**Facebook Data Analysis**

**1.Sanity check(using spark 2):**

**Code:**

from pyspark.sql import SparkSession

from pyspark.sql import Row

from pyspark.sql import functions

def parseInput(line):

fields = line.split(',')

return Row(value = str(fields[i]))

if \_\_name\_\_ == "\_\_main\_\_":

# Create a SparkSession (the config bit is only for Windows!)

spark = SparkSession.builder.appName("SanityCheck").getOrCreate()

# Get the raw data

lines = spark.sparkContext.textFile("hdfs:///tmp/facebook\_data/pseudo\_facebook.csv")

a=["userid","age","dob\_day","dob\_year","dob\_month","gender","tenure","friend\_count","friendships\_initiated","likes","likes\_received","mobile\_likes","mobile\_likes\_received","www\_likes",$

for i in range(15):

# Convert it to a RDD of Row objects with (value)

x = lines.map(parseInput)

# Convert that to a DataFrame

xDF = spark.createDataFrame(x)

# Compute count of Null Values

counts = xDF.filter(xDF["value"]=="NA").count()

# Print them out

print ("%s : %d"%(a[i],counts))

# Stop the session

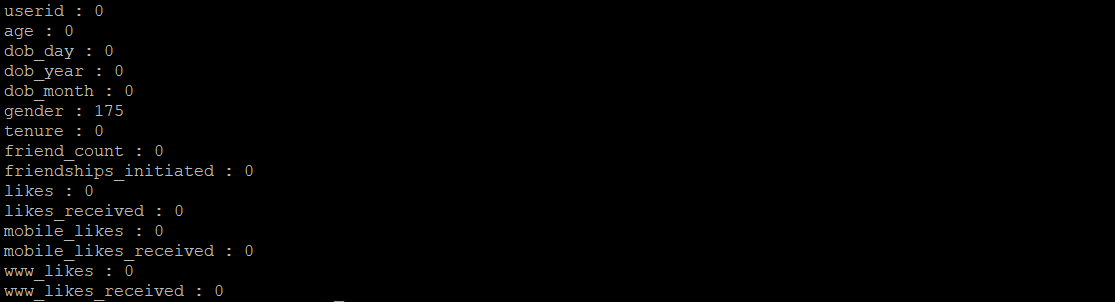
spark.stop()

**Command:**

export SPARK\_MAJOR\_VERSION=2

spark-submit SanityCheck.py

**Output:**



**Observation:** Gender has null values, we should not delete these as users might have kept it blank .

2: Facebook popularity based on ages(Using Mapreduce (python language))

**Code:**

from mrjob.job import MRJob

from mrjob.step import MRStep

class WhatAgeUsesFacebook(MRJob):

def steps(self):

return [

MRStep(mapper=self.mapper\_get\_ages,

reducer=self.reducer\_count\_ages),

MRStep(reducer=self.reducer\_sorted\_output)

]

def mapper\_get\_ages(self, \_, line):

(userid, age, dob\_day, dob\_year, dob\_month, gender, tenure, friend\_count, friendships\_initiated, likes, likes\_received, mobile\_likes, mobile\_likes\_received, www\_likes, www\_likes\_receved) = line.split(',')

yield age, 1

def reducer\_count\_ages(self, age, ones):

yield str(sum(ones)).zfill(5), age

def reducer\_sorted\_output(self, count, ages):

for age in ages:

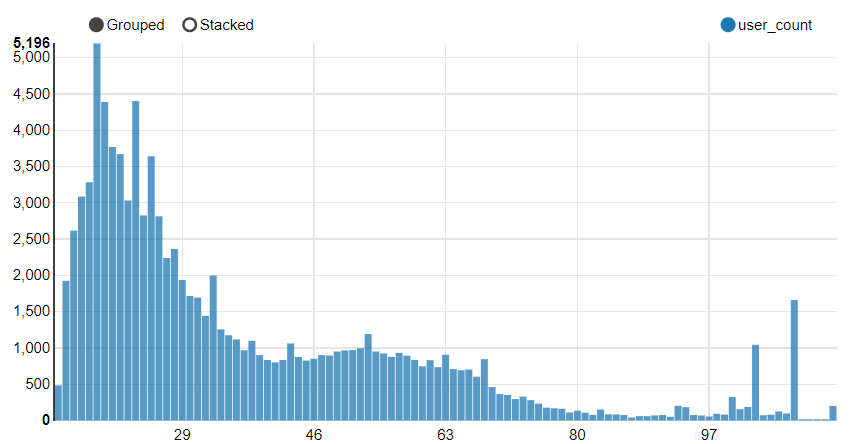
yield age, count

if \_\_name\_\_ == '\_\_main\_\_':

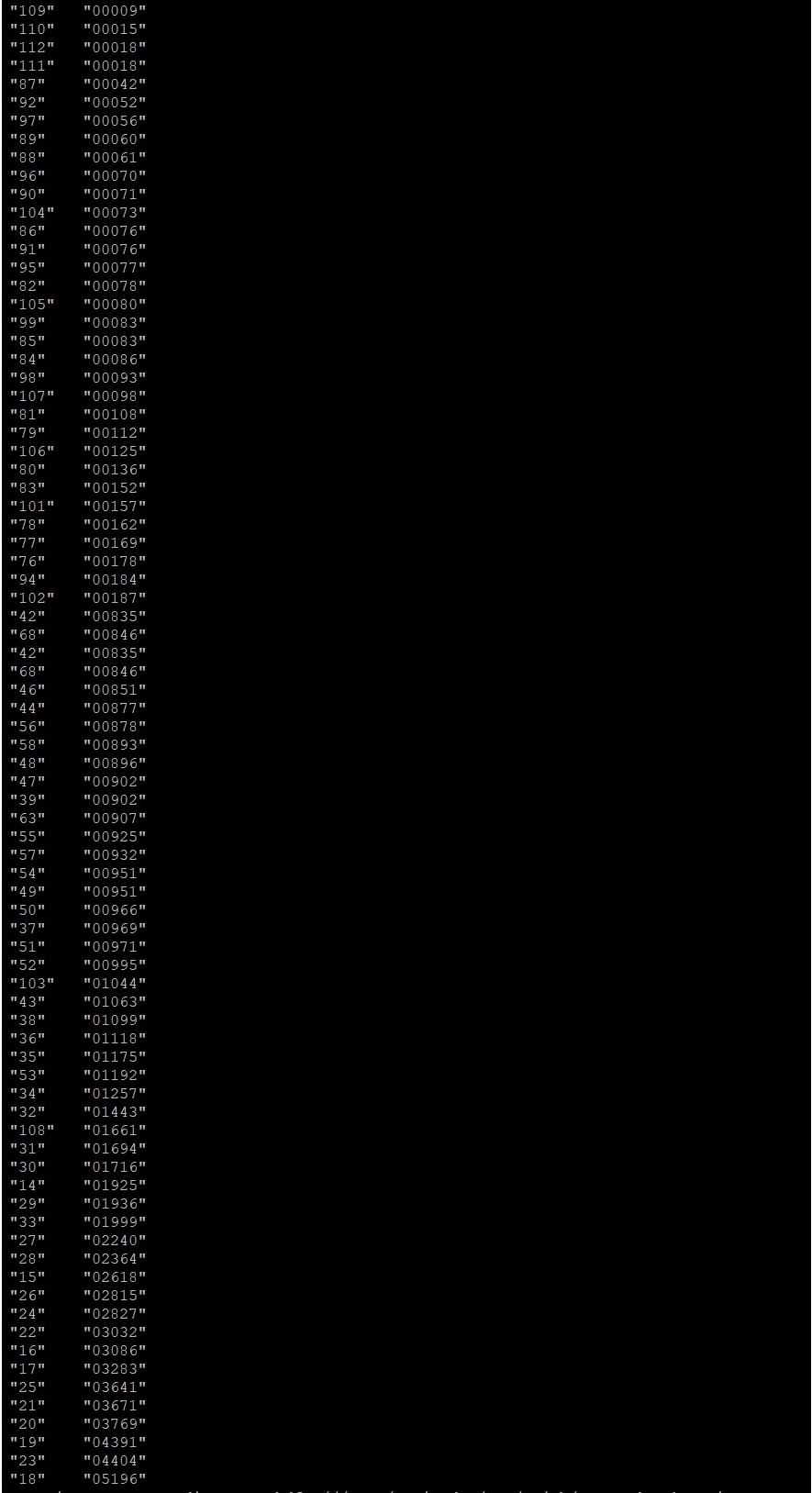
WhatAgeUsesFacebook.run()

**Command:** python map\_reduce1.py -r hadoop --hadoop-streaming-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar hdfs:///tmp/facebook\_data/pseudo\_facebook.csv

**Age wise distribution of users:**



**Output: (Age,Count)**

****

**Observation :** Facebook is most popular between age groups 16 and 26.

**3. Likes Given (Using Drill)**

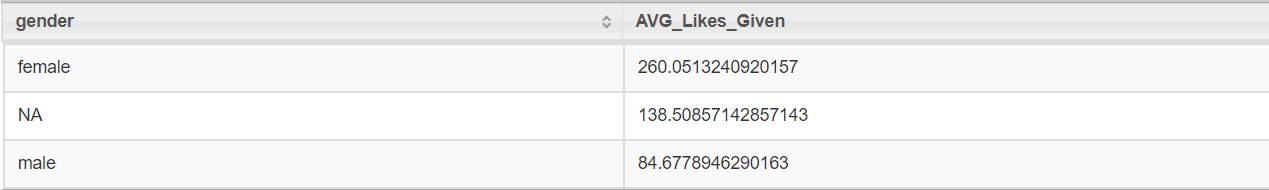
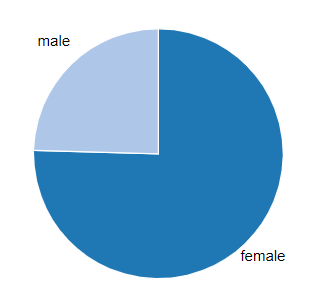
**CMD:** apache-drill-1.12.0/bin/drillbit.sh start -Ddrill.exec.http.port=8765

**Query 1:** SELECT gender,avg(likes) AS AVG\_Likes\_Given

FROM hive.facebook\_db.facebook

GROUP BY gender

ORDER BY AVG\_Likes\_Given DESC

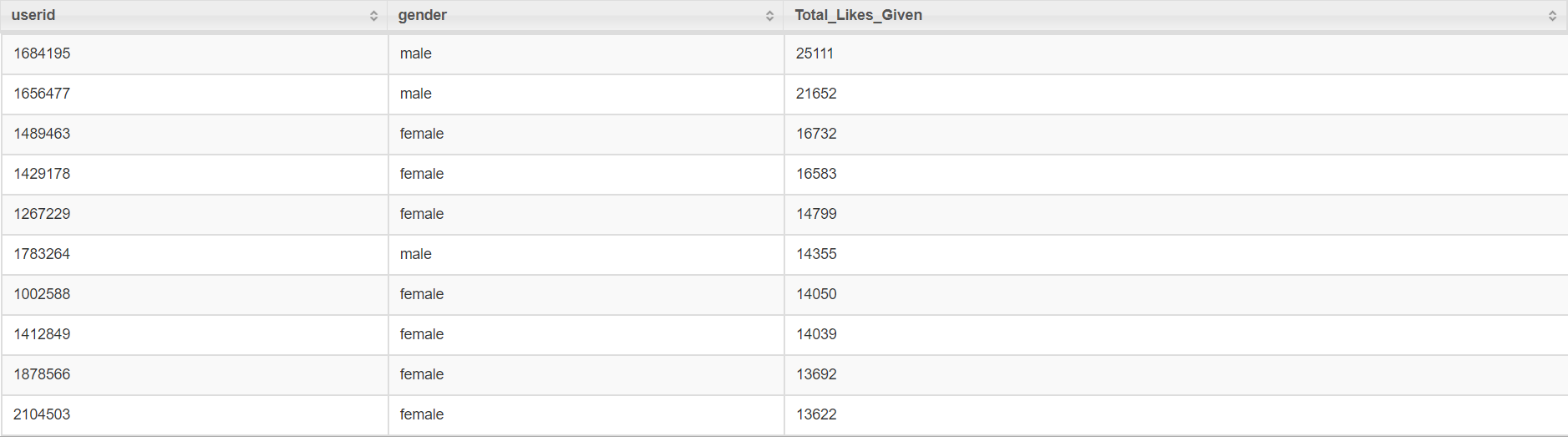
**Output: gender vs likes given : **

**Query 2:** SELECT userid, gender, likes AS Total\_Likes\_Given

FROM hive.facebook\_db.facebook

ORDER BY Total\_likes\_Given DESC LIMIT 10

**Output : Top 10 users with most likes given**



**Analysis Result:** Females give more likes then men

**4. Likes Received (Using Drill)**

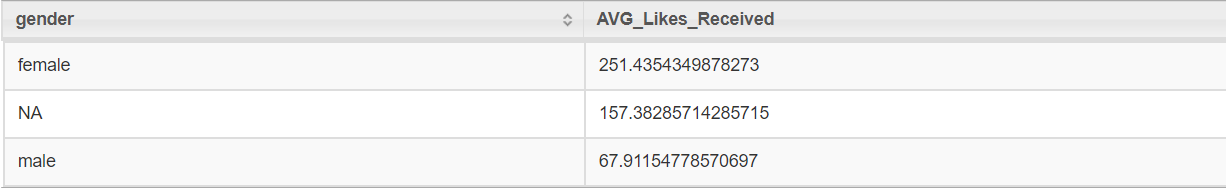
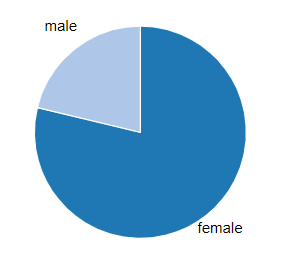
**CMD:** apache-drill-1.12.0/bin/drillbit.sh start -Ddrill.exec.http.port=8765

**Query 1:** SELECT gender,avg(likes\_received) AS AVG\_Likes\_Received

FROM hive.facebook\_db.facebook

GROUP BY gender

ORDER BY AVG\_Likes\_Received DESC

**Output: gender vs total likes received : **

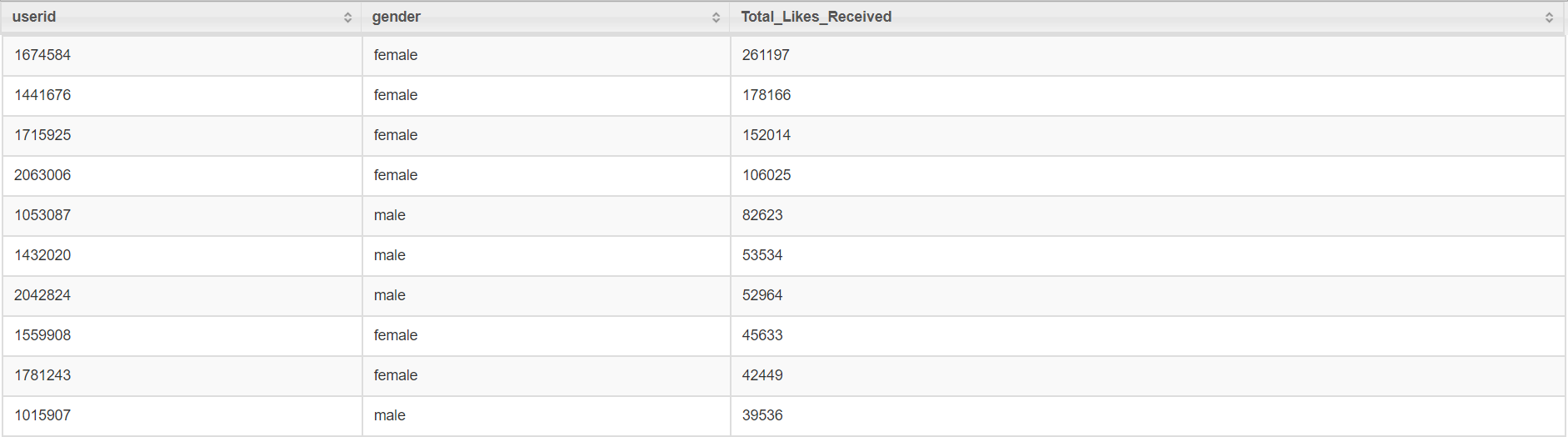
**Query 2:** SELECT userid, gender, likes\_received AS Total\_Likes\_Received

FROM hive.facebook\_db.facebook

ORDER BY likes\_received DESC

LIMIT 10

**Output : Top 10 users with most likes received**



**Analysis Result:** Females receive more likes then men

**5.Gender Count (Using Zeppelin(Spark code)):**

val x = fbDF.groupBy("gender").count().orderBy(desc("count")).cache()

x.show()

**Output:**

+------+-----+

|gender|count|

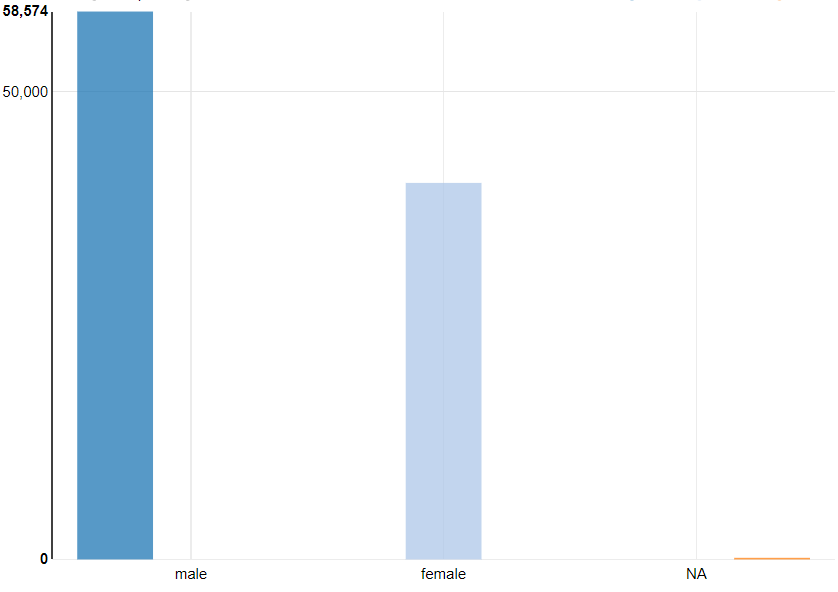
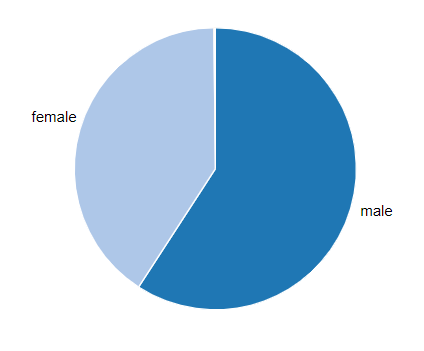
+------+-----+

| male |58574|

|female|40254|

| NA | 175 |

+------+-----+



**Analysis** : There are more male users than female .

**6.Likes Split Up (using Zeppelin-sql code)**

**Query 1:**

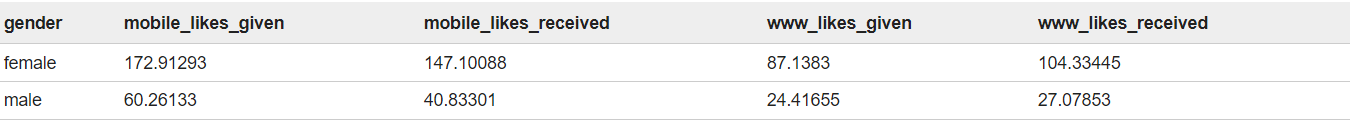
SELECT gender,avg(mobile\_likes) AS mobile\_likes\_given, avg(mobile\_likes\_received) AS mobile\_likes\_received, avg(www\_likes) AS www\_likes\_given, avg(www\_likes\_received) AS www\_likes\_received

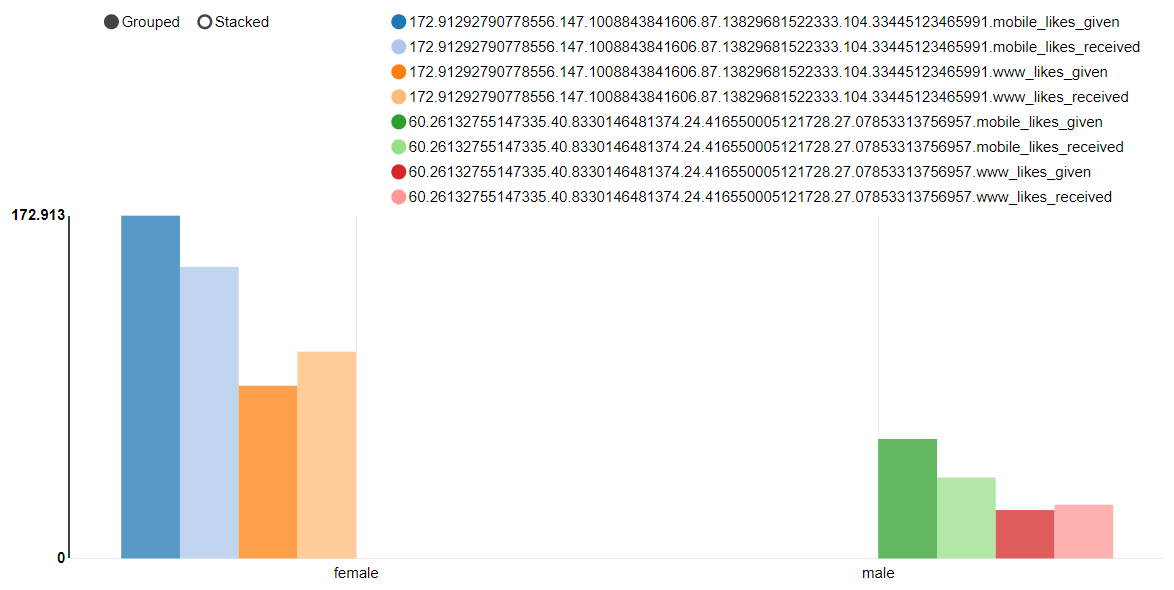
FROM fb

WHERE gender <> "NA"

GROUP BY gender

**Output:**





**Query2:**

%sql

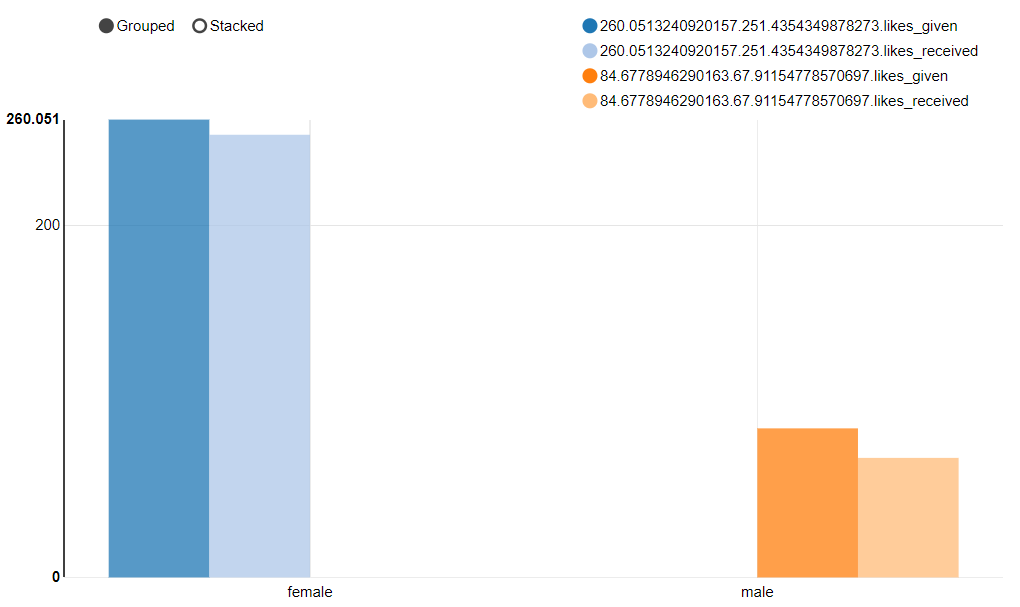
SELECT gender,avg(likes) AS likes\_given ,avg(likes\_received) AS likes\_received

FROM fb

WHERE gender <> "NA"

GROUP BY gender

**Output(Likes vs Likes Recived by gender):**

****

**Analysis:** Interesting obsservation for gender specific interaction with facebook: women like as well as are liked a lot more than men (nearly 2.5 as much).

### 7.Friends Counts & Friendships initiated (using Zepplin -sql code)

**Query :**

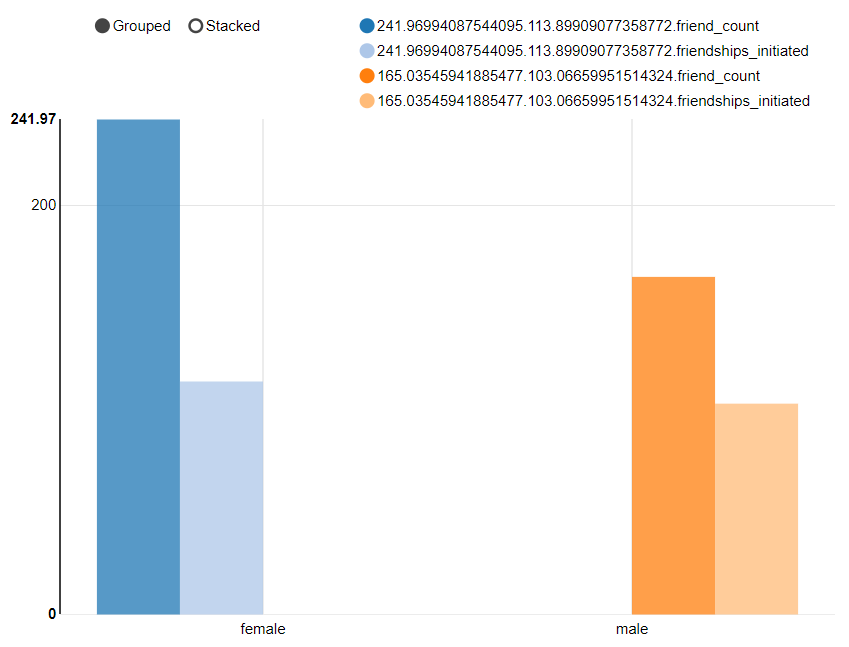
SELECT gender,avg(friend\_count) AS friend\_count ,avg(friendships\_initiated) AS friendships\_initiated

FROM fb

WHERE gender <> "NA"

GROUP BY gender

Output : (Friends Count vs Friendships Initiated)



**Analysis:** Women have more friends than men on facebook, the friendships initiated in proportion to friend count are more in case of men than women.

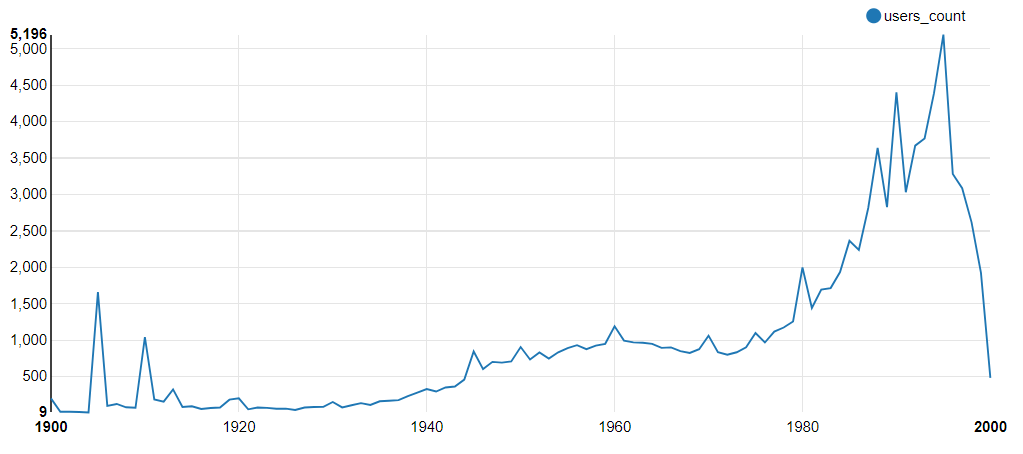
### 8. Users w.r.t birth year(using Zepplin -sql code)

**Query:** SELECT dob\_year,count(userid) AS users\_count

FROM fb

GROUP BY dob\_year

**Output:**

****

**Analysis:**

We see bumps between 1940 to 1980. After 1980 the no. users rocket. Since the data is till 2000 (we see miniscule value in 2000)