Yelp Dataset Analysis

STATS 140XP Final Project

Annaka Schone
Damien Ha
Sylvia Deng
Joyce Xu
Caitlin Ree
Meiyi Ye
Wenhong Sun

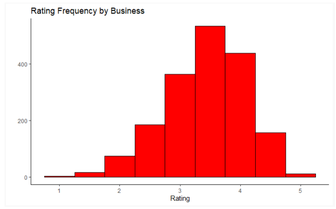
Introduction

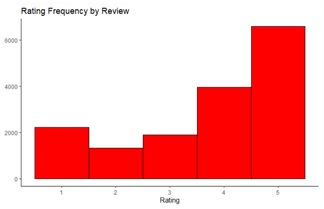
Yelp is a platform that crowd-sources reviews about businesses, with the majority being restaurants. The Yelp dataset being examined in this study was assembled for the **Yelp Dataset Challenge** in 2018. It contains Yelp business data, reviews and user data.
For the purpose of this study, the dataset has been reduced to business data and reviews of restaurants in the **Chinese Food** category, for the purpose of mitigating the effects of "Restaurant Category" in our analysis while keeping the dataset as large as possible. This study attempts to answer the following questions:

1. Do a restaurant's Yelp reviews accurately reflect its star rating?
2. What are the factors that contributed to the overall star rating of Chinese restaurants? Which factor has the greatest effect?

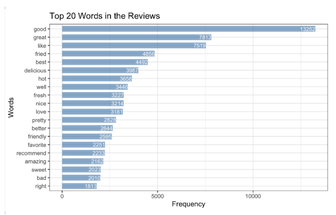
Exploratory Data Analysis

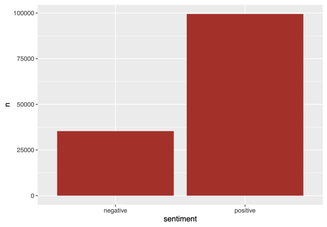
Distribution of Ratings by Business and Review





Distribution of Review Words and Sentiment





Word Cloud: Words that most commonly appear in the Yelp Reviews

Is Yelp ratings a good way to determine Chinese restaurant quality?

It probably IS!

Do a restaurant's Yelp reviews accurately reflect its star rating?

We first used **sentiment analysis** to Calculate the sentiment of each review based off lexicon and then created **2 indicators** based off the sentiment scores: overall sentiment per business (positive/negative) and sentiment percentage. Finally, we built a **linear regression model** to compare the businesses' original average stars and sentiment reflected through reviews.

```
##
## Call:
## lm(formula = stars_business ~ business_sentiment + sentiment_percentage,
##     data = yelp_dataset_final_perc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.58778 -0.29719  0.07715  0.31904  2.42215
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.07785    0.01850 112.298   <2e-16 ***
## business_sentimentpositive -0.02417    0.02081  -1.162    0.245
## sentiment_percentage            2.03410    0.03313  61.401   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4888 on 16003 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.3295, Adjusted R-squared:  0.3295
## F-statistic:  3933 on 2 and 16003 DF,  p-value: < 2.2e-16
```

The results indicate that:

1. A business' **percentage of positive review sentiments** is positively correlated with its average stars.
2. Sentiment percentage is a **more accurate reflection** of star rating than overall business sentiment.

Which aspects of a restaurant have the greatest effect on its Yelp star rating?

We applied both **linear regression** and **logistic regression** to test the effect of various potential variables on Yelp star ratings.

```
##
## Call:
## lm(formula = stars_business ~ stars_review + average_stars +
##     state + review_count + sentiment + business_sentiment + sentiment_percentage,
##     data = yelp_dataset_final_perc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.39589 -0.24846  0.01247  0.28313  2.03499
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.928e+00  3.096e-02  62.258  < 2e-16 ***
## stars_review                  1.056e-01  3.418e-03  30.903  < 2e-16 ***
## average_stars                 6.613e-03  5.207e-03   1.270 0.204074
## stateAZ                       8.029e-02  2.642e-02   3.039 0.002378 **
## stateCA                       1.040e-01  3.079e-02   3.377 0.000734 ***
## stateDE                      -9.356e-02  4.212e-02  -2.221 0.026344 *
## stateFL                       3.009e-02  2.466e-02   1.220 0.222437
## stateID                      -1.071e-01  3.596e-02  -2.977 0.002914 **
## stateIL                      -1.991e-01  5.263e-02  -3.783 0.000156 ***
## stateIN                       1.216e-02  2.651e-02   0.459 0.646403
## stateLA                       1.301e-01  2.841e-02   4.578 4.72e-06 ***
## stateMO                       6.014e-02  2.537e-02   2.371 0.017757 *
## stateNJ                       5.863e-02  2.722e-02   2.154 0.031233 *
## stateNV                       2.387e-02  2.542e-02   0.939 0.347727
## statePA                       3.570e-02  2.329e-02   1.533 0.125241
## stateTN                      -6.521e-02  2.690e-02  -2.424 0.015371 *
## review_count                  1.313e-04  3.445e-06  38.121  < 2e-16 ***
## sentimentpositive            -1.906e-01  1.116e-02 -17.110  < 2e-16 ***
## business_sentimentpositive -5.797e-02  1.906e-02  -3.041 0.002362 **
## sentiment_percentage          1.709e+00  3.223e-02  53.023  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.442 on 15862 degrees of freedom
##   (132 observations deleted due to missingness)
## Multiple R-squared:  0.4523, Adjusted R-squared:  0.4516
## F-statistic: 689.3 on 19 and 15862 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## glm(formula = binary ~ stars_review + average_stars + state +
##     review_count + sentiment, family = "binomial", data = yelp_dataset_final)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.6240284  0.2128889  -2.931  0.00338 **
## stars_review   0.2113596  0.0306044   6.906 4.98e-12 ***
## average_stars  0.0932458  0.0424371   2.197  0.02800 *
## stateAZ       -0.6344344  0.2231646  -2.843  0.00447 **
## stateCA       -1.4146165  0.2681666  -5.275 1.33e-07 ***
## stateDE       -0.1395539  0.3559806  -0.392  0.69504
## stateFL       -0.3902056  0.2007986  -1.943  0.05198 .
## stateID       -0.5804108  0.2862261  -2.028  0.04258 *
## stateIL        1.1888855  0.6319747   1.881  0.05994 .
## stateIN       -0.2298430  0.2299165  -1.000  0.31747
## stateLA       -0.1959170  0.2645303  -0.741  0.45892
## stateMO       -1.0312063  0.2003593  -5.147 2.65e-07 ***
## stateNJ       -0.1744060  0.2235313  -0.780  0.43526
## stateNV       -0.9646351  0.2345429  -4.113 3.91e-05 ***
## statePA       -0.7446485  0.1855902  -4.012 6.01e-05 ***
## stateTN       -1.3417298  0.2066399  -6.496 8.24e-11 ***
## review_count   0.0149429  0.0006237  23.959  < 2e-16 ***
## sentimentpositive  1.6210484  0.0877576  18.472  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear Regression Model

1. Most significant variables: **stars on particular reviews, review count, sentiment percentage**
2. Again, **sentiment percentage** is positively correlated with average stars, as does stars for each particular review

Logistic Regression Model

1. Similar to linear regression, stars on particular reviews and review count were significant variables
2. Location does not seem to have as much of an impact

Future Research

Future research likely involves re-running this analysis on other restaurant category subsets to see how if the results are consistent.