

BDM-2 Assignment



BY: AKANKSHA UTREJA (1191056) & SWAPNIL VERMANI(11910052)

**Question 1:**

*Note: All the outputs are shaded in Grey.*

Creating the tables,

tweets = LOAD 'tweet\_info.csv' using PigStorage(',') AS (id:chararray , tweet:chararray ,user : chararray);

>> DUMP tweets;

(396124548900589568,"Long nose bitch! @cyndigullett get yo old ass off twitter &amp; whip up some grits!",ContactSwad3)

(396124549051613184,"Calling it a night I'm tired and I have to work can't wait for a damn off day I'm gone sleep till the next damn morning #OnLife",Still\_Trill216)

(396124548967317504,"Walking dead just told me smooth seas do not make skilled sailors. Never been more inspired in my life",BMosgo)

(396124549831725056,"It's November yo",antcavallaro)

(396124549462245376,"@JonesOnTheNBA thoughts on @JJRedick so far this year?",YoungAY\_)

(396124549781409792,"Wearing baggy pants was the best option for the strip club",billie\_volta)

(396124549663580160,"You care and you show it. That's what wins me over.",k\_harps25)

(396124550121156608,"New IG Name : Under\_RAT\_ed , FOLLOW YA BOY 😎👌")

(396124550406340608,"Didn't the Houston Oilers have the same song that Miami is playing? Except it was Houston oilers number. Even sounds like the same dude.",ezotic)

(396124550595088384,"I'm so excited, growing beards is my god given tallent.")

(396124550905069569,"Halloween",kannnabliss)

(396124550900899840,"@YaBoyLouieD I knew you wouldn't like it",cameronsavage\_)

(396124551203262464,"@swalkietalkie he's more my age. I will be taking him, k thanks.")

(396124551375228928,"@Sheila413Sheila @NKOTB its @donniesstarsis but she's not online",DdubsJordanFan)

(396124548866654208,"Upgraded at Staples http://t.co/Oh7PHJATqs",blakegrimsley)

(396124552126025728,"Can I get one more tat please",Asap\_Gaydle)

(396124551735562240,"I'm just a lost girl waiting for someone to guide me",AlexaDiana21)

(396124551928492032,"@dannijayy\_ im tryin to with no shortcuts ",KingxJulien)

(396124552025346049,"@Deandra\_T favvvvv song 😍😍😍",madis0nbabyyy)

(396124551530037248,"Friend of a friend the one and only.. Will Ferrell! How can you not love him? http://t.co/9nWxNHp11G",Zachaouki8)

(396124552306388992,"Still up! Gonna be up ALL night!!",lane\_xoxo\_x3)

(396124552352505856,"Someone liked my Instagram picture and their name is ""greatesthotgirlz""...except I read it as greatesTHOTgirlz. 😂",laurenpascua)

users = LOAD 'user\_info.csv' using PigStorage(',') AS (login\_info:chararray , username:chararray ,state : chararray);

>> DUMP users;

(Everything\_OMAR,Claire Adams,MH)

(angelbebeko,Juan Perry,NV)

(SeattlePoker,Lucas Cook,DE)

(Bmayyyyy,Charlotte Cook,NE)

(\_maaesisonn,Parker Nelson,OH)

(iLoveJMA,Jace Sanchez,WA)

(Jihoon\_KK,Nevaeh Perry,MO)

(donahoe\_ryan,Michael Bryant,GU)

(lala251,Hailey Hughes,GU)

(Trobbbbbbbbb,Noah Martin,AZ)

(infinitepharaoh,Nolan King,WA)

(SheCrazyy\_,Daniel Walker,FL)

(TaeeJnes,Parker Wilson,VI)

(MiriamBustam,Caroline Kelly,AS)

(nancykJJJ99,Tristan Anderson,MT)

(JoshL1220,Abigail Sanders,OK)

(jcizzle40,Arianna Allen,KS)

(JoshuaRizzo,Caroline Cox,KY)

(MorganStarnes,Annabelle Long,MP)

(briana\_leavens,Emily Nelson,AS)

(AngeMarieH,Lily Barnes,NC)

2.

C = FOREACH users GENERATE login\_info, state;

DUMP C;

(ContactSwad3,IA)

(Still\_Trill216,NM)

(BMosgo,MP)

(antcavallaro,GU)

(YoungAY\_,GA)

(billie\_volta,MT)

(k\_harps25,ME)

(RDY\_4Life,KS)

(ezotic,NH)

(ElliottPeters11,MN)

(kannnabliss,IL)

(cameronsavage\_,MA)

(elisabethhh\_\_,IN)

(DdubsJordanFan,DE)

(blakegrimsley,PR)

D = FILTER C BY state=='LA';

grunt> DUMP D;

(Unodaone1,LA)

(OakvilleWx,LA)

(katedashh,LA)

(AnthonyGalla,LA)

(Pistol\_016,LA)

(bizcarr,LA)

(Gabbbss\_,LA)

(Phillyy\_Phill,LA)

(Shimi\_Souliki,LA)

(BigBoootyJuddy\_,LA)

(joisiler,LA)

(BixTweet004,LA)

(JustissMariee,LA)

(ambreacastle,LA)

(Novocane\_Feelin,LA)

(ayye\_its\_kaypee,LA)

(StallionK\_,LA)

(AhavaJadori,LA)

(gabbysroom\_,LA)

grunt> E= FOREACH D GENERATE login\_info;

grunt> DUMP E

(Unodaone1)

(OakvilleWx)

(katedashh)

(AnthonyGalla)

(Pistol\_016)

(bizcarr)

(Gabbbss\_)

(Phillyy\_Phill)

(Shimi\_Souliki)

(BigBoootyJuddy\_)

(joisiler)

(BixTweet004)

(JustissMariee)

(ambreacastle)

(Novocane\_Feelin)

(ayye\_its\_kaypee)

(StallionK\_)

(AhavaJadori)

(gabbysroom\_)

3.

grunt> D = FILTER tweets by tweet MATCHES '.\*favorite.\*';

(396124461256417280,"Honest Dm ? ❗️ favorite this",\_\_flyygirll)

(396124464506617856,"Sleeping Naked w/ my favorite cover &gt;",JessThee\_Best)

(396124504281579520,"I can't find my favorite leggings OR my Ugg boots so I'm currently having a white girl heart attack",Gracie\_bee\_)

4.

grunt> result =join users by login\_info, tweets by user;

grunt> DUMP result;

STORE result into 'resultJoin' USING PigStorage('\t');

5.

E = GROUP result by user;

grunt> tweetCount = FOREACH E GENERATE $0 AS USER ,COUNT($1) AS tweet\_count;

grunt> DUMP tweetCount;

(jorge\_aenriquez,1)

(kelseymariee004,1)

(laurenemilydean,2)

(lolololollollll,2)

(mydaughterworld,1)

(prfessorbigsock,1)

(rachelfischbach,2)

grunt> F = ORDER tweetCount BY $1 DESC;

grunt> Result = limit F 10;

grunt> dump Result;

(Latinaaa\_13,3)

(Murray619,3)

(Nita2blessed,2)

(\_\_flyygirll,2)

(lolololollollll,2)

(so\_nt\_average,2)

(KINGxKemmy,2)

(Blow1hoe,2)

(Enuff\_Said,2)

(trevorcook42,2)

6.

TweetMore = FILTER F by tweetCount>=5;

dump TweetMore;

0

max count=3

7.

NoTweet = FILTER F by tweetCount=0;

dump NoTweet;

0

**Question 2:**

hdfs dfs -mkdir /user/akanksha/order

hdfs dfs -put u.data /user/akanksha/order/

hdfs dfs -put u.item /user/akanksha/order/

hdfs dfs -put u.user /user/akanksha/order/

hdfs dfs -ls /user/akanksha/order/

hive

show databases;

use default;

-----------------------------------------------------------------------------------------------------------------

#Creating table 1:

CREATE TABLE Users (userid INT, age INT, gender STRING, occupation STRING, zipcode INT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '|';

#viewing the schema,

describe Users;

#Loading the data into table,

LOAD DATA INPATH '/user/akanksha/order/u.user' OVERWRITE INTO TABLE Users;

#Viweing data,

select \* from Users limit 10;

-------------------------------------------------------------------------------------------------------------------

#Creating table 2:

CREATE TABLE Movies (movieid INT, movietitle STRING,releaseDate STRING,Unknown STRING ,Url STRING, Action INT, Adventure INT, Animation INT,Children INT, Comedy INT, Crime INT,Documentary INT, Drama INT, Fantasy INT, FilmNoir INT, Horror INT, Musical INT,Mystery INT, Romance INT, SciFi INT,Thriller INT, War INT, Western INT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '|';

#viewing the schema,

describe Movies ;

#Loading the data into table,

LOAD DATA INPATH '/user/akanksha/order/u.item' OVERWRITE INTO TABLE Movies;

-------------------------------------------------------------------------------------------------------------------

#Creating table 3:

CREATE TABLE Rating (userid INT, itemid INT, Rating INT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t';

#viewing the schema,

describe Rating ;

#Loading the data into table,

LOAD DATA INPATH '/user/akanksha/order/u.data' OVERWRITE INTO TABLE Rating;

---------------------------------------------------------------------------------------------------------------------

(a)

a)

SELECT movies.movietitle, count(rating.itemid) as count

FROM movies JOIN Rating ON rating.itemid=movies.movieid

GROUP BY movietitle,movies.movieid

ORDER BY count DESC

LIMIT 30;

Star Wars (1977) 583

Contact (1997) 509

Fargo (1996) 508

Return of the Jedi (1983) 507

Liar Liar (1997) 485

English Patient, The (1996) 481

Scream (1996) 478

Toy Story (1995) 452

Air Force One (1997) 431

Independence Day (ID4) (1996) 429

Raiders of the Lost Ark (1981) 420

Godfather, The (1972) 413

Pulp Fiction (1994) 394

Twelve Monkeys (1995) 392

Silence of the Lambs, The (1991) 390

Jerry Maguire (1996) 384

Rock, The (1996) 378

Empire Strikes Back, The (1980) 367

Star Trek: First Contact (1996) 365

Titanic (1997) 350

Back to the Future (1985) 350

Mission: Impossible (1996) 344

Fugitive, The (1993) 336

Indiana Jones and the Last Crusade (1989) 331

Willy Wonka and the Chocolate Factory (1971) 326

Princess Bride, The (1987) 324

Forrest Gump (1994) 321

Monty Python and the Holy Grail (1974) 316

Saint, The (1997) 316

Full Monty, The (1997) 315

b)

SELECT rating.itemid, movies.movietitle, avg(rating) as avg

FROM movies JOIN Rating ON rating.itemid=movies.movieid

GROUP BY movies.movietitle, rating.itemid

ORDER BY avg DESC

LIMIT 30;

814 Great Day in Harlem, A (1994) 5.0

1189 Prefontaine (1997) 5.0

1599 Someone Else's America (1995) 5.0

1293 Star Kid (1997) 5.0

1653 Entertaining Angels: The Dorothy Day Story (1996) 5.0

1201 Marlene Dietrich: Shadow and Light (1996) 5.0

1500 Santa with Muscles (1996) 5.0

1467 Saint of Fort Washington, The (1993) 5.0

1536 Aiqing wansui (1994) 5.0

1122 They Made Me a Criminal (1939) 5.0

1449 Pather Panchali (1955) 4.625

1594 Everest (1998) 4.5

1642 Some Mother's Son (1996) 4.5

1398 Anna (1996) 4.5

119 Maya Lin: A Strong Clear Vision (1994) 4.5

408 Close Shave, A (1995) 4.491071428571429

318 Schindler's List (1993) 4.466442953020135

169 Wrong Trousers, The (1993) 4.466101694915254

483 Casablanca (1942) 4.45679012345679

114 Wallace & Gromit: The Best of Aardman Animation (1996) 4.447761194029851

64 Shawshank Redemption, The (1994) 4.445229681978798

603 Rear Window (1954) 4.3875598086124405

12 Usual Suspects, The (1995) 4.385767790262173

50 Star Wars (1977) 4.3584905660377355

178 12 Angry Men (1957) 4.344

1191 Letter From Death Row, A (1998) 4.333333333333333

513 Third Man, The (1949) 4.333333333333333

1639 Bitter Sugar (Azucar Amargo) (1996) 4.333333333333333

134 Citizen Kane (1941) 4.292929292929293

963 Some Folks Call It a Sling Blade (1993) 4.2926829268292686

c)

SELECT rating.itemid, movies.movietitle,count(rating.itemid) as count, avg(rating) as avg

FROM movies JOIN Rating ON rating.itemid=movies.movieid

GROUP BY movies.movietitle, rating.itemid

having count >=50

ORDER BY avg DESC

LIMIT 30;

408 Close Shave, A (1995) 112 4.491071428571429

318 Schindler's List (1993) 298 4.466442953020135

169 Wrong Trousers, The (1993) 118 4.466101694915254

483 Casablanca (1942) 243 4.45679012345679

114 Wallace & Gromit: The Best of Aardman Animation (1996) 67 4.447761194029851

64 Shawshank Redemption, The (1994) 283 4.445229681978798

603 Rear Window (1954) 209 4.3875598086124405

12 Usual Suspects, The (1995) 267 4.385767790262173

50 Star Wars (1977) 583 4.3584905660377355

178 12 Angry Men (1957) 125 4.344

513 Third Man, The (1949) 72 4.333333333333333

134 Citizen Kane (1941) 198 4.292929292929293

427 To Kill a Mockingbird (1962) 219 4.292237442922374

357 One Flew Over the Cuckoo's Nest (1975) 264 4.291666666666667

98 Silence of the Lambs, The (1991) 390 4.28974358974359

480 North by Northwest (1959) 179 4.284916201117318

127 Godfather, The (1972) 413 4.283292978208232

285 Secrets & Lies (1996) 162 4.265432098765432

272 Good Will Hunting (1997) 198 4.262626262626263

657 Manchurian Candidate, The (1962) 131 4.259541984732825

474 Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1963) 194 4.252577319587629

174 Raiders of the Lost Ark (1981) 420 4.252380952380952

479 Vertigo (1958) 179 4.251396648044692

313 Titanic (1997) 350 4.2457142857142856

511 Lawrence of Arabia (1962) 173 4.23121387283237

484 Maltese Falcon, The (1941) 138 4.2101449275362315

172 Empire Strikes Back, The (1980) 367 4.204359673024523

515 Boot, Das (1981) 201 4.203980099502488

488 Sunset Blvd. (1950) 65 4.2

223 Sling Blade (1996) 136 4.198529411764706

----------------------------

SELECT rating.itemid, movies.movietitle,count(rating.itemid) as count, avg(rating) as avg

FROM movies JOIN Rating ON rating.itemid=movies.movieid

GROUP BY movies.movietitle, rating.itemid

having count >=20

ORDER BY avg DESC

LIMIT 30;

408 Close Shave, A (1995) 112 4.491071428571429

318 Schindler's List (1993) 298 4.466442953020135

169 Wrong Trousers, The (1993) 118 4.466101694915254

483 Casablanca (1942) 243 4.45679012345679

114 Wallace & Gromit: The Best of Aardman Animation (1996) 67 4.447761194029851

64 Shawshank Redemption, The (1994) 283 4.445229681978798

603 Rear Window (1954) 209 4.3875598086124405

12 Usual Suspects, The (1995) 267 4.385767790262173

50 Star Wars (1977) 583 4.3584905660377355

178 12 Angry Men (1957) 125 4.344

513 Third Man, The (1949) 72 4.333333333333333

134 Citizen Kane (1941) 198 4.292929292929293

963 Some Folks Call It a Sling Blade (1993) 41 4.2926829268292686

427 To Kill a Mockingbird (1962) 219 4.292237442922374

357 One Flew Over the Cuckoo's Nest (1975) 264 4.291666666666667

98 Silence of the Lambs, The (1991) 390 4.28974358974359

480 North by Northwest (1959) 179 4.284916201117318

127 Godfather, The (1972) 413 4.283292978208232

285 Secrets & Lies (1996) 162 4.265432098765432

272 Good Will Hunting (1997) 198 4.262626262626263

251 Shall We Dance? (1996) 46 4.260869565217392

657 Manchurian Candidate, The (1962) 131 4.259541984732825

474 Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1963) 194 4.252577319587629

174 Raiders of the Lost Ark (1981) 420 4.252380952380952

479 Vertigo (1958) 179 4.251396648044692

313 Titanic (1997) 350 4.2457142857142856

511 Lawrence of Arabia (1962) 173 4.23121387283237

641 Paths of Glory (1957) 33 4.212121212121212

484 Maltese Falcon, The (1941) 138 4.2101449275362315

172 Empire Strikes Back, The (1980) 367 4.204359673024523

--------------------------------

SELECT rating.itemid, movies.movietitle,count(rating.itemid) as count, avg(rating) as avg

FROM movies JOIN Rating ON rating.itemid=movies.movieid

GROUP BY movies.movietitle, rating.itemid

having count >=100

ORDER BY avg DESC

LIMIT 30;

408 Close Shave, A (1995) 112 4.491071428571429

318 Schindler's List (1993) 298 4.466442953020135

169 Wrong Trousers, The (1993) 118 4.466101694915254

483 Casablanca (1942) 243 4.45679012345679

64 Shawshank Redemption, The (1994) 283 4.445229681978798

603 Rear Window (1954) 209 4.3875598086124405

12 Usual Suspects, The (1995) 267 4.385767790262173

50 Star Wars (1977) 583 4.3584905660377355

178 12 Angry Men (1957) 125 4.344

134 Citizen Kane (1941) 198 4.292929292929293

427 To Kill a Mockingbird (1962) 219 4.292237442922374

357 One Flew Over the Cuckoo's Nest (1975) 264 4.291666666666667

98 Silence of the Lambs, The (1991) 390 4.28974358974359

480 North by Northwest (1959) 179 4.284916201117318

127 Godfather, The (1972) 413 4.283292978208232

285 Secrets & Lies (1996) 162 4.265432098765432

272 Good Will Hunting (1997) 198 4.262626262626263

657 Manchurian Candidate, The (1962) 131 4.259541984732825

474 Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1963) 194 4.252577319587629

174 Raiders of the Lost Ark (1981) 420 4.252380952380952

479 Vertigo (1958) 179 4.251396648044692

313 Titanic (1997) 350 4.2457142857142856

511 Lawrence of Arabia (1962) 173 4.23121387283237

484 Maltese Falcon, The (1941) 138 4.2101449275362315

172 Empire Strikes Back, The (1980) 367 4.204359673024523

515 Boot, Das (1981) 201 4.203980099502488

223 Sling Blade (1996) 136 4.198529411764706

316 As Good As It Gets (1997) 112 4.196428571428571

187 Godfather: Part II, The (1974) 209 4.186602870813397

498 African Queen, The (1951) 152 4.184210526315789

From the above when we change the threshold, we can see that the list of movies change as the threshold changes though the top movies remains the same.

C)

Avg rating by Male==>

SELECT rating.itemid, users.gender, movies.movietitle, avg(rating) as avg

FROM Rating

JOIN MOVIES ON rating.itemid=movies.movieid

JOIN USERS on rating.userid= users.userid

GROUP BY movies.movietitle, rating.itemid, users.gender

having users.gender= 'M'

ORDER BY avg DESC

LIMIT 30;

1656 M Little City (1998) 5.0

814 M Great Day in Harlem, A (1994) 5.0

1306 M Delta of Venus (1994) 5.0

1293 M Star Kid (1997) 5.0

1612 M Leading Man, The (1996) 5.0

1653 M Entertaining Angels: The Dorothy Day Story (1996) 5.0

1122 M They Made Me a Criminal (1939) 5.0

1201 M Marlene Dietrich: Shadow and Light (1996) 5.0

1189 M Prefontaine (1997) 5.0

1191 M Letter From Death Row, A (1998) 5.0

1175 M Hugo Pool (1997) 5.0

1605 M Love Serenade (1996) 5.0

1536 M Aiqing wansui (1994) 5.0

1500 M Santa with Muscles (1996) 5.0

1144 M Quiet Room, The (1996) 5.0

1467 M Saint of Fort Washington, The (1993) 5.0

1473 M Little Princess, The (1939) 4.666666666666667

851 M Two or Three Things I Know About Her (1966) 4.666666666666667

1449 M Pather Panchali (1955) 4.625

1398 M Anna (1996) 4.5

1251 M A Chef in Love (1996) 4.5

1499 M Grosse Fatigue (1994) 4.5

1639 M Bitter Sugar (Azucar Amargo) (1996) 4.5

1429 M Sliding Doors (1998) 4.5

483 M Casablanca (1942) 4.473404255319149

169 M Wrong Trousers, The (1993) 4.472527472527473

408 M Close Shave, A (1995) 4.462365591397849

513 M Third Man, The (1949) 4.433333333333334

114 M Wallace & Gromit: The Best of Aardman Animation (1996) 4.423076923076923

641 M Paths of Glory (1957) 4.419354838709677

Avg Rating by Females==>

SELECT rating.itemid, users.gender, movies.movietitle, avg(rating) as avg

FROM Rating

JOIN MOVIES ON rating.itemid=movies.movieid

JOIN USERS on rating.userid= users.userid

GROUP BY movies.movietitle, rating.itemid, users.gender

having users.gender= 'F'

ORDER BY avg DESC

LIMIT 30;

119 F Maya Lin: A Strong Clear Vision (1994) 5.0

884 F Year of the Horse (1997) 5.0

1451 F Foreign Correspondent (1940) 5.0

1189 F Prefontaine (1997) 5.0

74 F Faster Pussycat! Kill! Kill! (1965) 5.0

1599 F Someone Else's America (1995) 5.0

1301 F Stripes (1981) 5.0

1368 F Mina Tannenbaum (1994) 5.0

1472 F Visitors, The (Visiteurs, Les) (1993) 5.0

1594 F Everest (1998) 5.0

883 F Telling Lies in America (1997) 5.0

318 F Schindler's List (1993) 4.632911392405063

408 F Close Shave, A (1995) 4.631578947368421

64 F Shawshank Redemption, The (1994) 4.5625

114 F Wallace & Gromit: The Best of Aardman Animation (1996) 4.533333333333333

251 F Shall We Dance? (1996) 4.529411764705882

1218 F Friday (1995) 4.5

1298 F Band Wagon, The (1953) 4.5

1172 F Women, The (1939) 4.5

1642 F Some Mother's Son (1996) 4.5

169 F Wrong Trousers, The (1993) 4.444444444444445

963 F Some Folks Call It a Sling Blade (1993) 4.416666666666667

223 F Sling Blade (1996) 4.411764705882353

285 F Secrets & Lies (1996) 4.407407407407407

483 F Casablanca (1942) 4.4

633 F Christmas Carol, A (1938) 4.4

83 F Much Ado About Nothing (1993) 4.381818181818182

272 F Good Will Hunting (1997) 4.3559322033898304

603 F Rear Window (1954) 4.351851851851852

427 F To Kill a Mockingbird (1962) 4.3478260869565215

From the above, we can say that Males give movies higher ratings than women.

The type of movies that males and females like are different as can be seen from above.

Avg rating according to Genre,

Popularity of Action genre in males and females,

select users.gender, avg(rating) as avg

from rating

join movies on

rating.itemid=movies.movieid

join users on rating.userid=users.userid

group by users.gender,movies.action

having movies.action=1

order by avg desc

limit 30;

F 3.5

M 3.125

Popularity of comedy genre in males and females,

select users.gender, avg(rating) as avg

from rating

join movies on

rating.itemid=movies.movieid

join users on rating.userid=users.userid

group by users.gender,movies.comedy

having movies.comedy=1

order by avg desc

limit 30;

F 3.4269713261648747

M 3.32

Popularity of romance genre in males and females

select users.gender, avg(rating) as avg

from rating

join movies on

rating.itemid=movies.movieid

join users on rating.userid=users.userid

group by users.gender,movies.romance

having movies.romance=1

order by avg desc

limit 30;

M 3.6642075807682524

F 3.5601217656012176

Similarly, we can find for other genres.

d)

SELECT users.occupation, avg(rating) as avg

FROM Users JOIN Rating ON users.userid=rating.userid

GROUP BY users.occupation

ORDER BY avg DESC

LIMIT 30;

none 3.779134295227525

lawyer 3.7353159851301116

doctor 3.688888888888889

educator 3.6706206312221985

artist 3.653379549393414

administrator 3.6356464768017114

scientist 3.611273080660836

salesman 3.582943925233645

programmer 3.5682604794257147

librarian 3.560781338896264

other 3.5523773797242804

engineer 3.541406727828746

technician 3.5322304620650313

student 3.5151432345038027

marketing 3.4856410256410255

retired 3.4667495338719703

entertainment 3.4410501193317424

writer 3.3757225433526012

executive 3.3491037320011756

homemaker 3.301003344481605

healthcare 2.896219686162625

e)

**By Average Rating**:

﻿

CREATE TABLE Temp2 (movieTitle STRING,genres INT,avgRating INT)ROW FORMAT DELIMITED

> FIELDS TERMINATED BY '|';

INSERT INTO Temp2 SELECT movies.movietitle, movies.Action + movies.Adventure + movies.Animation + movies.Children + movies.Comedy + movies.Crime + movies.Documentary + movies.Drama + movies.Fantasy +movies.FilmNoir+ movies.Horror + movies.Musical + movies.Mystery + movies.Romance + movies.SciFi + movies.Thriller + movies.War + movies.Western

AS genre , avg(rating.rating) AS avgrating

FROM movies JOIN Rating ON rating.itemid=movies.movieid

GROUP BY movies.movietitle, movies.Action + movies.Adventure + movies.Animation + movies.Children + movies.Comedy + movies.Crime + movies.Documentary + movies.Drama + movies.Fantasy +movies.FilmNoir+ movies.Horror + movies.Musical + movies.Mystery + movies.Romance + movies.SciFi + movies.Thriller + movies.War + movies.Western

'Til There Was You (1997) 2 9

1-900 (1994) 1 5

101 Dalmatians (1996) 2 109

12 Angry Men (1957) 1 125

187 (1997) 1 41

2 Days in the Valley (1996) 1 93

20,000 Leagues Under the Sea (1954) 4 72

2001: A Space Odyssey (1968) 4 259

3 Ninjas: High Noon At Mega Mountain (1998) 2 5

39 Steps, The (1935) 1 59

8 1/2 (1963) 1 38

8 Heads in a Duffel Bag (1997) 1 4

8 Seconds (1994) 1 4

A Chef in Love (1996) 1 8

Above the Rim (1994) 1 5

Absolute Power (1997) 2 127

Abyss, The (1989) 4 151

Ace Ventura: Pet Detective (1994) 1 103

Ace Ventura: When Nature Calls (1995) 1 37

Across the Sea of Time (1995) 1 4

Addams Family Values (1993) 1 87

Addicted to Love (1997) 2 54

Addiction, The (1995) 1 11

Adventures of Pinocchio, The (1996) 2 39

Adventures of Priscilla, Queen of the Desert, The (1994) 2 111

Adventures of Robin Hood, The (1938) 2 67

Affair to Remember, An (1957) 1 26

African Queen, The (1951) 4 152

Afterglow (1997) 2 18

Age of Innocence, The (1993) 1 65

SELECT movietitle ,corr(Temp2.genres,Temp2.avgRating) AS CORR From Temp2 GROUP BY movietitle;

**By Count:**

CREATE TABLE Temp1 (movieTitle STRING,genres INT, Ratingcount INT)ROW FORMAT DELIMITED

> FIELDS TERMINATED BY '|';

INSERT INTO Temp1 SELECT movies.movietitle, movies.Action + movies.Adventure + movies.Animation + movies.Children + movies.Comedy + movies.Crime + movies.Documentary + movies.Drama + movies.Fantasy +movies.FilmNoir+ movies.Horror + movies.Musical + movies.Mystery + movies.Romance + movies.SciFi + movies.Thriller + movies.War + movies.Western

AS genre , count(rating.itemid) AS Ratingcount

FROM movies JOIN Rating ON rating.itemid=movies.movieid

GROUP BY movies.movietitle, movies.Action + movies.Adventure + movies.Animation + movies.Children + movies.Comedy + movies.Crime + movies.Documentary + movies.Drama + movies.Fantasy +movies.FilmNoir+ movies.Horror + movies.Musical + movies.Mystery + movies.Romance + movies.SciFi + movies.Thriller + movies.War + movies.Western

﻿

'Til There Was You (1997) 2 9

1-900 (1994) 1 5

101 Dalmatians (1996) 2 109

12 Angry Men (1957) 1 125

187 (1997) 1 41

2 Days in the Valley (1996) 1 93

20,000 Leagues Under the Sea (1954) 4 72

2001: A Space Odyssey (1968) 4 259

3 Ninjas: High Noon At Mega Mountain (1998) 2 5

39 Steps, The (1935) 1 59

8 1/2 (1963) 1 38

8 Heads in a Duffel Bag (1997) 1 4

8 Seconds (1994) 1 4

A Chef in Love (1996) 1 8

Above the Rim (1994) 1 5

Absolute Power (1997) 2 127

Abyss, The (1989) 4 151

Ace Ventura: Pet Detective (1994) 1 103

Ace Ventura: When Nature Calls (1995) 1 37

Across the Sea of Time (1995) 1 4

Addams Family Values (1993) 1 87

Addicted to Love (1997) 2 54

Addiction, The (1995) 1 11

Adventures of Pinocchio, The (1996) 2 39

Adventures of Priscilla, Queen of the Desert, The (1994) 2 111

Adventures of Robin Hood, The (1938) 2 67

Affair to Remember, An (1957) 1 26

African Queen, The (1951) 4 152

Afterglow (1997) 2 18

Age of Innocence, The (1993) 1 65

SELECT movietitle ,corr(Temp1.genres,Temp1.Ratingcount) AS CORR From Temp2 GROUP BY movietitle;

**Question 3:**

1)

> show dbs

admin 0.000GB

config 0.000GB

local 0.000GB

> use local

switched to db local

> use test

switched to db test

> mongoimport --db test --collection business --drop --file "/Users/swapnilvermani/Downloads/business.json"

mongoimport --db test --collection business2 --drop --file "/Users/swapnilvermani/Downloads/business.json"

2)

> db.business2.stats().count

61184

3)

db.business2.aggregate([{$group : {\_id:"$state", count:{$sum:1}}}])

{ "\_id" : "OR", "count" : 1 }

{ "\_id" : "NW", "count" : 1 }

{ "\_id" : "KHL", "count" : 1 }

{ "\_id" : "MA", "count" : 1 }

{ "\_id" : "RP", "count" : 13 }

{ "\_id" : "BW", "count" : 934 }

{ "\_id" : "XGL", "count" : 1 }

{ "\_id" : "NTH", "count" : 1 }

{ "\_id" : "ELN", "count" : 10 }

{ "\_id" : "MLN", "count" : 123 }

{ "\_id" : "HAM", "count" : 1 }

{ "\_id" : "FIF", "count" : 4 }

{ "\_id" : "EDH", "count" : 2971 }

{ "\_id" : "SC", "count" : 189 }

{ "\_id" : "QC", "count" : 3921 }

{ "\_id" : "ON", "count" : 351 }

{ "\_id" : "CA", "count" : 3 }

{ "\_id" : "IL", "count" : 627 }

{ "\_id" : "WA", "count" : 1 }

{ "\_id" : "NV", "count" : 16485 }

4)

db.business2.find({"stars":"5"})

db.business2.find({"stars":5},{name:1,city:1,state:1,review\_count:1}).pretty()

{

"\_id" : ObjectId("5b91839bbbc11675198f762f"),

"city" : "Carnegie",

"review\_count" : 5,

"name" : "Alteration World",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7649"),

"city" : "Homestead",

"review\_count" : 3,

"name" : "American Buyers Discount Mattress Store",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7661"),

"city" : "West Mifflin",

"review\_count" : 4,

"name" : "Hunan Wok Chinese Restaurant",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7663"),

"city" : "McKeesport",

"review\_count" : 7,

"name" : "Minerva Bakery",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7672"),

"city" : "Bellevue",

"review\_count" : 3,

"name" : "Vivo",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7679"),

"city" : "Pittsburgh",

"review\_count" : 5,

"name" : "Paul's Clock Repair",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f767a"),

"city" : "Pittsburgh",

"review\_count" : 3,

"name" : "Brunos Garage",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f767f"),

"city" : "Pittsburgh",

"review\_count" : 3,

"name" : "Broadway Auto Parts",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7693"),

"city" : "Pittsburgh",

"review\_count" : 6,

"name" : "Ace Lock",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7695"),

"city" : "Pittsburgh",

"review\_count" : 64,

"name" : "Walter's Automotive",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7698"),

"city" : "Pittsburgh",

"review\_count" : 5,

"name" : "Hairdressers II",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7699"),

"city" : "Pittsburgh",

"review\_count" : 24,

"name" : "Food Glorious Food",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f769a"),

"city" : "Pittsburgh",

"review\_count" : 3,

"name" : "Carnegie Library of Pittsburgh - East Liberty",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f76a9"),

"city" : "Pittsburgh",

"review\_count" : 3,

"name" : "Block Harvey M DMD",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f76ca"),

"city" : "Pittsburgh",

"review\_count" : 7,

"name" : "North Shore Deli",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f76d0"),

"city" : "Pittsburgh",

"review\_count" : 4,

"name" : "Heinz USA",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f76dc"),

"city" : "Pittsburgh",

"review\_count" : 29,

"name" : "Children's Museum of Pittsburgh",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f76e7"),

"city" : "Pittsburgh",

"review\_count" : 3,

"name" : "Three Dogs & A Cat Grooming Salon",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f76e8"),

"city" : "Pittsburgh",

"review\_count" : 6,

"name" : "Tom Friday's Market",

"state" : "PA"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f76f7"),

"city" : "Pittsburgh",

"review\_count" : 5,

"name" : "West Liberty Cycles",

"state" : "PA"

}

5)

> db.business2.find({"state":"NC"},{name:1,city:1,state:1}).pretty()

{

"\_id" : ObjectId("5b91839bbbc11675198f788a"),

"city" : "Belmont",

"name" : "Carolina Auto Repair",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f788b"),

"city" : "Belmont",

"name" : "Wendy's",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f788c"),

"city" : "Belmont",

"name" : "Demetrios Restaurant",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f788d"),

"city" : "Belmont",

"name" : "Pizza Hut",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f788e"),

"city" : "Belmont",

"name" : "White's Restaurant",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f788f"),

"city" : "Belmont",

"name" : "Textile Lanes",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7890"),

"city" : "Belmont",

"name" : "Bi-Lo",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7891"),

"city" : "Belmont",

"name" : "Captain's Cap of Belmont Number 2",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7892"),

"city" : "Belmont",

"name" : "Catawba Fountain & Grill",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7893"),

"city" : "Belmont",

"name" : "Sub Corral Sandwich Shop",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7894"),

"city" : "Belmont",

"name" : "Belmont Post Office",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7895"),

"city" : "Belmont",

"name" : "McDonald's",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7896"),

"city" : "Belmont",

"name" : "Taco Bell",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7897"),

"city" : "Belmont",

"name" : "Shooters Express Gun Shop",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7898"),

"city" : "Belmont",

"name" : "New China",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f7899"),

"city" : "Huntersville",

"name" : "Huntersville Family Fitness & Aquatics",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f789a"),

"city" : "Huntersville",

"name" : "New Modern Nails",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f789b"),

"city" : "Huntersville",

"name" : "Latta Equestrian Center",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f789c"),

"city" : "Huntersville",

"name" : "Latta Plantation Park",

"state" : "NC"

}

{

"\_id" : ObjectId("5b91839bbbc11675198f789d"),

"city" : "Huntersville",

"name" : "Historic Latta Plantation",

"state" : "NC"

}

6)

> db.business2.aggregate([{$group : {\_id:"$state", count:{$sum:"$review\_count" }}}])

{ "\_id" : "OR", "count" : 4 }

{ "\_id" : "NW", "count" : 5 }

{ "\_id" : "KHL", "count" : 8 }

{ "\_id" : "MA", "count" : 4 }

{ "\_id" : "RP", "count" : 75 }

{ "\_id" : "BW", "count" : 8750 }

{ "\_id" : "XGL", "count" : 3 }

{ "\_id" : "NTH", "count" : 17 }

{ "\_id" : "ELN", "count" : 51 }

{ "\_id" : "MLN", "count" : 934 }

{ "\_id" : "HAM", "count" : 3 }

{ "\_id" : "FIF", "count" : 36 }

{ "\_id" : "EDH", "count" : 34346 }

{ "\_id" : "SC", "count" : 2419 }

{ "\_id" : "QC", "count" : 54569 }

{ "\_id" : "ON", "count" : 3098 }

{ "\_id" : "CA", "count" : 43 }

{ "\_id" : "IL", "count" : 13159 }

{ "\_id" : "WA", "count" : 9 }

{ "\_id" : "NV", "count" : 752904 }

7)

> db.business2.aggregate([

{$group : {\_id:"$state", avg\_rating:{$avg:"$stars"}}}

])

{ "\_id" : "OR", "avg\_rating" : 3 }

{ "\_id" : "NW", "avg\_rating" : 5 }

{ "\_id" : "KHL", "avg\_rating" : 3.5 }

{ "\_id" : "MA", "avg\_rating" : 5 }

{ "\_id" : "RP", "avg\_rating" : 4.230769230769231 }

{ "\_id" : "BW", "avg\_rating" : 3.766595289079229 }

{ "\_id" : "XGL", "avg\_rating" : 2.5 }

{ "\_id" : "NTH", "avg\_rating" : 2 }

{ "\_id" : "ELN", "avg\_rating" : 3.6 }

{ "\_id" : "MLN", "avg\_rating" : 3.8292682926829267 }

{ "\_id" : "HAM", "avg\_rating" : 4.5 }

{ "\_id" : "FIF", "avg\_rating" : 3.5 }

{ "\_id" : "EDH", "avg\_rating" : 3.7931672837428474 }

{ "\_id" : "SC", "avg\_rating" : 3.5767195767195767 }

{ "\_id" : "QC", "avg\_rating" : 3.6893649579188983 }

{ "\_id" : "ON", "avg\_rating" : 3.5868945868945867 }

{ "\_id" : "CA", "avg\_rating" : 3.3333333333333335 }

{ "\_id" : "IL", "avg\_rating" : 3.569377990430622 }

{ "\_id" : "WA", "avg\_rating" : 5 }

{ "\_id" : "NV", "avg\_rating" : 3.6584774037003336 }

8)

By stars :

> db.business2.aggregate([

... {$group : {\_id:"$state",sum\_stars:{$sum:"$stars"}}},

... {$sort : {sum\_stars :-1}},

... {$limit :10}

... ])

{ "\_id" : "AZ", "sum\_stars" : 92960 }

{ "\_id" : "NV", "sum\_stars" : 60310 }

{ "\_id" : "NC", "sum\_stars" : 17811.5 }

{ "\_id" : "QC", "sum\_stars" : 14466 }

{ "\_id" : "EDH", "sum\_stars" : 11269.5 }

{ "\_id" : "PA", "sum\_stars" : 11200 }

{ "\_id" : "WI", "sum\_stars" : 8406 }

{ "\_id" : "BW", "sum\_stars" : 3518 }

{ "\_id" : "IL", "sum\_stars" : 2238 }

{ "\_id" : "ON", "sum\_stars" : 1259 }

By reviews:

> db.business2.aggregate([

... ... {$group :{\_id:"$state",sum\_reviews:{$sum:"$review\_count"}}},

... ... {$sort : {sum\_reviews :-1}},

... ... {$limit :10}

... ... ])

{ "\_id" : "NV", "sum\_reviews" : 752904 }

{ "\_id" : "AZ", "sum\_reviews" : 636779 }

{ "\_id" : "NC", "sum\_reviews" : 102495 }

{ "\_id" : "PA", "sum\_reviews" : 72409 }

{ "\_id" : "QC", "sum\_reviews" : 54569 }

{ "\_id" : "WI", "sum\_reviews" : 47684 }

{ "\_id" : "EDH", "sum\_reviews" : 34346 }

{ "\_id" : "IL", "sum\_reviews" : 13159 }

{ "\_id" : "BW", "sum\_reviews" : 8750 }

{ "\_id" : "ON", "sum\_reviews" : 3098 }