Xiaodan Zhu
Oct 19, 2015

Yelp Data Analysis

1. Average ratings of different business categories vary a lot. For some business categories, ratings on average can be really high, such as above 4.5; while for other business categories, ratings on average can be much lower, such as lower than 3. This suggests people tend to be satisfied with some business categories more easily than others.

Example:

Business Categories With Average Higher Ratings

```
select category, avg(stars), min(stars), max(stars), count(*) from (select * from bcategories as A, business as B where A.business_id = B.business_id) as temp group by category having avg(stars)>4.5 and count(*)>20;
```

Category	Average	Min	Max	Count
Pilates	4.64	3	5	67
Acupuncture	4.73	2.5	5	67
Event Photography	4.88	1.5	5	51
Session Photography	4.55	1.5	5	63
Permanent Makeup	4.64	3	5	22
Yelp Events	4.80	3.5	5	33
Blow Dry/Out Services	4.53	2.5	5	58
Home Theatre Installation	4.77	4	5	24
Barre Classes	4.65	3.5	5	30
Spray Tanning	4.61	2	5	35
Photographers	4.52	1	5	165
Chiropractors	4.60	1	5	249
Boot Camps	4.74	3.5	5	48
Party Equipment Rentals	4.54	2	5	27
Dance Studios	4.56	3	5	70
Home Inspectors	4.67	3	5	23
Home Window Tinting	4.54	2.5	5	23
Hair Stylists	4.60	2	5	207
Life Coach	4.73	3	5	26
Naturopathic/Holistic	4.65	2	5	43
Dog Walkers	4.71	1	5	57
Martial Arts	4.52	2.5	5	98
Personal Injury Law	4.56	1	5	40

Business Categories With Average Lower Ratings

```
select category, avg(stars), min(stars), max(stars), count(*) from (select * from bcategories as A, business as B where A.business_id = B.business_id) as temp group by category having avg(stars)<3 and count(*)>20;
```

Category	Average	Min	Max	Count
Laboratory Testing	2.97	1	4.5	33
Television Service Providers	2.28	1	5	48
Taxis	2.85	1	5	78
Apartments	2.65	1	5	469
Car Rental	2.89	1	5	176
Airlines	2.73	1	5	39
Internet Service Providers	2.45	1	5	112
Post Offices	2.90	1.5	4.5	107
Property Management	2.67	1	5	104

2. People from different states have similar tastes and give same businesses similar ratings. For example, for businesses in both states AZ and NV, AZ people give average rating of 3.20, and NV people give average rating of 3.28.

Example:

Average Rating from AZ of 856 Businesses Also in NV

```
select state, avg(stars) from (select * from business where state = 'AZ' and name in (select name from business where state = 'NV')) as temp group by state;
```

Average Rating from NV of 856 Businesses Also in AZ

```
select state, avg(stars) from (select * from business where state = 'NV' and name in (select name from business where state = 'AZ')) as temp group by state;
```

State	Avg
AZ	3.2038
NV	3.2804

Notes:

These 856 businesses have 5135 satellites in AZ and 3476 satellites in NV.

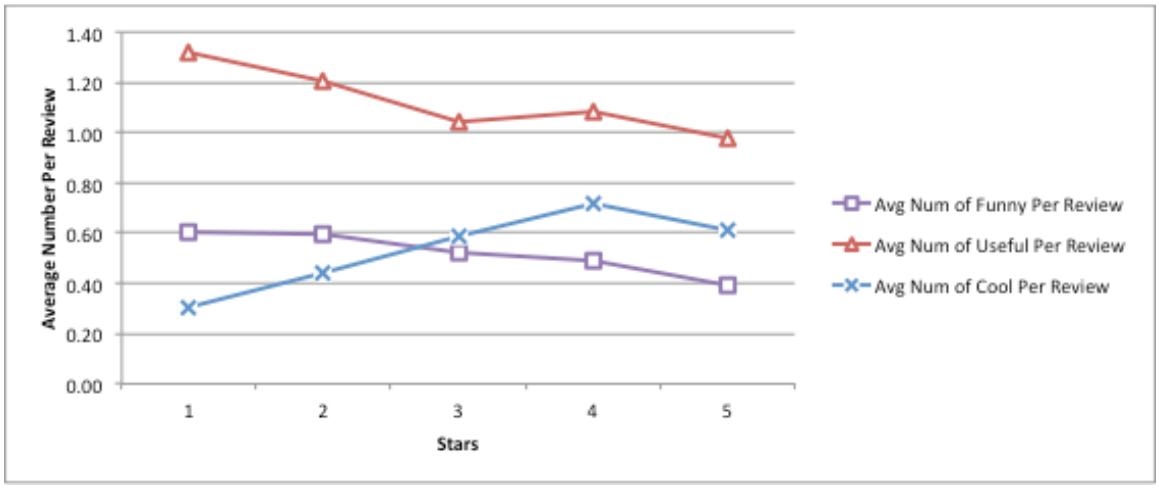
```
select count(distinct name) from business where state = 'AZ' and name in (select name from business where state = 'NV');
select count(*) from business where state = 'AZ' and name in (select name from business where state = 'NV');
select count(*) from business where state = 'NV' and name in (select name from business where state = 'AZ');
```

3. There are more high star reviews than low star reviews. The number of 4 and 5 star reviews is three times the number of 1 and 2 star reviews. While people on average think high star reviews are cooler, they think low star reviews are much more funny and useful.

Example:

select stars, avg(funny) as avg_funny, avg(userful) as avg_useful, avg(cool) as avg_cool, sum(funny) as sum_funny, sum(userful) as sum_useful, sum(cool) as sum_cool, count(*) from (select distinct * from review as A, reviewvotes as B where A.review_id = B.review_id) as temp group by stars;

Stars	Avg_Funny	Avg_Useful	Avg_Cool	Sum_Funny	Sum_Useful	Sum_Cool	Count
1	0.60	1.32	0.30	96319	210546	48622	159811
2	0.59	1.21	0.44	83176	169560	61511	140608
3	0.53	1.05	0.59	117209	232922	130922	222719
4	0.49	1.08	0.72	229532	504497	335295	466599
5	0.39	0.97	0.61	225263	564130	356079	579527



4. For people rated the same business multiple times, most of them kept their original ratings. Some of them changed their opinions over time; while among these people, slightly more of them increased their ratings.

Example:

```
create view multi_ratings as(select A.business_id, A.user_id, B.stars as first_rate, A.avg_stars from (select business_id, user_id, min(date) as first_rate_date, avg(stars) as avg_stars from review group by business_id, user_id having count(*)>1) as A, review as B where A.business_id=B.business_id and A.user_id=B.user_id and A.first_rate_date=B.date);
```

10956 Pairs/6837 Users Increased Ratings

```
select * from multi_ratings where first_rate<avg_stars;
select count(distinct user_id) from multi_ratings where first_rate<avg_stars;
```

10764/6485 Users Pairs Decreased Ratings

```
select * from multi_ratings where first_rate>avg_stars;
select count(distinct user_id) from multi_ratings where first_rate>avg_stars;
```

21060 Pairs/12778 Users Kept Ratings

```
select * from multi_ratings where first_rate=avg_stars;
select count(distinct user_id) from multi_ratings where first_rate=avg_stars;
```

Notes:

20412 users reviewed 17398 businesses more than once. 42780 pairs of business and user have more than one rating.

```
select distinct business_id from review group by business_id, user_id having count(*)>1;
select distinct user_id from review group by business_id, user_id having count(*)>1;
select business_id, user_id from review group by business_id, user_id having count(*)>1;
```

5. People joined Yelp later on average give more harsh ratings.

Example:

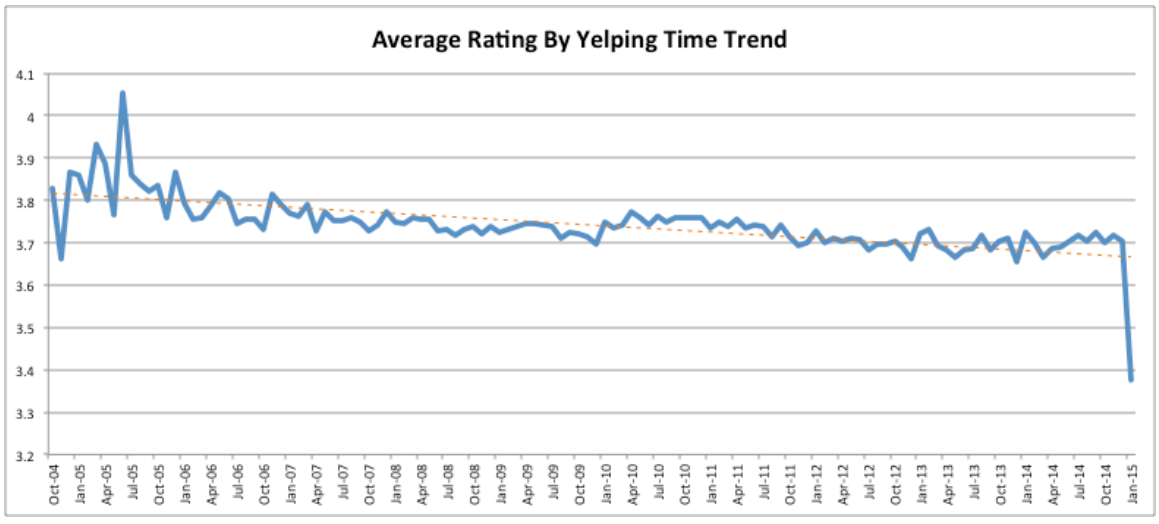
By Year (from 2004 to 2015)

select yelping_since_year as year, avg(average_stars) as average_rating from users group by yelping_since_year order by yelping_since_year ;

Year	Average_Rating
2004	3.8171
2005	3.8548
2006	3.7723
2007	3.7565
2008	3.7375
2009	3.7275
2010	3.7545
2011	3.7308
2012	3.6999
2013	3.6945
2014	3.7016
2015	3.3763

By Month (from Oct 2014 to Jan 2015)

select yelping_since_year as year, yelping_since_month as month, avg(average_stars) as average_rating from users group by yelping_since_year, yelping_since_month order by yelping_since_year, yelping_since_month;



6. From 2006 to 2015, average rating trend generally keeps constant. From 2004 to 2006, there are lots of turbulences in the average rating.

Example:

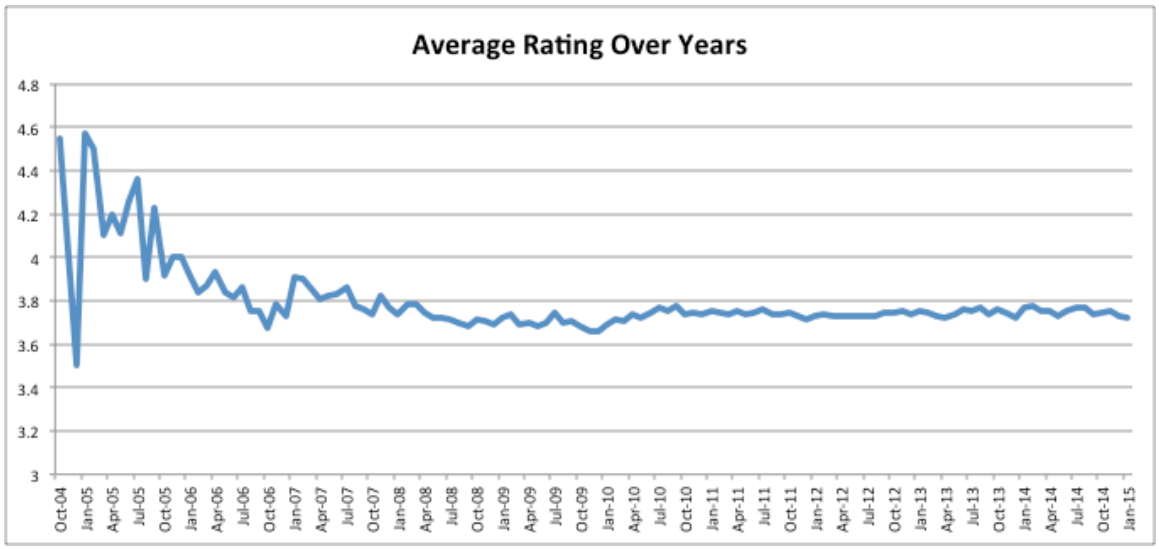
By Year

select EXTRACT(YEAR FROM date) as year, avg(stars) as avg_rating from review group by EXTRACT(YEAR FROM date) order by EXTRACT(YEAR FROM date);

Year	Average_Rating
2004	4.3846
2005	4.0515
2006	3.7988
2007	3.8149
2008	3.7195
2009	3.6976
2010	3.7388
2011	3.7414
2012	3.7341
2013	3.7452
2014	3.7529
2015	3.7197

By Month

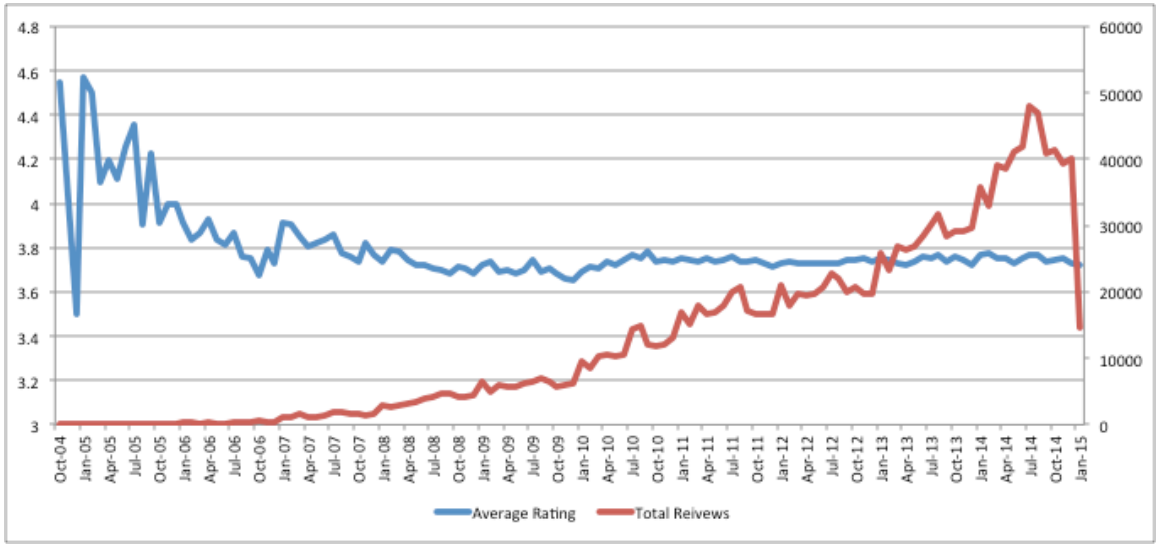
select EXTRACT(YEAR FROM date) as year, EXTRACT(MONTH FROM date) as month, avg(stars) as avg_rating from review group by EXTRACT(YEAR FROM date), EXTRACT(MONTH FROM date) order by EXTRACT(YEAR FROM date), EXTRACT(MONTH FROM date);



7. Number of reviews in the month is constantly increasing over years.

Example:

select EXTRACT(YEAR FROM date) as year, EXTRACT(MONTH FROM date) as month, count(*) as num_review from review group by EXTRACT(YEAR FROM date), EXTRACT(MONTH FROM date) order by EXTRACT(YEAR FROM date), EXTRACT(MONTH FROM date);



8. On average, people give slightly higher reviews after they become elite members.

Example:

Split the review into two parts: reviews before the user becomes an elite member, and reviews after the user becomes an elite member.

Before as Elite, Average Rating is 3.7532

```
select avg(stars) from review as A, elite as B where A.user_id=B.user_id and EXTRACT(YEAR FROM A.date)<=B.year;
```

After as Elite, Average Rating is 3.7729

```
select avg(stars) from review as A, elite as B where A.user_id=B.user_id and EXTRACT(YEAR FROM A.date)>B.year;
```

9. People give upper middle ratings have more fans. More specifically, users with average ratings in the range of 3.1 to 4.4 have more fans than users with average ratings out of that range. Users with average rating of 3.9 have most fans.

Example:

select round(average_stars,1) as stars, avg(fans) as avg_fans from users group by round(average_stars,1) order by round(average_stars,1);