**How to Increase Your Star Power: An Analysis of Yelp Business Data to Identify Restaurant Characteristics That Affect Yelp Users’ Restaurant Ratings**

# Statement of the Problem

In general, restaurants want to know how to attract new customers and encourage loyalty of existing customers. To this end, restaurants may benefit from better understanding of the characteristics that are most strongly tied to more positive perceptions among their existing customers. This information can shed light on changes that restaurants might want to make to better meet customers’ needs and thus improve customers’ perceptions of them.

Yelp is a website that both pulls information about restaurants from users by collecting their ratings and reviews of restaurants and pushes information to users about restaurants in terms of various characteristics, including objective information (e.g., hours, location) and subjective information about customers’ ratings and reviews. Yelp also has a “check-in” feature that allows users to check in when they are patronizing a restaurant. All of this information is available from data files collected and maintained by Yelp.

In light of these features, Yelp data have the potential to serve as a valuable source of insight to restaurants seeking to increase or at least maintain the number of customers served. By analyzing the associations between restaurant characteristics and the outcome of ratings, restaurants can better understand the drivers of their customers’ perceptions and the number of customers they serve. For example, how much of a boost in ratings might a restaurant expect if they provide wi-fi?

The client for this study is the National Restaurant Association (NRA). This organization is, by its own description, the “go-to resource for smart, relevant intelligence that helps our members run their businesses better.”[[1]](#footnote-1) Thus, the NRA should be interested in data-driven insights into changes that restaurants can make to attract more customers and increase transactions with existing customers. The NRA won’t make direct changes to restaurants based on these findings. However, they can disseminate the findings and recommendations to their members, including restaurant owners and managers. The findings will shed light on the restaurant modifications (e.g., offering wi-fi, changing the ambience of the restaurant) that would likely generate better ratings by customers and, ultimately, the greatest increases in revenue, and restaurant owners and managers should be very interested in this information.

# Proposed Approach

To better understand potential drivers of restaurant ratings, multivariate regression models will be estimated in which the following restaurant characteristics will be examined as predictors of overall ratings:

* Wi-fi
* Noise level
* Outdoor seating
* Parking
* Takes reservations
* Waiter service
* Price range
* Attire
* Ambience
* Music type
* Good for dancing
* Happy hour
* Good for groups
* Alcohol
* Good for which meal (dessert, lunch, dinner, etc.)
* Dietary restrictions (dairy-free, vegan, etc.)
* Take-out
* Drive-thru

The primary output of interest from this analysis will be the adjusted parameter estimates representing the strength and direction of association between each of the above covariates and the outcome of restaurant ratings independently of all other covariates. Based on these parameter estimates, statements can be made regarding the likely magnitude of change in restaurant ratings that can be expected from modifying certain restaurant characteristics (e.g., a restaurant that previously did not take reservations might experience a boost in restaurant ratings of .5 stars if they start taking reservations). Given the large number of restaurants in the data file, it is expected that many variables will evidence significant associations with the outcome due to high power to detect significant associations. Therefore, careful attention should be paid to the magnitude of association to identify restaurant characteristics whose modification is most likely to translate to meaningful gains in restaurant ratings.

# Data Exploration

## Important fields and information in the dataset

To identify modifiable restaurant characteristics associated with more positive perceptions, I will analyze the “business” data file from Yelp. This data set contains fields for several restaurant characteristics that can help to shed light on potential drivers of business such as overall ratings, which are indicated in this data file by the number of stars. This data file is organized such that there is one row of data for each business. Potentially modifiable restaurant characteristics that are candidates for inclusion in this analysis are:

* Wi-fi
* Noise level
* Outdoor seating
* Parking
* Takes reservations
* Waiter service
* Price range
* Attire
* Ambience
* Music type
* Good for dancing
* Happy hour
* Good for groups
* Alcohol
* Good for which meal (dessert, lunch, dinner, etc.)
* Dietary restrictions (dairy-free, vegan, etc.)
* Take-out
* Drive-thru

## Limitations of the dataset

**Inability to draw causal inferences.** One limitation of this dataset is that it does not permit causal inferences. That is, in the absence of an experimental design in which the independent variables are systematically manipulated to examine their effects on the dependent variable, we won’t know whether restaurant characteristics are causally related to restaurant ratings. If restaurant characteristics are significantly associated with restaurant ratings, we can conclude that our findings are consistent with a causal relationship, but we must acknowledge that there could be other variables not included in our analysis that are actually driving the observed associations between restaurant characteristics and ratings.

**Potentially limited generalizability of findings to customers who aren’t Yelp users.** Data on restaurant ratings are available only on Yelp users, who are likely different from the broader population of restaurant customers in ways that might limit the generalizability of findings to restaurant customers. For example, Yelp users might be younger, more comfortable with technology, better-educated, and wealthier than their non-Yelp-using counterparts. If this were true, then a restaurant characteristic like having wi-fi might be associated with restaurant perceptions (as indicated by ratings) among Yelp users but not among non-Yelp-users.

## Data cleaning and wrangling

I began the exploration of data by filtering the dataset, which contained information on many different types of businesses, to include only restaurants. This entailed parsing of a string variable called “categories” to extract businesses for which “Restaurants” was one of the categories. After subsetting the data frame to include only businesses categorized as “Restaurants”, there were 26,729 restaurants.

There were several subcategories of restaurants contained in the same string variable that seemed worthy of inclusion in the analysis as covariates (e.g., Mexican, Italian). I don’t view restaurant categories as modifiable characteristics—for example, I wouldn’t expect a restaurant to switch from offering Mexican to offering Italian cuisine based on findings from the model that I plan to run—but I think it would be worthwhile to include the categories as covariates in the model so that we can understand the associations between modifiable restaurant characteristics and restaurant ratings above and beyond (independently from) the associations between restaurant characteristics and restaurant ratings. To create binary indicators for each category of restaurant, I separated these categories into separate columns, generated a list of all unique categories of restaurants, and created a set of binary variables for all relevant categories that had at least 100 restaurants (because a categorical predictor with fewer than 100 observations in one category would be unlikely to be a significant predictor of ratings). In total, there were 67 subcategories of restaurants that met this criterion:

* + AmericanNew
  + AmericanTraditional
  + ArtsEntertainment
  + AsianFusion
  + Bagels
  + Bakeries
  + Barbeque
  + Bars
  + Breakfast Brunch
  + British
  + Buffets
  + Burgers
  + CajunCreole
  + CanadianNew
  + Caribbean
  + ChickenWings
  + Chinese
  + CocktailBars
  + CoffeeTea
  + ComfortFood
  + Delis
  + Desserts
  + Diners
  + Event (Event & Planning Services)
  + Fast Food
  + FishChips
  + Food
  + FoodTrucks
  + French
  + Gastropubs
  + German
  + GlutenFree
  + Greek
  + Hawaiian
  + HotDogs
  + IceCream
  + Indian
  + Italian
  + Japanese
  + JuiceBars
  + Korean
  + LatAmer
  + Lounges
  + Mediterranean
  + Mexican
  + MidEast
  + Nightlife
  + Pizza
  + Pubs
  + Salad
  + Sandwiches
  + Seafood
  + SoulFood
  + Soup
  + Southern
  + Specialty
  + SportsBars
  + Steakhouses
  + SushiBars
  + TapasBars
  + TapasSmall
  + TexMex
  + Thai
  + Vegan
  + Vegetarian
  + Vietnamese
  + WineBars

One feature of this data file that was somewhat challenging was that it included multiple data frames that were nested within each other. So that all variables to be analyzed were in the same data frame, I created new variables based on the old variables in the nested data frames and assigned them to the main data frame that subsumed all variables (“res\_data”).

## Univariate descriptive statistics and missingness

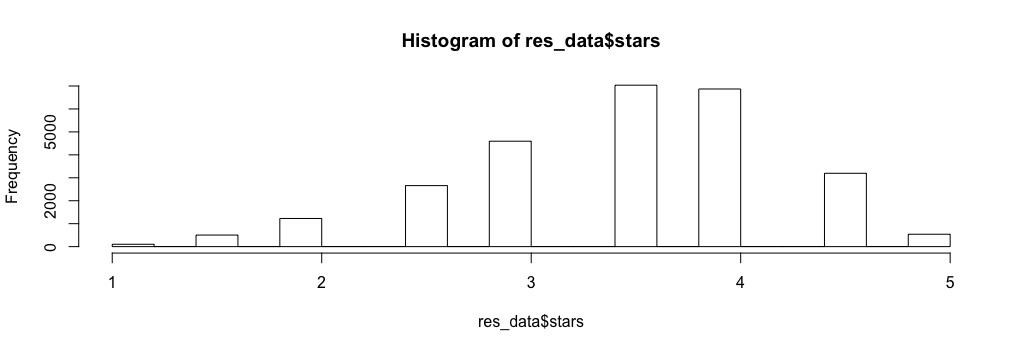
Next, I calculated basic univariate descriptive statistics for each variable of interest to understand how each variable “behaves”, including examination of frequency distributions and the amount of missingness. Appendix A contains the frequency distributions for each variable, with missingness shown as a separate category. This initial inspection revealed that several variables were characterized by a very large percentage of missing data (i.e. 85% or more), suggesting that these were too sparsely populated to be included in substantive analyses (i.e., model estimation). These variables and the percentage of restaurants that were missing values on them (in parentheses) were:

* Order at counter (98.6%)
* Drive-thru (86.3%)
* Coat check (90.4%)
* Dogs allowed (89.1%)
* Open 24 hours (99.1%)
* Music (all music indicators missing on >=92.3% cases)
* Happy hour (89.9%)
* BYOB (96.9%)
* BYOB/Corkage (95.2%)
* Smoking (89.4%)
* Dietary restrictions (all types missing 99.5% cases)

Excluding these variables leaves the following variables as possible candidates for inclusion in the analysis:

* Take-out (take\_out: 1 = yes, 0 = no)
* Wi-fi (wifi: 1 = yes (free or paid), 0 = no)
* Outdoor seating (outdoor: 1 = yes, 0 = no)
* Parking (any parking listed (parking): 1 = yes, 0 = no)
  + garage: 1 = yes, 0 = no
  + street: 1 = yes, 0 = no
  + validate: 1 = yes, 0 = no
  + lot: 1 = yes, 0 = no
  + valet: 1 = yes, 0 = no
* Takes reservations (takesres: 1 = yes, 0 = no)
* Waiter service (waiter: 1 = yes, 0 = no)
* Caters (caters: 1 = yes, 0 = no)
* Delivery (delivery: 1 = yes, 0 = no)
* Accepts credit cards (credit: 1 = yes, 0 = no)
* Wheelchair accessible (wheelch: 1 = yes, 0 = no) (although this has >50% missingness)
* Has tv (TV: 1 = yes, 0 = no)
* Noise level (noisier: 4 = very loud, 3 = loud, 2 = average, 1 = quiet)
* Good for kids (gfk: 1 = yes, 0 = no)
* Good for groups (gfg: 1 = yes, 0 = no)
* Price range (price: 1-4, higher scores indicate higher price range)
* Attire (attire: 1 = casual, 2 = dressy, 3 = formal)
* Ambience (romantic, intimate, classy, hipster, divey, touristy, trendy, upscale—for all, 1 = yes, 0 = no)
* Alcohol (alcohol: 1 = yes (beer and wine or full bar), 0 = no); 2 dummy indicators with none as the reference category: beer\_and\_wine (1 = yes, 0 = full bar or none), fullbar (1 = yes, 0 = beer and wine or none)
* Good for which meal (breakfast, brunch, lunch, dinner, dessert, latenight—for all, 1 = yes, 0 = no)

I also examined the shape of the distribution of the outcome variable, restaurant ratings (i.e., “stars”), because this has implications for the type of model that I will estimate. Examination of a histogram indicated that the ratings variable generally approximated the normal distribution, but the distribution was slightly skewed to the left. Of greater relevance to the type of model to run (e.g., linear regression vs. ordinal regression), however, is the distribution of residuals from a regression model that includes all relevant predictors. A key assumption of linear regression is that the residuals from the regression model are normally distributed. To determine whether this assumption is satisfied and linear regression is tenable for this analysis, I examine the distribution of residuals for a regression model that includes all relevant predictors, and how closely it approximates a normal distribution, in a later section.



## Handling of missing data

The variables with very high rates of missingness (noted above) were excluded from this analysis. For the remaining variables, I considered multiple imputation of missing data; however, because there wasn’t a compelling reason to expect these variables to correlate with each other (e.g., would wi-fi be expected to associate with the ambience or whether the restaurant is good for kids?), I looked to other strategies to handle missing data that didn’t depend on associations between the available variables.

The strategy that I settled on entails adding a category for missingness to categorical variables and, for continuous variables, assigning a nonmissing value (e.g., zero) to missing values and creating an indicator for missingness on that variable that is also included in the model. In this way, observations with missing data on one or more variables will not automatically be discarded from the analysis (as in listwise deletion), and missingness will be accounted for by including it as a separate category or a separate indicator in the model.

## Consideration of nested data structures

Other interesting bits of information gleaned from this inspection were that there were 298 cities represented by the restaurants in this dataset. This suggests the need to consider modeling a nested data structure, i.e., restaurants nested within cities.

The dataset also contained a field for neighborhoods, which was also considered as a possible level of analysis. However, examination of this variable revealed that, of the 26,729 restaurants in the dataset, 14,679 (54.9%) did not have a neighborhood listed. Given the high rate of missingness on this field, it was not considered further as an additional level of analysis.

## Regression diagnostics

After creating the derived variables of interest, I ran a linear regression model in base R using the lm function in which I regressed restaurant ratings (“stars) on all relevant restaurant characteristics and, where applicable, corresponding indicators of missingness, and other covariates, including restaurant subcategories and cities as a fixed-effect.

The initial regression model produced an error message indicating that there were singularities among 8 variables, all of which were indicators of missingness. Upon closer inspection of these indicators, I noticed that they were indicators of missingness for sets of variables—ambience, good for a particular type of meal (e.g., breakfast, brunch), and parking. The rates of missingness for indicators in each of these sets were nearly identical. Therefore, I left only one indicator of missingness for each of these three sets of predictors in the model (e.g., one indicator of missingness for all ambience variables, one indicator of missingness for whether the restaurant was good for a particular type of meal, and one indicator of missingness for the type of parking available). Then I reran the model and did not get any warnings.

Next, I examined the distribution of residuals to determine whether they approximated the normal distribution, thus indicating the tenability of a linear regression model for this analysis. The distribution of residuals did appear to approximate the normal distribution, thereby satisfying a key assumption of linear regression:



## Analytical approach

Based on my initial exploration of the data, a linear regression model seems appropriate for this analysis. In addition, a multilevel data structure in which restaurants (26,729 total) are nested within cities (298 total) may be warranted. One other consideration is that the number of reviews available for a given restaurant may affect the stability of restaurant ratings. That is, some restaurants may have very few reviews, in which case the number of “stars” may be a very unreliable indicator of users’ perceptions of the restaurant.

To address these issues, I will estimate multiple models to see how parameter estimates change as each of these considerations are taken into account. First, I will estimate a basic linear regression model in which the nesting of restaurants within cities is accounted for by modeling cities as a fixed effect. Next, I will estimate a multilevel linear regression model in which restaurants are nested within cities, with a random effect for cities such that the intercept is allowed to vary freely across cities; all other predictors in the model will be estimated as fixed effects. Then I will examine the effect of weighting the estimates by the number of reviews for the restaurant, so that restaurants with more reviews are given more influence in the estimation of model parameters. I will incorporate weights into both the linear regression model in which cities are estimated as a fixed effect (accomplished using the “lm” function in basic R) and the multilevel regression model in which a random effect is estimated for cities (accomplished using the “lmer” function in the “lme4” package in R). In all models, I will control for the subcategory of restaurants so that the parameter estimates of modifiable restaurant characteristics reflect the unique contribution of these characteristics to restaurant ratings above and beyond that accounted for by restaurant subcategories.

# Findings

Table 1 below juxtaposes the parameter estimates for the associations between potentially modifiable restaurant characteristics and restaurant ratings generated from four sets of models that differ in terms of whether 1) cities are modeled as a fixed effect in a linear regression model or as a random effect in a multilevel regression model, and 2) parameter estimates are weighted by the number of reviews for restaurants or not weighted.

Comparison of the parameter estimates across the four models indicates that the results for the two sets of models without weights are strikingly similar in terms of the statistical significance, magnitude, and direction of regression coefficients, as are the two sets of models with weights. Also worth noting is that the models without weights have more statistically significant predictors than do the models with weights.

Because there were so many predictors in the model, I have simplified the presentation of results by presenting in Table 1 only parameter estimates for restaurant characteristics whose associations with restaurant ratings are significant at p < .05.

In comparing the results between models with and without weights, it is apparent that there are several covariates that evidence significant associations of comparable magnitude and direction with restaurant ratings, regardless of whether weights are incorporated or not. These restaurant characteristics are the most robust predictors of restaurant ratings, i.e., they continue to be associated with restaurant ratings under different model specifications. I would be most confident in basing conclusions or recommendations on these restaurant characteristics.

Restaurant characteristics that evidenced a significant, positive association with restaurant ratings in models with and without weights are:

* having an ambience that is:
  + intimate
  + hipster
  + classy
  + upscale
  + divey
  + romantic
  + trendy
* offering catering
* having nicer attire
* having parking available in the form of:
  + a parking lot
  + street parking
* being good for:
  + dessert
  + brunch

Based on the magnitude of associations observed across all four sets of models, it appears that the covariate with the strongest association with restaurant ratings is having an intimate ambience, which is associated with .15-.20 more stars than not having an intimate ambience. The magnitudes of associations for other restaurant characteristics that were positively associated with restaurant ratings are shown in Table 1 below.

Restaurant characteristics that evidenced a significant, negative association with restaurant ratings in models with and without weights are:

* having an ambience that is touristy
* having a full bar (vs. serving no alcohol)
* being noisier
* being good for breakfast or late night
* offering delivery

Of the characteristics that were negatively associated with restaurant ratings, having a touristy ambience was associated with the greatest drop in ratings. Specifically, restaurants with a touristy ambience had roughly .4 less stars than those without a touristy ambience. The magnitudes of association for other restaurant characteristics that were negatively associated with restaurant ratings are shown in Table 1 below.

There were some restaurant characteristics that were significantly associated in models without weights but not in models without weights and vice versa. In general, it appeared that restaurant characteristics that were significantly associated with restaurant ratings in unweighted models but not weighted models and vice versa were of very small magnitude. Similarly, there were characteristics that were significantly associated with restaurant ratings in both unweighted and weighted models but that were opposite in the direction of association (i.e., positive in unweighted and negative in weighted models); for example, being good for lunch was significantly and positively associated with restaurant ratings in unweighted models and significantly and negatively associated with restaurant ratings in weighted models. These associations were also of very weak magnitude.

Table 1. Unstandardized coefficients from regression models predicting restaurant ratings from restaurant characteristics (N = 26,729)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor** | **Model 1: Unnweighted linear regression model, cities modeled as a fixed effect**  **(totalmodel2)** | **Model 2: Unweighted multilevel linear regression model with cities are modeled as a random effect**  **(totalmodel3)** | **Model 3: Weighted linear regression model with cities modeled as a fixed effect**  **(totalmodel4)** | **Model 4: Weighted multilevel linear regression model with cities modeled as a random effect (totalmodel5)** |
| touristy | -0.40 | -0.41 | -0.40 | -0.41 |
| credit | -0.28 | -0.29 |  |  |
| intimate | 0.20 | 0.20 | 0.16 | 0.15 |
| hipster | 0.19 | 0.19 | 0.12 | 0.12 |
| classy | 0.16 | 0.15 | 0.13 | 0.13 |
| upscale | 0.15 | 0.14 | 0.11 | 0.10 |
| caters | 0.15 | 0.14 | 0.09 | 0.09 |
| attire | 0.14 | 0.15 | 0.09 | 0.09 |
| lot | 0.14 | 0.13 | 0.16 | 0.16 |
| divey | 0.14 | 0.14 | 0.09 | 0.09 |
| fullbar | -0.12 | -0.12 | -0.14 | -0.14 |
| street | 0.11 | 0.12 | 0.15 | 0.16 |
| dessert | 0.11 | 0.11 | 0.07 | 0.07 |
| noisier | -0.11 | -0.11 | -0.09 | -0.09 |
| romantic | 0.11 | 0.10 | 0.07 | 0.06 |
| trendy | 0.08 | 0.08 | 0.09 | 0.09 |
| brunch | 0.08 | 0.08 | 0.06 | 0.06 |
| gfg | 0.08 | 0.08 | -0.05 | -0.05 |
| breakfast | -0.08 | -0.08 | -0.10 | -0.10 |
| delivery | -0.07 | -0.07 | -0.11 | -0.11 |
| wheelch | 0.06 | 0.05 |  |  |
| latenight | -0.05 | -0.04 | -0.14 | -0.14 |
| lunch | 0.04 | 0.04 | -0.02 | -0.03 |
| beer\_and\_wine | 0.04 | 0.04 |  |  |
| waiter | 0.03 | 0.03 |  |  |
| price | -0.03 | -0.03 |  |  |
| dinner | 0.03 | 0.03 |  |  |
| takesres | 0.03 | 0.03 | -0.05 | -0.05 |
| TV | 0.03 | 0.02 |  |  |
| valet |  |  | 0.06 | 0.06 |
| wifi |  |  | 0.02 | 0.02 |
| outdoor |  |  | -0.02 | -0.02 |
| gfk |  |  | -0.03 | -0.03 |
| takeout |  |  | -0.06 | -0.06 |

Note. All regression coefficients are unstandardized and interpreted as the change in restaurant ratings associated with a one-unit increase in the predictor. All coefficients shown in the table are statistically significant at p < .05. Blank cells indicate nonsignificant associations (magnitude and direction of coefficient not shown). Other covariates included in the models that are not shown include the effect of cities and restaurant subcategories.

# Conclusions

The aforementioned findings are consistent with the notion that restaurants may be able to boost their ratings by:

* having an ambience that is intimate, hipster, classy, upscale, divey, romantic, or trendy and NOT having a touristy ambience
* specifying nicer attire
* offering catering but NOT delivery
* having parking available in the form of:
  + a parking lot
  + street parking
* being good for dessert or brunch but NOT for breakfast or latenight meals
* being less noisy and not having a full bar

Looking holistically at the characteristics that are associated with higher vs. lower ratings, it appears that restaurants that are generally nicer and less tacky and have a better-behaved clientele have higher restaurant ratings. As noted earlier in the discussion of the limitations of this dataset, causal inferences cannot be made solely on the basis of this analysis. However, restaurants may wish to take note of these findings and consider modifying their characteristics along these lines to try to boost their ratings.

# Appendix A: Descriptive statistics and notes on restaurants from the business dataset (N = 26,729)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Frequencies** | **Notes** | | **Exclude from analysis because of high rate of missingness** |
| **Attributes** |  |  | |  |
| **Wi-fi** |  | Character variable | |  |
| free | 7455 |  | |  |
| paid | 156 |  | |  |
| no | 10771 |  | |  |
| NA | 8347 | 31.2% missing | |  |
| **Noise Level** |  | character variable—convert to numeric type and treat as ordinal variable where higher scores indicate more noise (louder) | |  |
| very\_loud | 570 |  | |  |
| loud | 1515 |  | |  |
| average | 13683 |  | |  |
| quiet | 4546 |  | |  |
| NA | 6415 | 24% missing | |  |
| **Has TV** |  |  | |  |
| TRUE | 10,788 |  | |  |
| FALSE | 10,960 |  | |  |
| NA | 4,981 | 18.6% missing | |  |
| **Outdoor seating** |  | Boolean (logical) variable | |  |
| TRUE | 10158 |  | |  |
| FALSE | 13581 |  | |  |
| NA | 2990 | 11.2% missing | |  |
| **Parking** |  |  | |  |
| **garage** |  |  | |  |
| TRUE | 1533 |  | |  |
| FALSE | 20415 |  | |  |
| NA | 4781 | 17.9% missing | |  |
| **street** |  |  | |  |
| TRUE | 3721 |  | |  |
| FALSE | 18225 |  | |  |
| NA | 4783 | 17.9% missing | |  |
| **validated** |  |  | |  |
| TRUE | 128 |  | |  |
| FALSE | 21623 |  | |  |
| NA | 4978 | 18.6% missing | |  |
| **lot** |  |  | |  |
| TRUE | 10104 |  | |  |
| FALSE | 11842 |  | |  |
| NA | 4783 | 17.9% missing | |  |
| **valet** |  |  | |  |
| TRUE | 671 |  | |  |
| FALSE | 21275 |  | |  |
| NA | 4783 | 17.9% missing | |  |
| **Takes Reservations** |  |  | |  |
| TRUE | 8604 |  | |  |
| FALSE | 15375 |  | |  |
| NA | 2750 | 10.3% missing | |  |
| **Waiter Service** |  |  | |  |
| TRUE | 13656 |  | |  |
| FALSE | 8373 |  | |  |
| NA | 4700 | 17.6% missing | |  |
| **Order at Counter** |  |  | |  |
| TRUE | 230 |  | |  |
| FALSE | 136 |  | |  |
| NA | 26,363 | 98.6% missing | | X |
| **Take-out** |  |  | |  |
| TRUE | 22,316 |  | |  |
| FALSE | 2,199 |  | |  |
| NA | 2,214 | 8.3% missing | |  |
| **Drive-Thru** |  |  | |  |
| TRUE | 1,625 |  | |  |
| FALSE | 2,046 |  | |  |
| NA | 23,058 | 86.3% missing | | X |
| **Caters** |  |  | |  |
| TRUE | 8,292 |  | |  |
| FALSE | 8,491 |  | |  |
| NA | 9,946 | 37.2% missing | |  |
| **Delivery** |  |  | |  |
| TRUE | 4,780 |  | |  |
| FALSE | 19,080 |  | |  |
| NA | 2,869 | 10.7% missing | |  |
| **Accepts Credit Cards** |  |  | |  |
| TRUE | 20,995 |  | |  |
| FALSE | 514 |  | |  |
| NA | 5,220 | 19.5% missing | |  |
| **Coat Check** |  |  | |  |
| TRUE | 171 |  | |  |
| FALSE | 2,387 |  | |  |
| NA | 24,171 | 90.4% missing | | X |
| **Wheelchair Accessible** |  |  | |  |
| TRUE | 11,159 |  | |  |
| FALSE | 1,118 |  | |  |
| NA | 14,452 | 54.1% missing | |  |
| **Good for Kids** |  |  | |  |
| TRUE | 20,257 |  | |  |
| FALSE | 4,164 |  | |  |
| NA | 2,308 | 8.63% missing | |  |
| **Dogs Allowed** |  |  | |  |
| TRUE | 676 |  | |  |
| FALSE | 2,232 |  | |  |
| NA | 23,821 | 89.1% missing | | X |
| **Open 24 Hours** |  |  | |  |
| TRUE | 18 |  | |  |
| FALSE | 222 |  | |  |
| NA | 26,489 | 99.1% missing | | X |
| **Price range** |  | integer | |  |
| 1 | 11,197 |  | |  |
| 2 | 12,071 |  | |  |
| 3 | 1,408 |  | |  |
| 4 | 296 |  | |  |
| NA | 1,757 | 6.6% missing | |  |
| **Attire** |  | Character var—this is an ordinal variable—convert to numeric where higher scores indicate greater formality of attire (1 = casual, 2 = dressy, 3 = formal) | |  |
| casual | 23,522 |  | |  |
| dressy | 770 |  | |  |
| formal | 61 |  | |  |
| NA | 2,376 | 8.9% missing | |  |
| **Ambience** |  |  | |  |
| romantic |  |  | |  |
| TRUE | 264 |  | |  |
| FALSE | 20,224 |  | |  |
| NA | 6,241 | 23.3% missing | |  |
| intimate |  |  | |  |
| TRUE | 227 |  | |  |
| FALSE | 20,261 |  | |  |
| NA | 6,241 | 23.3% missing | |  |
| classy |  |  | |  |
| TRUE | 423 |  | |  |
| FALSE | 20,065 |  | |  |
| NA | 6,241 | 23.3% missing | |  |
| hipster |  |  | |  |
| TRUE | 423 |  | |  |
| FALSE | 19,953 |  | |  |
| NA | 6,353 | 23.8% missing | |  |
| divey |  |  | |  |
| TRUE | 1,162 |  | |  |
| FALSE | 18,820 |  | |  |
| NA | 6,747 | 25.2% missing | |  |
| touristy |  |  | |  |
| TRUE | 107 |  | |  |
| FALSE | 20,381 |  | |  |
| NA | 6,241 | 23.3% missing | |  |
| trendy |  |  | |  |
| TRUE | 740 |  | |  |
| FALSE | 19,748 |  | |  |
| NA | 6,241 | 23.3% missing | |  |
| upscale |  |  | |  |
| TRUE | 157 |  | |  |
| FALSE | 20,225 |  | |  |
| NA | 6,347 | 23.7% missing | |  |
| **Music** |  |  | |  |
| dj |  |  | |  |
| TRUE | 290 |  | |  |
| FALSE | 1,779 |  | |  |
| NA | 24,660 | 92.3% missing | | X |
| background\_music |  |  | |  |
| TRUE | 590 |  | |  |
| FALSE | 323 |  | |  |
| NA | 25,816 | 96.6% missing | | X |
| Jukebox |  |  | |  |
| TRUE | 348 |  | |  |
| FALSE | 987 |  | |  |
| NA | 25,394 | 95.0% missing | | X |
| live |  |  | |  |
| TRUE | 355 |  | |  |
| FALSE | 962 |  | |  |
| NA | 25,412 | 95.1% missing | | X |
| video |  |  | |  |
| TRUE | 59 |  | |  |
| FALSE | 1,146 |  | |  |
| NA | 25,524 | 95.5% missing | | X |
| karaoke |  |  | |  |
| TRUE | 63 |  | |  |
| FALSE | 863 |  | |  |
| NA | 25,803 | 96.5% missing | | X |
| **Good For Dancing** |  |  | |  |
| TRUE | 385 |  | |  |
| FALSE | 2,266 |  | |  |
| NA | 24,078 | 90.1% missing | | X |
| **Good For Groups** |  |  | |  |
| TRUE | 21,841 |  | |  |
| FALSE | 2,782 |  | |  |
| NA | 2,106 | 7.9% missing | |  |
| ALCOHOL-RELATED VARIABLES | | | |  |
| **Happy Hour** |  | |  |  |
| TRUE | 2,266 | |  |  |
| FALSE | 444 | |  |  |
| NA | 24,019 | | 89.9% missing | X |
| **Alcohol** |  | |  |  |
| beer\_and\_wine | 3,361 | | Character variable |  |
| full\_bar | 8,285 | |  |  |
| none | 10,273 | |  |  |
| NA | 4,810 | | 18.0% missing |  |
| **BYOB** |  | |  |  |
| TRUE | 42 | |  |  |
| FALSE | 789 | |  |  |
| NA | 25,898 | | 96.9% missing | X |
| **BYOB/Corkage** |  | |  |  |
| no | 682 | |  |  |
| yes\_corkage | 144 | |  |  |
| yes\_free | 466 | |  |  |
| NA | 25,437 | | 95.2% missing | X |
| **Smoking** |  | |  |  |
| no | 1,242 | |  |  |
| outdoor | 1,279 | |  |  |
| yes | 318 | |  |  |
| NA | 23,890 | | 89.4% missing | X |
| **Good for which meal** |  | |  |  |
| breakfast |  | |  |  |
| TRUE | 1,880 | |  |  |
| FALSE | 21,095 | |  |  |
| NA | 3,754 | | 14.0% missing |  |
| brunch |  | |  |  |
| TRUE | 1,823 | |  |  |
| FALSE | 21,091 | |  |  |
| NA | 3,815 | | 14.3% missing |  |
| lunch |  | |  |  |
| TRUE | 7,717 | |  |  |
| FALSE | 15,251 | |  |  |
| NA | 3,761 | | 14.1% missing |  |
| dinner |  | |  |  |
| TRUE | 5,544 | |  |  |
| FALSE | 17,424 | |  |  |
| NA | 3,761 | | 14.1% missing |  |
| dessert |  | |  |  |
| TRUE | 713 | |  |  |
| FALSE | 22,200 | |  |  |
| NA | 3,816 | | 14.3% missing |  |
| latenight |  | |  |  |
| TRUE | 1,066 | |  |  |
| FALSE | 21,902 | |  |  |
| NA | 3,761 | | 14.1% missing |  |
| **Dietary Restrictions** |  | |  |  |
| dairy-free |  | |  |  |
| TRUE | 17 | |  |  |
| FALSE | 127 | |  |  |
| NA | 26,585 | | 99.5% missing | X |
| gluten-free |  | |  |  |
| TRUE | 9 | |  |  |
| FALSE | 135 | |  |  |
| NA | 26,585 | | 99.5% missing | X |
| vegan |  | |  |  |
| TRUE | 68 | |  |  |
| FALSE | 76 | |  |  |
| NA | 26,585 | | 99.5% missing | X |
| vegetarian |  | |  |  |
| TRUE | 52 | |  |  |
| FALSE | 92 | |  |  |
| NA | 26,585 | | 99.5% missing | X |
| kosher |  | |  |  |
| TRUE | 4 | |  |  |
| FALSE | 140 | |  |  |
| NA | 26,585 | | 99.5% missing | X |
| halal |  | |  |  |
| TRUE | 5 | |  |  |
| FALSE | 139 | |  |  |
| NA | 26,585 | | 99.5% missing | X |
| soy-free |  | |  |  |
| TRUE | 6 | |  |  |
| FALSE | 138 | |  |  |
| NA | 26,585 | | 99.5% missing | X |
| City | 298 cities | |  |  |
| State | 22 states | |  |  |
| Review\_count | M = 66.74, SD = 162.48; min = 3, max = 6,200 | |  |  |
| Stars | M = 3.48, SD = 0.76 | |  |  |
| 1 | 103 | |  |  |
| 1.5 | 504 | |  |  |
| 2 | 1,227 | |  |  |
| 2.5 | 2,659 | |  |  |
| 3 | 4,594 | |  |  |
| 3.5 | 7,036 | |  |  |
| 4 | 6,870 | |  |  |
| 4.5 | 3,196 | |  |  |
| 5 | 540 | |  |  |

1. http://www.restaurant.org/About-Us [↑](#footnote-ref-1)