A

1) Using the public data tables stackoverflow.users and stackoverflow.posts_questions, please write a query to find the 6 StackOverflow users who match the following criteria.
    a) He or she posted a question with tags including “bigdata” or “database” (see the *tags* field; note that tags field can include more than one tag) that was last active in 2016 (see the *last_activity_date* field). The poster of the question is identified by the *owner_user_id* field.
    b) He or she has a reputation score (see field *reputation*) of at least 200,00

```
SELECT TOP(6)
  Id
FROM Users
WHERE
  Id IN (
        SELECT
          OwnerUserId
        FROM Posts
        WHERE
           Tags LIKE '%<database>%'
          OR Tags LIKE '%<bigdata>%'
          AND YEAR(LastActivityDate)=2016
      )
  AND Reputation >= 200000
;
```

RESULT

Id
73070
18771
106224
18393
114251
9021

A2)For each of the users identified in 1), please calculate the percentage share of favorites (see field *favorite_count*) for each of three tiers defined as follows.

a) Tier 1: Top 3 questions posted by user ranked in terms of favorite count
b) Tier 2: Questions 4 to 10 in terms of favorite count
c) Tier 3: Remaining questions posted by user, if applicable
d)

```
SELECT TOP(6)
  Id
INTO #Ids
FROM Users
WHERE
  Id IN (
        SELECT
          OwnerUserId
        FROM Posts
        WHERE
           Tags LIKE '%<database>%'
          OR Tags LIKE '%<bigdata>%'
          AND YEAR(LastActivityDate)=2016
      )
  AND Reputation >= 200000
;

DECLARE @SmofCnts float;
DECLARE @T1Cnts float;
DECLARE @T2Cnts float;
DECLARE @T3Cnts float;
Declare @Id int;
CREATE TABLE #Results (user_id int, tier varchar(5), [share] float);

While (Select Count(*) From #Ids) > 0
Begin

  SET @Id = (SELECT TOP(1) Id FROM #Ids);

  SET @SmofCnts = (
              SELECT SUM(FavoriteCount)
              FROM Posts
              WHERE OwnerUserId=@Id
              );

  SET @T1Cnts = (
                SELECT SUM(FavoriteCount)
                FROM (
                    SELECT TOP(3) FavoriteCount
                    FROM Posts
                    WHERE OwnerUserId=@Id ORDER BY FavoriteCount DESC
```

```
            ) AS T1
  );

  INSERT #Results
  SELECT @Id,'Tier1', ISNULL(@T1Cnts/@SmofCnts,0);

  SET @T2Cnts = (
            SELECT SUM(FavoriteCount)
                        FROM (
                                    SELECT
                        ROW_NUMBER() OVER(ORDER BY
FavoriteCount DESC) AS Row#,
                        FavoriteCount
                        FROM Posts
                        WHERE OwnerUserId=@Id
                        ) as T2
                        WHERE Row# between 4 and 10
        );
        INSERT #Results
        SELECT @Id,'Tier2', ISNULL(@T2Cnts/@SmofCnts,0);

    SET @T3Cnts = (
            SELECT SUM(FavoriteCount)
                        FROM (
                                    SELECT
                        ROW_NUMBER() OVER(ORDER BY
FavoriteCount DESC) AS Row#,
                        FavoriteCount
                        FROM Posts
                        WHERE OwnerUserId=@Id
                        ) as T2
                        WHERE Row# > 10
        );
        INSERT #Results
        SELECT @Id,'Tier3', ISNULL(@T3Cnts/@SmofCnts,0);

  DELETE #Ids WHERE Id = @Id;
END

SELECT * from #Results;
```

RESULT

user_d	tier	share
148870	Tier1	1

148870	Tier2	0
148870	Tier3	0
28804	Tier1	0.776435045317221
28804	Tier2	0.129909365558912
28804	Tier3	0.093655589123867
3043	Tier1	0.258517034068136
3043	Tier2	0.344689378757515
3043	Tier3	0.396793587174349
1583	Tier1	0.923251847640705
1583	Tier2	0.05287094940307
1583	Tier3	0.023877202956225
31671	Tier1	0.22874251497006
31671	Tier2	0.212574850299401
31671	Tier3	0.558682634730539
3153	Tier1	0.493112947658402
3153	Tier2	0.216942148760331
3153	Tier3	0.289944903581267

A3)

```
dat <- read.table(text="user_d tier share
148870 Tier1 1
```

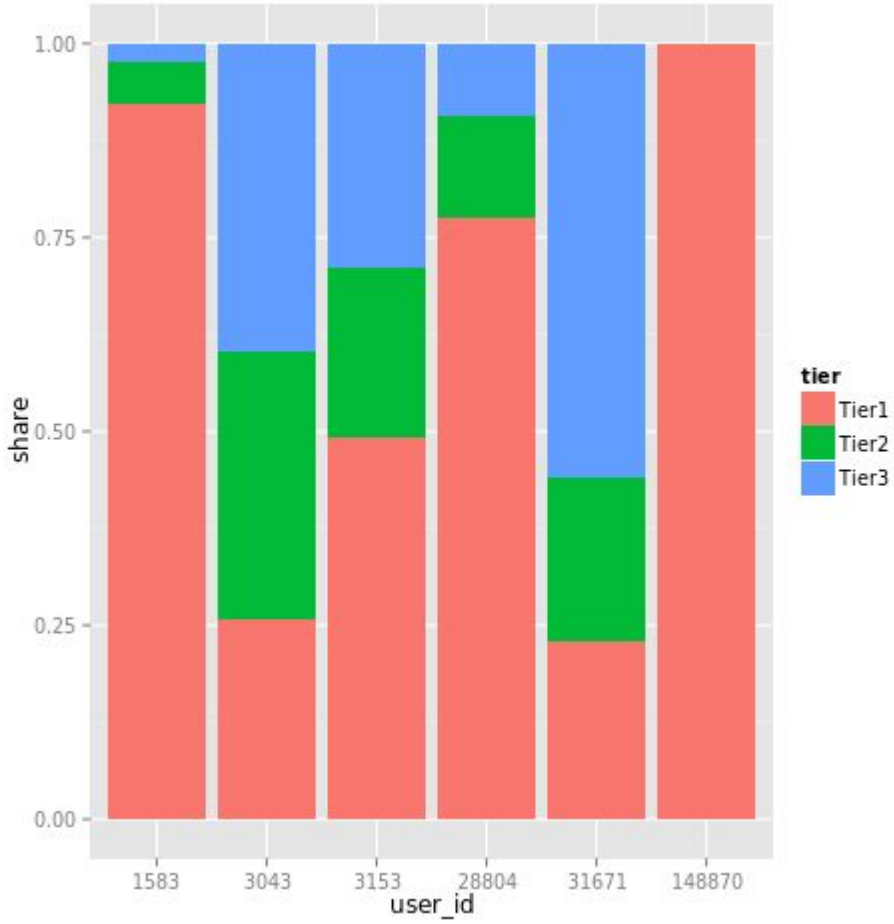
148870 Tier2 0
148870 Tier3 0
28804 Tier1 0.776435045317221
28804 Tier2 0.129909365558912
28804 Tier3 0.093655589123867
3043 Tier1 0.258517034068136
3043 Tier2 0.344689378757515
3043 Tier3 0.396793587174349
1583 Tier1 0.923251847640705
1583 Tier2 0.05287094940307
1583 Tier3 0.023877202956225
31671 Tier1 0.22874251497006
31671 Tier2 0.212574850299401
31671 Tier3 0.558682634730539
3153 Tier1 0.493112947658402
3153 Tier2 0.216942148760331
3153 Tier3 0.289944903581267",header=TRUE)

```
require('ggplot2')

dat$user_id <- as.factor(dat$user_id) # user_id as factor variable

ggplot(dat, aes(x = user_id, y = share, fill = tier)) +
  geom_bar(stat = 'identity')
```

a)



b) The 3 questions with the highest favorite count seem to have the greater share of the favorite count for many users.