-- (10 pts) Q1: What are the top 20 cities (with state information) with the most businesses on yelp?

-- Expected columns: state, city, business\_count

SELECT

state, city, COUNT(\*) AS business\_count

FROM

yelp\_business

GROUP BY state , city

ORDER BY business\_count DESC

LIMIT 20;

NV Las Vegas 26798

AZ Phoenix 17213

ON Toronto 17210

NC Charlotte 8554

AZ Scottsdale 8227

PA Pittsburgh 6354

QC Montreal 5972

AZ Mesa 5871

NV Henderson 4463

AZ Tempe 4263

AZ Chandler 3993

EDH Edinburgh 3704

OH Cleveland 3322

WI Madison 3213

AZ Glendale 3206

AZ Gilbert 3128

ON Mississauga 2726

BW Stuttgart 2000

AZ Peoria 1707

ON Markham 1564

-- (10 pts) Q2: What are the average stars for top 10 most reviewed restaurants in Toronto?

-- Expected columns: business\_id, review\_count, avg\_star

SELECT

business\_id, review\_count, stars AS avg\_star

FROM

yelp\_business

WHERE

city = 'Toronto'

ORDER BY review\_count DESC

LIMIT 10;

r\_BrIgzYcwo1NAuG9dLbpg 1494 4.5

aLcFhMe6DDJ430zelCpd2A 1270 4.0

RtUvSWO\_UZ8V3Wpj0n077w 1184 4.0

N93EYZy9R0sdlEvubu94ig 973 4.0

iGEvDk6hsizigmXhDKs2Vg 945 4.5

RwRNR4z3kY-4OsFqigY5sw 848 3.5

-av1lZI1JDY\_RZN2eTMnWg 808 3.5

zgQHtqX0gqMw1nlBZl2VnQ 759 3.0

uAAWlLdsoUf872F1FKiX1A 746 4.0

f5O7v\_X\_jCg2itqacRfxhg 701 4.0

-- (10 pts) Q3: How many users have left reviews but all the reviews are 1 star?

-- Expected columns: user\_count

SELECT

COUNT(DISTINCT s.user\_id) AS user\_count

FROM

(SELECT

user\_id

FROM

yelp\_review

WHERE

stars = 1) as s;

445215

-- (15 pts) Q4: Who wrote the second funniest review in this data?

-- What is his/her name and how many reviews has he written?

-- Expected columns: name, review\_count

SELECT

a.name, a.review\_count

FROM

(SELECT

\*

FROM

yelp\_user

ORDER BY funny DESC) a

LIMIT 1 , 1;

Rohlin 796

-- (15 pts) Q5: How many user have left tips to more business than reviews?

-- That is, for each user in this set, more businesses are getting tips from them than getting review from them.

-- Expected columns: user\_count

select count(\*) as user\_count from

(select user\_id,count(\*) as count

from yelp\_tip

group by user\_id) t1

left join

(select user\_id,count(\*) as count

from yelp\_review

group by user\_id) t2

on t1.user\_id = t2.user\_id

where t1.count > t2.count or (t1.count <> null and t2.count = null);

36506

-- (20 pts) Q6: We define "boring people" as -

-- users have left reviews but...

-- none of the reviews were marked as useful, funny, or cool...

-- and he/she has left no tips ever

-- How many "boring people" do we see in the data?

select count(\*) from

yelp\_user

where review\_count > 0 and

(useful + funny + cool) > 0 and

yelp\_user.user\_id not in (select user\_id from yelp\_tip);

472663

-- (20 pts) Q7: What is the distribution of reviews per user?

-- Please generate the histogram of this.

-- What is the distribution look like? How do you describe it?

-- Is mean or median a better representation of the distribution?

-- Expected columns: review\_count, user\_count

SELECT FLOOR(review\_count/10.00)\*10 as review\_count,

COUNT(\*) AS user\_count

FROM yelp\_user

GROUP BY FLOOR(review\_count/10.00)\*10

ORDER BY 1;

0 864530

10 184046

20 79687

30 44767

40 28314

50 20061

60 14718

70 11207

80 8893

90 7084

100 6244

110 5392

120 4557

130 3776

140 3470

150 3064

160 2700

170 2505

180 2141

190 1894

200 1792

……………..

Exponential distribution.

Mean