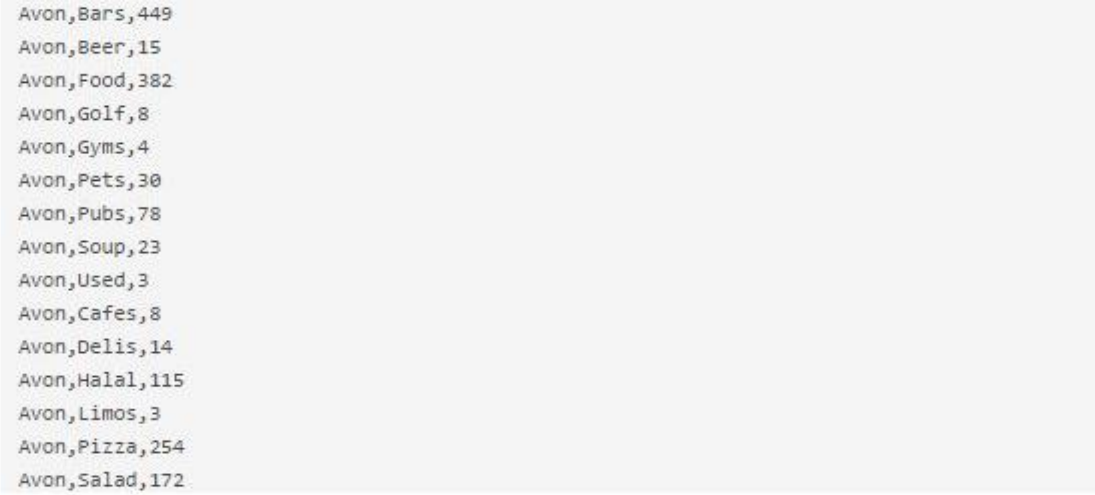
Nishchal Nagar
Nn1123

Hadoop: Map/Reduce with Pig on Yelp dataset

1. Summarize the number of reviews by US city, by business category

i. Loaded *yelp_academic_dataset_business.json*
ii. Generated Custom Data view including name, latitude, longitude, city, state, number of reviews and flattened categories.
iii. Filtered the above table according to US cities latitude and longitude values and state not in Canada
iv. Group the table by city and categories
v. In the final output , generated city , categories and sum of reviews
vi. Stored data

Data generated :

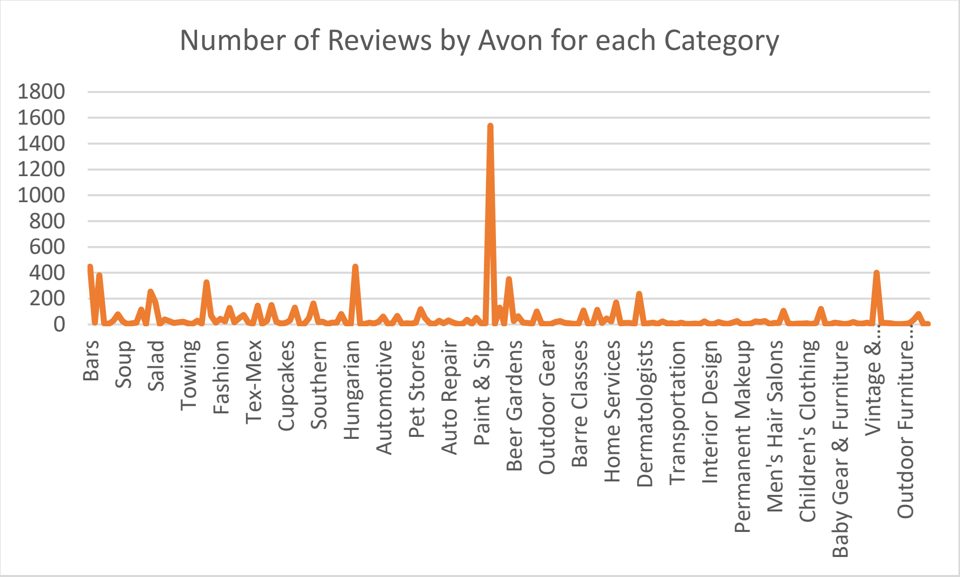


```
Avon,Bars,449
Avon,Beer,15
Avon,Food,382
Avon,Golf,8
Avon,Gyms,4
Avon,Pets,30
Avon,Pubs,78
Avon,Soup,23
Avon,Used,3
Avon,Cafes,8
Avon,Delis,14
Avon,Halal,115
Avon,Limos,3
Avon,Pizza,254
Avon,Salad,172
```

Figure 1: Sample output

Visualisation :

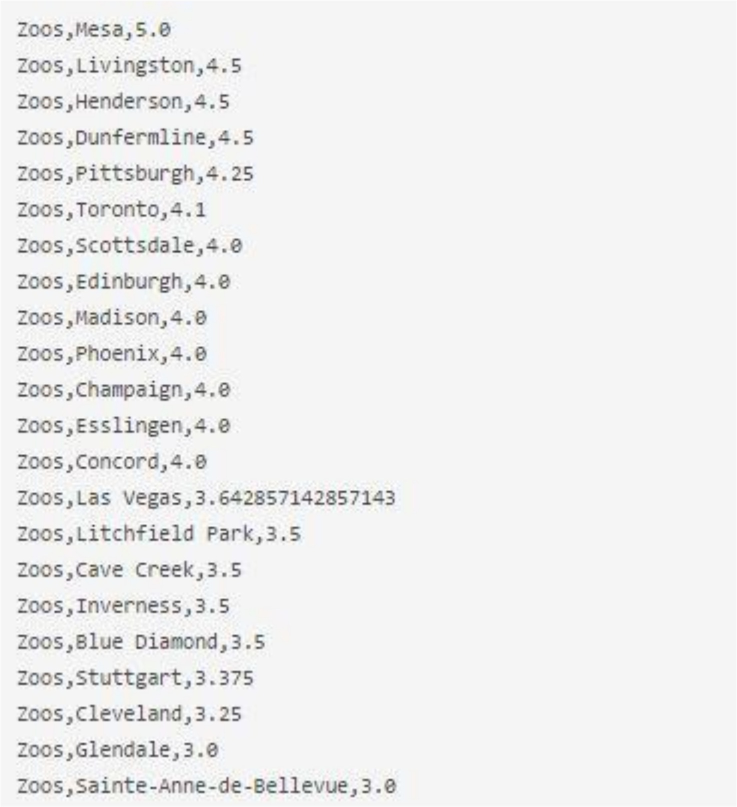
Below is graph for Avon . Y axis : Number of reviews ; X-axis: categories



2. Rank all ***cities*** by # of stars descending, for **each category**

i. Loaded *yelp_academic_dataset_business.json*
ii. Generated Custom Data view including name, city, number of reviews ,city, number of reviews, stars and flattened categories.
iii. Group the table by city and categories
iv. Computed average rating for each category in grouped data
v. Ordered data by category and rating in descending order.
vi. Stored data

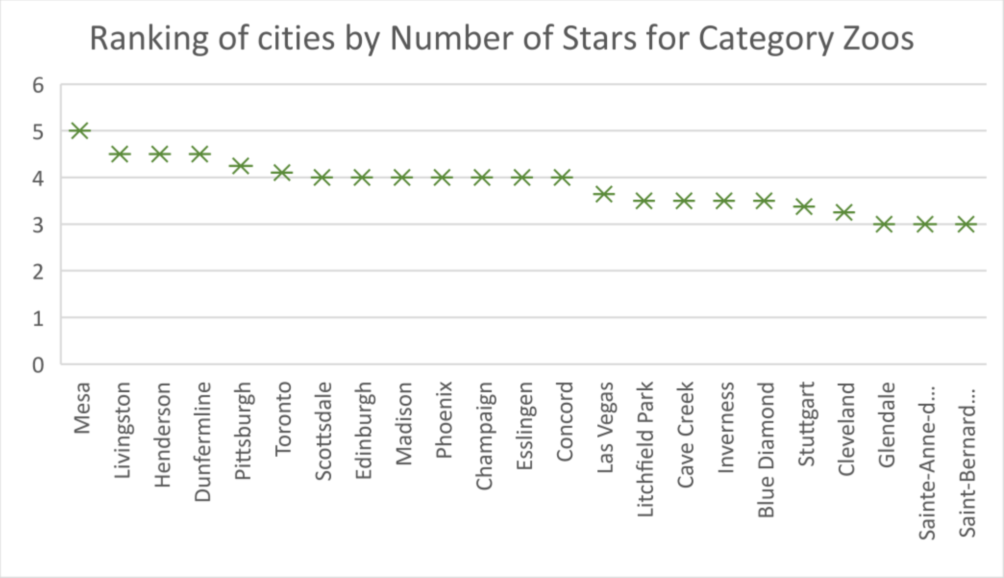
Data generated :



```
Zoos,Mesa,5.0
Zoos,Livingston,4.5
Zoos,Henderson,4.5
Zoos,Dunfermline,4.5
Zoos,Pittsburgh,4.25
Zoos,Toronto,4.1
Zoos,Scottsdale,4.0
Zoos,Edinburgh,4.0
Zoos,Madison,4.0
Zoos,Phoenix,4.0
Zoos,Champaign,4.0
Zoos,Esslingen,4.0
Zoos,Concord,4.0
Zoos,Las Vegas,3.642857142857143
Zoos,Litchfield Park,3.5
Zoos,Cave Creek,3.5
Zoos,Inverness,3.5
Zoos,Blue Diamond,3.5
Zoos,Stuttgart,3.375
Zoos,Cleveland,3.25
Zoos,Glendale,3.0
Zoos,Sainte-Anne-de-Bellevue,3.0
```

Figure 2: Cities ranked for Zoos category

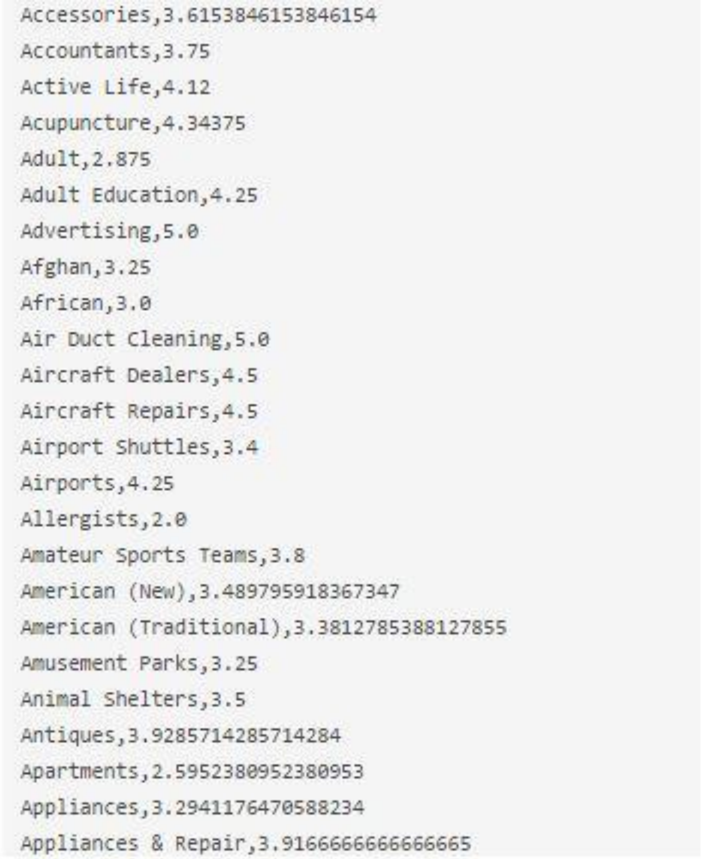
Visualization:



3. What is the average rank (# stars) for businesses within 10 miles of the University of Wisconsin - Madison, by type of business?

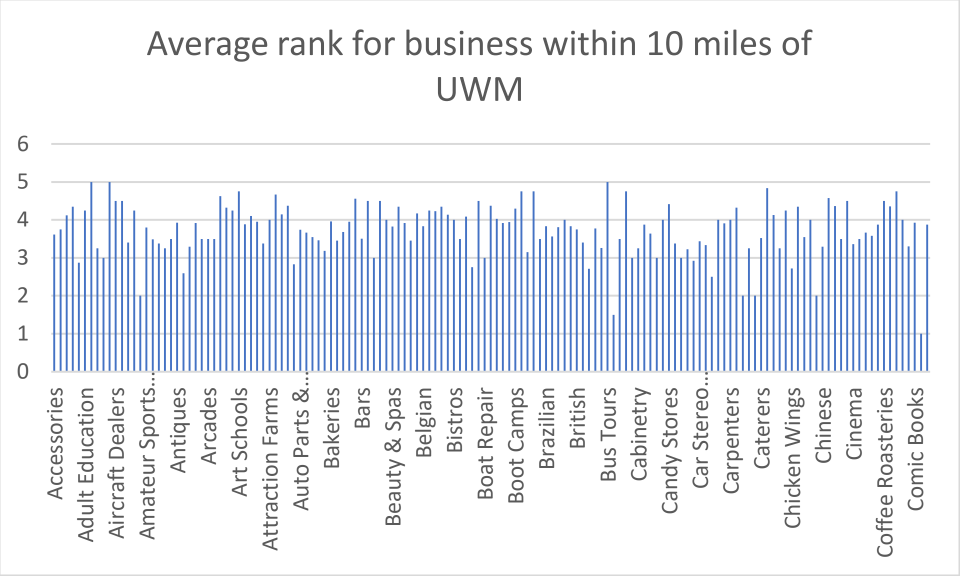
   i. Loaded *yelp_academic_dataset_business.json*
   ii. Generated Custom Data view including latitude, longitude, stars and flattened categories.
   iii. Filter the table by latitude and longitude which are +- 10 minutes of UWM
   iv. Group data by categories
   v. Computed average rating for each category in grouped data
   vi. Ordered data by category.
   vii. Stored data

Ouput generated:



```
Accessories,3.6153846153846154
Accountants,3.75
Active Life,4.12
Acupuncture,4.34375
Adult,2.875
Adult Education,4.25
Advertising,5.0
Afghan,3.25
African,3.0
Air Duct Cleaning,5.0
Aircraft Dealers,4.5
Aircraft Repairs,4.5
Airport Shuttles,3.4
Airports,4.25
Allergists,2.0
Amateur Sports Teams,3.8
American (New),3.489795918367347
American (Traditional),3.3812785388127855
Amusement Parks,3.25
Animal Shelters,3.5
Antiques,3.9285714285714284
Apartments,2.5952380952380953
Appliances,3.2941176470588234
Appliances & Repair,3.9166666666666665
```

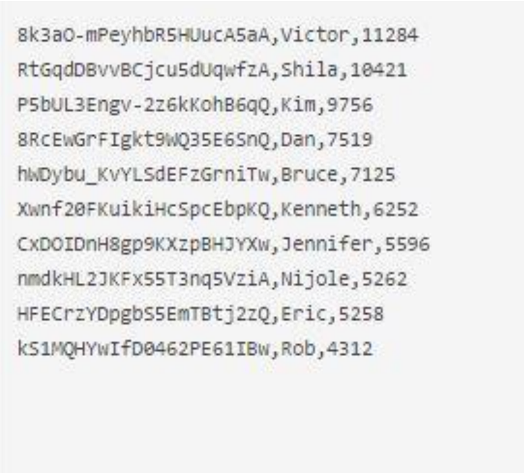
Visualization:


Average rank for business within 10 miles of UWM

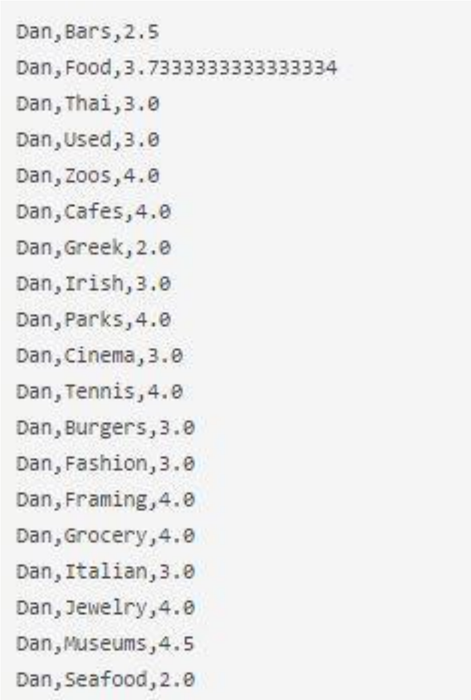
4. Rank reviewers by number of reviews. For the top 10 reviewers, show their average number of stars, by category.

i. Loaded yelp_academic_dataset_review.json , yelp_academic_dataset_user.json and yelp_academic_dataset_business.json
ii. Generated ReviewData including user_id, starRating, businessID
iii. Generate userdata including user_id ,username, numberOfReviews
iv. Order userdata by review count using DESC and limit it to 10 to get top 10
v. Join top10 userdata and reviewData bu userId
vi. Customized joined data to generate username, starRating and businessID
vii. Generated business data including business Id and category
viii. Join business data with previously joined data by businessID
ix. For this joined data , generated username, starRating and Flattened category.
x. Group data by userName and category
xi. In final output , generate Flattened username, category and Average of rating
xii. Stored topTen user data and final output into separate files.

Result:



```
8k3aO-mPeyhbR5HUucA5aA,Victor,11284
RtGqdDBvvBCjcu5dUqwfzA,Shila,10421
P5bUL3Engv-2z6kKohB6qQ,Kim,9756
8RcEwGrFIgkt9WQ35E6SnQ,Dan,7519
hWDybu_KvYLSdEFzGrniTw,Bruce,7125
Xwnf20FKuikiHcSpcEbpKQ,Kenneth,6252
CxDOIDnH8gp9KXzpBHJYXw,Jennifer,5596
nmdkHL2JKFx55T3nq5VziA,Nijole,5262
HFECrzYDpgbS5EmTBtj2zQ,Eric,5258
kS1MQHYwIfD0462PE61IBw,Rob,4312
```

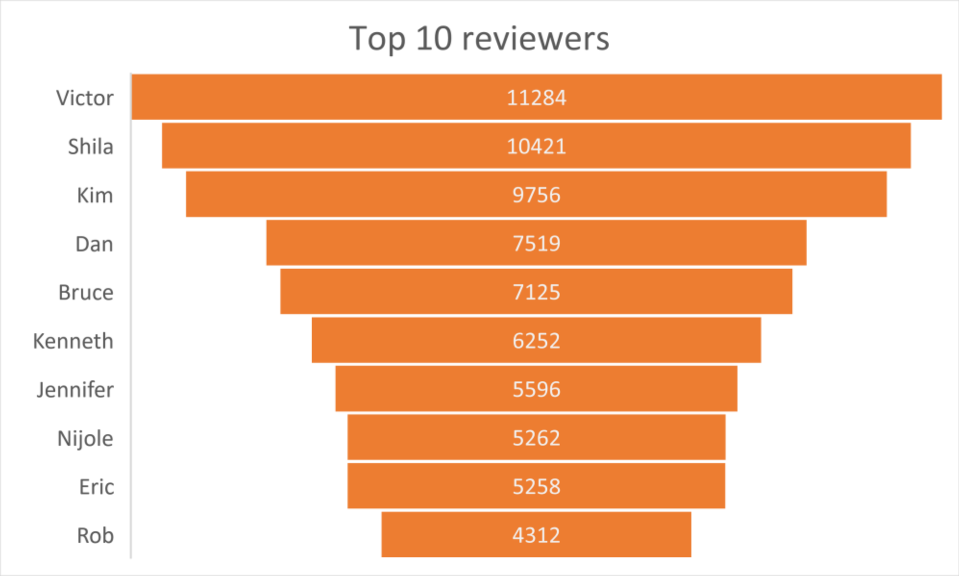


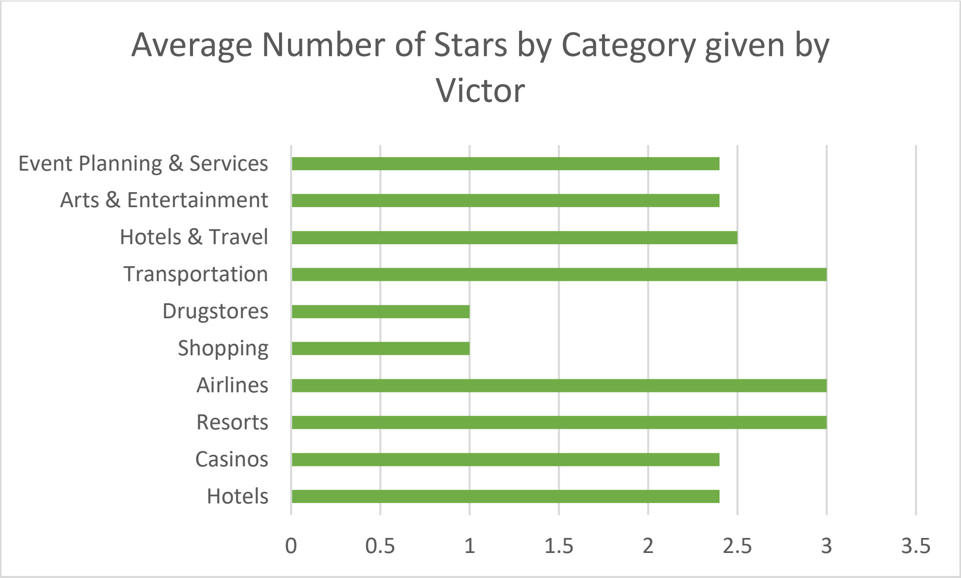
```
Dan,Bars,2.5
Dan,Food,3.7333333333333334
Dan,Thai,3.0
Dan,Used,3.0
Dan,Zoos,4.0
Dan,Cafes,4.0
Dan,Greek,2.0
Dan,Irish,3.0
Dan,Parks,4.0
Dan,Cinema,3.0
Dan,Tennis,4.0
Dan,Burgers,3.0
Dan,Fashion,3.0
Dan,Framing,4.0
Dan,Grocery,4.0
Dan,Italian,3.0
Dan,Jewelry,4.0
Dan,Museums,4.5
Dan,Seafood,2.0
```

Figure 3a: Top 10 reviewers

Figure 3b: Average number of stars by Category of top 10 reviewers

Visualization:





5. For the top 10 and bottom 10 food business near UWM (in terms of stars), summarize star rating for reviews in January through May.

i. Loaded *yelp_academic_dataset_business.json* and *yelp_academic_dataset_review.json*

ii. Generated Business Data view including business_id, name, latitude, longitude, stars, and flattened categories.

iii. Filter the businessData by latitude and longitude which are +- 10 minutes of UWM

iv. Filter the Data by category==’Food’

v. Order the data by rating in descending order and limit 10 tuples and store it as topTenFoodBusiness

vi. Order the data by rating in ascending order and limit 10 tuples and store it as bottomTenFoodBusiness

vii. Generated Review data view including business ID , stars and date as datetime.

viii. Filter reviewData by 1<= GetMonth(date) <=5 and store it as reviewsInJanToMay

ix. Join topTenFoodBusiness and reviewsInJanToMay by businessID

x. Create custom view for top ten including business name, businessID, rating and month

xi. Grouped custom view for top ten by businessID, businessName, month.

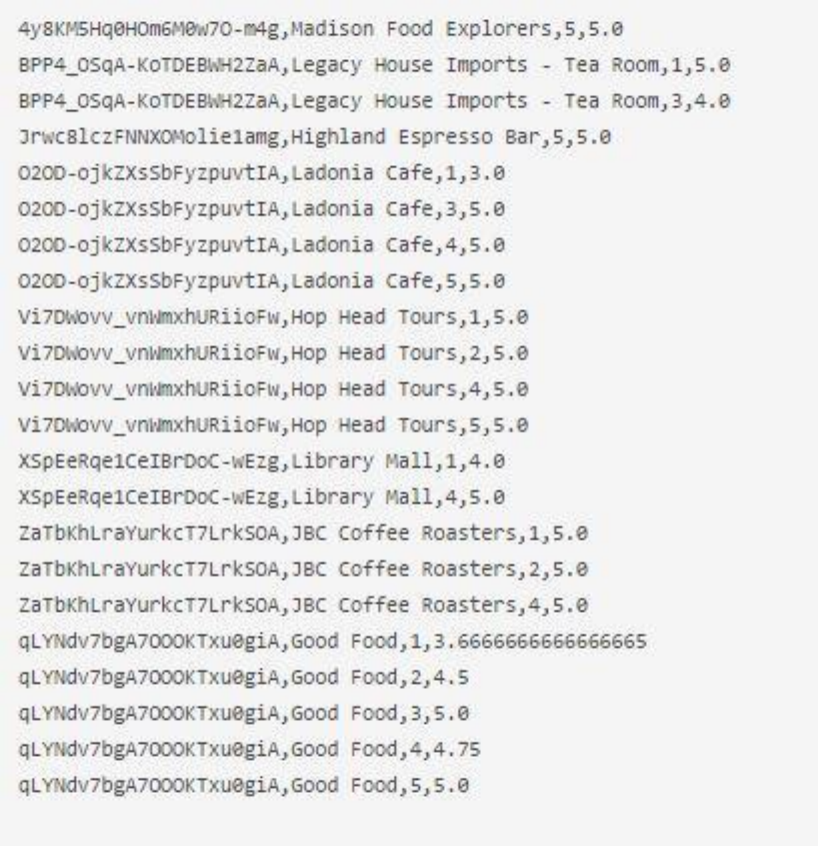
xii. Flattened group and generated flatten group and average of ratings from customView

xiii. Similarly generated for bottom ten business by performing ix to xii for bottomTenFoodBusiness

xiv. Stored final outputs of both in 2 files.

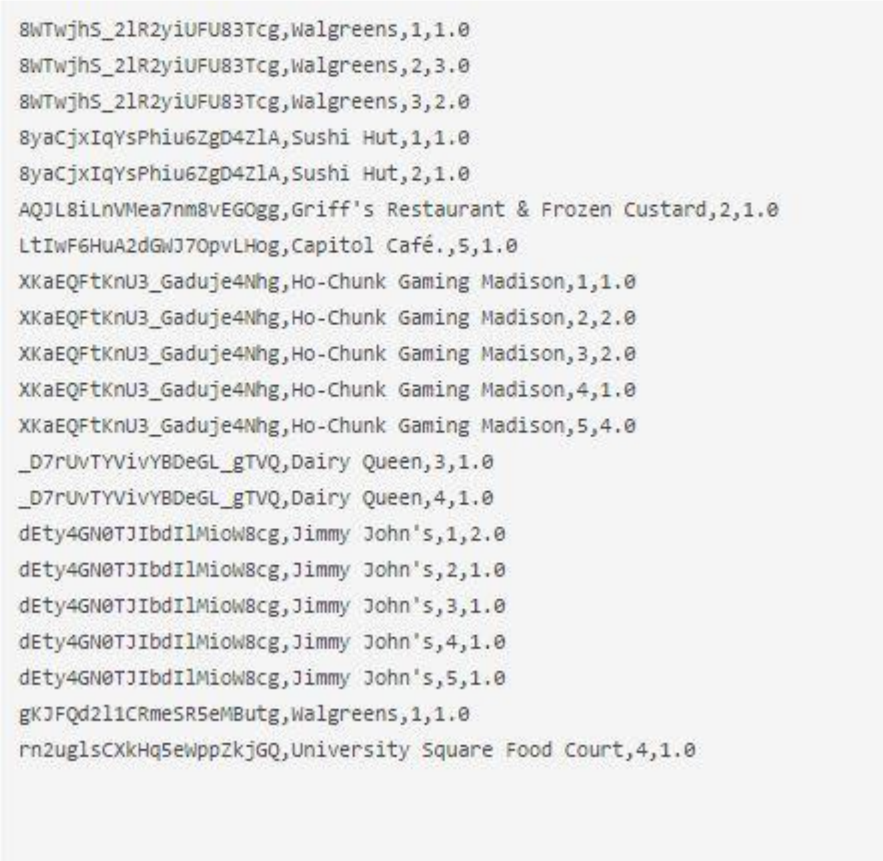
Assumption – Summarize star rating means show average ratings of all selected food business in month of Jan to May

Output generated :



```
4y8KM5Hq0HOm6M0w7O-m4g,Madison Food Explorers,5,5.0
BPP4_OSqA-KoTDEBWH2ZaA,Legacy House Imports - Tea Room,1,5.0
BPP4_OSqA-KoTDEBWH2ZaA,Legacy House Imports - Tea Room,3,4.0
Jrwc8lczFNNXOMolie1amg,Highland Espresso Bar,5,5.0
O2OD-ojkZXsSbFyzpuvtIA,Ladonia Cafe,1,3.0
O2OD-ojkZXsSbFyzpuvtIA,Ladonia Cafe,3,5.0
O2OD-ojkZXsSbFyzpuvtIA,Ladonia Cafe,4,5.0
O2OD-ojkZXsSbFyzpuvtIA,Ladonia Cafe,5,5.0
Vi7DWovv_vnWmxhURiioFw,Hop Head Tours,1,5.0
Vi7DWovv_vnWmxhURiioFw,Hop Head Tours,2,5.0
Vi7DWovv_vnWmxhURiioFw,Hop Head Tours,4,5.0
Vi7DWovv_vnWmxhURiioFw,Hop Head Tours,5,5.0
XSpEeRqe1CeIBrDoC-wEzg,Library Mall,1,4.0
XSpEeRqe1CeIBrDoC-wEzg,Library Mall,4,5.0
ZaTbKhLraYurkcT7LrkSOA,JBC Coffee Roasters,1,5.0
ZaTbKhLraYurkcT7LrkSOA,JBC Coffee Roasters,2,5.0
ZaTbKhLraYurkcT7LrkSOA,JBC Coffee Roasters,4,5.0
qLYNdv7bgA7OOOKTxu0giA,Good Food,1,3.6666666666666665
qLYNdv7bgA7OOOKTxu0giA,Good Food,2,4.5
qLYNdv7bgA7OOOKTxu0giA,Good Food,3,5.0
qLYNdv7bgA7OOOKTxu0giA,Good Food,4,4.75
qLYNdv7bgA7OOOKTxu0giA,Good Food,5,5.0
```

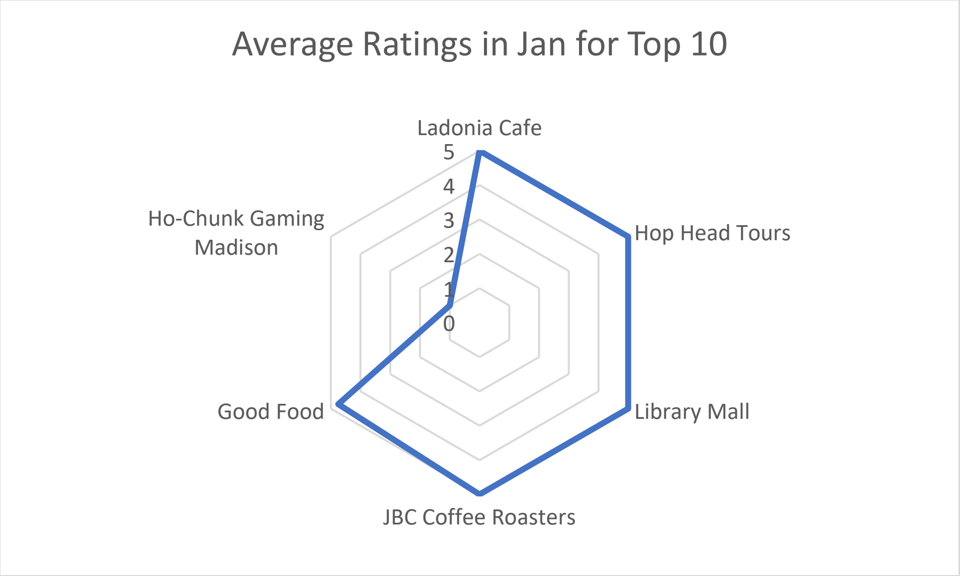
Figure 4: Star rating for Top 10 food business near UWM in month of Jan-May

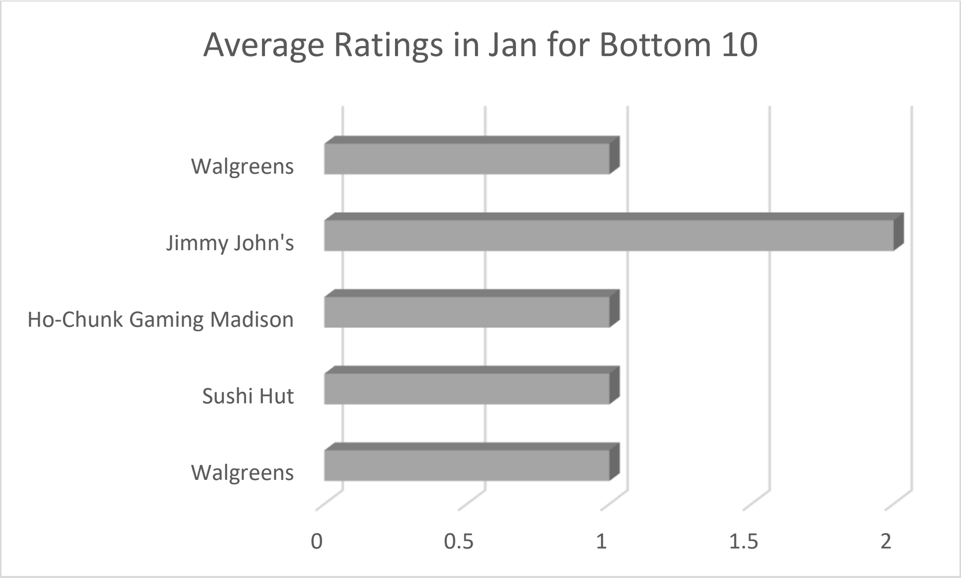


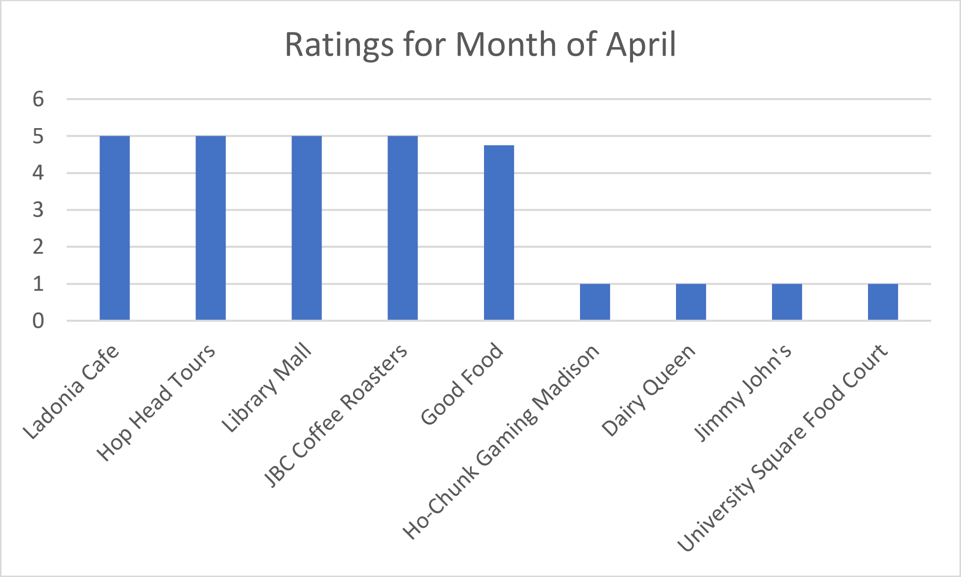
```
8WTwjhS_2lR2yiUFU83Tcg,Walgreens,1,1.0
8WTwjhS_2lR2yiUFU83Tcg,Walgreens,2,3.0
8WTwjhS_2lR2yiUFU83Tcg,Walgreens,3,2.0
8yaCjxIqYsPhiu6ZgD4ZlA,Sushi Hut,1,1.0
8yaCjxIqYsPhiu6ZgD4ZlA,Sushi Hut,2,1.0
AQJL8iLnVMea7nm8vEGOgg,Griff's Restaurant & Frozen Custard,2,1.0
LtIwF6HuA2dGWJ7OpvLHog,Capitol Café.,5,1.0
XKaEQFtKnU3_Gaduje4Nhg,Ho-Chunk Gaming Madison,1,1.0
XKaEQFtKnU3_Gaduje4Nhg,Ho-Chunk Gaming Madison,2,2.0
XKaEQFtKnU3_Gaduje4Nhg,Ho-Chunk Gaming Madison,3,2.0
XKaEQFtKnU3_Gaduje4Nhg,Ho-Chunk Gaming Madison,4,1.0
XKaEQFtKnU3_Gaduje4Nhg,Ho-Chunk Gaming Madison,5,4.0
_D7rUvTYVivYBDeGL_gTVQ,Dairy Queen,3,1.0
_D7rUvTYVivYBDeGL_gTVQ,Dairy Queen,4,1.0
dEty4GN0TJIbdIlMioW8cg,Jimmy John's,1,2.0
dEty4GN0TJIbdIlMioW8cg,Jimmy John's,2,1.0
dEty4GN0TJIbdIlMioW8cg,Jimmy John's,3,1.0
dEty4GN0TJIbdIlMioW8cg,Jimmy John's,4,1.0
dEty4GN0TJIbdIlMioW8cg,Jimmy John's,5,1.0
gKJFQd2l1CRmeSR5eMButg,Walgreens,1,1.0
rn2uglsCXkHq5eWppZkjGQ,University Square Food Court,4,1.0
```

Figure 5: Star rating for Bottom 10 food business near UWM in month of Jan-May

Visualization:




Average Ratings in Jan for Bottom 10
Walgreens
Jimmy John's
Ho-Chunk Gaming Madison
Sushi Hut
Walgreens
0
0.5
1
1.5
2


Ratings for Month of April
6
5
4
3
2
1
0
Ladonia Cafe
Hop Head Tours
Library Mall
JBC Coffee Roasters
Good Food
Ho-Chunk Gaming Madison
Dairy Queen
Jimmy John's
University Square Food Court