



APRIL 12, 2020

PROJECT REPORT

PART 1

CHINMAY SUNIL KARANDIKAR
BIG DATA MIS 6346.502
THE UNIVERSITY OF TEXAS AT DALLAS




Table of contents

Setting up the project and loading data	2
Question 1	4
Explore the dataset and provide basic exploratory analysis over time and per product category	4
Question 2	6
Provide detailed analysis of Music/Digital_Music_Purchase and Digital_Video_Games/Video_Games over time.	6
1. Do you see correlation (maybe negative) between the categories over time?	6
2. Are there same users reviewing in both categories?	7
3. Can you identify similar items in both categories? Do they get same rating?	10
4. You should cover additional questions and not limit yourself to the above questions	11
Question 3	13
You should demonstrate your ability to use Hive advanced functions:	13
1. Window functions: moving average, rank, aggregation functions using relevant ordering and partitioning	13
2. Analytical Aggregate functions: percentile, min, max, average, standard deviation, correlation	17
Percentile	17
Min	19
Max.....	20
Standard deviation	21
References	22

Setting up the project and loading data

```
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;
```

```
create view temporary
```

```
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 z.* from(
```

```
select *,row_number() over(partition by customer_id,product_id) as row1 from temporary)z where  
row1=1;
```

Question 1

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

Query

Select year, product_category, count(review_id) as NoOfReviews, count(Distinct(customer_id)) as NoOfUsers, avg(star_rating) as AvgRating ,avg(length(review_body)) as AvgLenReview, sum(case when verified_purchase='Y' then 1 else 0 end) as VerifiedPurchases, sum(case when verified_purchase='N' then 1 else 0 end) as Nonverifiedpurchases, sum(helpful_votes) as TotHelpfulVotes from amazon_review.filtered_reviews where year>=2005 group by year,product_category order by year;

Output

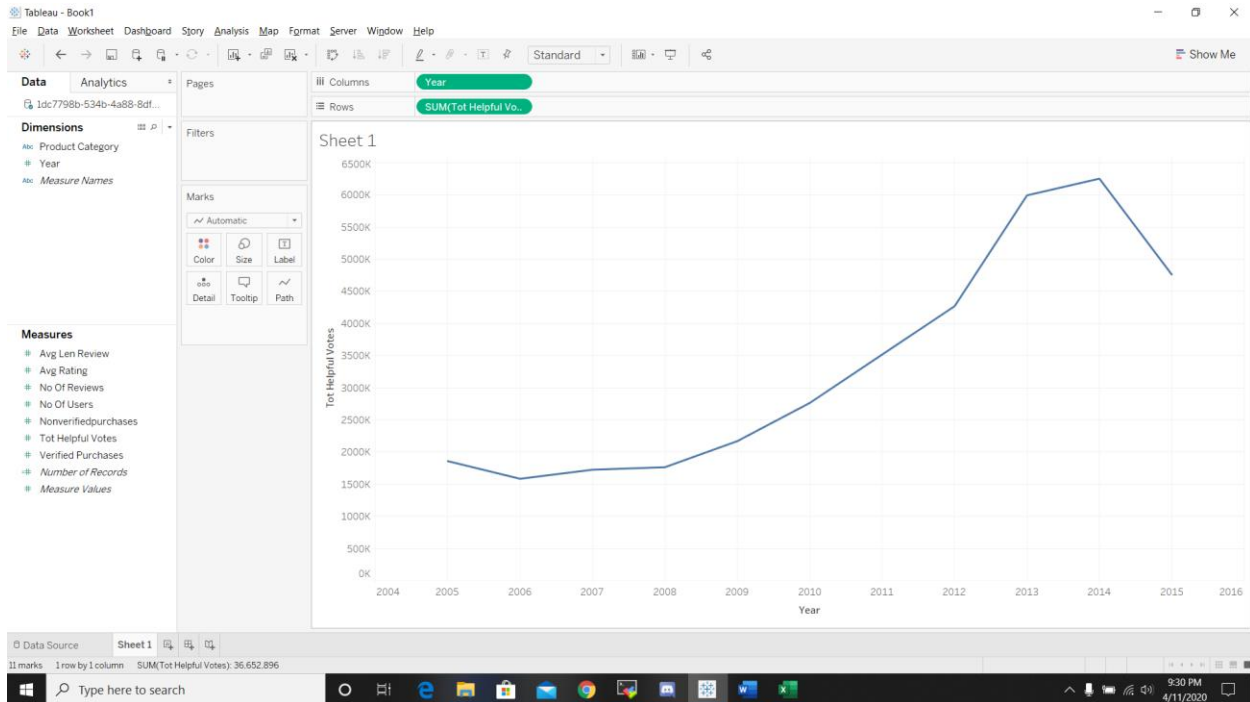
```

-----
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: 141.12 s
-----
OK
2005 Sports 4514 3975 3.6896322552060257 588.4319893664156 650 3864 47138
2005 Toys 36102 26901 3.822696803501191 573.5330729599468 2663 33439 262397
2005 Video Games 27101 17557 3.765912696948452 1238.6441697416974 1211 25890 178556
2005 Music 255095 129479 4.271765428725612 928.4475509124052 20057 235038 1243814
2005 Wireless 11835 10585 3.414026193493874 903.4652302492607 2139 9696 119942
2005 Digital_Music_Purchase 8 7 4.629 881.25 2 0 3
2005 Automotive 660 593 3.693039393939304 588.1969696969697 109 551 7844
2006 Toys 26049 21078 3.8100115460538566 568.53000111736 3133 23716 163919
2006 Video Games 24702 16992 3.751882438668934 1228.6540361104364 1934 22768 173006
2006 Digital_Video_Games 1 1 4.0 281.0 0 1 4
2006 Wireless 19055 17982 3.5994434651221356 817.3607151050919 4977 14878 154455
2006 Sports 9529 8352 3.8969461643484344 553.166019519362 2107 7422 96100
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 Music 204397 109017 4.317171974148349 927.6406356257675 25600 178797 971414
2007 Digital_Music_Purchase 2235 1913 4.405360127516779 554.7131991051455 460 1769 4219
2007 Wireless 47737 42009 3.761044891803004 587.6516329052926 18055 28682 197757
2007 Automotive 8885 7833 4.000787844682049 398.7304445694992 3862 5023 60131
2007 Music 223643 121546 4.384586148459822 820.064437518724 50885 172758 804830
2007 Toys 48108 37768 4.076847925500956 451.89240874698595 14203 33905 258254
2007 Video Games 43493 31071 3.9552111834088244 832.539075253489 7664 35829 211110
2007 Sports 29543 26146 4.067494838032698 433.07358705189724 10139 19404 189130
2008 Digital_Music_Purchase 22041 16114 4.4610408616215235 582.3237148949685 5068 16973 37953
2008 Automotive 13851 11984 3.9724929607970543 417.80716193776624 6924 6927 88545
2008 Sports 40923 35753 4.0453290325733695 454.6303301321995 16014 24909 235646
2008 Wireless 63654 55672 3.769692399534906 586.3079932133095 27705 35869 203106
2008 Digital_Video_Games 5 5 2.0 807.2 0 5 40
2008 Music 193916 187849 4.3734864580540025 838.8960323026465 50772 143144 612645
2008 Toys 65293 49910 4.078706752638108 457.32367941433233 22790 42503 279951
2008 Video Games 60169 43361 3.790722797453838 848.9686881949176 11509 48660 306375
2009 Digital_Video_Games 1561 1145 3.89237668161435 620.6816143497758 716 845 7782

```

Visualization

I have taken a line graph for the total helpful votes' vs year. We can clearly see that the number of total helpful votes kept on increasing until the year 2014 but since then dropped significantly.



Question 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?

A.

Correlation for Music/Digital_Music_Purchase

Query

```
select corr(MUSIC,DIGMUSICPURCHASE) 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 DIGMUSICPURCHASE
from amazon_review.filtered_reviews where year>=2005 group by year order by year)r;
```

Output

```
ec2-34-239-169-87.compute-1.amazonaws.com (hadoop)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split Multitex Tunneling Packages Settings Help
Quick connect...
/home/hadoop/
Name
.
.and
.ssh
.bash_profile
.bashrc
2012 Automotive 219819 125099.0
2013 Automotive 610000 311413.0
2014 Automotive 1175374 668398.0
2015 Automotive 1305043 1030139.0
Time taken: 48.251 seconds, Fetched: 86 row(s)
hive> Display all 574 possibilities? (y or n)
hive>
>;
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 DIGMUSICPURCHASE from amazon_review.filtered_reviews where year>=2005 grou
p by year order by year)r;
FAILED: SemanticException [Error 10004]: Line 1:18 Invalid table alias or column reference 'Digital_Music_Purchase': (possible column names are: year, music,
digmusicpurchase)
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 DIGMUSICPURCHASE from amazon_review.filtered_reviews where year>=2005 grou
p by year order by year)r;
FAILED: SemanticException [Error 10004]: Line 1:18 Invalid table alias or column reference 'Digital_Music_Purchase': (possible column names are: year, music,
digmusicpurchase)
hive> select corr(MUSIC,DIGMUSICPURCHASE) 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 DIGMUSICPURCHASE from amazon_review.filtered_reviews where year>=2005 grou
p by year order by year)r;
Query ID = hadoop_20200411221439_b3391d9f-72dd-4982-9337-3c73c21eba31
Total jobs = 1
Launching job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586631058862_0011)
-----
VERTICES    MODE    STATUS    TOTAL    COMPLETED    RUNNING    PENDING    FAILED    KILLED
-----
Map 1 ..... container    SUCCEEDED    28        28            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: 48.76 s
-----
OK
0.9558776790048999
Time taken: 49.255 seconds, Fetched: 1 row(s)
hive>
```

Correlation between Digital_Video_Games/Video_Games based on count of reviews.

Query

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

Output

```
> 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 DIGMUSICPURCHASE from amazon_review.filtered_reviews where year>=2005 grou
p by year order by year);
Query ID = hadoop_20200411221439_b3391d9f-72dd-4982-9337-3c73c21eba31
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586631058862_0011)

-----
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: 48.76 s
-----
OK
0.9558770790048999
Time taken: 49.255 seconds, Fetched: 1 row(s)
hive> select corr(VidGames,DigVidGames) from (
> Select year,sum(case when product_category='Video_Games' then 1 else 0 end) as VidGames,
> sum(case when product_category='Digital_Video_Games' then 1 else 0 end) as DigVidGames from amazon_review.filtered_reviews where year>=2005 group by yea
r order by year);
Query ID = hadoop_20200411221826_d26cbe58-0564-4ed3-a543-6cdf83fae5a4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586631058862_0011)

-----
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: 47.93 s
-----
OK
0.9662987902060765
Time taken: 48.39 seconds, Fetched: 1 row(s)
hive>
```

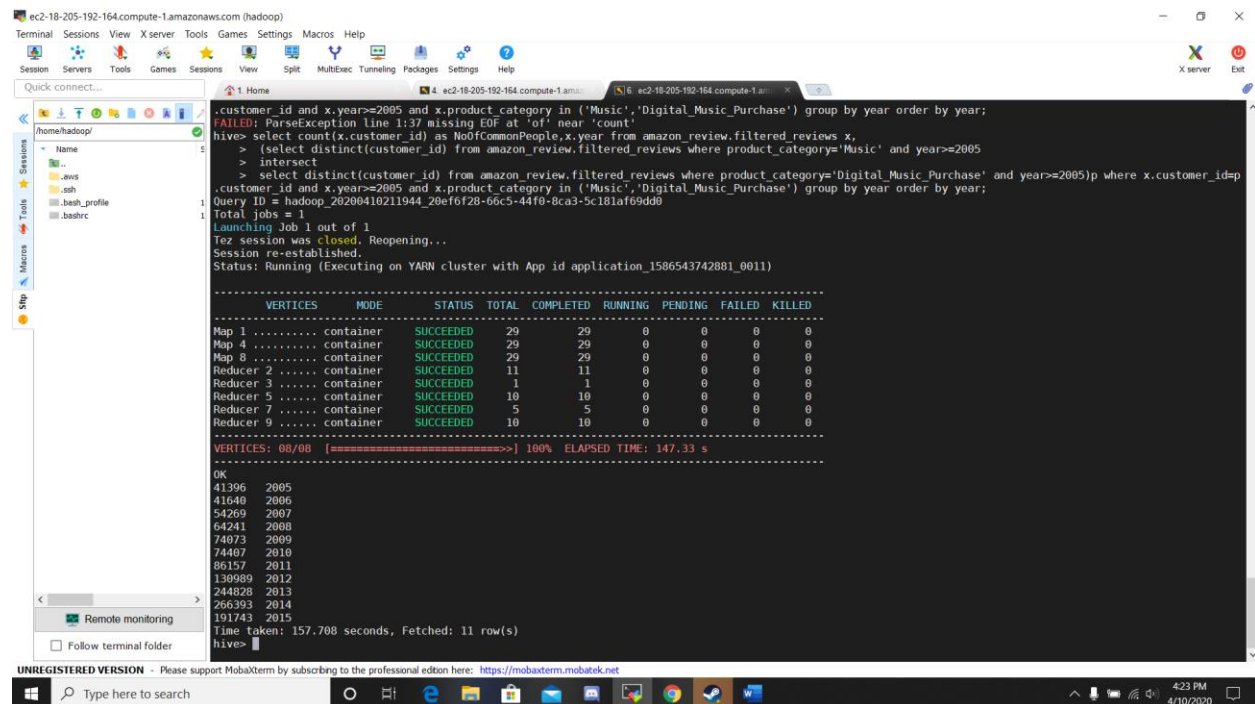
2. Are there same users reviewing in both categories?

A. Providing detailed analysis over time for the Music/Digital_Music_Purchase category.

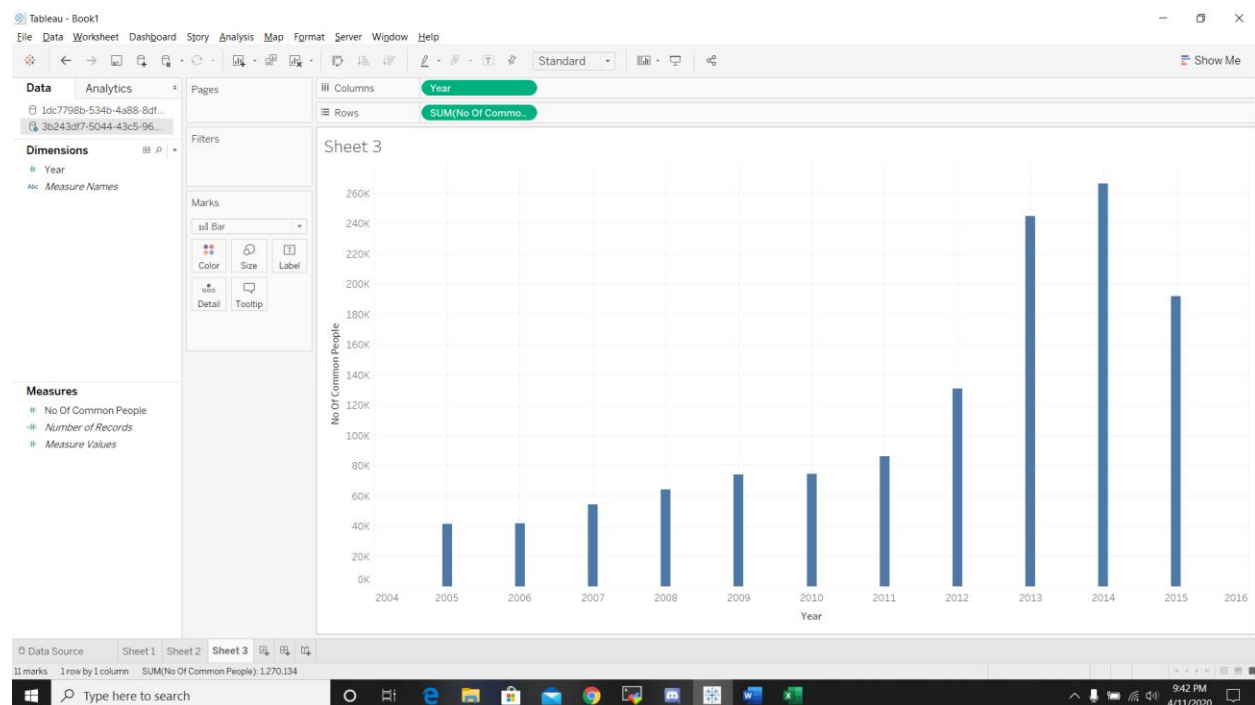
Query

```
select count(x.customer_id) as NoOfCommonPeople,x.year from amazon_review.filtered_reviews x,
(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)p where x.customer_id=p.customer_id and
x.year>=2005 and x.product_category in ('Music','Digital_Music_Purchase') group by year order by year;
```


Output



Visualization



The number of common people dropped after the year 2014 significantly.

Providing detailed analysis over time for the Digital_Video_Games/Video_Games category.

Query

```
select count(x.customer_id) as NoOfCommonPeople,x.year from amazon_review.filtered_reviews x,
(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)p where x.customer_id=p.customer_id and
x.year>=2005 and x.product_category in ('Video_Games','Digital_Video_Games') group by year order by
year;
```

Output

```
> intersect
> select distinct(customer_id) from amazon_review.filtered_reviews where product_category='Video_Games' and year>=2005)p where r.customer_id=p.customer_id and x.year>=2005 and x.product_category in ('Video_Games','Digital_Video_Games') group by year order by year;
FAILED: NullPointerException null
hive> select count(x.customer_id) as NoOfCommonPeople,x.year from amazon_review.filtered_reviews x,
> (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)p where x.customer_id=p.customer_id and x.year>=2005 and x.product_category in ('Video_Games','Digital_Video_Games') group by year order by year;
Query ID = hadoop_20200410212501_ede96c6c-4eb2-4124-b789-bd18c43233df
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586543742881_0011)

-----
VERTICES    MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  29      29          0         0         0         0
Map 4 ..... container  SUCCEEDED  29      29          0         0         0         0
Map 6 ..... container  SUCCEEDED  29      29          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: 114.55 s
-----
OK
1882  2005
1484  2006
2158  2007
3883  2008
5842  2009
8063  2010
13001 2011
18913 2012
38266 2013
44352 2014
29989 2015
Time taken: 115.403 seconds, Fetched: 11 row(s)
hive>
```

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

create view MUSIC as

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

create view DIGMUSICPURCHASE as

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

Query

Select x.product_id,AvgRatingForMusic, AvgRatingForDigMusicPur from MUSIC x inner join DIGMUSICPURCHASE p on p.product_id=x.product_id;

Output

```

ec2-18-205-192-164.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
29989 2015
Time taken: 115.403 seconds, Fetched: 11 row(s)
hive> create view MUSIC as
> select product_id,round(avg(star_rating),2) as AvgRatingForMusic from filtered_reviews where product_category='Music' and year>=2005 group by product_id;
FAILED: SemanticException [Error 10001]: Line 2:70 Table not found 'filtered_reviews'
hive> select product_id,round(avg(star_rating),2) as AvgRatingForMusic from filtered_reviews where product_category='Music' and year>=2005 group by product_id;
FAILED: SemanticException [Error 10001]: Line 1:70 Table not found 'filtered_reviews'
hive> create view MUSIC as
> select product_id,round(avg(star_rating),2) as AvgRatingForMusic from amazon_review.filtered_reviews where product_category='Music' and year>=2005 group by product_id;
OK
Time taken: 0.135 seconds
hive> create view DIGMUSICPURCHASE as
> select product_id,round(avg(star_rating),2) as AvgRatingForDigMusicPur from amazon_review.filtered_reviews where product_category='Digital_Music_Purchase' and year>=2005 group by product_id;
OK
Time taken: 0.135 seconds
hive> Select x.product_id,AvgRatingForMusic, AvgRatingForDigMusicPur from MUSIC x inner join DIGMUSICPURCHASE p on p.product_id=x.product_id;
Query ID = hadoop_20200410213706_05766141-ad11-4f46-b6ae-7c310e811b0f
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_1586543742881_0012)
-----
VERTICES  MODE  STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  29  29  0  0  0  0
Map 4 ..... container  SUCCEEDED  29  29  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: 84.88 s
-----
OK
00019H1ZJS 3.0 5.0
Time taken: 93.25 seconds, Fetched: 1 row(s)
hive>

```

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

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

Query-

create view music as

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

create view DigMusic as

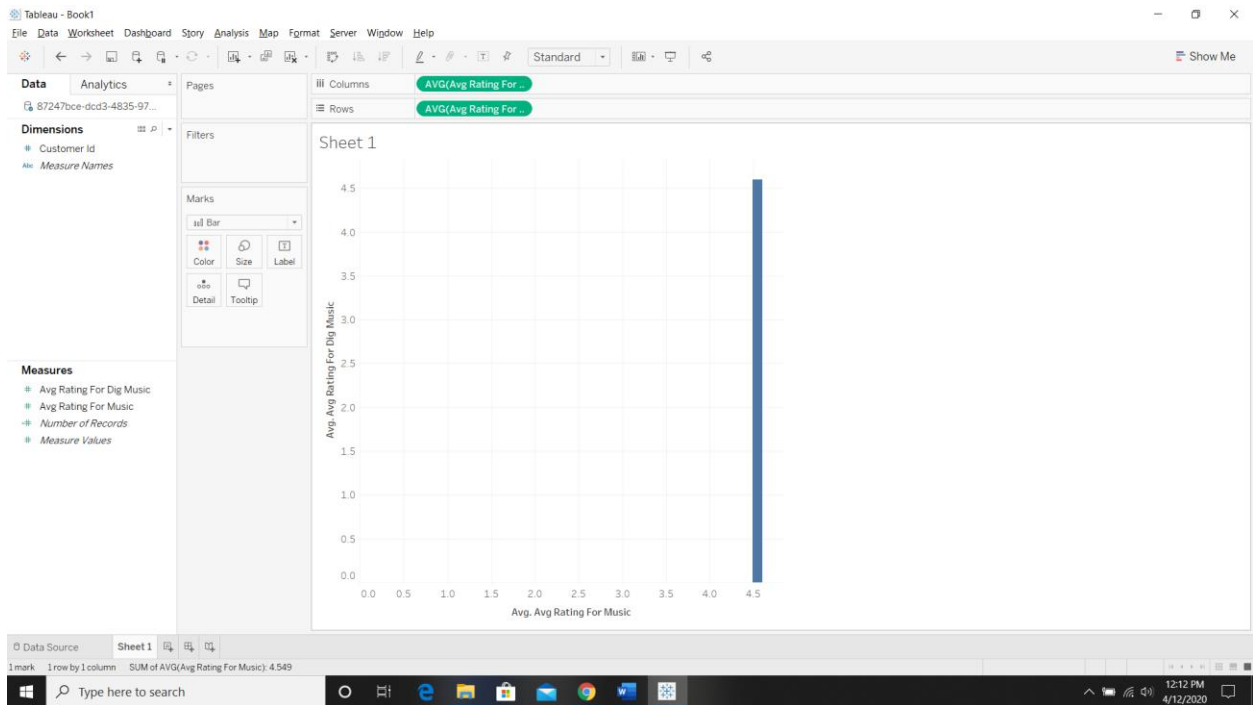
```
select customer_id,product_category,round(avg(star_rating),2) as AvgRatingForDigMusic from
amazon_review.filtered_reviews where product_category='Digital_Music_Purchase' and year>=2005
group by customer_id,product_category;
```

```
select r.customer_id, AvgRatingForMusic, AvgRatingForDigMusic from DigMusic r inner join music u on
u.customer_id=r.customer_id;
```

Output

```
ec2-34-239-169-87.compute-1.amazonaws.com (hadoop)
Terminal Sessions View X.server Tools Games Sessions View Split Multitask Tunneling Packages Settings Help
Quick connect...
/home/hadoop
Name
.aws
.ssh
.bash_profile
.bashrc
9441081 5.0 5.0
9467658 5.0 5.0
95329 5.0 5.0
9545631 5.0 5.0
9549762 3.33 5.0
957981 5.0 5.0
9666260 4.5 4.0
9689272 5.0 3.0
9708745 5.0 5.0
9733207 5.0 3.0
9754075 4.5 4.5
9777559 5.0 5.0
9803534 5.0 3.67
9810843 5.0 5.0
9837953 4.25 4.67
9839088 5.0 5.0
9869141 5.0 3.5
9880558 4.33 5.0
9882154 5.0 5.0
9882351 3.0 5.0
9883776 4.67 5.0
9887566 5.0 5.0
9892098 4.5 5.0
9894213 5.0 1.0
9895347 4.38 5.0
9897631 5.0 5.0
9900029 5.0 5.0
9905314 4.67 5.0
9913310 4.93 4.77
9915480 5.0 5.0
9921197 4.79 4.5
9929476 5.0 5.0
9939633 4.0 5.0
9943631 5.0 5.0
9945164 4.0 4.0
9951642 5.0 4.5
9956480 5.0 5.0
9979390 5.0 5.0
9986615 3.85 3.0
9991404 4.33 5.0
Time taken: 92.848 seconds, Fetched: 140830 row(s)
hive>
```

Visualization



People are fond of both digital music and music categories equally.

Question 3

You should demonstrate your ability to use Hive advanced functions:

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

Calculating three year Moving average based on number of reviews per product category over time.

Query

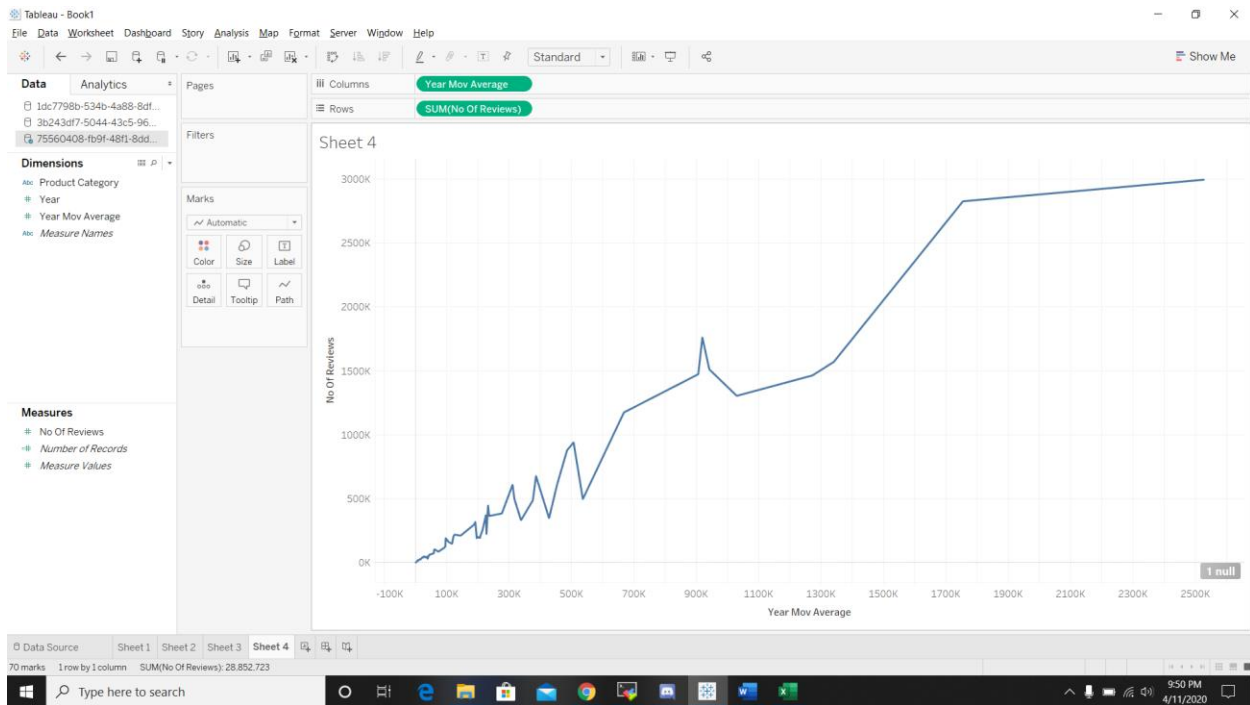
```
select year,product_category,NoOfReviews,(case when row_number() over (Partition by
product_category order by year) > 2
then round(AVG(NoOfReviews) OVER (PARTITION BY product_category order by year ROWS 2
PRECEDING))
end) as 3YearMovAverage from
(Select year,product_category,count(review_id) as NoOfReviews,count(Distinct(customer_id)) as
NumberOfUser,avg(star_rating) as average_review_stars,avg(length(review_body)) as AvgLenOfReview
from amazon_review.filtered_reviews group by year,product_category order by product_category,year)
as x where year>=2005;
```

Output

year	product_category	NoOfReviews	3YearMovAverage
2008	Music	193914	207319.0
2009	Music	204796	207452.0
2010	Music	190796	190582.0
2011	Music	201766	199119.0
2012	Music	251184	214555.0
2013	Music	498752	317207.0
2014	Music	613323	454393.0
2015	Music	497941	536672.0
2005	Sports	4514	NULL
2006	Sports	9529	NULL
2007	Sports	29542	14528.0
2008	Sports	40924	26665.0
2009	Sports	60362	43609.0
2010	Sports	98591	66626.0
2011	Sports	206668	121874.0
2012	Sports	372950	226070.0
2013	Sports	941257	506958.0
2014	Sports	1512128	942112.0
2015	Sports	1571116	1341500.0
2005	Wireless	11834	NULL
2006	Wireless	19855	NULL
2007	Wireless	47739	26476.0
2008	Wireless	63054	43740.0
2009	Wireless	93972	68455.0
2010	Wireless	162003	106543.0
2011	Wireless	319852	191942.0
2012	Wireless	677736	386530.0
2013	Wireless	1762025	919871.0
2014	Wireless	2026124	1755205.0
2015	Wireless	2995050	2527735.0
2005	Automotive	600	NULL
2006	Automotive	2190	NULL
2007	Automotive	8885	3912.0
2008	Automotive	13851	8309.0
2009	Automotive	23951	15562.0
2010	Automotive	51059	29620.0
2011	Automotive	104420	59810.0
2012	Automotive	219819	125099.0
2013	Automotive	610600	311413.0
2014	Automotive	1175374	668398.0
2015	Automotive	1305043	1030139.0

Time taken: 48.251 seconds, Fetched: 86 row(s)

Visualization



The number of reviews is increasing over the years.

Ranking top 10 products in each category based on average length of reviews.

Query

```
select product_id,product_category, AvgLenOfReview,ranking from(
select product_id,product_category, AvgLenOfReview,rank() over (Partition by product_category order
by AvgLenOfReview desc) as ranking from
(Select product_id,product_category,count(Distinct(customer_id)) as NoOfUsers,avg(star_rating) as
AvgReviewRating,avg(length(review_body)) as AvgLenOfReview
from amazon_review.filtered_reviews group by product_category,product_id)as x)as z where
ranking<=10;
```

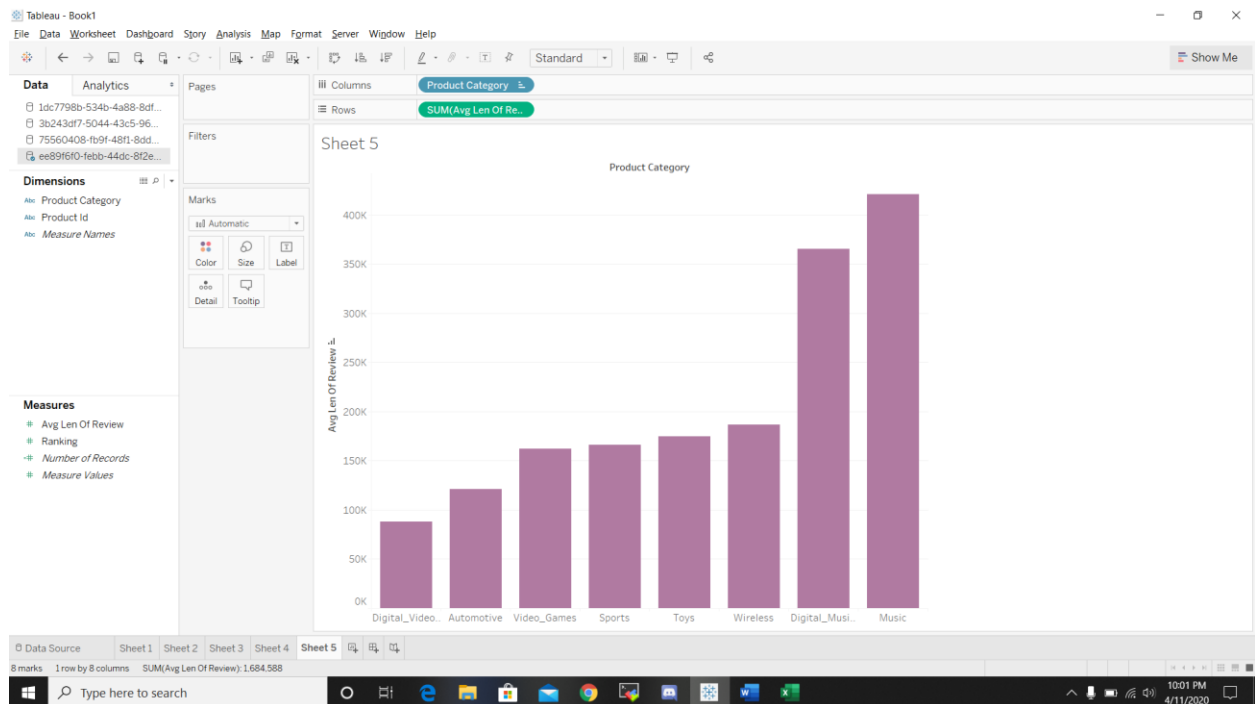

Output

```

B000E7Z60Q Video Games 19551.0 1
B000F56M9B Video Games 18969.0 3
B00ZS3JUTCM Video Games 17739.0 4
B0045Y0ORG Video Games 16786.5 5
B00065W8NG Video Games 15156.0 6
B000AF41ZU Video Games 14646.5 7
B000G9X1SG Video Games 14100.0 8
B001B8HVAU Video Games 12851.0 9
B001BHGJHS Video Games 12676.5 10
B002C9ERUG Sports 24432.0 1
B005TRIGDC Sports 18936.0 2
B005TRIGPA Sports 18071.0 3
B006Z29TH4 Sports 17450.0 4
B0077079MC Sports 16629.0 5
B001IXCI4Y Sports 14452.0 6
B00M2495G2 Sports 14077.5 7
B000450WIE Sports 13957.0 8
B000YA19BU Sports 13802.0 9
B00BE582MU Sports 13753.0 10
B0014LURX4 Digital_Music_Purchase 40588.0 1
B0014LRFJI Digital_Music_Purchase 39842.0 2
B0014LWUD0 Digital_Music_Purchase 39579.0 3
B000B90GJ1 Digital_Music_Purchase 35848.0 4
B001PW56M Digital_Music_Purchase 35364.0 5
B001P21WAG Digital_Music_Purchase 33376.0 6
B004W4RWU Digital_Music_Purchase 29399.0 7
B0010LYK3W Digital_Music_Purchase 27876.0 8
B0010LWPE Digital_Music_Purchase 27876.0 8
B0010M4YCI Digital_Music_Purchase 27876.0 8
B0010LWPS Digital_Music_Purchase 27876.0 8
B00UW94BK0 Automotive 19074.0 1
B00ZUCPUQW Automotive 16276.0 2
B00B810JSA Automotive 11969.0 3
B0002101HC Automotive 11474.0 4
B00DHBPE8 Automotive 10942.0 5
B00HCBEYW0 Automotive 10725.0 6
B00279R3M0 Automotive 10379.0 7
B00UW943UW Automotive 10136.0 8
B00416ANEC Automotive 10086.0 9
B000BNSMO Automotive 10060.0 10
Time taken: 110.613 seconds, Fetched: 81 row(s)
hive>

```

Visualization



The category music has received the lengthiest reviews.

Using aggregate function average to find out top 5 products in each marketplace based on average star rating

Query

```
SELECT v.marketplace,v.product_id,
       v.Ranking
FROM
  (SELECT z.product_id,
         z.marketplace,z.AvgRating,
         Row_number()
         OVER (partition by z.marketplace
              ORDER BY z.AvgRating desc) AS Ranking
  FROM
    (SELECT product_id,
             marketplace,
             avg(star_rating) AS AvgRating
     FROM amazon_review.filtered_reviews
     WHERE year>= 2005
     GROUP BY product_id,marketplace)as z)as v
 WHERE v.Ranking <=5 order by v.product_id;
```

Output

The screenshot shows the Mobaxterm interface with a terminal window titled "ec2-34-239-169-87.compute-1.amazonaws.com (hadoop)". The terminal output indicates a Hadoop job is running on a YARN cluster. The job has completed successfully, as shown by the "100%" progress bar and the "ELAPSED TIME: 94.01 s". Below the progress bar, a list of vertices is displayed, each with its ID, name, and status. All vertices are in a "SUCCEEDED" state.

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	28	28	0	0	0	0
Reducer 2	container	SUCCEEDED	20	20	0	0	0	0
Reducer 3	container	SUCCEEDED	10	10	0	0	0	0
Reducer 4	container	SUCCEEDED	1	1	0	0	0	0

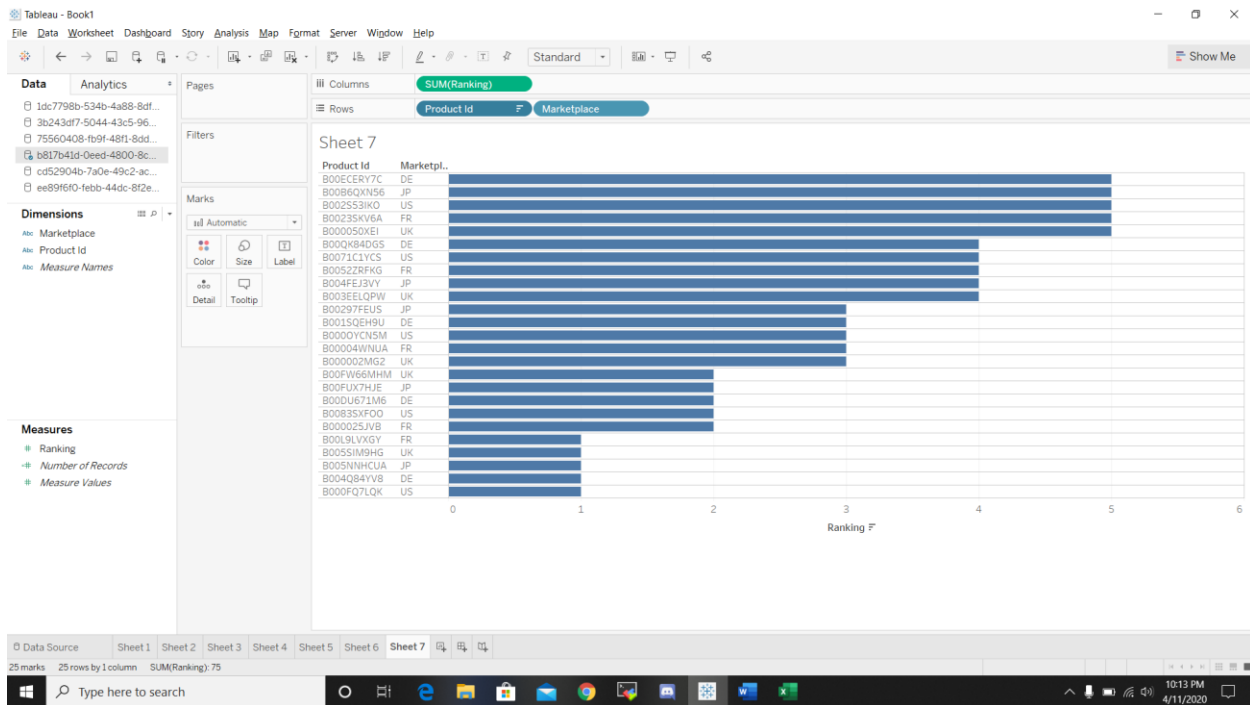
VERTICES: 04/04 [=====] 100% ELAPSED TIME: 94.01 s

OK
US 0298770164 2
US 0510599125 3
US 0545747325 4
US 0563494255 5
DE 1589949811 1
JP B00000BQPY 3
JP B00000BCWA 4
JP B00000BMFU 5
UK B0000021G6 5
UK B000002KDZ 4
UK B00000SYXZ 2
UK B00000SMHJ 1
FR B00000W3NC 1
JP B00641OZY4 1
JP B006A9X9WI 2
FR B00DGBRSMC 2
FR B00D0UNFSS 3
FR B00D0UZ5FG 4
FR B00E1ULLOY 5
UK B00JXEX4AM 3
DE B00NQGYZC4 5
DE B0006HD6Z8 4
DE B000C7ATGW 3
DE B00SDPN2EU 2
US B01C69JNCS 1

Time taken: 94.492 seconds, Fetched: 25 row(s)
hive>

At the bottom of the screen, there is a notification: "UNREGISTERED VERSION - Please support MobaxTerm by subscribing to the professional edition here: https://mobaxterm.mobatek.net"

Visualization



2. Analytical Aggregate functions: percentile, min, max, average, standard deviation, correlation

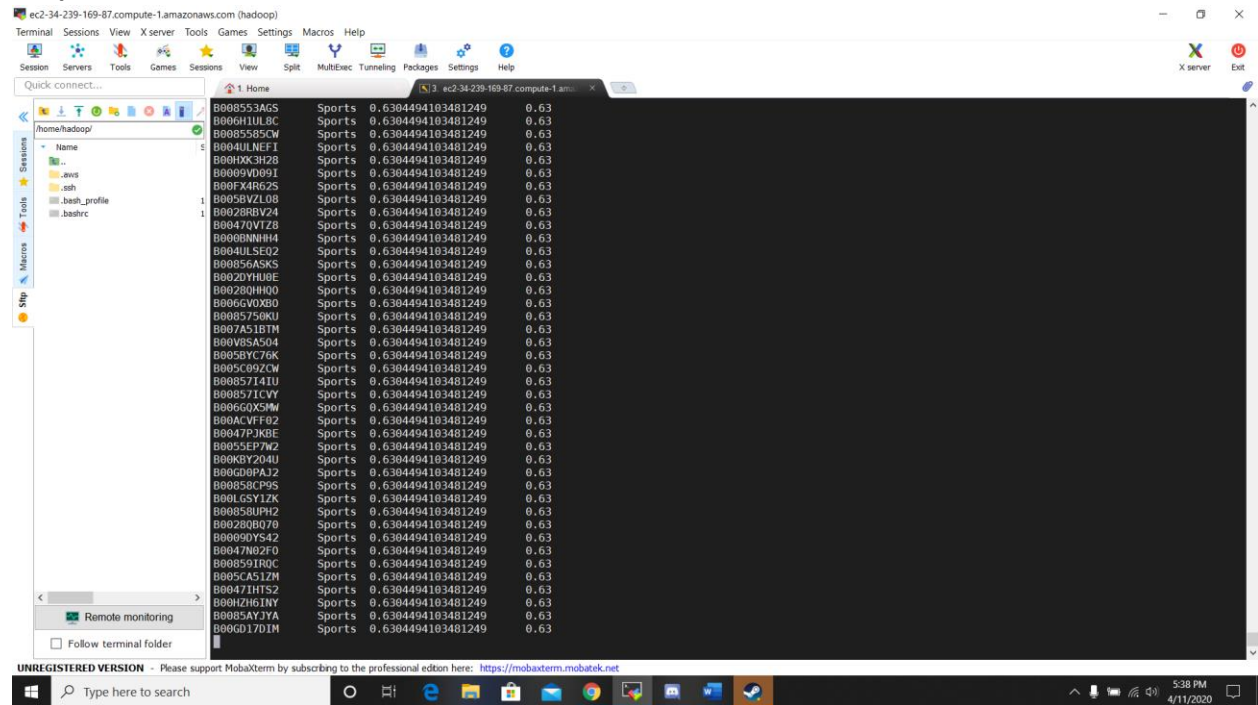
Percentile

Products having highest Percentile of star ratings given by customers:

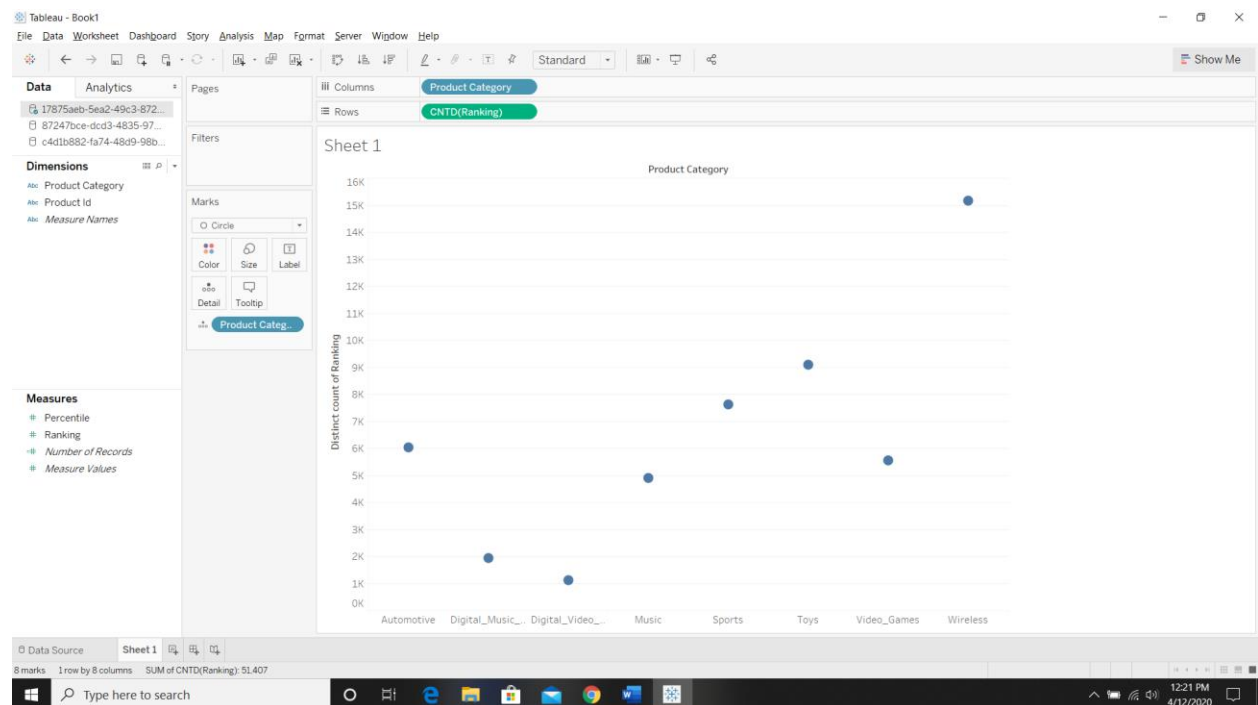
Query

```
SELECT a.product_id, a.product_category, a.Ranking, round(a.Ranking, 2) as Percentile from
(SELECT b.product_id, b.product_category, PERCENT_RANK() OVER (partition by b.product_category
ORDER BY b.AvgRating desc) AS Ranking
FROM
(SELECT product_id, product_category, avg(star_rating) AS AvgRating
FROM amazon_review.filtered_reviews
WHERE year >= 2005
GROUP BY product_id, product_category) as b) as a order by a.Ranking desc;
```

Output



Visualization



Among all the categories, the product category wireless has the maximum number of reviews and users.

Min

Query**Product category which has got minimum number of reviews.**

SELECT product_category,count(*) as NoOfReviews

FROM amazon_review.filtered_reviews WHERE year>=2005 group by product_category having count(*) in (

(SELECT min(NoOfReviews)

FROM

(SELECT product_category,
count(*) as NoOfReviews

FROM amazon_review.filtered_reviews

WHERE year>= 2005

GROUP BY product_category) as x));

Output

```

ec2-34-239-169-87.compute-1.amazonaws.com (hadoop)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultExec Tunneling Packages Settings Help
Quick connect...
/home/hadoop/
Name
.
.andi
.ssh
.bash_profile
.bashrc
1
1
1
Digital_Music_Purchase 1
Digital_Music_Purchase 1
Digital_Music_Purchase 1
Music 1
Music 1
Digital_Music_Purchase 1
Sports 1
Sports 1
Music 1
Sports 1
Music 1
Music 1
Time taken: 128.401 seconds, Fetched: 2707925 row(s)
hive> SELECT product_category,count(*) as NoOfReviews
> FROM amazon_review.filtered_reviews WHERE year>=2005 group by product_category having count(*) in (
> (SELECT min(NoOfReviews)
> FROM
> (SELECT product_category,
> count(*) as NoOfReviews
> FROM amazon_review.filtered_reviews
> WHERE year>= 2005
> GROUP BY product_category) as x));
Query ID = hadoop_20200411213927_a95f245e-3ec6-4bec-9197-bbf724ab560b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586631058862_0010)

-----
VERTICES      MODE        STATUS      TOTAL   COMPLETED   RUNNING   PENDING   FAILED   KILLED
-----
Map 1 ..... container  SUCCEEDED    20         20           0           0           0           0
Map 3 ..... container  SUCCEEDED    20         20           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: 63.30 s
-----
OK
Digital_Video_Games 145422
Time taken: 63.875 seconds, Fetched: 1 row(s)
hive>

```

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

Max

Top product which has got maximum number of reviews.

Query

```
SELECT product_id,product_category,count(*) as NoOfReviews
```

```
FROM amazon_review.filtered_reviews WHERE year>=2005 group by product_id,product_category
having count(*) in (
```

```
(SELECT max(NoOfReviews)
```

```
FROM
```

```
(SELECT product_id,product_category,
count(*) as NoOfReviews
```

```
FROM amazon_review.filtered_reviews
```

```
WHERE year>= 2005
```

```
GROUP BY product_id,product_category) as x));
```

Output

```
ec2-34-239-169-87.compute-1.amazonaws.com (hadoop)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split Multitask Tunneling Packages Settings Help
Quick connect...
1 Home
Name
/home/hadoop/
DE B005DPN2EU 2
DE B000C7M10W 3
DE B000GHJ6Z8 4
DE B00NQGZVC4 5
US B005S7JLB0 1
US B005648729 2
US B0042880742 3
US B312591551 4
US B393731138 5
Time taken: 92.897 seconds, Fetched: 25 row(s)
hive> SELECT product_id,product_category,count(*) as Number_of_reviews
> FROM amazon_review.filtered_reviews WHERE year>=2005 group by product_id,product_category having count(*) in (
> (SELECT max(Number_of_reviews)
> FROM
> (SELECT product_id,product_category,
> count(*) AS Number_of_reviews
> FROM amazon_review.filtered_reviews
> WHERE year>= 2005
> GROUP BY product_id,product_category) AS x));
Query ID = hadoop_20200411212337_a86dc99-4823-4242-b281-e1eda3abb866
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_1586631058862_0010)
-----
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-----
Map 1 ..... container SUCCEEDED 28 28 0 0 0 0
Map 3 ..... container SUCCEEDED 28 28 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: 127.36 s
-----
OK
B004S8F7QM Toys 24277
Time taken: 134.189 seconds, Fetched: 1 row(s)
hive> Display all 574 possibilities? (y or n)
```

Standard deviation

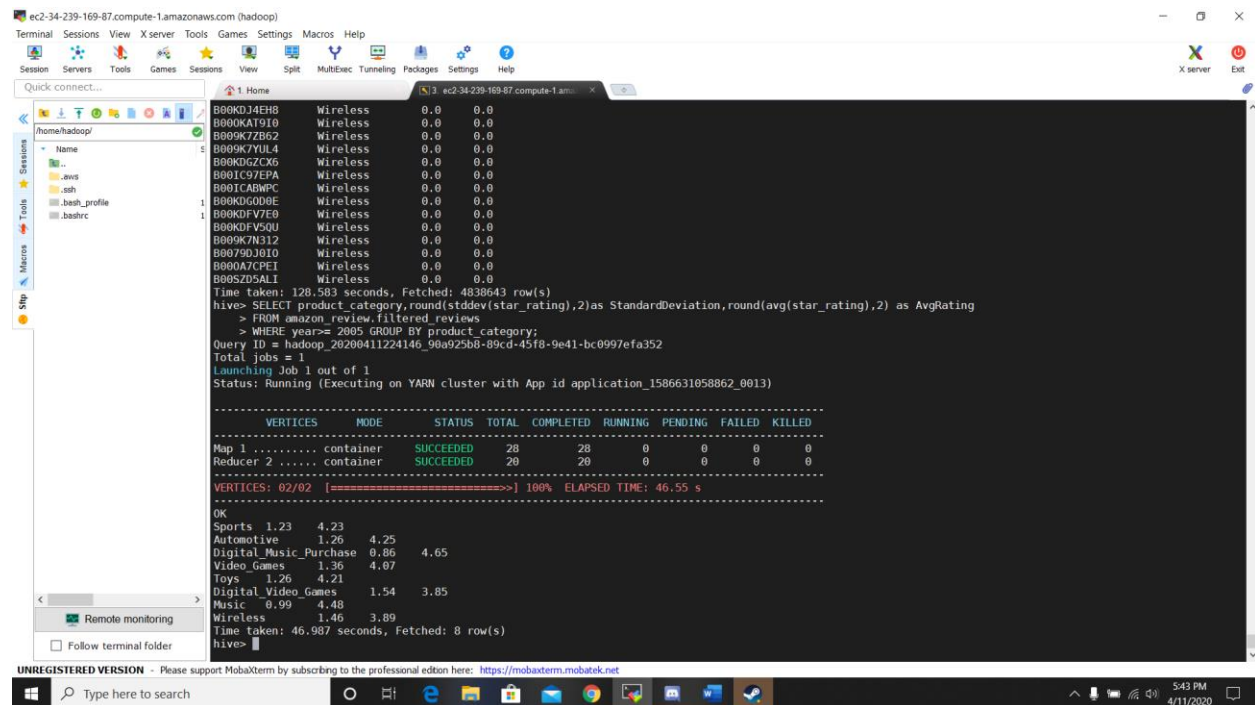
Calculating Standard Deviation to analyze normal distribution of star rating of product categories.

```
SELECT product_category,round(stddev(star_rating),2)as StandardDeviation,round(avg(star_rating),2) as AvgRating
```

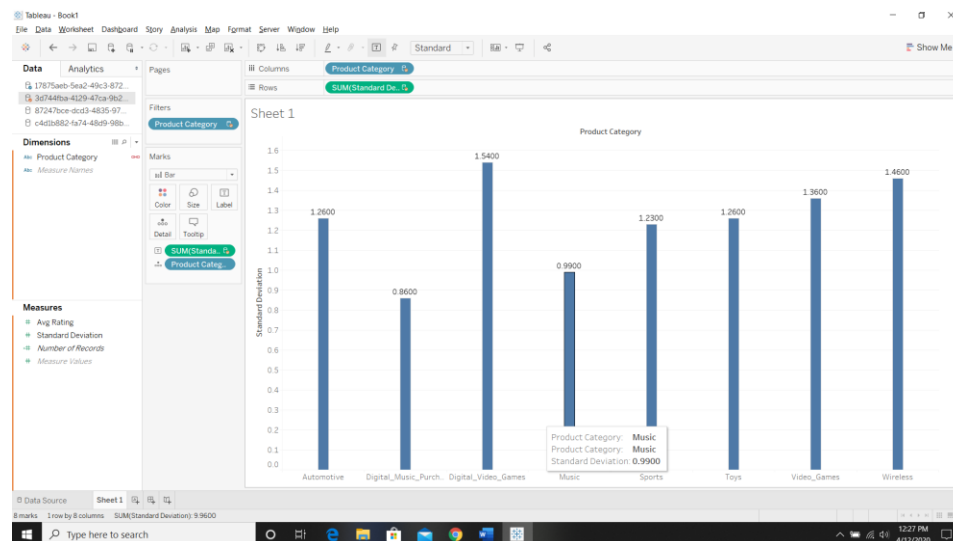
FROM amazon_review.filtered_reviews

```
WHERE year>= 2005 GROUP BY product_category;
```

Output



Visualization



The standard deviation for digital video games category is the highest which means the points of data are spread-out from the mean significantly.

References

<https://www.w3schools.com/>

https://www.pluralsight.com/courses/aws-athena-get-started?aid=701j0000001heloAAI&promo=&oid=7014Q0000022aAOQAY&utm_source=non_branded&utm_medium=digital_paid_search_google&utm_campaign=US_Dynamic&utm_content=&gclid=Cj0KCQjw-Mr0BRDyARIsAKEFbefZsQ7ZU8ISJP85zZBCMAmla0xhAxMLSjTxR4MUagHvCcMCBk9A3ugaApqpEALw_wcB

<https://www.practicefusion.com/ehr-training/>

<https://www.kaggle.com/learn/advanced-sql>

<http://www.sqlcourse2.com/>