Yelp Project Analysis

Executive Summary

We are a group of data analysts trying to evaluate the yelp rating system to (1)find what makes “internet-famous” restaurants, and (2)sort & invite 100 influencers to a series of forthcoming events (online/offline) to get them to share their opinions about the customer rating rationales. Key Recommendations are based on:

- Trend: Yelp could focus on increasing user growth by implementing targeted marketing campaigns to attract new users and businesses, especially during the contraction periods when growth slows down. Yelp could also consider incentivizing users to leave reviews and ratings, as it has a positive impact on the platform's activeness and growth. Additionally, given that food and inside are the most commonly assigned labels to photo reviews, Yelp could encourage users to post more photos of these categories to enhance the user experience and attract more users to the platform.
- Correlation (business owners' side): Yelp could suggest high-rated restaurants (Star>=4) focus on showcasing their menu and interior through photos while improving the representation of drinks and outside areas. Yelp could also advise low-rated restaurants (Star <=3) to prioritize improving their menu and interior, as these are the photo labels that have a moderate to high positive impact on high-rated restaurants. Additionally, Yelp could encourage both high and low-rated restaurants to increase the number of reviews. Also, given that the labels have a relatively lower correlation to business ratings as of now, Yelp could try implementing new marketing campaigns suggesting users and businesses actively use labels when posting photos to improve engagement and increase bonding among businesses and users.
- User: Yelp could suggest to businesses that improving their overall rating can lead to increased ratings from popular users, which in turn can lead to increased user popularity and engagement. Yelp could also encourage users to leave detailed and high-quality reviews, as it positively impacts both user popularity and restaurant ratings. Additionally, Yelp could consider implementing programs to incentivize general users while promoting “Yelp Influencers” to leave more reviews and ratings, as it can drive user engagement and ultimately benefit businesses on the platform.

SQL Analysis

a. Trend:

- How many new users/ businesses enter Yelp each year? What’s the YoY change?

Yelp's total user count increased over time, but new user entrance and growth rate varied. During the expansion stage from 2005 to 2011, Yelp had a strong bump in new users, reaching over 7,000 users in 2011. In the peak stage from 2012 to 2015, Yelp steadily gained new users, reaching almost 2,000 in 2015. However, in 2016 and 2017, the company experienced a contraction period with a significant decrease in new user growth. Despite this, Yelp expanded its active users from 13 in 2005 to over 13,000 in 2017, showing impressive growth overall. (See Appendix 1).

- Are there any patterns in the photo uploaded? What label do users add most to their photos?

Yelp users can assign five different labels to photo reviews: food, drink, menu, inside, and outside. Among the top 100 restaurants with the most photo reviews, food is the most commonly assigned label, followed by inside. However, a small number of restaurants receive more photo reviews for their outside and inside areas than for their food, indicating that users value not only the taste of the food but also the restaurant's ambiance and surroundings.

- How does the number of reviews change YoY?

According to the review growth chart, it is excellent that the number of reviews increases rapidly from 2005 to 2017. The activeness of existing users improved every year, but the growth rate fluctuated and declined during recent years. The highest growth rate is 347.5% in 2007 while in 2017 the number of reviews grew to only 29.2%. Based on such a large sample, it is not easy for Yelp to change the habits of users and enhance the review growth rate every year. (Appendix 2)

- **Business Suggestions for Yelp:**

  Yelp could focus on increasing user growth by implementing targeted marketing campaigns to attract new users and businesses, especially during the contraction periods when growth slows down. Yelp could also consider incentivizing users to leave reviews and ratings, as it has a positive impact on the platform's activeness and growth. Additionally, given that food and inside are the most commonly assigned labels to photo reviews, Yelp could encourage users to post more photos of these categories to enhance the user experience and attract more users to the platform.

b. Correlation(Business Owner Side)

- **Findings:**
  - Types of labels attached to the uploaded photos

"Menu" and "inside" are the photo labels that have a moderate to a high positive impact on high-rated restaurants. “Outside” and “drink” are the labels that have a moderate to a high negative impact on high-rated restaurants.
All labels of the photo give the low-rated restaurant positive impacts, indicating that photos might not be the determining factor for a restaurant’s low rating.
However, the overall impact of photo labels on ratings is quite low, suggesting that they are not the only or primary factor driving ratings. (See Appendix 3)

  - Number of photos uploaded

The R-square values for the relationship between photos and ratings are quite small, with values of 0.0003 and 0.0057 for businesses with ratings greater than or equal to 4 and less than or equal to 3, respectively. The regression lines' slopes are also quite small, with values of 0.0001 and 0.0024, respectively. These results suggest that photos are not a strong predictor of the rating for businesses in either category, but the positive slopes indicate that photos may have some impact on the rating, particularly for businesses with lower ratings. As the number of photos uploaded increases, the ratings tend to slightly increase.

- Number of tips

The R-square values and slopes for the relationship between tips and ratings are quite small for businesses with ratings greater than or equal to 4 (R-square = 0.0026, slope = -0.011) or less than or equal to 3 (R-square = 0.0023, slope = 0.019), indicating a weak correlation. Based on the results, it can be inferred that tips have a weak influence on the rating for businesses with ratings greater than or equal to 4 or less than or equal to 3. However, there is a slight negative effect of tips on the rating for higher-rated businesses and a slight positive effect for lower-rated businesses–as number of tips increase, ratings will increase for the low-rating business group but decrease for the high-rating business group.

- Length of business hours

Business hours have a weak and possibly negligible relationship with ratings for businesses with ratings greater than or equal to 4 (R-square = 0.0029, slope = -0.0010) or less than or equal to 3 (R-square = 0.0006, slope = -0.0005). The small R-square values and slopes suggest that business hours are not strong predictors of the rating for businesses in either category and that other factors may be more important in determining the rating. The positive and negative signs of the slopes indicate that business hours may have a slightly negative effect on the rating for higher-rated businesses and a slightly weaker negative effect for lower-rated businesses. Overall, ratings tend to decrease as business hours increase.

- Number of reviews

The R-square values for the relationship between the number of reviews and the rating for businesses with ratings greater than or equal to 4 or less than or equal to 3 are 0.015 and 0.005, respectively, indicating a weak correlation. The slopes of the regression lines are also small, suggesting that the number of reviews is not a strong predictor of the rating for businesses in either category. From the slopes, we could see that the number of reviews gives a slightly positive effect on lower-rated businesses and a slightly negative effect on higher-rated businesses. Overall, the relationship between the number of reviews and the rating for businesses with ratings in these categories is weak. (Summarization of R Square and Slope statistics in Appendix 4)

Overall, we can conclude from our analysis that the review number is most correlated with the rating of higher-rated businesses, and it has the largest absolute value of the slope, suggesting the increase in review number will drive ratings down most severely for the higher-rated businesses group among the factors we analyzed. The number of photos uploaded is most correlated with the rating of lower-rated businesses, but the review number has the largest slope among the four factors, suggesting the increase in review number will drive ratings up most strongly for the lower-rated businesses group among the factors we analyzed. (Appendix 5)

- What could be the contributing factors for low-rating restaurants?

It can be observed that the factors of photos, tips, and business hours have a weak and possibly negligible relationship with ratings for both categories of restaurants, and the number of reviews has a weak positive relationship. However, for restaurants with lower ratings, the slopes for tips and reviews are larger, suggesting a slightly stronger positive relationship with the rating.

- **Business Suggestions for Yelp:**

Yelp could suggest high-rated restaurants (Star>=4) focus on showcasing their menu and interior through photos while improving the representation of drinks and outside areas. Yelp could also advise low-rated restaurants (Star <=3) to prioritize improving their menu and interior, as these are the photo labels that have a moderate to high positive impact on high-rated restaurants. Additionally, Yelp could encourage both high and low-rated restaurants to increase the number of reviews. Also, given that the labels have a relatively lower correlation to business ratings as of now, Yelp could try implementing new marketing campaigns suggesting users and businesses actively use labels when posting photos to improve engagement and increase bonding among businesses and users.

c. Popular users (influencers) analysis

- What are the top 100 users who have the most fans? How many of them are elite yelpers?

The top 100 users with the most fans have a range of followers from around 30 to 500, with 91% of them as elite yelpers for at least one year from 2005 to 2017. Therefore, we think it is highly likely that joining the elite community could build up a more differentiated profile for the yelp user which further boosts the user's profile popularity and attracts increasing fans.

- Correlation between user popularity (fans amount) and
    1) review amount
    2) Sign up time
    3) Compliment amount
    4) User activeness

Since we perform the regression analysis on the user group as a separate and unified dataset, the sample size is evenly distributed and observed, therefore, we combine the four relevant factors together into the regression model for any potential correlations and seek their coefficient significance at the same time.

	Slope	R Square
Review amount	0.080149637	0.50593437
Sign up time	-0.068919147	
Compliment amount	0.019870921	
Weekly activeness	0.010196215	

Under the multiple regression model, the R-square value for all the variables is the same, equal to 0.506, which is relatively large and indicates that more than 50% of the variances can be explained by the three different variables. Moreover, with the identical correlation matching, the three slopes of regression lines reveal diverse significances on the dependent variable. Both the review amount and compliment amount are positively correlated with the fans amount, while the review amount shows a stronger connection of 0.08 slope than the compliment amount with a slightly weaker connection of 0.02. In the contrast, the sign-up time which is calculated as the total year length of user account has a negative effect on the user popularity, with a slope of negative 0.07, showing that the followers prefer users with a more recent account. However, though the user's weekly activeness is calculated as the week duration per review, it shows a modestly positive relationship with the user popularity, with a positive slope of 0.01.

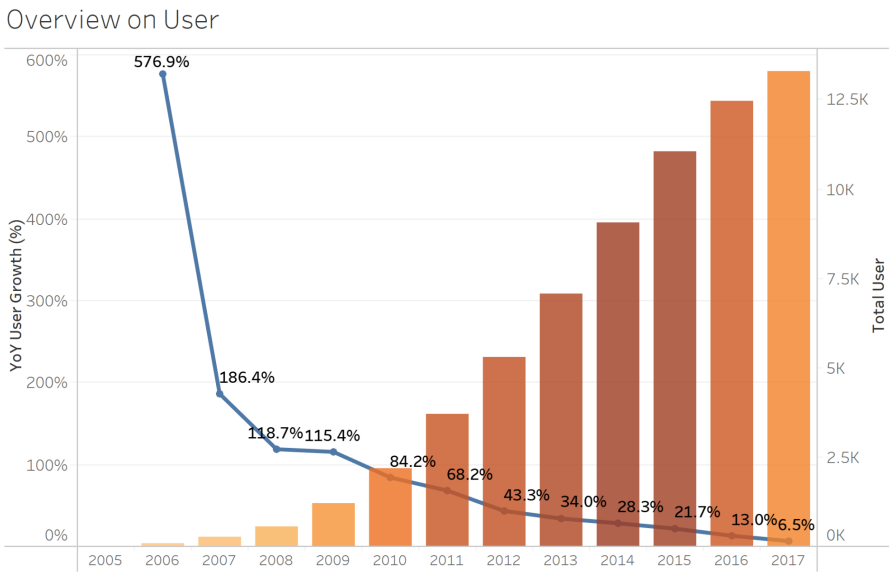
- Would the stores they gave high scores also have higher overall ratings?

After analyzing the user data, we would like to figure out if the popular user's rating for a restaurant is a factor correlating with a restaurant's general rating. Therefore, we first filter the most popular users with the largest number of fans on Yelp and generate the table of users' ratings for each business and the corresponding business average star in SQL. The coefficient is about 0.28 based on the regression model, indicating a positive correlation between the two variables. Then we perform the visualization on Tableau by setting the business rating as the column and the popular user's rating as the row. According to the graph, there is a positive relationship between the business stars and the stars given by popular users. (Appendix 6)

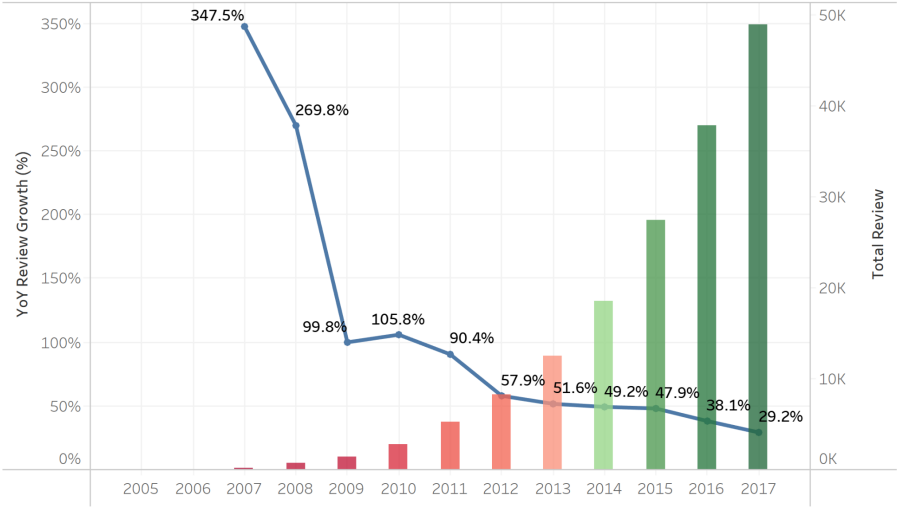
- **Business Suggestions for Yelp:**

  Yelp could suggest to businesses that improving their overall rating can lead to increased ratings from popular users, which in turn can lead to increased user popularity and engagement. Yelp could also encourage users to leave detailed and high-quality reviews, as it positively impacts both user popularity and restaurant ratings. Additionally, Yelp could consider implementing programs to incentivize general users while promoting "Yelp Influencers" to leave more reviews and ratings, as it can drive user engagement and ultimately benefit businesses on the platform.

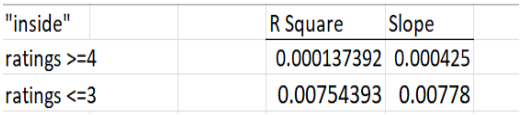
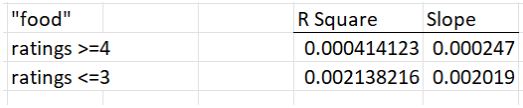
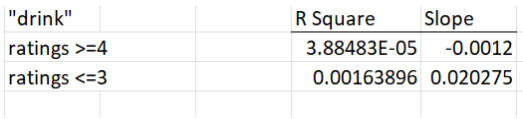
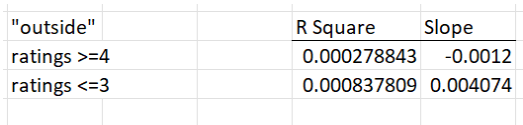
Appendix

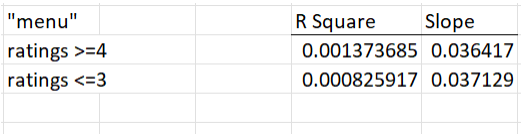


Appendix 1



Appendix 2

Label	Findings	
Inside	The R square value for ratings >=4 is 0.000137392, indicating a very weak relationship between ratings and the "inside" label. The positive slope of 0.000425002 suggests a weak positive relationship between higher ratings and the "inside" label. For ratings <=3, the R square value is 0.007543931, indicating a moderate relationship between lower ratings and the "inside" label. The positive slope of 0.007782737 suggests a weak positive relationship between lower ratings and the "inside" label, but slightly stronger than for ratings >=4.	
Food	The R square value for both ratings >=4 and ratings <=3 is very low (0.000414123 and 0.002138216, respectively), indicating that the variation in the "food" label is not well-explained by ratings. The positive slope is very small for both (0.000246983 and 0.002018875, respectively), suggesting a weak positive relationship between higher ratings and the "food" label.	
Drink	The R square value for ratings >=4 is very low (3.88483E-05), indicating that the variation in the "drink" label is not well-explained by ratings. The negative slope of -0.001195617 suggests a weak negative relationship between higher ratings and the "drink" label. For ratings <=3, the R square value is moderate (0.00163896), indicating a moderate relationship between lower ratings and the "drink" label. The positive slope of 0.020275457 suggests a moderate positive relationship between lower ratings and the "drink" label.	
Outside	The R square value for ratings >=4 is very low (0.000278843), indicating that the variation in the "outside" label is not well-explained by ratings. The negative slope of -0.001203506 suggests a weak negative relationship between higher ratings and the "outside" label. For ratings <=3, the R square value is moderate (0.000837809),	

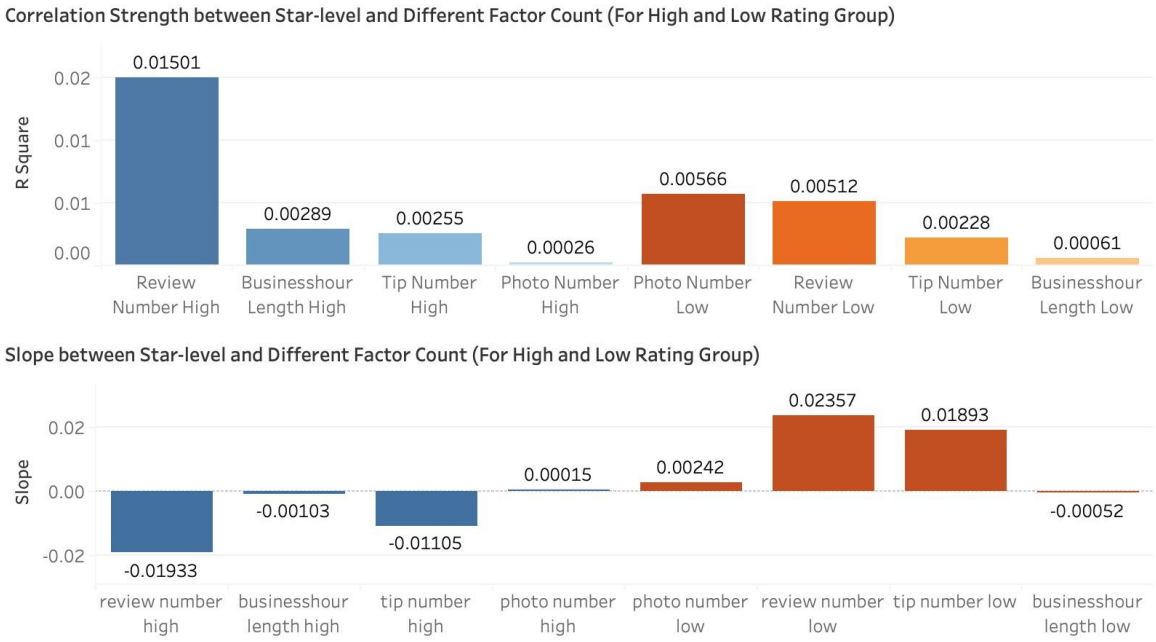
	indicating a moderate relationship between lower ratings and the "outside" label. The positive slope of 0.004074005 suggests a moderate positive relationship between lower ratings and the "outside" label.	
Menu	The R square value for both ratings >=4 and ratings <=3 is moderate (0.001373685 and 0.000825917, respectively), indicating a moderate relationship between ratings and the "menu" label. The positive slope is moderate to high for both (0.036416559 and 0.037128906, respectively), suggesting a moderate to a high positive relationship between ratings and the "menu" label.	

"menu"	R Square	Slope
ratings >=4	0.001373685	0.036417
ratings <=3	0.000825917	0.037129

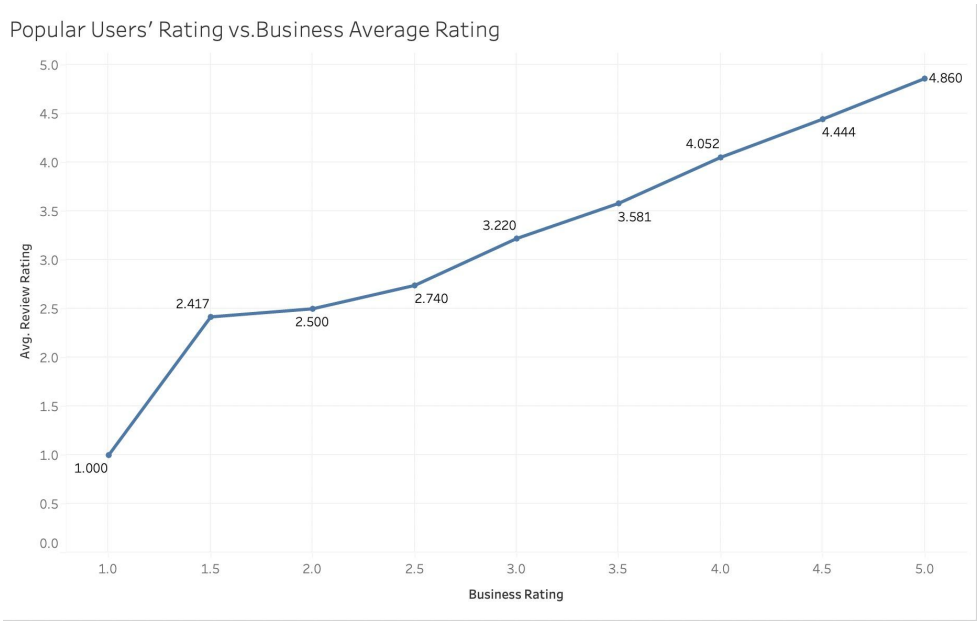
Appendix 3

review		R Square	Slope
--number of reviews of business with ratings >=4		0.015007107	-0.019329
--number of reviews of business with ratings <=3		0.005118426	0.0235697
photo	ratings >=4	0.000262282	0.0001479
	ratings <=3	0.005661169	0.0024222
tip	ratings >=4	0.00255276	-0.011052
	ratings <=3	0.002275722	0.0189261
businesshour	ratings >=4	0.002891581	-0.001031
	ratings <=3	0.000614316	-0.000516

Appendix 4



Appendix 5



Appendix 6

SQL Codes

- --number of reviews of business with ratings >=4

```
select business_id, count(business_id) as review_no, business.stars
from review join business using (business_id)
where business.stars>=4
group by business_id
order by business.stars desc;
```

- --number of reviews of business with ratings <=3

```
select business_id, count(business_id) as review_no, business.stars
from review join business using (business_id)
where business.stars<=3
group by business_id
order by business.stars desc;
```

- --types of label

```
--sort labels for businesses with ratings >=4
SELECT
business_id,
name,
stars,
SUM(CASE WHEN photos.label = 'inside' THEN 1 ELSE 0 END) AS inside_count,
SUM(CASE WHEN photos.label = 'food' THEN 1 ELSE 0 END) AS food_count,
SUM(CASE WHEN photos.label = 'drink' THEN 1 ELSE 0 END) AS drink_count,
SUM(CASE WHEN photos.label = 'outside' THEN 1 ELSE 0 END) AS outside_count,
SUM(CASE WHEN photos.label = 'menu' THEN 1 ELSE 0 END) AS menu_count
FROM business
INNER JOIN photos ON business.business_id = photos.business_id
WHERE business.stars >= 4
GROUP BY business.business_id, business.name, business.stars
ORDER BY business.stars DESC;

--sort labels for businesses with ratings<=3
SELECT
business_id,
name,
stars,
SUM(CASE WHEN photos.label = 'inside' THEN 1 ELSE 0 END) AS inside_count,
SUM(CASE WHEN photos.label = 'food' THEN 1 ELSE 0 END) AS food_count,
SUM(CASE WHEN photos.label = 'drink' THEN 1 ELSE 0 END) AS drink_count,
SUM(CASE WHEN photos.label = 'outside' THEN 1 ELSE 0 END) AS outside_count,
SUM(CASE WHEN photos.label = 'menu' THEN 1 ELSE 0 END) AS menu_count
FROM business
```

```
INNER JOIN photos ON business.business_id = photos.business_id
WHERE business.stars<= 3
GROUP BY business.business_id, business.name, business.stars
ORDER BY business.stars DESC;
```

- --no of photos of business with ratings >=4

```
select business_id, count(photo_id) as photo_no, business.stars
from photos join business using (business_id)
where business.stars>=4
group by business_id
order by business.stars desc;
```

- --no of photos of business with ratings <=3

```
select business_id, count(photo_id) as photo_no, business.stars
from photos join business using (business_id)
where business.stars<=3
group by business_id
order by business.stars desc;
```

- --no of tips of business with ratings >=4

```
select business_id, count(text) as tip_no, business.stars
from tip join business using (business_id)
where business.stars>=4
group by business_id
order by business.stars desc;
```

- --no of tips of business with ratings >=4

```
select business_id, count(text) as tip_no, business.stars
from tip join business using (business_id)
where business.stars<=3
group by business_id
order by business.stars desc;
```

- --length of business hours

```
create temporary table t1 as

select business_id,"hours.Monday",
    (CAST (Substr("hours.Monday",1,instr("hours.Monday", '-')-1)as int))
    +case when Substr("hours.Monday",instr("hours.Monday", ':')+1, 1)='3' then 0.5 else 0
    end as Monstarttime,
```

```
	(CAST (Substr("hours.Monday", instr("hours.Monday",'-')+1, length("hours.Monday"))as
int))
	+case when (Substr("hours.Monday",instr("hours.Monday",'-')+3, 1)='3'
		or Substr("hours.Monday",instr("hours.Monday",'-')+4, 1)='3') then 0.5 else 0
	end as Monendtime
from business
where "hours.Monday" is not "N/A" and "hours.Monday" is not null;

create temporary table t2 as

select business_id,"hours.Tuesday",
	(CAST (Substr("hours.Tuesday",1,instr("hours.Tuesday", '-')-1)as int))
	+case when Substr("hours.Tuesday",instr("hours.Tuesday", ':')+1, 1)='3' then 0.5 else 0
	end as Tuestarttime,
	(CAST (Substr("hours.Tuesday", instr("hours.Tuesday",'-')+1, length("hours.Tuesday"))as
int))
	+case when (Substr("hours.Tuesday",instr("hours.Tuesday",'-')+3, 1)='3'
		or Substr("hours.Tuesday",instr("hours.Tuesday",'-')+4, 1)='3') then 0.5 else 0
	end as Tueendtime
from business
where "hours.Tuesday" is not "N/A" and "hours.Tuesday" is not null;

create temporary table t3 as

select business_id,"hours.Wednesday",
	(CAST (Substr("hours.Wednesday",1,instr("hours.Wednesday", '-')-1)as int))
	+case when Substr("hours.Wednesday",instr("hours.Wednesday", ':')+1, 1)='3' then 0.5 else
0
	end as Wedstarttime,
	(CAST (Substr("hours.Wednesday", instr("hours.Wednesday",'-')+1,
length("hours.Wednesday"))as int))
	+case when (Substr("hours.Wednesday",instr("hours.Wednesday",'-')+3, 1)='3'
		or Substr("hours.Wednesday",instr("hours.Wednesday",'-')+4, 1)='3') then 0.5 else 0
	end as Wedendtime
from business
where "hours.Wednesday" is not "N/A" and "hours.Wednesday" is not null;

create temporary table t4 as

select business_id,"hours.Thursday",
	(CAST (Substr("hours.Thursday",1,instr("hours.Thursday", '-')-1)as int))
	+case when Substr("hours.Thursday",instr("hours.Thursday", ':')+1, 1)='3' then 0.5 else 0
	end as Thustarttime,
	(CAST (Substr("hours.Thursday", instr("hours.Thursday",'-')+1,
length("hours.Thursday"))as int))
	+case when (Substr("hours.Thursday",instr("hours.Thursday",'-')+3, 1)='3'
```

```
            or Substr("hours.Thursday",instr("hours.Thursday",'-')+4, 1)='3') then 0.5 else 0
    end as Thuendtime
from business
where "hours.Thursday" is not "N/A" and "hours.Thursday" is not null;

create temporary table t5 as

select business_id,"hours.Friday",
    (CAST (Substr("hours.Friday",1,instr("hours.Friday", '-')-1)as int))
    +case when Substr("hours.Friday",instr("hours.Friday", ':')+1, 1)='3' then 0.5 else 0
    end as Fristarttime,
    (CAST (Substr("hours.Friday", instr("hours.Friday",'-')+1, length("hours.Friday"))as int))
    +case when (Substr("hours.Friday",instr("hours.Friday",'-')+3, 1)='3'
            or Substr("hours.Friday",instr("hours.Friday",'-')+4, 1)='3') then 0.5 else 0
    end as Friendtime
from business
where "hours.Friday" is not "N/A" and "hours.Friday" is not null;

create temporary table t6 as

select t1.business_id, business.stars,
    case when (Monendtime-Monstarttime)>0 then (Monendtime-Monstarttime)
        when (Monendtime-Monstarttime)<=0 then (Monendtime-Monstarttime+24)
    end as Monhours,
    case when (Tueendtime-Tuestarttime)>0 then (Tueendtime-Tuestarttime)
        when (Tueendtime-Tuestarttime)<=0 then (Tueendtime-Tuestarttime+24)
    end as Tuehours,
    case when (Wedendtime-Wedstarttime)>0 then (Wedendtime-Wedstarttime)
        when (Wedendtime-Wedstarttime)<=0 then (Wedendtime-Wedstarttime+24)
    end as Wedhours,
    case when (Thuendtime-Thustarttime)>0 then (Thuendtime-Thustarttime)
        when (Thuendtime-Thustarttime)<=0 then (Thuendtime-Thustarttime+24)
    end as Thuhours,
    case when (Friendtime-Fristarttime)>0 then (Friendtime-Fristarttime)
        when (Friendtime-Fristarttime)<=0 then (Friendtime-Fristarttime+24)
    end as Frihours
from t1,t2,t3,t4,t5,business
where (Monhours+Tuehours+Wedhours+Thuhours+Frihours)<>0
    and t1.business_id=t2.business_id
    and t1.business_id=t3.business_id
    and t1.business_id=t4.business_id
    and t1.business_id=t5.business_id
    and t1.business_id=business.business_id;

----length of hours of business with ratings >=4
select business_id,stars,Monhours,Tuehours,Wedhours,Thuhours,Frihours,
```

```
    (Monhours+Tuehours+Wedhours+Thuhours+Frihours)as totalhours
from t6
where stars>=4
group by business_id
order by stars desc;

----length of hours of business with ratings <=3
select business_id,stars,Monhours,Tuehours,Wedhours,Thuhours,Frihours,
    (Monhours+Tuehours+Wedhours+Thuhours+Frihours)as totalhours
from t6
where stars<=3
group by business_id
order by stars desc;
```

- What are the top 100 users who have the most fans?

```
select user_id, name, fans, elite, review_count, yelping_since
from user
group by user_id, name
order by fans desc
limit 100;
```

- How many of the top 100 users are elite yelpers?

```
select user_id, name, fans, review_count
from
(select user_id, name, fans, elite, review_count
from user
group by user_id, name
order by fans desc
limit 100)
where elite is not null
group by user_id, name
order by fans desc;
```

- Do users who start using yelp earlier have more fans?

```
select user_id, name, fans, 2023 - strftime("%Y", yelping_since) as Sign_Up_Period,
review_count
from user
group by user_id, name
order by fans desc;Do users who start using yelp earlier have more fans? (correlation)
select user_id, name, fans, 2023 - strftime("%Y", yelping_since) as Sign_Up_Period,
review_count
from user
group by user_id, name
```

```
order by fans desc;
```

- Calculate user's weekly activeness

```
select user_id,
name,
fans,
review_count,
strftime("%J", date('now','localtime')) - strftime("%J", date(yelping_since))+1 as
Days_Difference,
round(((strftime("%J", date('now','localtime')) - strftime("%J", date(yelping_since))+1) /
7) / review_count, 0) as Daily_Activeness
from user
group by user_id, name;
```

- Find the correlation between popular user review and business' overall rating

```
select user.user_id, user.name, user.fans, review.review_id, business.business_id,
review.stars, business.stars as Average_Stars
from user join review on user.user_id = review.user_id
join business on review.business_id = business.business_id
where user.fans >= 30
group by user.user_id, user.name, review.review_id, review.stars, business.stars
order by user.fans desc;
```

- Find the overview of photo reviews regarding labels

```
select name, label, count(*) as Num_Of_Photo
from photos join business using (business_id)
group by name, label;
```