Big Data Class Project – Part 1

Amazon Customer Review Data Analysis

Introduction:

Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency. Big data has one or more of the following characteristics: high volume, high velocity or high variety. Big data analytics is the often complex process of examining large and varied data sets, or big data, to uncover information -- such as hidden patterns, unknown correlations, market trends and customer preferences -- that can help organizations make informed business decisions.

Problem description:

In this project we are practicing and demonstrating out Big Data (BD) and Analytics skills. We are using Amazon customer review data which has more than 160 Million (160,796,570) observations. We will be using HDFS, Hive and AWS to handle and analyze the data and systematically extract information from.

Data Set:
Amazon reviews dataset: *https://registry.opendata.aws/amazon-reviews/*

Project Environment:
AWS EMR - AWS Educate Class account.
Project tools:
1. AWS S3
2. AWS EMR - Hive and HDFS
3. AWS Athena - AWS alternative to Hive for the files stored in S3

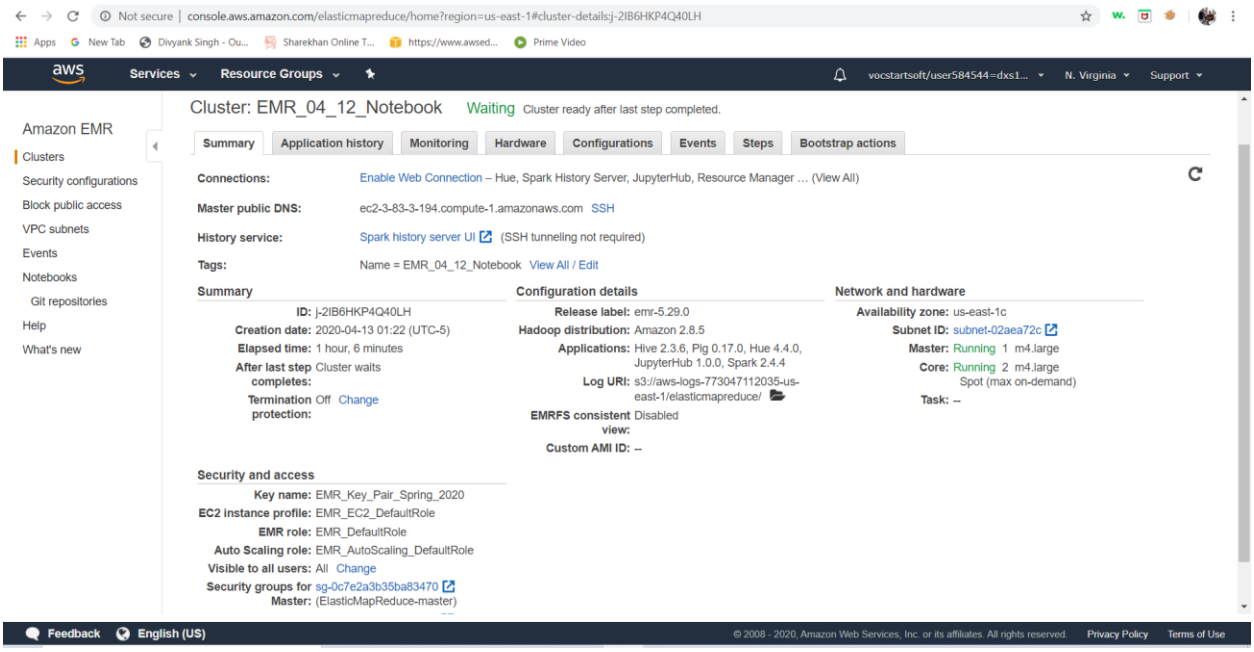
Dataset Requirements:
A)Use the following product categories:
- Wireless
- Automotive
- Music
- Digital_Music_Purchase
- Sports
- Toys
- Digital_Video_Games
- Video_Games

B)Start your analysis from year 2005.

C)Exclude multiple reviews by the same users for the same product. Each user should be allowed to review the product only once. To improve performance of your queries, create external table to point to HDFS/S3 file that will include all review-ids to be excluded.

Step 1:- AWS EMR:

Provisioned EMR with below details:-



Step 2:-

A) Created directory for below each product category:-

1. **Wireless**
2. **Automotive**
3. **Music**
4. **Digital_Music_Purchase**
5. **Sports**
6. **Toys**
7. **Digital_Video_Games**
8. **Video_Games**

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product_category=Wireless/

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product_category=Automotive/

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product_category=Music/

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product_category=Sports/

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product_category=Digital_Music_Purchase/

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product_category=Toys/

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product_category=Digital_Video_Games/

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product_category=Video_Games/

B)Copying dataset from S3 in HDFS for each of the product category mentioned:-

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product_category=Wireless/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product_category=Wireless/

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product_category=Automotive/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product_category=Automotive/

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product_category=Music/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product_category=Music/

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product_category=Digital_Music_Purchase/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product_category=Digital_Music_Purchase/

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product_category=Sports/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product_category=Sports/

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product_category=Toys/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product_category=Toys/

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product_category=Digital_Video_Games/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product_category=Digital_Video_Games/

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product_category=Video_Games/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product_category=Video_Games/

Step3:- Creating database and tables:

A)Creating Database:-

```
create database amazon_review;
```

B)Dropping table:-

```
drop table amazon_review.amazon_reviews_parquet;
```

C)Creating external table:-

```
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;

D)Creating table which has year>=2005 and product category ('Wireless', 'Automotive', 'Music', 'Digital_Music_Purchase', 'Sports', 'Toys', 'Digital_Video_Games', 'Video_Games')

Query:-

```
create table amazon_review.dat
as
(select * from amazon_review.amazon_reviews_parquet where year >= 2005);
```

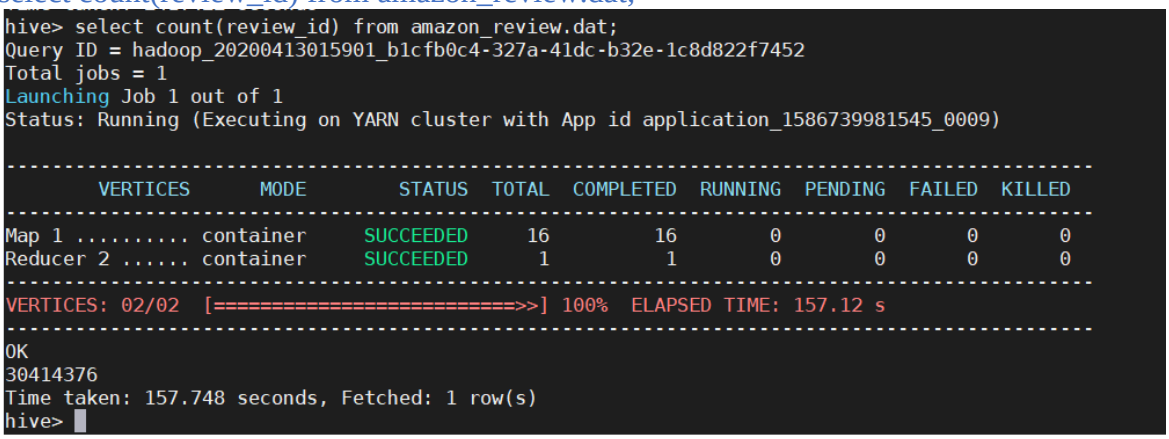
```
hive> create table amazon_review.dat
    > as
    > (select * from amazon_review.amazon_reviews_parquet where year >= 2005);
Query ID = hadoop_20200413065538_78b9a2f3-f03f-4165-a70d-aba0e27eaeeb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586759670360_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     13         13        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 246.70 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-92-232.ec2.internal:8020/user/hive/warehouse/amazon_review.db/dat
OK
Time taken: 253.586 seconds
hive>
```

E) Checking obs now:- 30414376

Query:-

```
select count(review_id) from amazon_review.dat;
```



F)Creating a table which has obs where a customer gave multiple reviews for a product:-

Query:-

```
create table amazon_review.excludedID
as
select distinct * from amazon_review.dat where
review_id in (select review_id from
              (select * from amazon_review.dat
              ) tab1
            inner join
              (select customer_id,product_id,count(*) as CNT
               from amazon_review.dat
               group by customer_id,product_id
               having count(*)>1) tab2
            on tab1.customer_id=tab2.customer_id and tab1.product_id=tab2.product_id);
```

```
    > ;
hive> create table amazon_review.excludedID
    >  as
    > select distinct * from amazon_review.dat where
    > review_id in (select review_id from
    >                 (select * from amazon_review.dat
    >                 ) tab1
    >             inner join
    >                 (select customer_id,product_id,count(*) as CNT
    >                  from amazon_review.dat
    >                  group by customer_id,product_id
    >                  having count(*)>1) tab2
    >             on tab1.customer_id=tab2.customer_id and tab1.product_id=tab2.product_id);
Query ID = hadoop_20200413070234_8cd2a1ea-bf18-4e6c-8f1e-5e697e03fbc7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586759670360_0009)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     17         17        0        0       0       0
Map 4 .......... container     SUCCEEDED     17         17        0        0       0       0
Map 7 .......... container     SUCCEEDED     17         17        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     90         90        0        0       0       0
Reducer 3 ...... container     SUCCEEDED     64         64        0        0       0       0
Reducer 5 ...... container     SUCCEEDED     68         68        0        0       0       0
Reducer 6 ...... container     SUCCEEDED     64         64        0        0       0       0
Reducer 8 ...... container     SUCCEEDED     58         58        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 08/08  [==========================>>] 100%  ELAPSED TIME: 820.28 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-92-232.ec2.internal:8020/user/hive/warehouse/amazon_review.db/excludedid
OK
Time taken: 824.023 seconds
```

G)Checking obs now:- 790413

Query:-

```
select count(review_id) from amazon_review.excluded_reviewid;
```

```
hive> select count(review_id) from amazon_review.excludedid
    > ;
Query ID = hadoop_20200413022644_1c81ea27-91a6-442e-b9fa-d0355e580b25
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586739981545_0010)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     13         13        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 23.18 s
--------------------------------------------------------------------------------
OK
790413
Time taken: 23.804 seconds, Fetched: 1 row(s)
hive>
```

H)Saving the obs where customer's has only 1 review for a product:-

Query:-

```
create table amazon_review.final
as
SELECT *
  FROM amazon_review.dat
  WHERE NOT EXISTS
   (SELECT * FROM amazon_review.excludedID
  WHERE dat.review_id = excludedid.review_id);
```

```
hive> create table amazon_review.final1
    > as
    > SELECT *
    >   FROM amazon_review.dat
    >   WHERE NOT EXISTS
    >    (SELECT * FROM amazon_review.excludedID
    >   WHERE dat.review_id = excludedid.review_id);
Query ID = hadoop_20200413043502_a0d3f266-85b9-4e80-be05-4bd097bd2e02
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_1586739981545_0014)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     16         16        0        0       0       0
Map 3 .......... container     SUCCEEDED     13         13        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     59         59        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 810.09 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-25-39.ec2.internal:8020/user/hive/warehouse/amazon_review.db/final1
OK
Time taken: 822.785 seconds
```

I) Checking obs now:- 28845525

Query:-

select count(review_id) from amazon_review.final;

```
hive> select count(*) from final1;
OK
28845525
Time taken: 0.15 seconds, Fetched: 1 row(s)
hive>
```

Step4:- Performing Exploratory Data Analysis

1. Explore the dataset and provide basic exploratory analysis:

1. **Number of reviews**

   Query:-

   select count(review_id) from amazon_review.final;

   Output:-

   28845525

```
hive> select count(review_id) from amazon_review.final1;
Query ID = hadoop_20200413045328_6f1386c3-9c34-4f85-93c5-3d7cb7a9946f
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_1586739981545_0015)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     15         15        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 684.03 s
--------------------------------------------------------------------------------
OK
28845525
Time taken: 692.902 seconds, Fetched: 1 row(s)
```

2. **Number of users**

Query:-

select count(customer_id) from amazon_review.final;

Output:-

28845525

```
hive> select count( customer_id) from amazon_review.final1;
Query ID = hadoop_20200413050550_424bcd8b-8906-4755-9387-12b1ab4f0504
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_1586739981545_0016)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     15         15        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 125.77 s
--------------------------------------------------------------------------------
OK
28845525
Time taken: 134.132 seconds, Fetched: 1 row(s)
```

3. **Average review stars**

Query:-

select round(avg(star_rating),2) from amazon_review.final;

Output:-

4.17

```
hive> select round(avg(star_rating),2) from amazon_review.final1;
Query ID = hadoop_20200413050755_fc1fc670-2c7c-4681-bf7f-419cc9cf1f28
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586739981545_0017)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     15         15        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 149.05 s
--------------------------------------------------------------------------------
OK
4.17
Time taken: 154.352 seconds, Fetched: 1 row(s)
```

4. **Average length of the review**

Query:-

select round(avg(length(review_body)),2) from amazon_review.final;

Output:-

301.83

```
hive> select round(avg(length(review_body)),2) from amazon_review.final1;
Query ID = hadoop_20200413050848_83699ddf-8690-4b7e-a572-b2e4de7d0d9d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586739981545_0016)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     15         15        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 212.14 s
--------------------------------------------------------------------------------
OK
301.83
Time taken: 212.852 seconds, Fetched: 1 row(s)
```

5. **Number of verified versus unverified reviews**

Query:-

```
select verified_purchase,count(*) as count_ver_pur from amazon_review.final
group by verified_purchase;
```

Output:-

	verified_purchase	count_ver_pur
1	Y	23850537
2	N	4994988

(Athena)

```
hive> select verified_purchase,count(*) as count_ver_pur from amazon_review.final1
    > group by verified_purchase;
Query ID = hadoop_20200413051241_5f0d416d-b75c-4fbd-9482-b10e044ae6cc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586739981545_0017)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     15         15        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     54         54        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 142.49 s
--------------------------------------------------------------------------------
OK
Y       23850537
N       4994988
Time taken: 143.23 seconds, Fetched: 2 row(s)
```

(Terminal)

6. **At least two more additional metrics:-**

A) Count of each product_category:-

Query:-

```
Select product_category, count_prod_cat, rank() over
 (ORDER BY count_prod_cat   DESC)
```

```
From
(select product_category,count(*) as count_prod_cat from amazon_review.final
group by product_category) r
;
```

Output:-

```
hive> Select product_category, count_prod_cat, rank() over
    >  (ORDER BY count_prod_cat   DESC)
    > From
    > (select product_category,count(*) as count_prod_cat from amazon_review.final
    > group by product_category) r
    > ;
Query ID = hadoop_20200413091733_8934ee15-7fdd-4249-9472-8d70f8dc7b96
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_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     54         54        0        0       0       0
Reducer 3 ...... container     SUCCEEDED     27         27        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 78.05 s
--------------------------------------------------------------------------------
OK
Wireless        8978302 1
Sports  4847051 2
Toys    4795936 3
Automotive      3514995 4
Music   3331361 5
Digital_Music_Purchase  1636189 6
Video_Games     1596278 7
Digital_Video_Games     145413  8
Time taken: 87.884 seconds, Fetched: 8 row(s)
```

(Terminal)

	product_category	count_prod_cat
1	Wireless	8978302
2	Sports	4847051
3	Toys	4795936
4	Automotive	3514995
5	Music	3331361
6	Digital_Music_Purchase	1636189
7	Video_Games	1596278
8	Digital_Video_Games	145413

num	product_category	count_prod_cat	average
1	Wireless	8978302	31.13
2	Sports	4847051	16.8
3	Toys	4795936	16.63
4	Automotive	3514995	12.19
5	Music	3331361	11.55
6	Digital_Music_Purchase	1636189	5.67
7	Video_Games	1596278	5.53
8	Digital_Video_Games	145413	0.5
	Sum	28845525	

Interpretation:-
We can see that more that 31% of the reviews are of wireless products, followed by Sports(16.8%) and Toys(16.63%), whereas Video games and digital video category constitutes only 6%.

B) Count of each marketplace:-
Query:-

```
Select marketplace, count_mar_plc, rank() over
 (ORDER BY count_mar_plc DESC)
from
(select marketplace,count(*) as count_mar_plc from amazon_review.final
group by marketplace) as r
;
```

Output:-

```
hive> Select marketplace, count_mar_plc, rank() over
    >  (ORDER BY count_mar_plc DESC)
    > from
    > (select marketplace,count(*) as count_mar_plc from amazon_review.final
    > group by marketplace) as r
    > ;
Query ID = hadoop_20200413092125_c66b904c-bd71-46c9-a50d-2a3a61066fdf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     54         54        0        0       0       0
Reducer 3 ...... container     SUCCEEDED     27         27        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 82.91 s
--------------------------------------------------------------------------------
OK
US      28125769        1
UK      368706  2
DE      187164  3
FR      82134   4
JP      81752   5
Time taken: 83.549 seconds, Fetched: 5 row(s)
```

	marketplace	count_mar_plc
1	JP	81752
2	FR	82134
3	DE	187164
4	UK	368706
5	US	28125769

	marketplace	count_mar_plc	per
1	JP	81752	0.28341311
2	FR	82134	0.28473741
3	DE	187164	0.64884934
4	UK	368706	1.27820866
5	US	28125769	97.5047915
	sum	28845525	

Interpretation:-
I have grouped dataset based on marketplace and found that more than 97% of the customers are from just US.

C) Finding mean,min,max,standard deviation per category:-
Query:-

```
select product_category,round(avg(star_rating),2) mean_rat ,min(star_rating) min_rat
,max(star_rating) as max_rat, round(stddev(star_rating),2) as std_dev_rat from
amazon_review.final
group by product_category
order by product_category;
```

Output:-

```
hive> select product_category,round(avg(star_rating),2) mean_rat ,min(star_rating) min_rat
    > ,max(star_rating) as max_rat, round(stddev(star_rating),2) as std_dev_rat from amazon_review.final
    > group by product_category
    > order by product_category;
Query ID = hadoop_20200413100655_13728321-bc8e-49eb-adc0-7a5227a4b20d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     54         54        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 78.54 s
--------------------------------------------------------------------------------
OK
Automotive      4.25    1       5       1.26
Digital_Music_Purchase  4.65    1       5       0.86
Digital_Video_Games     3.85    1       5       1.54
Music   4.48    1       5       0.99
Sports  4.23    1       5       1.23
Toys    4.21    1       5       1.26
Video_Games     4.07    1       5       1.36
Wireless        3.89    1       5       1.46
Time taken: 79.13 seconds, Fetched: 8 row(s)
```

	product_category	mean_rat	min_rat	max_rat	std_dev_rat
1	Automotive	4.25	1	5	1.26
2	Digital_Music_Purchase	4.65	1	5	0.86
3	Digital_Video_Games	3.85	1	5	1.54
4	Music	4.48	1	5	0.99
5	Sports	4.23	1	5	1.23
6	Toys	4.21	1	5	1.26
7	Video_Games	4.07	1	5	1.36
8	Wireless	3.89	1	5	1.46

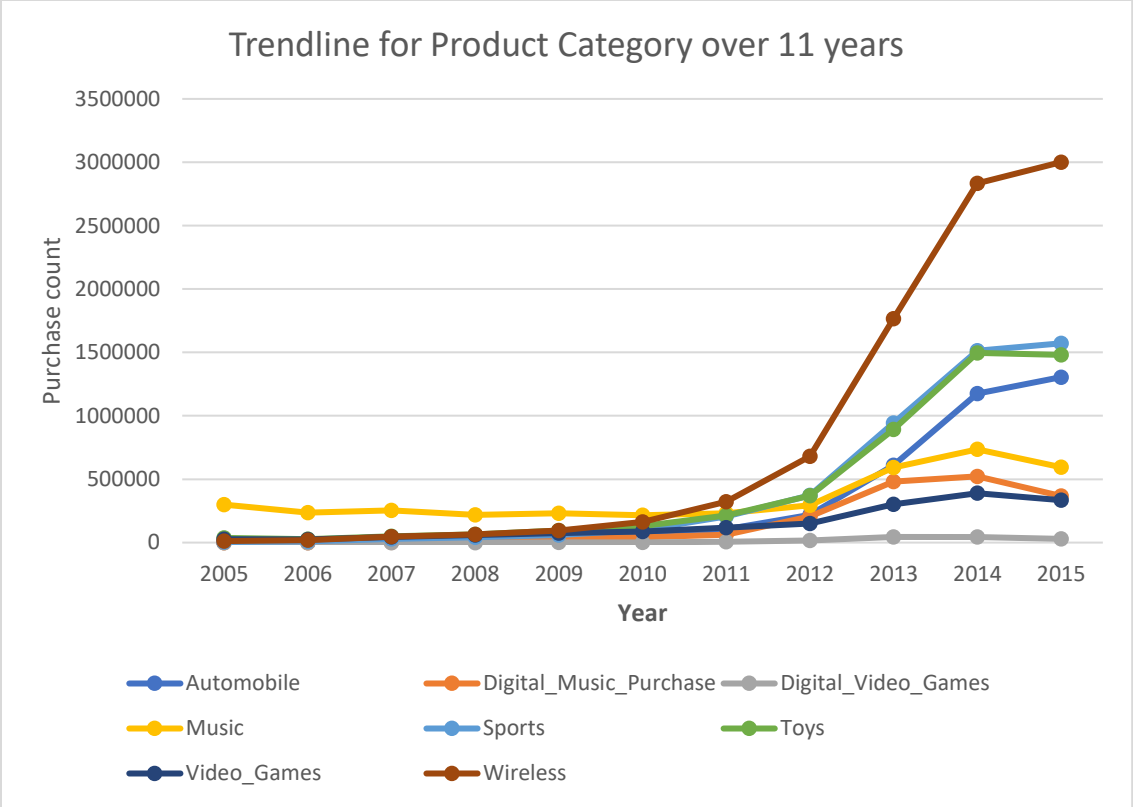
Interpretation:-

The output shows the minimum, maximum, mean and standard deviation for each of the product category.

7. **Provide trending (over time) analysis of each of the metrics above:-**

Query:-

```
select product_category,year,count(*) as count_prod from amazon_review.final
group by product_category,year
order by product_category,year
;
```

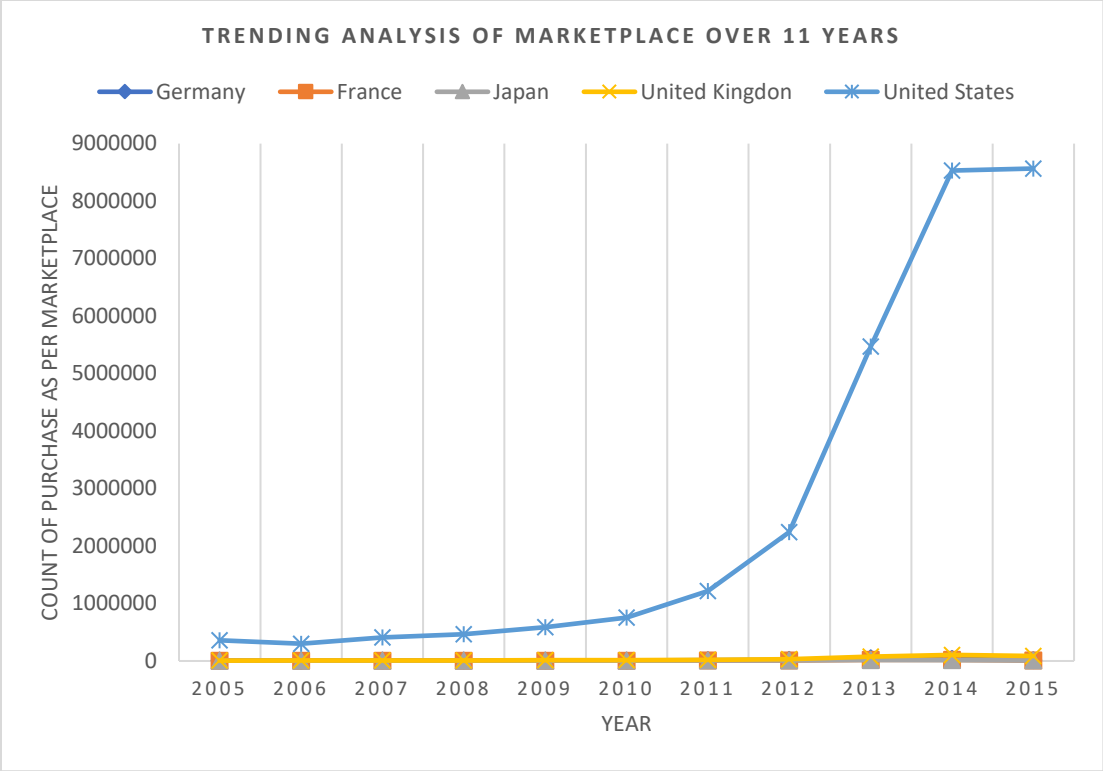


Interpretation:-

I have grouped dataset based on product category and year and imported the output in excel to get the above trendline, which shows that purchase count has definitely increased for all the product categories, though wireless category has noticed increase in purchase count rate highest.

Query:-

```
select marketplace,year,count(*) as count_marketplace from amazon_review.final
group by marketplace,year
order by marketplace,year
;
```



Interpretation:-

I have grouped dataset based on marketplace and year and imported the output in excel to get the above trendline, which shows that purchase count rate has significantly increased in US only.

2. Provide detailed analysis of Music/Digital_Music_Purchase and Digital_Video_Games/Video_Games over time.

1. Do you see correlation (maybe negative) between the categories over time?
Part 1:-Checking between Music/Digital_Music_Purchase
Query:-

```
select year,count(*) as count from amazon_review.final
where product_category in ('Music')
group by year
order by year;
```

Output:-

```
hive> select year,count(*) as count from amazon_review.final
    > where product_category in ('Music')
    > group by year
    > order by year;
Query ID = hadoop_20200413092524_4194aa9c-f754-4020-b243-f73890bc85a6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     27         27        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 84.28 s
--------------------------------------------------------------------------------
OK
2005    254481
2006    204103
2007    223403
2008    193683
2009    204559
2010    190525
2011    201344
2012    250574
2013    498207
2014    612780
2015    497702
Time taken: 84.911 seconds, Fetched: 11 row(s)
```

	year	count
1	2005	299887
2	2006	235827
3	2007	254733
4	2008	219129
5	2009	230984
6	2010	215737
7	2011	230394
8	2012	294877
9	2013	591170
10	2014	735349
11	2015	594836

Interpretation:-
Here I have performed grouping dataset based on year for product category “Music”

Query:-

```
select year,count(*) as count from amazon_review.final
where product_category in ('Digital_Music_Purchase')
group by year
order by year;
```

Output:-

```
hive> select year,count(*) as count from amazon_review.final
    > where product_category in ('Digital_Music_Purchase')
    > group by year
    > order by year;
Query ID = hadoop_20200413092752_364333d0-1da3-45fe-80ae-b2a8713139f0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     27         27        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 77.91 s
--------------------------------------------------------------------------------
OK
2005    8
2006    21
2007    2235
2008    22040
2009    36180
2010    40807
2011    57237
2012    192623
2013    448558
2014    487772
2015    348708
Time taken: 78.604 seconds, Fetched: 11 row(s)
```

	year	count
1	2005	8
2	2006	21
3	2007	2313
4	2008	23109
5	2009	37695
6	2010	42821
7	2011	61701
8	2012	207079
9	2013	479899
10	2014	522272
11	2015	367254

Here I have performed grouping dataset based on year for product category “Music”

Query:- (using corr function to find the correlation between music product category and year)

```
select corr(count_prod,year) as corr_music from
```

(select product_category,year,count(*) as count_prod from amazon_review.final
where product_category in ('Music')
group by product_category,year
order by product_category,year) as r;

Output:-

```
hive> select corr(count_prod,year) as corr_music from
    > (select product_category,year,count(*) as count_prod from amazon_review.final
    > where product_category in ('Music')
    > group by product_category,year
    > order by product_category,year) as r;
Query ID = hadoop_20200413093041_64dc7799-e97a-485b-9aa8-f88c51c6f9ba
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     27         27        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 81.49 s
--------------------------------------------------------------------------------
OK
0.7401744556901582
Time taken: 82.054 seconds, Fetched: 1 row(s)
```

	corr_music
1	0.7386732

I did a correlation test between music category purchase count and years and found that there is a high positive correlation between them, which states as year increased so does the purchase count.

Query:- (using corr function to find the correlation between digital_music_category product category and year)

select round(corr(count_prod,year),2) as corr_music from
(select product_category,year,count(*) as count_prod from amazon_review.final
where product_category in ('Digital_Music_Purchase')
group by product_category,year
order by product_category,year) as r;

Output:-

```
hive> select round(corr(count_prod,year),2) as corr_music from
    > (select product_category,year,count(*) as count_prod from amazon_review.final
    > where product_category in ('Digital_Music_Purchase')
    > group by product_category,year
    > order by product_category,year) as r;
Query ID = hadoop_20200413093312_03a81a87-abac-4d38-89c7-32886bdd24aa
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     27         27        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 71.75 s
--------------------------------------------------------------------------------
OK
0.86
Time taken: 72.341 seconds, Fetched: 1 row(s)
```

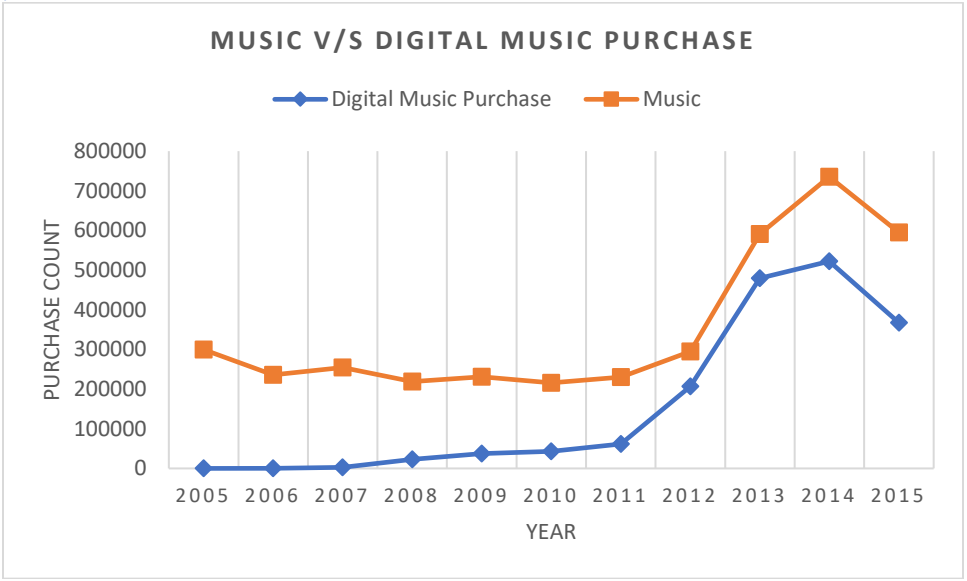
	corr_music
1	0.854252

I did a correlation test between digital music category purchase count and years and found that there is a high positive correlation between them, which states as year increased so does the purchase count.

Detailed analysis:-

Query:-

```
select product_category,year,count(*) as count_prod from amazon_review.final
where product_category in ('Music', 'Digital_Music_Purchase')
group by product_category,year
order by product_category,year
;
```



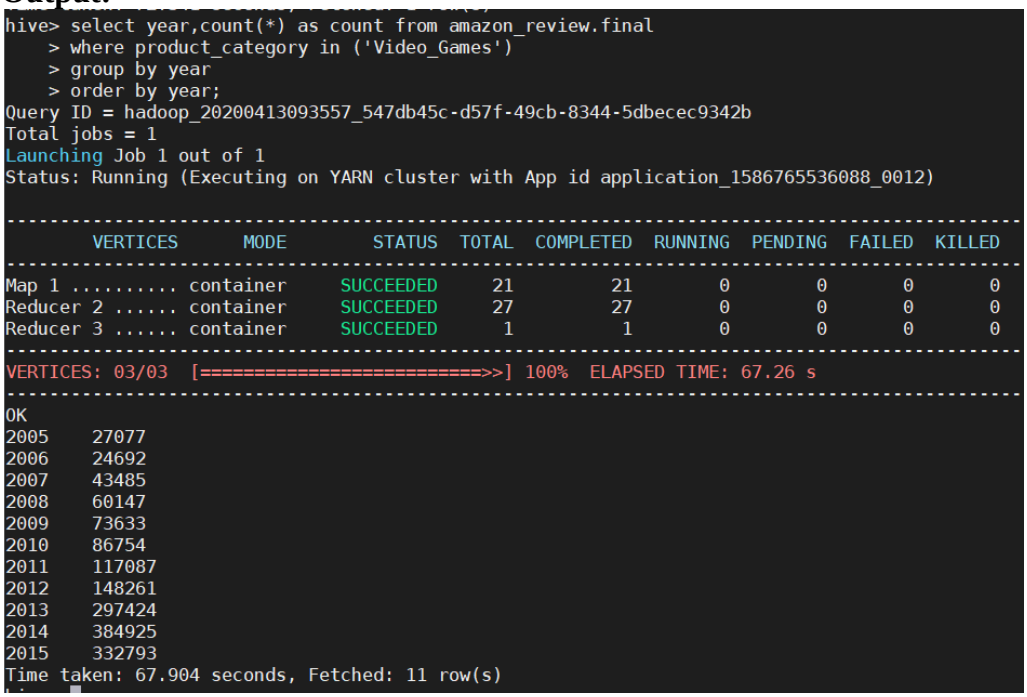
The above trendline between year and music/digital_music_category even shows positive correlation

Part2:-Checking between Digital_Video_Games/Video_Games

Query:-

```
select year,count(*) as count from amazon_review.final
where product_category in ('Video_Games')
group by year
order by year;
```

Output:-



```
hive> select year,count(*) as count from amazon_review.final
    > where product_category in ('Video_Games')
    > group by year
    > order by year;
Query ID = hadoop_20200413093557_547db45c-d57f-49cb-8344-5dbecec9342b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     27         27        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 67.26 s
--------------------------------------------------------------------------------
OK
2005    27077
2006    24692
2007    43485
2008    60147
2009    73633
2010    86754
2011    117087
2012    148261
2013    297424
2014    384925
2015    332793
Time taken: 67.904 seconds, Fetched: 11 row(s)
```

Here I have performed grouping dataset based on year for product category “Video_Games”

Query:-

```
select year,count(*) as count from amazon_review.final
where product_category in ('Digital_Video_Games')
group by year
order by year;
```

Output:-

```
hive> select year,count(*) as count from amazon_review.final
    > where product_category in ('Digital_Video_Games')
    > group by year
    > order by year;
Query ID = hadoop_20200413093742_7219b5f7-b5a1-43fb-a727-1115533ce734
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     27         27        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 69.24 s
--------------------------------------------------------------------------------
OK
2006    1
2008    5
2009    1561
2010    2551
2011    7644
2012    16624
2013    43262
2014    43748
2015    30017
Time taken: 69.779 seconds, Fetched: 9 row(s)
```

Here I have performed grouping dataset based on year for product category "Digital_Video_Games"

Query:- (using corr function to find the correlation between Video_Games product category and year)

```
select round(corr(count_prod,year),2) as corr_vid from
(select product_category,year,count(*) as count_prod from amazon_review.final
where product_category in ('Video_Games')
group by product_category,year
order by product_category,year) as r;
```

Output:-

```
hive> select round(corr(count_prod,year),2) as corr_vid from
    > (select product_category,year,count(*) as count_prod from amazon_review.final
    > where product_category in ('Video_Games')
    > group by product_category,year
    > order by product_category,year) as r;
Query ID = hadoop_20200413093947_62569215-537a-40a3-b11d-8f02de35dc9a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     27         27        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 79.40 s
--------------------------------------------------------------------------------
OK
0.91
Time taken: 80.073 seconds, Fetched: 1 row(s)
```

I did a correlation test between video games category purchase count and years and found that there is a high positive correlation between them, which states as year increased so does the purchase count.

Query:- (using corr function to find the correlation between Digital_Video_Games product category and year)

```
select round(corr(count_prod,year),2) as corr_vid from
(select product_category,year,count(*) as count_prod from amazon_review.final
where product_category in ('Digital_Video_Games')
group by product_category,year
order by product_category,year) as r;
```

Output:-

```
hive> select round(corr(count_prod,year),2) as corr_vid from
    > (select product_category,year,count(*) as count_prod from amazon_review.final
    > where product_category in ('Digital_Video_Games')
    > group by product_category,year
    > order by product_category,year) as r;
Query ID = hadoop_20200413094142_f5b17d16-f667-4bc7-90c6-73d8e7f7a3e2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     27         27        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 88.63 s
--------------------------------------------------------------------------------
OK
0.85
Time taken: 89.228 seconds, Fetched: 1 row(s)
```

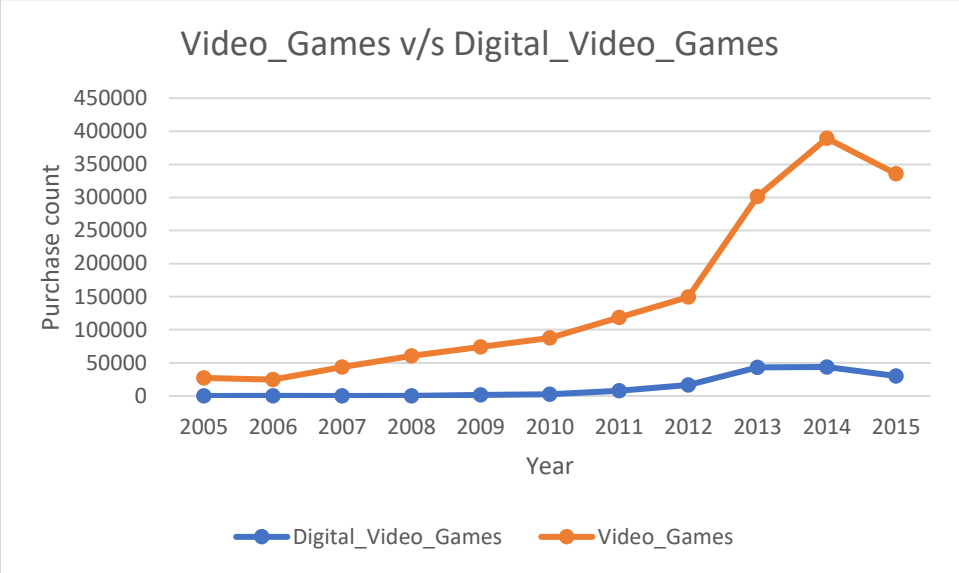
Interpretation:-

I did a correlation test between digital video games category purchase count and years and found that there is a high positive correlation between them, which states as year increased so does the purchase count.

Detailed Analysis:-

Query:-

```
select product_category,year,count(*) as count_prod from amazon_review.final
where product_category in ('Video_Games', 'Digital_Video_Games')
group by product_category,year
order by product_category,year
;
```



Interpretation:-

The above trendline between year and music/digital_music_category even shows positive correlation

2. Are there same users reviewing in both categories?

Query:-

```
select count(distinct dmp.customer_id) as common_cus from
   (select customer_id from amazon_review.final
    where product_category = 'Digital_Music_Purchase') as dmp
inner join
   (select customer_id from amazon_review.final
    where product_category = 'Music') as mus
on dmp.customer_id=mus.customer_id;
```

Output:-

```
hive> select count(distinct dmp.customer_id) as common_cus from
    >     (select customer_id from amazon_review.final
    >      where product_category = 'Digital_Music_Purchase') as dmp
    > inner join
    >     (select customer_id from amazon_review.final
    >      where product_category = 'Music') as mus
    > on dmp.customer_id=mus.customer_id;
Query ID = hadoop_20200413094415_bb8e92da-0c6f-4bc0-b134-621a7bd56781
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Map 5 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     54         54        0        0       0       0
Reducer 3 ...... container     SUCCEEDED     30         30        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 175.31 s
--------------------------------------------------------------------------------
OK
140797
Time taken: 176.096 seconds, Fetched: 1 row(s)
```

	common_cus
1	140797

Interpretation:-
For Music and Digital_Music_Purchase there are 140797 common customers.

Query:-

```
select count(distinct dvs.customer_id) as common_cus from
   (select customer_id from amazon_review.final
    where product_category = 'Digital_Video_Games') as dvs
inner join
   (select customer_id from amazon_review.final
    where product_category = 'Video_Games') as vs
on dvs.customer_id=vs.customer_id;
```

Output:-

```
Time taken: 176.096 seconds, Fetched: 1 row(s)
hive> select count(distinct dvs.customer_id) as common_cus from
    >     (select customer_id from amazon_review.final
    >      where product_category = 'Digital_Video_Games') as dvs
    > inner join
    >     (select customer_id from amazon_review.final
    >      where product_category = 'Video_Games') as vs
    > on dvs.customer_id=vs.customer_id;
Query ID = hadoop_20200413094854_399c0d99-119b-4140-9955-e825453f18bf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Map 5 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     54         54        0        0       0       0
Reducer 3 ...... container     SUCCEEDED     30         30        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 149.88 s
--------------------------------------------------------------------------------
OK
29762
Time taken: 150.577 seconds, Fetched: 1 row(s)
```

	common_cus
1	29762

Interpretation:-
For Video_Games and Digital_Video_Games there are 29,762 common customers.

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

Part1:-Video_Games and Digital_Video_Games
A) Finding similar items in both categories:-
Query:-

```
select distinct vg.product_id from
(select product_id from final
where product_category='Video_Games') vg
inner join
(select product_id from final
where product_category='Digital_Video_Games') dvg
on vg.product_id=dvg.product_id
;
```

Output:-

```
hive> select distinct vg.product_id from
    > (select product_id from final
    > where product_category='Video_Games') vg
    > inner join
    > (select product_id from final
    > where product_category='Digital_Video_Games') dvg
    > on vg.product_id=dvg.product_id
    > ;
Query ID = hadoop_20200413095157_b805fd4b-482c-43d1-8be8-7cdb04e9c9a5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Map 4 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     54         54        0        0       0       0
Reducer 3 ...... container     SUCCEEDED     30         30        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 115.06 s
--------------------------------------------------------------------------------
OK
B00NBBME0Y
B004YNII9Y
B00B4WVTUS
B0047T7MEW
Time taken: 115.685 seconds, Fetched: 4 row(s)
```

	product_id
1	B00B4WVTUS
2	B0047T7MEW
3	B00NBBME0Y
4	B004YNII9Y

Interpretation:-
There is 4 similar product between Video_Games and Digital_Video_Games

B) Do they get similar ratings:-

```
select vg.product_id,vg_avg_rat,dvg_avg_rat,round((vg_avg_rat-dvg_avg_rat),2) as diff
from
(select product_id, round(avg(star_rating),2) as vg_avg_rat from final
where product_category='Video_Games'
group by product_id) vg
inner join
(select product_id, round(avg(star_rating),2) as dvg_avg_rat from final
where product_category='Digital_Video_Games'
group by product_id) dvg
on vg.product_id=dvg.product_id;
```

Output:-

```
hive> select vg.product_id,vg_avg_rat,dvg_avg_rat,round((vg_avg_rat-dvg_avg_rat),2) as diff  from
    > (select product_id, round(avg(star_rating),2) as vg_avg_rat from final
    > where product_category='Video_Games'
    > group by product_id) vg
    > inner join
    > (select product_id, round(avg(star_rating),2) as dvg_avg_rat from final
    > where product_category='Digital_Video_Games'
    > group by product_id) dvg
    > on vg.product_id=dvg.product_id;
Query ID = hadoop_20200413095627_d6b596af-6498-4018-a949-e68f1edb1df8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Map 4 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     27         27        0        0       0       0
Reducer 3 ...... container     SUCCEEDED     27         27        0        0       0       0
Reducer 5 ...... container     SUCCEEDED     27         27        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 129.71 s
----------------------------------------------------------------------------------------------
OK
B004YNII9Y      4.0     2.82    1.18
B0047T7MEW      4.5     3.69    0.81
B00B4WVTUS      5.0     4.64    0.36
B00NBBME0Y      3.47    5.0     -1.53
Time taken: 130.573 seconds, Fetched: 4 row(s)
```

	product_id	vg_avg_rat	dvg_avg_rat	diff
1	B004YNII9Y	4.0	2.82	1.18
2	B00NBBME0Y	3.47	5.0	-1.53
3	B00B4WVTUS	5.0	4.64	0.36
4	B0047T7MEW	4.5	3.69	0.81

Interpretation:-

The above table shows that they don't get similar ratings.

Part2:- Music v/s Digital_Music_Purchase

Query:-

```
select distinct m.product_id from
(select product_id from final
where product_category='Music') m
inner join
(select product_id from final
where product_category='Digital_Music_Purchase') dmp
on m.product_id=dmp.product_id;
;
```

Output:-

```
hive> select distinct m.product_id from
    > (select product_id from final
    > where product_category='Music') m
    > inner join
    > (select product_id from final
    > where product_category='Digital_Music_Purchase') dmp
    > on m.product_id=dmp.product_id;
Query ID = hadoop_20200413095918_672a666b-603f-4e3c-b192-c4a73b24615c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Map 4 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     54         54        0        0       0       0
Reducer 3 ...... container     SUCCEEDED     30         30        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 136.62 s
--------------------------------------------------------------------------------
OK
B0019M1ZJS
Time taken: 137.348 seconds, Fetched: 1 row(s)
```

	product_id
1	B0019M1ZJS

Interpretation:-
There is only 1 similar product between Video_Games and Digital_Video_Games

Query:-

```
select m.product_id,mus_avg_rat,dmp_avg_rat from
(select product_id, avg(star_rating) as mus_avg_rat from final
where product_category='Music'
group by product_id) m
inner join
(select product_id, avg(star_rating) as dmp_avg_rat from final
where product_category='Digital_Music_Purchase'
group by product_id) dmp
on m.product_id=dmp.product_id;
```

Output:-

```
hive> select m.product_id,mus_avg_rat,dmp_avg_rat from
    > (select product_id, avg(star_rating) as mus_avg_rat from final
    > where product_category='Music'
    > group by product_id) m
    > inner join
    > (select product_id, avg(star_rating) as dmp_avg_rat from final
    > where product_category='Digital_Music_Purchase'
    > group by product_id) dmp
    > on m.product_id=dmp.product_id;
Query ID = hadoop_20200413100215_bf660d70-9a63-4811-8246-e525b085a699
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Map 4 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     27         27        0        0       0       0
Reducer 3 ...... container     SUCCEEDED     27         27        0        0       0       0
Reducer 5 ...... container     SUCCEEDED     27         27        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 138.83 s
--------------------------------------------------------------------------------
OK
B0019M1ZJS      3.0     5.0
Time taken: 139.457 seconds, Fetched: 1 row(s)
```

	product_id	mus_avg_rat	dmp_avg_rat
1	B0019M1ZJS	3.0	5.0

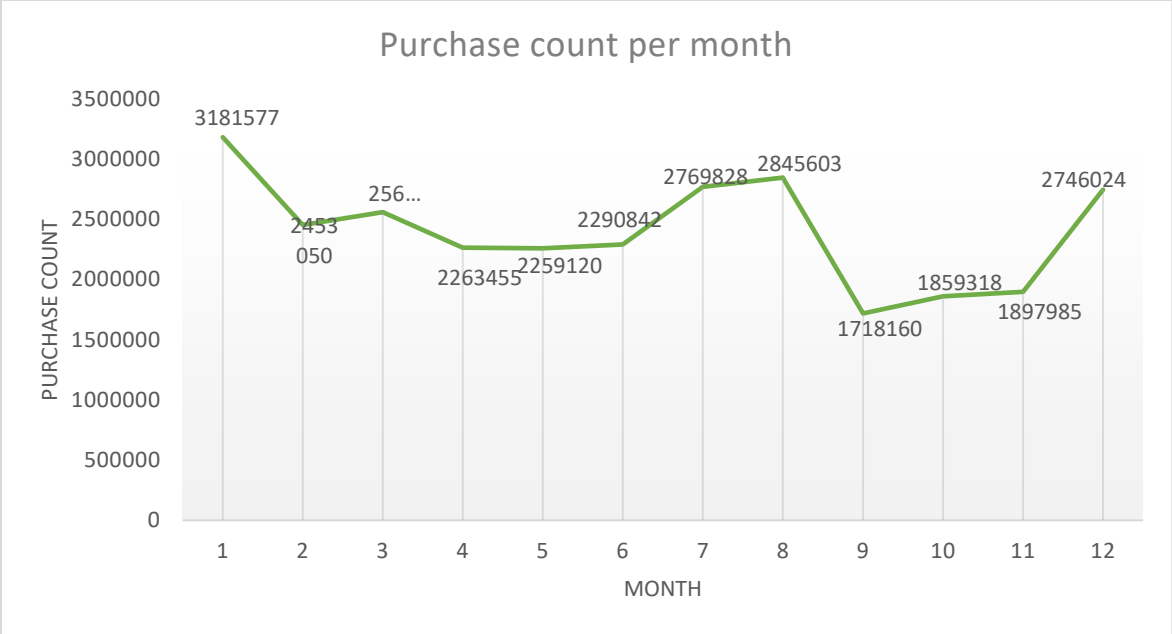
Interpretation:-

The above table shows that they don't get similar ratings.

4. You should cover additional questions and not limit yourself to the above questions

Query1:-

```
select month(review_date) as month,count(*)  from final
group by month(review_date)
order by month(review_date)
;
```

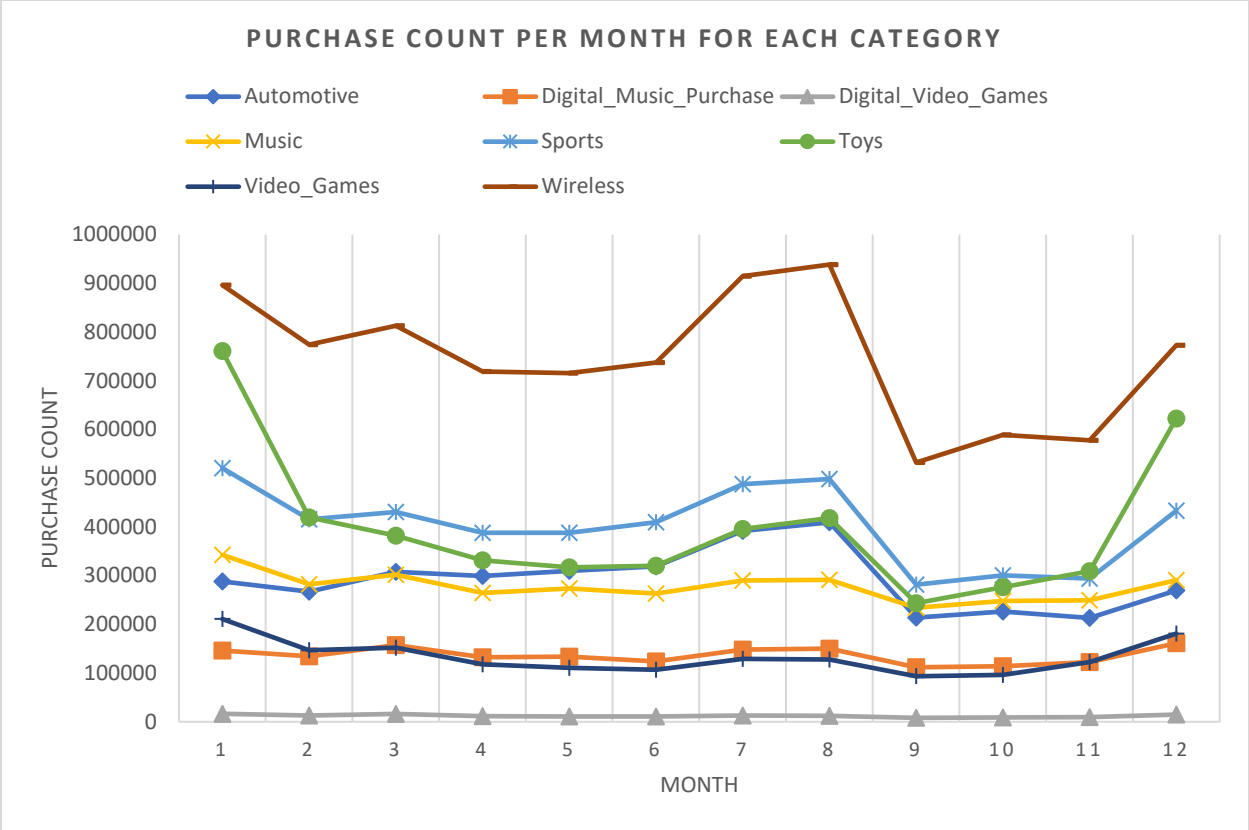


Interpretation:-

We can see that January month sees the highest purchase(3181577), then follows August July, December.

Query2:-

```
select product_category,month(review_date) as month,count(*)  from final
group by product_category,month(review_date)
order by product_category,month(review_date)
;
```



Interpretation:-

Here I have grouped the dataset based on Product category and month, to find the purchase count for each category per month. We can see that in August month wireless category saw maximum purchase whereas other categories saw in January and December.

Query3:-

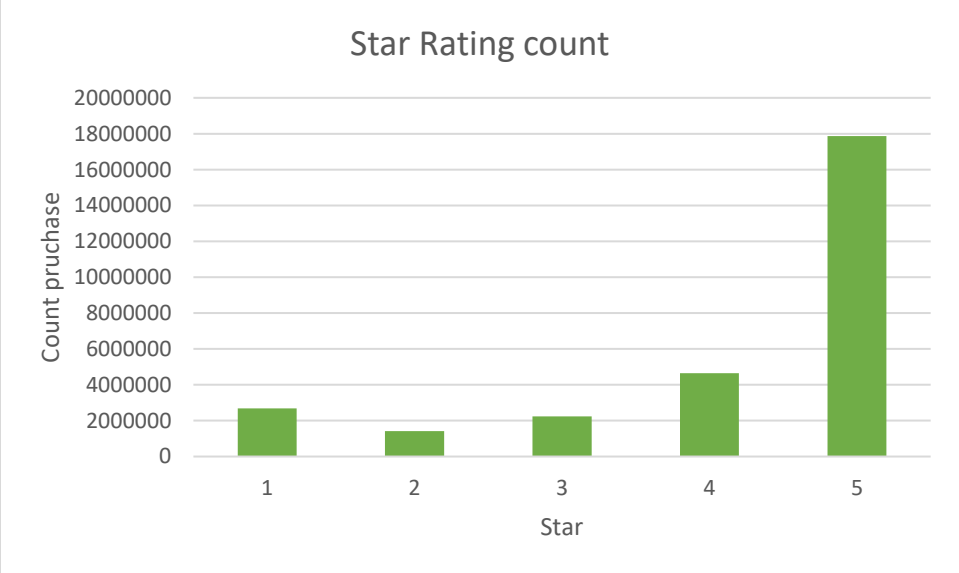
```
select star_rating, count(*) as cnt
from final
group by star_rating
order by star_rating;
```

Output:-

```
hive> select star_rating, count(*) as cnt
    > from final
    > group by star_rating
    > order by star_rating;
Query ID = hadoop_20200413100507_cb321f40-3d01-40c8-91ec-8379abe67f6c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586765536088_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     54         54        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 72.85 s
--------------------------------------------------------------------------------
OK
1       2676666
2       1412985
3       2242062
4       4654896
5       17858916
Time taken: 73.381 seconds, Fetched: 5 row(s)
hive>
```

	star_rating	cnt
1	1	2676666
2	2	1412985
3	3	2242062
4	4	4654896
5	5	17858916

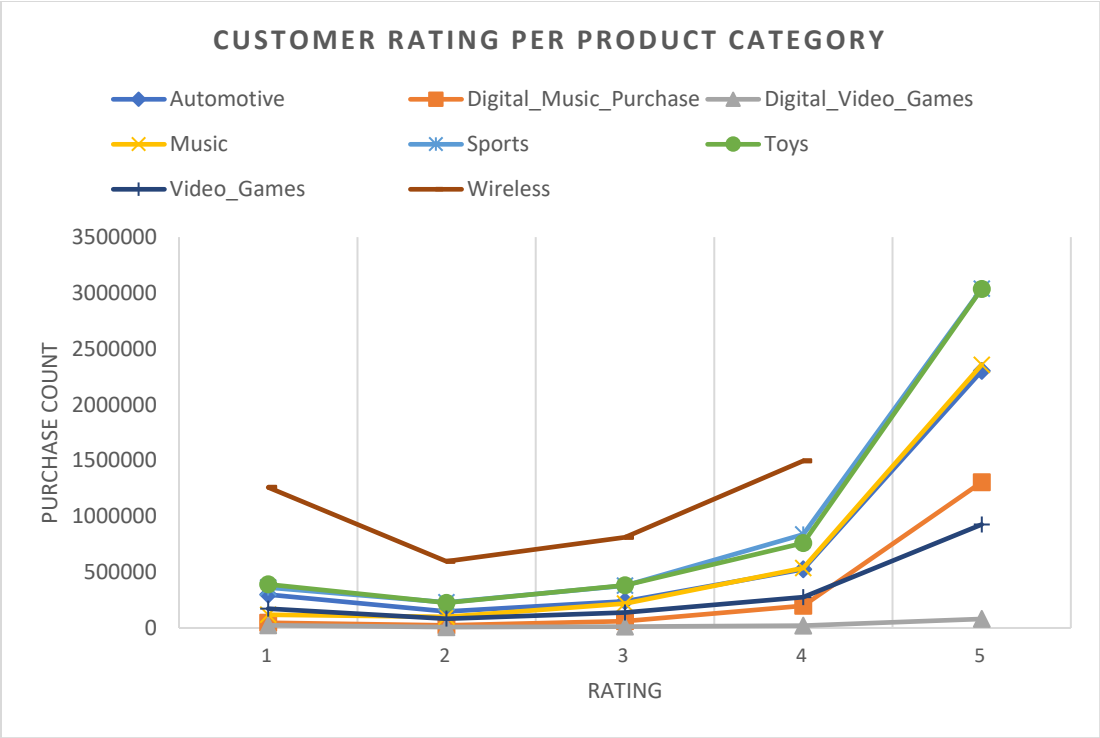


Interpretation:-

In the above visualization we can say that we have most data we star rating 5. That shows most of the customers are highly satisfied with the products.

Query4:-

```
select product_category,star_rating as month,count(*)  from final
group by product_category,star_rating
order by product_category,star_rating
;
```



Interpretation:-

We can see that for wireless category, though purchase count is highest, customers are not highly satisfied with the products as their highest rating is 4, even after that rating 1 got maximum count.

Whereas for other product categories there is maximum purchase count for star rating 5 than other ratings, which shows that customers buying products other than wireless category are more *satisfied*.

Conclusion:-

1) We can see that more that 31% of the reviews are of wireless products, followed by Sports(16.8%) and Toys(16.63%), whereas Video games and digital video category constitutes only 6%.
2) 97% of the reviews are from just US.
3) There is a positive correlation between number of customers and years and this can be seen for all the product categories.
4) Wireless product category has noticed highest increase in reviews over the years compared to any other product category.
5) The number of customers have rapidly increased in the US than other countries.
6) There is high positive correlation between music category/ digital music category and years.
7) There is high positive correlation between video games/digital video games and years.
8) For all the product categories January and December are highest sale months whereas only wireless product sees August as the maximum sale month
9) Most of the customers are highly satisfied with the purchased products since most of the reviews are with 5 star rating.
10) We can see that for wireless category, though purchase count is highest, customers are not highly satisfied with the products.
11) Whereas for other product categories there is maximum purchase count for star rating 5 than other ratings, which shows that customers buying products other than wireless category are more *satisfied*.

References:-

https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

https://www.ibm.com/analytics/hadoop/big-data-analytics

https://aws.amazon.com/emr/faqs/