Milestone 3: Beyond Descriptive Stats

Dive Deeper

Look deeper into the features you are investigating, consider:

- Relationships / Correlation, Pearson Correlation
- Linear Regression for future prediction (if the relationship is linear)
- Textual Analysis for TF-IDF (Term Frequency-Inverse Document Frequency; Row-based and column-based, stop-word removal?

Specify 1-2 correlations you discovered. List the fields that you found to be correlated and describe what you learned from these correlations.

1. The data is read using Yelp's business dataset, and filtered to "coffee" related company only to match our customer business category. We determine and plot the distribution number of reviews distributed with respect to the star rating. With this, we discover there is a left-skew distribution, with most review ratings being at 4 to 4.5. What I can induce from this correlation is that coffee related company generally having higher ratings.
2. In addition, we attempt to find relationship between star ratings and the length of the review. The result shown that there is no correlation between these two properties.
3. Besides, we determine and plot the distribution number of reviews distributed with respect to whether the company is open or not. We discovered that closed company having lower rating than the company that are currently operating. The result indicated that company having lower review rating tends to close.

Go Broader

Expand the features you are investigating. Look for connections/relationships that you may have initially missed.

1. What jumps out at you now?
2. Use the descriptive stats to point you to features that you may now want to consider.

What key terms did you discover in any text analysis, for whom? Any themes? If you are not analyzing text, summarize what other things you are considering in your analysis?

To suggest the company location for CoffeeKing, rather than the location with numerous coffee companies, which come with huge competition, I suggest that we choose a location that has lowest review rating for other coffee companies. In addition, we need to understand the context of the review to apply or avoid in CoffeeKing.

- Location:
    - To make sure the sample size is sufficient to make sense of our analysis, we put a condition that only those cities with more than 500 reviews will be extracted.
    - We extracted city with lowest review rating which is Bryn Mawr with average rating of 3.09. Hence, CoffeeKing can open a new shop at Bryn Mawr with features that customer want.
- Operating hour:
    - We extracted operating hours with most review and highest average review ratings.
    - The result indicates that 6am to 6pm is a suitable operating hour for CoffeeKing.

New Metric

Create 1 or 2 new metrics to track relationships of data you discovered. Explain why you created them.

After identify the location and operating hours, we would like to know the features for good and bad coffee shops in order to provide insights to CoffeeKing. Topic modelling helps to pick specific topics from the huge volume of text. In our case, it helps us to understand the reviews content.

Hence, the following are the metrics we are going to analyze:

1. Frequency of occurrence of the word in the high rating review.
   - Results: ‘coffee’, ‘great’, ‘place’, ‘food’, ‘breakfast’, ‘delicious’, ‘fresh’, ‘service’.
2. Frequency of occurrence of the word in the low rating review.
   - Results: ‘food’, ‘order’, ‘coffee’, ‘time’, ‘never’, ‘place’, ‘drive’, ‘minutes’