JMP Instructional data set 010

Coffee Shops Yelp Reviews

JMP STATISTICAL DISCOVERY

JMP platforms and statistical methods

Distribution: Bar chart; frequency table
Tabulate: Table of summary statistics
Text Explorer: Analysis of text data
Graph Builder: Scatter plot

Problem statement

Reading customer reviews of cafés and restaurants have become a common way people decide where to go. Platforms like Yelp are among the most widely used sources for sharing and reading these experiences. To better understand the perceptions and sentiments of both residents and visitors toward coffee shops in Austin, Texas, a dataset was compiled consisting of 4,466 reviews across 72 establishments [1].

Some goals of the analyses will be to:

1. Develop a general assessment of how patrons feel about these establishments.
2. Identify common attributes that patrons comment on and positive/negative sentiments associated with them.
3. Identify the coffee shops that receive some of the best/worst ratings and uncover some of the reasons for those ratings.
4. Develop a thorough understanding of the experience and sentiment that patrons have for a specific coffee shop.

Data

IDS_010.jmp

ID	ID of the reviewer
Coffee Shop	Name of the coffee shop
Rating	Numerical rating (1,…, 5) of the coffee shop by the reviewer
Rating (Ordinal)	Rating as an ordinal variable
Reviews	Written reviews posted to Yelp

Exercises

1. Create a Bar Chart and Frequency Table of the Ratings. Provide a brief sentence or two describing what is learned from these results.

   Instructions:
   Analyze > Distribution. Place 'Rating (Ordinal)' in the Y, Columns role. Click OK.

2. Create a Bar Chart of the average rating for each coffee shop. How many coffee shops have an average rating of 4.5 or higher? Do any coffee shops have a perfect score of 5? If so, how many reviews does that coffee shop have? How many shops have an average rating of 4.0 or lower? What coffee shop has the lowest average rating? How many reviews does that shop have? Are there any people who gave that shop a rating of 5? If so, what percentage did so? What is the most common rating given to that shop?

   Instructions:
   Bar chart of the average rating for each coffee shop:
   Use Graph > Graph Builder. Place 'Rating' on the X axis, and 'Coffee Shop' on the Y.
   Select the bar chart symbol in the graph palette. 

   Right click on the Y axis and choose Order > 'Rating' ascending.
   Choose to Label by Value in the Label drop down menu in the controls on the left.
   Choose Label Format of Fixed Dec: 1.

   Bar chart of the number of reviews for each coffee shop:
   Use Graph > Graph Builder. Place 'Coffee Shop' on the Y axis. Select the bar chart symbol in the graph palette. Choose to order in ascending order and add the data values.

   To view the distribution of ratings for each coffee shop:
   Return to the Distribution report created in Exercise 1. Under the top red triangle, choose Local Data Filter. Select 'Coffee Shop' as the variable and click the "+" symbol.
   Now individual (or groups) of coffee shops can be selected and the Bar chart and Frequency table will be updated using only the ones selected.

3. Examine the relationship between the number of reviews the coffee shops received vs. their average rating. Does there appear to be a correlation between these two variables? In other words, do shops that receive high or low average rating tend to have more/less number of reviews? Describe any relationship you see in a few brief sentences citing some of the shops that receive the least and most number of reviews. What is a potential reason for these results?

Instructions:
Table of the number of reviews and average rating for each coffee shop:
Analyze > Tabulate. Place 'Rating' in the 'Resulting Cells' area. Place the statistics 'N' and 'Mean' on top of the 'Sum' label to replace it. This small table shows the total number of reviews and the average rating for all those reviews. Finish by placing the variable 'Coffee Shop' in the Row Drop Zone.

New JMP data table for analysis:
Under the drop red triangle, select 'Make into Data Table'

Scatterplot to examine relationship:
Graph > Graph Builder. Place 'Mean(Rating)' on the Y axis, and 'N(Rating)' on the X. Choose to add a linear regression line to the graph by selecting the Line of Fit icon in the graph palette. 

4. Analyze the written reviews by performing the following analyses.

   a. Create a table showing the frequency of words and phrases.

Instructions:
Analyze > Text Explorer. Place 'Reviews' in the Text Columns. Click OK.

As many of the words/phrases used in the reviews don't provide any actual assessment of how people feel about these establishments, it's important in analyzing text data to focus on only those words/phrases that do. As a result, a common first step in analyzing text data is to remove those noninformative words. For example, note that the word 'coffee' is the most frequently used words (8,259). This word is not helpful in understanding customers' sentiment as reviewers use this word simply because they are reviewing coffee shops. To remove this word, right-click on the word and select 'Add Stop Word'. Choose to remove other words that you feel are not helpful in understanding customers' sentiment, targeting only those words that occur at least 750 times.

Similar words can also be combined. For example, the word 'food' occurred 5,701 times and the word 'foods' occurred 875. Highlight both words, then right-click and choose Recode. In the dialog box that appears, highlight the two words and select Group, the Recode.

Add a word cloud by selecting Display Options > Show Word cloud. Chose centered layout by selecting that option under the red triangle next to the term Word Cloud.

b. Examine the 10 words used the most. What does this reveal about the themes reviewers comment on the most, regardless of if it's expressing positive or negative sentiment.

c. Examine the most common words used for those who gave a rating of a 1. Do the same for those that gave a rating of 5. What are some similarities and differences regarding themes these two groups of reviewers comment on?

> ***Instructions:***
> *Choose Local Data Filter under the red triangle. Select 'Rating (Ordinal)' as the variable and select the "+" symbol. Now the data can be filtered to only use reviewers that provided a select rating.*

d. Select one of the coffee shops with the top 5 highest average rating. Examine the words used most frequently. Do the same for one of the 5 lowest rated shops. What are reviewers predominantly commenting on in these shops and what insights does this provide regarding what patrons are and/or are not happy about?

> ***Instructions:***
> *From the open report, click the 'And' button in the Local Data Filter. Add 'Coffee Shop'. Now the data can be filtered to select shops and/or reviewers that provided a select rating.*
>
> *To see the full review for those that contain a certain word, right-click on a word and choose Show Text.*

5. Perform an analysis of the ratings and reviews for one coffee shop (not examined thus far and with at least 50 reviews). Create 2-3 presentation slides containing 2-3 graphs and 5-10 bullet points summarizing some of the key findings.
   (e.g., Summarize the distribution of the ratings? What are some things patrons are most pleased with? What are some common complaints? How do patrons generally feel about the quality of the coffee, food, vibe, service, seating, etc.? Provide some recommendation on ways they can improve).

Supporting materials

1. Data is derived from data.world *'Austin Coffee Yelp Reviews'*
   https://data.world/rdowns26/austin-coffee-yelp-reviews

** Please adhere to any citation requirements for distribution and use of these data.*