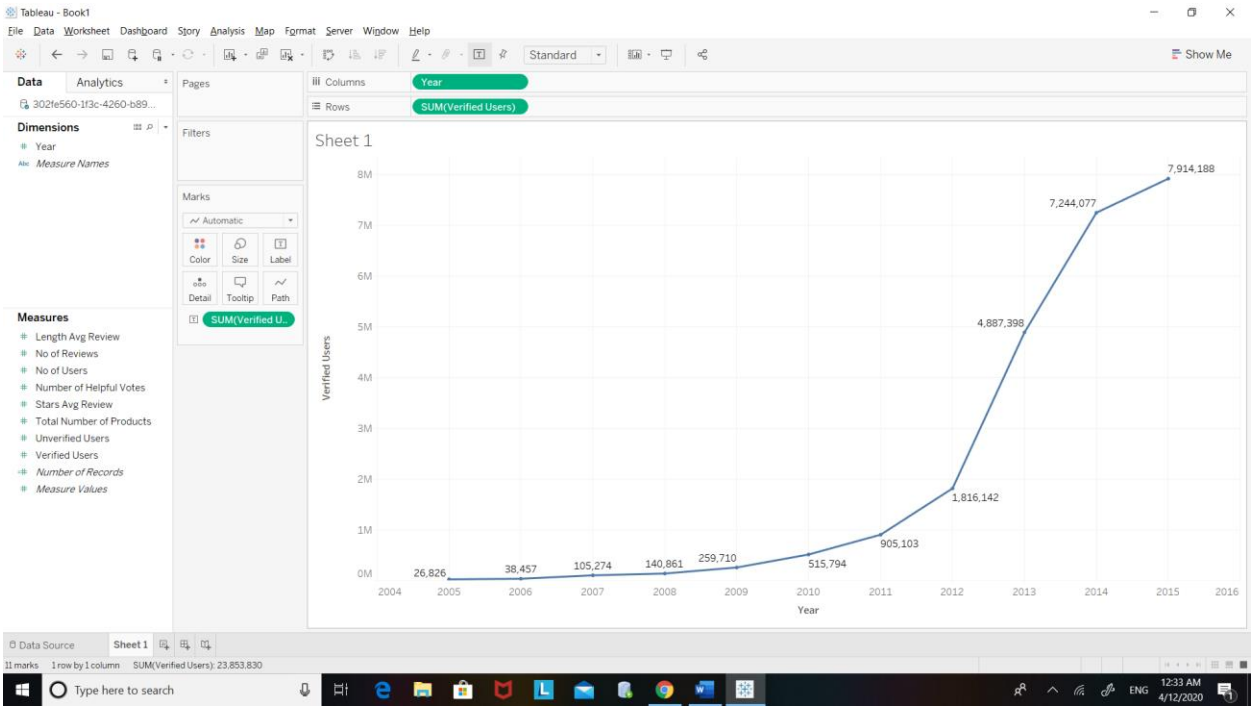
From the above visualization we can see that the number of Users increased over the years.



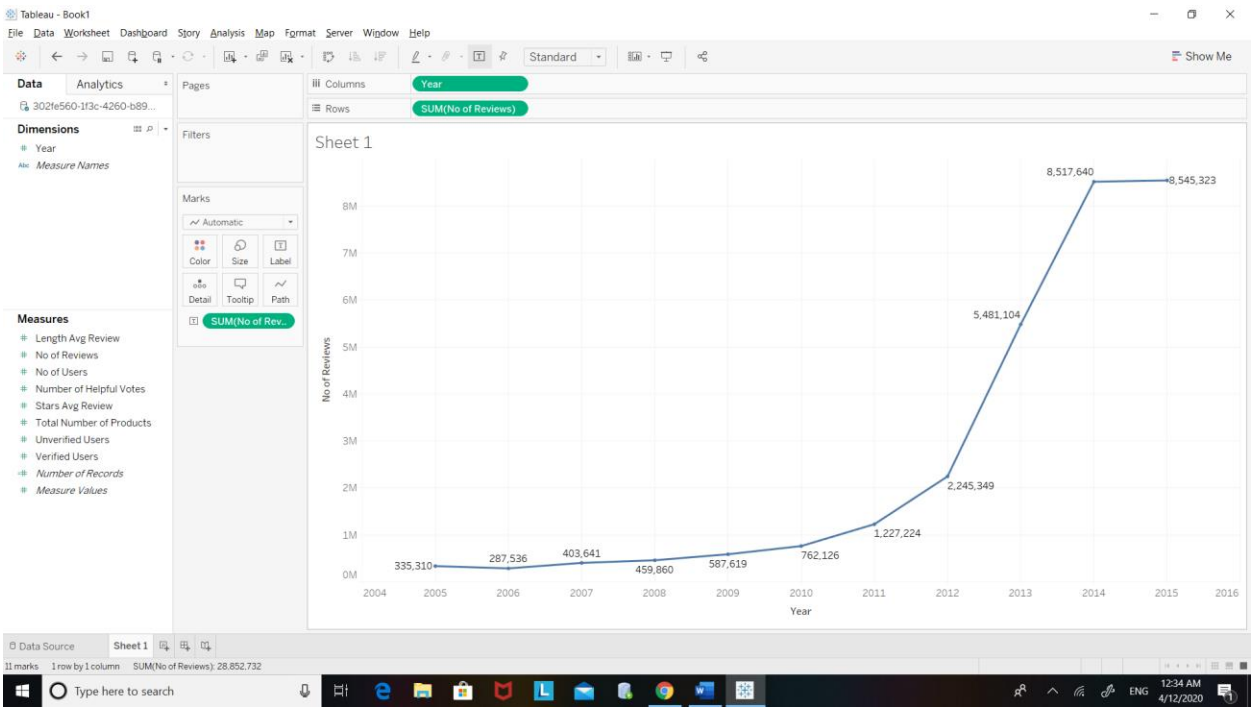
The above graph shows us an upward trend in the Number of Helpful Votes up until 2014 and then a decrease in the number in 2015.



The above Visulaization helps us analyse that the Number of Unverified Users increased over the year up until 2014 and then saw a decrease in Number in 2015



This graph shows increase in the number of verified users over the years.



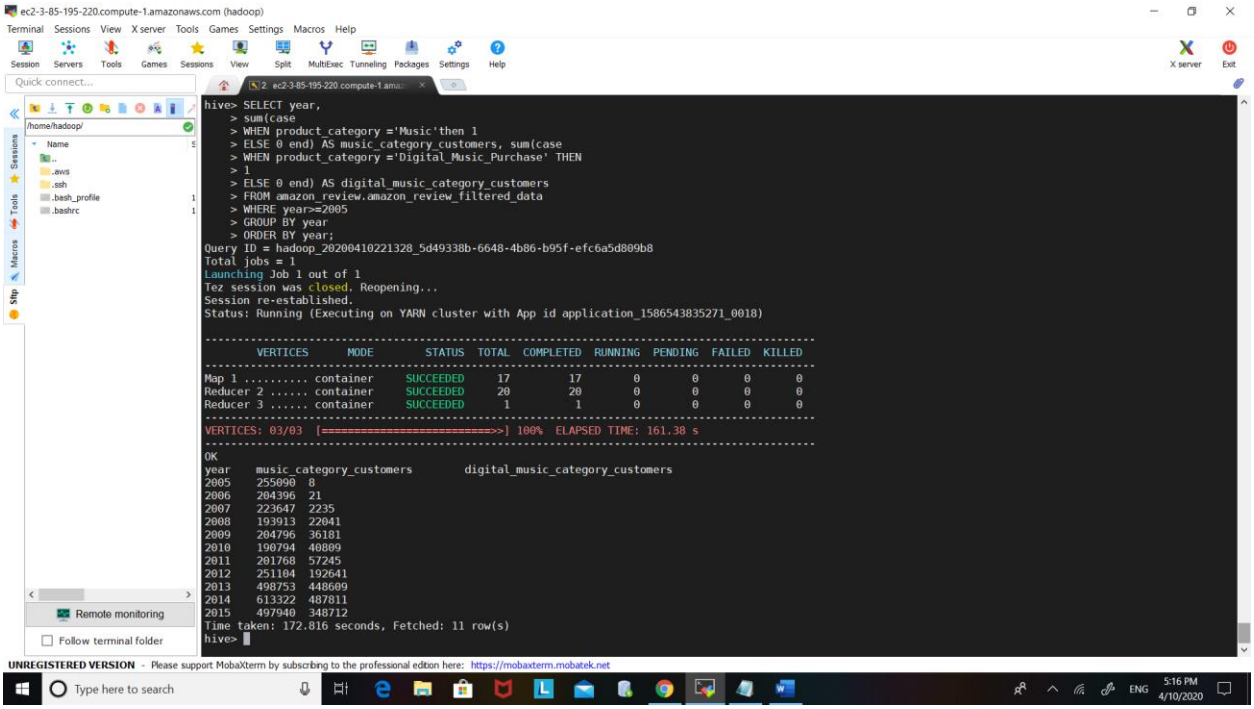
This graph helps us conclude that the number of reviews increased over the years.

Detailed Analysis of Music/Digital_Music_Purchase and Digital_Video_Games/Video_Games over time

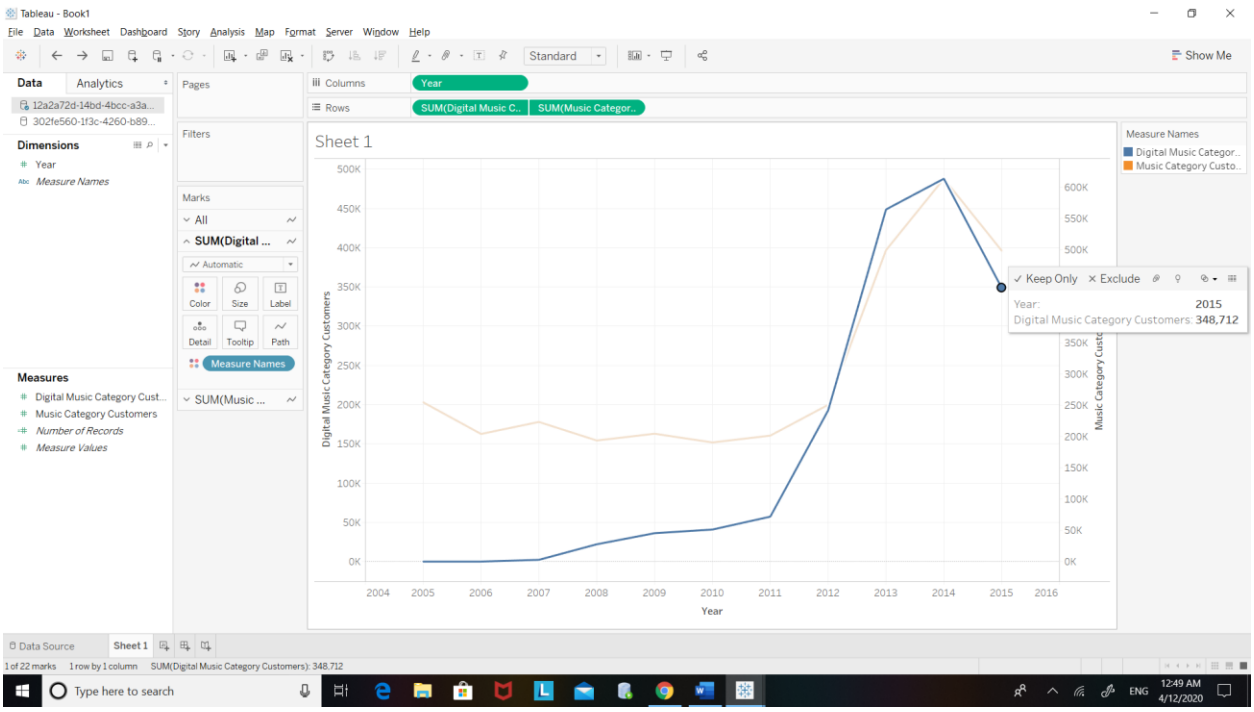
Analysis on Music Related Category

```
SELECT year, sum(case
WHEN product_category ='Music'then 1
ELSE 0 end) AS music_category_customers, sum(case
WHEN product_category ='Digital_Music_Purchase' THEN
1
ELSE 0 end) AS digital_music_category_customers
FROM amazon_review.amazon_review_filtered_data
WHERE year>=2005
GROUP BY year
ORDER BY year;
```

Output:



Visualization:



Digital Music Category saw a steep rise and did overtake the Music Category for a short period of time around 2013 but then experienced a decline in the number of customers for 2015.

Both the categories had an increase in the number of users over the years.

Co-relation between Various Categories over time.

Analysis on Game Related Category

SELECT year, sum(case

WHEN product_category ='Video_Games'then 1

ELSE 0 end) AS video_game_category_customers, sum(case

WHEN product_category ='Digital_Video_Games' THEN

1

ELSE 0 end) AS digital_video_game_category_customers

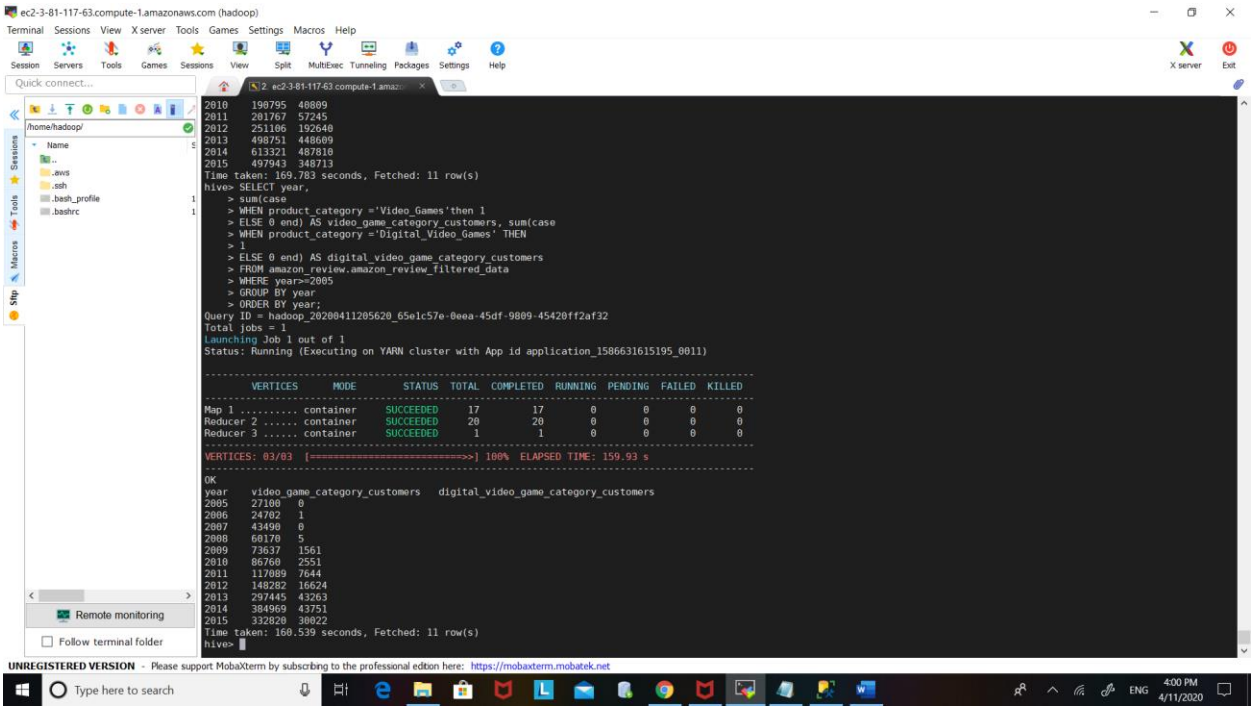
FROM amazon_review.amazon_review_filtered_data

WHERE year>=2005

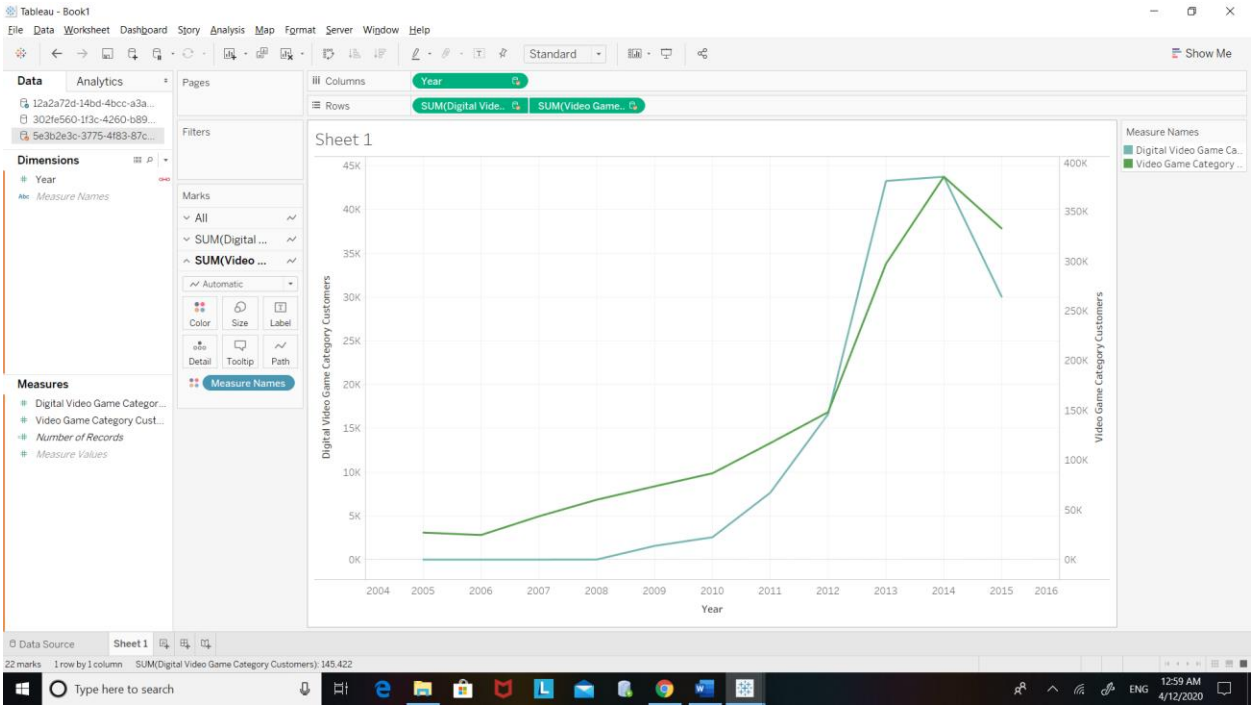
GROUP BY year

ORDER BY year;

Output:



Visualization:



Digital Video Game Category witnessed a steep increase in the number of customers after the year 2010. It did overtake the Video Game category for some time in 2012-2014.

On an individual basis both of them increased and had a decline in 2015, similar to the Music Category.

Are there Users reviewing in both categories?

Music Category:

```
SELECT count(y.customer_id) AS count,y.year

FROM amazon_review.amazon_review_filtered_data y,

(SELECT distinct(customer_id)

FROM amazon_review.amazon_review_filtered_data

WHERE product_category='Music'

AND year>=2005 intersect SELECT distinct(customer_id)

FROM amazon_review.amazon_review_filtered_data

WHERE product_category='Digital_Music_Purchase'

AND year>=2005)z

WHERE y.customer_id=z.customer_id
```

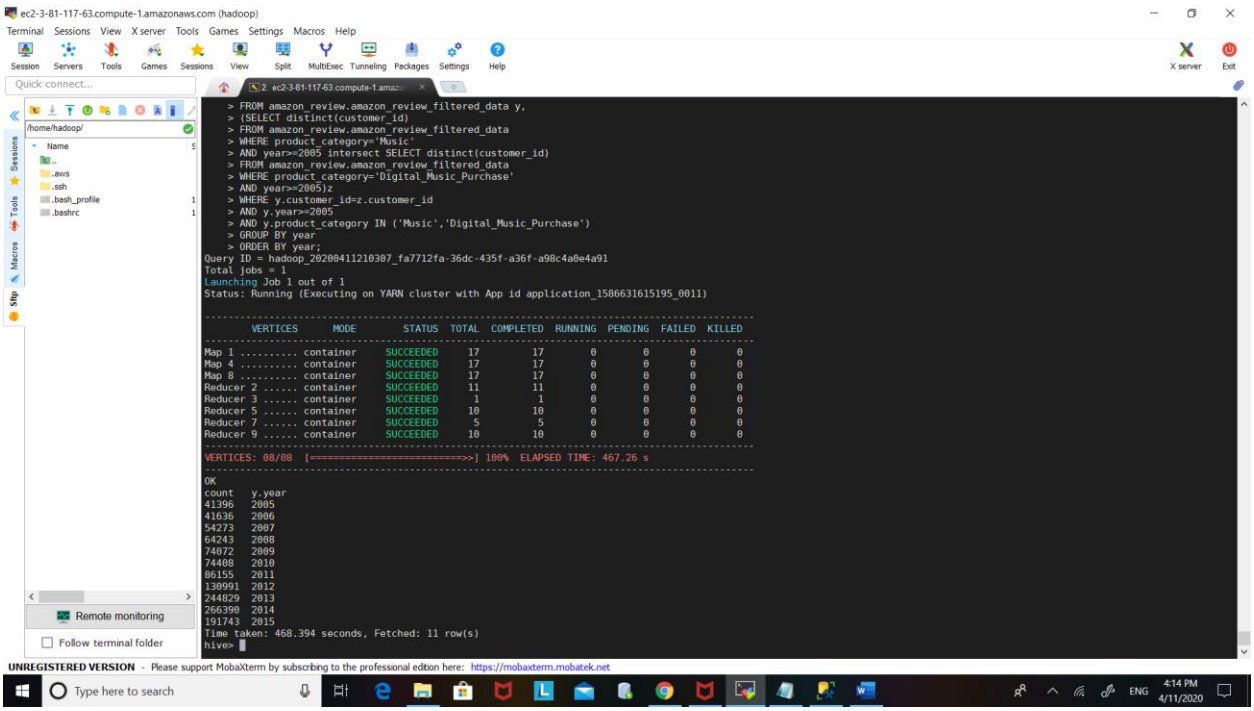
AND y.year>=2005

AND y.product_category IN ('Music','Digital_Music_Purchase')

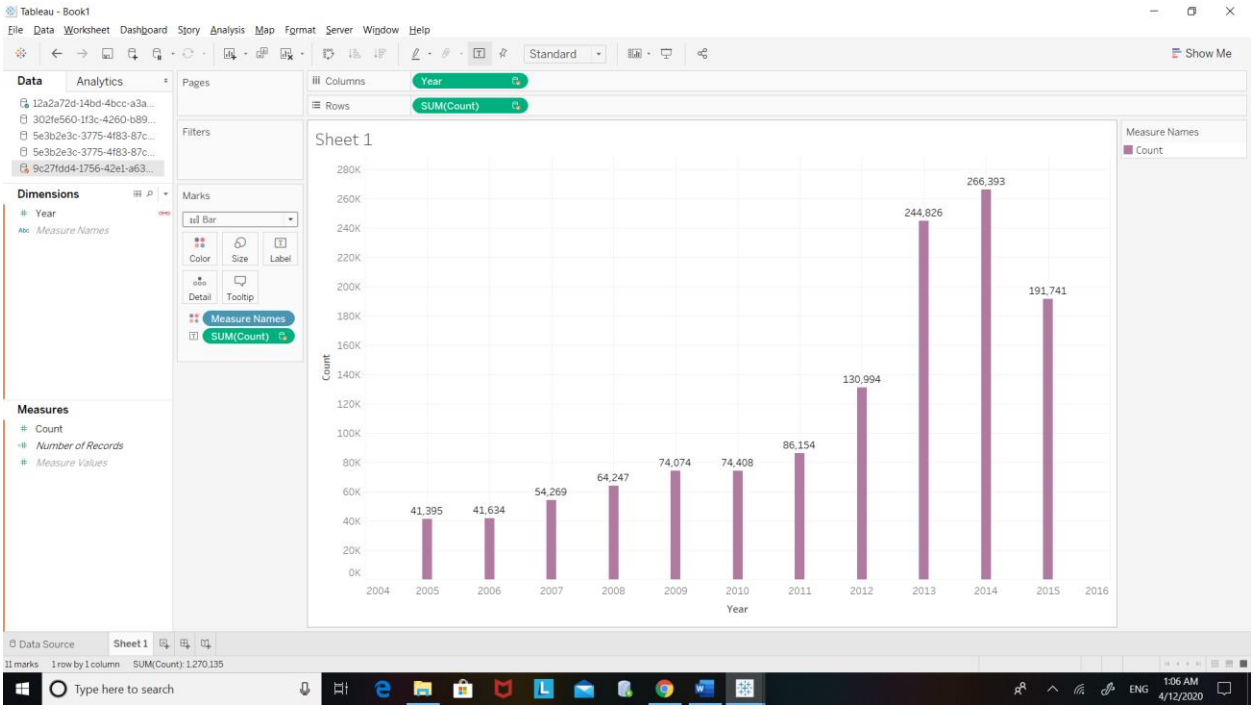
GROUP BY year

ORDER BY year;

Output:



Visualization:



This Bar Graph shows that customers who reviewd both Music and Digital Music Categories increased over the years with the most being in the year 2014 and then there was a decrease in 2015.

Game Category

SELECT count(y.customer_id) AS count,

y.year

FROM amazon_review.amazon_review_filtered_data y,

(SELECT distinct(customer_id)

FROM amazon_review.amazon_review_filtered_data

WHERE product_category='Video_Games'

AND year>=2005 intersect SELECT distinct(customer_id)

FROM amazon_review.amazon_review_filtered_data

WHERE product_category='Digital_Video_Games'

AND year>=2005)z

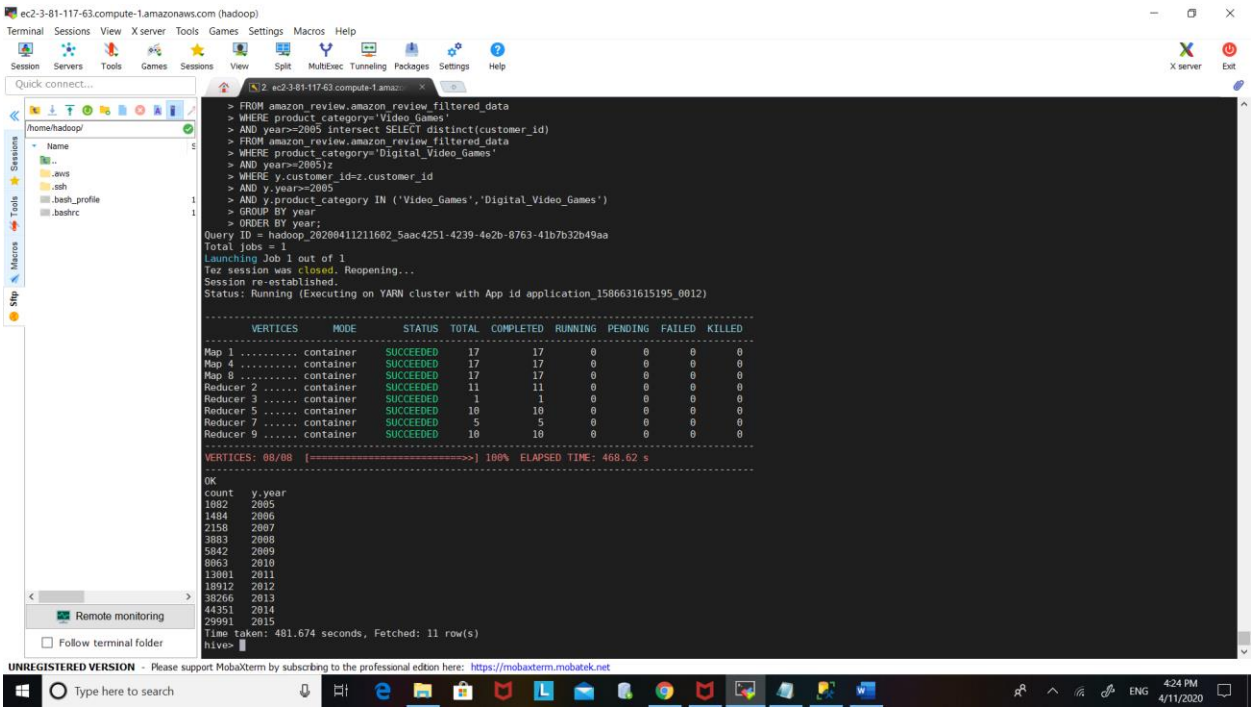
WHERE y.customer_id=z.customer_id

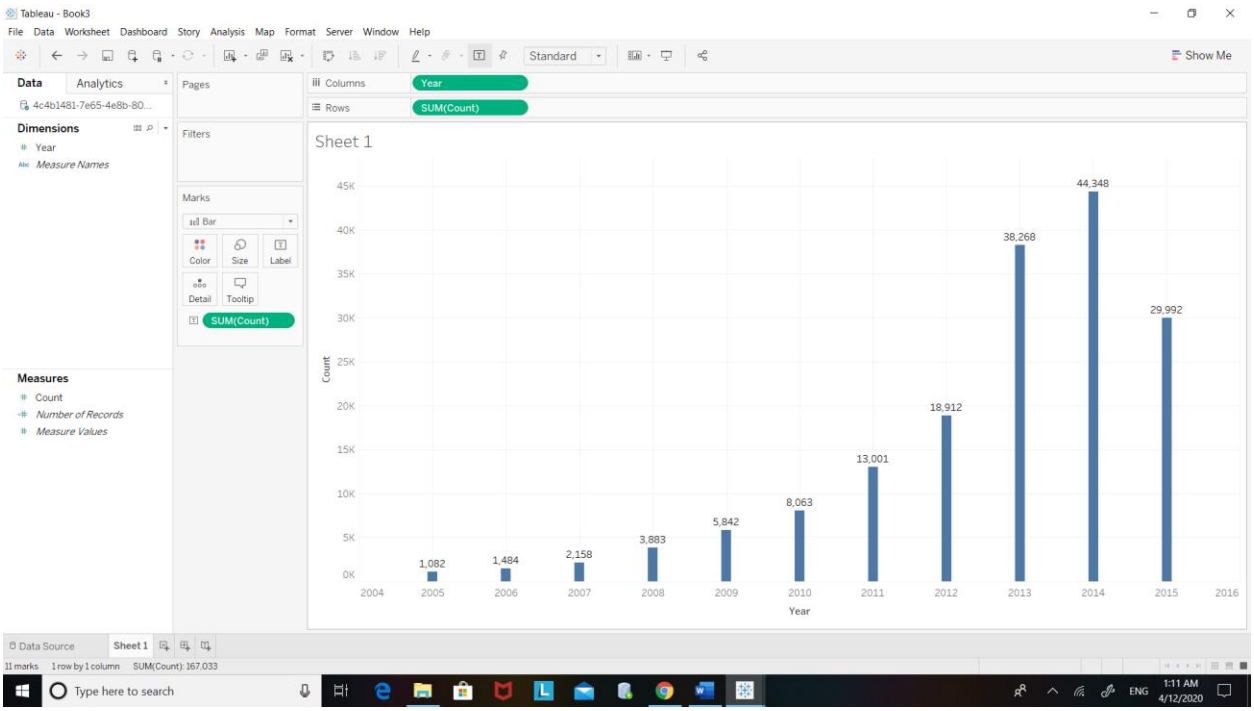
AND y.year>=2005

```
AND y.product_category IN ('Video_Games','Digital_Video_Games')

GROUP BY year

ORDER BY year;
```





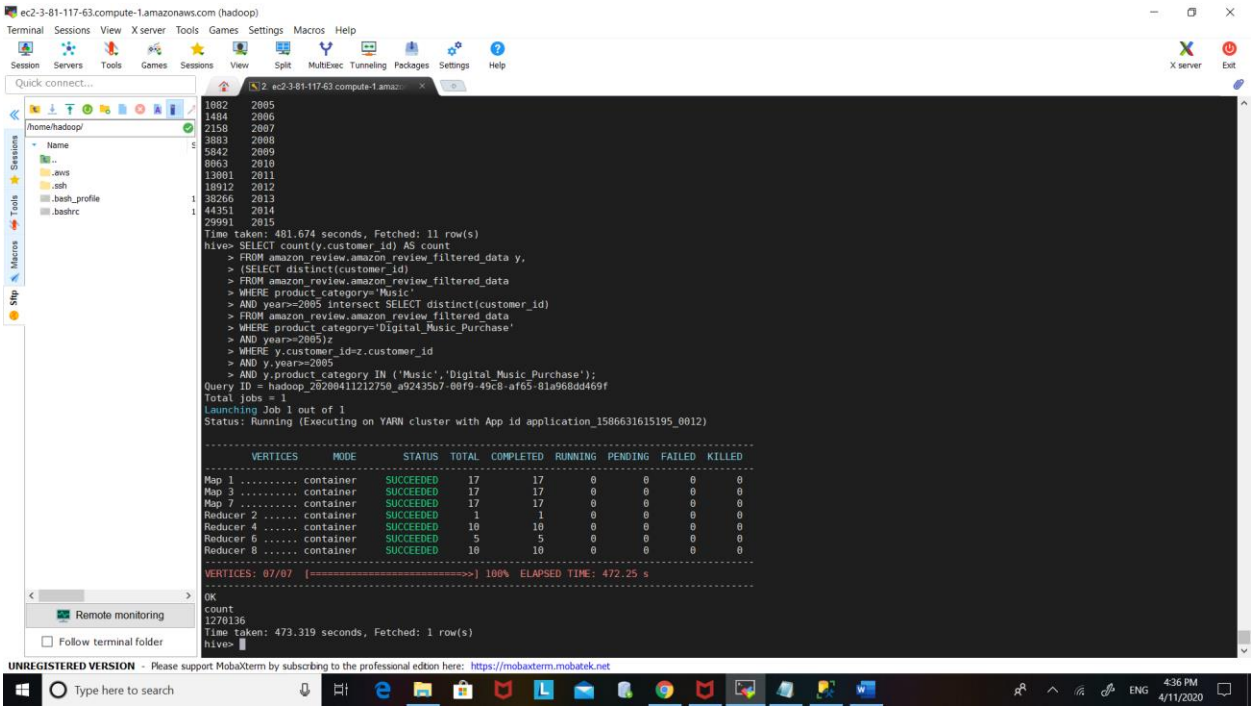
The number of customers reviewing Game Category shows a similar pattern to the Music Category with the highest number being in the year 2014 and then a decrease in number in 2015.

Total Number of Same Users reviewing in both categories

Users reviewing Music and Digital_Music_Purchase categories:

```
SELECT count(y.customer_id) AS count
FROM amazon_review.amazon_review_filtered_data y,
(SELECT distinct(customer_id)
FROM amazon_review.amazon_review_filtered_data
WHERE product_category='Music'
AND year>=2005 intersect SELECT distinct(customer_id)
FROM amazon_review.amazon_review_filtered_data
WHERE product_category='Digital_Music_Purchase'
AND year>=2005)z
WHERE y.customer_id=z.customer_id
AND y.year>=2005
AND y.product_category IN ('Music','Digital_Music_Purchase');
```

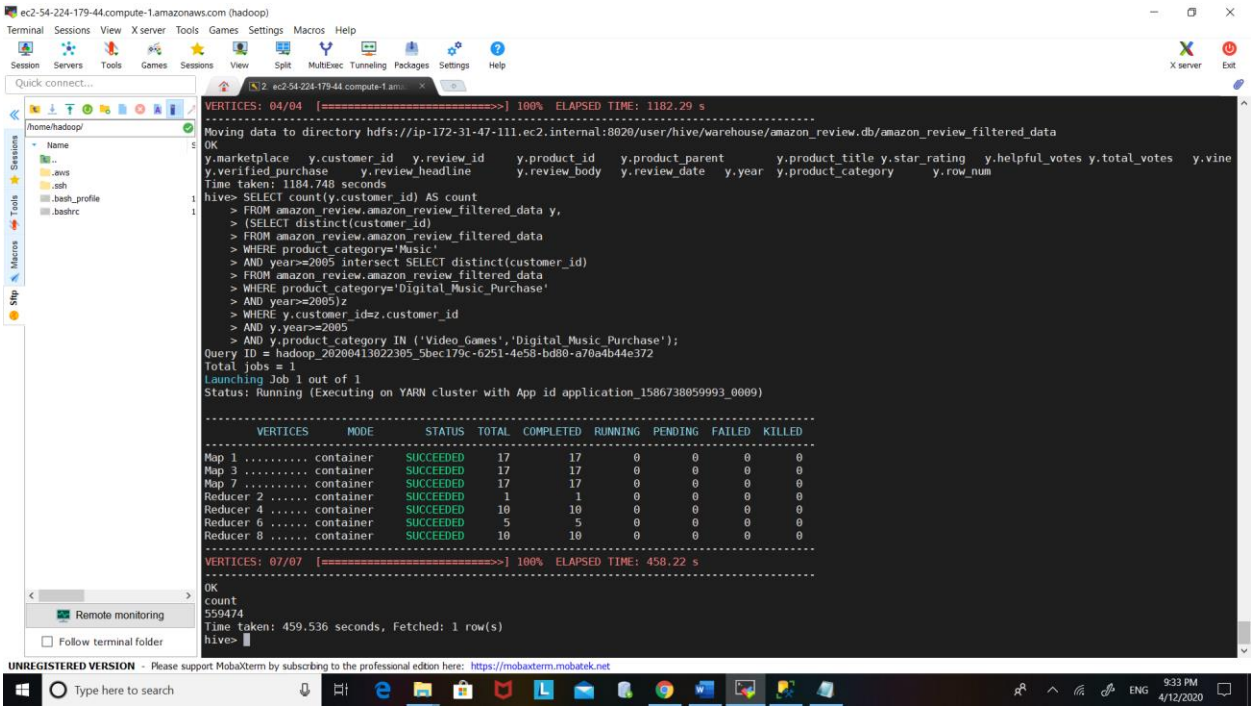
Output:



Users reviewing video_games and Digital_Video_Games categories:

```
SELECT count(y.customer_id) AS count
FROM amazon_review.amazon_review_filtered_data y,
(SELECT distinct(customer_id)
FROM amazon_review.amazon_review_filtered_data
WHERE product_category='Video_Games'
AND year>=2005 intersect SELECT distinct(customer_id)
FROM amazon_review.amazon_review_filtered_data
WHERE product_category='Digital_Video_Games'
AND year>=2005)z
WHERE y.customer_id=z.customer_id
AND y.year>=2005
AND y.product_category IN ('Video_Games','Digital_Video_Games');
```

Output:



Can you identify similar items in both categories? Do they get same rating?

Create Views to secure data

```
CREATE view music_category AS
(SELECT product_id,
round(avg(star_rating),
2) AS Music_Ranking
FROM amazon_review.amazon_review_filtered_data
WHERE product_category='Music'
AND year>= 2005
GROUP BY product_id);

CREATE view digital_music_category AS
(SELECT product_id,
round(avg(star_rating),
2) AS Digital_Music_Ranking
FROM amazon_review.amazon_review_filtered_data
WHERE product_category='Digital_Music_Purchase'
AND year>= 2005
GROUP BY product_id);

CREATE view video_game_category AS
(SELECT product_id,
round(avg(star_rating),
2) AS Game_Ranking
FROM amazon_review.amazon_review_filtered_data
WHERE product_category='Video_Games'
AND year>= 2005
GROUP BY product_id);

CREATE view digital_video_game_category AS
(SELECT product_id,
```

round(avg(star_rating),

2) AS Digital_Game_Ranking

FROM amazon_review.amazon_review_filtered_data

WHERE product_category='Digital_Video_Games'

AND year>= 2005

GROUP BY product_id);

Analysis between Music and Digital Music Category

SELECT x.product_id ,

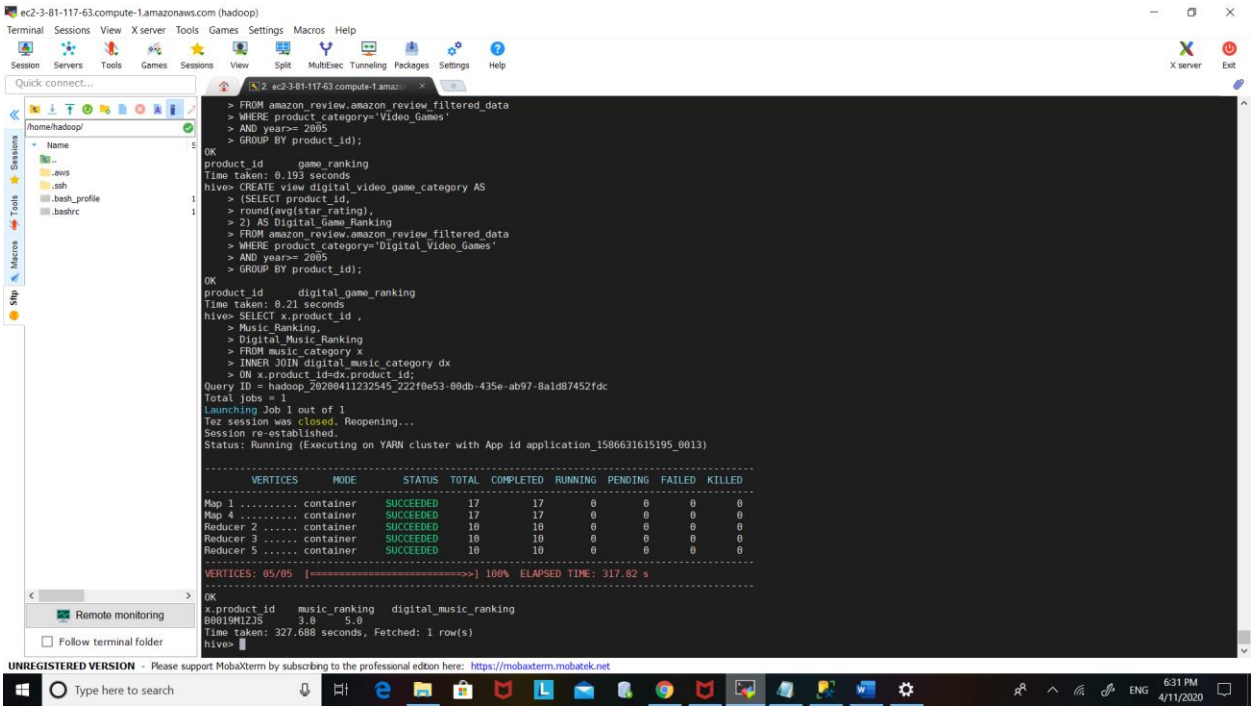
Music_Ranking,

Digital_Music_Ranking

FROM music_category x

INNER JOIN digital_music_category dx

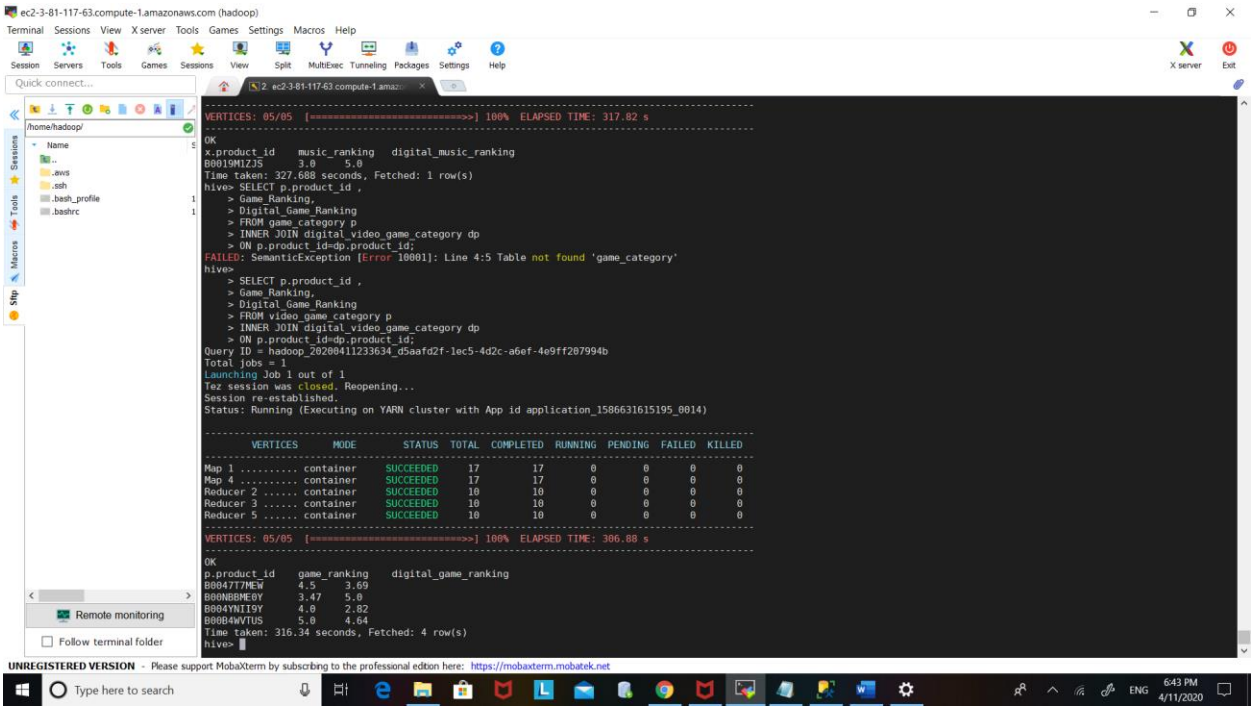
ON x.product_id=dx.product_id;



From this we can interpret that the same product has different rating in Music and Digital Music Category.

Analysis between Video Games and Digital Video Games

```
SELECT p.product_id ,
Game_Ranking,
Digital_Game_Ranking
FROM video_game_category p
INNER JOIN digital_video_game_category dp
ON p.product_id=dp.product_id;
```



We can see that the products have different ratings in Video Game and Digital Video Game Category.

Hive Advanced Functions

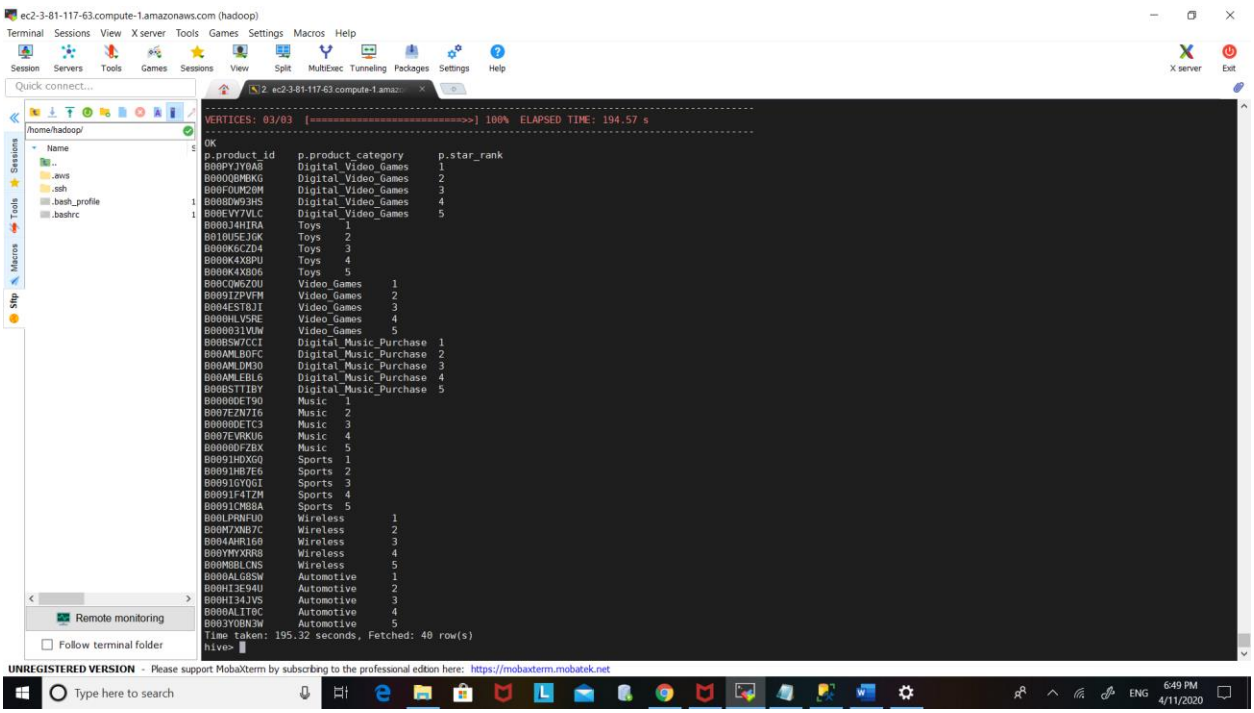
Ranking based on products under different product categories

To calculate rank to find out popular products

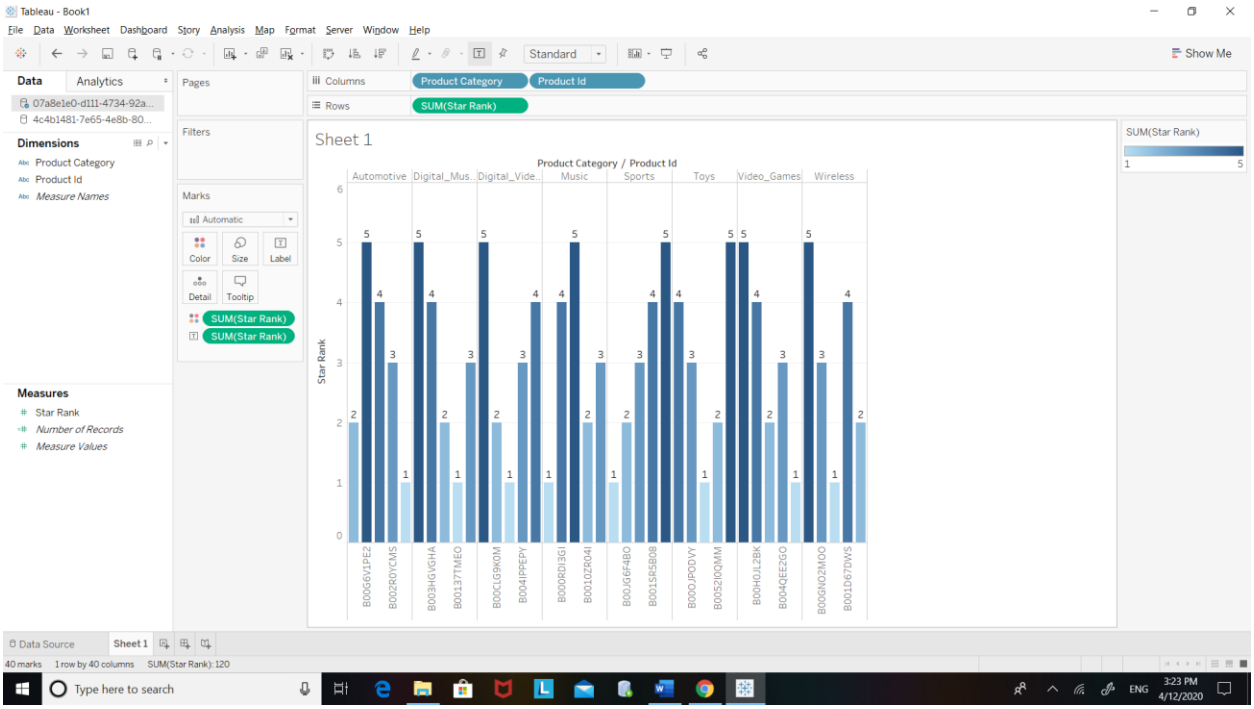
```
SELECT p.product_id,
p.product_category,
p.star_rank
```

```
FROM
(SELECT y.product_id,
y.product_category,
Row_number()
OVER (partition by y.product_category
ORDER BY y.avg_rating desc) AS star_rank
FROM
(SELECT product_id,
product_category,
avg(star_rating) AS avg_rating
FROM amazon_review.amazon_review_filtered_data
WHERE year>= 2005
GROUP BY product_id,product_category)as y)as p
WHERE p.star_rank<=5;
```

Output:



Visualization:

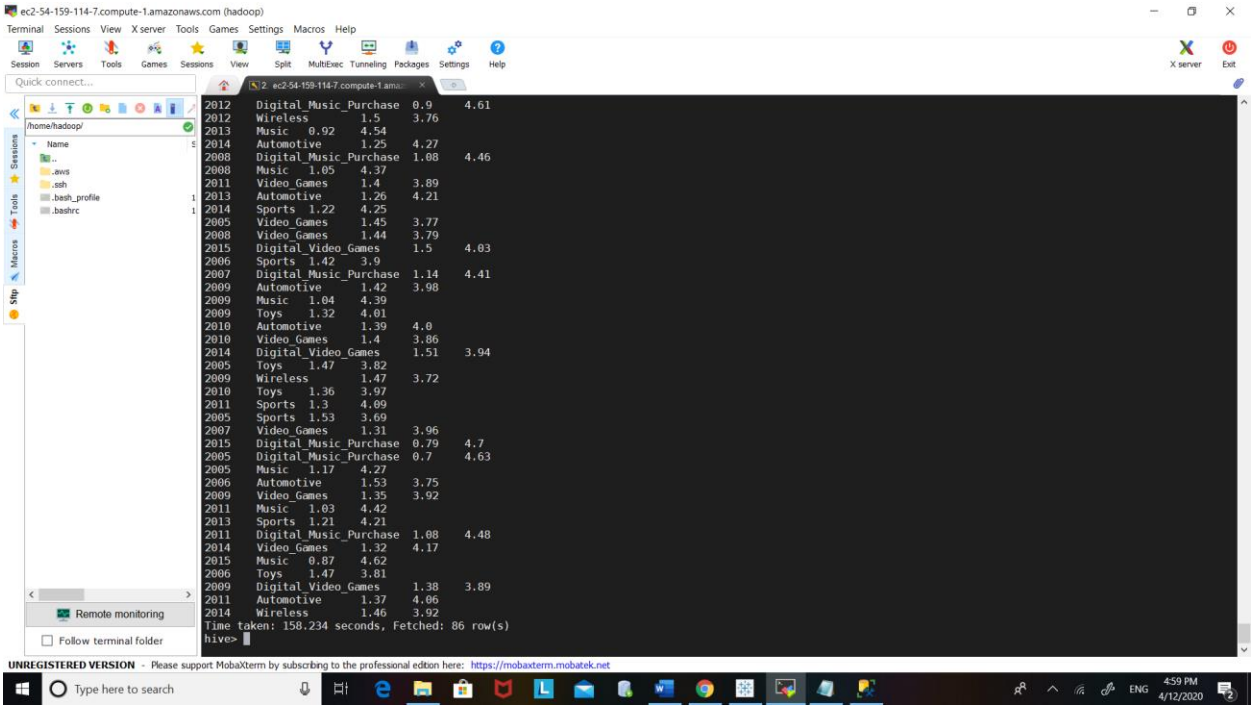


The Visualization shows the ratings for various products. The light shade of blue is to depict a low rating and the darker shade depicts a higher rating.

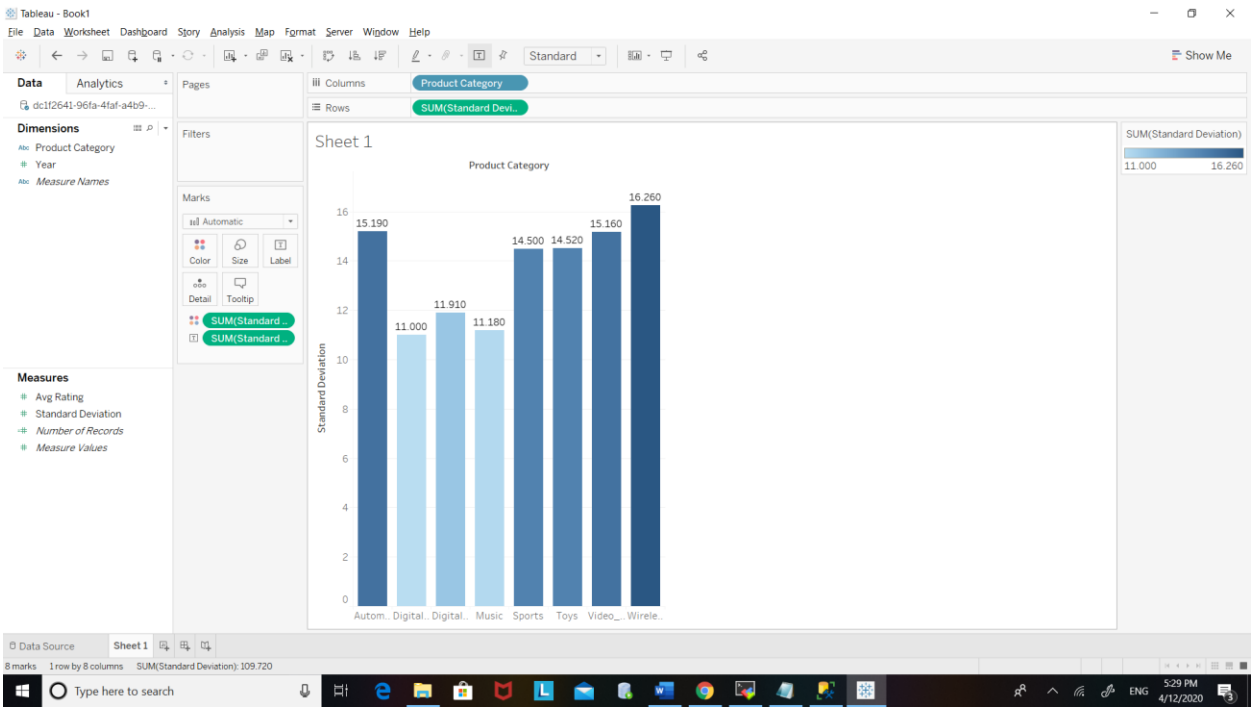
Calculate Standard Deviation

```
SELECT year,
product_category,
round(stddev(star_rating),2)as Standard_Deviation,
round(avg(star_rating),2) as Average_Rating
FROM amazon_review.amazon_review_filtered_data
WHERE year>= 2005
GROUP BY product_category,year;
```

Output:



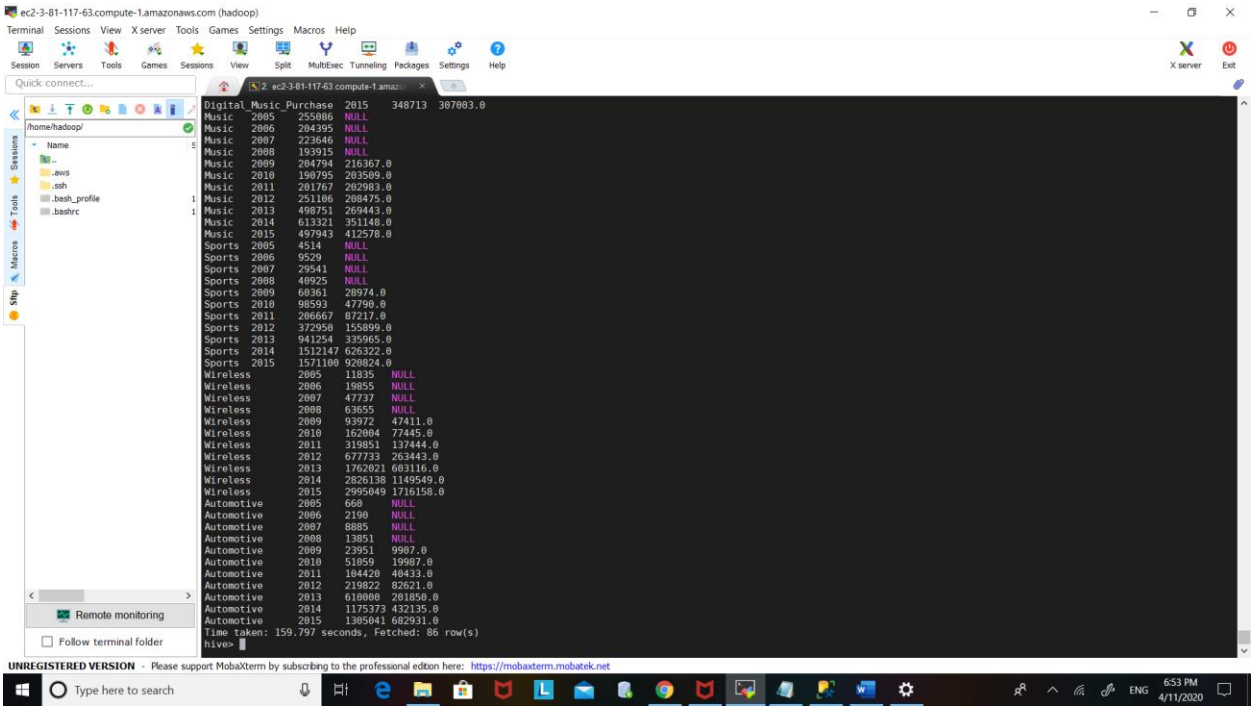
Visualization:

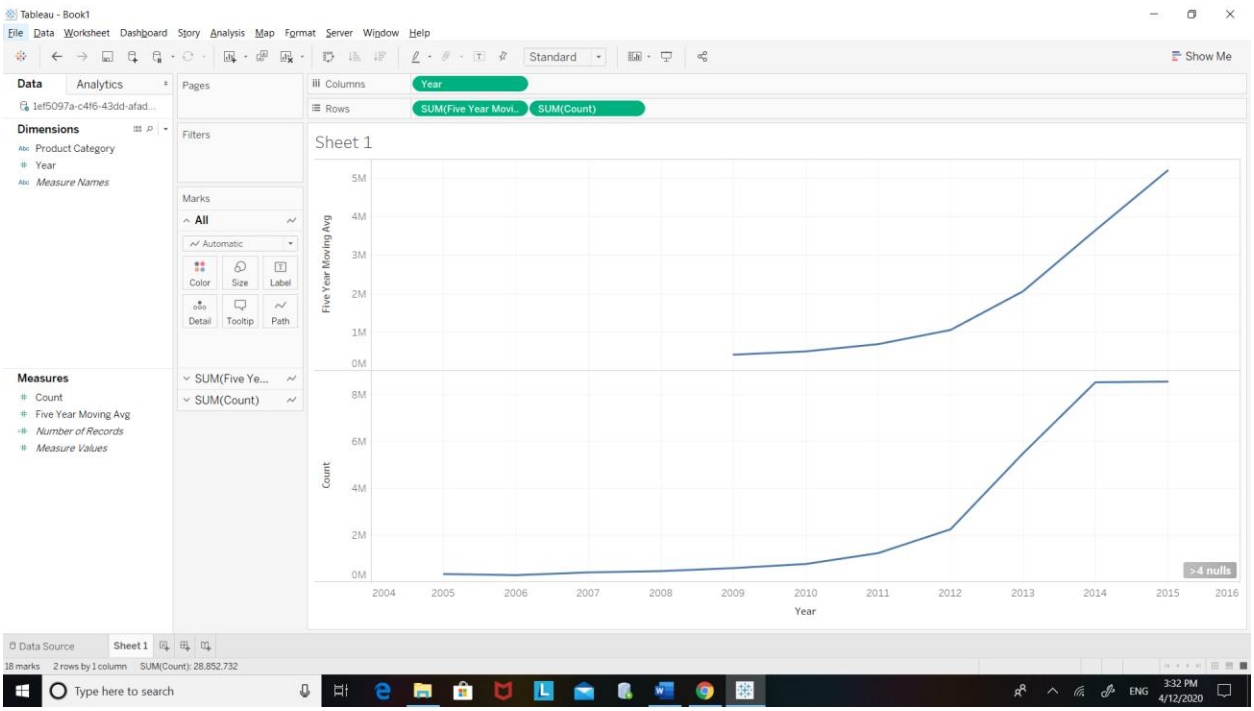


The graph shows the standard deviation for customers for various products.

Compared the growth of the products by calculating moving average

```
SELECT y.product_category,y.year,y.count,
(case
WHEN row_number()
OVER (partition by y.product_category order by y.year) > 4 THEN
round(AVG(y.count)
OVER (PARTITION BY y.product_category
ORDER BY y.year ROWS 4 PRECEDING)) end) AS five_year_moving_avg
FROM
(SELECT product_category,
year,
count(review_id) AS count
FROM amazon_review.amazon_review_filtered_data
GROUP BY product_category,year
ORDER BY year desc) AS y
WHERE y.year>= 2005;
```

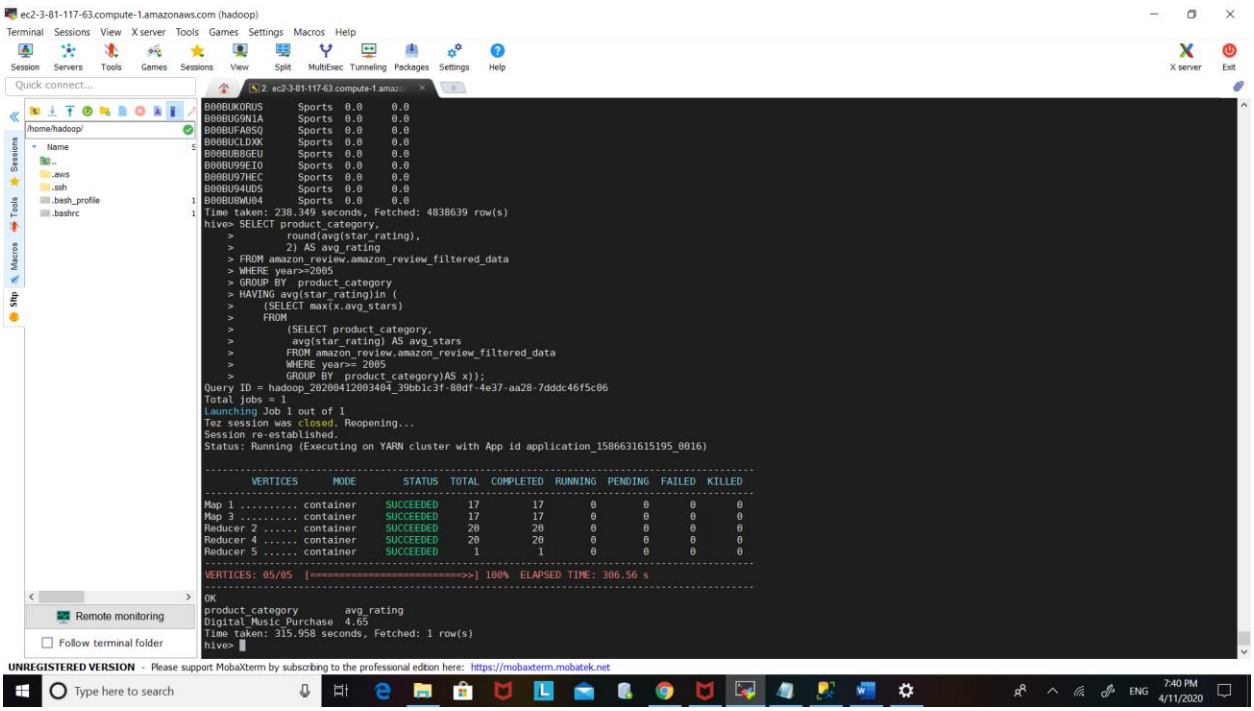




The Growth of Products with respect to moving average is as shown in the visualization above added to that increase in count of customers.

Maximum Rating of a Product in specific product category

```
SELECT product_category,
round(avg(star_rating),2) AS avg_rating
FROM amazon_review.amazon_review_filtered_data
WHERE year>=2005
GROUP BY  product_category
HAVING avg(star_rating)in (
(SELECT max(x.avg_stars)
FROM
(SELECT product_category,
 avg(star_rating) AS avg_stars
 FROM amazon_review.amazon_review_filtered_data
 WHERE year>= 2005
 GROUP BY  product_category)AS x));
```



Minimum rating of a product in a specific product category

SELECT p.product_id, p.product_category,

round(min(p.avg_rating),

2) AS Min_Rating

FROM

(SELECT product_id,

product_category,

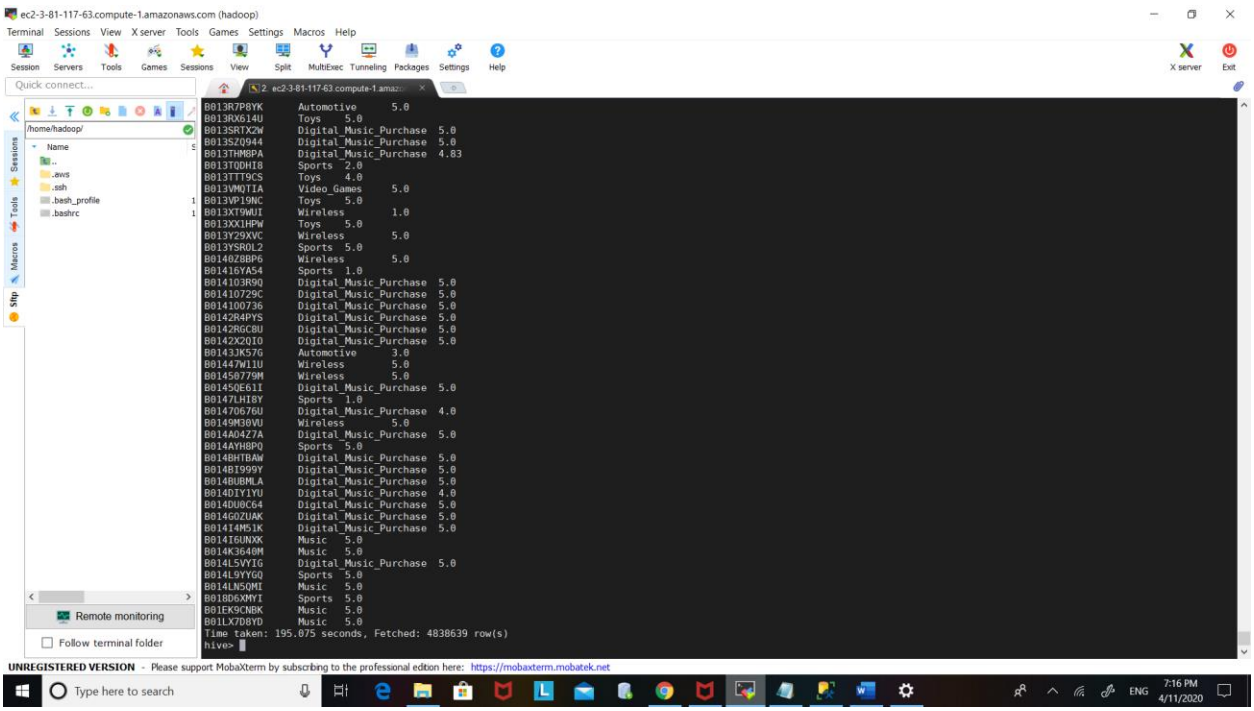
avg(star_rating) AS avg_rating

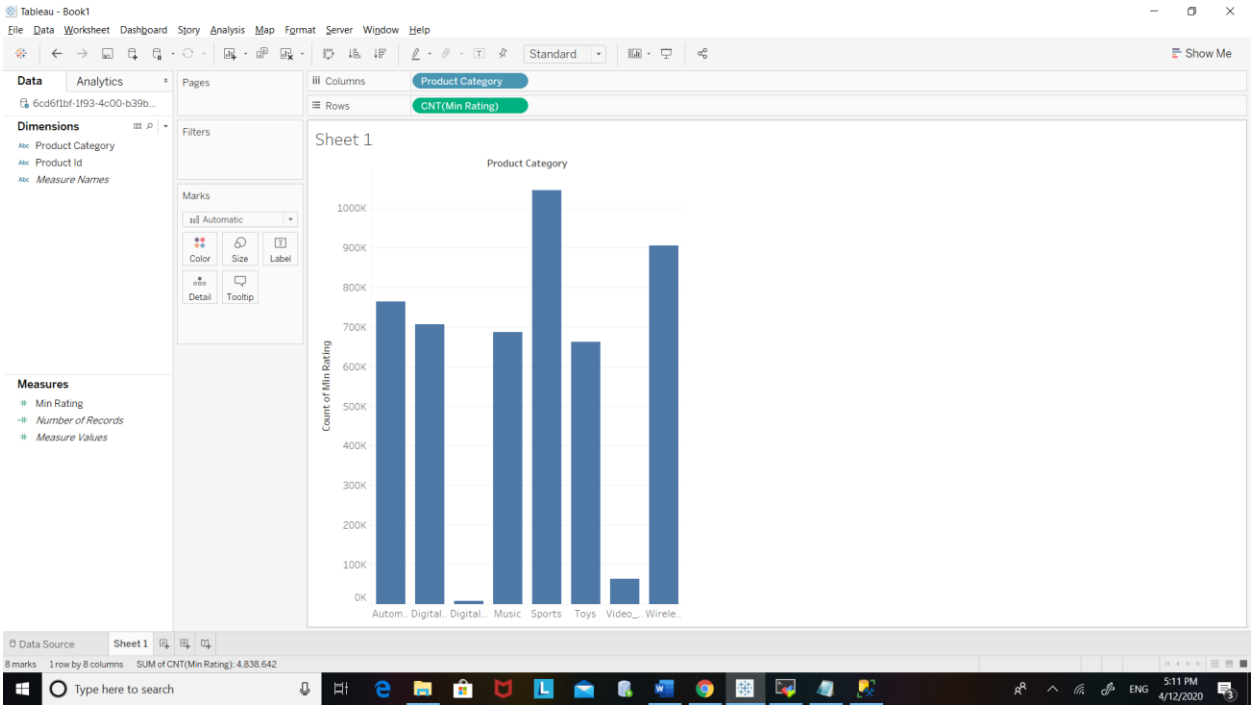
FROM amazon_review.amazon_review_filtered_data

WHERE year>= 2005

GROUP BY product_id, product_category) AS p

GROUP BY p.product_id, p.product_category;





The above graph shows the minimum rating per product.

Products having highest percentile of star ratings given by customers:

SELECT x.product_id, x.product_category,x.star_rank,

round(x.star_rank,

2) AS Rank_Percentile from

(SELECT y.product_id,

y.product_category,

PERCENT_RANK()

OVER (partition by y.product_category

ORDER BY y.avg_rating desc) AS star_rank

FROM

(SELECT product_id,

product_category,

avg(star_rating) AS avg_rating

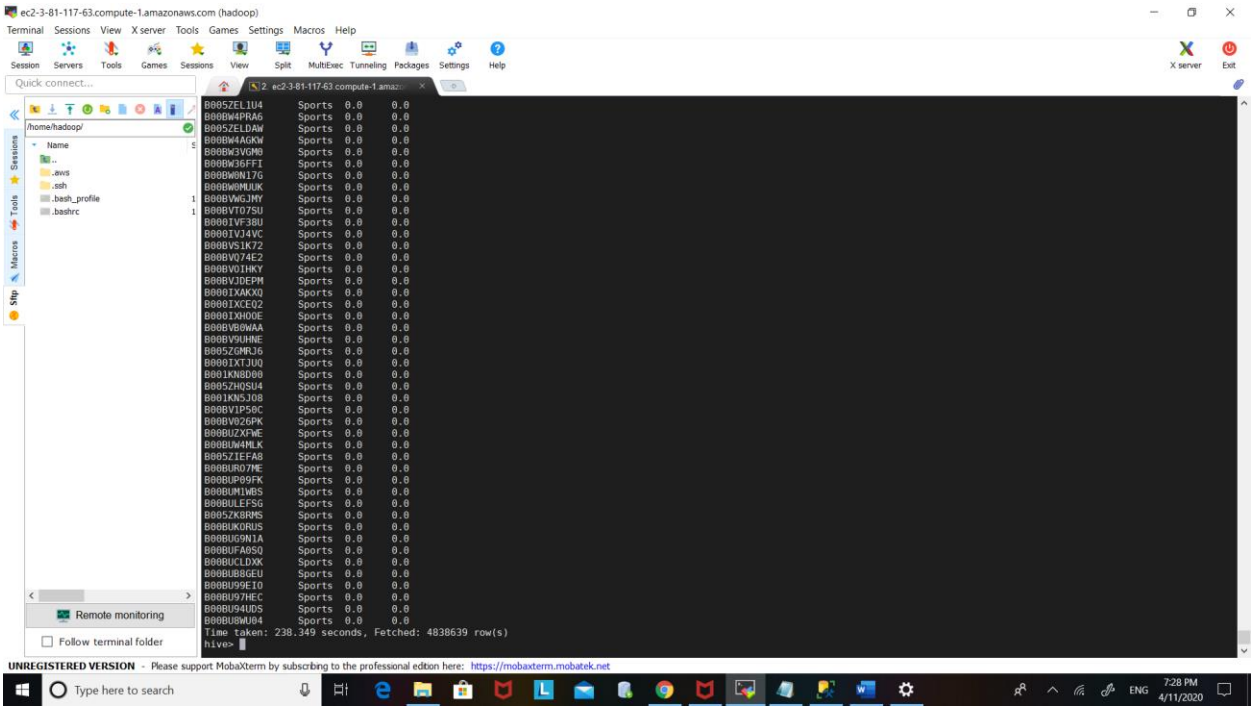
FROM amazon_review.amazon_review_filtered_data

WHERE year>= 2005

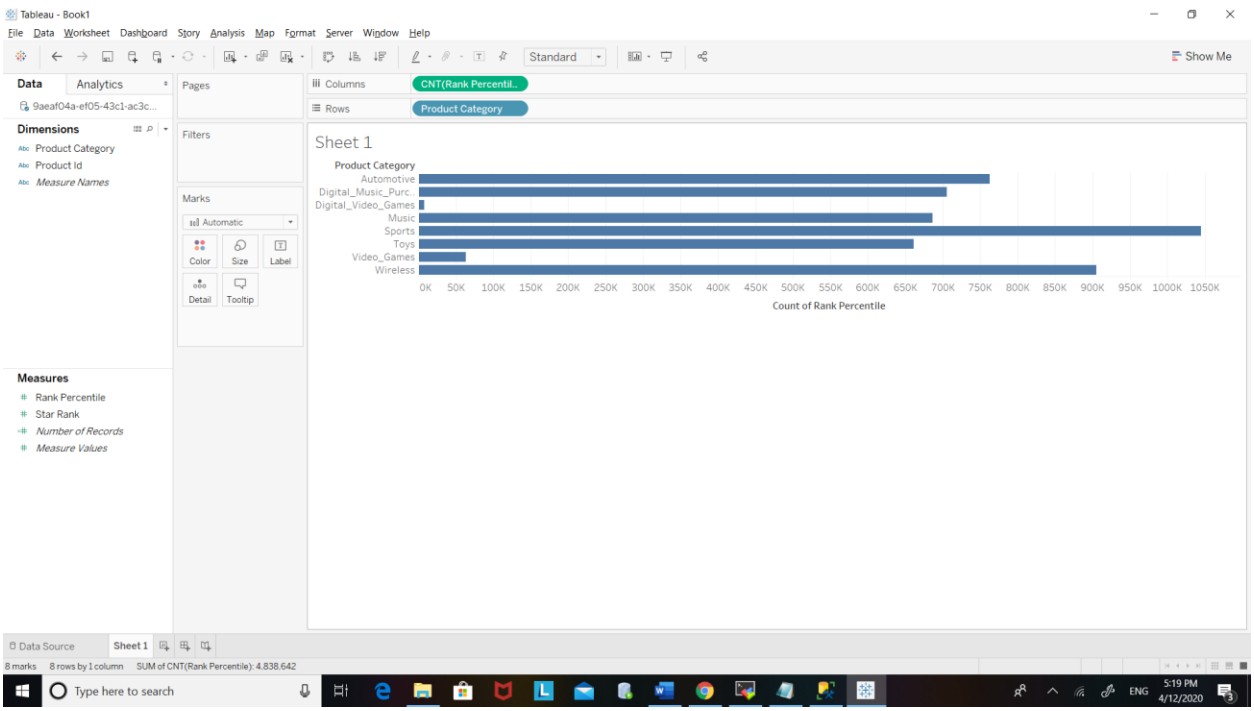
GROUP BY product_id,product_category)as y)as x order by x.star_rank

desc;

Output:



Visualization:



From the above graph we can conclude that Digital Video Game has the least percentile count and Sports Category has the highest.

Conclusion:

After running various queries and analyzing them we could draw several conclusions.

Some being as simple as the Number of Customers increasing over time to the same product having different ratings over two categories.

Visualizations using Tableau were completed for most of the queries.

These Visualizations made it easy to read the output and draw conclusions.

Digital Video Game has the least percentile of star ratings given by customers and Sports has the highest.

The Year 2014 had the most no of customers, the most number of reviews and verified and unverified users.

The Year 2015 saw a drop in the number for the aforementioned categories.

Similarly Digital Video Game had the minimum rating for the star rating among the various categories and Sports had the maximum star rating among all the categories.

The least rated products and highest rated products could also be separated using the queries.

We could come to a conclusion that there are Users who are reviewing products in more than one category.

We used Aggregate function to show various results on the various categories.

Standard deviation, Max and Min rating were found for each categories and their visualizations were completed to improve the knowledge of the business by making it easy to understand the various effects.

We could co-relate various categories like, Music and Video Games.

The dataset was cleaned, partitioned and various inferences were thus drawn with the help of hive Queries and Visualizations were completed as well to complete the Project Part 1 on the Amazon dataset.

References:

Class Notes

Class Mates : Prathamesh Namjoshi

Anuja Ghavate

W3 Schools

https://learnsql.com/course/window-functions?gclid=Cj0KCQjw-Mr0BRDyARIsAKEFbed8JEr8VChIef6mrHaKvCLkHe0Xucmkufhs5HdXV9xcGiVVlSxt5soaAkhLEALw_wcB

https://mode.com/sql-tutorial/sql-window-functions/

https://data-flair.training/blogs/hive-join/

https://data-flair.training/blogs/hive-internal-tables-vs-external-tables/

https://acadgild.com/blog/join-in-hive-optimization