EDA Approach

- In **RAW_recipes** file the top five rows are:
    1. arriba baked winter squash mexican style
    2. a bit different breakfast pizza
    3. all in the kitchen chili
    4. alouette potatoes
    5. amish tomato ketchup for canning
- Now, we have extracted seven individual columns from an array for nutrition alongwith some test cases.
- We have standardize the nutrition values based on calories.
- Converted tag column from string to array of strings.
- Joined Recipe Data to Review Data

- Now, we read the second file i.e **RAW_review** file.
- We join both the data frames using recipe_id.
- Converted date to Datetype().

- Setting up EDA
- Defined Custom Functions
- Bucketing and cleaning numerical functions which resulted in low ratings for recipe older than 6 years, recipes less than 2 steps are high rated and recipes more than 29 steps are again low rated.

- Feature Extraction – Here we observed top and most rated tags, bottom and least rated tags.

What can be done to improve the recommendation?

- The count and percentage of positive, negative reviews can be used for better recommendation results.
- A more sophisticated formula like the Bayesian average can be used for calculating the popularity score, as the used formula is biased towards the number of votes.
- For better recommendations, more advance methods like word embeddings or CountVectorizer can be used for encoding text to vector (or feature extraction)