QTM100_final_project

Zoe Howard-Barr, Arin Nelson, Katie Silver, Haoxiang Zhang

2023-04-24

Load the data set

We would like to start by loading the data set.

```
# Load data set into environnment
zomato <- read.csv("zomato.csv")

# Have a look of the data set
head(zomato)
```

	Restaurant.ID <int>	Restaurant.Name <chr>	Country.Code <int>	City <chr>
1	6317637	Le Petit Souffle	162	Makati City
2	6304287	Izakaya Kikufuji	162	Makati City
3	6300002	Heat - Edsa Shangri-La	162	Mandaluyong City
4	6318506	Ooma	162	Mandaluyong City
5	6314302	Sambo Kojin	162	Mandaluyong City
6	18189371	Din Tai Fung	162	Mandaluyong City

6 rows | 1-5 of 22 columns

Load packages

Then we would like to load all packages that we are going to be using.

Research question 1

Then we would want to start the operations for research question 1.

The data are rating text and having booking or not. Before all operations, it is necessary to recode the variables.

```
zomato$Rating_text <- factor( #Turn Rating text into a new factored variable
  zomato$Rating.text, # Then, specify different levels of rating text
  levels = c("Poor",
             "Average",
             "Good",
             "Very Good",
             "Excellent",
             "Not Rated"))

# Using a similar method, factorize the has table booking variable
zomato$Has_booking <- factor(zomato$Has.Table.booking,
                             levels = c("Yes",
                                        "No"))
```

Then, we can start with our operations.

We would like to start by visualizations. A bar plot is one of the good options in this case.

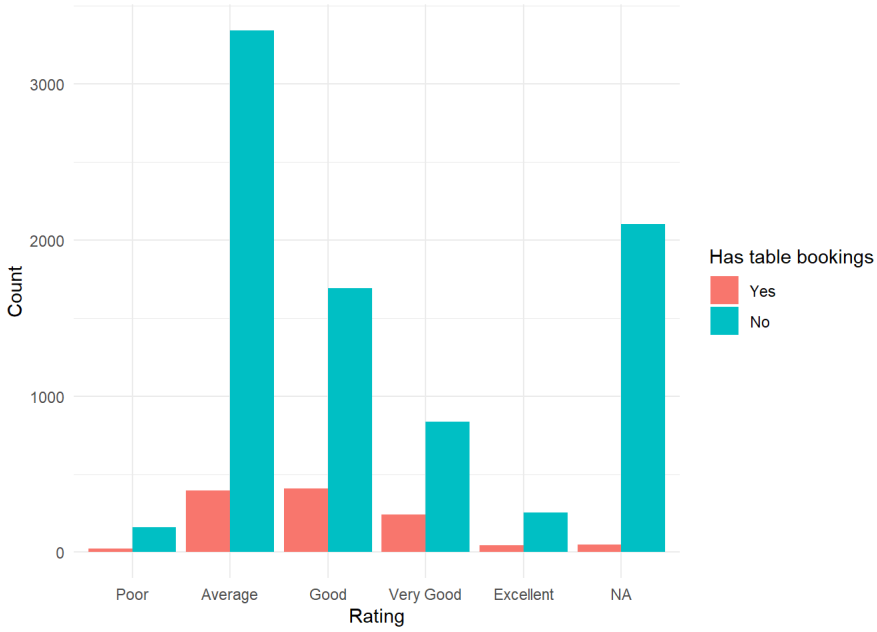
```
# Operate data into summarized form in order to place into bar plot
zomato_sum <- zomato %>%
  group_by(Rating_text, Has_booking) %>% # Group by the two variables we are interested in
  summarise(n = n()) # Return the count for each category
```

```
## `summarise()` has grouped output by 'Rating_text'. You can override using the
## `.groups` argument.
```

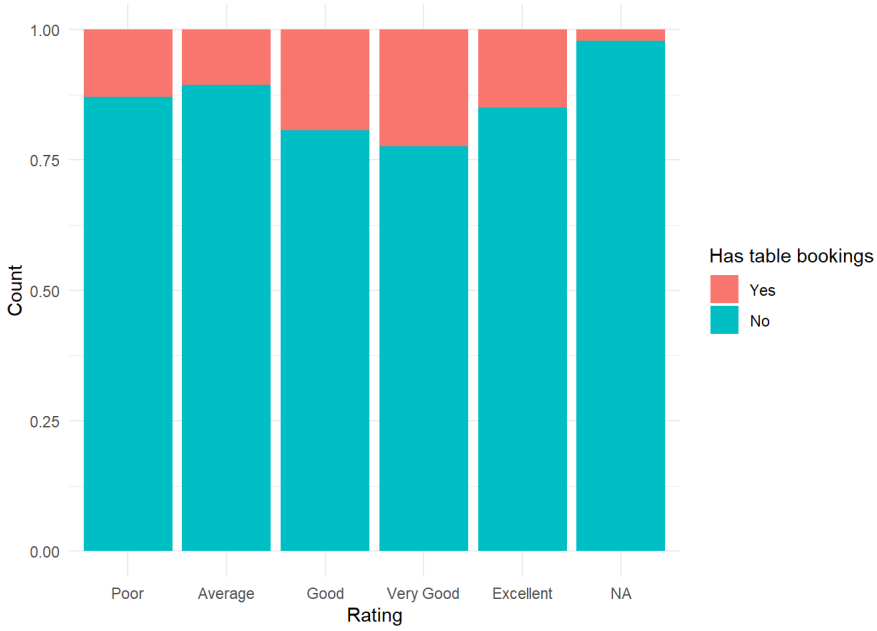
```
# Remove NA values
zomato_sum$Rating_text[complete.cases(zomato_sum$Rating_text)]
```

```
##  [1] Poor      Poor      Average   Average   Good      Good      Very Good
##  [8] Very Good Excellent Excellent
## Levels: Poor Average Good Very Good Excellent Not Rated
```

```
# Generate bar plot
ggplot(zomato_sum, aes(x = Rating_text, # Rating text is the first category
                       y = n, # Show exact count on the vertical axis
                       fill = Has_booking) # We are also interested of table bookings
       ) +

  geom_bar(stat = "identity", # Show exact value
           position = "dodge" # Produce side by side bar chart
           ) +

  #Add labels
  labs(x = "Rating", y = "Count", fill = "Has table bookings") +

  #Add a theme
  theme_minimal()
```




```
#Similar process with slight difference
ggplot(zomato_sum, aes(x = Rating_text, y = n, fill = Has_booking)) +
  geom_bar(stat = "identity", position = "fill" #Show proportion instead
           ) +
  labs(x = "Rating", y = "Count", fill = "Has table bookings") +
  theme_minimal()
```



Some tables would also help us to visualize the data.

```
#comparing variables
table(zomato$Has.Table.booking, zomato$Rating.text)
```

```
##
##       Average Excellent Good Not rated Poor Very Good
##   No     3343       256 1694      2101  162       837
##   Yes     394        45  406        47   24       242
```

Then, we would want to focus on the hypothesis testing of the research question. We would like to start by performing a chi-square test.

```
#running chi square test
chisq.test(zomato$Has.Table.booking, zomato$Rating.text, correct= F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  zomato$Has.Table.booking and zomato$Rating.text
## X-squared = 420.13, df = 5, p-value < 2.2e-16
```

Chi-squared test results indicate that there is an association between restaurant rating and whether a restaurant has table booking. Therefore we can reject the null hypothesis. Restaurants that have high ratings (Very Good or Excellent) are more likely to have table service than those who have the lowest ratings.

Despite from a chi-squared test, we would also like to perform a 2 sample t test. Though there are 5 categories, it can be divided into 2 categories, which in this case we can calculate each proportion using values in the summary table and produce a 2 sample t test.

```
# Make a two-sample test for proportions without continuity
prop.test(c(693, 418), c(2787, 3505), correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(693, 418) out of c(2787, 3505)
## X-squared = 178.78, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1100927 0.1486999
## sample estimates:
##    prop 1    prop 2
## 0.2486545 0.1192582
```

These results indicate that there is a association between restaurant rating and whether a restaurant has a table booking.($\chi 2$ =178.78, P < 2.2e -16) Therefore, we can reject the null hypothesis.

Restaurants who have above average service are more likely to have table bookings (62.38%) than those who have below average service (37.62%).

Across both tests we find restaurants that have higher ratings are more likely to have table booking options. This indicates that perhaps one factor that influences restaurant rating may be table booking. However, a majority of restaurants that have high ratings do not have table bookings. This may be due to other factors like locality, price range, or cuisine quality.

Further research could be done into other factors that influence restaurant rating or why the majority of restaurants choose not to have table booking services.

Research Question 2

We would want to proceed to research question 2.

The data are price range and aggregate rating. We would still start by recoding the variables.

```
# Recode price range into a categorical variable
zomato$Price_range2 <- factor(zomato$Price.range, c(1, 2, 3, 4))
```

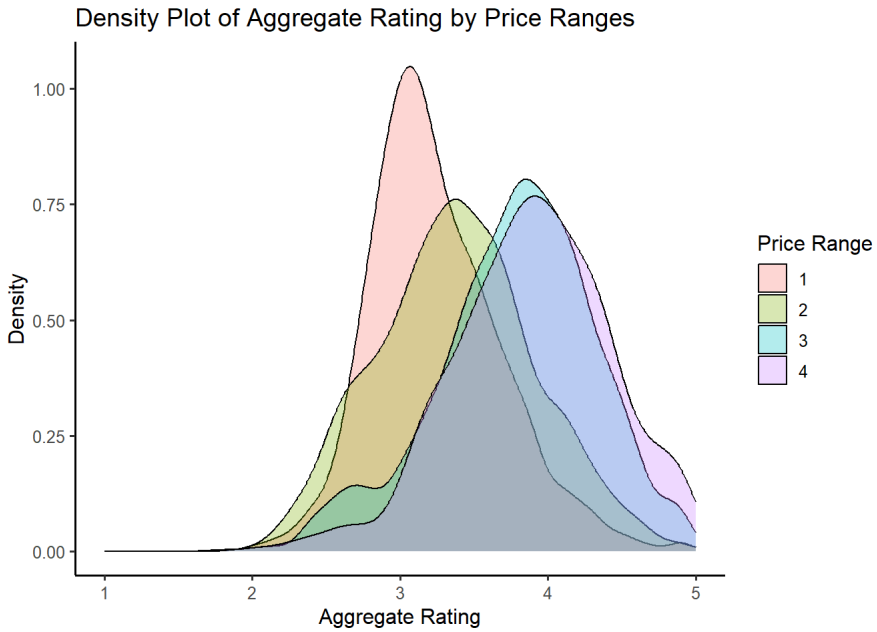
Then we would want to perform the visualizations. Firstly, we would want to start by a bar chart showing different rating levels for different price ranges.

```
# Generate bar plot
ggplot(data = zomato, aes(x = Price_range2, # Independent Variable
                          y = Aggregate.rating) # Dependent Variable
       )+

  # Specify plot type
  geom_bar(stat = "summary",
           fun = mean # Specify the stat being used is the mean
           ) +

  # Set limits
  ylim(0, 5) +

  # Set the theme
  theme_minimal() +

  # Add labels
  labs(x = "Price Range",
       y = "Aggregate Rating",
       title = "Aggregate Rating accross Different Price Ranges")
```

Despite a bar chart, we would also want to visualize the variables in more detail. Therefore, we can include a density plot for different price ranges.

```
ggplot(zomato, aes(x = Aggregate.rating, # Dependent Variable
                   # We don't need to include y value because it is density
                   fill = Price_range2) # Group by different price ranges
       ) +

  # Specify density plot
  geom_density(alpha = 0.3) +

  # Add labels
  labs(x = "Aggregate Rating", y = "Density",
       fill = "Price Range",
       title = "Density Plot of Aggregate Rating by Price Ranges") +

  # Add limits
  xlim(1, 5) +

  # Add a theme
  theme_classic2()
```

```
## Warning: Removed 2148 rows containing non-finite values (`stat_density()`).
```



Then, we would want to proceed on to the hypothesis testing.

```
#Perform Chi Square test
chisq.test(zomato$Price.range, zomato$Aggregate.rating, correct= F)
```

```
## Warning in chisq.test(zomato$Price.range, zomato$Aggregate.rating, correct =
## F): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  zomato$Price.range and zomato$Aggregate.rating
## X-squared = 3258, df = 96, p-value < 2.2e-16
```

Chi squared test results indicate that there is an association between restaurant rating and a restaurant's price range. Therefore, we can reject the null hypothesis. Restaurants that have higher price ranges (3s or 4s) are more likely to have 5 ratings (41.86% and 24.58% respectively). Restaurants with lower price ranges (1s) are less likely to have 5 ratings (10.63%).

Despite from a chi squared test, an ANOVA test may also be applicable in this situation.

```
# Perform ANOVA test
anova <- aov(Aggregate.rating ~ Price_range2, # Specify Variables
             data = zomato)

# Show results
anova(anova)
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
Price_range2	3	4442.282	1480.760598	807.0346	0
Residuals	9547	17516.996	1.834817	*NA*	*NA*

2 rows

The ANOVA test results indicate that the mean price range of a restaurant is not equal across all rating categories. Therefore, we can reject our null hypothesis and conclude that there is an association between price range and a restaurant's aggregate rating.

Therefore, we can perform a post-hoc test in order to see which group have differences.

```
# Perform Tukey test
tukey <- glht(anova, linfct=mcp(Price_range2 = "Tukey")) # Specify Tukey test

#Show results
summary(tukey)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = Aggregate.rating ~ Price_range2, data = zomato)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 2 - 1 == 0  0.94117    0.03166   29.73   <1e-04 ***
## 3 - 1 == 0  1.68349    0.04142   40.64   <1e-04 ***
## 4 - 1 == 0  1.81803    0.05953   30.54   <1e-04 ***
## 3 - 2 == 0  0.74233    0.04350   17.06   <1e-04 ***
## 4 - 2 == 0  0.87686    0.06100   14.38   <1e-04 ***
## 4 - 3 == 0  0.13454    0.06659    2.02     0.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Because the ANOVA test produced significant results, we can conduct a Tukey test to see pairwise comparisons of variables.All pairwise comparisons produced p-values that were below 0.05 except for price ranges of 3 and 4, allowing us to reject the null hypothesis that any two compared categories except of price range of 3 and 4 have equal means.

Across both tests we find restaurants that have higher ratings are more likely to have higher price ranges. This indicates that perhaps one factor that influences restaurant rating may be the price range of that restaurant. However, restaurants that have high ratings do not all have high price ranges. This may mean rating is due to a mix of other factors like location, cuisine type, publicity, etc. .

Further research could be done into other factors that influence restaurant rating or why some restaurants with low price ranges are able to achieve high ratings.

Extra credit

Since the extra credit question is also based on research question 2, we would also be including them here.

We would want to start by turning the categorical variable, in this case price range, to a dichotomous variable.

```
# Recode price range into a dichotomous variable
zomato$Price.range2[zomato$Price.range < 2.5] <- "Low"
zomato$Price.range2[zomato$Price.range > 2.5] <- "High"
```

Then we would be able to perform our two sample t test.

```
# Perform a two-sample t-test comparing the aggregate ratings of the low and high price ranges
t.test(Aggregate.rating ~ Price.range2, data = zomato, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Aggregate.rating by Price.range2
## t = 53.71, df = 6429.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group High and group Low is not equal to 0
## 95 percent confidence interval:
##  1.286594 1.384068
## sample estimates:
## mean in group High  mean in group Low
##           3.722919           2.387588
```

2-Sample T test results indicate that there is an association between a restaurant's aggregate rating and the pricing. Therefore, we can reject the null hypothesis. Restaurants that have higher price ranges are more likely to have high ratings. Restaurants with lower price ranges are more likely to have low ratings.

In the two-sample t-test, we find that there is a an association between a restaurant's aggregate rating and the pricing (high or low). Restaurants with high pricing tend to have higher ratings while restaurants with low pricing tend to have lower ratings.

This supports our research in question two, which found a positive association between an increase in restaurant pricing and higher aggregate rating.

Aligning with question 2, further research could be done into other factors that influence restaurant rating or assessing outliers in which low priced restaurants did in fact have a high rating.

Research Question 3

We would now want to move on to research question 3. Because both variables, aggregate rating and votes are numerical, there is no need to recode them into factors. We can start directly with the visualization. In this case, a scatter plot with a line of best fit would be a good option.

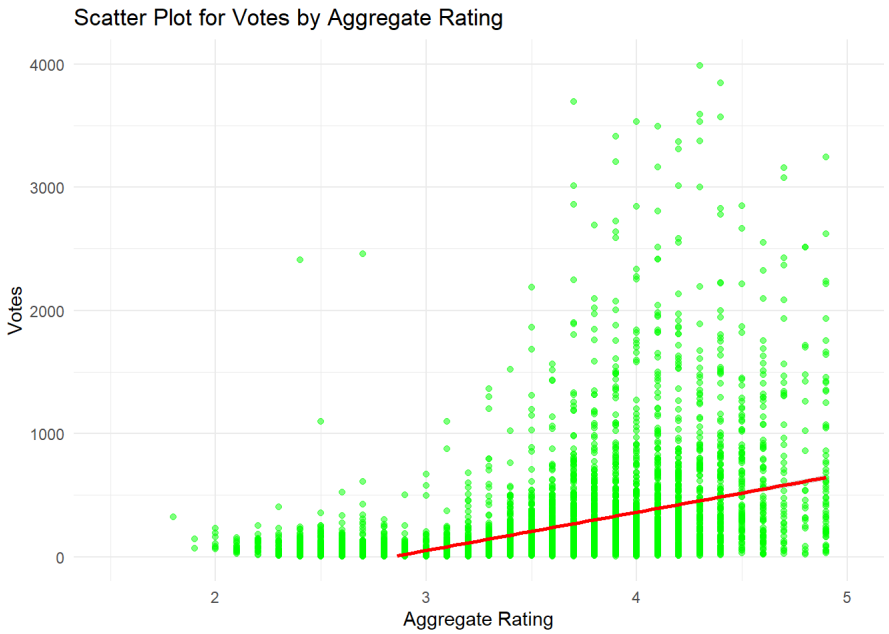
```
ggplot(zomato, aes(x = Aggregate.rating, # Independent Variable
                   y = Votes) # Dependent Variable
       ) +

  # Generate scatter plots
  geom_point(alpha = 0.5, color = "green") +

  # Generate line of best fit
  geom_smooth(method = "lm", color = "red") +

  # Add theme
  theme_minimal() +

  # Add limits
  ylim(0, 4000) +
  xlim(1.5, 5) +

  # Add labels
  labs(x = "Aggregate Rating", y = "Votes",
       title = "Scatter Plot for Votes by Aggregate Rating")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 2167 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 2167 rows containing missing values (`geom_point()`).
```

```
## Warning: Removed 27 rows containing missing values (`geom_smooth()`).
```



Then, we would want to perform the regressions. We would start with a linear regression.

```
# perform a linear regression
myfit <- lm(Votes ~ Aggregate.rating, data = zomato)

# display the regression
stargazer(myfit, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                Votes
## ------------------------------------------------
## Aggregate.rating             88.988***
##                               (2.757)
##
## Constant                    -80.366***
##                               (8.455)
##
## ------------------------------------------------
## Observations                   9,551
## R2                             0.098
## Adjusted R2                    0.098
## Residual Std. Error     408.478 (df = 9549)
## F Statistic         1,042.192*** (df = 1; 9549)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

The results indicate that there is a statistically significant positive relationship between aggregate rating and Votes. The regression coefficient for Aggregate.rating is 88.988, which is statistically significant at the 0.01 level.

The intercept is also significant at the 0.01 level, suggesting expected -80.366 votes for a rating of zero. However, this can be unrealistic because we cannot get a negative number of votes.

The R-squared value for the model is 0.098, which means that 9.8% of the variation in Votes is explained by the Aggregate rating variable. The low value of R-squared suggests that a linear regression with a single variable of aggregate rating may not fully explain the votes variable.

In order to perform a better regression, we added a squared variable into the regression.

```
# Generate the squared variable
zomato$rating_sq <- zomato$Aggregate.rating ^ 2

# perform the regression
myfit <- lm(Votes ~ Aggregate.rating + rating_sq, data = zomato)

# display the regression coefficients
stargazer(myfit, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                Votes
## ------------------------------------------------
## Aggregate.rating            -269.676***
##                               (9.564)
##
## rating_sq                    92.291***
##                               (2.371)
##
## Constant                      7.625
##                              (8.174)
##
## ------------------------------------------------
## Observations                   9,551
## R2                             0.222
## Adjusted R2                    0.222
## Residual Std. Error     379.499 (df = 9548)
## F Statistic         1,361.231*** (df = 2; 9548)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Adding a squared variable of aggregate rating improves both the R-squared score to 0.222 and the F statistic from 1042.192 to 1361.231, and the coefficient for the squared variable is also statistically significant at the 0.01 level.

This means that a squared variable may better explain the relationship between aggregate rating and votes.

The results from our linear regression test indicate that there is evidence that the correlation between the number of votes a restaurant receives and its aggregate rating are not equal to zero. This result may indicate that an increase in votes leads to an increase in rating.

This may mean restaurants try to encourage patrons to vote on the restaurant in hopes of increases its overall rating.

Discussion

Our results look at how restaurants are perceived by the public by looking at restaurant price, rating, and the number of votes a restaurant received. We explored how these factors were interconnected and found statistically that there are intrinsic relationships between them.

If we were to re-explore this topic, perhaps a more extensive sample size or more time to delve deeper into the facets of our data set would have allowed us to pull more conclusions. We could have explored cuisine type and location as factors that influence the ways in which a restaurant is regarded. Perhaps it would be possible to statistically discern what makes up a quality restaurant.

Work Cited

Alexandrov, Alexei, and Martin Lariviere. “Are Reservations Recommended?” Kellogg Insight, 10 May 2019, https://insight.kellogg.northwestern.edu/article/are_reservations_recommended (https://insight.kellogg.northwestern.edu/article/are_reservations_recommended).

Christen. “Importance of Restaurant Online Reviews.” Revolving Kitchen, 31 May 2022, https://revolvingkitchen.com/2022/04/29/how-important-are-online-restaurant-reviews/# (https://revolvingkitchen.com/2022/04/29/how-important-are-online-restaurant-reviews/#):~:text=Restaurants%20Serve%20More%20Customers%20Than%20Other%20Businesses&text=Many%20customers%20have%20opinions%20about,m

Finder, Chuck. “Price Is Ripe: Study Finds Increase in Menu Prices Means Decrease in Restaurant Ratings - the Source - Washington University in St. Louis.” The Source, 9 Feb. 2021, https://source.wustl.edu/2021/01/price-is-ripe-study-finds-increase-in-menu-prices-means-decrease-in-restaurant-ratings/ (https://source.wustl.edu/2021/01/price-is-ripe-study-finds-increase-in-menu-prices-means-decrease-in-restaurant-ratings/).

Masset, Philippe, and Jean-Philippe Weisskopf. “Online Customer Reviews: Their Impact on Restaurants.” Hospitality News & Business Insights by EHL, https://hospitalityinsights.ehl.edu/online-customer-reviews-restaurants (https://hospitalityinsights.ehl.edu/online-customer-reviews-restaurants).

Montti, Roger. “Research Exposes Role of Pricing on User Ratings.” Search Engine Journal, 6 Jan. 2021, https://www.searchenginejournal.com/how-consumer-ratings-may-be-biased-by-price/391813/ (https://www.searchenginejournal.com/how-consumer-ratings-may-be-biased-by-price/391813/).