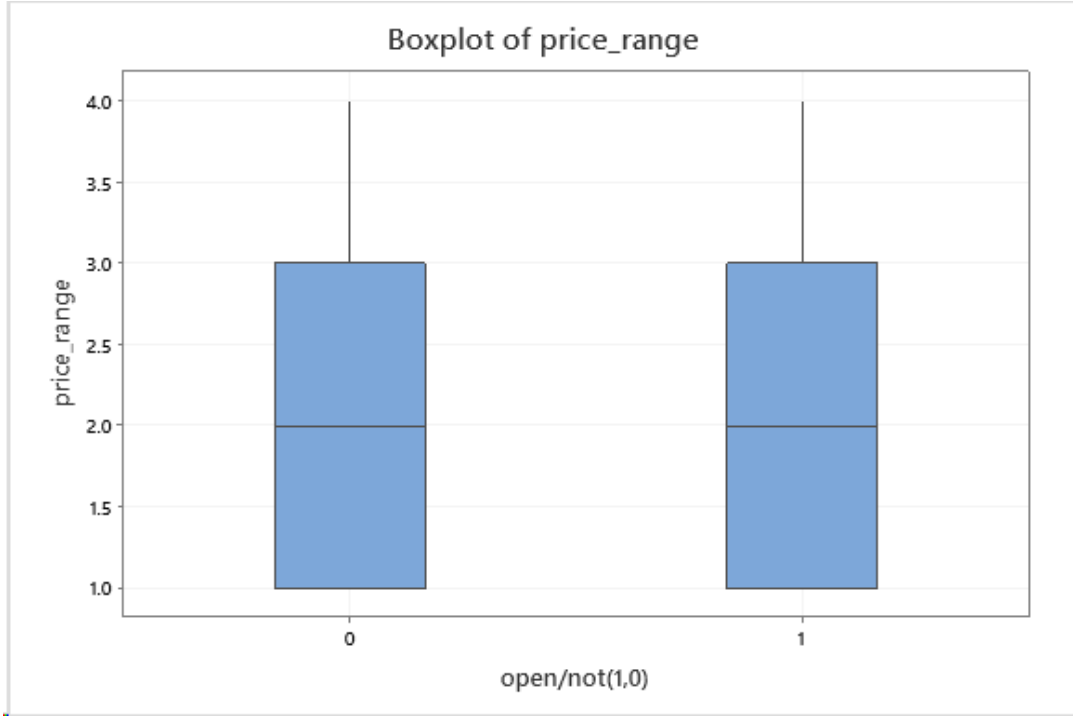
# Can Yelp star ratings, price range, and review count predict restaurants that permanently closed during the COVID-19 pandemic?

*By Rina Battulga*

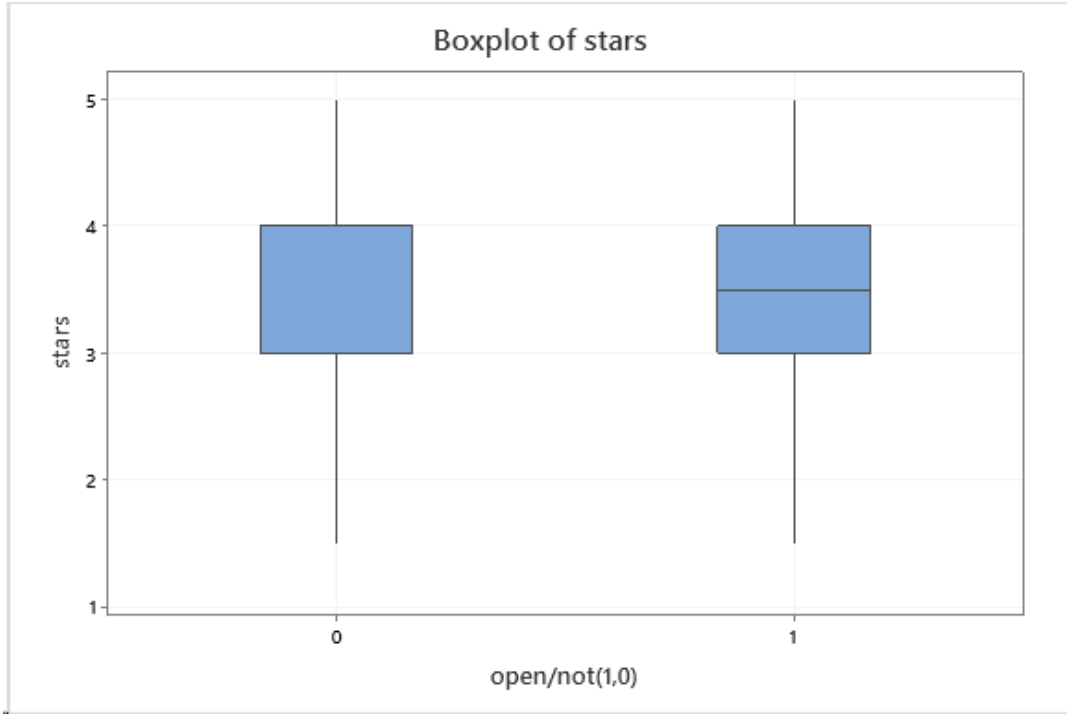
Over the past year, numerous businesses had to close their doors as a result of the COVID-19 pandemic not only in the U.S., but all over the world. The food and drinks industry has been particularly hit hard by the regulations prohibiting indoor/outdoor dining set by the government. More than 110,000 chained locations as well as independent restaurants have closed down in the U.S. alone.[1] In my study, I will be looking at two subgroups, restaurants that have permanently closed vs. restaurants that stayed open after the pandemic, from Yelp restaurant data.[2] The subgroups will be represented as binary response variables; 1 indicating that it's open, and 0 for closed. The predicting variables are the restaurant's price range, Yelp star ratings, and review count. This is going to be a retrospective case controlled study.

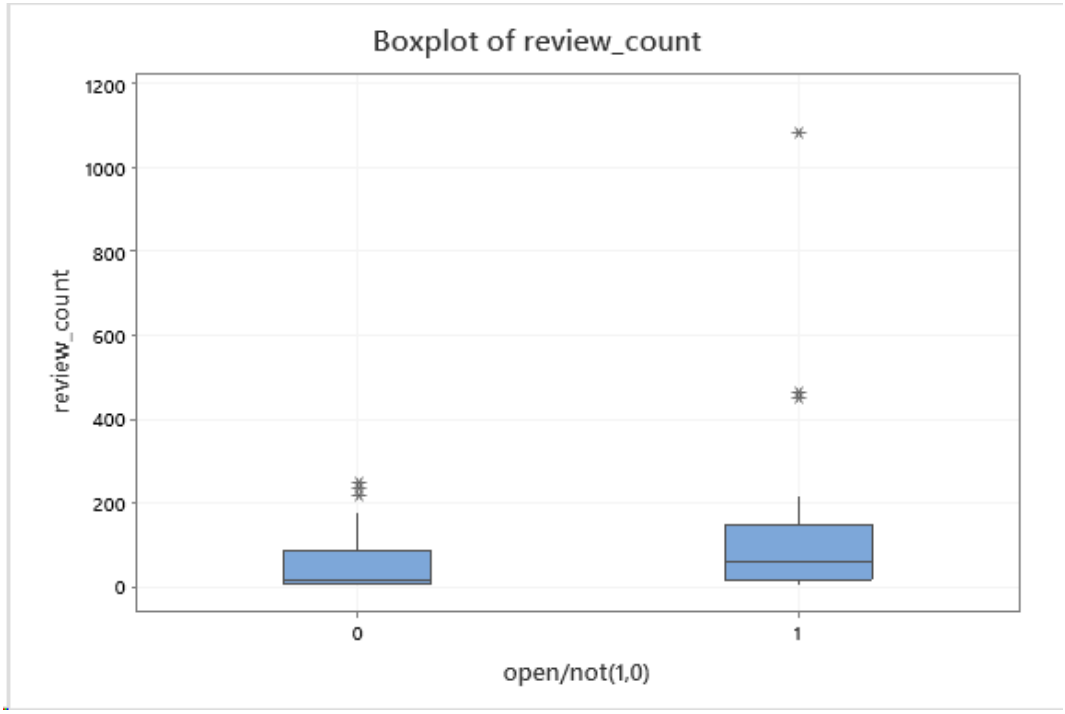
First, let's look at side-by-side box-plots of the predicting variables and binary response variable:



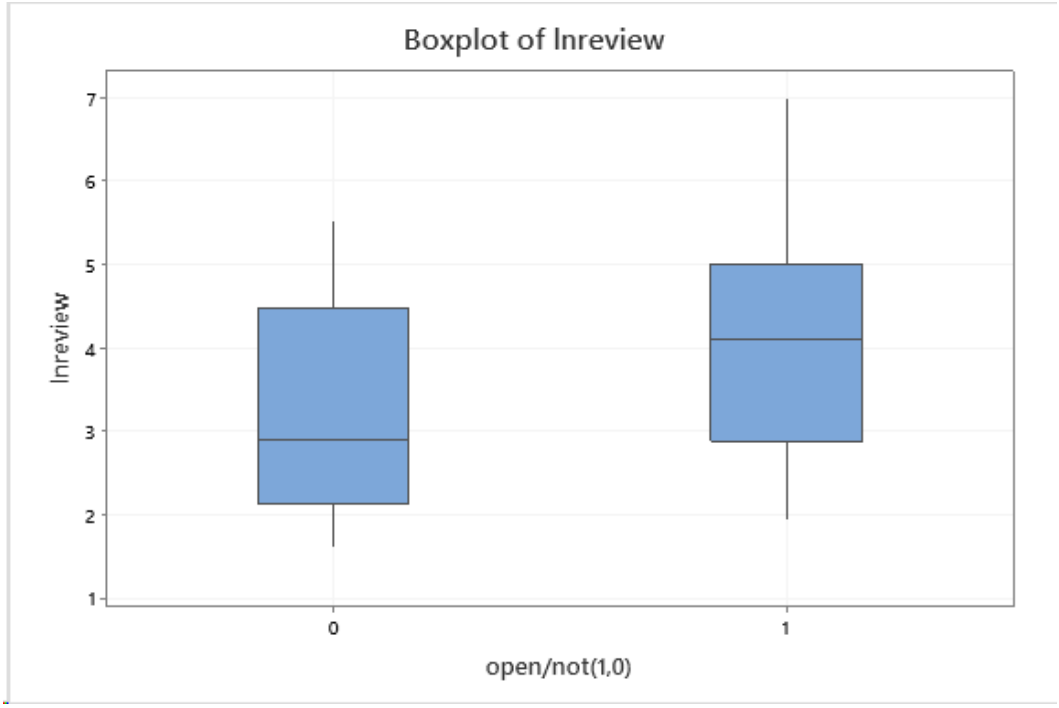

[1] https://fortune.com/2021/01/26/restaurants-bars-closed-2020-jobs-lost-how-many-have-closed-us-covid-pandemic-stimulus-unemployment/

[2] https://www.yelp.com/dataset




Looking at the boxplots, variables price_range and stars don't seem to have that big of a difference. The review count boxplot, on the other hand, suggests more variability for the restaurants that stayed open. The boxplots also seem to suggest short left tail and long right tail for both open and closed restaurants. We could try taking the natural log for these data points.




The natural logged boxplot looks more suggestive. There is definitely quite some discrepancy between the two. Restaurants that closed now had lower numbers of views on average than for those that stayed open. I'll use the natural log of the review count variable.

Now let's try fitting the data to a binary logistic regression model:

## Method

| | |
|---|---|
| Link function | Logit |
| Residuals for diagnostics | Pearson |
| Rows used | 66 |

## Response Information

| Variable | Value | Count |
|---|---|---|
| open/not(1,0) | 1 | 33 (Event) |
| | 0 | 33 |
| | Total | 66 |

## Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = 0.52 - 0.189 price_range - 0.670 stars + 0.608 lnreview

## Coefficients

| Term | Coef | SE Coef | Z-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 0.52 | 1.43 | 0.37 | 0.714 | |
| price_range | -0.189 | 0.258 | -0.73 | 0.463 | 1.09 |
| stars | -0.670 | 0.383 | -1.75 | 0.081 | 1.30 |
| lnreview | 0.608 | 0.233 | 2.61 | 0.009 | 1.26 |

## Odds Ratios for Continuous Predictors

| | Odds Ratio | 95% CI |
|---|---|---|
| price_range | 0.8276 | (0.4994, 1.3717) |
| stars | 0.5116 | (0.2413, 1.0848) |
| lnreview | 1.8359 | (1.1637, 2.8961) |

## Model Summary

| Deviance R-Sq | Deviance R-Sq(adj) | AIC | AICc | BIC | Area Under ROC Curve |
|---|---|---|---|---|---|
| 9.41% | 6.13% | 90.88 | 91.54 | 99.64 | 0.6974 |

## Goodness-of-Fit Tests

| Test | DF | Chi-Square | P-Value |
|---|---|---|---|
| Deviance | 62 | 82.88 | 0.039 |
| Pearson | 62 | 64.69 | 0.383 |
| Hosmer-Lemeshow | 8 | 9.14 | 0.331 |

## Analysis of Variance

| Source | DF | Adj Dev | Adj Mean | Likelihood Ratio Chi-Square | P-Value |
|---|---|---|---|---|---|
| Regression | 3 | 8.6120 | 2.8707 | 8.61 | 0.035 |
| price_range | 1 | 0.5455 | 0.5455 | 0.55 | 0.460 |
| stars | 1 | 3.3509 | 3.3509 | 3.35 | 0.067 |
| lnreview | 1 | 7.8592 | 7.8592 | 7.86 | 0.005 |
| Error | 62 | 82.8835 | 1.3368 | | |
| Total | 65 | 91.4954 | | | |

## Measures of Association

| Pairs | Number | Percent | Summary Measures | Value |
|---|---|---|---|---|
| Concordant | 757 | 69.5 | Somers' D | 0.40 |
| Discordant | 325 | 29.8 | Goodman-Kruskal Gamma | 0.40 |
| Ties | 7 | 0.6 | Kendall's Tau-a | 0.20 |
| Total | 1089 | 100.0 | | |

*Association is between the response variable and predicted probabilities*

When looking at the coefficients, there is little to no collinearity. The VIF's are all below 1.4. The Analysis of Variance table for this model shows that there might be some evidence against the null. The p-value for the regression is around 0.035, which could be suggestive of some association between the predicting and response variables. It is marginally statistically significant. From the three predictors, lnreview has the lowest p-value of 0.005, with pretty strong evidence against the null. The star variable's p-value is also pretty low equalling 0.067. However, the p-value for price range is quite high equalling 0.460, with weak evidence against the null hypothesis. The Area under the ROC is equal to 0.6974, we should keep this in mind when checking other models.

Moreover, the odds ratio table tells us that holding everything else fixed, 1 point higher in price_range is associated with multiplying the odds of a restaurant by 0.8276. Furthermore, one point increase in star ratings is also associated with 48.84% decrease in the odds of restaurants staying open. In addition, holding all else fixed, a 1% increase in review counts is associated with a 83.59 % increase in the odds of the restaurant staying open.

Now looking at the goodness-of-fit test, we can see that Hosmer-Lemeshow's test has a high p-value of 0.331, so we do not reject the null. This suggests that the model fits the data.

In addition, the Measures of Association tells us that 69.5% of the pairs are concordant, and 29.8% of the pairs are discordant. This means that 69.5% of the restaurants that stayed open had a higher estimated probability of the not shutting down. The Somers' D statistic is 0.4 which shows us that there isn't that big of a separation between the groups.

Looking at the Model Summary $AIC_c$ value, it is 91.54 for this model.

Although $AIC_c$ is not applicable for a binary logistic model, looking at subsets as if the data is for a least squares regression sometimes works.

## Response is open/not(1,0)

| Vars | R-Sq | R-Sq (adj) | R-Sq (pred) | Mallows Cp | S | price_range | stars | lnreview |
|---|---|---|---|---|---|---|---|---|
| 1 | 7.6 | 6.2 | 2.3 | 3.3 | 0.48796 | | | X |
| 1 | 1.1 | 0.0 | 0.0 | 7.9 | 0.50498 | | X | |
| 2 | 11.6 | 8.8 | 4.0 | 2.5 | 0.48128 | | X | X |
| 2 | 7.8 | 4.8 | 0.0 | 5.2 | 0.49151 | X | | X |
| 3 | 12.3 | 8.1 | 1.2 | 4.0 | 0.48305 | X | X | X |

When looking at the best subsets, the Mallows Cp test suggests that a 2 predictor model with variables stars and lnreview would work the best because it has the lowest value. We can try remodelling the data with only these two predictors.

## Method

| | |
|---|---|
| Link function | Logit |
| Residuals for diagnostics | Pearson |
| Rows used | 66 |

## Response Information

| Variable | Value | Count |
|---|---|---|
| open/not(1,0) | 1 | 33 (Event) |
| | 0 | 33 |
| | Total | 66 |

## Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = 1.23 - 0.473 stars + 0.00557 review_count

## Coefficients

| Term | Coef | SE Coef | Z-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 1.23 | 1.18 | 1.04 | 0.297 | |
| stars | -0.473 | 0.340 | -1.39 | 0.165 | 1.09 |
| review_count | 0.00557 | 0.00326 | 1.71 | 0.087 | 1.09 |

## Odds Ratios for Continuous Predictors

| | Odds Ratio | 95% CI |
|---|---|---|
| stars | 0.6232 | (0.3199, 1.2140) |
| review_count | 1.0056 | (0.9992, 1.0120) |

## Model Summary

| Deviance R-Sq | Deviance R-Sq(adj) | AIC | AICc | BIC | Area Under ROC Curve |
|---|---|---|---|---|---|
| 6.28% | 4.10% | 91.75 | 92.14 | 98.32 | 0.6327 |

## Goodness-of-Fit Tests

| Test | DF | Chi-Square | P-Value |
|---|---|---|---|
| Deviance | 63 | 85.75 | 0.030 |
| Pearson | 63 | 64.07 | 0.439 |
| Hosmer-Lemeshow | 8 | 5.77 | 0.673 |

## Analysis of Variance

| Source | DF | Adj Dev | Adj Mean | Likelihood Ratio Chi-Square | P-Value |
|---|---|---|---|---|---|
| Regression | 2 | 5.747 | 2.874 | 5.75 | 0.056 |
| stars | 1 | 2.028 | 2.028 | 2.03 | 0.154 |
| review_count | 1 | 5.026 | 5.026 | 5.03 | 0.025 |
| Error | 63 | 85.748 | 1.361 | | |
| Total | 65 | 91.495 | | | |

## Measures of Association

| Pairs | Number | Percent | Summary Measures | Value |
|---|---|---|---|---|
| Concordant | 683 | 62.7 | Somers' D | 0.26 |
| Discordant | 395 | 36.3 | Goodman-Kruskal Gamma | 0.27 |
| Ties | 11 | 1.0 | Kendall's Tau-a | 0.13 |
| Total | 1089 | 100.0 | | |

*Association is between the response variable and predicted probabilities*

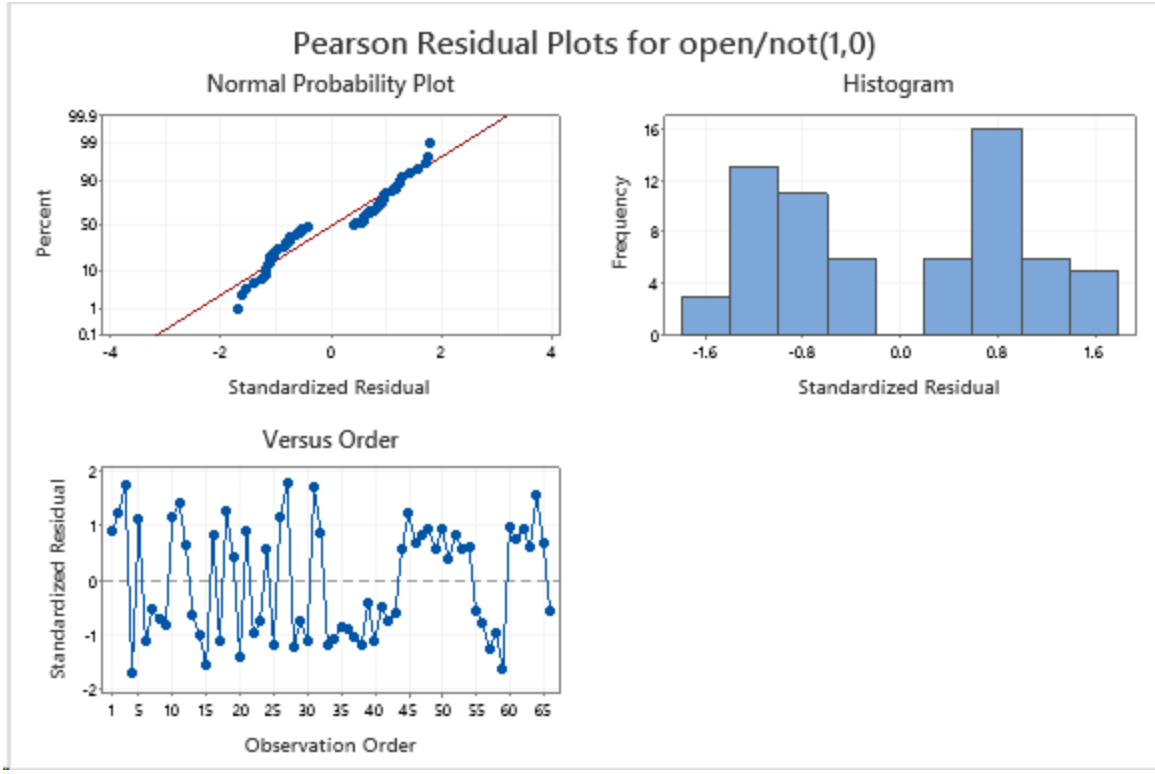
When looking at the coefficients, there is little to no collinearity. The VIF's are all below 1.1. The Analysis of Variance table for this model shows that there might be some evidence against the null. The p-value for the regression is around 0.056, which is slightly higher than the previous model. From the two predictors, lnreview has the lowest p-value of 0.025, which is less statistically significant than the variable's p-value from the 3 predictor model. The star variable's p-value is also higher equalling 0.154, compared to previous 0.067. This model overall has slightly weaker evidence against rejecting the null.

Now looking at the goodness-of-fit test, we can see that Hosmer-Lemeshow's test is quite high with a p-value of 0.673. We strongly do not reject the null. This suggests that this model fits the data pretty well.

In addition, the Measures of Association tells us that 62.7% of the pairs are concordant, and 36.3% of the pairs are discordant. The Somers' D statistic is 0.26, way less than the previous model; this shows us that there isn't a clear separation between the groups.

In the Model Summary, the $AIC_c$ value is 92.14. It is higher than the three predictor model. It seems that our original three variable model works better than the two predictor model with this data.

Now let's look at the standardized residual plot for the 3 predictor model:




From the standardized residual plots, there doesn't seem to be any unusual outliers in the data. For the histogram, we can notice a pretty strong bimodal model.

Now let's look at the classification matrix table with a cut-off at 0.5 because we are using a retrospective study with an equal number of restaurants that stayed open and closed:

**Rows: open/not(1,0)  Columns: predict**

| | 0 | 1 | All |
|---|---|---|---|
| 0 | 18 | 15 | 33 |
| | 27.27 | 22.73 | 50.00 |
| 1 | 11 | 22 | 33 |
| | 16.67 | 33.33 | 50.00 |
| All | 29 | 37 | 66 |
| | 43.94 | 56.06 | 100.00 |

*Cell Contents*
*Count*
*% of Total*

60.6% of the restaurants were classified correctly.

$$C_{pro} = (1.25)[(.4394)(.5000) + (.5606)(.5000)] = 1.25[.2197 + .2803] = 0.625$$

Comparing the $C_{pro}$ and the model's performance, it is unfortunate that the base rate is not that different from the result we got. The model we have is less than satisfactory.

Furthermore, we have to fix the intercept of the model because we are doing a retrospective study. I am going to be using the National Restaurant Association of U.S.'s statistics recognizing a 30% failure rate for the restaurant industry to correct the intercept.[3]

$$ß_0 = 0.52 + \ln [(.7)(33)/(0.3)(33)] = 0.52 + 0.8329 = 1.3529$$

Now we can adjust the probabilities with the new intercept:

**Restaurant newprob**
Jimmy John's 0.130851
Arby's 0.072664
Whistle Stop Restaurant 0.037736
Proper Eats 0.235041
Thai Yummy 0.091106
Boston Super Dog 0.119386
Cafe Ventana 0.028631
Abaee Noodle 0.050034
Cascade Restaurant 0.069130
Mr G's Pizza & Subs 0.078313
Everything POP Shopping & Dining 0.055263
Blake's On The Park 0.225132
Viuda Bistro 0.040054
Sam's Club 0.102349
Lustre Pearl Bar 0.207234
Mizu Teppanyaki & Sushi 0.145986
CaveMan Cuisine 0.107516
Geng Shi Ji 0.068696
Tuscan Kitchen 0.381608
Ristorante L 0.175373
Mikado 0.121123
Prelog's European Kitchen & Bar 0.091379

[3] https://stars.library.ucf.edu/dickpope-pubs/15/

Anthony's Restaurant  0.056717
Barley Swine  0.275129
White Egret Farm  0.119706
Crystal  0.084722
Dtox  0.038243
Journeyman  0.128452
The Roadhouse  0.053047
Roe  0.105624
The Road House  0.041645
Castagna Restaurant  0.142390
Noah Halal  0.129674
Teriyaki House  0.108947
First China Restaurant  0.071477
Bawarchi Tikka Kabob & Curry House  0.076072
Gigi's Roast Beef & Pizza  0.104573
B Street Coffee House  0.132927
Whole Time Chicken  0.018164
Olmecas Mexican Restaurant  0.121083
El Pollo Rey  0.023767
Hot Pink Taco  0.054473
Home Run Ice Cream  0.037736
Butterfly Belly Asian Cuisine  0.272949
The Smoothie Factory  0.072287
Cicis  0.207234
Can Font  0.143128
Chicken Express  0.116646
Daybreak Diner  0.264577
Broadway's Pastry & Coffee Shop  0.118724
Taco Bell  0.432382
Bridge City Cafe - Pioneer Place  0.140025
Crown Pizza & Grille  0.256743
Max Noodle House  0.237011
J Kendall's  0.031436
Saigon Subs  0.062378
Duck Walk On Wakefield Square  0.149149
French American Brasserie  0.090458
Thai House  0.223033
Pollo Regio  0.112629
Norwell Pizzeria Seafood & Grill  0.176360
Don Pedro  0.117964

Daikanyama 0.247772
Papa Murphy's 0.047121
Bronx Pizza 0.193647
Dabakh Restaurant 0.031878

The final logistic regression that's most ideal is:

$$y = 1.3529 - 0.189 \text{ price\_range} - 0.670 \text{ stars} + 0.608 \text{ lnreview}$$

Although the model did not perform better than the base rate, we can see that the variable of review counts definitely had some association with the outcome of a restaurant staying open or not. The other two variables for star ratings and price range did not have strong associations. For further studies, other variables including categorical ones for types of cuisine and regions could have more significant results.