What Do Successful Recipes Have in Common?

Capstone 1 Project Report

By Anna Kantur

Springboard

Introduction

What do successful recipes have in common?

- Is it nutritional value?
- Number of steps in a recipe?
- Time to cook?
- A specific ingredient?
- Something else?

Who might care?

- Grocery store owners who decide what products to put on display for advertisement
- Professional and amateur cooks who want to create the best recipes
- Meal kit services who are want to offer best recipes

Dataset

- 180K+ recipes and 700K+ recipe reviews covering 18 years of user interactions and uploads on Food.com (formerly GeniusKitchen)
- Kaggle:

https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions

Data Wrangling

CSV
'RAW_recipes.csv' and
'RAW_interactions.csv'

pandas
recipes_df
reviews_df

Remove Irrelevant Data and Missing Values
210,244 recipes, but only 18,762 with more than 10 reviews
Keep only > 10 reviews

0 star ratings
present in 58% of unique recipes and represent 6% of all the reviews. Allowed on Food.com
Keep them

Recipe Tags
Wrangle the data to have each tag as a separate column and 0/1 if present/not present in the recipe
Create tags_matrix data frame

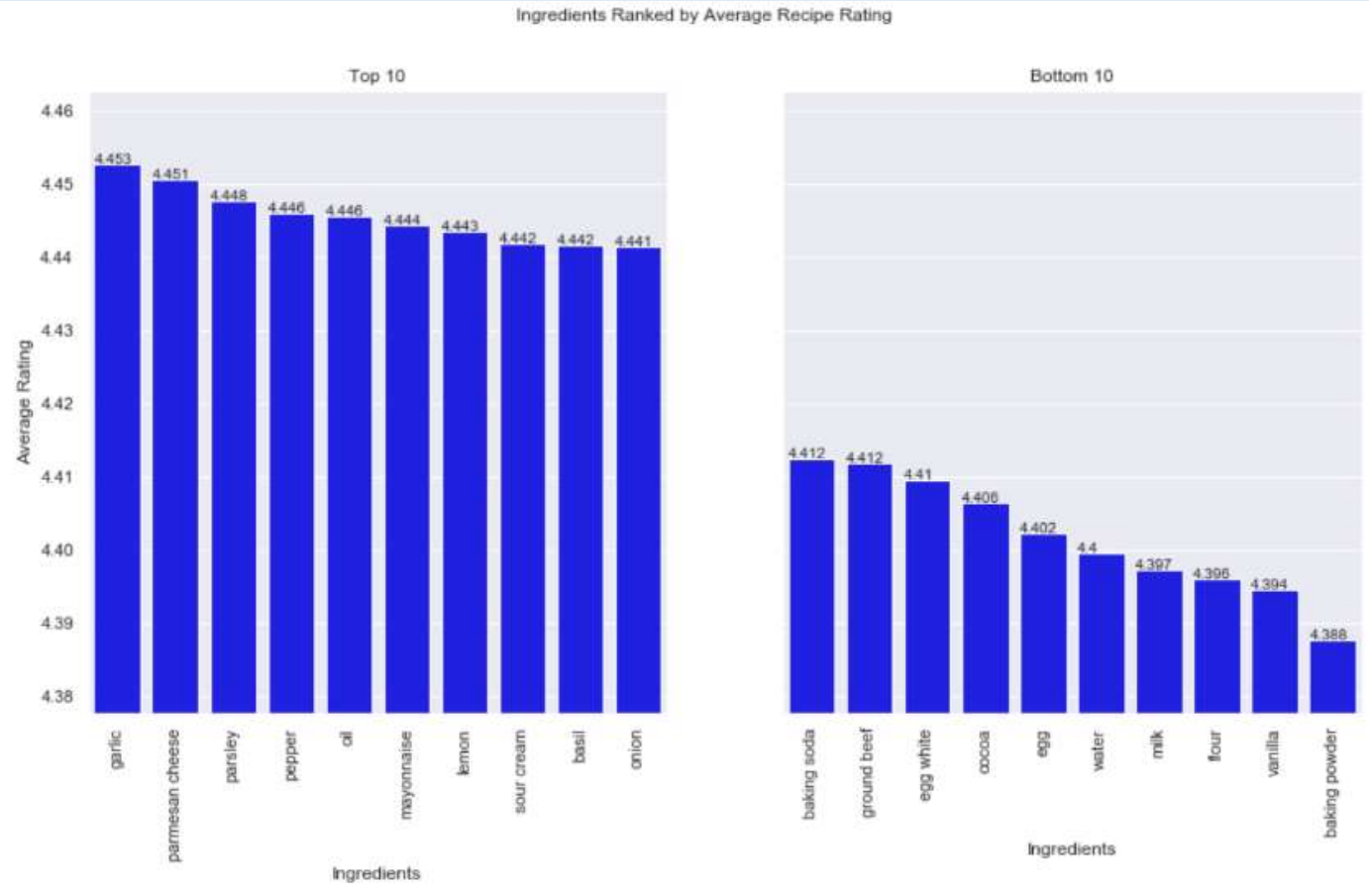
Recipe Ingredients
Remove duplicate names, e.g. cilantro and fresh cilantro
For ingredients in > 100 recipes have each ingredient as a separate column and 0/1 if present/not present in the recipe
Create ingredients_matrix data frame

Data frames for further use:
- recipes_df
- reviews_df
- ingredients_matrix
- tags_matrix

Data Summary

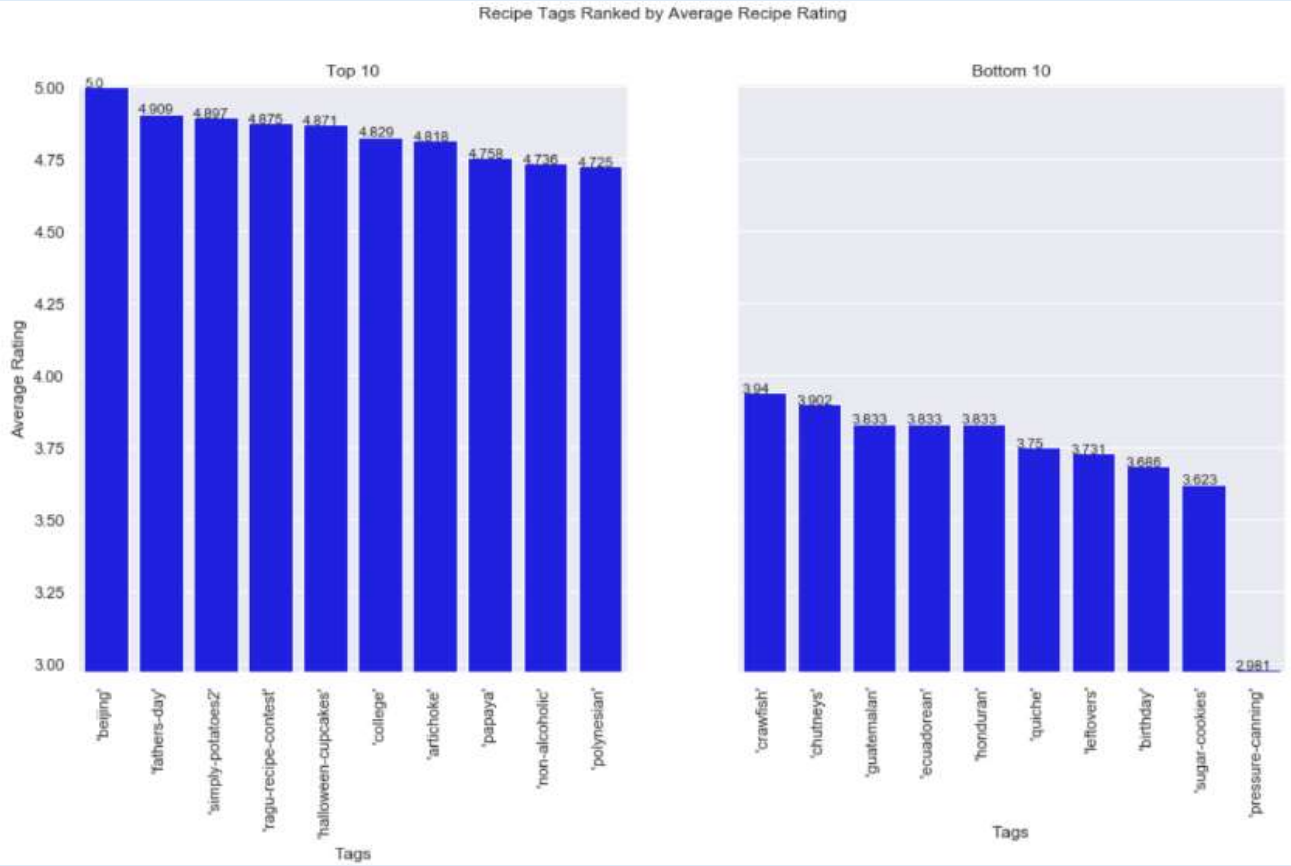
- Average recipe rating = 4.435
- 75% of reviews are 5 stars
- 495 unique recipe tags
- 149 unique recipe ingredients

Best and Worst Ingredients



- Garlic is the best ingredient (average recipe ranking of 4.453)
- Baking powder is the worst ingredient (average recipe ranking of 4.388)

Best and Worst Tags

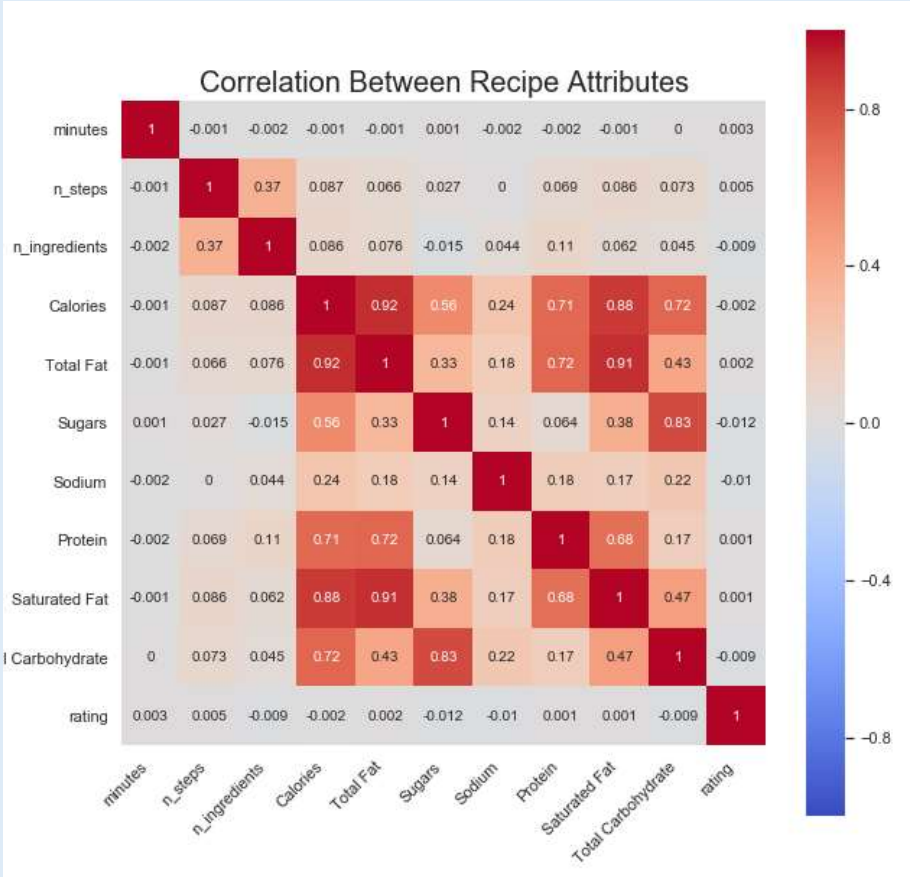


- Beijing is the best tag (average recipe ranking of 5.000)
- Pressure canning the worst tag (average recipe ranking of 2.981)

Nutritional Values Correlation Matrix

- There is a strong correlation between calories content and total fats (saturated fats specifically), protein and carbohydrates.

- However, there is no strong correlation between the recipe rating and any of the mentioned recipe attributes



Statistical Analysis

t-test results:

- **Garlic**

p-value = 9.475783077096418e-18

- **Baking Powder**

p-value = 9.475783077096418e-18

- **‘Beijing’ recipe tag**

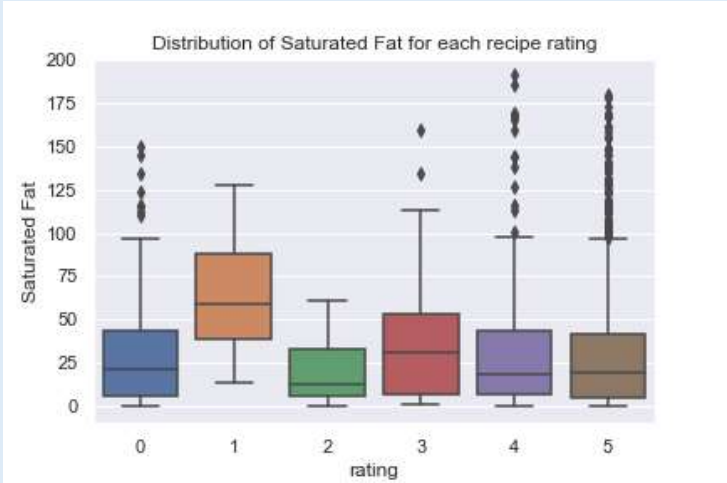
p-value = 0.13

- **'pressure-canning' recipe tag**

p-value = 2.792736995754129e-10

- **Saturated Fats Content**

p-value = 0.03

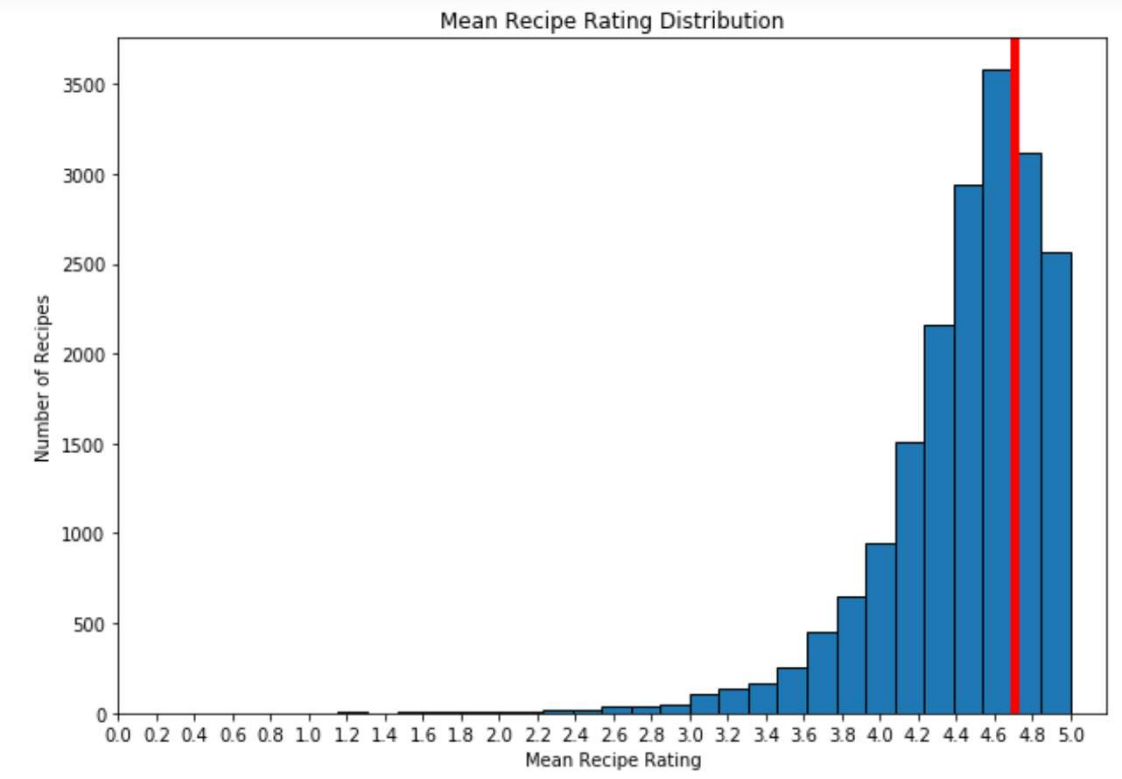


Machine Learning

Predictive Modelling

- We choose 4.7 average recipe rating as a divide for “good” vs “bad” recipes

- There are 5,519 “good” recipes and 13,302 “bad” recipes in our dataset.



Machine Learning

Predictive Modelling – Random Forest

- Top 20 Features:

0 ('Calories', 0.03599724473589182)
1 ('Sugars', 0.034046714261403835)
2 ('Protein', 0.032015894747727876)
3 ('minutes', 0.031239016224030726)
4 ('Saturated Fat', 0.03111097618770412)
5 ('Sodium', 0.030477369009240 14)
6 ('Total Fat', 0.030016450811690142)
7 ('Total Carbohydrate', 0.028400197936175275)
8 ('n_steps', 0.02699280475047596)
9 (" 'easy'", 0.007238298792856107)
10 (" 'equipment'", 0.006913044424659004)
11 (" 'occasion'", 0.006421651924366628)
12 (" 'number-of-servings'", 0.0061877163666479535)
13 (" 'cuisine'", 0.006035790961877401)
14 (" 'oven'", 0.005836153491809918)
15 ("'60-minutes-or-less'", 0.005774535718058847)
16 (" 'low-in-something'", 0.0056769836799905275)
17 ('salt', 0.005489661715824159)
18 (" 'taste-mood'", 0.005294530153120957)
19 (" 'inexpensive'", 0.005267690999395954)

Machine Learning

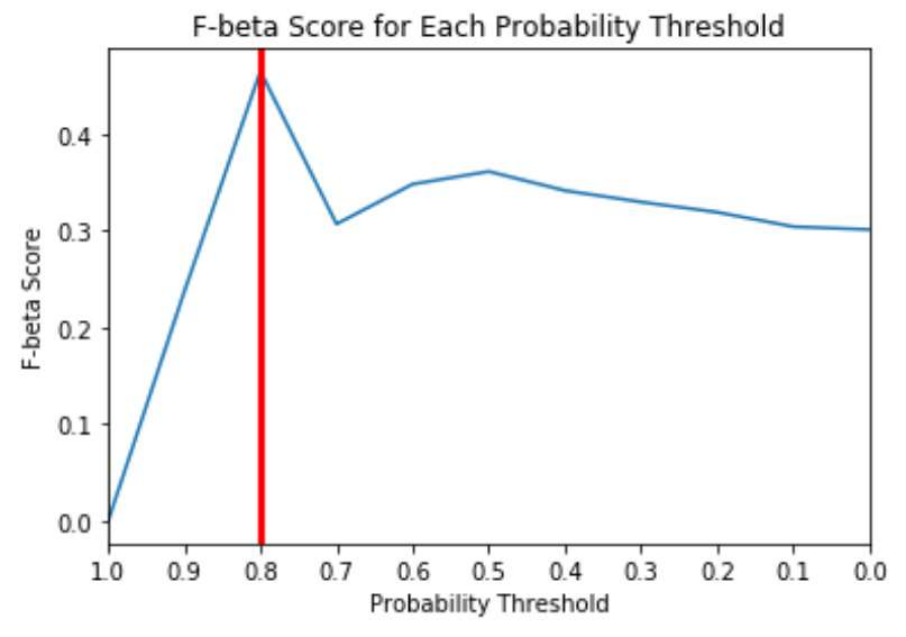
Predictive Modelling – Best Model

Model	**Best Score**	**Best Parameters**
Random Forest	62% ✓	{'max_depth': 11, 'max_features': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0, 'n_estimators': 100}
Random Forest with 20 Best Features	57%	{'max_depth': 7, 'max_features': 5, 'min_samples_leaf': 3, 'min_samples_split': 5, 'min_weight_fraction_leaf': 0, 'n_estimators': 100}
Logistic Regression	54%	{'C': 8.483428982440725e-05, 'class_weight': None}
KNN	54%	{'n_neighbors': 20, 'weights': 'distance'}

Machine Learning

Predictive Modelling – Thresholding Probability

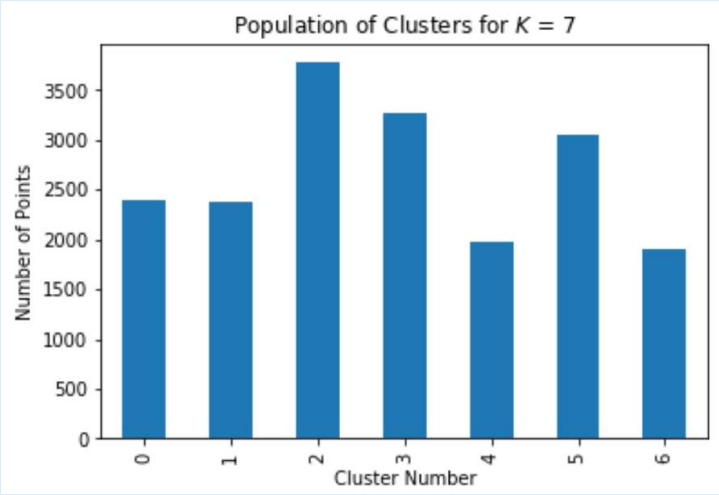
- The optimal probability threshold = 0.8, at which the F-beta score = 46%.

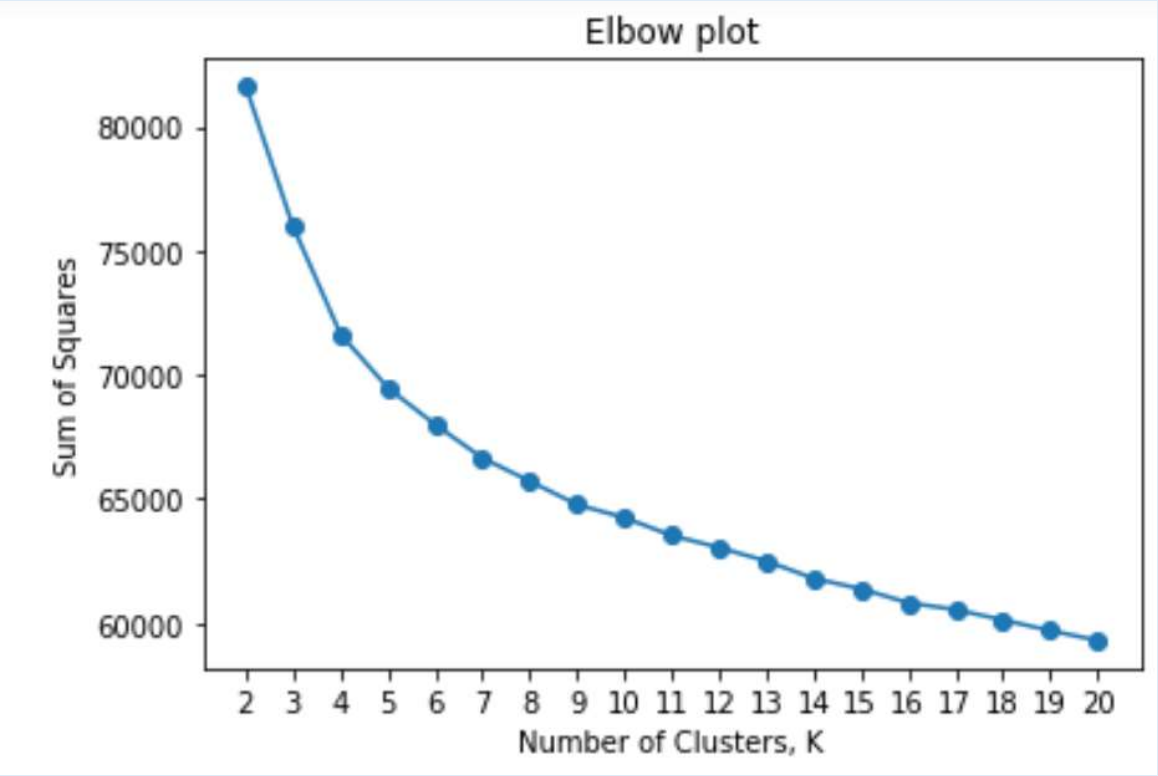


Machine Learning

Clustering – The Elbow Method

- The best K is somewhere between 3 and 7.
- We try K=7:



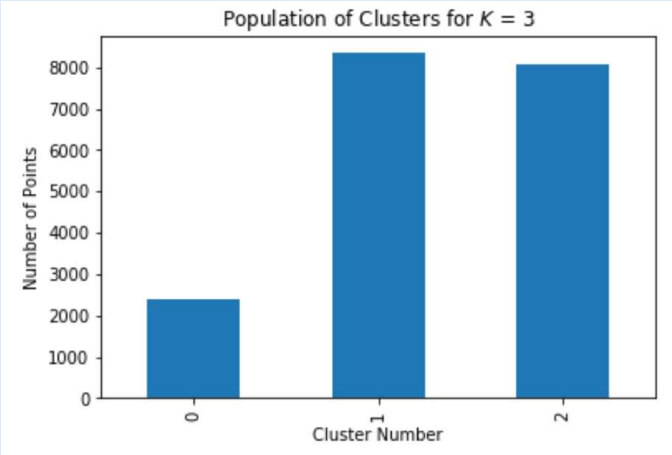


Machine Learning

Clustering – The Silhouette Method

For K between 4 and 7 we get the following results:

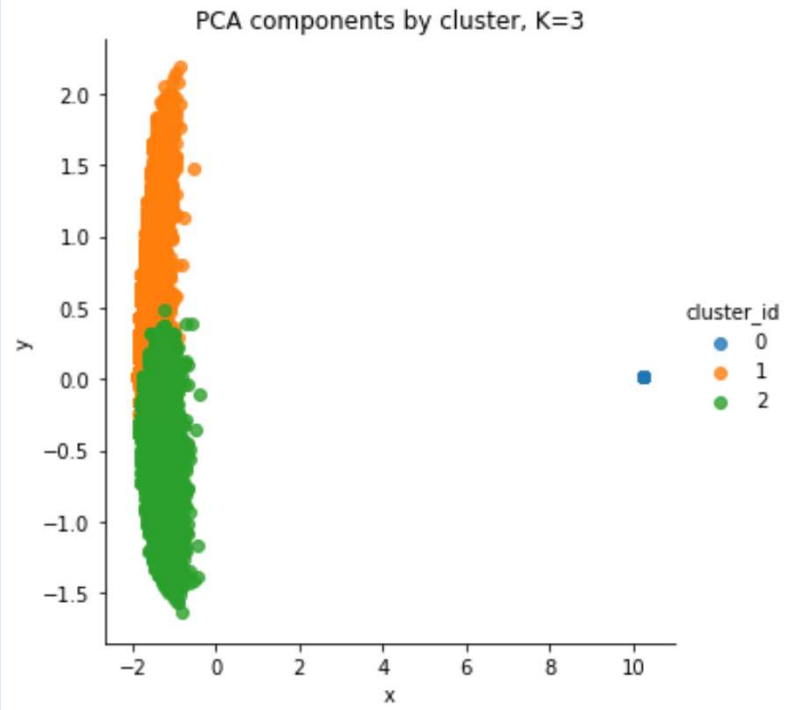
- For n_clusters = 3 The average silhouette_score = 0.19172585856463387
- For n_clusters = 4 The average silhouette_score = 0.18957449729577003
- For n_clusters = 5 The average silhouette_score = 0.18111405590654364
- For n_clusters = 6 The average silhouette_score = 0.16267324366196434
- For n_clusters = 7 The average silhouette_score = 0.1632525460602517



Machine Learning

Clustering – PCA Dimension Reduction

- We use PCA to reduce the dimensionality of our data from 149 dimensions (e.g. ingredients) to 2 dimensions



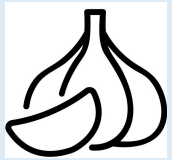
Machine Learning

Clustering – Results

- Cluster 1 Desserts/Baking
- Cluster 0 Mid Day Meal
- Cluster 1 Dinner

	recipe_name	top_10_ingredients	Calories	Sugars	Sodium	Protein	Saturated Fat	Total Carbohydrate
0	[da best chicago style italian beef, i hate m...	[sesame seeds, thyme, ginger, tabasco sauce, f...	1285145.0	297944.0	74806.0	77959.0	119762.0	48137.0
1	[chile rellenos, chinese candy, healthy for t...	[sugar, salt, butter, egg, flour, milk, water,...	3850483.8	911326.0	226026.0	224000.0	358662.0	148227.0
2	[chicken lickin good pork chops, grilled ve...	[salt, garlic, onion, pepper, oil, butter, wat...	3384196.1	224985.0	322173.0	365498.0	329119.0	79829.0

Recommendations



- Use garlic in recipes, people love it
- Avoid recipes requiring pressure-canning or high amounts of saturated fats
- For grocery store owners it might be wise to arrange the products around the three recipe clusters that we found: Desserts/Baking Cluster, Mid Day Meal Cluster, and Dinner Cluster
- To get an idea of how much people would like your recipe, think about nutritional values and how long it takes to make the recipe

Ideas for Further Research

- Further exploration of descriptive data not used in this Capstone project (recipe description, detailed descriptions of each recipe step) can yield more insights and predictive value
- More data on lower rated recipes can be collected (the current data has 75% of all reviews with 5 star ratings)

Thank You!

Anna Kantur

https://www.linkedin.com/in/annakantur/

https://github.com/a-kantur