BIG DATA ANALYTICS

Project Part 1 -Analysis of Amazon Review Dataset**

Report by Vallabhan Kudlu Ramakrishnan

The University of Texas at Dallas**

April 2020

When purchasing a product only variable we look into are rating and reviews. Improving it is important to maintain current customers and will draw new customers. This project aims to identify relationships between different aspects of a review including review length, overall review star rating, review helpfulness, reviewer frequency, product price, etc.

This project uses advanced tools to research the different statistical patterns related to these aspects of the analysis. The Relation were analyzed using Hive,AWS Athena and Power BI

The workflow of this project involves four stages:

• Data collection and preparation • Analysis using Hive • Conclusions derived from analysis • Visualizations

The Amazon Test Dataset is available online free of charge on the SNAP datasets website.
The list contains customer reviews of items that are offered on the Amazon website. There are two pieces of this dataset. The first component includes about 35 million reviews spanning 13 years of data from product feedback. It includes user-related information that offers analysis, time-related information, and analysis features such as review length, description, etc.

In order to understand the structure and schema of the dataset, let us look at a sample Amazon Review.

• Summary : The title of the review

• Review text : The actual content of the review.

• Rating : User rating of the product on a scale of 1 to 5.

• Helpfulness : The number of people who found the review useful.

Analysis

```
create database amazon_review;
drop table amazon_review.amazon_reviews_parquet;

CREATE EXTERNAL TABLE amazon_review.amazon_reviews_parquet(
  `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');

Msck repair table amazon_review.amazon_reviews_parquet;

set hive.cli.print.header=true;

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

**Final view**
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

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');

Msck repair table amazon_review.amazon_reviews_v2;

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_r
ating,
helpful_votes,total_votes,vine,verified_purchase,review_headline,review_body,rev
iew_date,year
from amazon_review.amazon_reviews_include where product_category='Wireless';

insert overwrite table amazon_review.amazon_reviews_v2
partition(product_category='Automotive')
select
marketplace,customer_id,review_id,product_id,product_parent,product_title,star_r
ating,
helpful_votes,total_votes,vine,verified_purchase,review_headline,review_body,rev
iew_date,year
from amazon_review.amazon_reviews_include where product_category='Automotive';

insert overwrite table amazon_review.amazon_reviews_v2
partition(product_category='Music')
select
marketplace,customer_id,review_id,product_id,product_parent,product_title,star_r
ating,
helpful_votes,total_votes,vine,verified_purchase,review_headline,review_body,rev
iew_date,year
from amazon_review.amazon_reviews_include where product_category='Music';

insert overwrite table amazon_review.amazon_reviews_v2
partition(product_category='Digital_Music_Purchase')
select
marketplace,customer_id,review_id,product_id,product_parent,product_title,star_r
ating,
helpful_votes,total_votes,vine,verified_purchase,review_headline,review_body,rev
iew_date,year
from amazon_review.amazon_reviews_include where
product_category='Digital_Music_Purchase';

insert overwrite table amazon_review.amazon_reviews_v2
partition(product_category='Sports')
select
marketplace,customer_id,review_id,product_id,product_parent,product_title,star_r
ating,
helpful_votes,total_votes,vine,verified_purchase,review_headline,review_body,rev
iew_date,year
from amazon_review.amazon_reviews_include where product_category='Sports';

insert overwrite table amazon_review.amazon_reviews_v2
partition(product_category='Toys')
select
marketplace,customer_id,review_id,product_id,product_parent,product_title,star_r
ating,
helpful_votes,total_votes,vine,verified_purchase,review_headline,review_body,rev
iew_date,year
from amazon_review.amazon_reviews_include where product_category='Toys';

insert overwrite table amazon_review.amazon_reviews_v2
partition(product_category='Digital_Video_Games')
select
marketplace,customer_id,review_id,product_id,product_parent,product_title,star_r
ating,
```

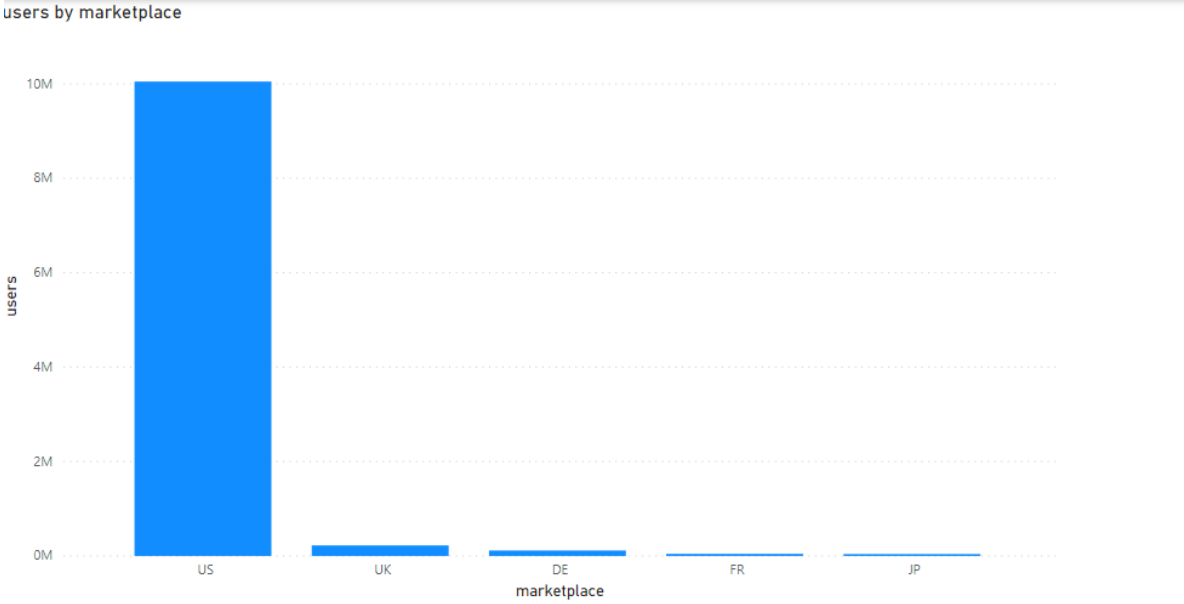
```
helpful_votes,total_votes,vine,verified_purchase,review_headline,review_body,rev
iew_date,year
from amazon_review.amazon_reviews_include where
product_category='Digital_Video_Games';

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

1)How many unique customer from different marketplace?

US seem to have majority of business with new customers

```
select marketplace,count(distinct(customer_id))as users from amazon_reviews_v2
group by marketplace limit 10;
```

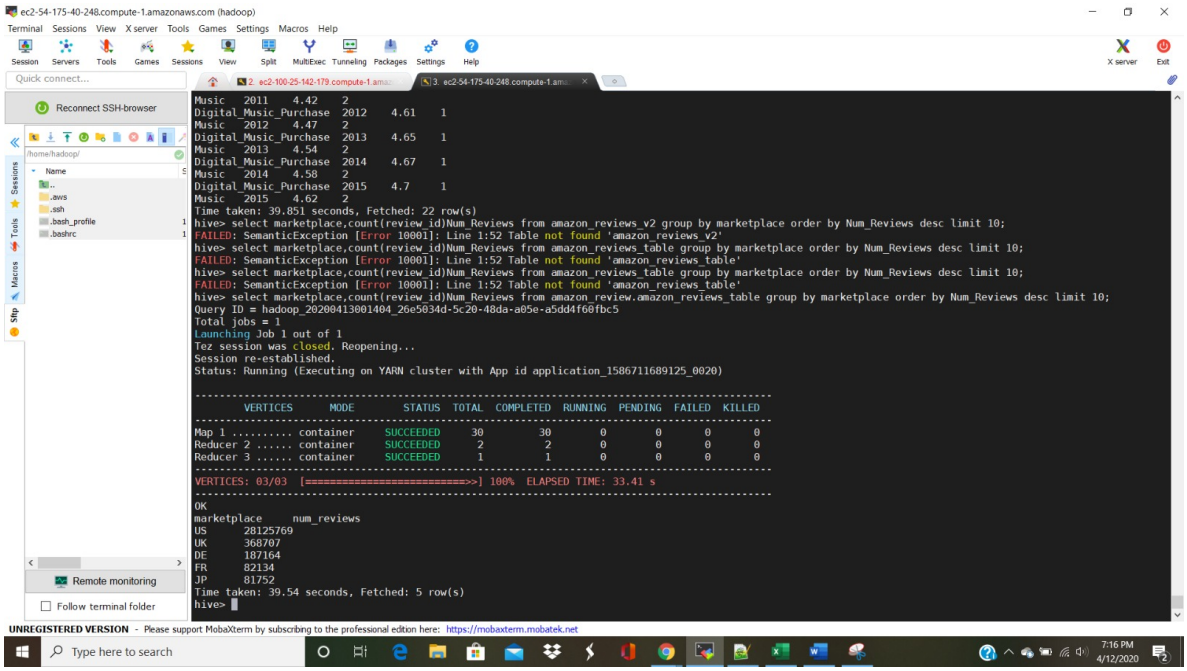
```
hive> select marketplace,count(distinct(customer_id)) as users from amazon_reviews_v2 group by marketplace limit 10;
Query ID = hadoop_20200412210656_ae05be0c-078a-4e5c-a589-1e92f6865c29
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0010)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED     26         26        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 77.92 s
--------------------------------------------------------------------------------
OK
FR      45790
JP      41114
UK      227365
US      11407866
DE      117430
Time taken: 79.978 seconds, Fetched: 5 row(s)
```

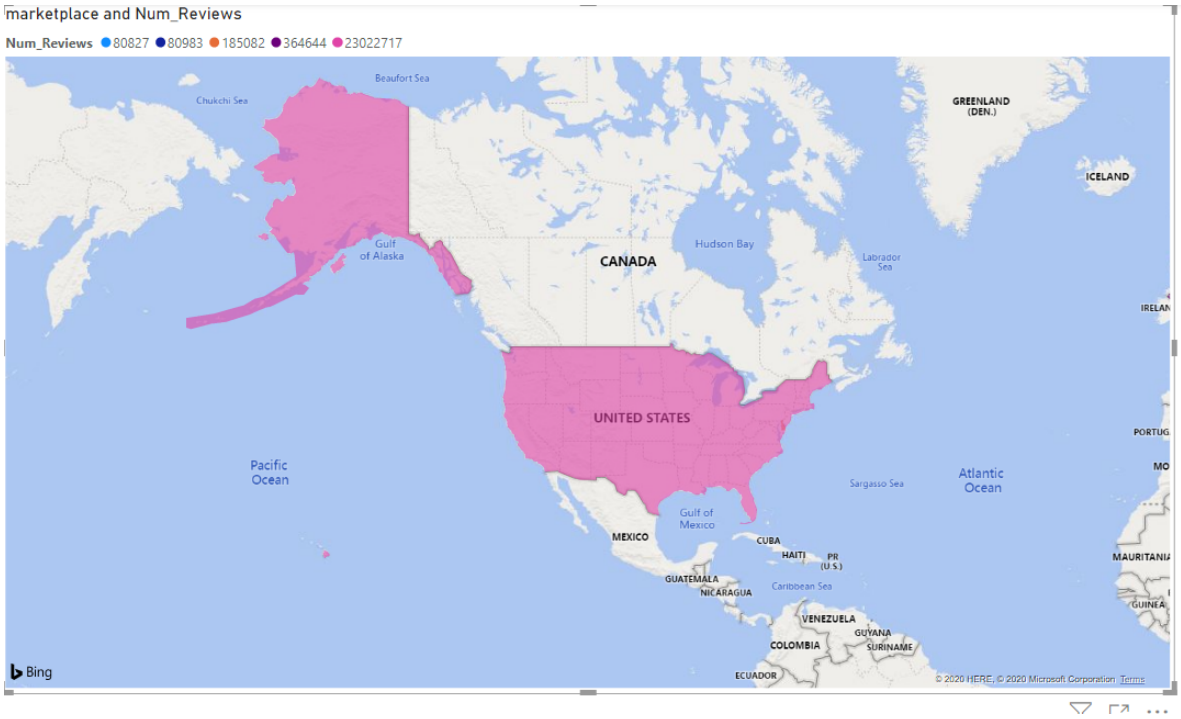


2) Which marketplace is reviewing the product most?

US seem to review lot of products

```
select marketplace,count(review_id)Num_Reviews from amazon_reviews_v2 group by marketplace order by Num_Reviews desc limit 10;
```



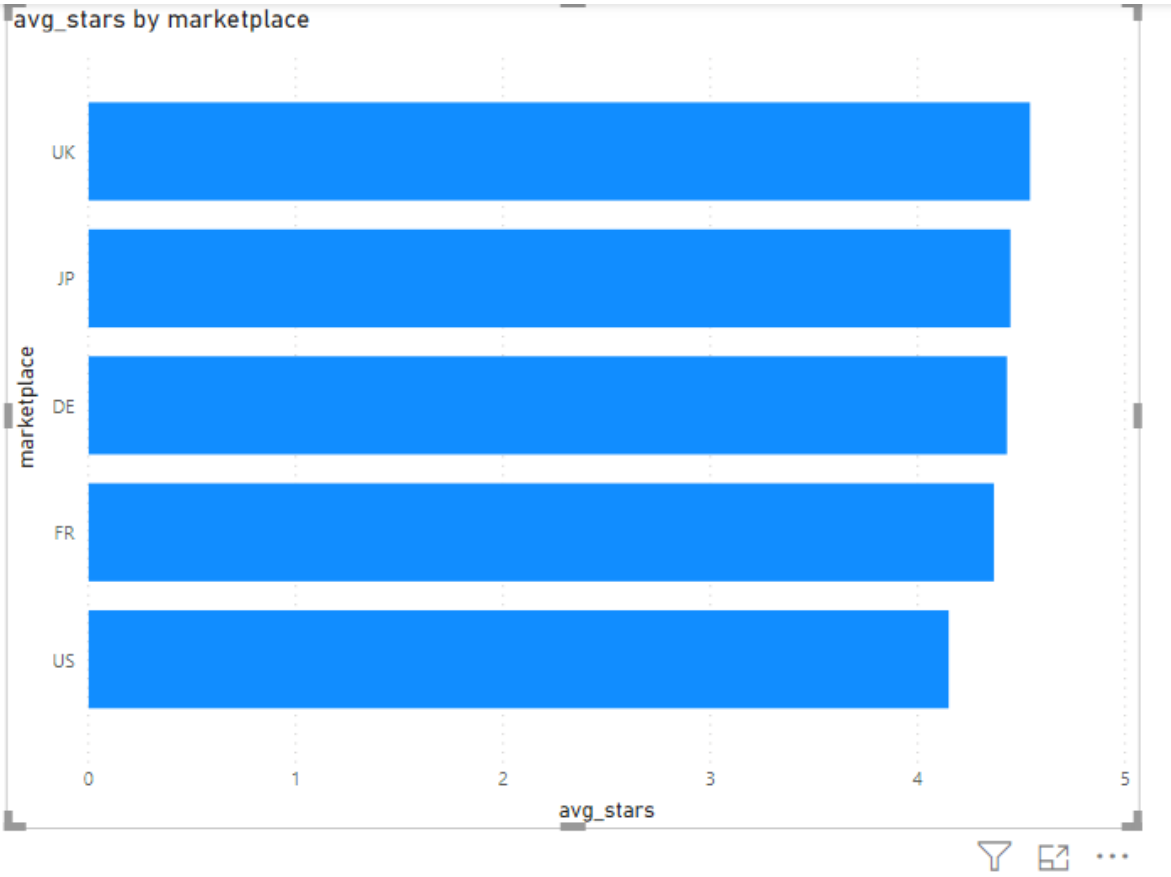


3)Trend of star rating

UK is appears to give 4.5 average star rating

```
select marketplace,avg(star_rating) as avg_stars from amazon_reviews_v2 group by marketplace limit 10;
```

```
hive> select marketplace,avg(star_rating) as review_stars from amazon_reviews_v2 group by marketplace limit 10;
Query ID = hadoop_20200412212615_21ecf006-1a56-476f-ba54-3c7b5e0bfa72
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     24         24        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 32.06 s
--------------------------------------------------------------------------------
OK
DE      4.4272883674210854
UK      4.535368192087484
FR      4.363211337570312
JP      4.4378608474410415
US      4.157076309629081
Time taken: 36.072 seconds, Fetched: 5 row(s)
```

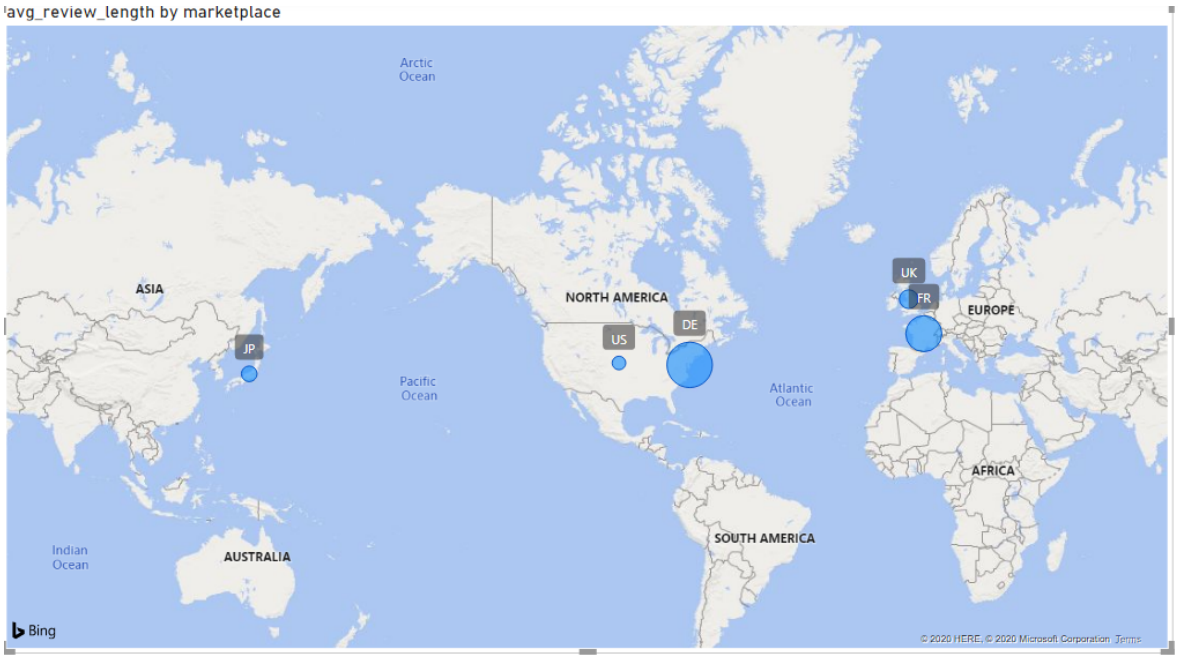


4) what is the average length of reviews from different marketplace?

US reviewers likely to write a long review body when compared to other marketplace

```
select marketplace,avg(length(review_body)) as avg_review_length from
amazon_reviews_v2 group by marketplace order by avg_review_length desc limit 10;
```

```
hive> select marketplace,avg(length(review_body))avg_review_length from amazon_reviews_v2 group by marketplace order by avg_review_length desc limit 10;
Query ID = hadoop_20200412213525_049794fc-db09-496e-abaa-0e8f5812f422
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1586723064686_0013)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     24         24        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 59.95 s
--------------------------------------------------------------------------------
OK
DE      768.0271312859311
FR      625.5894148586457
UK      374.37413108040306
JP      324.7476025051375
US      296.7596532062387
Time taken: 67.025 seconds, Fetched: 5 row(s)
```

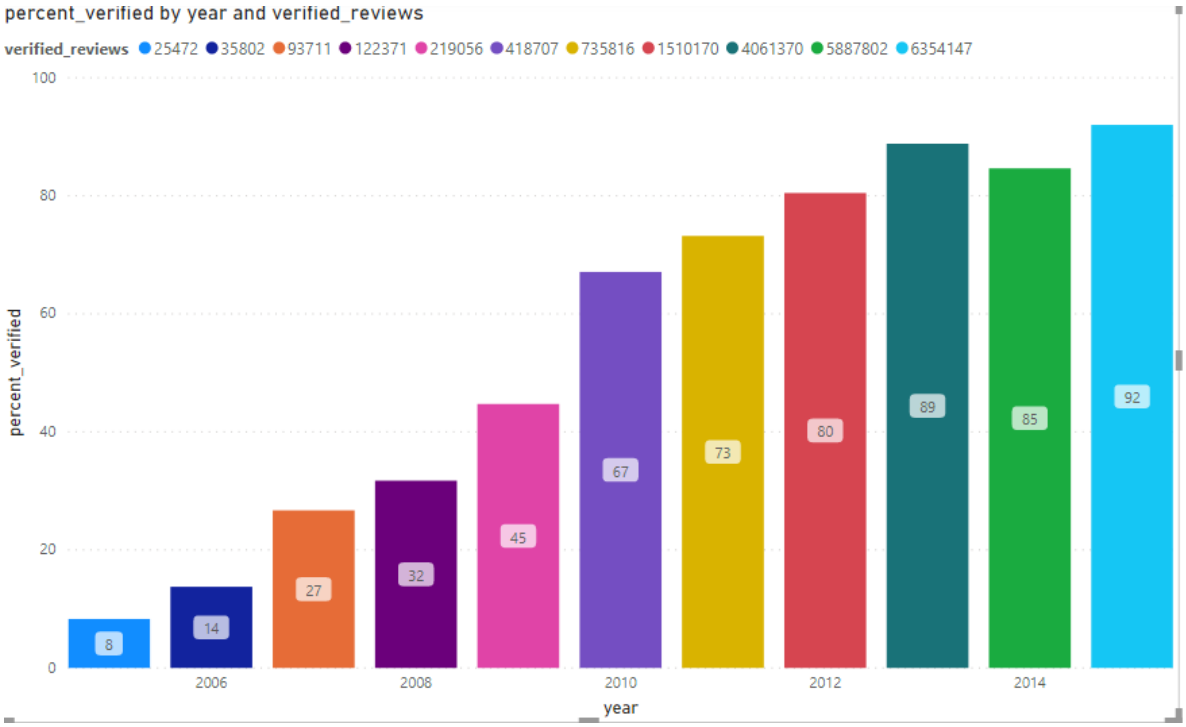


5)Has verification purchases changed over time for reviews?

Verification purchases seems to be increased from time to time

```
select year,sum(case when verified_purchase = 'Y' then 1 else 0 end)
verified_reviews,
       (sum(case when verified_purchase = 'Y' then 1 else 0
end)*100.00)/count(*) percent_verified
       from amazon_reviews_v2 group by year order by year;
```

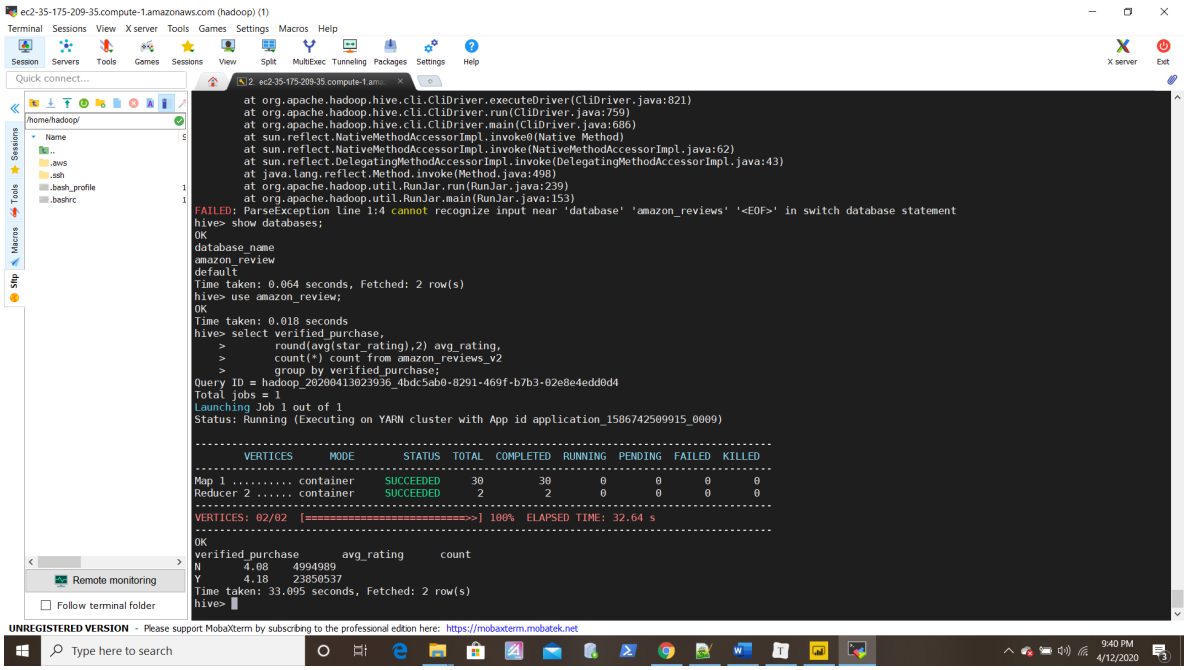
```
hive> select year,sum(case when verified_purchase = 'Y' then 1 else 0 end) verified_reviews,
    >        (sum(case when verified_purchase = 'Y' then 1 else 0 end)*100.00)/count(*) percent_verified
    >        from amazon_reviews_v2 group by year order by year;
Query ID = hadoop_20200412213826_6e147585-53f5-4bde-84c3-a81cdff22d02
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     24         24        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 33.44 s
----------------------------------------------------------------------------------------------
OK
2005    26792   8.006024240395879
2006    38434   13.381613077311422
2007    105235  26.088368507036807
2008    140802  30.636570145434607
2009    259658  44.207114121154236
2010    515739  67.696764530995475
2011    904989  73.771205403541550
2012    1815921 80.898664668465879
2013    4886874 89.174303489128100
2014    7242922 85.052409429305456
2015    7913171 92.615791230516928
Time taken: 34.028 seconds, Fetched: 11 row(s)
```

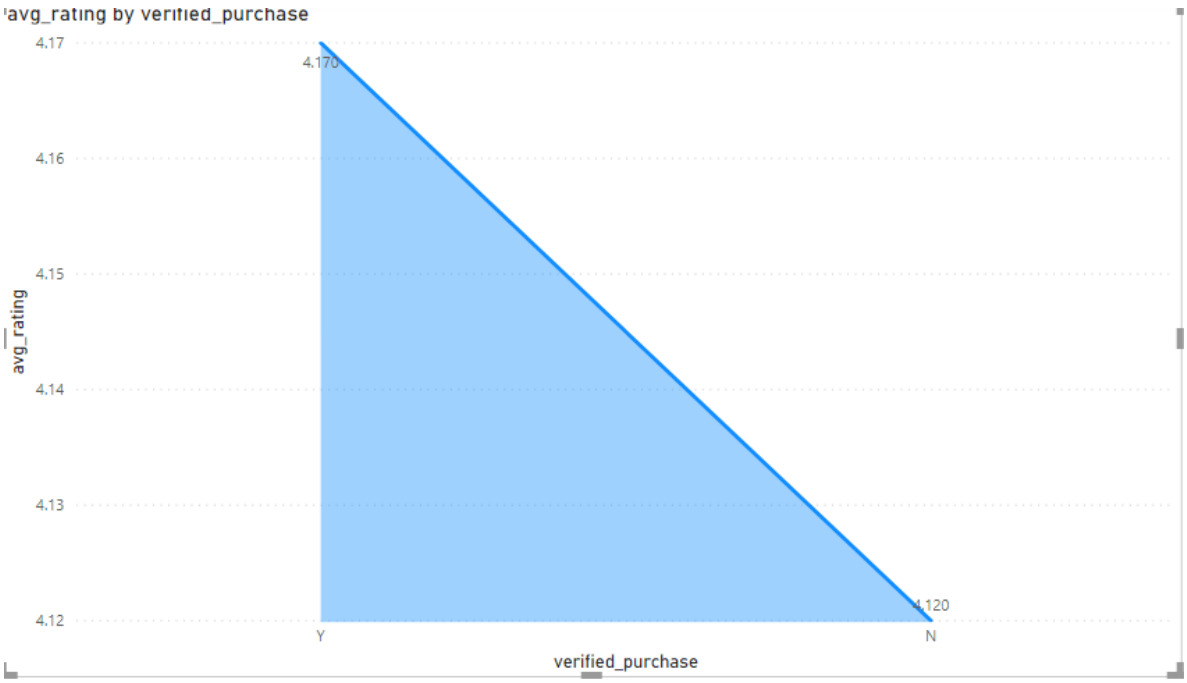


6) How do Ratings Vary with verified Purchase?

> **As seen from the graph verified_purchase are on average rated 0.05 higher**

```
select verified_purchase,
       round(avg(star_rating),2) avg_rating,
       count(*) count from amazon_reviews_v2
       group by verified_purchase;
```

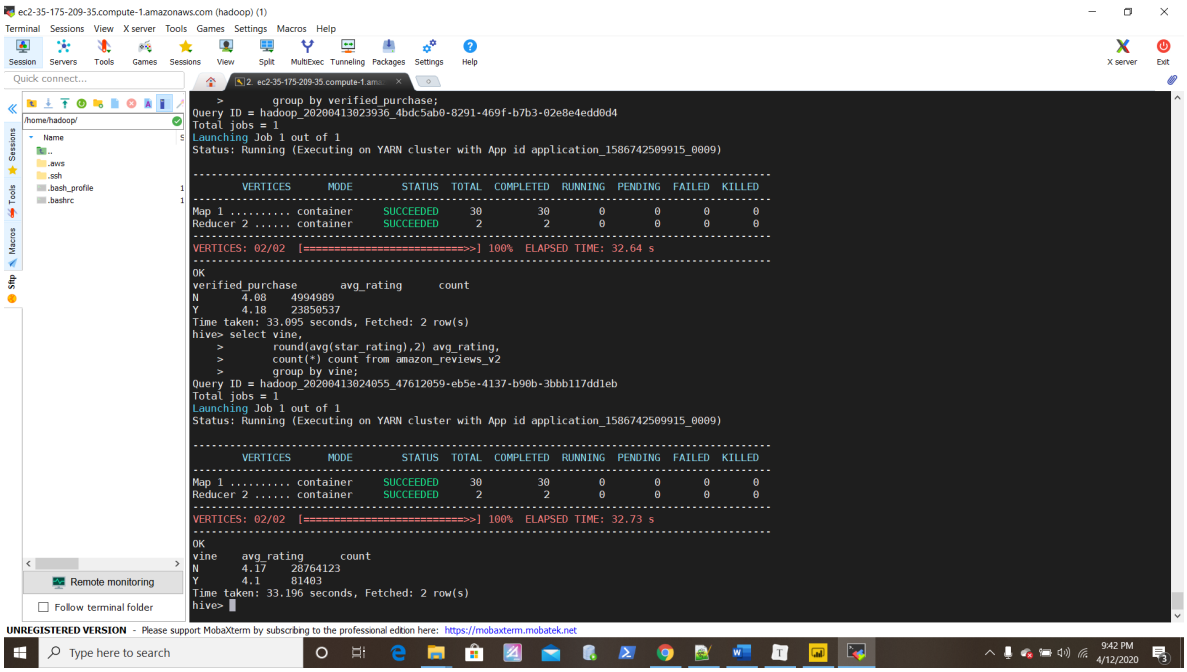


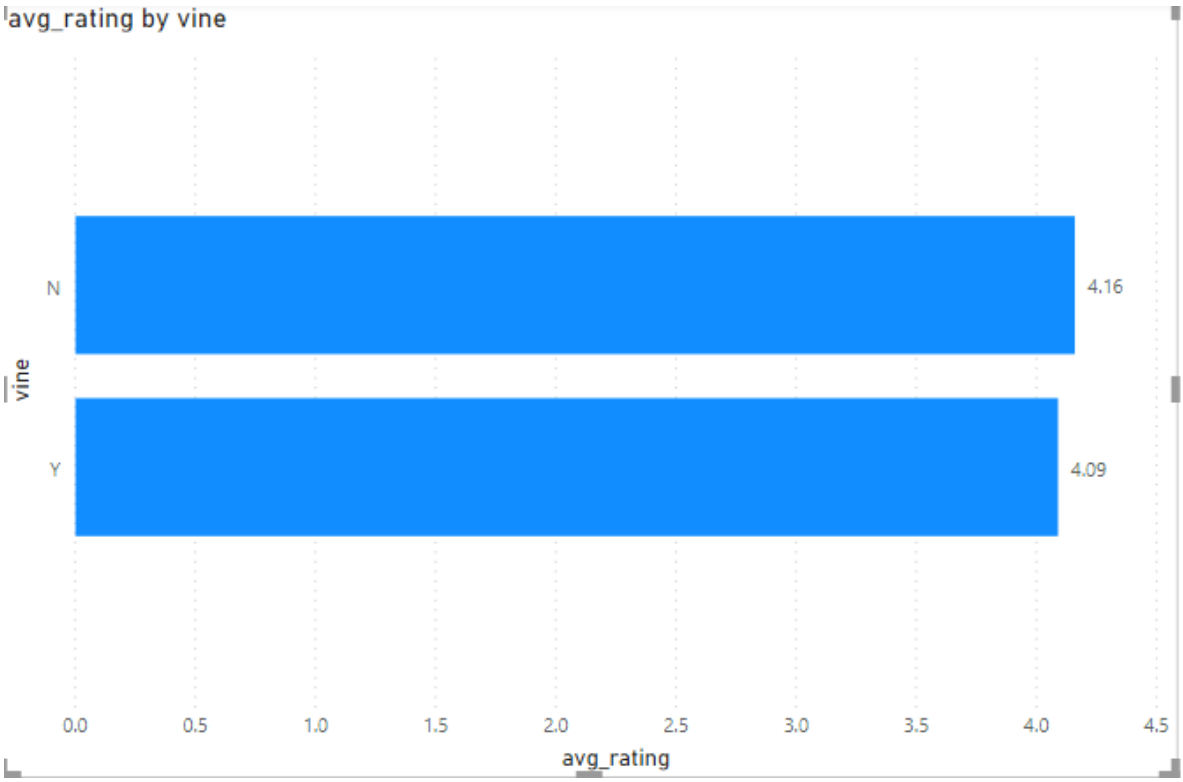


7) How do Ratings vary with vine membership?

Vine Reviews are on average 0.16 lower than non-Vine

```
select vine,
       round(avg(star_rating),2) avg_rating,
       count(*) count from amazon_reviews_v2
       group by vine;
```

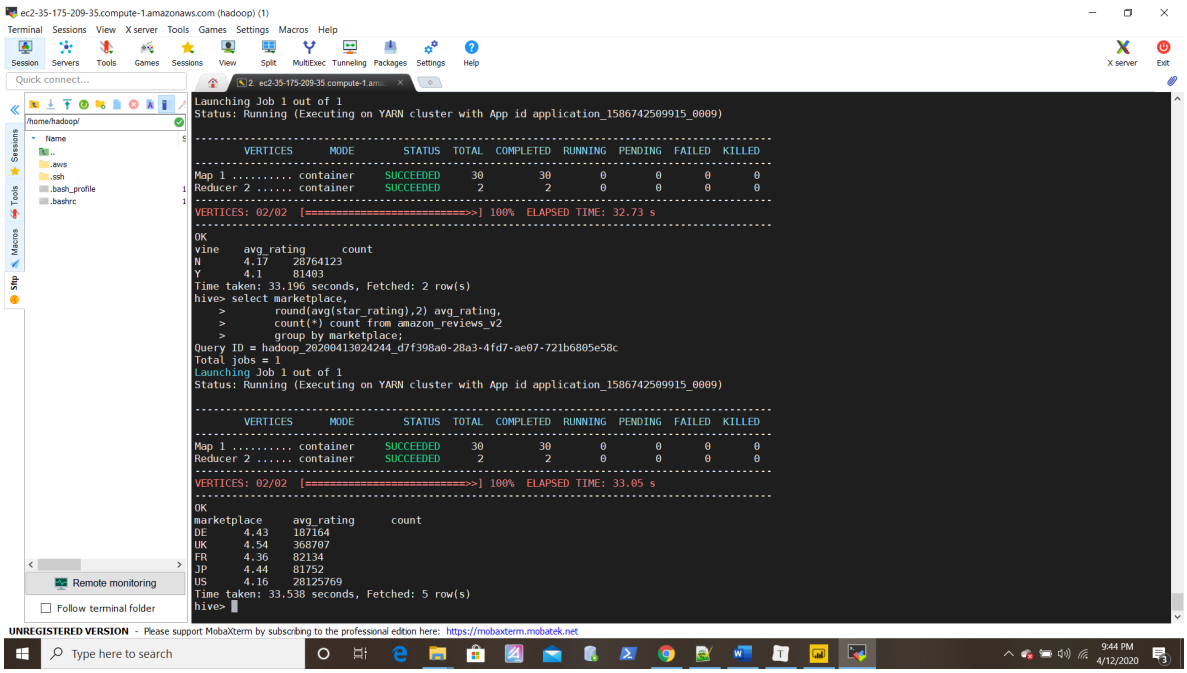


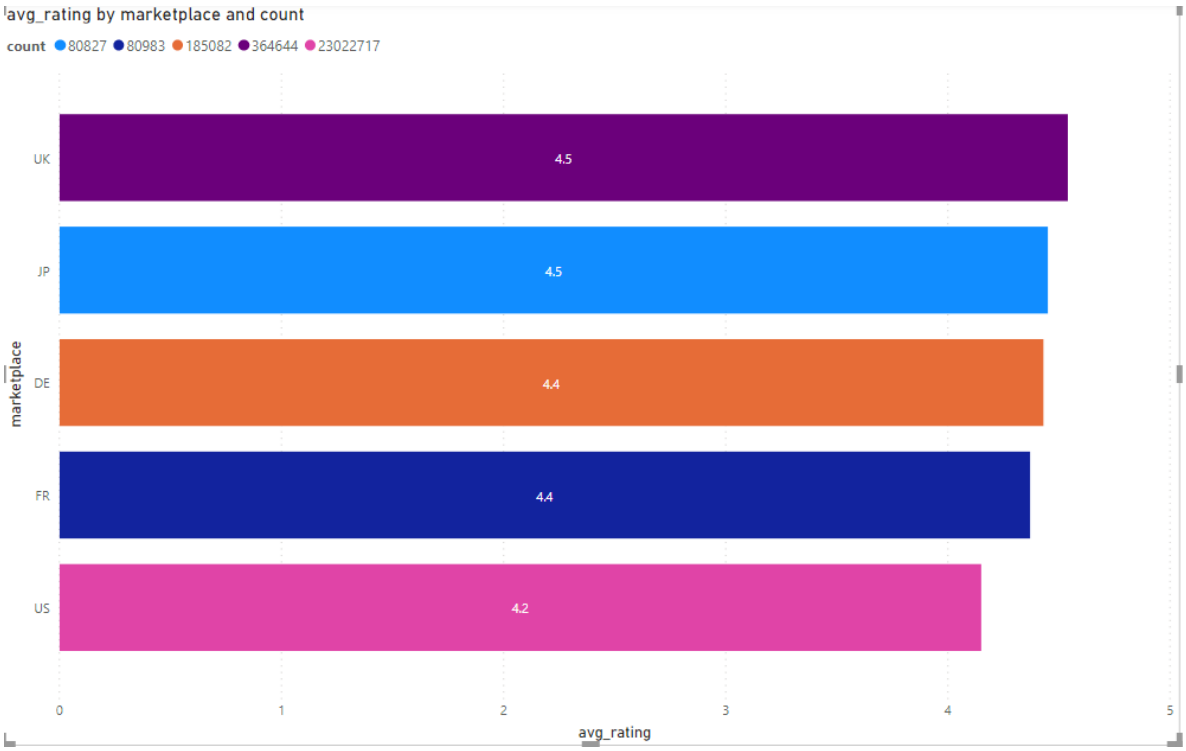


8)How do Ratings vary with Marketplace?

UK reviewers are the most positive ,with average rating of 0.3 higher than the US marketplace

```
select marketplace,
       round(avg(star_rating),2) avg_rating,
       count(*) count from amazon_reviews_v2
       group by marketplace;
```

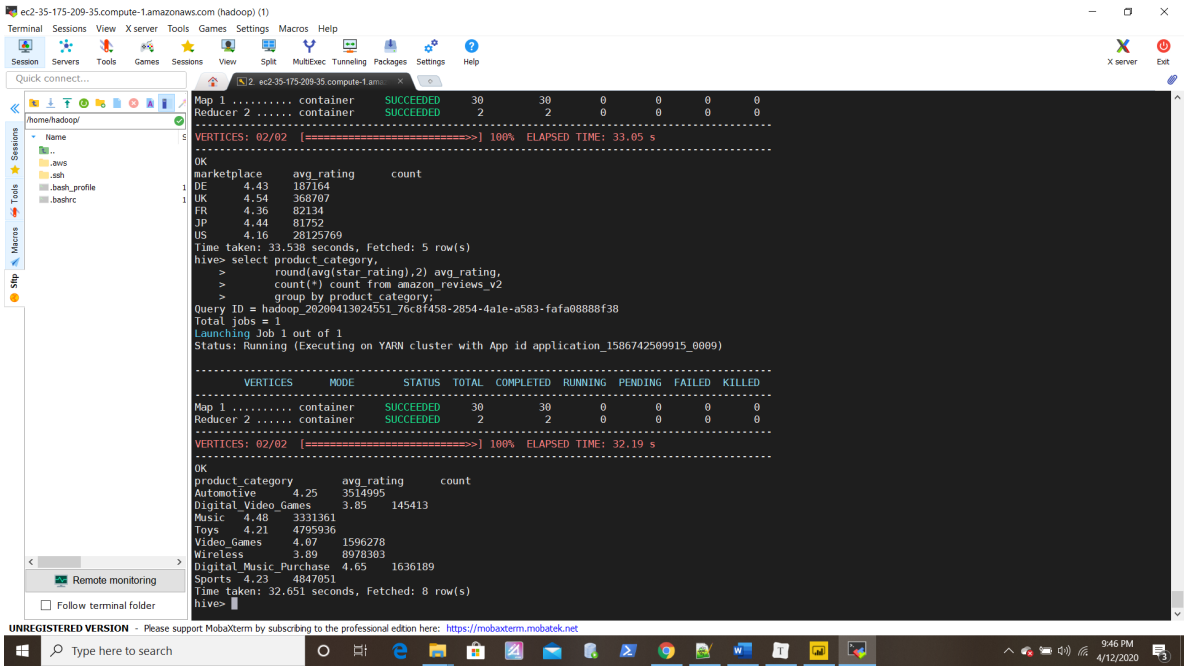


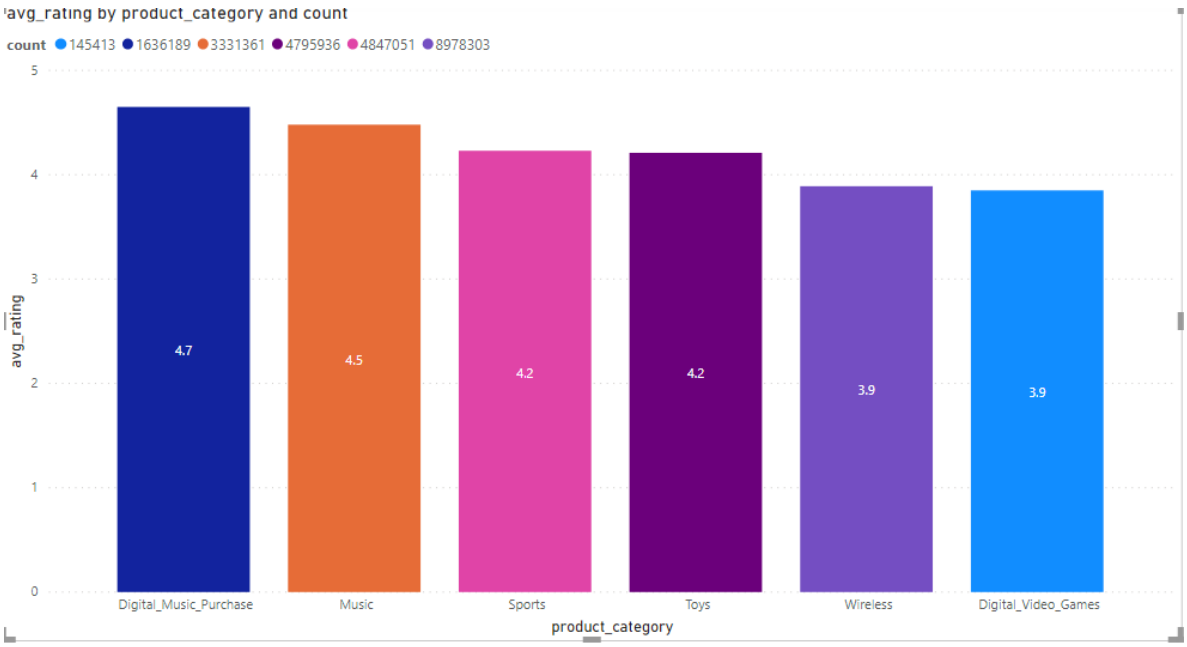


9)How do Ratings vary with product_category?

Digit Music purchase has highest average rating when compared to other product category

```
select product_category,
       round(avg(star_rating),2) avg_rating,
       count(*) count from amazon_reviews_v2
       group by product_category;
```

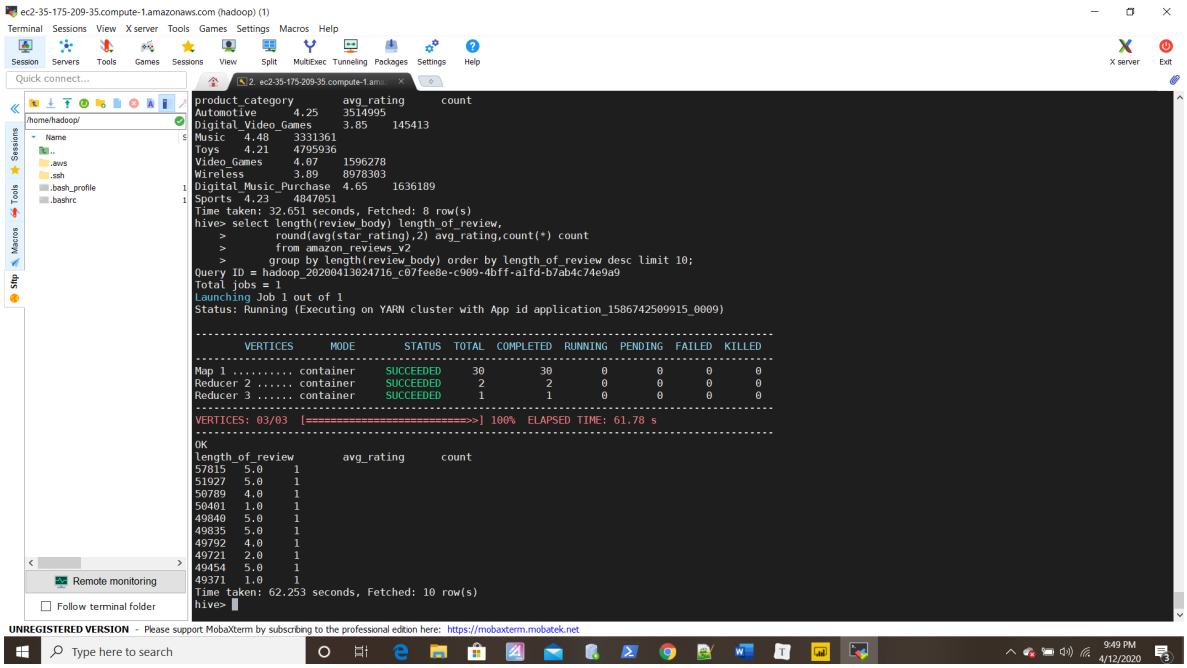


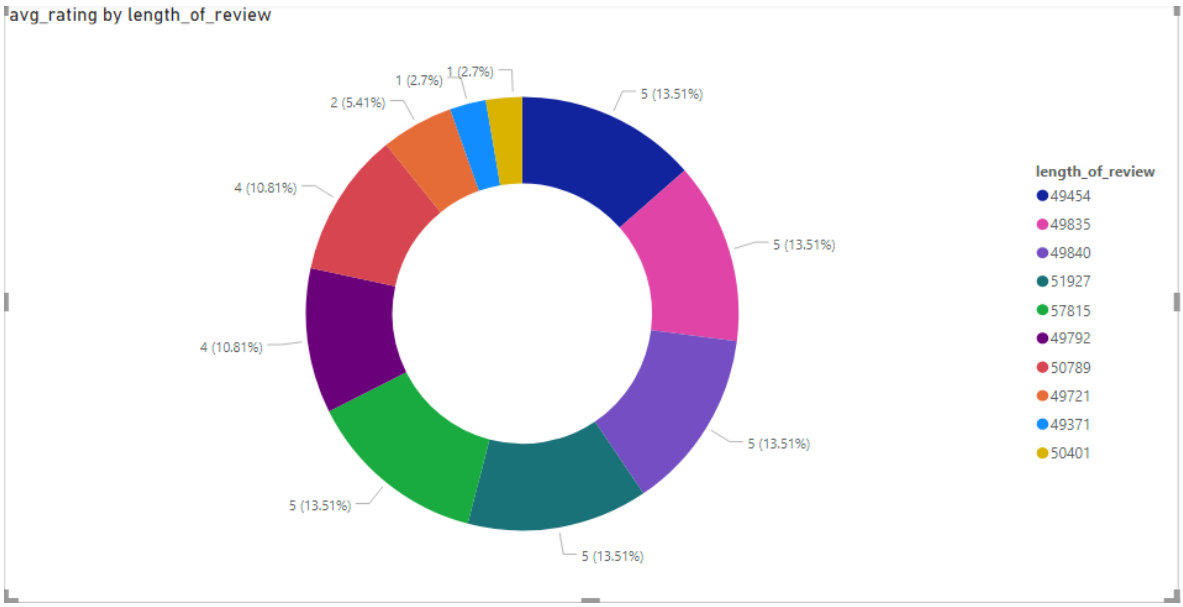


10)Does length of review influence star rating?

As analyzed length of review has little influence on star rating

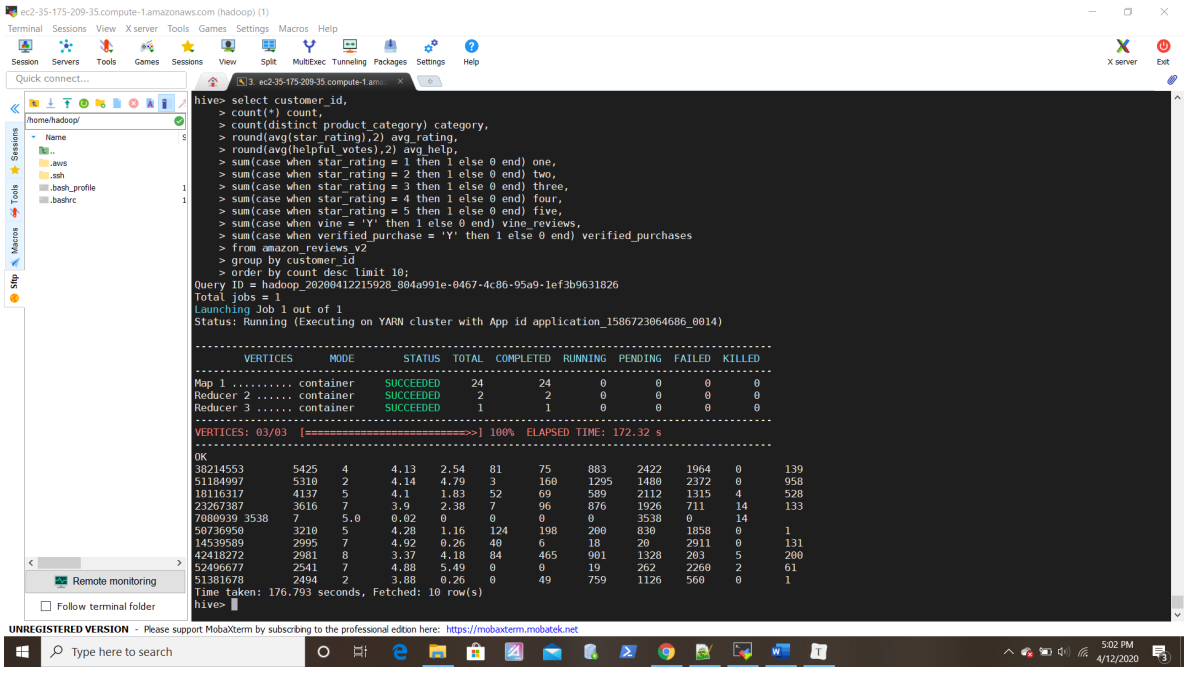
```
select length(review_body) length_of_review,
       round(avg(star_rating),2) avg_rating,count(*) count
       from amazon_reviews_v2
      group by length(review_body) order by length_of_review desc limit 10;
```





11) **Looking at the Table we can easily check how helpful the average review is ,how many Vine Reviews they have and how many of their purchases are verified**

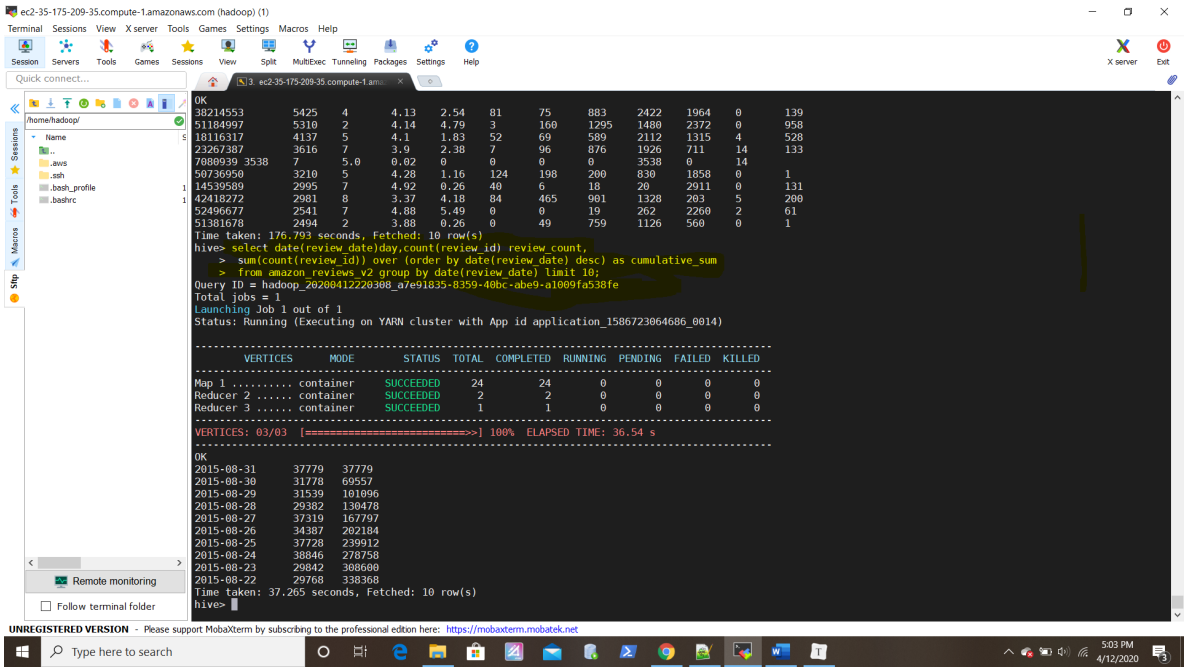
```
select customer_id,
count(*) count,
count(distinct product_category) category,
round(avg(star_rating),2) avg_rating,
round(avg(helpful_votes),2) avg_help,
sum(case when star_rating = 1 then 1 else 0 end) one,
sum(case when star_rating = 2 then 1 else 0 end) two,
sum(case when star_rating = 3 then 1 else 0 end) three,
sum(case when star_rating = 4 then 1 else 0 end) four,
sum(case when star_rating = 5 then 1 else 0 end) five,
sum(case when vine = 'Y' then 1 else 0 end) vine_reviews,
sum(case when verified_purchase = 'Y' then 1 else 0 end) verified_purchases
from amazon_reviews_v2
group by customer_id
order by count desc;
```

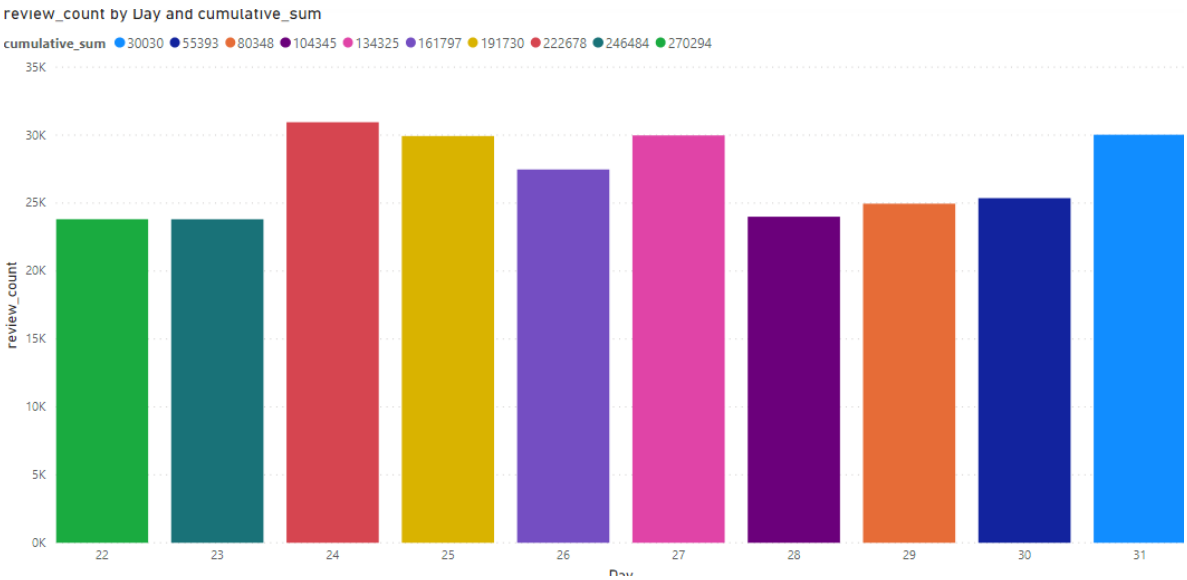


12)How many reviews per day?

We can see that there is variation in number of reviews that are given within 30 day period as seen from the visualization the cumulative sum hovers each day

```
select date(review_date)day,count(review_id) review_count,
 sum(count(review_id)) over (order by date(review_date) desc) as cumulative_sum
 from amazon_reviews_v2 group by date(review_date) limit 10;
```

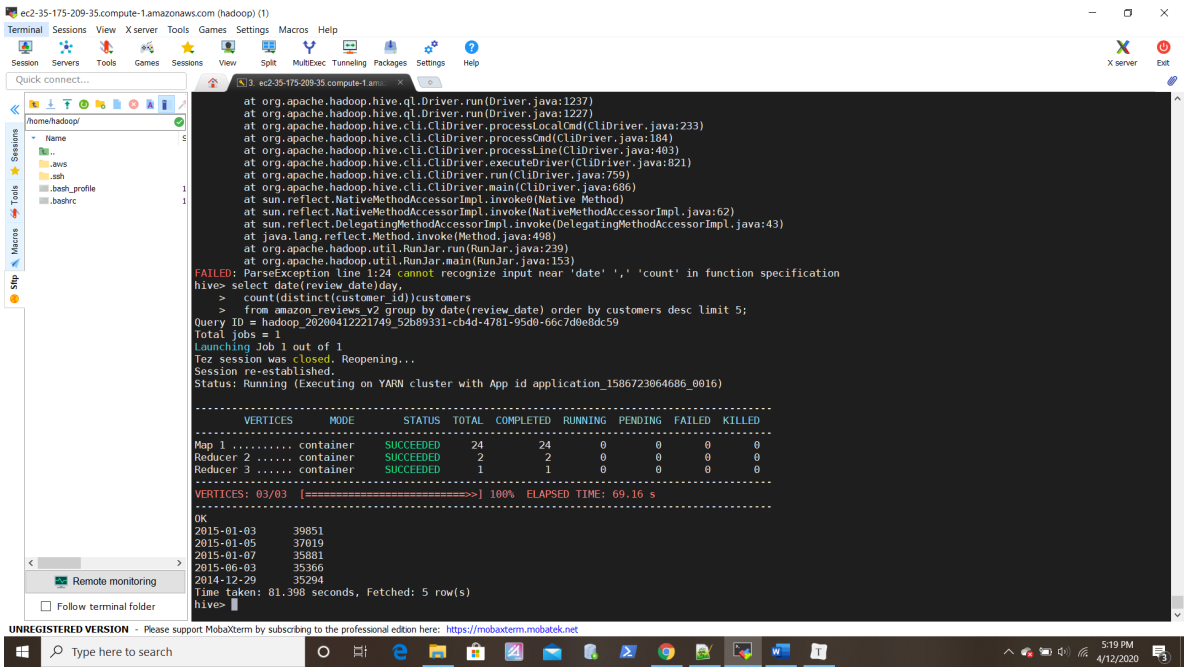


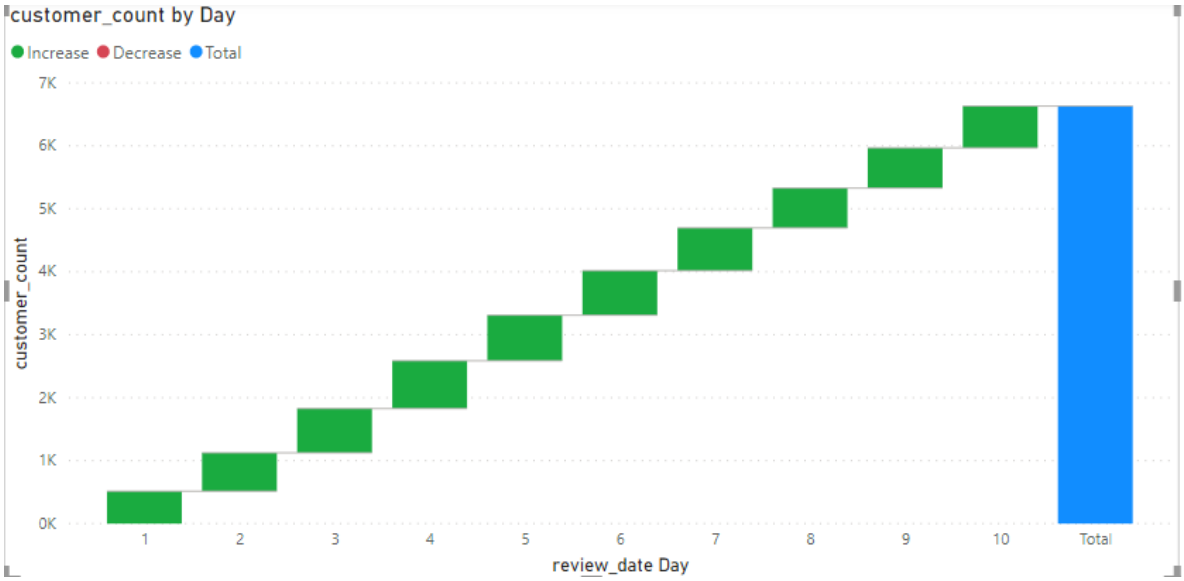


13)How many unique customer per day?

There seems to be increase in the new customers per day

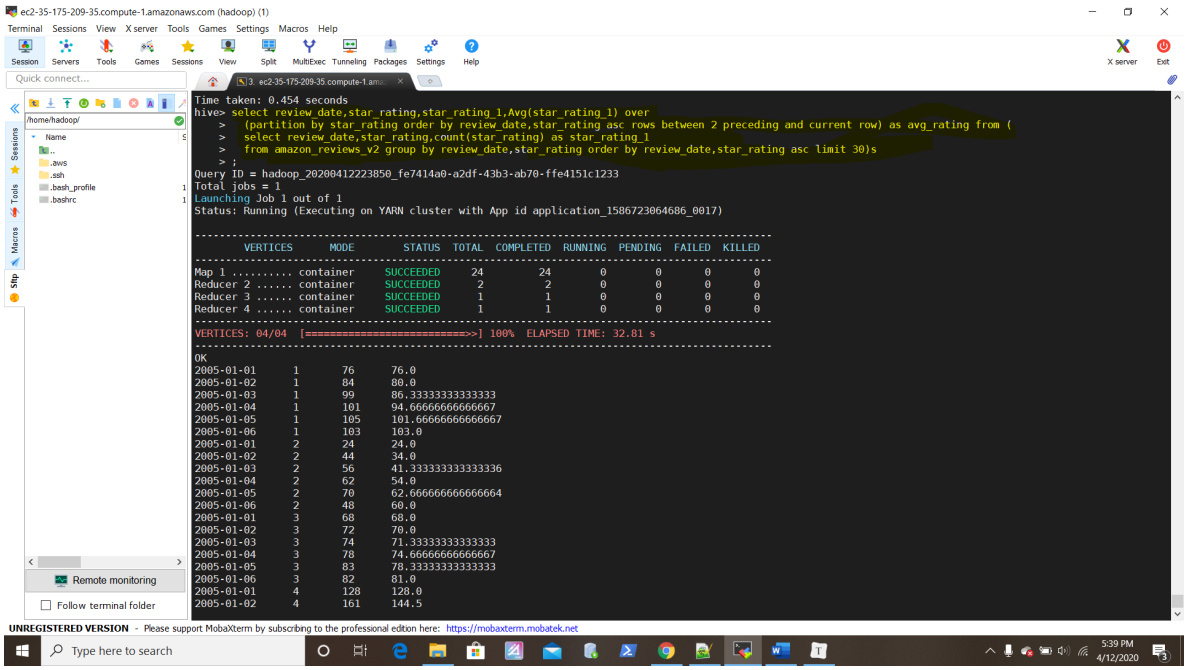
```
select review_date,count(distinct(customer_id)) customer_count,
 sum(count(distinct(customer_id))) over (order by review_date desc) as
cumulative_sum_customer
 from amazon_reviews_v2 group by review_date order by review_date limit 10;
```





14)What is the moving average star rating for 2 days?

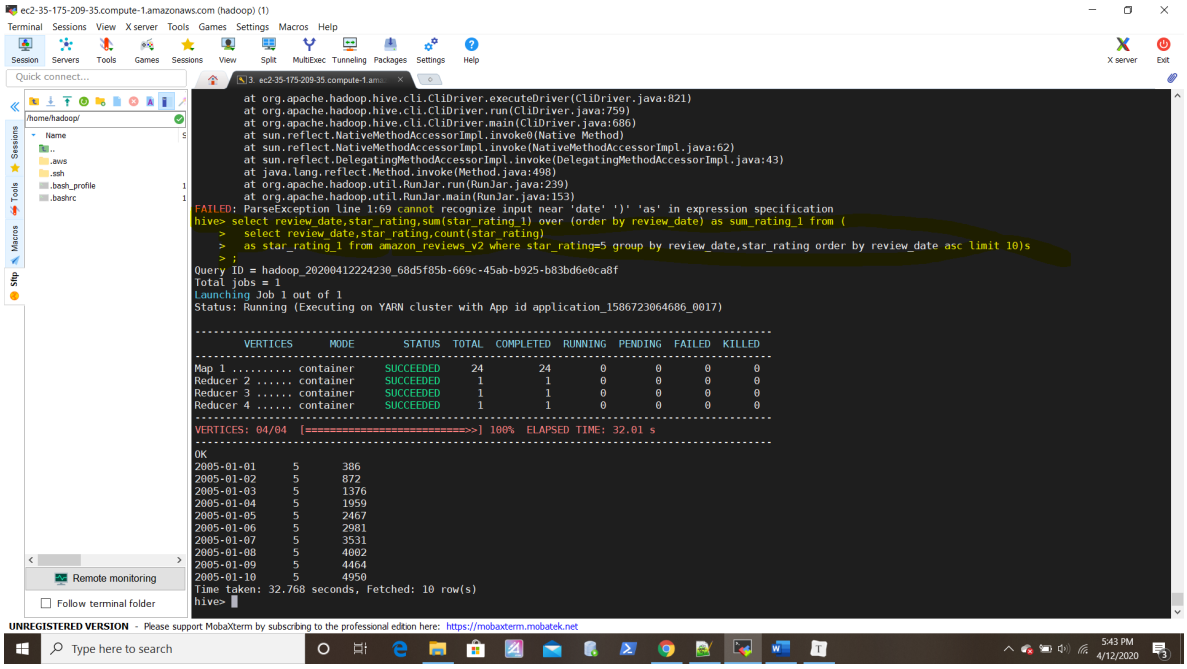
```
select review_date,star_rating,star_rating_1,Avg(star_rating_1) over
  (partition by star_rating order by review_date,star_rating asc rows between 2
preceding and current row) as avg_rating from (
  select review_date,star_rating,count(star_rating) as star_rating_1
  from amazon_reviews_v2 group by review_date,star_rating order by
review_date,star_rating asc limit 30)s;
```



15)Number of five star rating per day for time period of 10 days?

As analyzed the trend of five star rating seems to be increasing for the given time interval

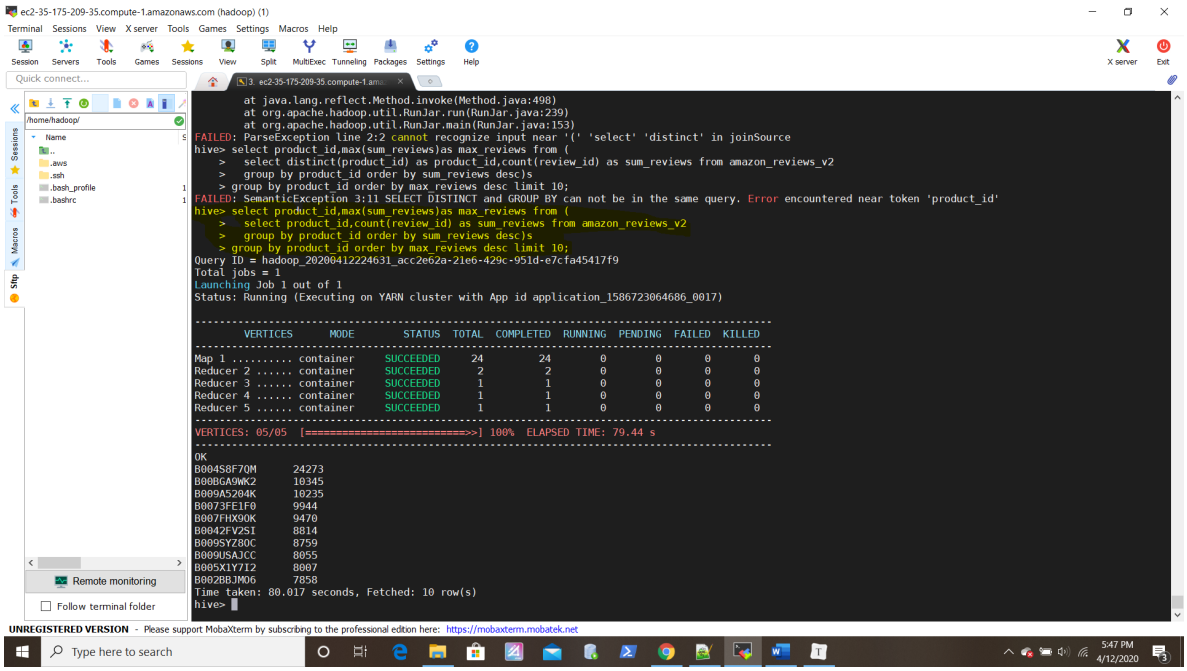
```
select review_date,star_rating,sum(star_rating_1) over (order by review_date) as
sum_rating_1 from (
  select review_date,star_rating,count(star_rating)
  as star_rating_1 from amazon_reviews_v2 where star_rating=5 group by
review_date,star_rating order by review_date asc limit 10)s
```

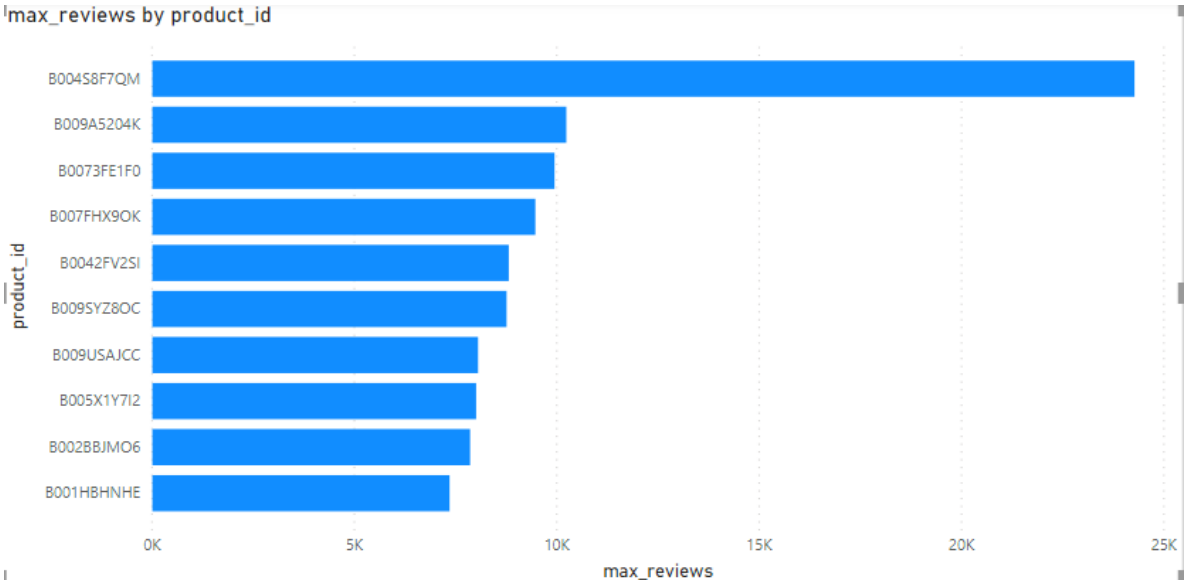


16) which product id has max reviews?

As we have top 10 product id with maximum reviews we can say that product id B00458F70M has the highest reviews

```
select product_id,max(sum_reviews)as max_reviews from (
  select product_id,count(review_id) as sum_reviews from amazon_reviews_v2
  group by product_id order by sum_reviews desc)s
group by product_id order by max_reviews desc limit 10;
```





17)what is standard deviation for the star rating?

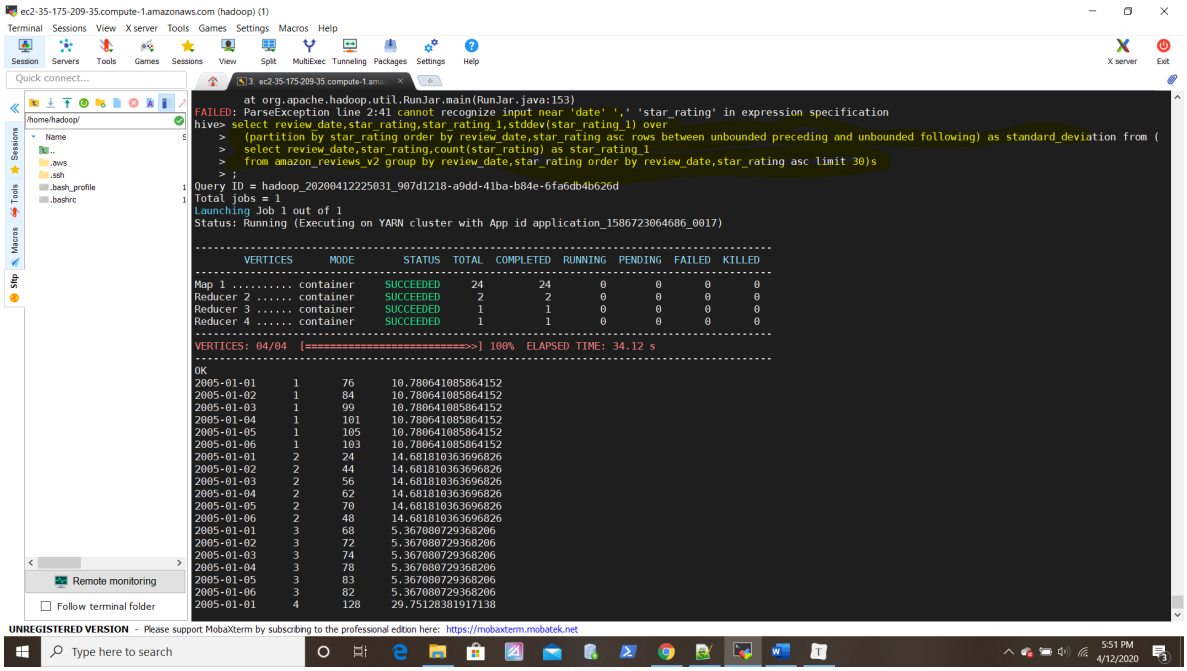
The standard deviation for one star rating - 10.70

The standard deviation for two star rating - 14.60

The standard deviation for three star rating - 5.36

The standard deviation for four star rating -29.75

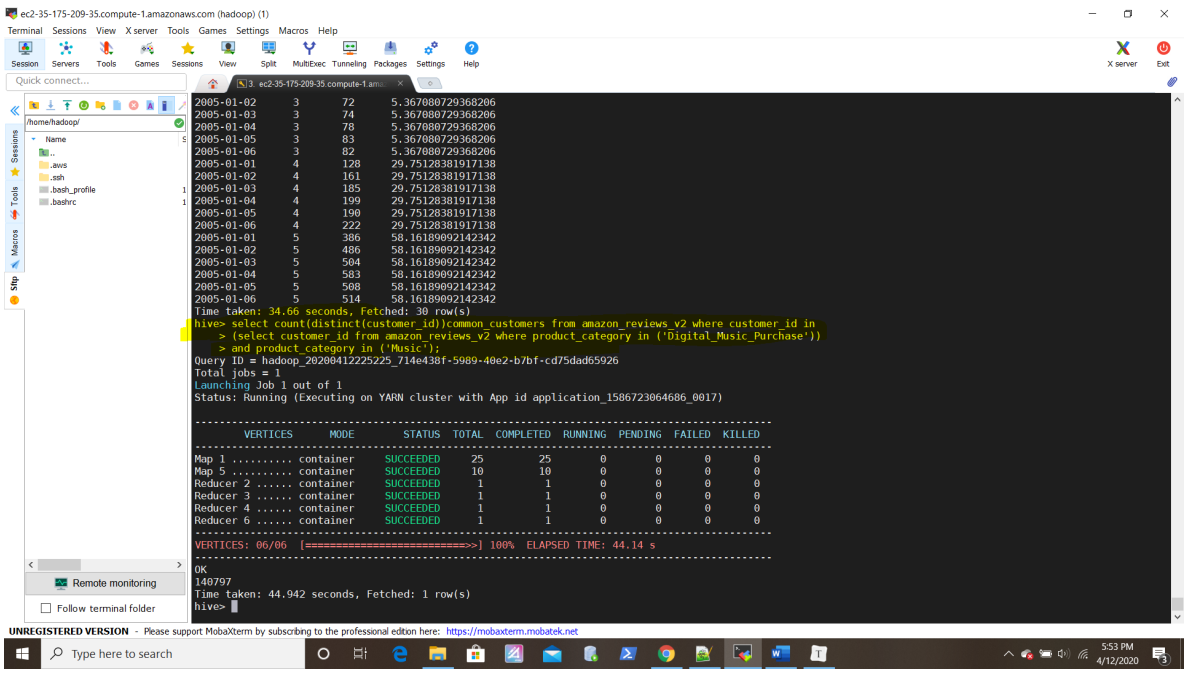
```
select review_date,star_rating,star_rating_1,stddev(star_rating_1) over
  (partition by star_rating order by review_date,star_rating asc rows between
unbounded preceding and unbounded following) as standard_deviation from (
  select review_date,star_rating,count(star_rating) as star_rating_1
  from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



18) Are they any common users between music and digital music purchase?

From the analysis there are 140797 users are common between two categories

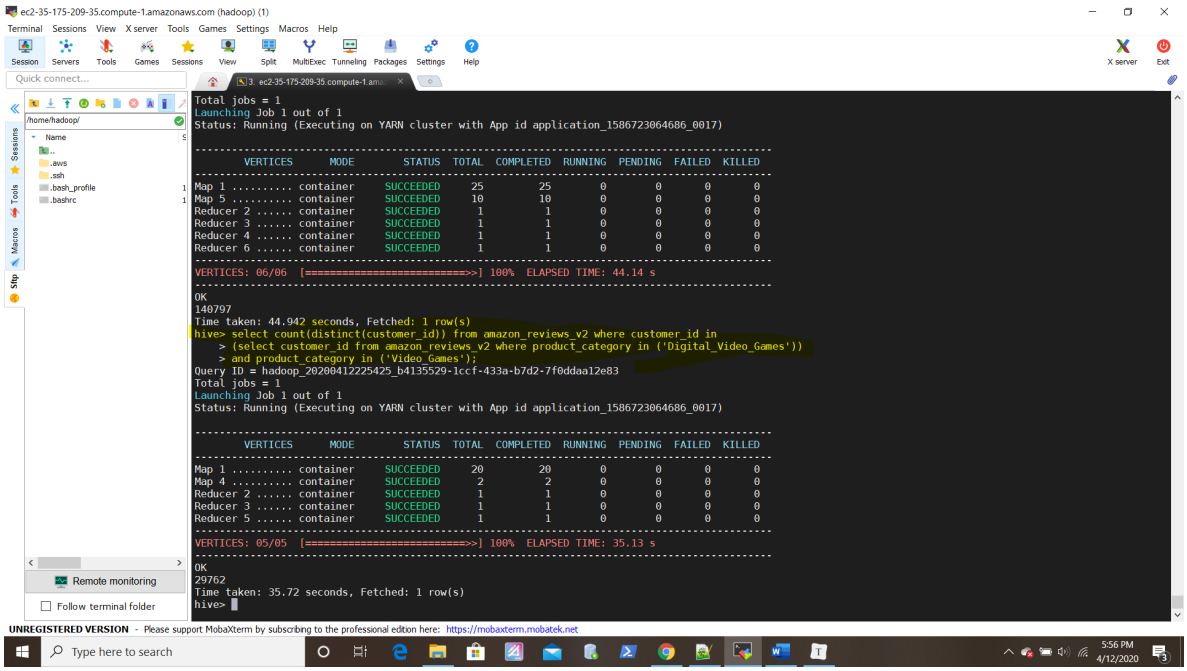
```
select count(distinct(customer_id))common_customers from amazon_reviews_v2 where
customer_id in
(select customer_id from amazon_reviews_v2 where product_category in
('Digital_Music_Purchase'))
and product_category in ('Music');
```



19)Are they any common users between Video Games and digital video Games?

From the analysis there 29762 common users

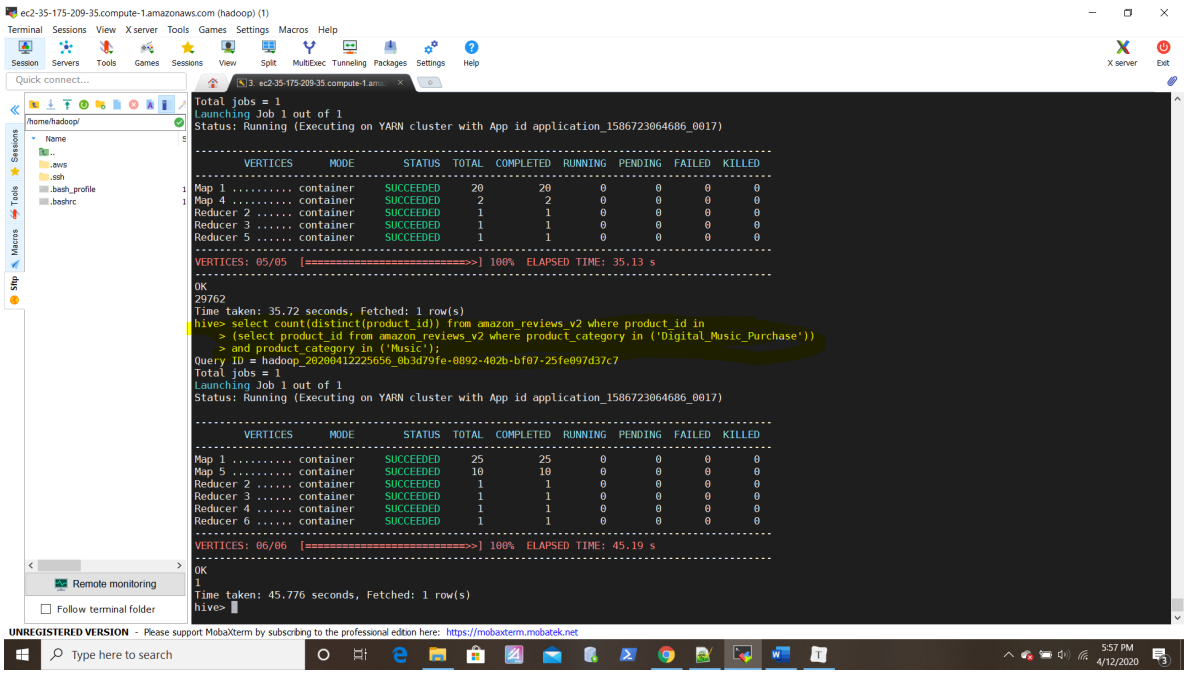
```
select count(distinct(customer_id)) from amazon_reviews_v2 where customer_id in
(select customer_id from amazon_reviews_v2 where product_category in
('Digital_Video_Games'))
and product_category in ('Video_Games');
```



20)Any common products between Digital Music Purchase and Music?

yes ,there seem to one common product between the above the category

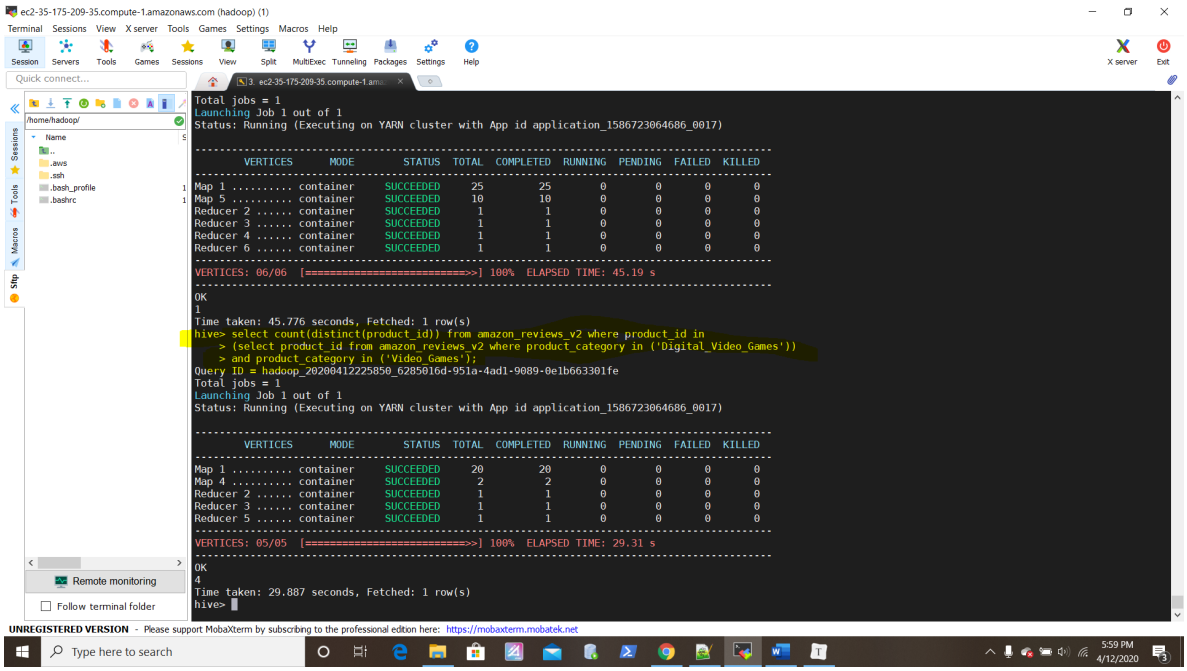
```
select count(distinct(product_id)) from amazon_reviews_v2 where product_id in
(select product_id from amazon_reviews_v2 where product_category in
('Digital_Music_Purchase'))
and product_category in ('Music');
```



21)Any common product between Digital Video Games and Video Games?

Yes there are 4 common product between video games and Digital Video Games

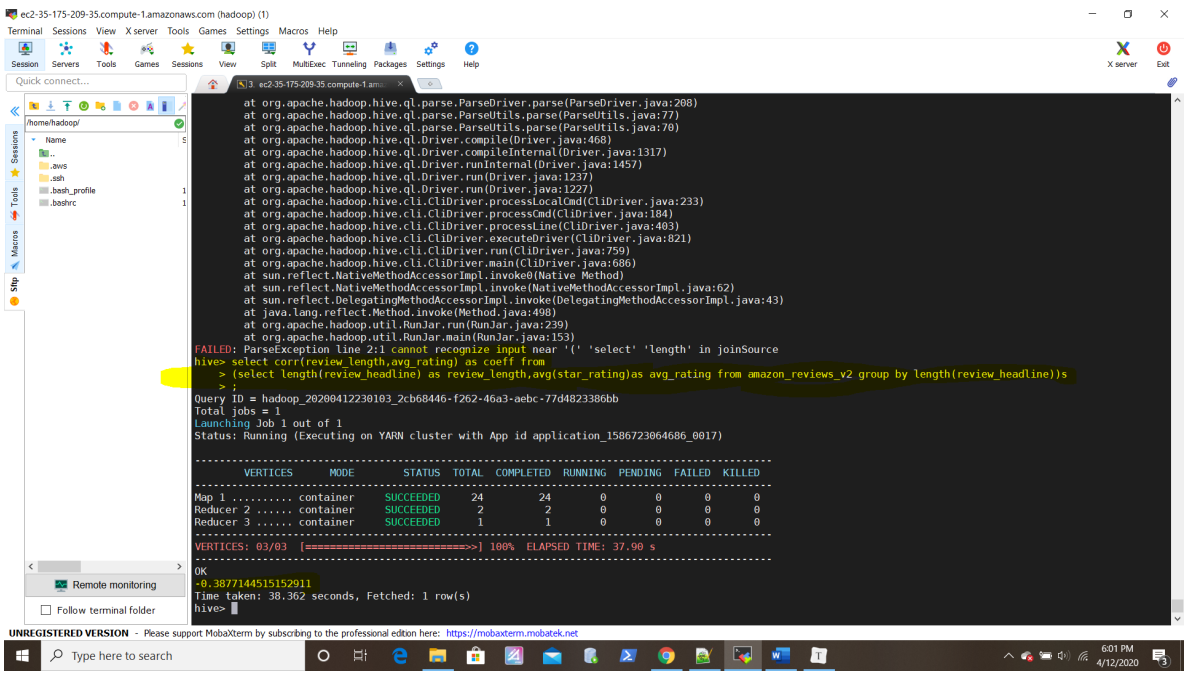
```
select count(distinct(product_id)) from amazon_reviews_v2 where product_id in
(select product_id from amazon_reviews_v2 where product_category in
('Digital_Video_Games'))
and product_category in ('Video_Games');
```



22)Is there correlation between average rating and length of review?

It seems to negatively correlated

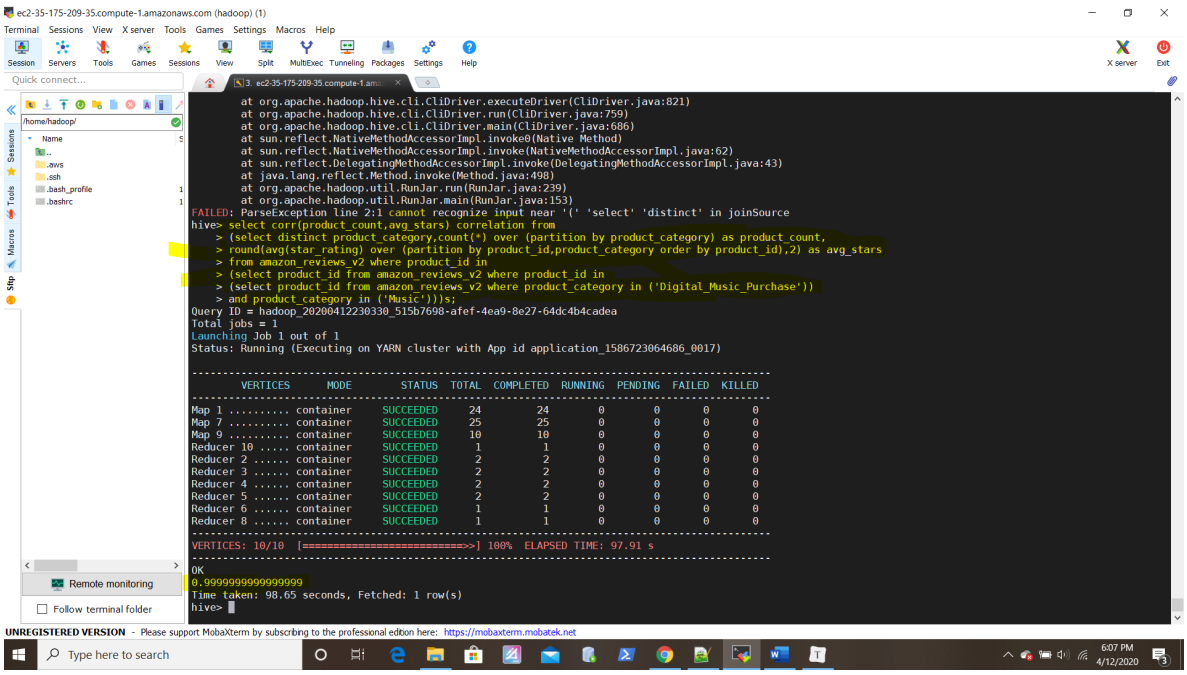
```
select corr(review_length,avg_rating) as coeff from
(select length(review_headline) as review_length,avg(star_rating)as avg_rating
from amazon_reviews_v2 group by length(review_headline))s;
```



23)Correlation Between products and star rating of Music and Digital Music Purchase ?

From the analysis there is strong correlation

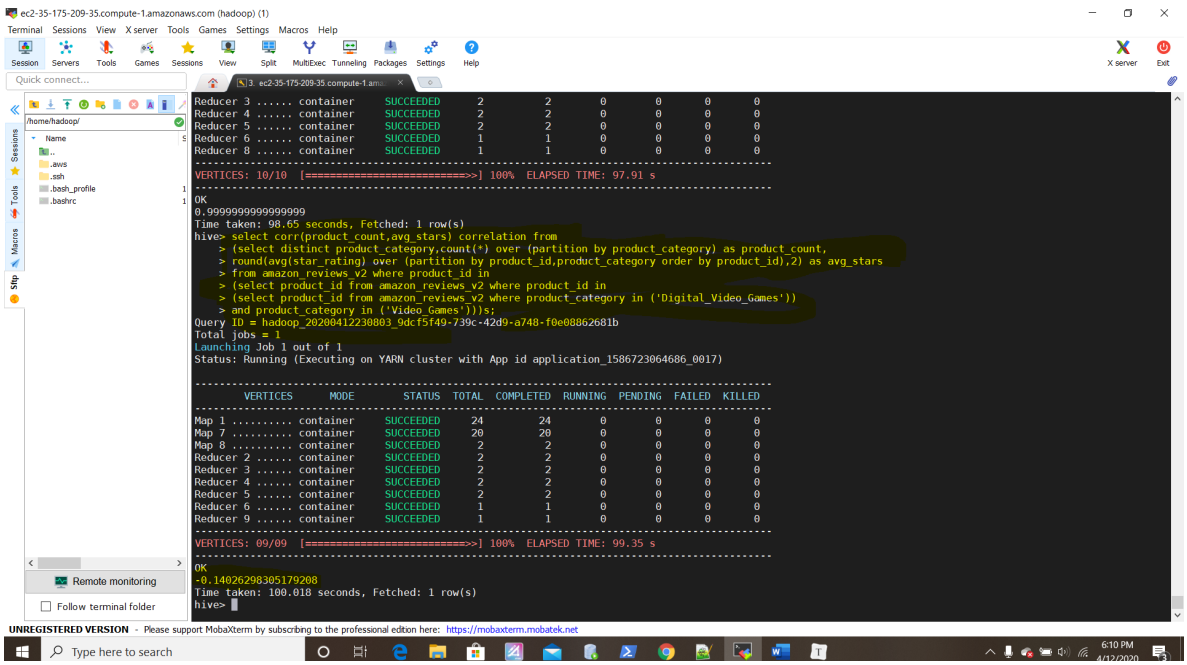
```
select corr(product_count,avg_stars) correlation from
(select distinct product_category,count(*) over (partition by product_category)
as product_count,
round(avg(star_rating) over (partition by product_id,product_category order by
product_id),2) as avg_stars
from amazon_reviews_v2 where product_id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



24)Correlation Between products and star rating of Digital Video Games and Video Games ?

It is weakly correlated

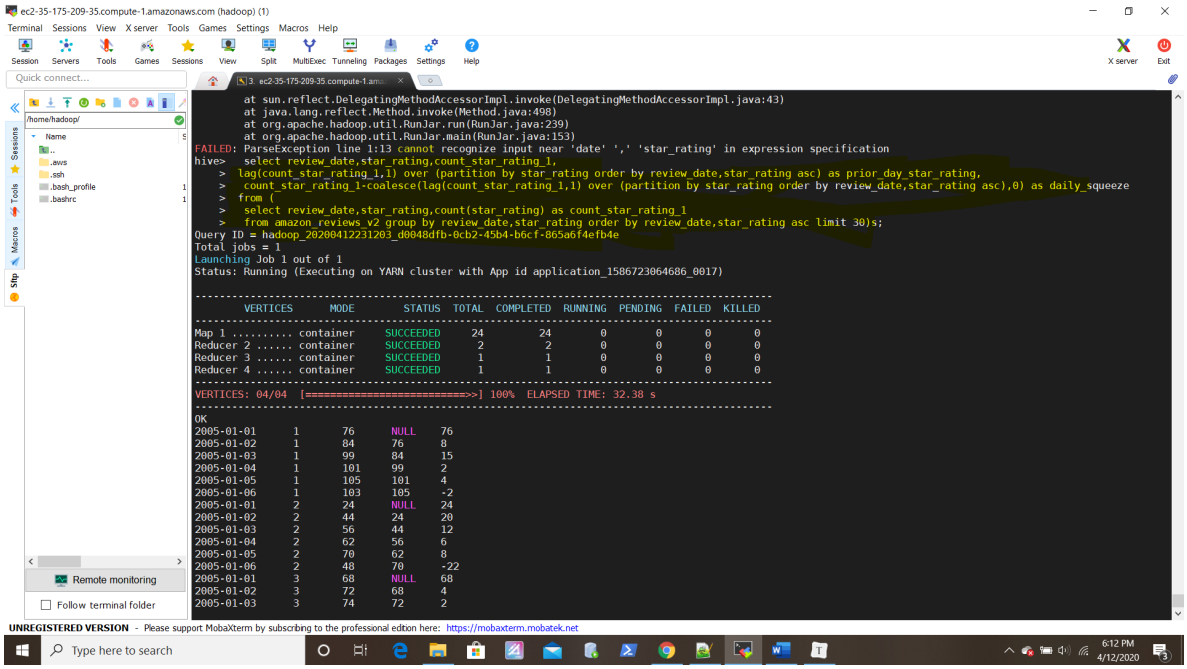
```
select corr(product_count,avg_stars) correlation from
(select distinct product_category,count(*) over (partition by product_category)
as product_count,
round(avg(star_rating) over (partition by product_id,product_category order by
product_id),2) as avg_stars
from amazon_reviews_v2 where product_id 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;
```



25)What is the prior day rating?

We can analyze the prior day rating with the help of following query which is one of the KPI's for the business purpose

```
 select review_date,star_rating,count_star_rating_1,
 lag(count_star_rating_1,1) over (partition by star_rating order by
review_date,star_rating asc) as prior_day_star_rating,
  count_star_rating_1-coalesce(lag(count_star_rating_1,1) over (partition by
star_rating order by review_date,star_rating asc),0) as daily_squeeze
 from (
  select review_date,star_rating,count(star_rating) as count_star_rating_1
  from amazon_reviews_v2 group by review_date,star_rating order by
review_date,star_rating asc limit 30)s;
```



References:

Article 'Amazon Review Data: Spotting Trends and Fake Reviews' by Mark Chopping

'A Study of Amazon User Review Data using Visualization' by Preeti Bamane