AXG180147

Project Part 1:

Introduction and problem description

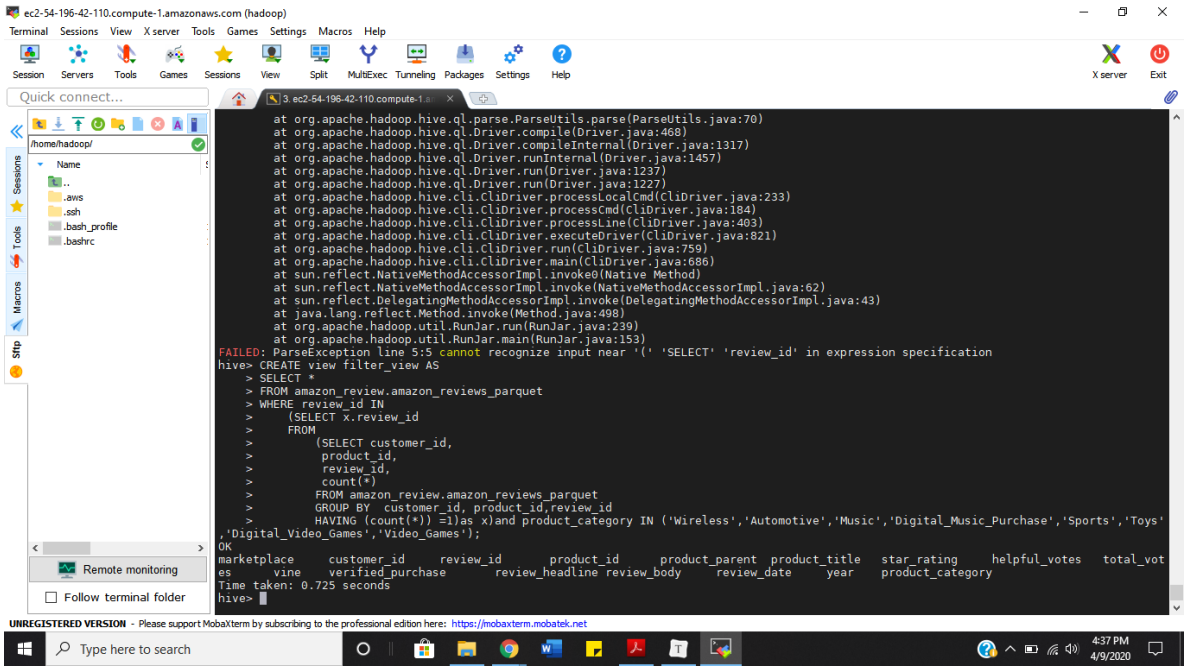
Initially, it was difficult to find out products having a high response rate as compared to other products. Due to this, it was difficult to decide on which products we should move our marketing focus to improve response and sales. Since customer rating is a very important parameter which can impact the purchase decision of future customers. Hence if some products are getting poor rating then those products need more focus and promotion. To find out these products, I have decided to analyze the amazon reviews database to gather information related to products amazon is offering and ratings those products are getting from customers. After analyzing Amazon Reviews Database, I realized for few product categories we are having very low ratings as compared to other product categories. My analysis involves the below steps:

Data Cleaning

In the Amazon Review database there are many records with multiple reviews by the same users for the same product. It is not appropriate to use this data for analysis, it may cause misinterpretation. Hence it is better to exclude such data from the database by creating a filter_view and excluded records having multiple reviews by the same users for the same product. Also, it is unreliable to use old data for analysis. So for analysis, I have included the data after 2005 and the data having product categories such as Wireless, Automotive, Music, Digital_Music_Purchase, Sports, Toys, Digital_Video_Games, Video_Games.

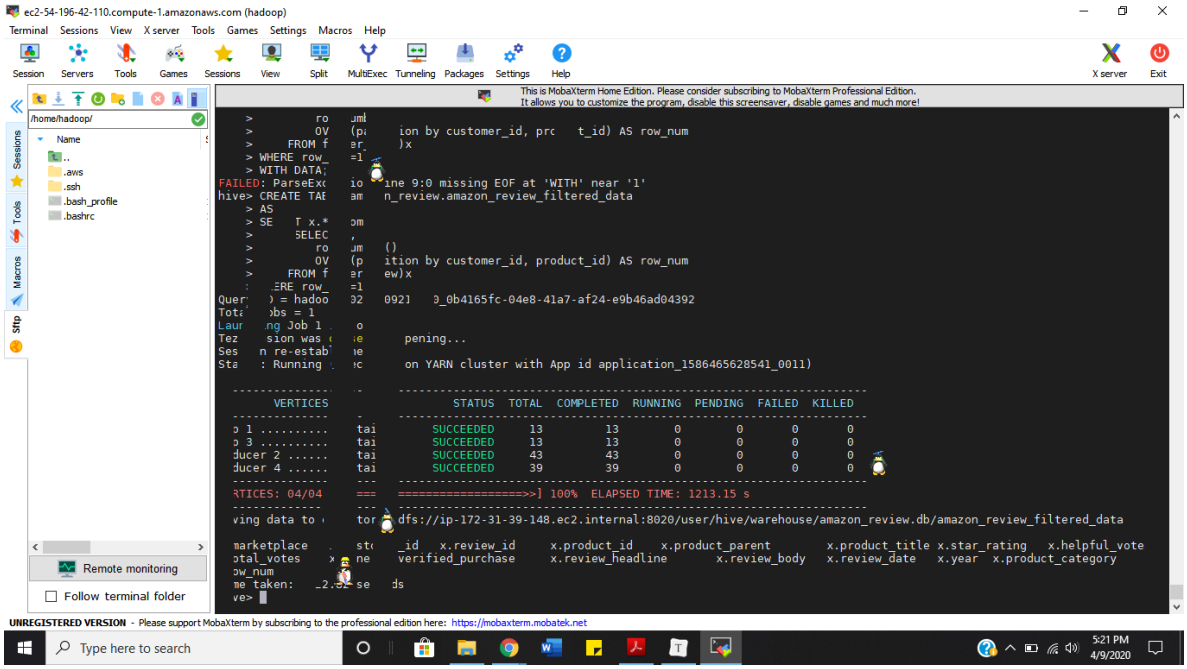
Create a View with Excluded data:

```
CREATE view filter_view 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','Digit
al_Video_Games','Video_Games');
```



Create a Table with Excluded data:

```
CREATE TABLE amazon_review.amazon_review_filtered_data
AS
SELECT x.* from
    (SELECT *,
        row_number()
        OVER (partition by customer_id, product_id) AS row_num
    FROM filter_view)x
WHERE row_num=1;
```

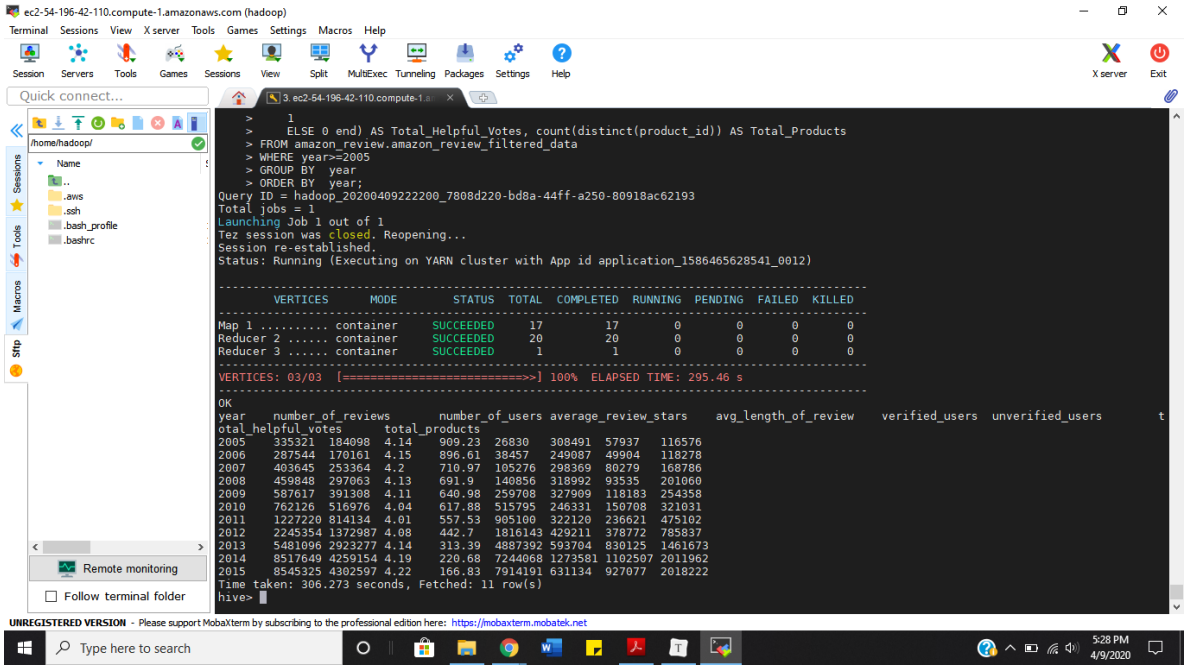


Basic Exploratory Analysis:

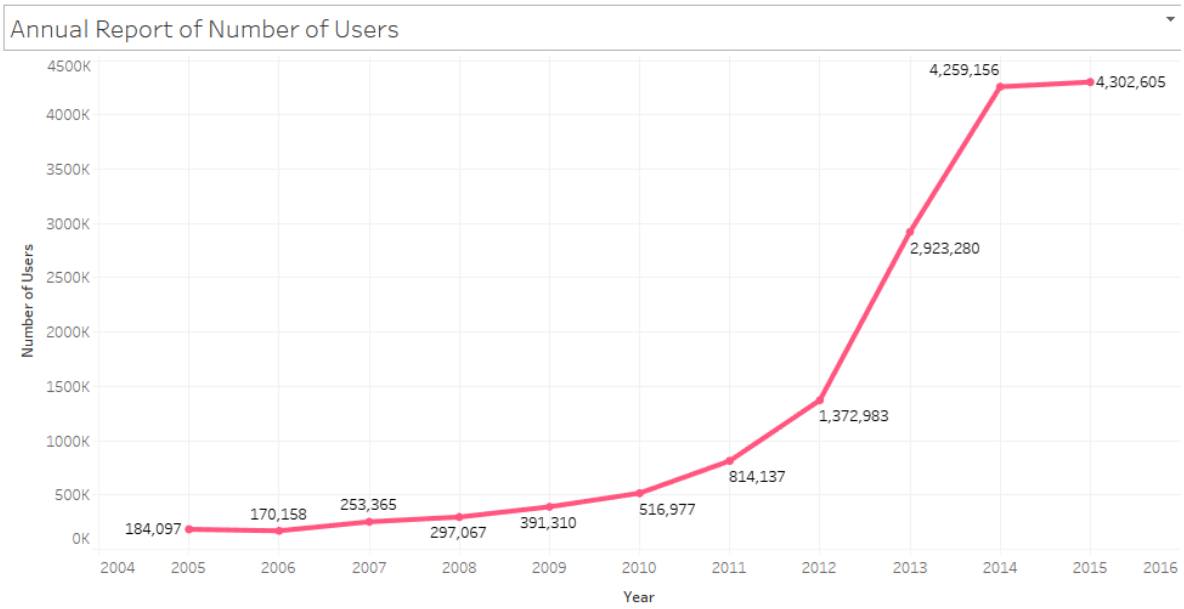
Carried out basic exploratory analysis to understand a basic overview of the Amazon Review Database.

Calculated different parameters like Number_of_Reviews, Number_of_Users, Average_Review_Stars, Avg_Length_of_Review, Verified_Users, Unverified_Users, Total_Helpful_Votes, Total_Products.

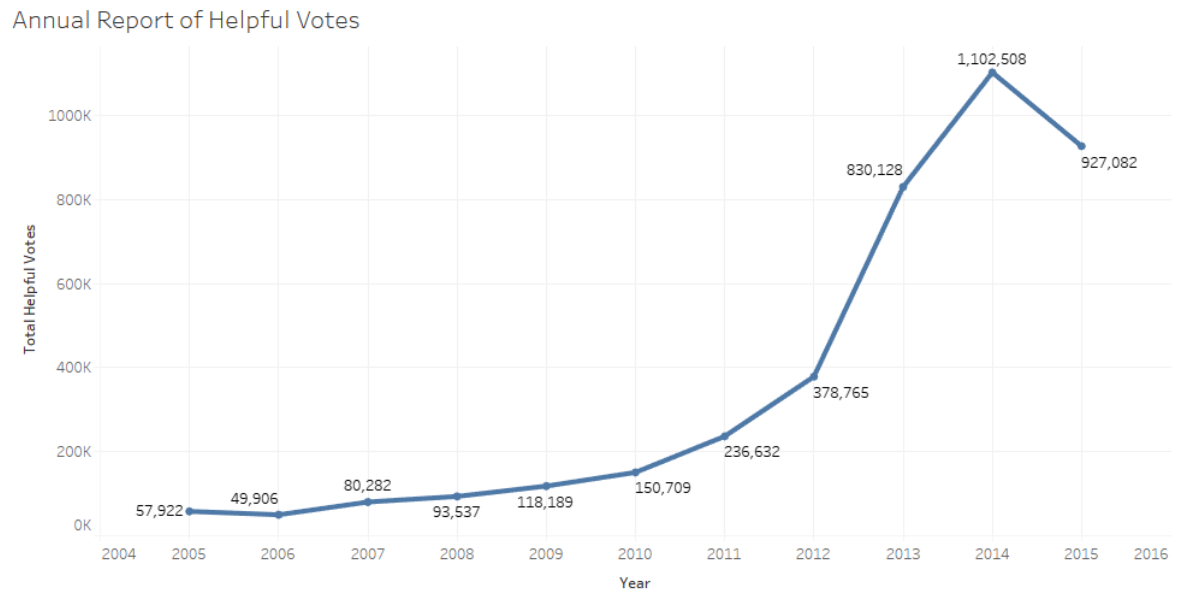
```
SELECT year,
        count(review_id) AS Number_of_Reviews,
         count(distinct(customer_id)) AS Number_of_Users,
         round(avg(star_rating),
        2) AS Average_Review_Stars,
         round(avg(length(review_body)),
        2) AS Avg_Length_of_Review,
         sum(case
    WHEN verified_purchase ='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 Total_Helpful_Votes, count(distinct(product_id)) AS
Total_Products
FROM amazon_review.amazon_review_filtered_data
WHERE year>=2005
GROUP BY  year
ORDER BY  year;
```



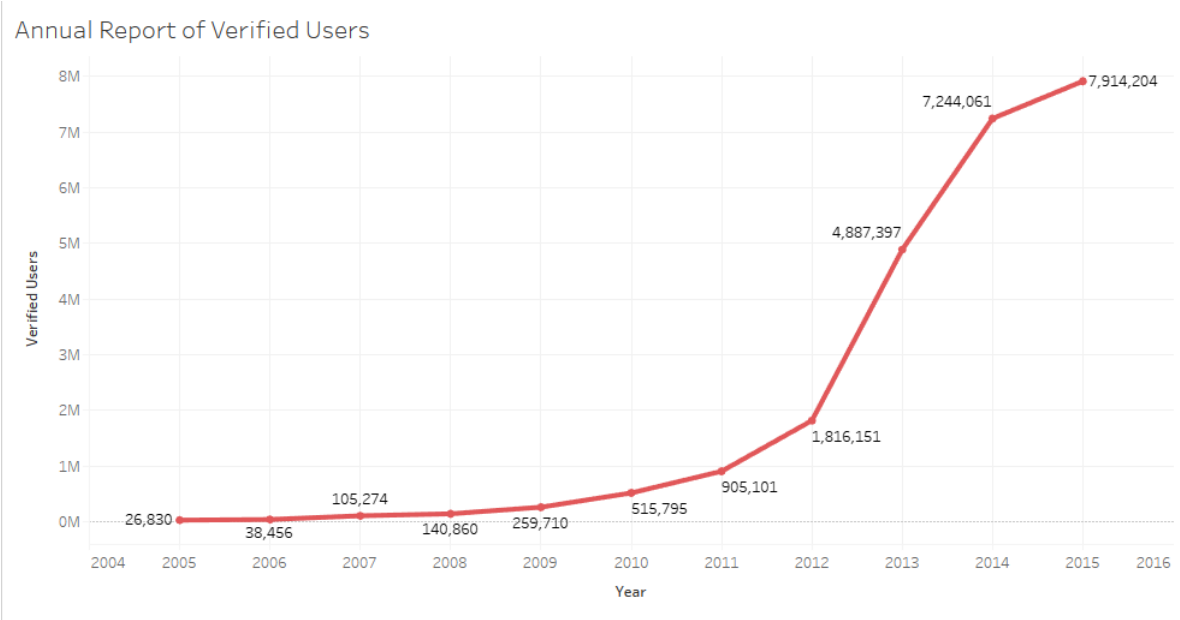
VISUALIZATION:



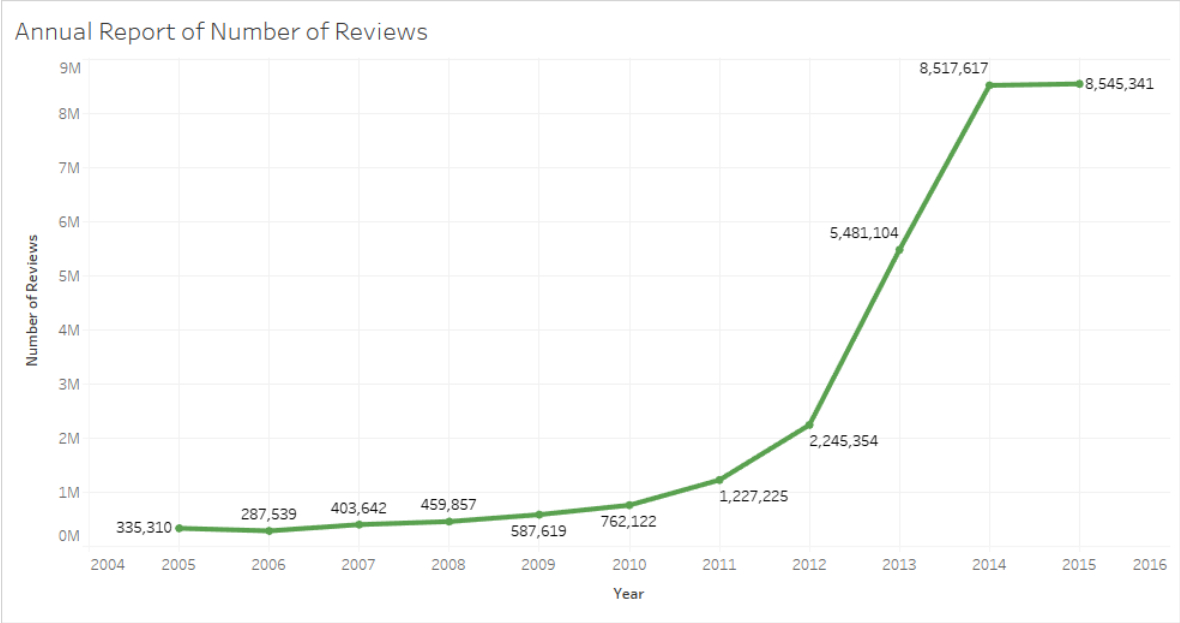
From this graph, we can analyze the number of customers increased gradually from 2008 till 2014 but after 2014, there is no significant growth in customer count.



from this graph, number of helpful votes increased from 2005 to 2014. But after 2014, number of helpful votes decreased drastically.



From this graph, we can see number of verified users increased gradually from 2005 to 2015.



From this graph, we can interpret number of reviews increased gradually from 2005 to 2012. After 2012, count increased drastically till 2014. After 2014, there is slight decrease in number of reviews.

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

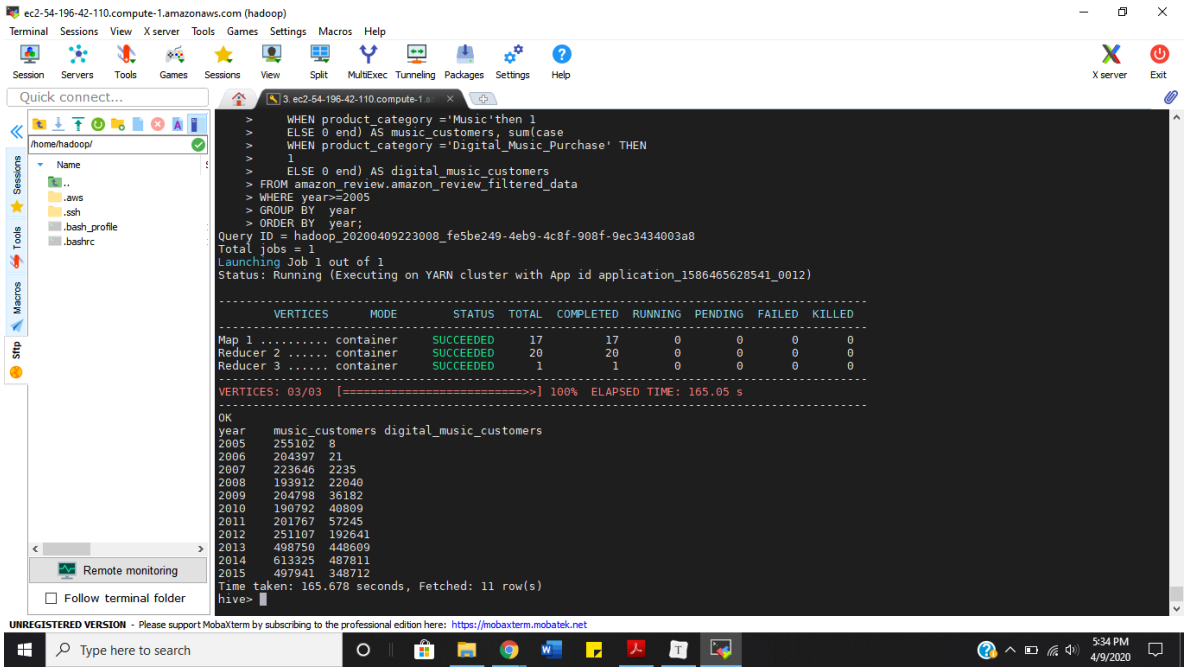
Performed detailed analysis to find out if there is a correlation between different product categories like

1. music and Digital_Music_Purchase
2. Video_Games and Digital_Video_Games.

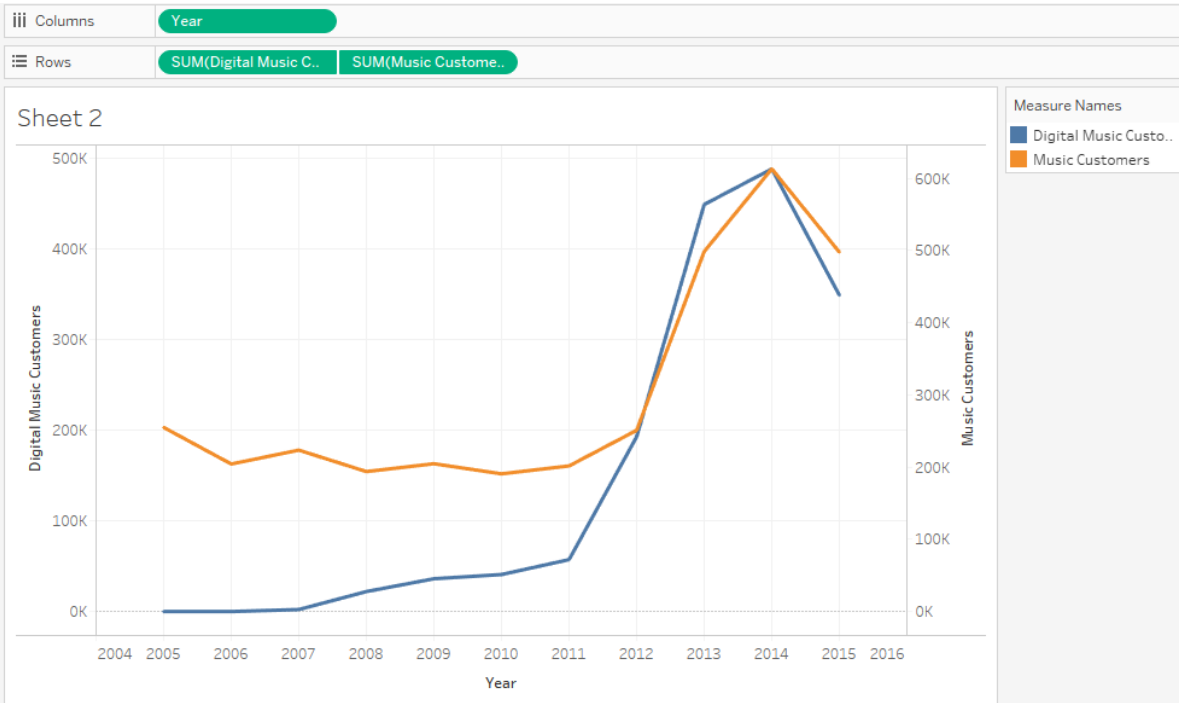
Correlation between the categories over time

Analysis of Music Related Category

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



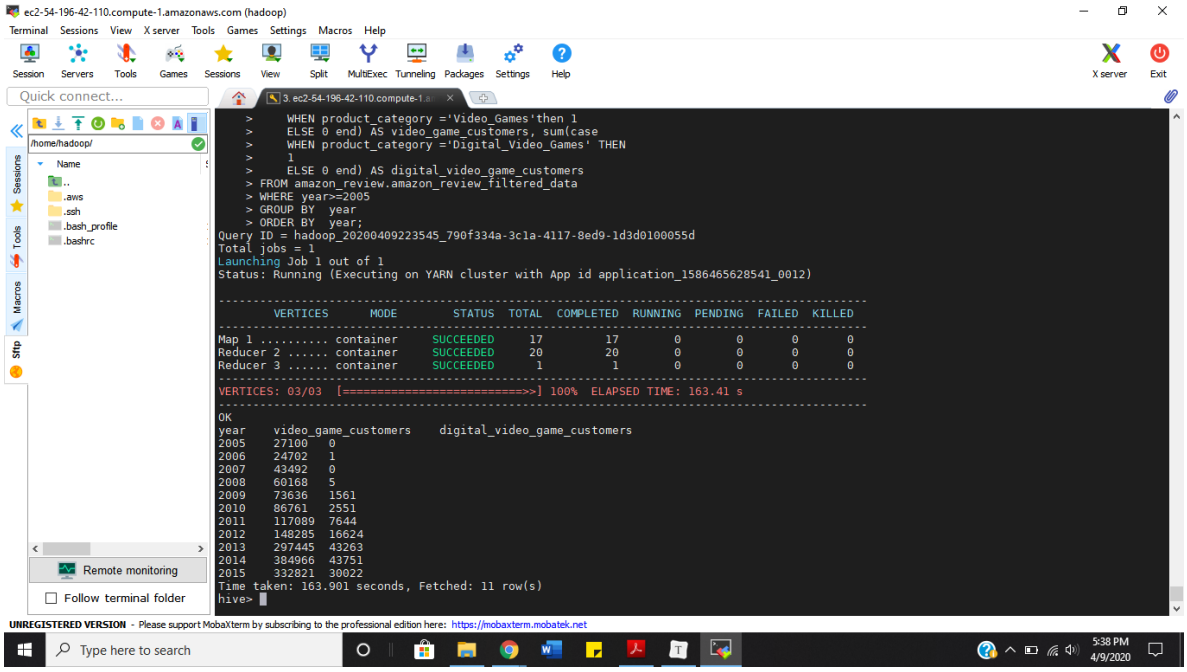
VISUALIZATION:



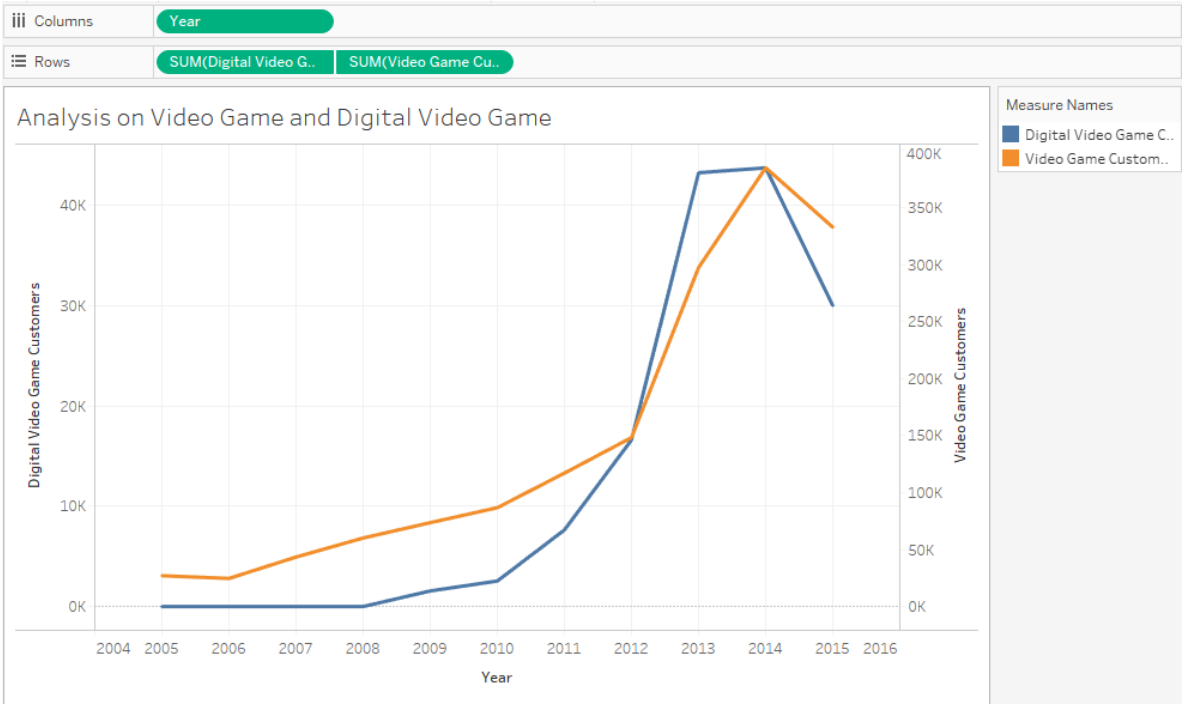
For the music category, there was no drastic change in the number of customers from 2005 to 2012 but after 2012 to 2014 the number of customers increased but after 2014, the number of customers decreased drastically. Similarly, for the Digital Music category, initially, in 2005 there was not a single customer who reviewed digital music products but from 2007 number of customers started increasing till 2014.

Analysis on Game Related Category

```
SELECT year,
         sum(case
    WHEN product_category ='Video_Games'then 1
    ELSE 0 end) AS video_game_customers, sum(case
    WHEN product_category ='Digital_Video_Games' THEN
    1
    ELSE 0 end) AS digital_video_game_customers
FROM amazon_review.amazon_review_filtered_data
WHERE year>=2005
GROUP BY  year
ORDER BY  year;
```



VISUALIZATION:



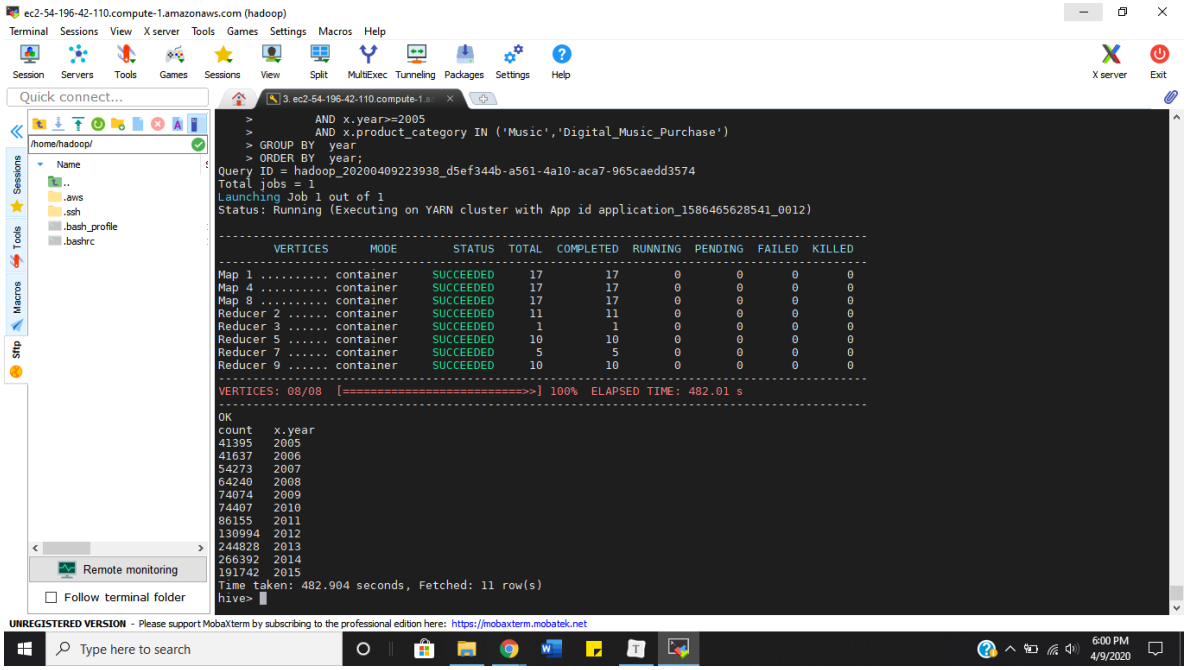
For the Video_Game category, the number of customers gradually increased from 2005 to 2012. After 2012 to 2014 the number of customers increased drastically. Similarly, for the Digital Music category, initially, from 2005 to 2008 there was not a single customer reviewing digital music products but after 2008 till 2014 number of customers increased gradually. After 2014, the number of customers started decreasing.

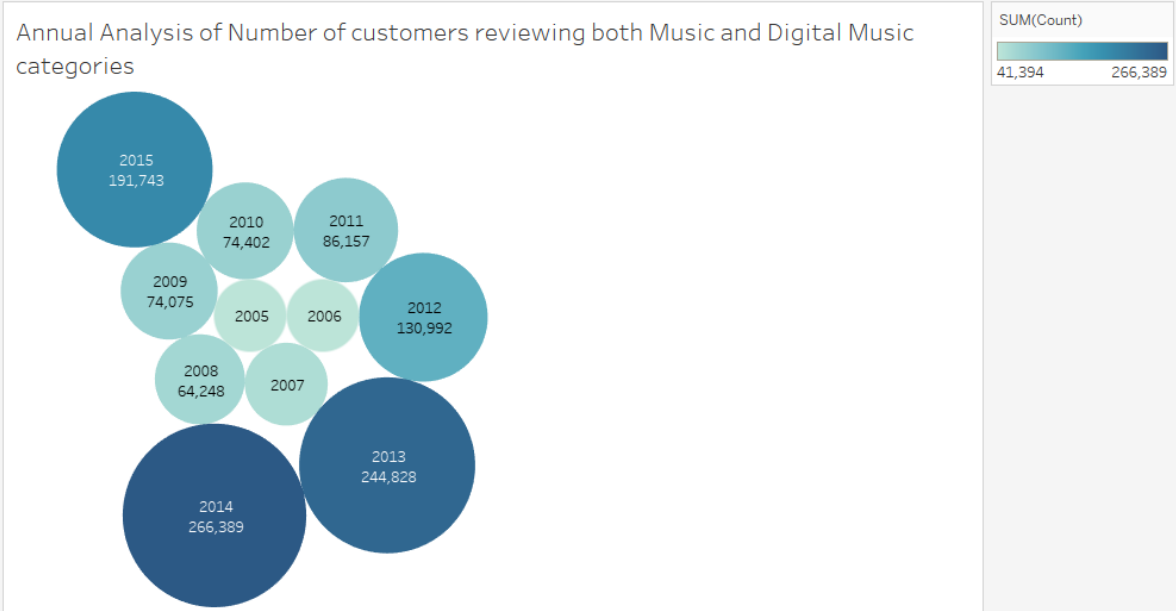
Are there the same users reviewing in both categories?

To find out whether there are same customer who are reviewing for products in both categories:

Analysis of music related categories:

```
SELECT count(x.customer_id) AS count,
        x.year
FROM amazon_review.amazon_review_filtered_data x,
    (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)y
WHERE x.customer_id=y.customer_id
        AND x.year>=2005
        AND x.product_category IN ('Music','Digital_Music_Purchase')
GROUP BY  year
ORDER BY  year;
```

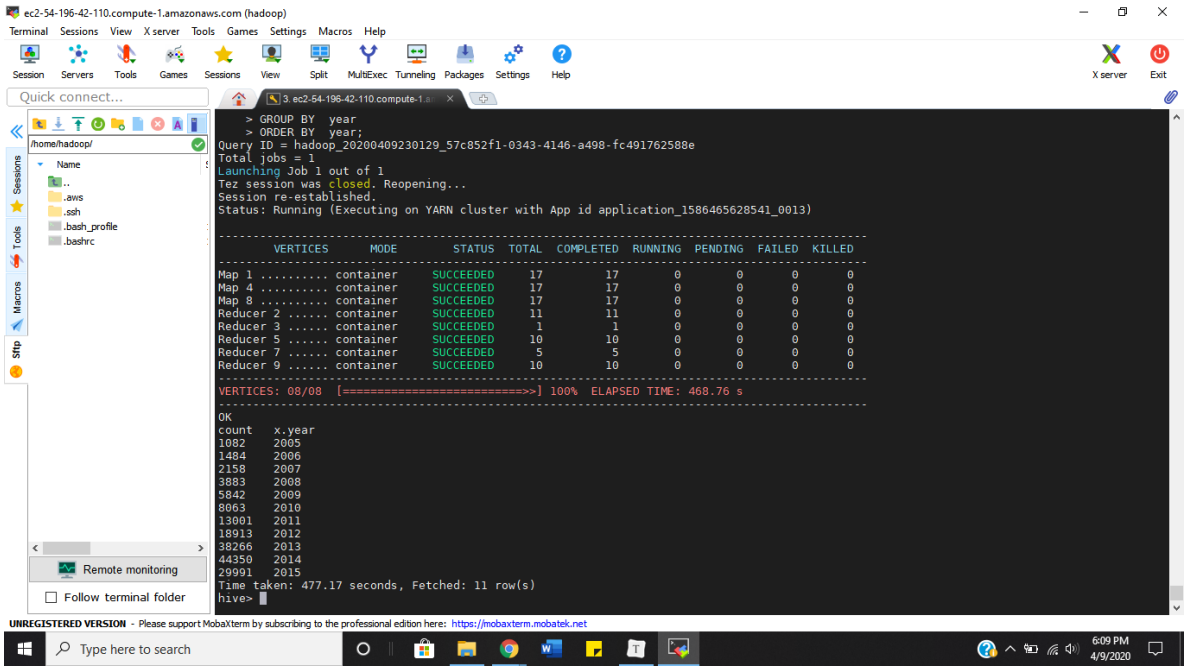


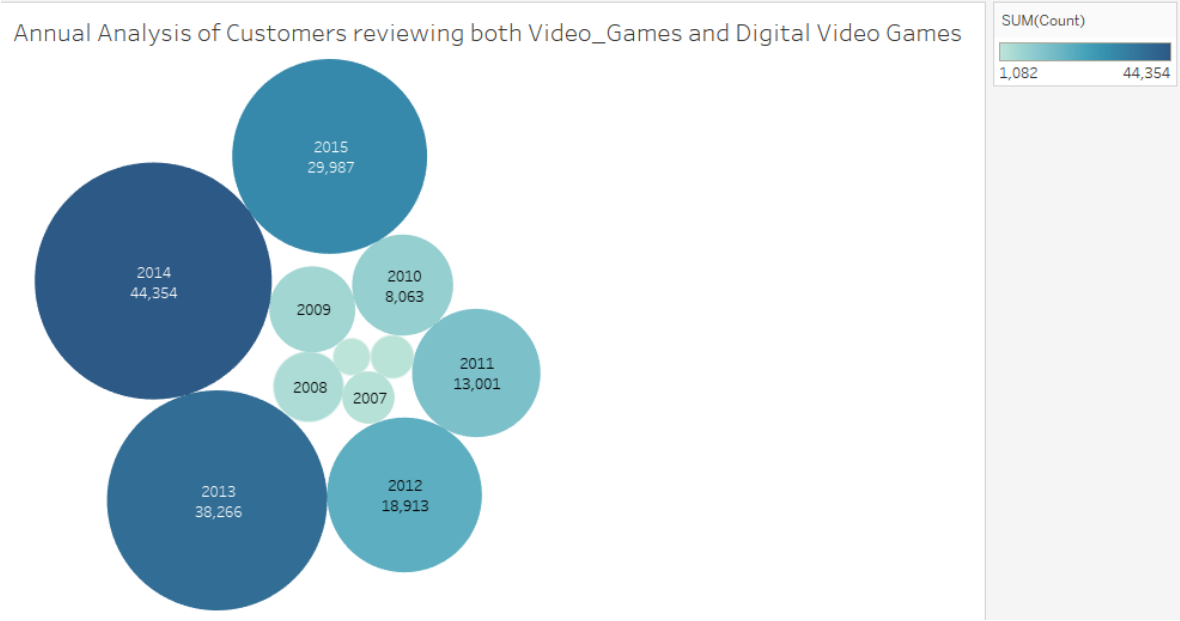


From the above bubble chart, we can interpret, in 2014 the number of customers reviewing both Music and Digital music categories was more as compared to other years.

Analysis of game related categories:

```
SELECT count(x.customer_id) AS count,
        x.year
FROM amazon_review.amazon_review_filtered_data x,
    (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)y
WHERE x.customer_id=y.customer_id
        AND x.year>=2005
        AND x.product_category IN ('Video_Games','Digital_Video_Games')
GROUP BY  year
ORDER BY  year;
```



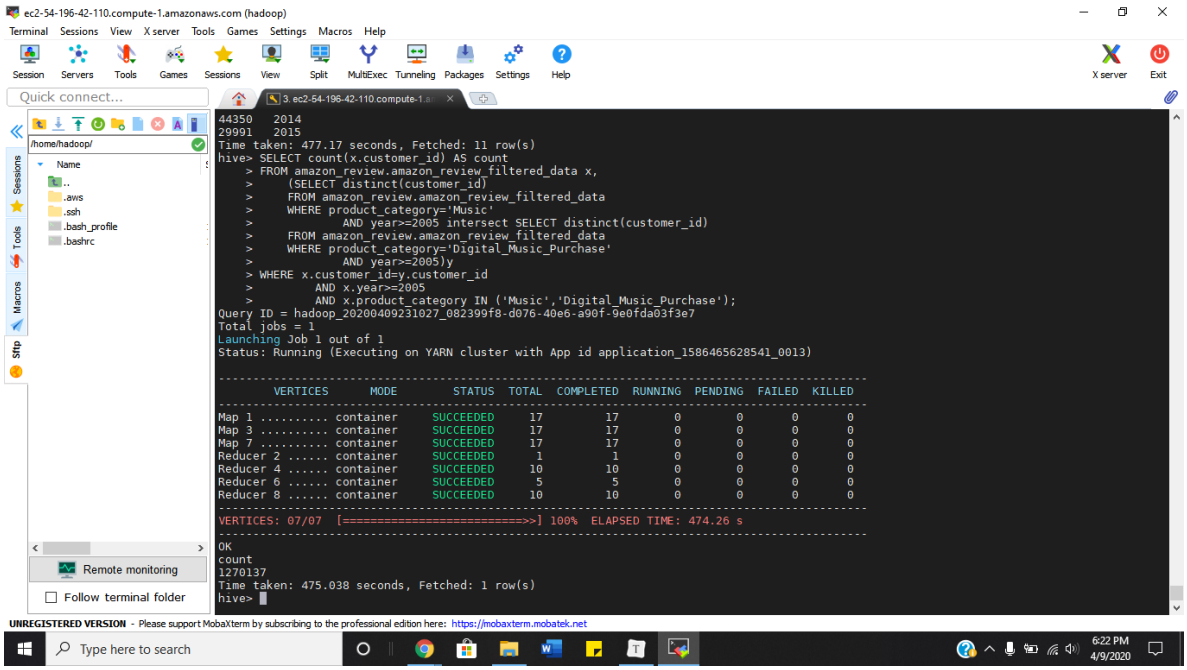


From the above bubble chart, we can conclude that in 2014 the number of customers reviewing both Video_Games and Digital video game categories was more as compared to other years.

Total Number of same users reviewing in both categories

Users reviewing Music and Digital_Music_Purchase categories:

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

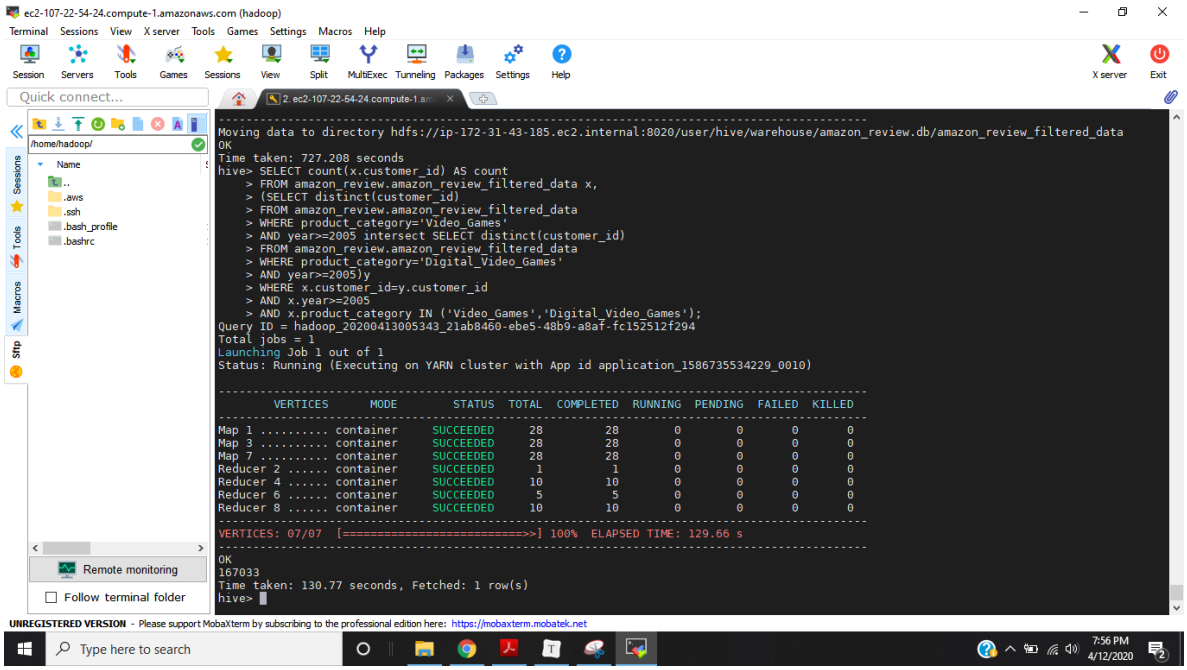


Results

	count
1	1270134

Users reviewing video_games and Digital_Video_Games categories:

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



Results

	count
1	167033

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

Performed analysis to find out whether there are similar items in both categories having the same rating.

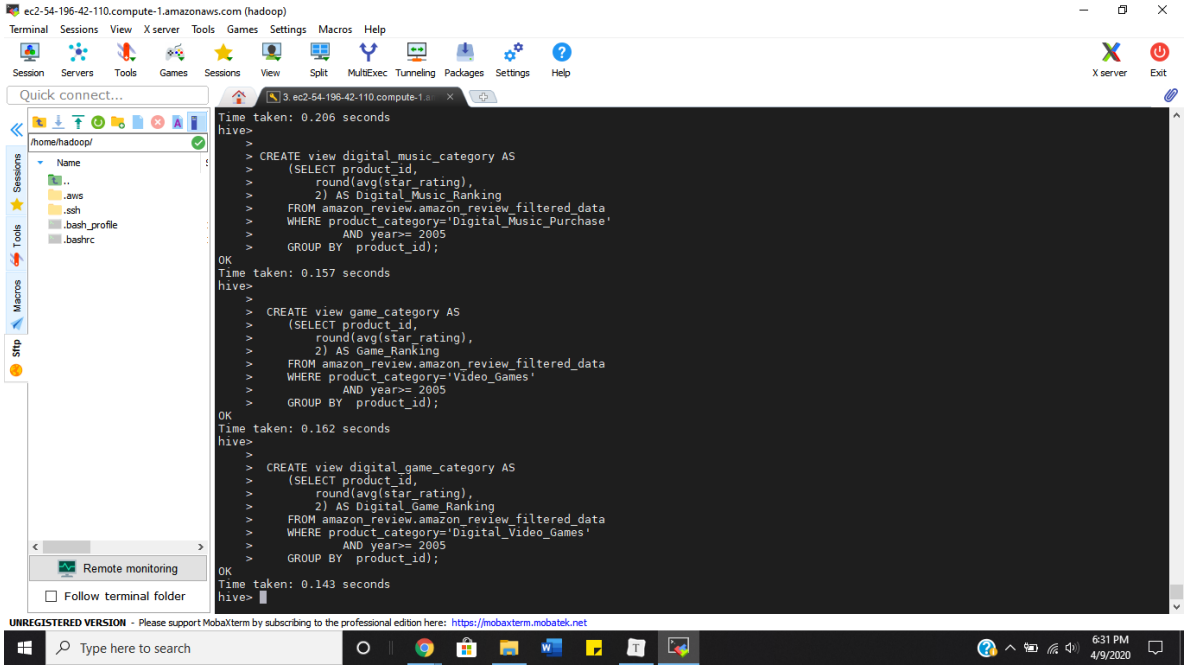
```
CREATE view music_category AS
    (SELECT product_id,
         round(avg(star_rating),
         2) AS Music_Ranking
    FROM amazon_review.amazon_review_filtered_data
    WHERE product_category='Music'
            AND year>= 2005
    GROUP BY  product_id);


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 game_category AS
    (SELECT product_id,
        round(avg(star_rating),
```

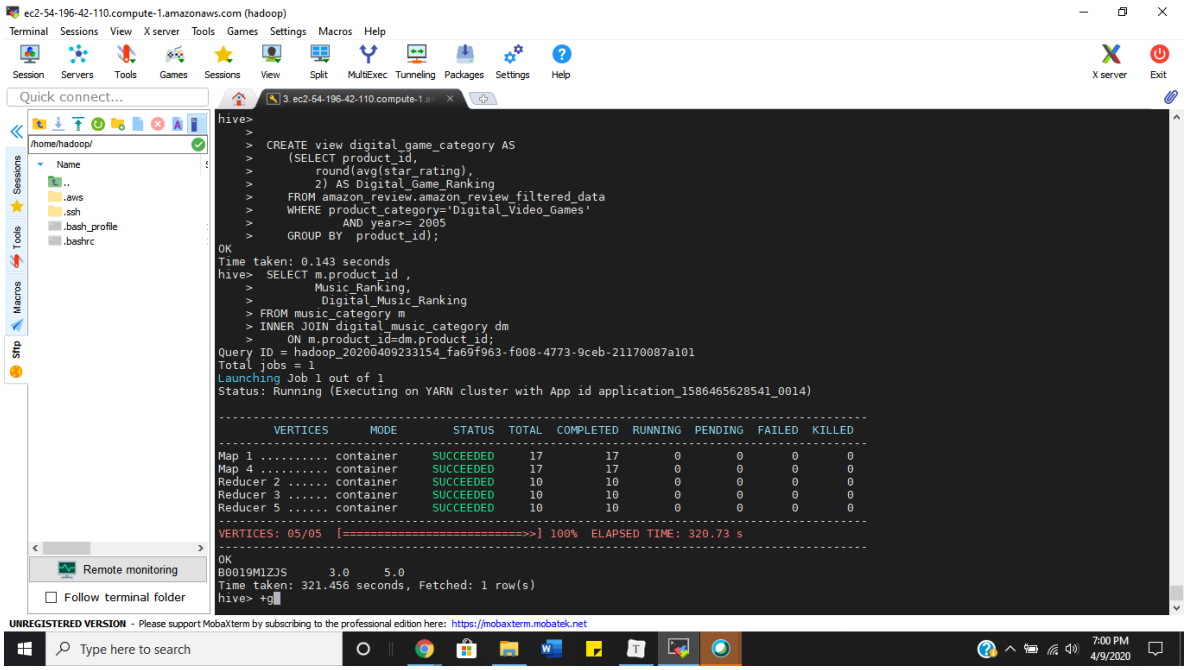
```
        2) AS Game_Ranking
    FROM amazon_review.amazon_review_filtered_data
    WHERE product_category='Video_Games'
            AND year>= 2005
    GROUP BY  product_id);


CREATE view digital_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:

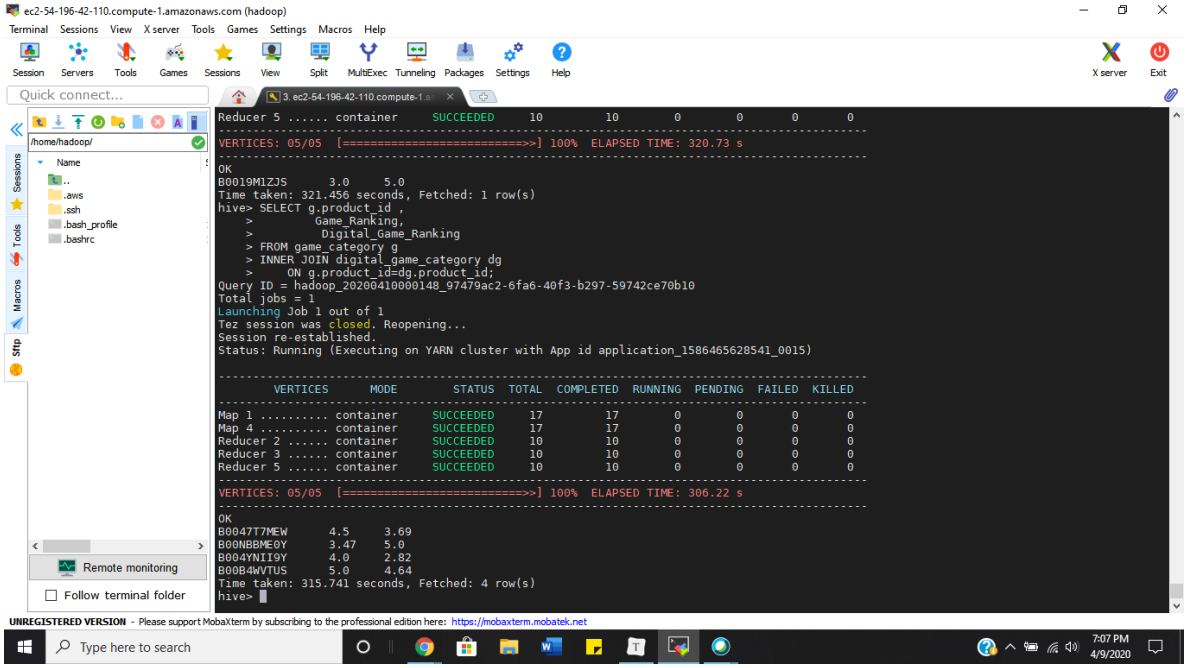
```
 SELECT m.product_id ,
        Music_Ranking,
         Digital_Music_Ranking
FROM music_category m
INNER JOIN digital_music_category dm
    ON m.product_id=dm.product_id;
```



From this, we can interpret that we have the same item in both Music and Digital_Music_Purchase categories with different ratings.

Analysis between Video_Games and Digital_Game_Ranking category:

```
 SELECT g.product_id ,
        Game_Ranking,
         Digital_Game_Ranking
FROM game_category g
INNER JOIN digital_game_category dg
    ON g.product_id=dg.product_id;
```



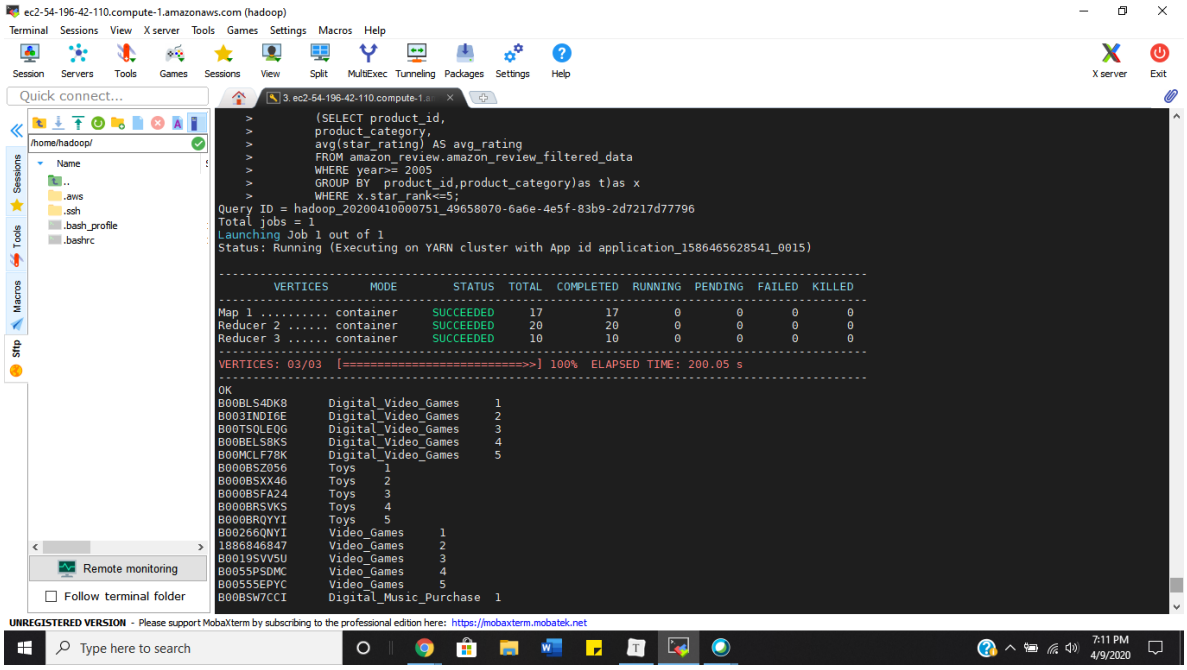
From this, we can interpret that we have the same items in both Music and Digital_Music_Purchase categories with a different rating.

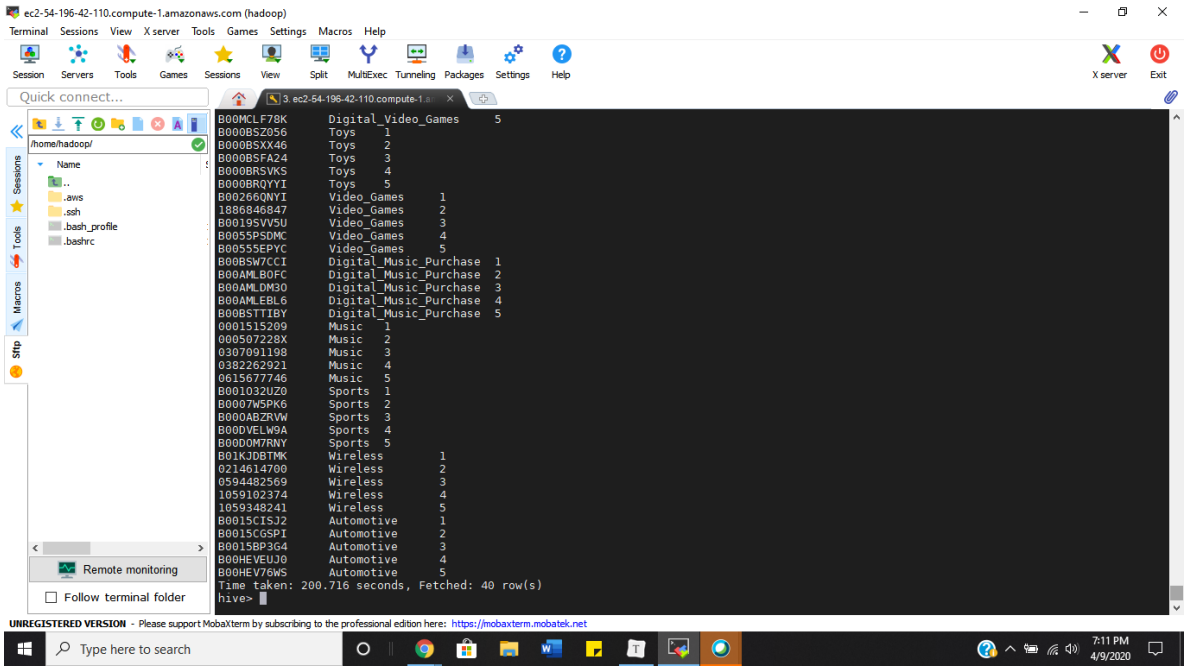
Hive Advanced Functions

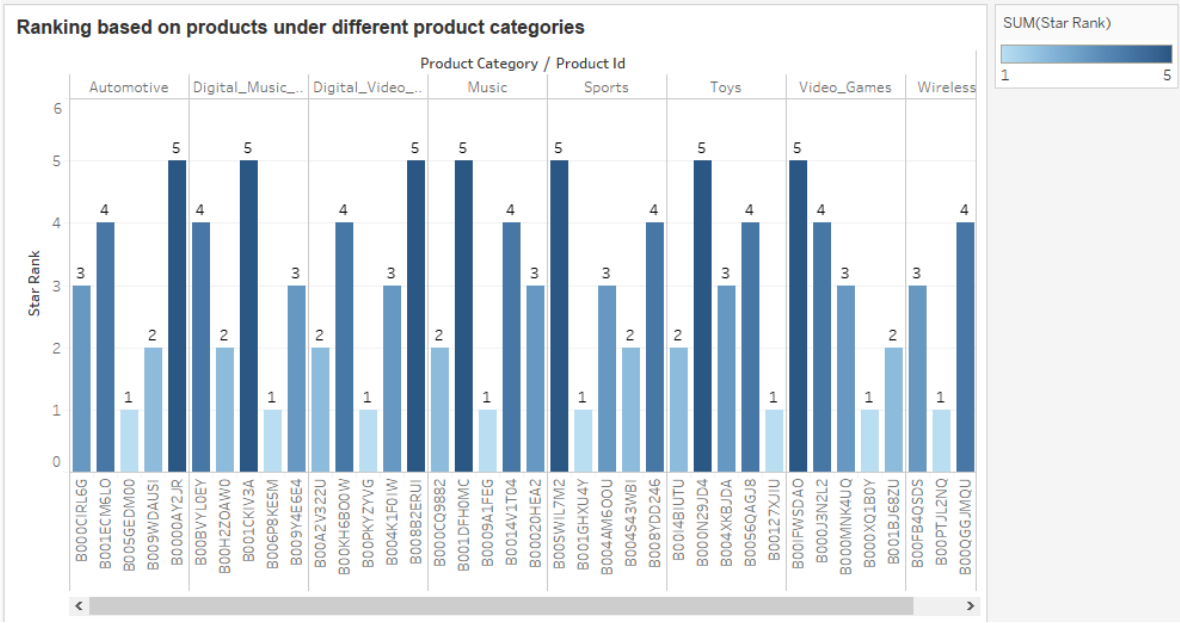
Ranking based on products under different product categories

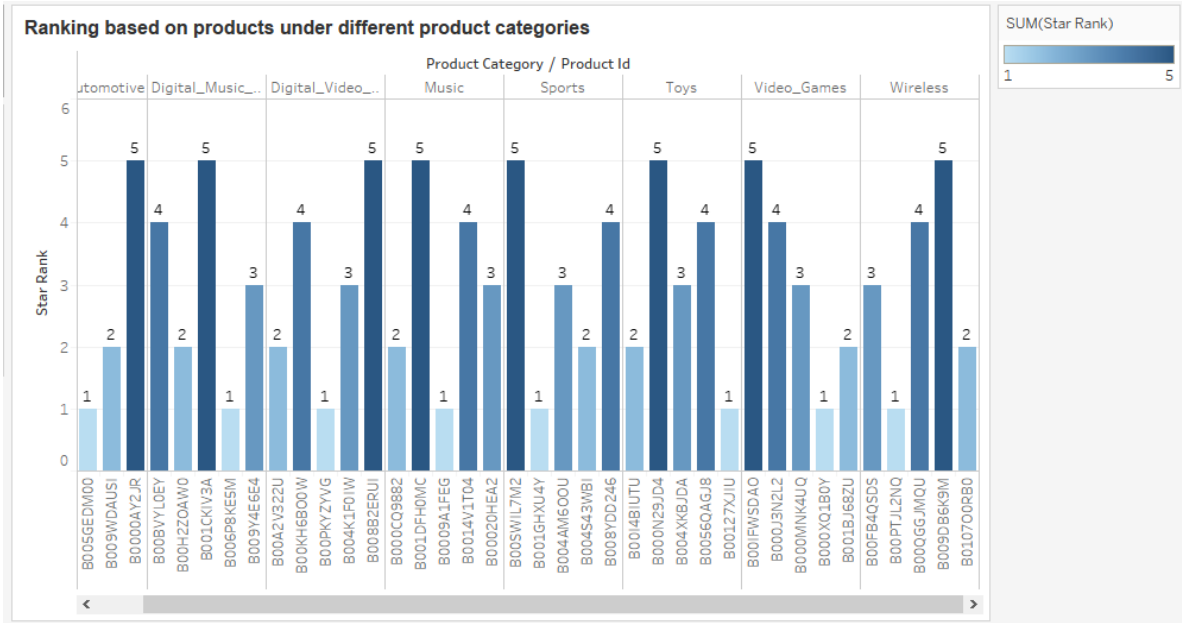
Calculated rank of different products in a specific product category to find out popular products and least rated products in that category.

```
SELECT x.product_id,
        x.product_category,
        x.star_rank
FROM
    (SELECT t.product_id,
        t.product_category,
        Row_number()
        OVER (partition by t.product_category
        ORDER BY  t.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 t)as x
        WHERE x.star_rank<=5;
```







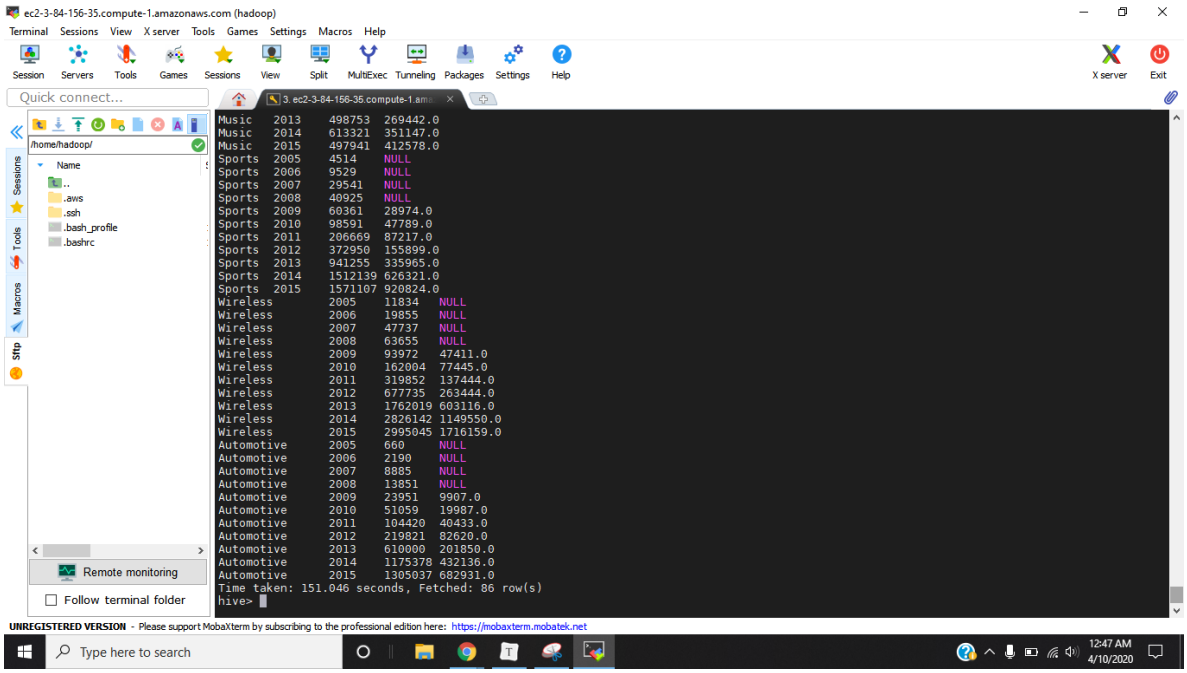


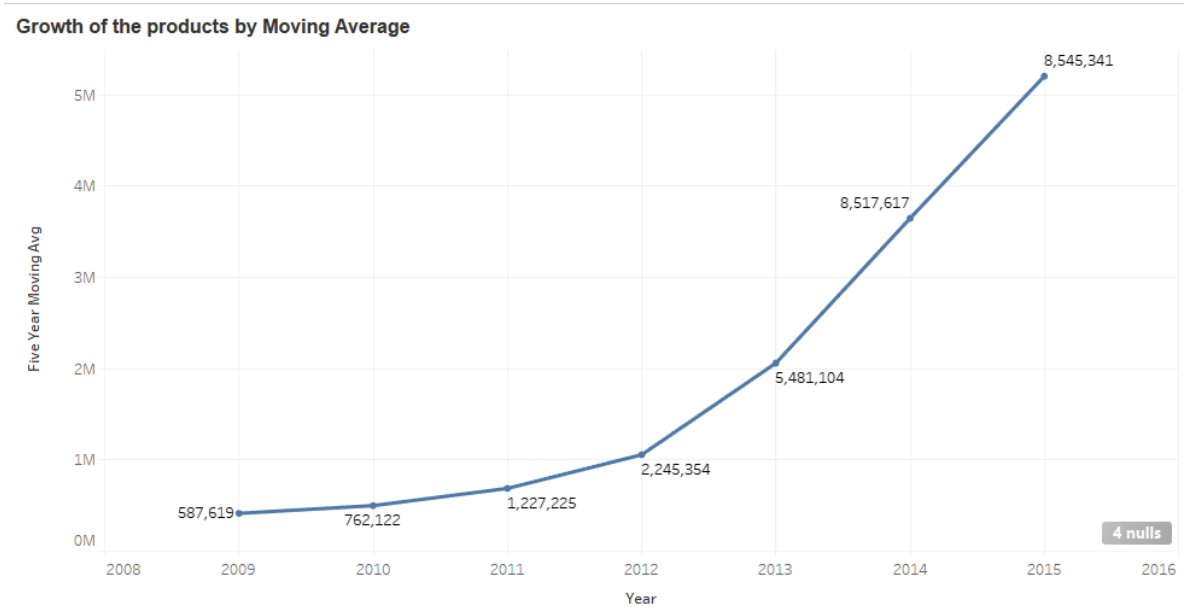
From the above bar graph, we can see products represented in light color have a low rating as compared to products represented in a darker shade. From this, we can find out the products on which we need to focus more.

Compared the growth of the products by calculating moving average:

Calculated moving average for product categories to evaluate growth.

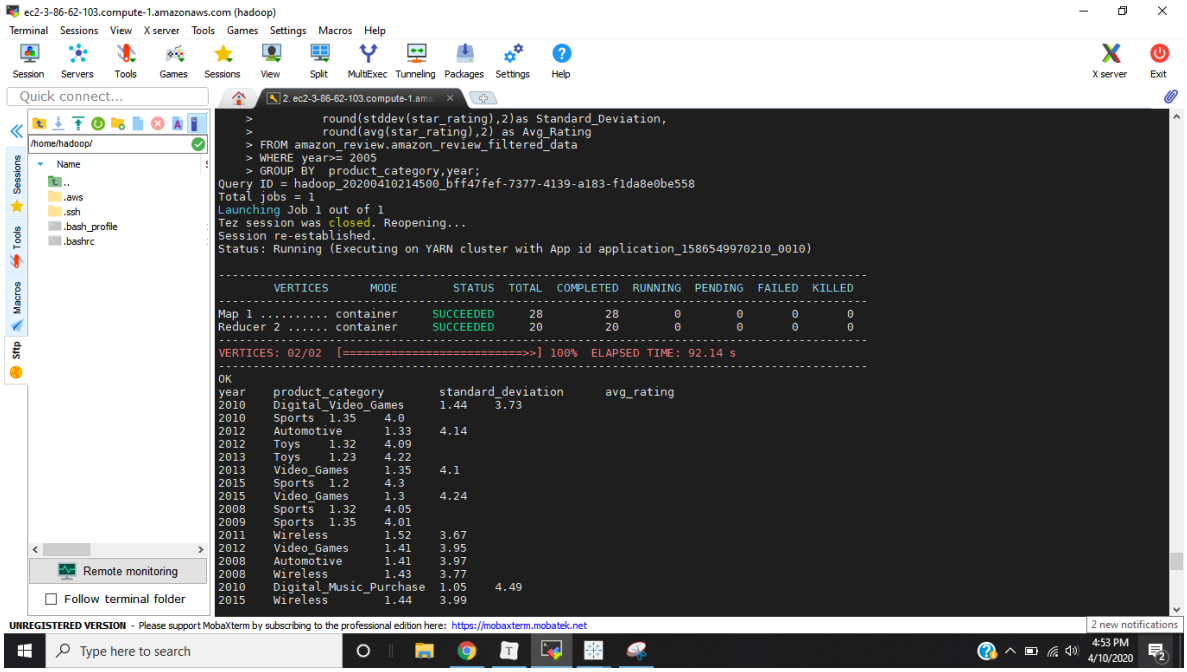
```
SELECT x.product_category,x.year,x.count,
         (case
    WHEN row_number()
    OVER (partition by x.product_category order by x.year) > 4 THEN
    round(AVG(x.count)
    OVER (PARTITION BY x.product_category
ORDER BY  x.year ROWS 4 PRECEDING)) end) AS five_year_moving_avg
FROM
    (SELECT product_category,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

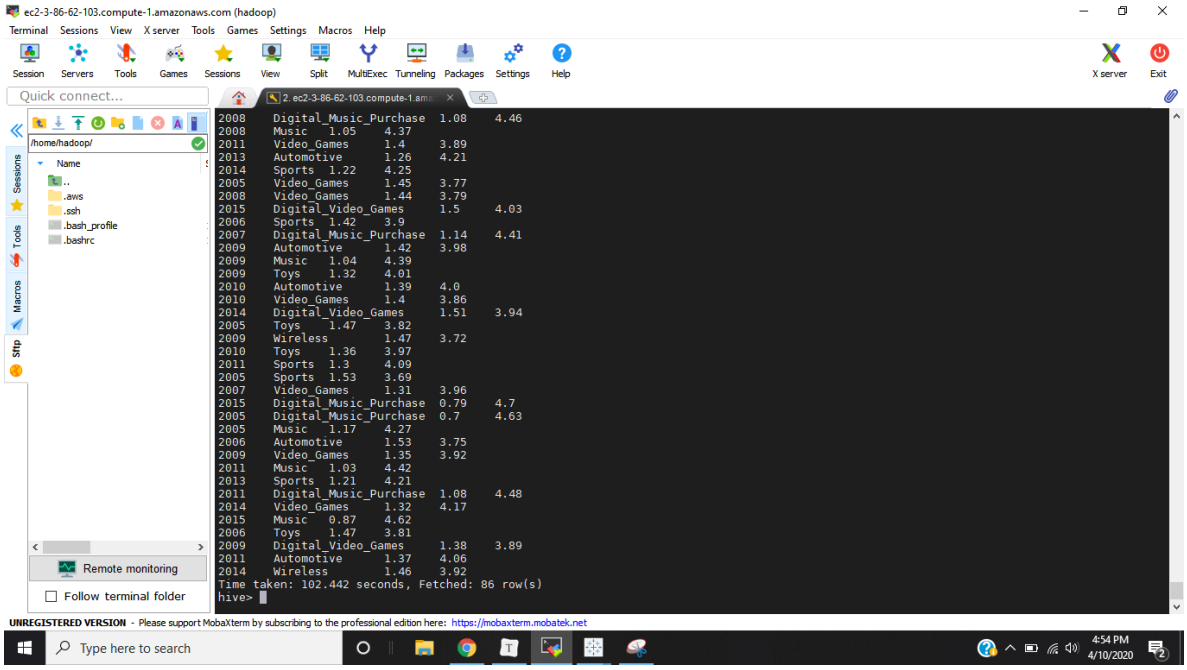


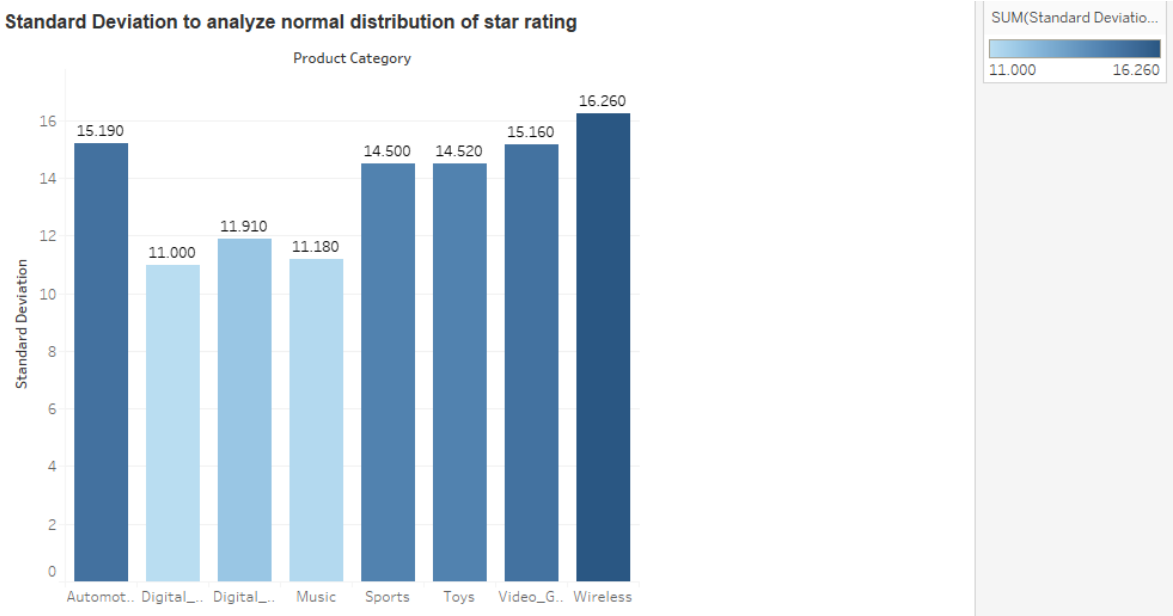


Calculated Standard Deviation to analyze normal distribution of star rating of particular product category

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



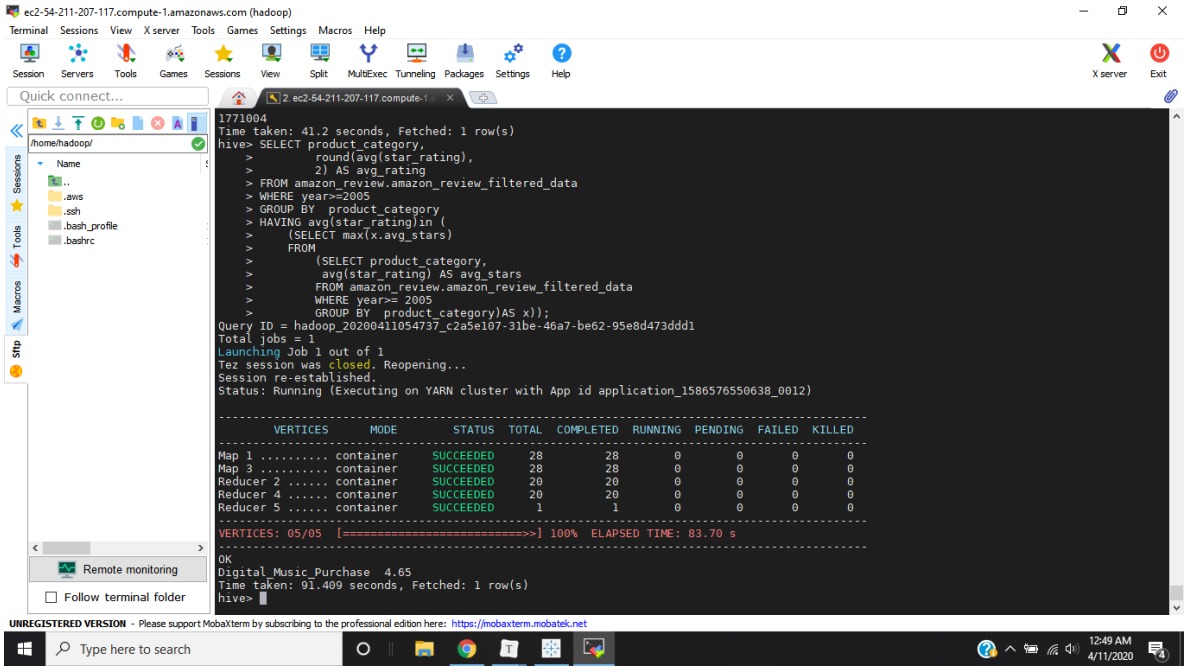




Maximum Rating of a product in specific product category:

Calculated maximum rating of a product in a specific product category to find out the most popular product in that category.

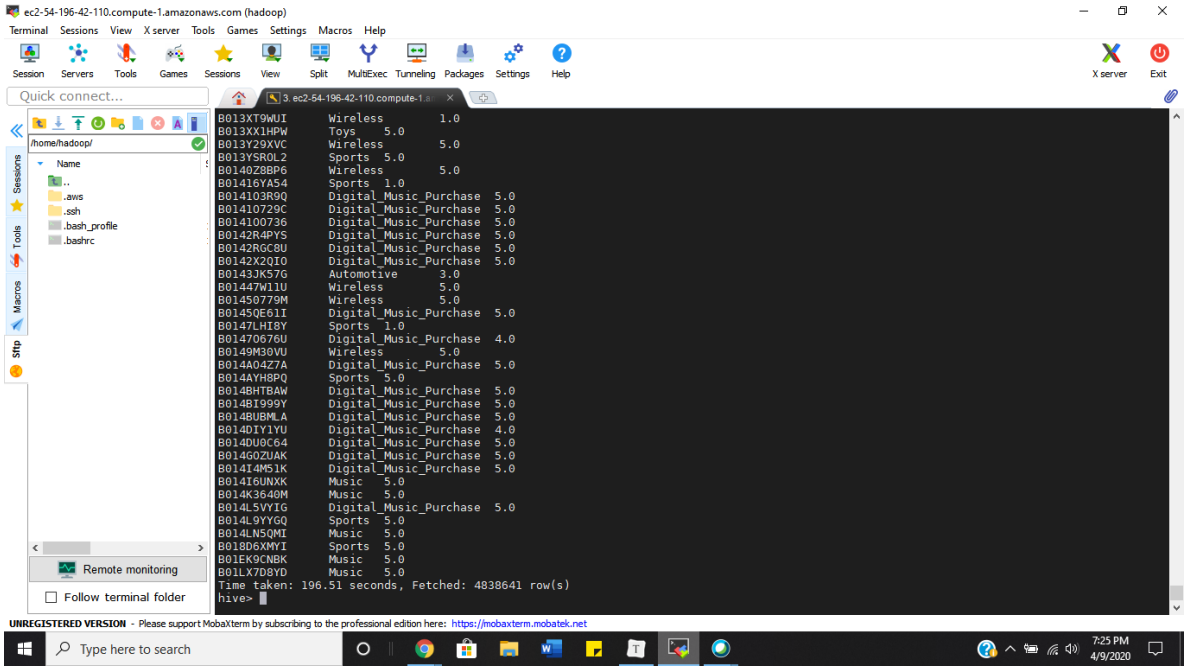
```
SELECT product_category,
        round(avg(star_rating),
        2) AS avg_rating
FROM amazon_review.amazon_review_filtered_data
WHERE year>=2005
GROUP BY  product_category
HAVING avg(star_rating)in (
    (SELECT max(x.avg_stars)
    FROM
        (SELECT product_category,
         avg(star_rating)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

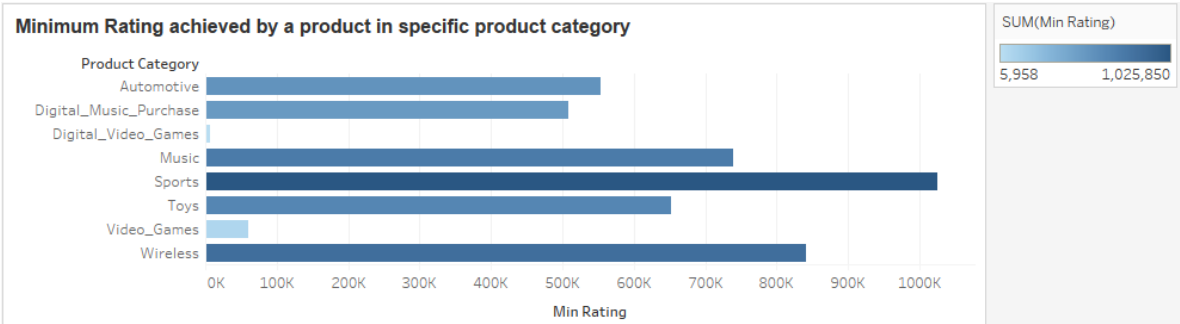


Minimum Rating of a product in the specific product category:

Calculated minimum rating of a product in a specific product category to find out the least popular product in that category.

```
SELECT x.product_id,
         x.product_category,
        round(min(x.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 x
GROUP BY  x.product_id, x.product_category;
```



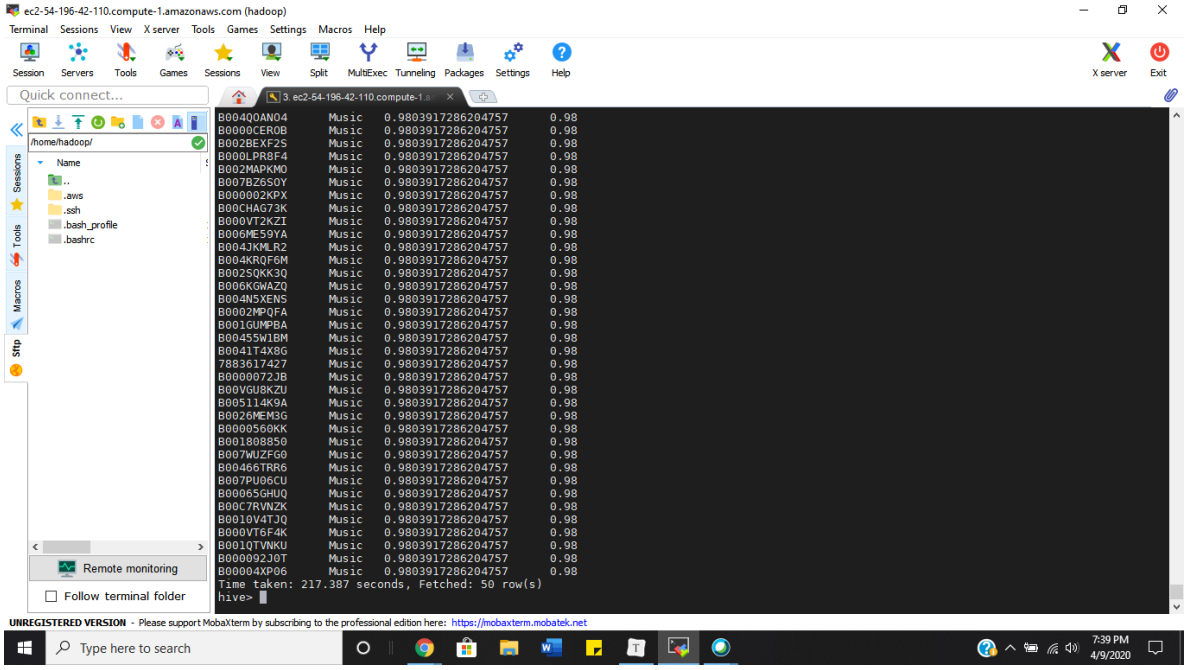


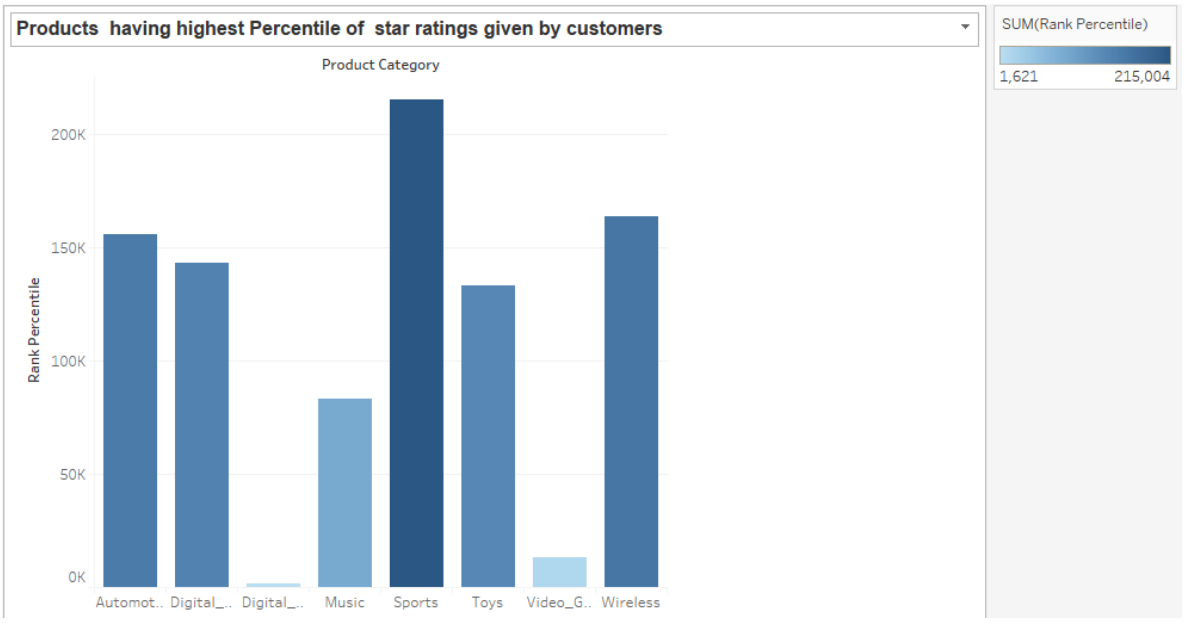
From the above diagram, we can interpret that the Digital_Video_Games product category is the least popular product category as compared to others.

Products having highest Percentile of star ratings given by customers:

Calculated percentile of product categories to find out product category having the highest percentile.

```
SELECT y.product_id,
        y.product_category,y.star_rank,
        round(y.star_rank,
        2) AS Rank_Percentile from
    (SELECT x.product_id,
        x.product_category,
         PERCENT_RANK()
        OVER (partition by x.product_category
    ORDER BY  x.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 x)as y order by y.star_rank
desc;
```





From the above diagram, we can interpret that the sports product category is the most popular product category since it has the highest percentile of rating as compared to others.

Conclusion:

Analyzed different parameters like Total Number_of_Reviews, Number_of_Users, Average_Review_Stars, Avg_Length_of_Review, Verified_Users, Unverified_Users, Total_Helpful_Votes, Total_Products by performing exploratory analysis on cleaned amazon_reviews database. Realized few product categories are having low average ratings after analyzing these parameters. Found correlation between Music and Digital Music Purchase categories from the detailed analysis of these categories. On the other hand, Video_Games and Digital_Video_Games categories are not correlated. In 2014 the number of customers reviewing both Music/Digital_Music_Purchase and Video_Games/Digital_Video_Games categories was more as compared to other years. Hence moved to focus on 2014 to find out the reasons behind this result. After interpreting the results of this analysis I realized there are the same products in both Music and Digital Music Purchase and Video_Games/Digital_Video_Games categories but they don't have the same rating. Also, the calculated rank of different products in a specific product category to find out popular products and least rated products in that category and discovered positive annual growth for those products by calculating and comparing moving average. Also, discovered that the sports category has the highest percentile of products with a maximum star rating and the Digital_Video_Games category has the lowest percentile of products with maximum star rating.

References:

Amazon reviews dataset:

https://registry.opendata.aws/amazon-reviews/

Documentation:

https://s3.amazonaws.com/amazon-reviews-pds/readme.html

Code References:

https://www.tutorialspoint.com/index.htm