BEER CHALLENGE

DAYOUNG KIM

1. RANK TOP 3 BREWERIES WHICH PRODUCE THE STRONGEST BEERS?

- Sort beer by ABV value
- The top threes are breweryId = 65133, 35, 16866

2. WHICH YEAR DID BEERS ENJOY THE HIGHEST RATINGS?

- If the data (row) doesn't have 'review_overall', fill with median value
- Group by 'review_time' and to have the value of average of 'review_overall'
- Sort by 'review_overall'
- The times would be '898560001, 904089601, 973014566'

3. BASED ON THE USER'S RATINGS WHICH FACTORS ARE IMPORTANT AMONG TASTE, AROMA, APPEARANCE, AND PALETTE?

- Get the data33 = beerdata[['review_taste', 'review_aroma', 'review_appearance', 'review_palette', 'review_overall']]
- Fill NaN values with median
- Use Linear Regression, assuming linear relationship
- Get the coefficients for each feature
- Aroma is the most important feature because it has the highest coefficient value, which is 0.5531015

4. IF YOU WERE TO RECOMMEND 3 BEERS TO YOUR FRIENDS BASED ON THIS DATA WHICH ONES WILL YOU RECOMMEND?

- Check if beer ID or beer name is null in any row
- Fill all NaN value with median values
- ‘review_count' column will show how many times the beer was rated
- average review_overall for each beer
- Get the mean of all reviews
- Get the minimum number of ratings
- Use review_count and avg_review_overall to calculate the score of each beer
- Group by ‘beer_name’ and get the scores
- According to both tables, I recommend Sierra Nevada Celebration Ale, Sierra Nevada Pale Ale, Founders Breakfast Stout

5. WHICH BEER STYLE SEEMS TO BE THE FAVORITE BASED ON REVIEWS WRITTEN BY USERS?

- Drop NaN values
- Approach 1
  - Round Review for Random Forest classification
  - Check the accuracy to move forward
  - Calculate score as question #4
- Approach 2
  - Sentiment Analysis with LSTM
  - After checking distribution, set threshold for being positive as 4.5
- American IPA 43369 American Double / Imperial IPA 26106 American Double / Imperial Stout

6. HOW DOES WRITTEN REVIEW COMPARE TO OVERALL REVIEW SCORE FOR THE BEER STYLES?

- Calculate ‘score’ as question #4 for ‘beer_style’
- Compare with two outputs of question #5
- It seems like both are similar to ‘overall review score output’

7. HOW DO FIND SIMILAR BEER DRINKERS BY USING WRITTEN REVIEWS ONLY?

- Check the distinct number of users, styles, and beer names
- Approach 1. Naïve Bayes
  - Use Naïve Bayes to see if the text reviews can predict the beer_name or beer_style
  - The performance is low
- Approach 2. LSTM
  - Use LSTM to see if the text reviews can predict beer_style
  - The performance is low (stopped at epoch 1)

7. HOW DO FIND SIMILAR BEER DRINKERS BY USING WRITTEN REVIEWS ONLY?

- Approach 3. Kmeans (Unsupervised Learning)
  - Instead, used unsupervised learning to make clusters of users after combining all 'review_text' for each user
  - Use Elbow Method to decide the K value
  - Check the performance and visualize the clusters
  - If there are multiple users to check the similarity, use the Kmeans model to see if they can exist in the same cluster
    - 1) Conduct 'user_profileName' clustering
      - Clustering upon users
    - 2) Conduct 'review_text' clustering
      - Clustering possible depending on text review inputs
      - If able to have more clusters visualized via scatter plot and calculated 'silhouette score' or else, better performance and output could have happened
      - (Tried Elbow Method to 30 clusters, but hard to find "elbow")

THANK YOU FOR READING