

Project Part 1

SHRESHTHA JHA

Data Cleaning

In Amazon Review dataset there are many records with multiple reviews by the same customer for the same product. This may cause ambiguity in analysis and accuracy may be less. Hence it is better to filter such records from the dataset. Also previous years data may be less effective due to changing trends. So for analysis, I have included the data after 2005 and the data with product categories as Wireless, Automotive, Music, Digital_Music_Purchase, Sports, Toys, Digital_Video_Games, Video_Games have been included in my analysis.

Sql queries to create hive tables.

```
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;

Creating temp view to filter data including only required product categories.

```
create view temp
as
select * from amazon_review.amazon_reviews_parquet where review_id in (select review_id from
(select customer_id, product_id,review_id,count(*)
from amazon_review.amazon_reviews_parquet
group by customer_id, product_id,review_id
having count(*)=1) as t) and product_category in
('Wireless','Automotive','Music','Digital_Music_Purchase','Sports','Toys','Digital_Video_Games','Video_G
ames');
```

Creating table to filter reviews that are reviewed multiple times by same customer for same product.

```
create table amazon_review.filtered_reviews
AS
select t.* from(
select *,row_number() over(partition by customer_id,product_id) as row1 from temp)t where row1=1;
```

Exploratory Analysis.

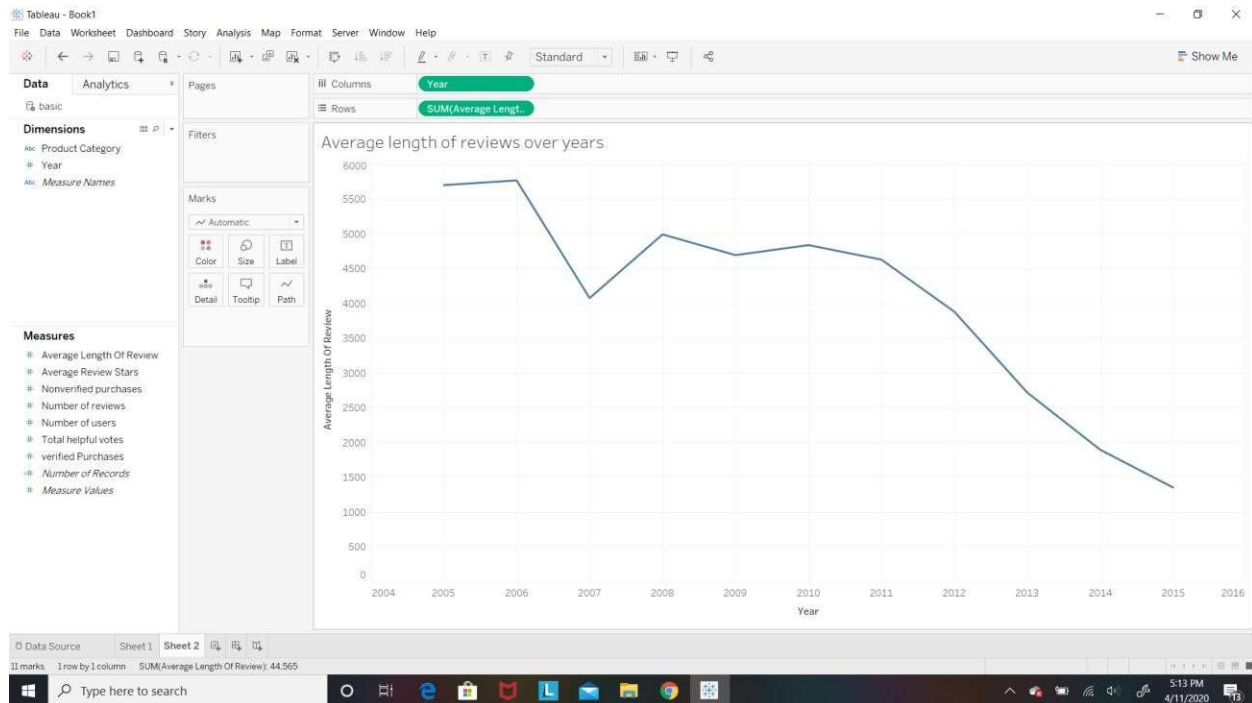
1. Explore the dataset and provide basic exploratory analysis over time and per product category

Query-

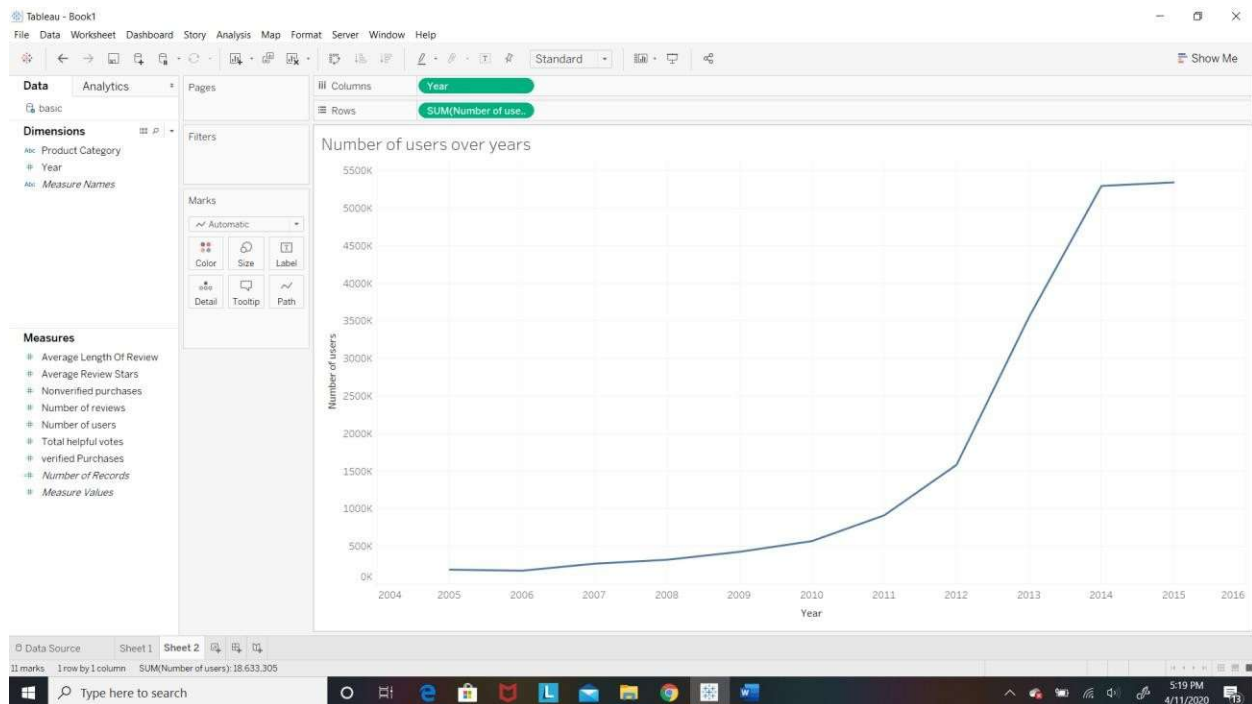
Select year,product_category,count(review_id) as Number_of_reviews,count(Distinct(customer_id)) as Number_of_users,avg(star_rating) as average_review_stars,avg(length(review_body)) as average_length_of_review, sum(case when verified_purchase='Y' then 1 else 0 end) as verified_Purchases, sum(case when verified_purchase='N' then 1 else 0 end) as Nonverified_purchases,sum(helpful_votes) as Total_helpful_votes from amazon_review.filtered_reviews where year>=2005 group by year,product_category order by year;

Year	Product Category	Number of Reviews	Number of Users	Average Review Stars	Average Length of Review	Verified Purchases	Nonverified Purchases	Total Helpful Votes
2005	Digital_Music_Purchase	8	7	4.625	881.25	2	6	3
2005	Automotive	660	593	3.693939393939394	588.1969696969697	109	551	7844
2005	Sports	4514	3975	3.6896322552060257	588.4319893664156	658	3864	47124
2005	Music	255091	129479	4.271718719986201	928.5174427049241	20053	235038	1241364
2005	Wireless	11835	10585	3.414826193493874	903.4652302492607	2139	9696	119942
2005	Toys	36104	26902	3.822762020828717	573.549135829825	2663	33441	262463
2005	Video Games	27101	17557	3.765949595959596	1238.4387453874538	1211	25890	178568
2006	Digital_Video_Games	1	1	4.0	281.0	0	1	4
2006	Sports	9529	8352	3.8969461643484344	553.166019519362	2187	7422	96096
2006	Wireless	19855	17982	3.5094434651221356	817.3586502140519	4977	14878	154452
2006	Music	204483	109019	4.317216479210188	927.6133520545197	25604	178799	971330
2006	Digital_Music_Purchase	21	20	3.857142857142857	899.6666666666666	5	16	19
2006	Automotive	2190	1986	3.752054794520548	493.67625570776255	698	1492	25684
2006	Toys	26849	21070	3.8100115460530566	568.5423665605073	3133	23716	163919
2006	Video Games	24782	16902	3.751882438668934	1228.654926726581	1934	22768	173021
2007	Sports	29543	20146	4.067494838032698	433.07358765189724	10139	19404	189130
2007	Toys	48107	37767	4.076828735942795	451.89265595443493	14203	33904	258250
2007	Video Games	43495	31072	3.9553281986435223	832.4826301873778	7666	35829	211109
2007	Music	223640	121546	4.384594483655777	820.0850227591819	58886	172760	804875
2007	Digital_Music_Purchase	2235	1913	4.405360127516779	554.7131991051455	466	1769	4219
2007	Wireless	47736	42099	3.7610608345902463	587.6265501927267	18054	29682	197763
2007	Automotive	8885	7833	4.008787844682849	398.7304445694992	3862	5023	60131
2008	Digital_Video_Games	5	5	2.0	887.2	0	5	40
2008	Music	193910	107846	4.373472229384766	838.8939714300449	50768	143142	612573
2008	Wireless	63655	55672	3.769696017594847	586.3092608953190	27786	35869	203106
2008	Toys	65294	49910	4.078720862560113	457.32777896897113	22790	42504	279957
2008	Video Games	60166	43360	3.790679121098295	848.9850413855002	11508	48658	306388
2008	Digital_Music_Purchase	22041	16114	4.4610408616215235	582.3237148949685	5068	16973	37953
2008	Automotive	13851	11984	3.9724929607970543	417.88716193776624	6924	6927	88545
2008	Sports	40923	35753	4.0453208325733695	454.6303381321095	10014	24909	235646
2008	Music	60361	52104	4.012905684133795	482.38458323884626	32670	27691	383667
2009	Digital_Music_Purchase	36181	26817	4.461264199441696	524.1902656090213	10633	25548	66114
2009	Toys	93156	70907	4.009081540641505	439.0045729743656	46167	46989	374284
2009	Video Games	73636	53556	3.920473681351513	832.7485469064045	25601	48035	283002
2009	Wireless	93972	79971	3.722638145149619	572.164517890197	54238	39734	294607
2009	Digital_Video_Games	1561	1145	3.89237668161435	620.6016143497750	716	845	7782
2009	Music	204796	121448	4.390780990253716	786.7226996621027	74681	130115	585788
2009	Automotive	23951	20175	3.9797928754874534	438.2716796793453	15002	8949	174620
2010	Digital_Video_Games	2551	1828	3.733829870638965	683.3833790670326	1738	813	10842
2010	Automotive	51059	42110	4.004955051998068	428.0769501948726	43266	7793	229781
2010	Digital_Music_Purchase	40809	30127	4.49207282707246	543.2398980617021	15373	25436	73507

Visualizations



From this visualization we can infer that average length of reviews has decreased over the years.



We can see that Number of users have increased exponentially over the years.

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

i Do you see correlation (maybe negative) between the categories over time?

Correlation for Music/Digital_Music_Purchase

Query-

```
select corr(Music,Digital_Music_Purchase) from (
Select year,sum(case when product_category='Music' then 1 else 0 end) as Music,
sum(case when product_category='Digital_Music_Purchase' then 1 else 0 end) as
Digital_Music_Purchase from amazon_review.filtered_reviews where year>=2005 group by year order
by year)r;
```

```
hive> select corr(Music,Digital_Music_Purchase) from (
> Select year,sum(case when product_category='Music' then 1 else 0 end) as Music,
> sum(case when product_category='Digital_Music_Purchase' then 1 else 0 end) as Digital_Music_Purchase from amazon_review.filtered_reviews where year>=200
5 group by year order by year)r;
Query ID = hadoop_20200411032529_4ef5b5fc-deff-4f60-8c4d-499d977ccdd
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_1586545402644_0018)

-----
VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED   20      20           0         0         0         0
Reducer 2 ..... container  SUCCEEDED   20      20           0         0         0         0
Reducer 3 ..... container  SUCCEEDED    1        1           0         0         0         0
-----
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 49.04 s
-----
OK
0.9558821088253973
Time taken: 55.631 seconds, Fetched: 1 row(s)
hive>
```

Correlation between Digital_Video_Games/Video_Games based on count of reviews.

Query-

```
select corr(Video_Games,Digital_Video_Games) from (
Select year,sum(case when product_category='Video_Games' then 1 else 0 end) as Video_Games,
sum(case when product_category='Digital_Video_Games' then 1 else 0 end) as Digital_Video_Games
from amazon_review.filtered_reviews where year>=2005 group by year order by year)r;
```

Output-

```

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_1586545402644_0018)

-----
VERTICES    MODE    STATUS    TOTAL    COMPLETED    RUNNING    PENDING    FAILED    KILLED
-----
Map 1 ..... container    SUCCEEDED    29        29          0          0          0          0
Reducer 2 ..... container    SUCCEEDED    20        20          0          0          0          0
Reducer 3 ..... container    SUCCEEDED     1         1          0          0          0          0
-----
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 49.04 s
-----
OK
0.9558821088253973
Time taken: 55.631 seconds, Fetched: 1 row(s)
hive> select corr(Video_Games,Digital_Video_Games) from (
> select year,sum(case when product_category='Video_Games' then 1 else 0 end) as Video_Games,
> sum(case when product_category='Digital_Video_Games' then 1 else 0 end) as Digital_Video_Games from amazon_review.filtered_reviews where year>=2005 group
p by year order by year);
Query ID = hadoop_20200411042805_99001de0-1273-42b9-a9ad-cf5578f663da
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_1586545402644_0019)

-----
VERTICES    MODE    STATUS    TOTAL    COMPLETED    RUNNING    PENDING    FAILED    KILLED
-----
Map 1 ..... container    SUCCEEDED    29        29          0          0          0          0
Reducer 2 ..... container    SUCCEEDED    20        20          0          0          0          0
Reducer 3 ..... container    SUCCEEDED     1         1          0          0          0          0
-----
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 48.49 s
-----
OK
0.9662947732101295
Time taken: 55.12 seconds, Fetched: 1 row(s)
hive>

```

ii Are there same users reviewing in both categories?

Finding count of users that have given reviews in Music and Digital_Music_Purchase category for each year.

Query-

```

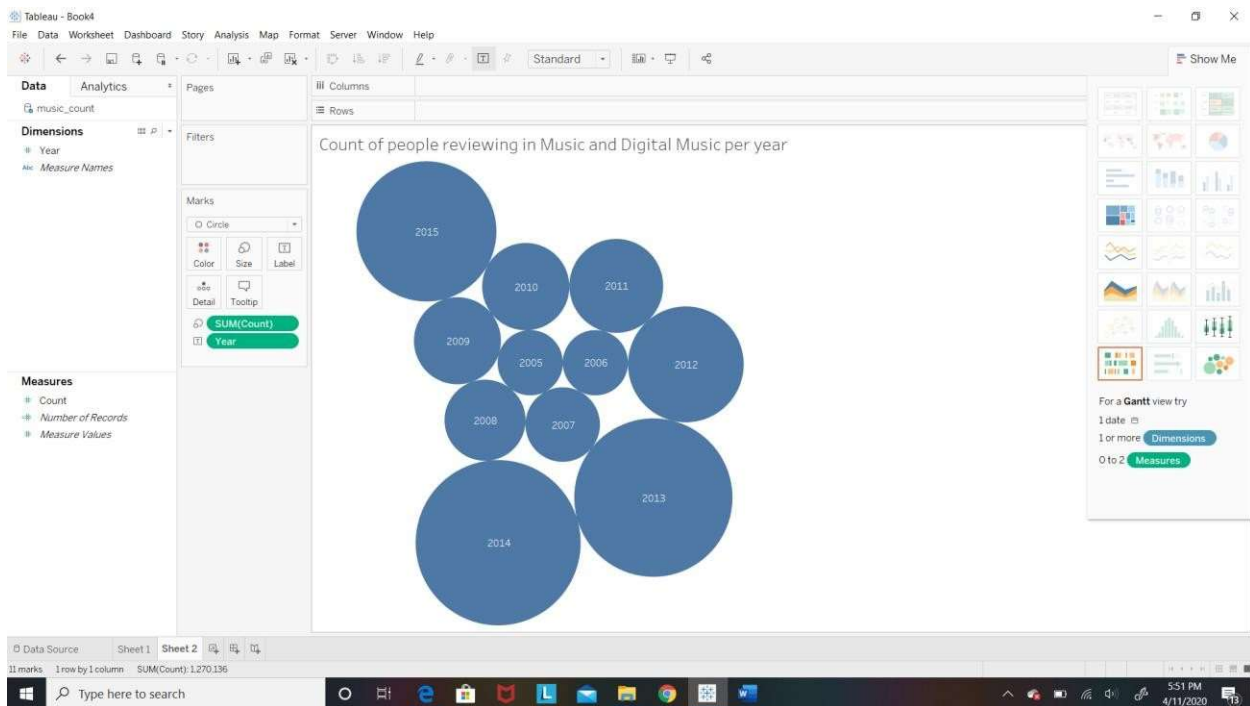
select count(r.customer_id) as count,r.year from amazon_review.filtered_reviews r,
(select distinct(customer_id) from amazon_review.filtered_reviews where product_category='Music'
and year>=2005
intersect
select distinct(customer_id) from amazon_review.filtered_reviews where
product_category='Digital_Music_Purchase' and year>=2005)u where r.customer_id=u.customer_id and
r.year>=2005 and r.product_category in ('Music','Digital_Music_Purchase') group by year order by year;

```

```
ec2-18-234-166-153.compute-1.amazonaws.com (hadoop)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split Multiterm Tunneling Packages Settings Help
Quick connect...
/home/hadoop
Name
aws
ssh
bash_profile
bashrc
Remote monitoring
Follow terminal folder
filtered_reviews
Time taken: 0.017 seconds, Fetched: 2 row(s)
hive> select count(r.customer_id) as count,r.year from amazon_review.filtered_reviews r,
> (select distinct(customer_id) from amazon_review.filtered_reviews where product_category='Music' and year=>2005
> intersect
> select distinct(customer_id) from amazon_review.filtered_reviews where product_category='Digital_Music_Purchase' and year=>2005)u where r.customer_id=u
customer_id and r.year=>2005 and r.product_category in ('Music','Digital_Music_Purchase') group by year order by year;
Query ID = hadoop_20200405224342_b3ff6cd6-6c8a-47d7-b40b-3085c927c6ee
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_1506115072005_0016)

-----
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 8 ..... container    SUCCEEDED    17        17            0            0            0            0
Reducer 2 ..... container    SUCCEEDED    11        11            0            0            0            0
Reducer 3 ..... container    SUCCEEDED    1         1            0            0            0            0
Reducer 5 ..... container    SUCCEEDED    10        10            0            0            0            0
Reducer 7 ..... container    SUCCEEDED    5         5            0            0            0            0
Reducer 9 ..... container    SUCCEEDED    10        10            0            0            0            0
-----
VERTICES: 08/08 [=====] 100% ELAPSED TIME: 458.44 s
OK
41393 2005
41640 2006
54272 2007
64240 2008
74073 2009
74400 2010
86157 2011
130992 2012
244828 2013
266390 2014
191743 2015
Time taken: 468.408 seconds, Fetched: 11 row(s)
hive>
```

Visualization



From this we can infer that 2014 was the year when there were maximum people who had reviewed in both these categories.

Total count of people that have reviewed in both these categories(Music/Digital_Music_Purchase)

Query-

```
select count(r.customer_id) as count from amazon_review.filtered_reviews r,  
(select distinct(customer_id) from amazon_review.filtered_reviews where product_category='Music'  
and year>=2005  
intersect  
select distinct(customer_id) from amazon_review.filtered_reviews where  
product_category='Digital_Music_Purchase' and year>=2005)u where r.customer_id=u.customer_id and  
r.year>=2005 and r.product_category in ('Music','Digital_Music_Purchase');
```

```
Time taken: 468.408 seconds, Fetched: 11 row(s)
hive> select count(r.customer_id) as count from amazon_review.filtered_reviews r,  
> (select distinct(customer_id) from amazon_review.filtered_reviews where product_category='Music' and year>=2005  
> intersect  
> select distinct(customer_id) from amazon_review.filtered_reviews where product_category='Digital_Music_Purchase' and year>=2005)u where r.customer_id=u  
,customer_id and r.year>=2005 and r.product_category in ('Music','Digital_Music_Purchase');  
Query ID = hadoop_20200405225958_8487b00f-f83e-4a7a-ad5e-34887ccb5c47  
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_1586115072005_0017)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	17	17	0	0	0	0
Map 3	container	SUCCEEDED	17	17	0	0	0	0
Map 7	container	SUCCEEDED	17	17	0	0	0	0
Reducer 2	container	SUCCEEDED	1	1	0	0	0	0
Reducer 4	container	SUCCEEDED	10	10	0	0	0	0
Reducer 6	container	SUCCEEDED	5	5	0	0	0	0
Reducer 8	container	SUCCEEDED	10	10	0	0	0	0

```
VERTICES: 07/07 [=====] 100% ELAPSED TIME: 449.51 s  
OK  
12/7/136  
Time taken: 460.76 seconds, Fetched: 1 row(s)
```

Finding count of users that have given reviews for Digital_Video_Games and Video_Games category over years.

Query-

```
select count(r.customer_id) as count,r.year from amazon_review.filtered_reviews r,  
(select distinct(customer_id) from amazon_review.filtered_reviews where  
product_category='Digital_Video_Games' and year>=2005  
intersect  
select distinct(customer_id) from amazon_review.filtered_reviews where  
product_category='Video_Games' and year>=2005)u where r.customer_id=u.customer_id and  
r.year>=2005 and r.product_category in ('Video_Games','Digital_Video_Games') group by year order by  
year;
```



```

Time taken: 382.521 seconds, Fetched: 4 row(s)
hive> select count(r.customer_id) as count, r.year from amazon_review.filtered_reviews r,
> (select distinct(customer_id) from amazon_review.filtered_reviews where product_category='Digital_Video_Games' and year>=2005
> intersect
> select distinct(customer_id) from amazon_review.filtered_reviews where product_category='Video_Games' and year>=2005)u where r.customer_id=u.customer_id and r.year>=2005 and r.product_category in ('Video_Games', 'Digital_Video_Games');
Query ID = hadoop_20200406024225_05275c35-7753-42d4-9b9f-106c13c24d0d
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_1586115072805_0023)

-----
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 8 ..... container  SUCCEEDED   17      17          0         0         0         0
Reducer 2 ..... container  SUCCEEDED   11      11          0         0         0         0
Reducer 3 ..... container  SUCCEEDED    1         1          0         0         0         0
Reducer 5 ..... container  SUCCEEDED   10      10          0         0         0         0
Reducer 7 ..... container  SUCCEEDED    5         5          0         0         0         0
Reducer 9 ..... container  SUCCEEDED   10      10          0         0         0         0
-----
VERTICES: 08/08 [=====] 100% ELAPSED TIME: 446.12 s
-----
OK
1082    2005
1484    2006
2158    2007
3883    2008
5842    2009
8063    2010
13001   2011
18912   2012
38267   2013
44350   2014
29991   2015
Time taken: 455.11 seconds, Fetched: 11 row(s)

```

Total Count of people that have reviewed in both these categories.(Digital_Video_Games and Video_Games category)

Query-

```

select count(r.customer_id) as count from amazon_review.filtered_reviews r,
(select distinct(customer_id) from amazon_review.filtered_reviews where
product_category='Digital_Video_Games' and year>=2005
intersect
select distinct(customer_id) from amazon_review.filtered_reviews where
product_category='Video_Games' and year>=2005)u where r.customer_id=u.customer_id and
r.year>=2005 and r.product_category in ('Video_Games', 'Digital_Video_Games');

```

```

Time taken: 455.11 seconds, Fetched: 11 row(s)
hive> select count(r.customer_id) as count from amazon_review.filtered_reviews r,
> (select distinct(customer_id) from amazon_review.filtered_reviews where product_category='Digital_Video_Games' and year>=2005
> intersect
> select distinct(customer_id) from amazon_review.filtered_reviews where product_category='Video_Games' and year>=2005)u where r.customer_id=u.customer_id and r.year>=2005 and r.product_category in ('Video_Games', 'Digital_Video_Games');
Query ID = hadoop_20200406025405_c4532571-e93e-437a-962d-860314b39d0c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586115072805_0023)

-----
VERTICES      MODE        STATUS      TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED   17      17          0         0         0         0
Map 3 ..... container  SUCCEEDED   17      17          0         0         0         0
Map 7 ..... container  SUCCEEDED   17      17          0         0         0         0
Reducer 2 ..... container  SUCCEEDED    1         1          0         0         0         0
Reducer 4 ..... container  SUCCEEDED   10      10          0         0         0         0
Reducer 6 ..... container  SUCCEEDED    5         5          0         0         0         0
Reducer 8 ..... container  SUCCEEDED   10      10          0         0         0         0
-----
VERTICES: 07/07 [=====] 100% ELAPSED TIME: 439.06 s
-----
OK
167033
Time taken: 439.834 seconds, Fetched: 1 row(s)
hive>

```

Want to learn more about the professional edition? <https://facebook.com/stacks24x7>

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

For Music/Digital_Music_Purchase

Creating view with product_category music.

create view music as

```
select product_id,round(avg(star_rating),2) as Average_rating_by_customer_for_Music_products from
filtered_reviews where product_category='Music' and year>=2005 group by product_id;
```

Creating view with product_category digital_Music_Purchase .

create view Digital_Music_Purchase as

```
select product_id,round(avg(star_rating),2) as
Average_rating_by_customer_for_Digital_Music_purchase_products from filtered_reviews where
product_category='Digital_Music_Purchase' and year>=2005 group by product_id;
```

To find similar items in both categories and their average rating in respective categories using inner join on product ids from both views generated above.

Query-

select

```
r.product_id,Average_rating_by_customer_for_Music_products,Average_rating_by_customer_for_Digital_Music_purchase_products from music r inner join Digital_Music_Purchase u on
u.product_id=r.product_id;
```

```
Time taken: 0.131 seconds
hive> select r.product_id,Average_rating_by_customer_for_Music_products,Average_rating_by_customer_for_Digital_Music_purchase_products from music r inner join
Digital_Music_Purchase u on u.product_id=r.product_id;
Query ID = hadoop_20200406014302_36f1ea14-9f4f-4d83-889a-0024fd19c98a
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_1586115072805_0021)

-----
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
Reducer 2 ..... container  SUCCEEDED   10      10           0         0         0         0
Reducer 3 ..... container  SUCCEEDED   10      10           0         0         0         0
Reducer 5 ..... container  SUCCEEDED   10      10           0         0         0         0
-----
VERTICES: 05/05  [=====] 100%  ELAPSED TIME: 308.69 s
-----
OK
B0019M1ZJS      3.0      5.0
Time taken: 317.15 seconds, Fetched: 1 row(s)
hive>
```

We can see that this product is common in both product categories but it does not get same average ratings.

For Digital_Video_Games/Video_Games category

Creating view for Digital Video_games category.

```
create view Digital_Video_Games as
select product_id,round(avg(star_rating),2) as Average_rating_by_customer_for_Digital_Video_Games
from filtered_reviews where product_category='Digital_Video_Games' and year>=2005 group by
product_id;
```

Creating view for Video_games Category

```
create view Video_Games as
select product_id,round(avg(star_rating),2) as Average_rating_by_customer_for_Video_Games from
filtered_reviews where product_category='Video_Games' and year>=2005 group by product_id;
```

To find similar items in both categories and their average rating in respective categories using inner join on product ids from both views generated above.

Query-

```
select
r.product_id,Average_rating_by_customer_for_Digital_Video_Games,Average_rating_by_customer_for
_Video_Games from Video_Games r inner join Digital_Video_Games u on u.product_id=r.product_id;
```

```
Time taken: 0.197 seconds
hive> select r.product_id,Average_rating_by_customer_for_Digital_Video_Games,Average_rating_by_customer_for_Video_Games from Video_Games r inner join Digital_Video_Games u on u.product_id=r.product_id;
Query ID = hadoop_20200406022255_143d4705-5b0e-4e92-afee-f81f3f889ddd
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_1586115072805_0022)

-----
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
Reducer 2 ... container  SUCCEEDED  10      10          0         0         0         0
Reducer 3 ... container  SUCCEEDED  10      10          0         0         0         0
Reducer 5 ... container  SUCCEEDED  10      10          0         0         0         0
-----
VERTICES: 05/05  [=====] 100% ELAPSED TIME: 294.15 s
-----
OK
B0047T7MEW      3.69    4.5
B00NB8ME0Y      5.0     3.47
B004YNII9Y      2.82    4.0
B00B4WVTUS      4.64    5.0
Time taken: 302.521 seconds, Fetched: 4 row(s)
hive>
```

ort MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

We can see that these products are common in both categories and they do not get same average ratings.

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

List of customers who have given reviews for products in both Digital_Video_Games and Video_Games category and their ratings in both categories.

Query-

create view video_games as

```
select customer_id,product_category,round(avg(star_rating),2) as  
Average_rating_by_customer_for_video_games_products from  
amazon_review.filtered_reviews where product_category='Video_Games' and year>=2005  
group by customer_id,product_category;
```

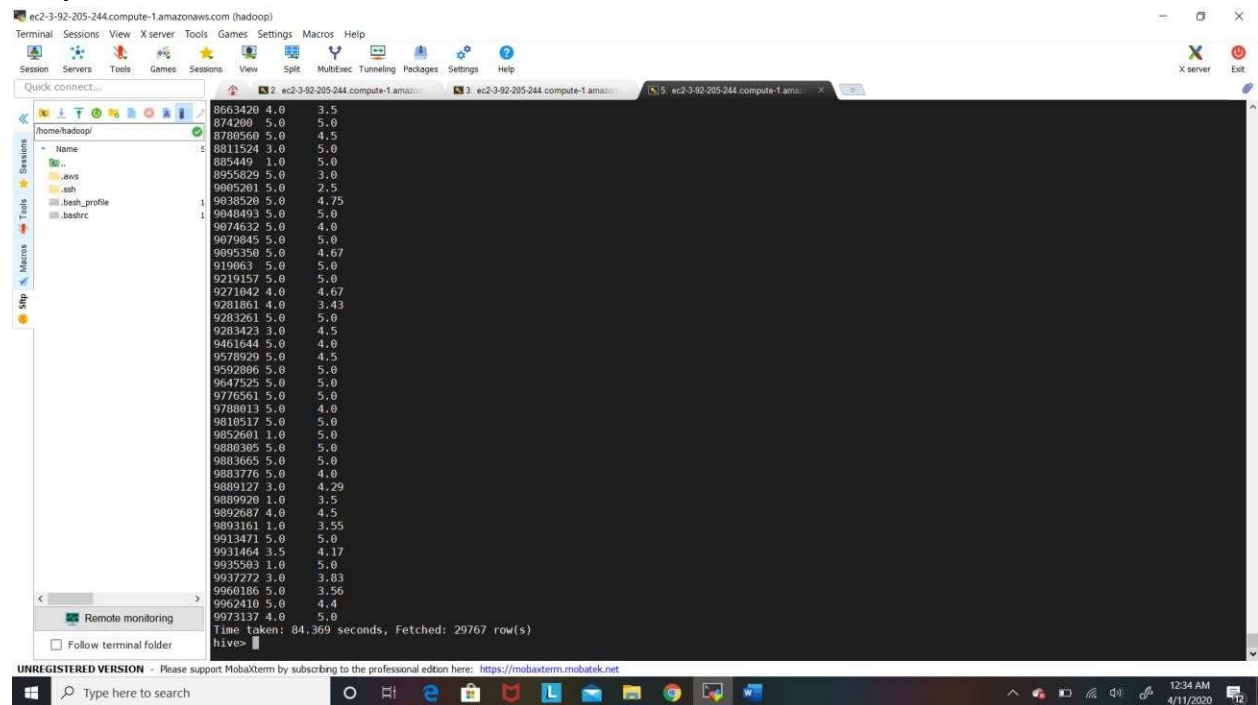
create view Digital_Video_Games as

```
select customer_id,product_category,round(avg(star_rating),2) as  
Average_rating_by_customer_for_Digital_video_games_products from  
amazon_review.filtered_reviews where product_category='Digital_Video_Games' and  
year>=2005 group by customer_id,product_category;
```

select

```
r.customer_id,Average_rating_by_customer_for_Digital_video_games_products,Average_ratin  
g_by_customer_for_video_games_products from Digital_Video_Games r inner join  
video_games u on u.customer_id=r.customer_id;
```

Output-



8663420	4.0	3.5
874209	5.0	5.0
8780560	5.0	4.5
8811524	3.0	5.0
885449	1.0	5.0
8955829	5.0	3.0
9095281	5.0	2.5
9038520	5.0	4.75
9048493	5.0	5.0
9074632	5.0	4.0
9079045	5.0	5.0
9095350	5.0	4.67
919063	5.0	5.0
9219157	5.0	5.0
9271042	4.0	4.67
9281861	4.0	3.43
9283261	5.0	5.0
9283423	3.0	4.5
9461644	5.0	4.0
9578929	5.0	4.5
9592806	5.0	5.0
9647525	5.0	5.0
9776561	5.0	5.0
9788013	5.0	4.0
9810517	5.0	5.0
9852601	1.0	5.0
9880385	5.0	5.0
9883665	5.0	5.0
9883776	5.0	4.0
9889127	3.0	4.29
9889920	1.0	3.5
9892687	4.0	4.5
9893161	1.0	3.55
9913471	5.0	5.0
9931464	3.5	4.17
9935503	1.0	5.0
9937272	3.0	3.83
9960186	5.0	3.56
9962410	5.0	4.4
9973137	4.0	5.0

Time taken: 84.369 seconds, Fetched: 29767 row(s)
hive>

3. You should demonstrate your ability to use Hive advanced functions:

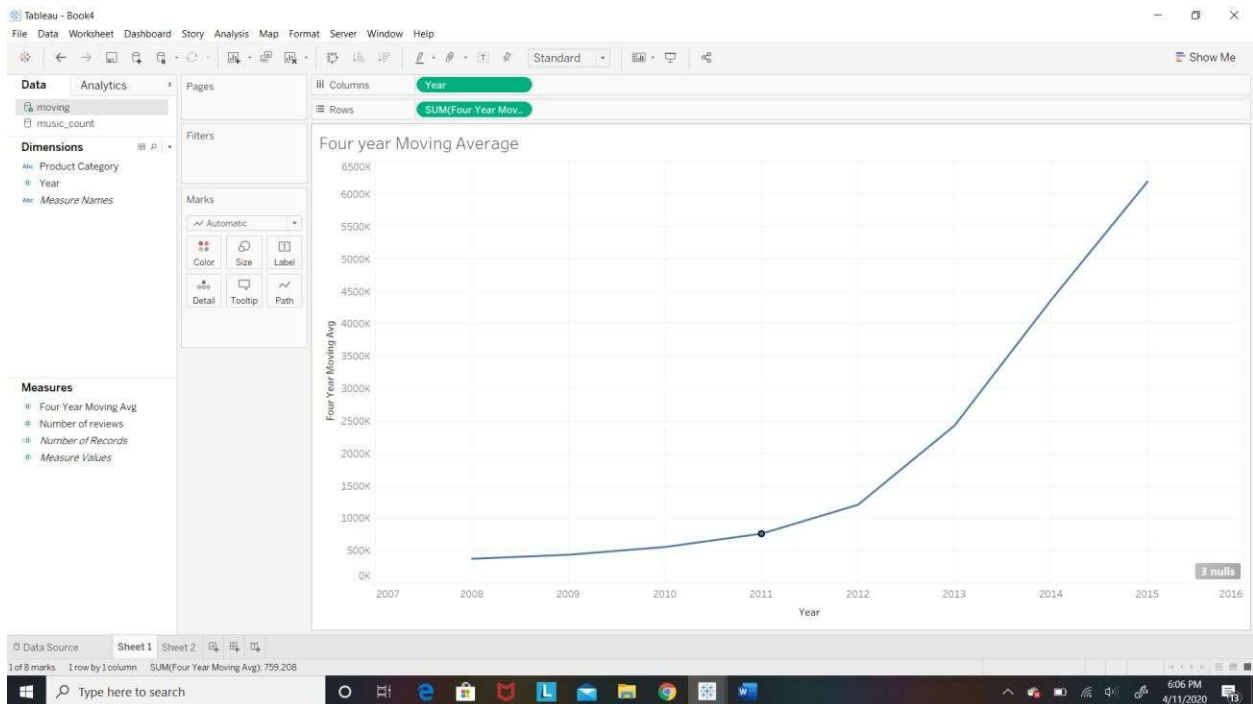
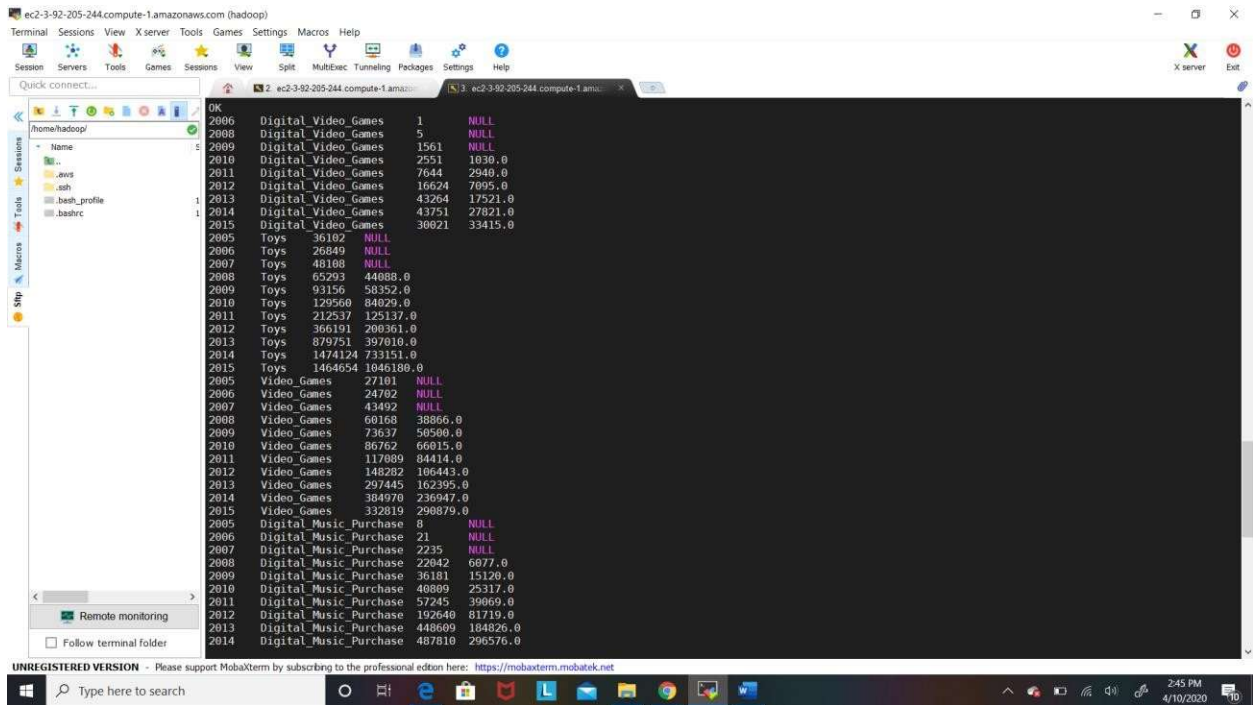
i. Window functions: moving average, rank, aggregation functions using relevant ordering and partitioning

Calculating four year Moving average(current year and previous three years) based on number of reviews per product category over the years.

Query-

```
select year,product_category,Number_of_reviews,(case when row_number() over (Partition by
product_category order by year) > 3
then round(AVG(Number_of_reviews) OVER (PARTITION BY product_category order by year ROWS 3
PRECEDING))
end) as four_year_moving_avg from
(Select year,product_category,count(review_id) as Number_of_reviews,count(Distinct(customer_id)) as
Number_of_users,avg(star_rating) as average_review_stars,avg(length(review_body)) as
average_length_of_review
from amazon_review.filtered_reviews group by year,product_category order by product_category,year)
as x where year>=2005;
```

Output-



We can infer that Number of reviews are increasing over the years.

Ranking Top five products in each product categories based on number of reviews per product.

Query-

```
select product_id,product_category,Number_of_reviews,rank from(
select product_id,product_category,Number_of_reviews,rank() over (Partition by product_category
order by Number_of_reviews desc) as rank from
(Select product_id,product_category,count(review_id) as
Number_of_reviews,count(Distinct(customer_id)) as Number_of_users,avg(star_rating) as
average_review_stars,avg(length(review_body)) as average_length_of_review
from amazon_review.filtered_reviews group by product_category,product_id)as x)as z where rank<=5;
```

Output-

Product ID	Product Category	Number of Reviews	Number of Users	Average Review Stars
B002V0WIP6	Digital_Video_Games	5125	1	
B004R9K4BC	Digital_Video_Games	5073	2	
B00GAC1D2G	Digital_Video_Games	3666	3	
B004R9K4PB	Digital_Video_Games	3589	4	
B004R9K4QG	Digital_Video_Games	3582	5	
B0008B0WZG	Music	3669	1	
B0008AGMEC	Music	2738	2	
B0003GZMIE	Music	2255	3	
B004AB2JVI	Music	2088	4	
B00NP211Z5	Music	2059	5	
B009A5204K	Wireless	18252	1	
B0073FE1F0	Wireless	9945	2	
B007F8X9DK	Wireless	9471	3	
B0042FV2SI	Wireless	8839	4	
B009SVZ80C	Wireless	8780	5	
B004S8F7QM	Toys	24277	1	
B005JFNE8G	Toys	6659	2	
B008JNP0VK	Toys	3963	3	
B0053X62GK	Toys	3951	4	
B49900606	Toys	3647	5	
B00BG99WK2	Video_Games	18353	1	
B007FTE2VW	Video_Games	3972	2	
B0017863BA	Video_Games	3721	3	
B00505Y1LE	Video_Games	3564	4	
B005CPGHAA	Video_Games	3482	5	
B001B8H8HE	Sports	7378	1	
7245456313	Sports	3693	2	
B00F8X54DC	Sports	3851	3	
B000UVXZ8	Sports	3006	4	
B002021RS6	Sports	2922	5	
B00PB8S0S0	Digital_Music_Purchase	980	1	
B00L6GKD36	Digital_Music_Purchase	754	2	
B002HP8EKE	Digital_Music_Purchase	749	3	
B001L185C	Digital_Music_Purchase	734	4	
B001KS28HY	Digital_Music_Purchase	721	5	
B005NL0AHS	Automotive	4894	1	
B000CTTKBS	Automotive	4422	2	
B001LHV0VK	Automotive	3694	3	
B00068XCQU	Automotive	2688	4	
B001A725HY	Automotive	2483	5	

Time taken: 183.788 seconds, Fetched: 40 row(s)

Ranking top five products in each category based on Average star rating for each product.

Query-

```
SELECT v.product_id,
       v.product_category,
       v.star_rank
FROM
  (SELECT z.product_id,
         z.product_category,z.avg_rating,
```



```

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

```

Output-

```

VERTICES: 03/03 [=====] 100% ELAPSED TIME: 186.27 s
OK
+-----+-----+-----+
| product_id | product_category | star_rank |
+-----+-----+-----+
| B006H7K6JA | Digital_Video_Games | 1 |
| B0019J7F10 | Digital_Video_Games | 2 |
| B00Y00GMW5 | Digital_Video_Games | 3 |
| B009ED35G0 | Digital_Video_Games | 4 |
| B00X024FHA | Digital_Video_Games | 5 |
| B000J4H1RA | Toys | 1 |
| B010U5EJGK | Toys | 2 |
| B000K6CZ04 | Toys | 3 |
| B000K4X8PU | Toys | 4 |
| B000K4X806 | Toys | 5 |
| B0026G0NY1 | Video_Games | 1 |
| B006046847 | Video_Games | 2 |
| B0019SVVU0 | Video_Games | 3 |
| B0055PSDKC | Video_Games | 4 |
| B0055SEPYC | Video_Games | 5 |
| B00BSW7CCI | Digital_Music_Purchase | 1 |
| B00AML80FC | Digital_Music_Purchase | 2 |
| B00AMLDM00 | Digital_Music_Purchase | 3 |
| B00AMLEB16 | Digital_Music_Purchase | 4 |
| B00BSTT1BY | Digital_Music_Purchase | 5 |
| B0000DET90 | Music | 1 |
| B007EZN7I6 | Music | 2 |
| B00000ETC3 | Music | 3 |
| B007EVR0U6 | Music | 4 |
| B0000DFZBX | Music | 5 |
| B00IB0A04E | Sports | 1 |
| B00IB12UII | Sports | 2 |
| B00IB132SA | Sports | 3 |
| B00IB13E4W | Sports | 4 |
| B00IB14KZY | Sports | 5 |
| B01KJDB1PK | Wireless | 1 |
| 0214614700 | Wireless | 2 |
| 0504482569 | Wireless | 3 |
| 1059102374 | Wireless | 4 |
| 1059340241 | Wireless | 5 |
| B0015CISJ2 | Automotive | 1 |
| B0015CGSPI | Automotive | 2 |
| B0015BP3G4 | Automotive | 3 |
| B00HEVEUJ0 | Automotive | 4 |

```

- ii. **Analytical Aggregate functions: percentile, min, max, average, standard deviation, correlation**

Using Max function to find out category which has got maximum average star rating.

Query-

```
SELECT product_category,round(avg(star_rating),2) as average_star_rating
```

```

FROM amazon_review.filtered_reviews WHERE year>=2005 group by product_category having
avg(star_rating)in ((SELECT max(x.avg_rating)
FROM
(SELECT product_category,
      avg(star_rating) AS avg_rating
FROM amazon_review.filtered_reviews
WHERE year>= 2005
GROUP BY product_category)AS x));

```

Output-

```

to DAG TERMINATED, failedTasks:0 killedTasks:20, Vertex vertex_1586545402644_0013_1_04 [Reducer 2] killed/failed due to:DAG_TERMINATED]DAG did not succeed due to DAG KILL, failedVertices:0 killedVertices:5
hive> SELECT product_category,round(avg(star_rating),2) as average_star_rating
> FROM
> FROM amazon_review.filtered_reviews WHERE year>=2005 group by product_category having avg(star_rating)in ((SELECT max(x.avg_rating)
> FROM
> (SELECT product_category,
>      avg(star_rating) AS avg_rating
> FROM amazon_review.filtered_reviews
> WHERE year>= 2005
> GROUP BY product_category)AS x));
Query ID = hadoop_20200411002329_e30b454c-f4f5-4449-ae8c-c40bd4e7e317
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586545402644_0013)

-----
VERTICES    MODE      STATUS    TOTAL    COMPLETED    RUNNING    PENDING    FAILED    KILLED
-----
Map 1 ..... container    SUCCEEDED    29         29           0           0           0           0
Map 3 ..... container    SUCCEEDED    29         29           0           0           0           0
Reducer 2 ... container    SUCCEEDED    20         20           0           0           0           0
Reducer 4 ... container    SUCCEEDED    20         20           0           0           0           0
Reducer 5 ... container    SUCCEEDED     1           1           0           0           0           0
-----
VERTICES: 05/05 [=====] 100% ELAPSED TIME: 83.75 s
-----
OK
Digital_Music_Purchase 4.65
Time taken: 84.368 seconds, Fetched: 1 row(s)
hive>

```

Please support MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

From this we can infer that **Digital music purchase category** has highest average star rating of **4.65**.

Using Minimum function to find out count of distinct products in each category that have got minimum average star rating.

Query-

```

select product_category,count(distinct(product_id)) from (
SELECT product_id,product_category,avg(star_rating)
FROM amazon_review.filtered_reviews WHERE year>=2005 group by product_id,product_category
having avg(star_rating) in (
(SELECT min(avg_rating)
FROM
(SELECT product_id,product_category,
      avg(star_rating) AS avg_rating
FROM amazon_review.filtered_reviews
WHERE year>= 2005

```

GROUP BY product_id,product_category) AS x)))z group by product_category;

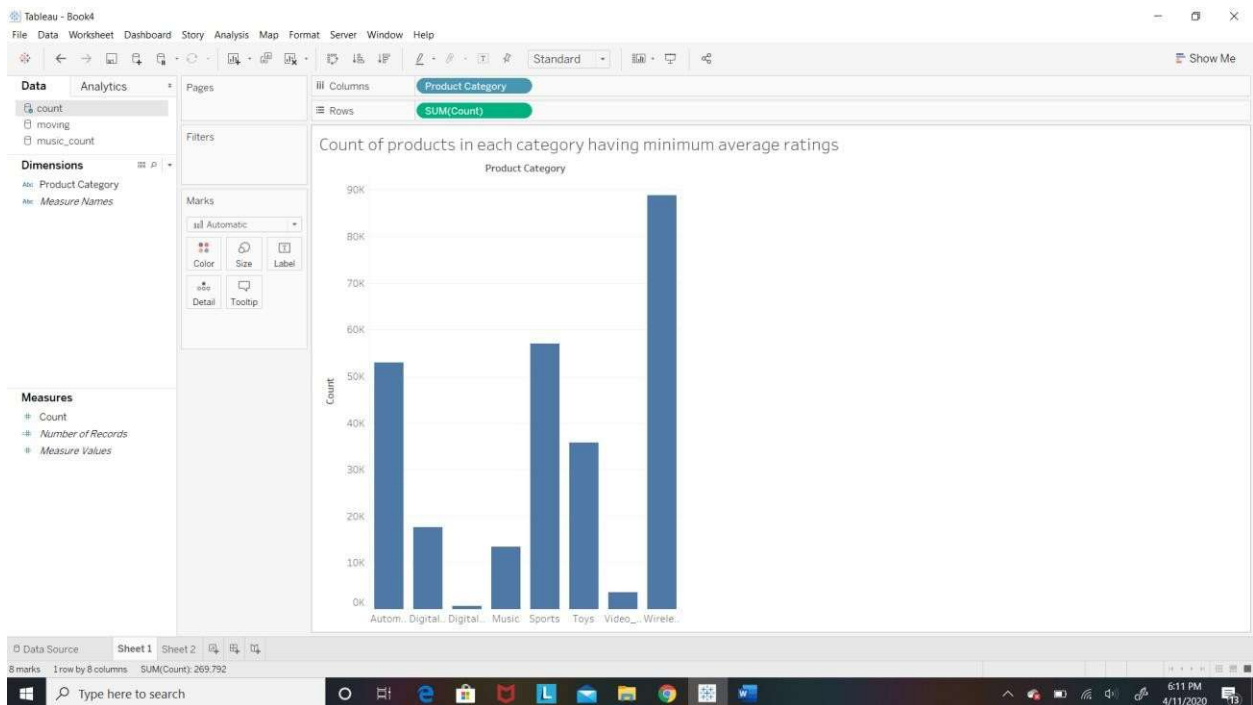
Output-

```

hive> select product_category,count(distinct(product_id)) as count from (
> SELECT product_id,product_category,avg(star_rating)
> FROM amazon_review.filtered_reviews WHERE year>=2005 group by product_id,product_category having avg(star_rating) in (
> (SELECT min(avg_rating)
> FROM
> (SELECT product_id,product_category,
> avg(star_rating) AS avg_rating
> FROM amazon_review.filtered_reviews
> WHERE year>= 2005
> GROUP BY product_id,product_category) AS x)))z group by product_category order by count desc ;
Query ID = hadoop_20200411005752_7a6295ba-c77f-4afb-93b5-c28a16b890b1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586545402644_0014)

-----
VERTICES      MODE        STATUS      TOTAL   COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED    29      29          0         0         0         0
Map 5 ..... container  SUCCEEDED    29      29          0         0         0         0
Reducer 2 ..... container  SUCCEEDED    20      20          0         0         0         0
Reducer 3 ..... container  SUCCEEDED    11      11          0         0         0         0
Reducer 4 ..... container  SUCCEEDED     1       1          0         0         0         0
Reducer 6 ..... container  SUCCEEDED    20      20          0         0         0         0
Reducer 7 ..... container  SUCCEEDED     1       1          0         0         0         0
Reducer 8 ..... container  SUCCEEDED     1       1          0         0         0         0
-----
VERTICES: 00/00 [=====] 100% ELAPSED TIME: 149.09 s
-----
OK
Wireless      88835
Sports 56990
Automotive    52888
Toys 35720
Digital_Music_Purchase 17651
Music 13460
Video Games   3598
Digital Video Games 643
Time taken: 149.755 seconds, Fetched: 8 row(s)
hive>
  
```

Visualization



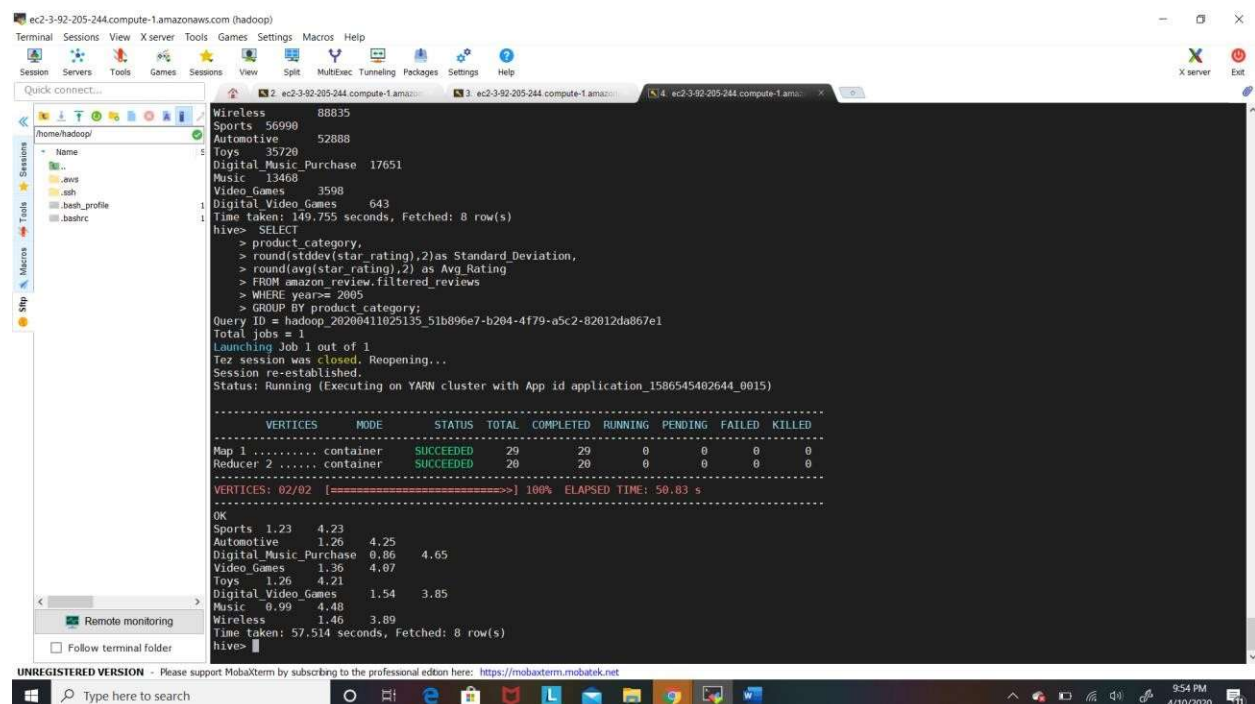
Interpretation- we can see that **Wireless category** has highest number of products that have received minimum average ratings.

Calculating Standard Deviation to analyze normal distribution of star rating of product categories.

Query-

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

Output



```
ec2-3-92-205-244.compute-1.amazonaws.com (hadoop)
Terminal Sessions View Xserver Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split Multitab Tunneling Packages Settings Help
Quick connect...
/home/hadoop
Name
. . . . .
Wireless 88835
Sports 56990
Automotive 52888
Toys 35720
Digital_Music_Purchase 17651
Music 13468
Video_Games 3598
Digital_Video_Games 643
Time taken: 149.755 seconds, Fetched: 8 row(s)
hive> SELECT
> product_category,
> round(stddev(star_rating),2)as Standard_Deviation,
> round(avg(star_rating),2) as Avg_Rating
> FROM amazon_review.filtered_reviews
> WHERE year>= 2005
> GROUP BY product_category;
Query ID = hadoop_20200411025135_51b896e7-b204-4f79-a5c2-82612da867e1
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_1586545482644_0015)

-----
VERTICES  MODE  STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  29      29          0         0         0         0
Reducer 2 ..... container  SUCCEEDED  20      20          0         0         0         0
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 50.83 s
-----
OK
Sports 1.23 4.23
Automotive 1.26 4.25
Digital_Music_Purchase 0.86 4.65
Video_Games 1.36 4.07
Toys 1.26 4.21
Digital_Video_Games 1.54 3.85
Music 0.99 4.48
Wireless 1.46 3.89
Time taken: 57.514 seconds, Fetched: 8 row(s)
hive>
```

We can see that Digital Music purchase category has least standard deviation which means it is least spread out in terms of average ratings from the mean.

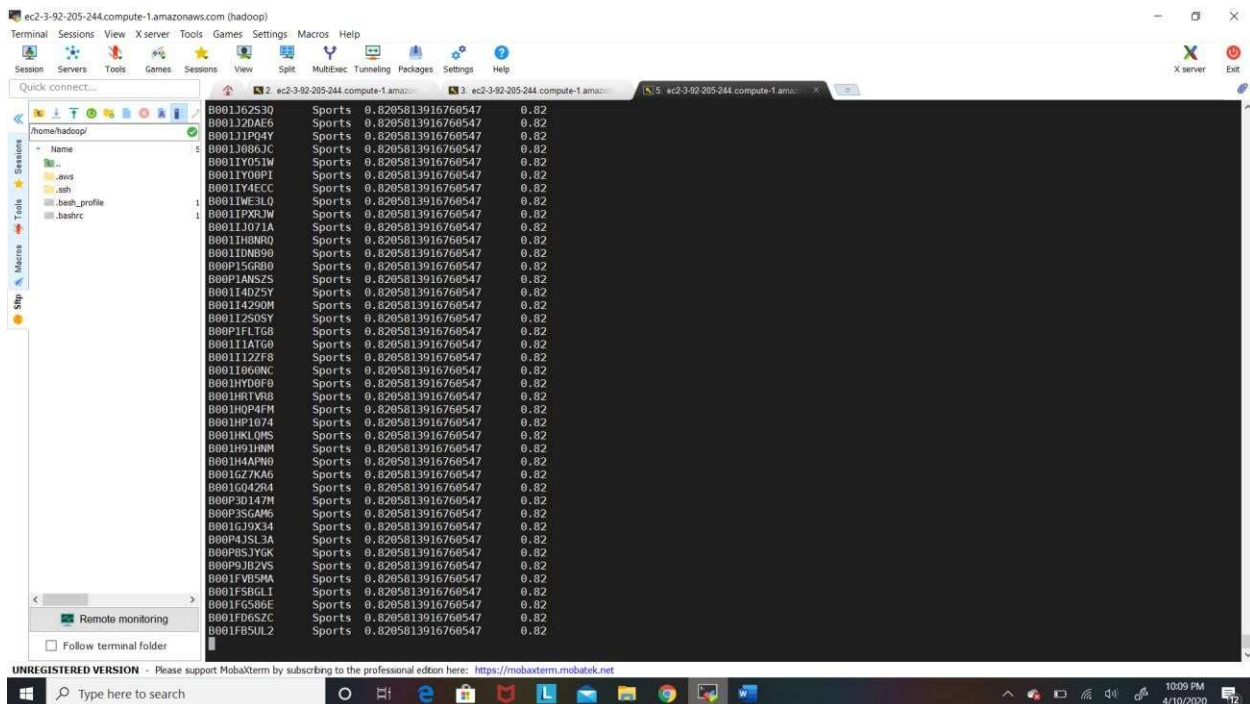
Products having highest Percentile of star ratings given by customers:

Query-

```

SELECT y.product_id,
y.product_category,y.star_rank,
round(y.star_rank,
2) AS Rank_Percentile from
(SELECT x.product_id,
x.product_category,
PERCENT_RANK()
OVER (partition by x.product_category
ORDER BY x.avg_rating desc) AS star_rank
FROM
(SELECT product_id,
product_category,
avg(star_rating) AS avg_rating
FROM amazon_review.filtered_reviews
WHERE year>= 2005
GROUP BY product_id,product_category)as x)as y order by y.star_rank
desc;

```



References- <https://towardsdatascience.com/converting-thumbs-up-thumbs-down-to-percentiles-with-skewness-intact-5ee70574a694>

<https://dzone.com/articles/100-shades-of-grey>

Database management lectures and Advanced sql slides and code for Moving average.