Data Mining Project - Final Report

Exploratory data analytics and predictive modelling on data from *Food.com*

Submitted by :

Anit Gupta

Susanth Sampath Kumar Dasari

Sonam Dawani

In this analysis, we are performing exploratory data analytics and predictive modelling to solve some business needs we identified in Food.com and also solutions that are helpful to their customers.

We are using data from the following kaggle project: https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions (https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions)

Food.com is a place where you can find recipies for all ocassions. It is a social networking platform for people who like to try new recipes and people who like to make new recipes.

The website has a lot of features that attract people and retain them. There are sections where you can find ratings and reviews for the recipes which makes it perferct for people to double-check that is the recipe they want.

The data from kaggle website has Recipes, Interactions and User information. We are only considering Recipes and Interactions for our analysis. Interations being the reviews and ratings posted for each recipe.

Let's start with importing libraries

Importing necessary Libraries

Toggle code

Reading in the data

We will read the recipes data which is in the csv format directly into a dataframe and explore it a bit.

```
Index(['name', 'id', 'minutes', 'contributor_id', 'submitted', 'tags',
       'nutrition', 'n_steps', 'steps', 'description', 'ingredients',
       'n_ingredients'],
      dtype='object')
Number of columns:  12
```

	name	id	minutes	contributor_id	submitted	tags	nutrition	n_steps	steps	description	ingredients	n_ingredients
0	arriba baked winter squash mexican style	137739	55	47892	2005-09-16	['60-minutes-or-less', 'time-to-make', 'course...	[51.5, 0.0, 13.0, 0.0, 2.0, 0.0, 4.0]	11	['make a choice and proceed with recipe', 'dep...	autumn is my favorite time of year to cook! th...	['winter squash', 'mexican seasoning', 'mixed ...	7
1	a bit different breakfast pizza	31490	30	26278	2002-06-17	['30-minutes-or-less', 'time-to-make', 'course...	[173.4, 18.0, 0.0, 17.0, 22.0, 35.0, 1.0]	9	['preheat oven to 425 degrees f', 'press dough...	this recipe calls for the crust to be prebaked...	['prepared pizza crust', 'sausage patty', 'egg...	6
2	all in the kitchen chili	112140	130	196586	2005-02-25	['time-to-make', 'course', 'preparation', 'mai...	[269.8, 22.0, 32.0, 48.0, 39.0, 27.0, 5.0]	6	['brown ground beef in large pot', 'add choppe...	this modified version of 'mom's' chili was a h...	['ground beef', 'yellow onions', 'diced tomato...	13
3	alouette potatoes	59389	45	68585	2003-04-14	['60-minutes-or-less', 'time-to-make', 'course...	[368.1, 17.0, 10.0, 2.0, 14.0, 8.0, 20.0]	11	['place potatoes in a large pot of lightly sal...	this is a super easy, great tasting, make ahea...	['spreadable cheese with garlic and herbs', 'n...	11
4	amish tomato ketchup for canning	44061	190	41706	2002-10-25	['weeknight', 'time-to-make', 'course', 'main-...	[352.9, 1.0, 337.0, 23.0, 3.0, 0.0, 28.0]	5	['mix all ingredients& boil for 2 1 / 2 hours ...	my dh's amish mother raised him on this recipe...	['tomato juice', 'apple cider vinegar', 'sugar...	8

```
name              object
id                 int64
minutes            int64
contributor_id     int64
submitted         object
tags              object
nutrition         object
n_steps            int64
steps             object
description       object
ingredients       object
n_ingredients      int64
dtype: object
```

The attributes **id** and **contributor_id** are clearly identifiers, so let's convert them into string objects.

Also let's set the **recipe id** as the index for each row in our dataset.

```
Number of total recipes:  231637
```

```
Number of contributors:  27926
```

Let's describe the numerical fields in the data and look at their distributions.

	minutes	n_steps	n_ingredients
count	2.32e+05	231637.00	231637.00
mean	9.40e+03	9.77	9.05
std	4.46e+06	6.00	3.73
min	0.00e+00	0.00	1.00
25%	2.00e+01	6.00	6.00
50%	4.00e+01	9.00	9.00
75%	6.50e+01	12.00	11.00
max	2.15e+09	145.00	43.00

Interactions data

Reading in the data

```
Index(['user_id', 'recipe_id', 'date', 'rating', 'review'], dtype='object')
Number of columns:  5
```

	user_id	recipe_id	date	rating	review
0	38094	40893	2003-02-17	4	Great with a salad. Cooked on top of stove for...
1	1293707	40893	2011-12-21	5	So simple, so delicious! Great for chilly fall...
2	8937	44394	2002-12-01	4	This worked very well and is EASY. I used not...
3	126440	85009	2010-02-27	5	I made the Mexican topping and took it to bunk...
4	57222	85009	2011-10-01	5	Made the cheddar bacon topping, adding a sprin...

```
Total number of reviews:  1132367

Total number of contributors:  226570
```

Summarize the interactions data based on recipe_id, so that we might have the mean rating for each recipe and also the number of reviews posted for each recipe.

	mean_rating	review_count
recipe_id		
38	4.25	4
39	3.00	1
40	4.33	9
41	4.50	2
43	1.00	1

Joining Interations data with the original recipe data

```
Index(['name', 'minutes', 'contributor_id', 'submitted', 'tags', 'nutrition',
       'n_steps', 'steps', 'description', 'ingredients', 'n_ingredients',
       'id_copy', 'mean_rating', 'review_count'],
      dtype='object')
```

	name	minutes	contributor_id	submitted	tags	nutrition	n_steps	steps	description	ingredients	n_ingredients	id_copy
id												
137739	arriba baked winter squash mexican style	55	47892	2005-09-16	['60-minutes-or-less', 'time-to-make', 'course...	[51.5, 0.0, 13.0, 0.0, 2.0, 0.0, 4.0]	11	['make a choice and proceed with recipe', 'dep...	autumn is my favorite time of year to cook! th...	['winter squash', 'mexican seasoning', 'mixed ...	7	137739
31490	a bit different breakfast pizza	30	26278	2002-06-17	['30-minutes-or-less', 'time-to-make', 'course...	[173.4, 18.0, 0.0, 17.0, 22.0, 35.0, 1.0]	9	['preheat oven to 425 degrees f', 'press dough...	this recipe calls for the crust to be prebaked...	['prepared pizza crust', 'sausage patty', 'egg...	6	31490
112140	all in the kitchen chili	130	196586	2005-02-25	['time-to-make', 'course', 'preparation', 'mai...	[269.8, 22.0, 32.0, 48.0, 39.0, 27.0, 5.0]	6	['brown ground beef in large pot', 'add choppe...	this modified version of 'mom's' chili was a h...	['ground beef', 'yellow onions', 'diced tomato...	13	112140
59389	alouette potatoes	45	68585	2003-04-14	['60-minutes-or-less', 'time-to-make', 'course...	[368.1, 17.0, 10.0, 2.0, 14.0, 8.0, 20.0]	11	['place potatoes in a large pot of lightly sal...	this is a super easy, great tasting, make ahea...	['spreadable cheese with garlic and herbs', 'n...	11	59389
44061	amish tomato ketchup for canning	190	41706	2002-10-25	['weeknight', 'time-to-make', 'course', 'main-...	[352.9, 1.0, 337.0, 23.0, 3.0, 0.0, 28.0]	5	['mix all ingredients& boil for 2 1 / 2 hours ...	my dh's amish mother raised him on this recipe...	['tomato juice', 'apple cider vinegar', 'sugar...	8	44061

Pre-processing of the data

The data in it's original format has features like nutritional values, ingredients, steps as lists and because of reading in from the CSV format, the lists are read and understood as strings by pandas rather than a python list object!

Let's convert the necessary fields to a more usable formats.

Converting ingredients to usable strings

```
name              object
minutes            int64
contributor_id    object
submitted         object
tags              object
nutrition         object
n_steps            int64
steps             object
description       object
ingredients       object
n_ingredients      int64
id_copy           object
mean_rating      float64
review_count       int64
ingr_str          object
dtype: object
```

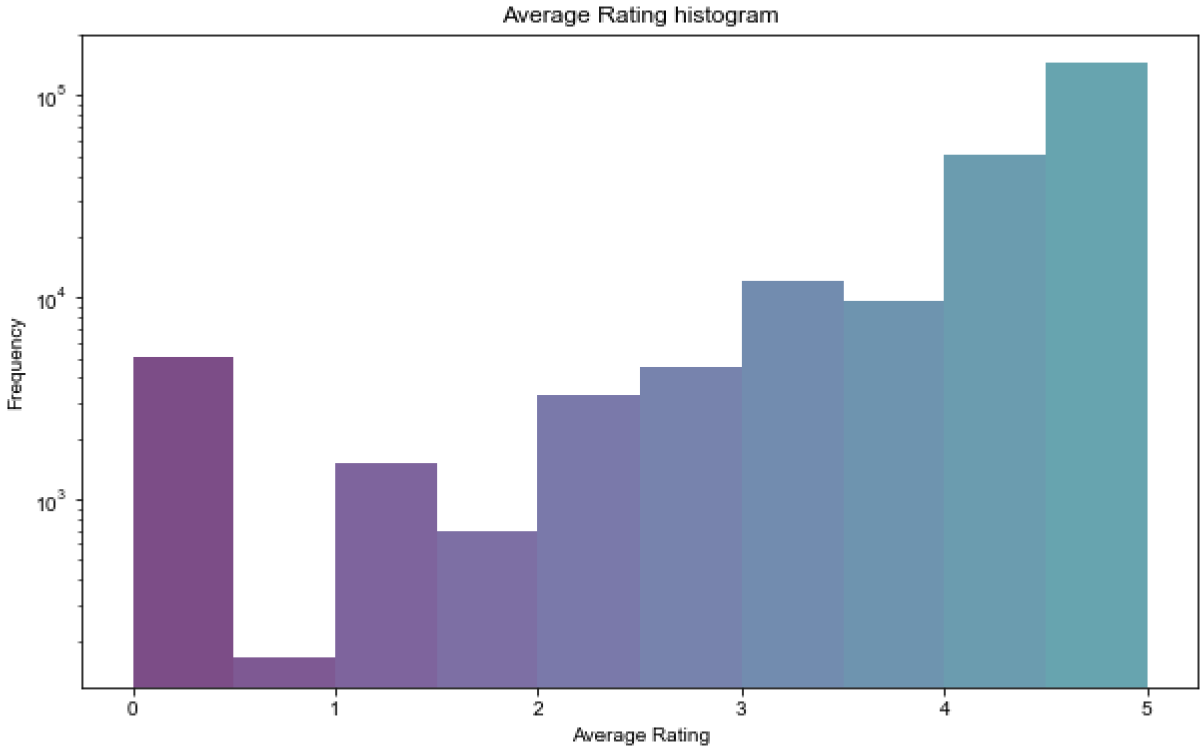
Flattening the nutritional values to columns

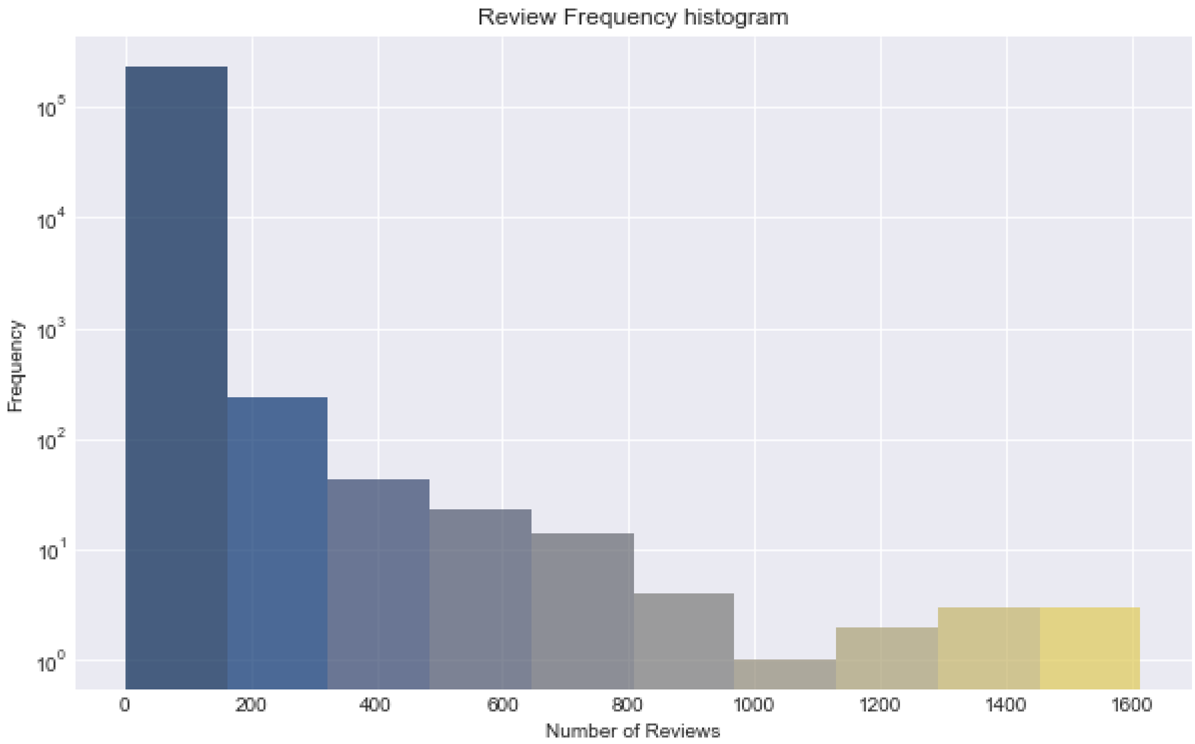
```
Index(['name', 'minutes', 'contributor_id', 'submitted', 'tags', 'nutrition',
       'n_steps', 'steps', 'description', 'ingredients', 'n_ingredients',
       'id_copy', 'mean_rating', 'review_count', 'ingr_str', 'cal', 'totalFat',
       'sugar', 'sodium', 'protein', 'satFat', 'carbs'],
      dtype='object')
```

	name	minutes	contributor_id	submitted	tags	nutrition	n_steps	steps	description	ingredients	...	mean_rating	revie
id													
137739	arriba baked winter squash mexican style	55	47892	2005-09-16	['60-minutes-or-less', 'time-to-make', 'course...	[51.5, 0.0, 13.0, 0.0, 2.0, 0.0, 4.0]	11	['make a choice and proceed with recipe', 'dep...	autumn is my favorite time of year to cook! th...	['winter squash', 'mexican seasoning', 'mixed ...	...	5.0	
31490	a bit different breakfast pizza	30	26278	2002-06-17	['30-minutes-or-less', 'time-to-make', 'course...	[173.4, 18.0, 0.0, 17.0, 22.0, 35.0, 1.0]	9	['preheat oven to 425 degrees f', 'press dough...	this recipe calls for the crust to be prebaked...	['prepared pizza crust', 'sausage patty', 'egg...	...	3.5	
112140	all in the kitchen chili	130	196586	2005-02-25	['time-to-make', 'course', 'preparation', 'mai...	[269.8, 22.0, 32.0, 48.0, 39.0, 27.0, 5.0]	6	['brown ground beef in large pot', 'add choppe...	this modified version of 'mom's' chili was a h...	['ground beef', 'yellow onions', 'diced tomato...	...	4.0	
59389	alouette potatoes	45	68585	2003-04-14	['60-minutes-or-less', 'time-to-make', 'course...	[368.1, 17.0, 10.0, 2.0, 14.0, 8.0, 20.0]	11	['place potatoes in a large pot of lightly sal...	this is a super easy, great tasting, make ahea...	['spreadable cheese with garlic and herbs', 'n...	...	4.5	
44061	amish tomato ketchup for canning	190	41706	2002-10-25	['weeknight', 'time-to-make', 'course', 'main-...	[352.9, 1.0, 337.0, 23.0, 3.0, 0.0, 28.0]	5	['mix all ingredients& boil for 2 1 / 2 hours ...	my dh's amish mother raised him on this recipe...	['tomato juice', 'apple cider vinegar', 'sugar...	...	5.0	

5 rows × 22 columns

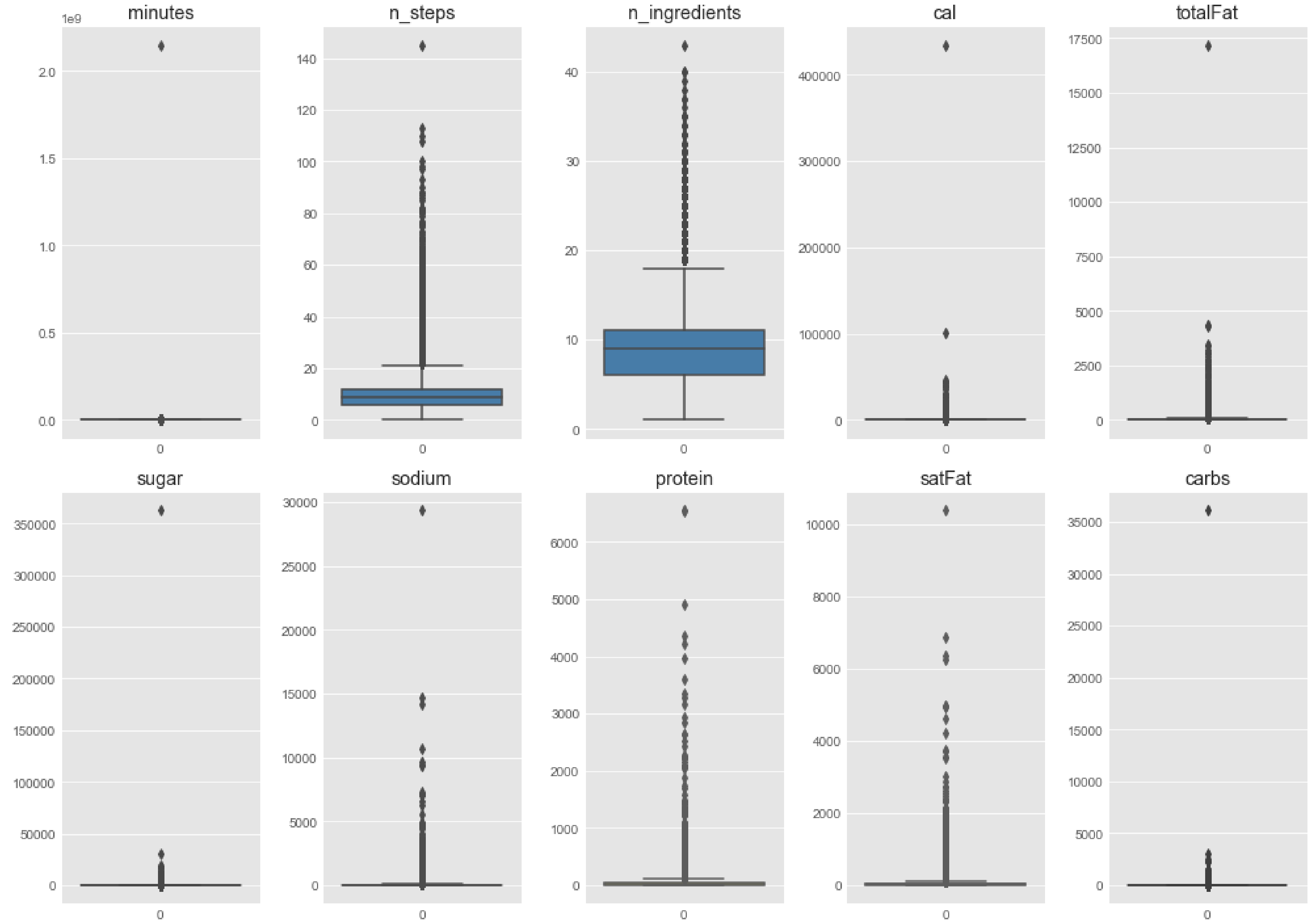
Exploring the data





	name	minutes	contributor_id	submitted	tags	nutrition	n_steps	steps	description	ingredients	...	mean_rating	review_count
id													
2886	best banana bread	65	1762	1999-09-26	['time-to-make', 'course', 'main-ingredient', ...	[272.8, 16.0, 97.0, 14.0, 7.0, 31.0, 14.0]	13	['remove odd pots and pans from oven', 'prehea...	you'll never need another banana bread recipe ...	['butter', 'granulated sugar', 'eggs', 'banana...	...	4.19	1613

1 rows × 22 columns



The above boxplots represent the distributions of the numeric features in our data. In all of the features there are few extreme values that are completely skewing the distributions. Such values can be called as outliers.

We will need to handle these outliers before moving forward with our analysis.

Performing clamping technique to remove outliers

We can see outliers in above box plot. But how is the boundary for the outlier is decided (the two horizontal lines which we see before the outliers)? So those values are decided by the the Inter Quartile Range (IQR) which is differenec of first and third quartile. So using that I can set my lower and upper bound as : lower bound = Q1 - 1.5 *IQR upper bound = Q3 + 1.5* IQR

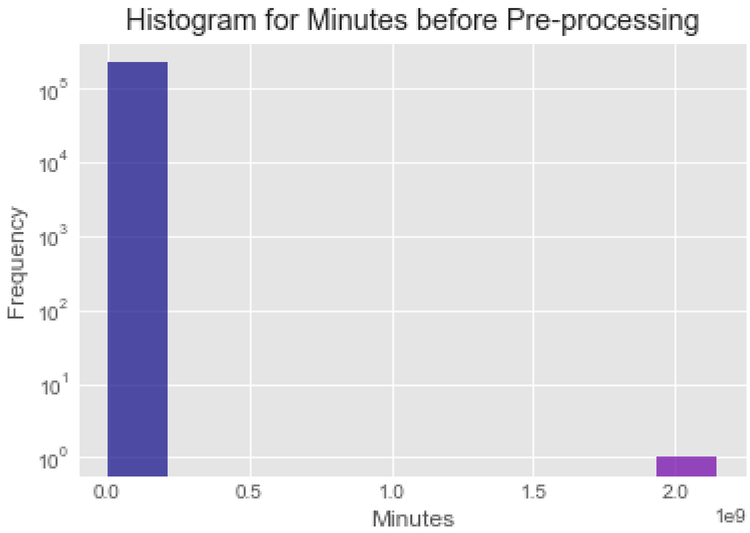
Tha values out of these range is considered as outliers and we can remove them. So let's do the same for our data.

	minutes	n_steps	n_ingredients	mean_rating	review_count	cal	totalFat	sugar	sodium	protein	satFat	carb
count	177201.00	177201.00	177201.00	177201.00	177201.00	177201.00	177201.00	177201.00	177201.00	177201.00	177201.00	177201.0
mean	40.92	9.17	8.88	4.37	4.86	298.81	22.17	35.91	18.01	26.42	27.30	8.9
std	30.37	4.83	3.52	0.96	16.90	190.97	19.46	39.98	18.25	27.35	27.49	7.2
min	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
25%	20.00	6.00	6.00	4.00	1.00	151.90	7.00	8.00	4.00	6.00	6.00	3.0
50%	35.00	8.00	9.00	4.71	2.00	264.30	17.00	20.00	12.00	15.00	18.00	7.0
75%	55.00	12.00	11.00	5.00	4.00	409.90	32.00	49.00	27.00	42.00	40.00	13.0
max	176.00	26.00	23.00	5.00	1613.00	1338.20	105.00	183.00	91.00	138.00	131.00	37.0

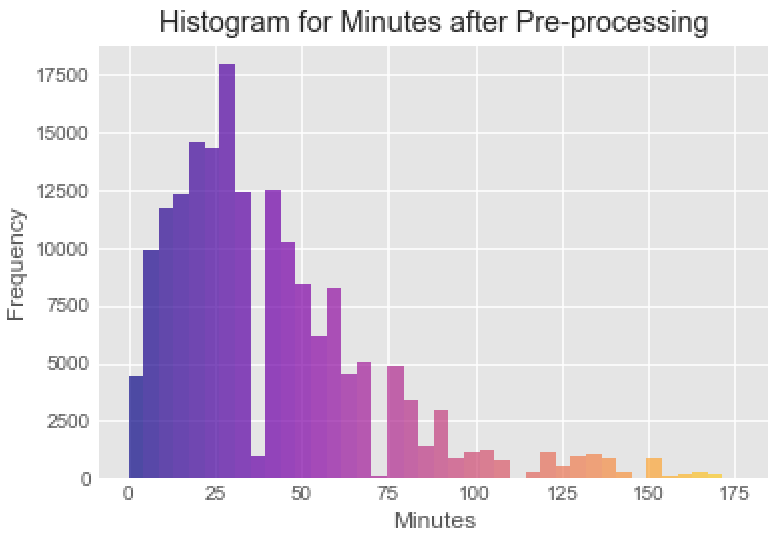
```
The number of recipes remaining after handling outliers: 177201
```

Let's look at distributions of some features before and after handling outliers through histograms.

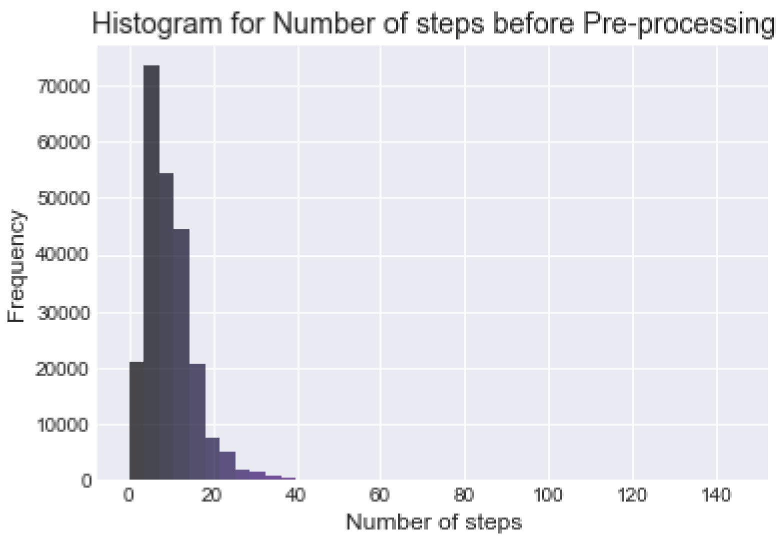
Minutes feature before



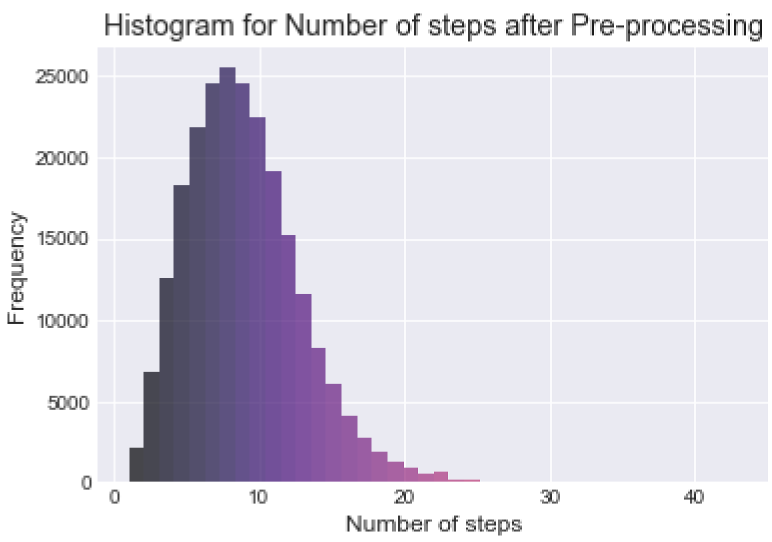
Minutes feature after



n_steps feature before



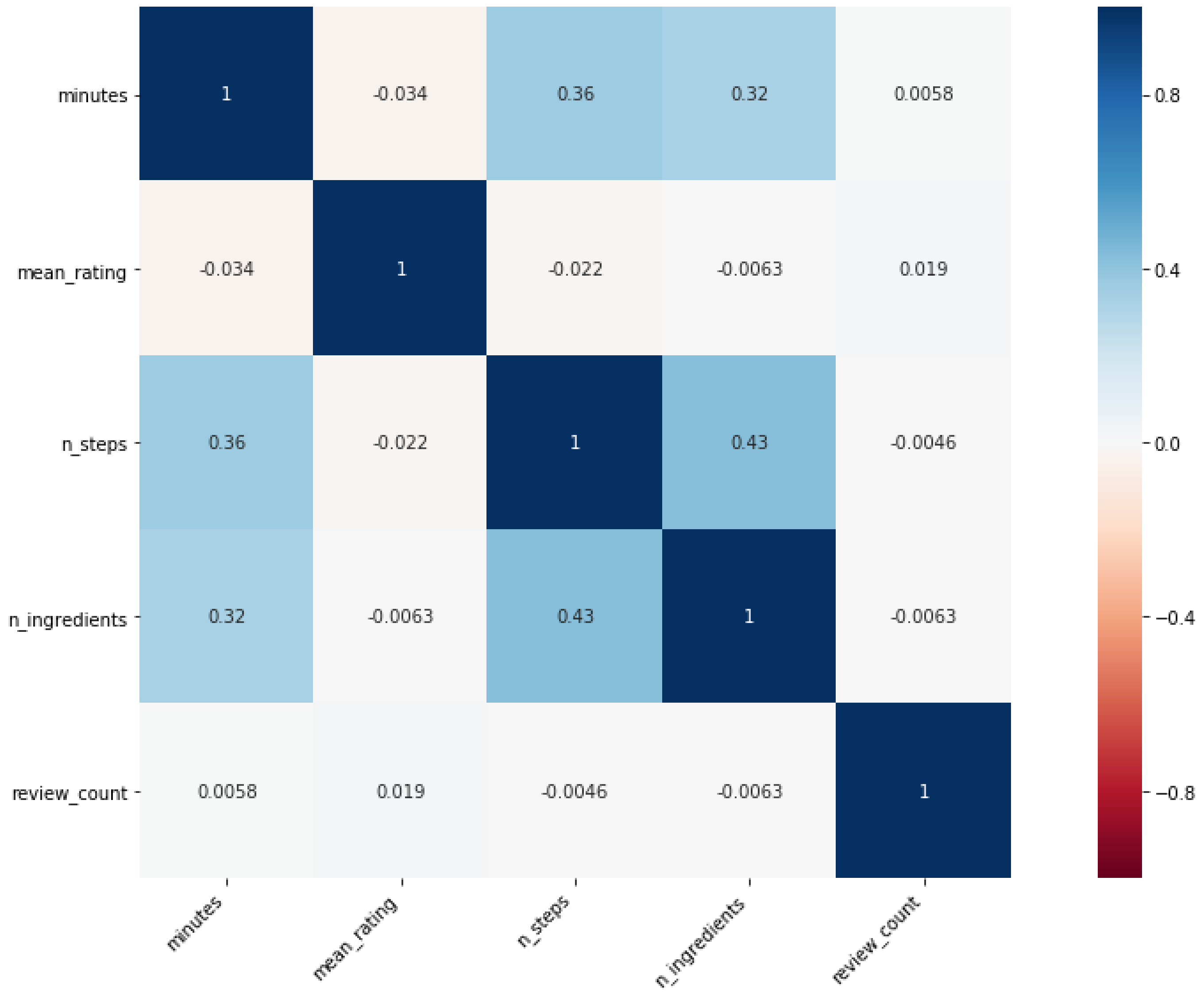
n_steps feature after



Checking to see if there are any Null values that we need to handle.

```
Checking if Null values exist:
name              False
minutes           False
contributor_id    False
submitted         False
tags              False
nutrition         False
n_steps           False
steps             False
description        True
ingredients       False
n_ingredients     False
id_copy           False
mean_rating       False
review_count      False
ingr_str          False
cal               False
totalFat          False
sugar             False
sodium            False
protein           False
satFat            False
carbs             False
dtype: bool
```

Let's look at the correlations between all the numerical fields in original data**



There are no significant correlations between any of these fields, making them very independent of each other. This raises two situations:

1. Since there are no correlations, predictive models are more relaible.
2. Since there are no correlations, it will be hard to extract insights through relationships between various fields.

Feature Engineering Cuisine

It was to our suprise to learn that Food.com doesn't contain the information about a recipe's cuisine.

We can try to introduce the recipe information using the basic instincts of **Data Engineering** and the concepts of **Data Mining**.

Why ?

We are using the dataset from one of the famous website in its domain, food.com, It provides recipies for thousands of dishes (to be precise: 231637). So basically this website have recipies for every event you can think of such as pool parties, christmas holidays and so on.

But it was to our suprise that they dont have any filter for cuisines. Even in their dataset they dont have any field which they can leverage to have this extended feature on their website.

Hence we move ahead to fix this problem using **data engineering** basic instincts and the skills we have learnt in the **Data Minning**.

What Is Data Engineering ?

Data engineering is the aspect of data science that focuses on **practical applications of data collection and analysis**. For all the work that data scientists do to answer questions using large sets of information, there have to be mechanisms for **collecting and validating that information**.

Ian Buss, principal solutions architect at Cloudera, notes that data scientists focus on finding new insights from a data set, while data engineers are concerned with the production readiness of that data and all that comes with it: **formats, scaling, resilience and security**

So lets start..

First we have analysed the data set and we found that ingridents would be the best field in the exsisting dataset to use and leverage and predict cuisine for every recipie.

Then using one similar dataset where we had ingridients and cuisines we trained our model upto the accuracy of ~75%

Major steps and strategy

1. We have 3 files in total which are as follows :
   - Train.json : this is with ingridients and cuisines
   - Test.json : This is with ingridients only
   - RAW_recipes.csv : This is the food.com data set in which we intend to add cuisine for each recipie.
2. So using Train.csv we split this dataset into test and train
3. We apply multiple model and check and get maximum accuracy.(in our case random forest classifier performs best).
4. Having done that we can now proceed on the dummy data set Test.csv this is just an extra step that where we are predicting cuisines from the ingridients and checking manually that every thing is working good before we scale our solution to an entire dataset.
5. After we have predicted cuisine now its time to predict the cuisines of entire data set. so we run the predict function giving tf-idf matrix for the ingridients.
6. Once we have the predictions we can add this column to the main dataframe.

Reading CSVs

Reading train.json which has all the data bot ingridients and cuisines

```
cuisine        39774
id             39774
ingredients    39774
dtype: int64
```

Reading test.json which has only ingridients. this is our dummy test file to see our model works correctly.

```
id             9944
ingredients    9944
dtype: int64
```

```
name              231636
minutes           231637
contributor_id    231637
submitted         231637
tags              231637
nutrition         231637
n_steps           231637
steps             231637
description       226658
ingredients       231637
n_ingredients     231637
id_copy           231637
mean_rating       231637
review_count      231637
ingr_str          231637
dtype: int64
```

Here comes the important part and we must take care since we are dealing with categorical data we need to vectorize our data. For that wew are using TF-IDF.

What is TF-IDF ?

Tf-idf is a very common technique for determining roughly what each document in a set of documents is “about”. It cleverly accomplishes this by looking at two simple metrics: tf (term frequency) and idf (inverse document frequency).

Term frequency : It is the proportion of occurrences of a specific term to total number of terms in a document.

Inverse document frequency : It is the inverse of the proportion of documents that contain that word/phrase.

```
TF-IDF Matrix looks like below :
 [[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]

Cuisine looks like below :
 0          greek
1    southern_us
2       filipino
3         indian
4         indian
Name: cuisine, dtype: object
```

Split and Train

Now that we have data ready which can be further used to train our model we will move ahead straight to train our model. The only thing is since we are using Random Forest Classifier we can pass mulitple parameters with different configuration. So in order to get the best suitable model we are using **GRID SEARCH**

What is Grid Search?

Grid search is the process of performing hyper parameter tuning in order to determine the optimal values for a given model. This is significant as the performance of the entire model is based on the hyper parameter values specified.

You can change these values and experiment more to see which value ranges give better performance. A cross validation process is performed in order to determine the hyper parameter value set which provides the best accuracy levels.

```
GridSearchCV(cv=5, error_score='raise-deprecating',
             estimator=RandomForestClassifier(bootstrap=True, class_weight=None,
                                              criterion='gini', max_depth=None,
                                              max_features='auto',
                                              max_leaf_nodes=None,
                                              min_impurity_decrease=0.0,
                                              min_impurity_split=None,
                                              min_samples_leaf=1,
                                              min_samples_split=2,
                                              min_weight_fraction_leaf=0.0,
                                              n_estimators='warn', n_jobs=None,
                                              oob_score=False,
                                              random_state=None, verbose=0,
                                              warm_start=False),
             iid='warn', n_jobs=None, param_grid={'n_estimators': [100]},
             pre_dispatch='2*n_jobs', refit=True, return_train_score=False,
             scoring=None, verbose=0)
```

We are here checking the model score and the best parameters to use.

```
best param {'n_estimators': 100}
best score 0.7395581256481977
best estimator RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=None, max_features='auto', max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)
```

How accurate is the model ?

To answer the above question we are evaluation our model on 3 basic parameters whcih are :

- The Score of the model (grid.score)
- The Accuracy of the model (accuracy.score)
- Classification rate (Using classification report)

```
model score :  0.7458202388434947
model accuracy :  0.7458202388434947
              precision    recall  f1-score   support

     italian       0.84      0.41      0.55        90
     mexican       0.70      0.21      0.32       170
 southern_us       0.80      0.71      0.75       293
      indian       0.69      0.88      0.78       551
     chinese       0.80      0.50      0.61       134
      french       0.61      0.50      0.55       537
cajun_creole       0.84      0.54      0.66       237
        thai       0.84      0.89      0.86       608
    japanese       0.85      0.36      0.51       155
       greek       0.71      0.93      0.81      1556
     spanish       0.92      0.53      0.67       102
      korean       0.86      0.61      0.72       270
  vietnamese       0.92      0.59      0.72       171
    moroccan       0.83      0.93      0.88      1300
     british       0.82      0.62      0.71       154
    filipino       0.78      0.25      0.38        85
       irish       0.63      0.79      0.70       845
    jamaican       0.79      0.26      0.39       189
     russian       0.74      0.74      0.74       318
   brazilian       0.87      0.38      0.53       190

    accuracy                           0.75      7955
   macro avg       0.79      0.58      0.64      7955
weighted avg       0.76      0.75      0.73      7955
```

Now as we have disscussed multiple times earlier our model is ready to be deployed and we can start predicting the cuisine given the ingridents. We just have to make sure that since we trained our model with the TF_IDF vectorizer we must use the same for predictions.

Using our dummy test dataset we first convert the ingredients to the vector and then pass it to grid.predict() this will give us the cuisine.

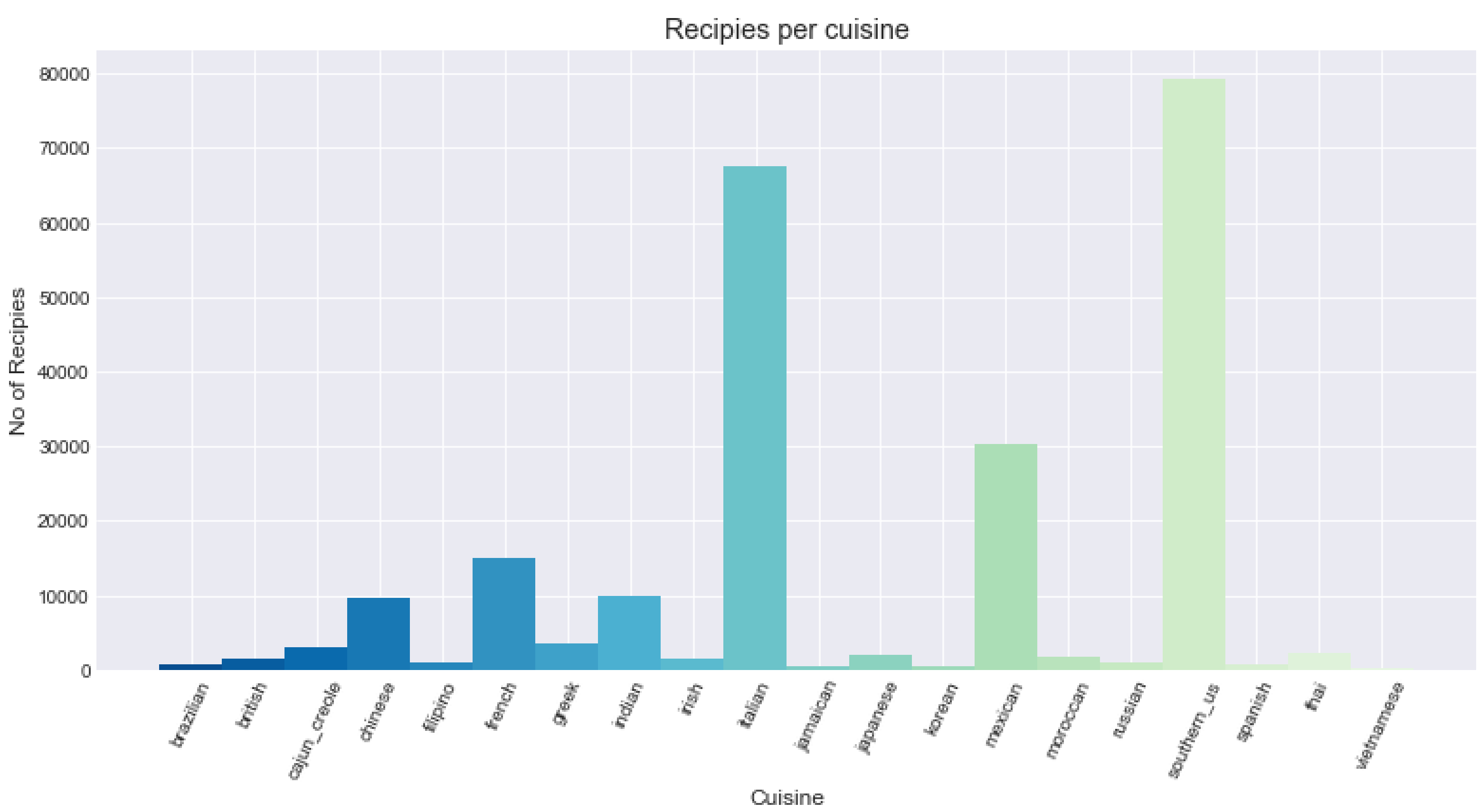
	id	ingredients	ingredient_list	cuisine
7	41217	[ground ginger, white pepper, green onions, or...	ground ginger,white pepper,green onions,orange...	chinese

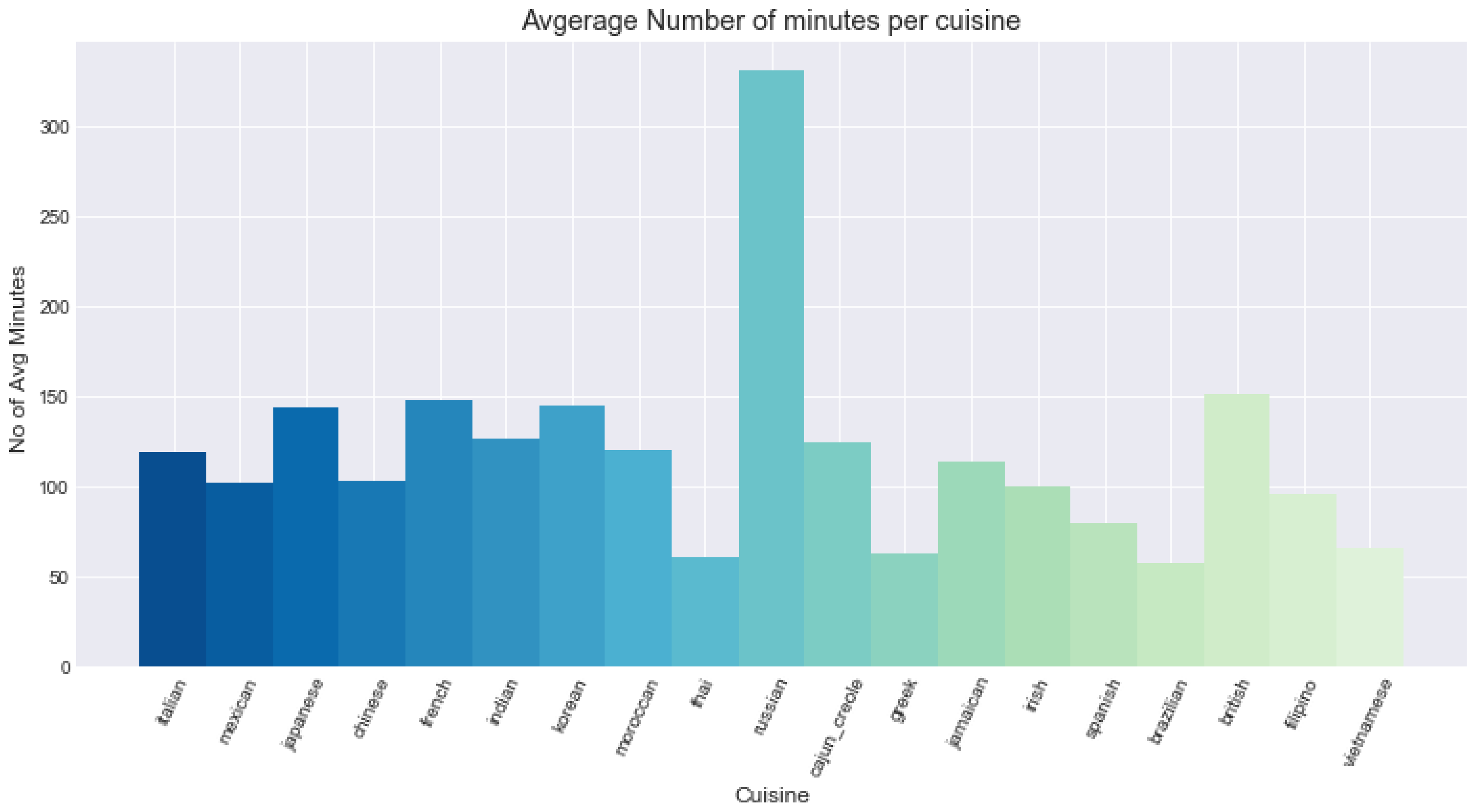
Now we can see from the above modified Dataframe that our model is predicting things quite nicely. So we will move on to applying the same model on the entire data set.

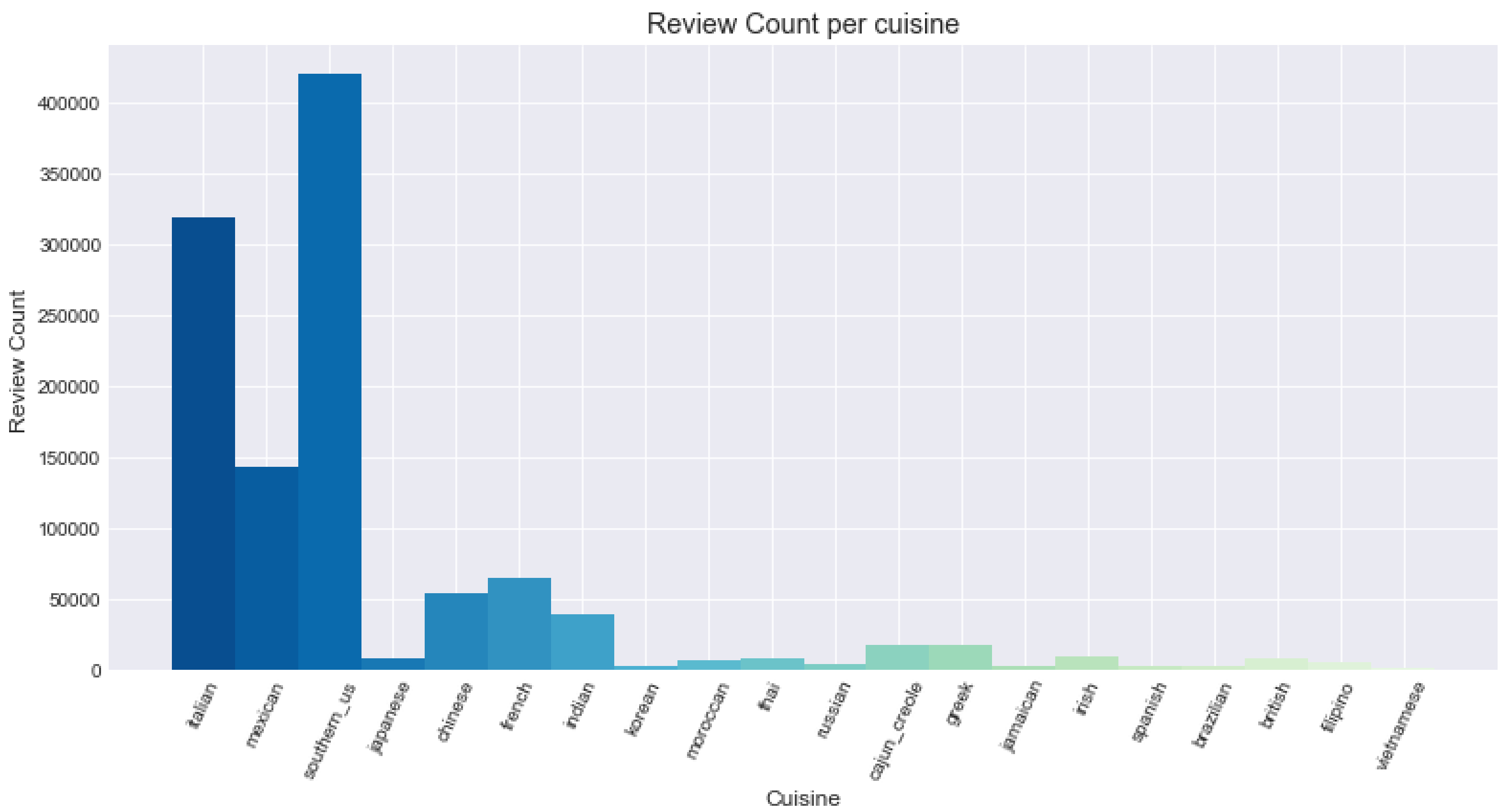
We just need to keep in mind the same thing that since we trained our model with the TF_IDF vectorizer we must use the same for predictions.

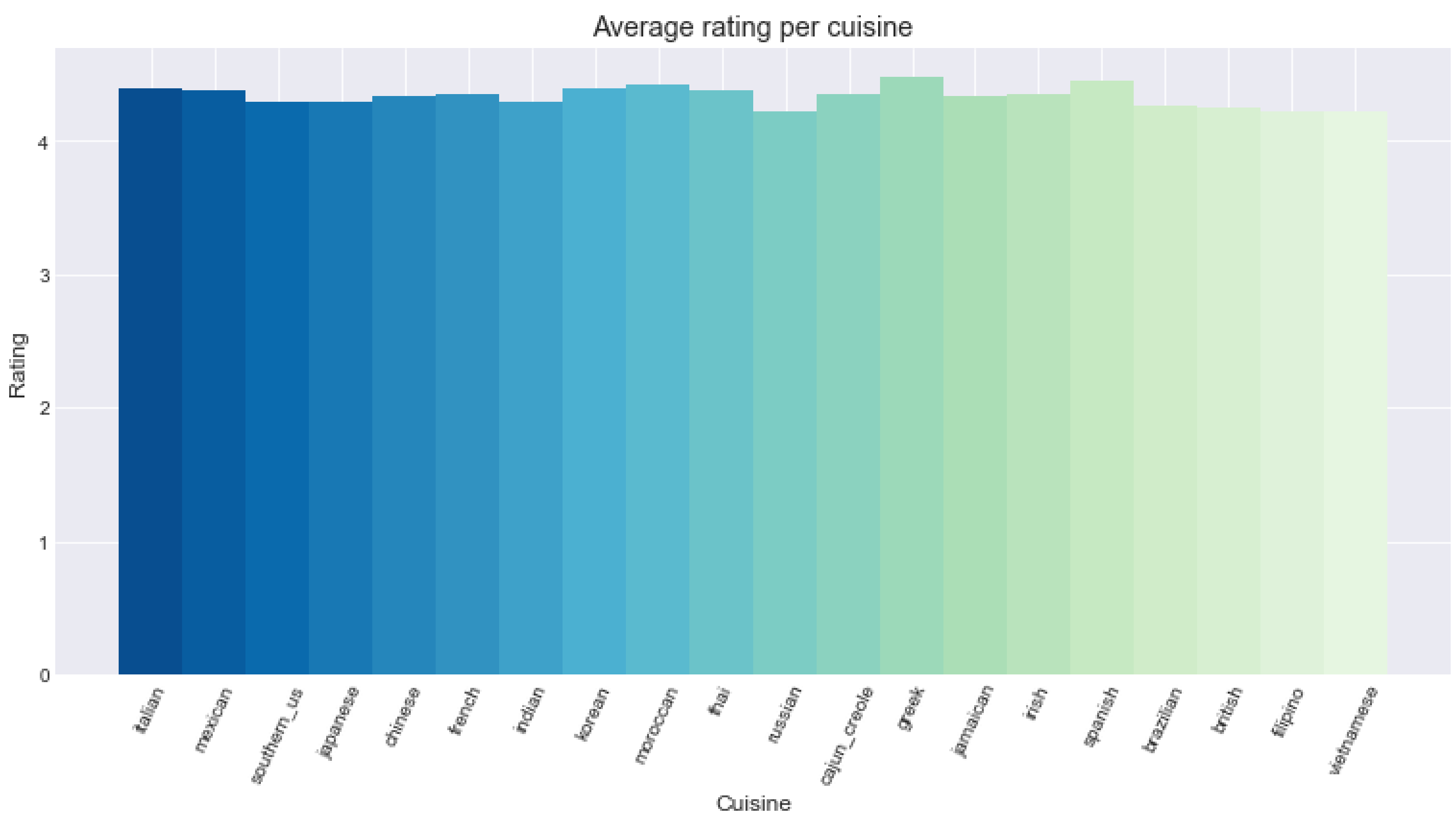
id	name	minutes	contributor_id	submitted	tags	nutrition	n_steps	steps	description	ingredients	n_ingredients	id_copy
137739	arriba baked winter squash mexican style	55	47892	2005-09-16	['60-minutes-or-less', 'time-to-make', 'course...	[51.5, 0.0, 13.0, 0.0, 2.0, 0.0, 4.0]	11	['make a choice and proceed with recipe', 'dep...	autumn is my favorite time of year to cook! th...	['winter squash', 'mexican seasoning', 'mixed ...	7	137739
31490	a bit different breakfast pizza	30	26278	2002-06-17	['30-minutes-or-less', 'time-to-make', 'course...	[173.4, 18.0, 0.0, 17.0, 22.0, 35.0, 1.0]	9	['preheat oven to 425 degrees f', 'press dough...	this recipe calls for the crust to be prebaked...	['prepared pizza crust', 'sausage patty', 'egg...	6	31490
112140	all in the kitchen chili	130	196586	2005-02-25	['time-to-make', 'course', 'preparation', 'mai...	[269.8, 22.0, 32.0, 48.0, 39.0, 27.0, 5.0]	6	['brown ground beef in large pot', 'add choppe...	this modified version of 'mom's' chili was a h...	['ground beef', 'yellow onions', 'diced tomato...	13	112140
59389	alouette potatoes	45	68585	2003-04-14	['60-minutes-or-less', 'time-to-make', 'course...	[368.1, 17.0, 10.0, 2.0, 14.0, 8.0, 20.0]	11	['place potatoes in a large pot of lightly sal...	this is a super easy, great tasting, make ahea...	['spreadable cheese with garlic and herbs', 'n...	11	59389
44061	amish tomato ketchup for canning	190	41706	2002-10-25	['weeknight', 'time-to-make', 'course', 'main-...	[352.9, 1.0, 337.0, 23.0, 3.0, 0.0, 28.0]	5	['mix all ingredients& boil for 2 1 / 2 hours ...	my dh's amish mother raised him on this recipe...	['tomato juice', 'apple cider vinegar', 'sugar...	8	44061

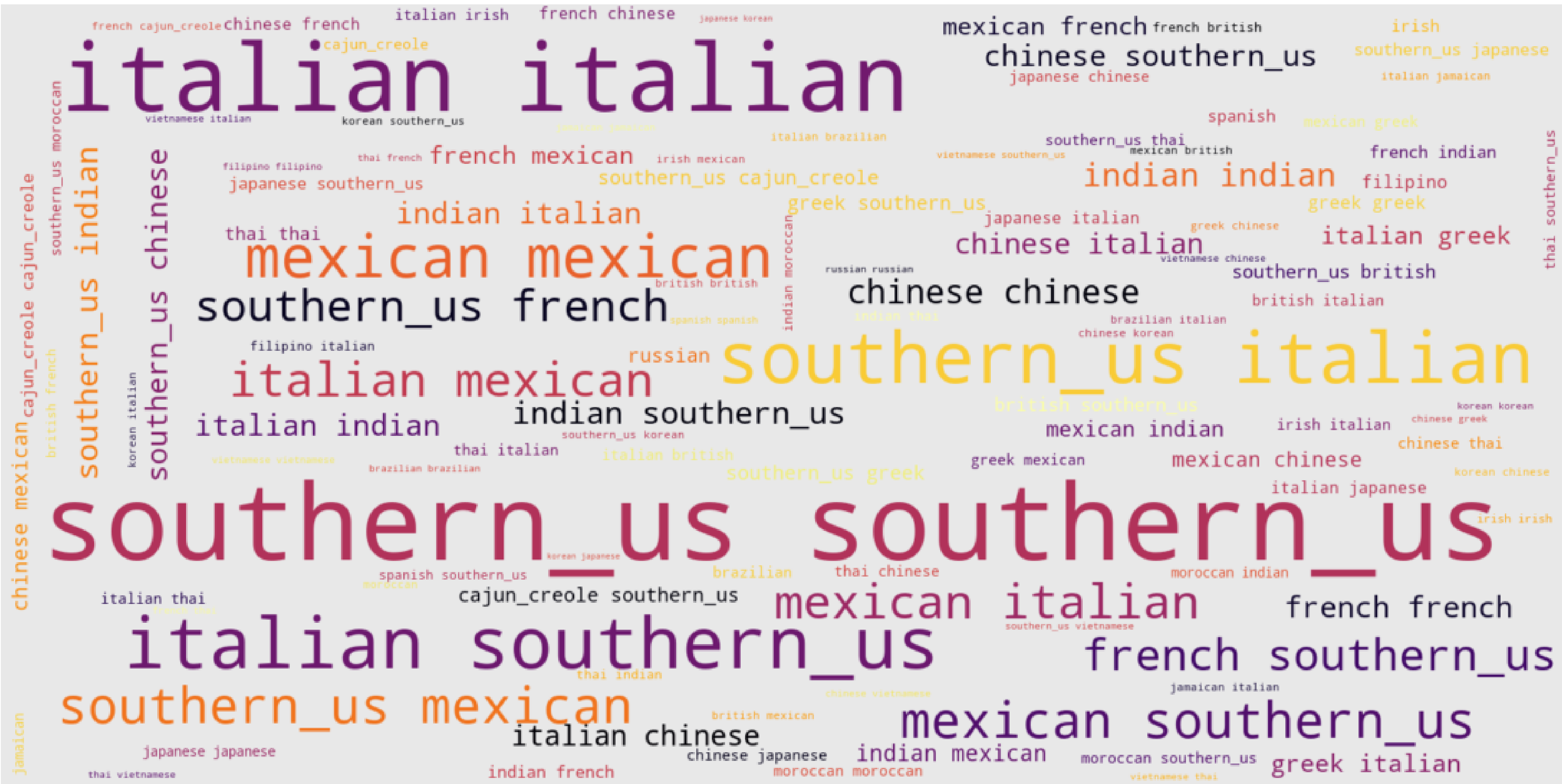
Graph to classify recipe on the basis of the cuisines.




Avgerage Number of minutes per cuisine
No of Avg Minutes
0
50
100
150
200
250
300
italian
mexican
japanese
chinese
french
indian
korean
moroccan
thai
russian
cajun_creole
greek
jamaican
irish
spanish
brazilian
british
filipino
vietnamese
Cuisine


Review Count per cuisine
Review Count
0
50000
100000
150000
200000
250000
300000
350000
400000
italian
mexican
southern_us
japanese
chinese
french
indian
korean
moroccan
thai
russian
cajun_creole
greek
jamaican
irish
spanish
brazilian
british
filipino
vietnamese
Cuisine


Average rating per cuisine
Rating
0
1
2
3
4
italian
mexican
southern_us
japanese
chinese
french
indian
korean
moroccan
thai
russian
cajun_creole
greek
jamaican
irish
spanish
brazilian
british
filipino
vietnamese
Cuisine



- After successfully integrating cuisine into the data, we can see that the Average rating for each cuisine is more or less the same. There's not much of a trend.
- The number of ratings per cuisine doesn't specifically give us any information given the high volumes of recipes for specific cuisines.
- But what if we want to understand what kind of recipes become more popular based on some other values available in the data.

> In such senarios, ***Clustering*** comes very handy

Clustering:

- Clustering is the process of grouping similar data points based on similarity/distance metrics.
- It can uncover patterns which are previously undiscovered.
- Each group can be assumed as a different class of the data.
- While the data points in a cluster are selected to maximize similarity between them, the clusters should be very dissimilar.
- Most clustering algorithms deal with noise in the data.
- For example, in social networks, users can be clustered based on their likes and dislikes.

For this analysis, we will use K-Means clustering.

K-Means clustering algorithm:

- K-means algorithm is a partition based clustering algorithm.
- **K** here is the number of clusters to be aggregated from the data.
- When we define the **K** value, the algorithm finds the very number of means from the data.
- It will find distance of each of the data points from the means cluusters them to the nearest mean.
- These means are called cluster centers.

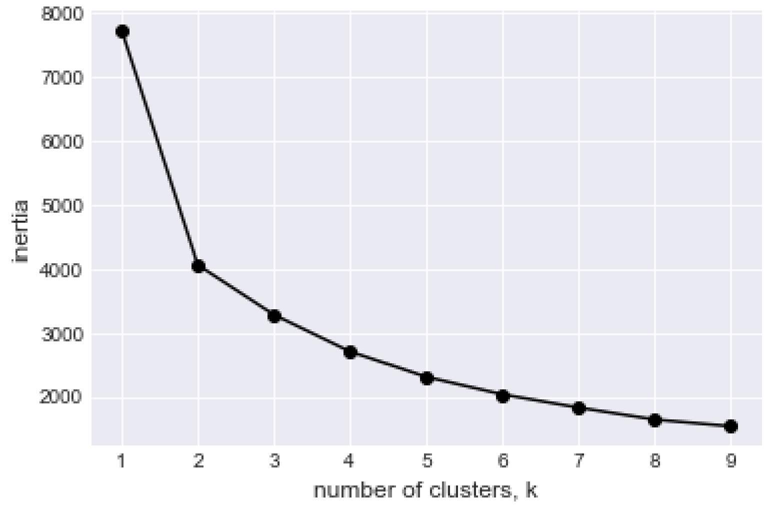
First let's try to cluster the data by nutrition values

Clustering by nutritional values

Since our nutrition values have differeng ranges and scales of values, it is important we normalize them. We will use the **Normalizer** function from sklearn for this.

Now we apply K-means clustering algorithm on the normalized data.

But to find the optimal value for **K**, we will use the elbow plot. Elbow plot shows us the inertia score for each K value. In this plot, the point where the line bends like an elbow of a hand is considered the optimal value for K.



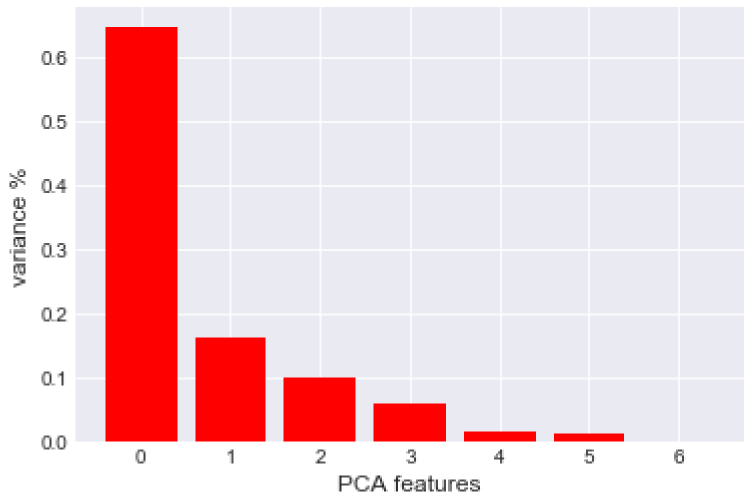
From the above graph we can see that there is slight bend near the value 4, thus 4 can be considered as the value of K.

Now to visualize the 4 clusters on a 2D graph, we will use PCA for dimensionality reduction.

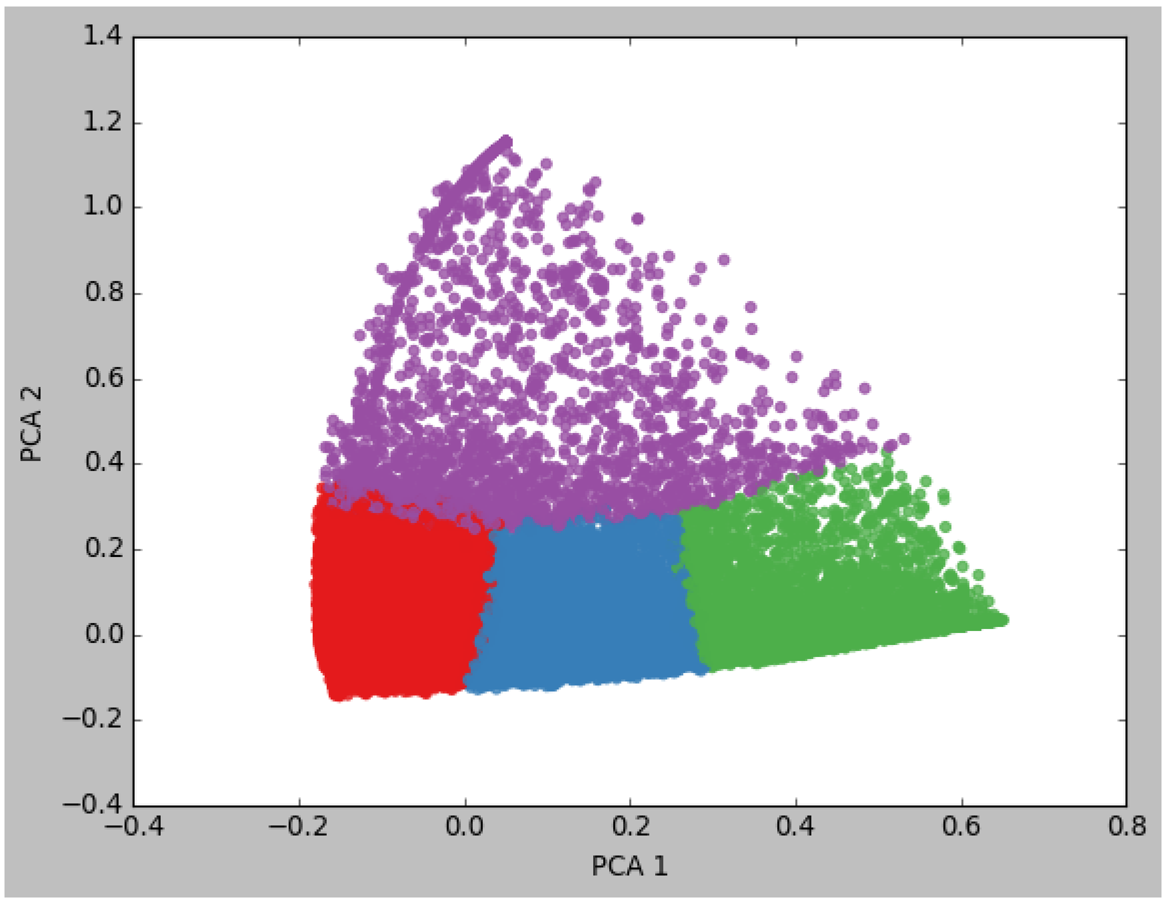
Principal Component Analysis is a statistical procedure to perform feature extraction, when we have too many features to work with.

- The algorithm is mainly used for reducing features to either limit over fitting the data or to visualize in a 2 dimensional or 3 dimensional plot.
- It mainly works on the variation of the features available for a data point and brings out strong underlying structures.
- These help us to understand and visualize the data more easily.
- PCA itself doesn't remove any features, but computes new features as a function of one or more existing features.

```
The explained variance ratios are : [0.64729439 0.16230638 0.0998635  0.06000448 0.01652446 0.01197393
 0.00203287]
```



The above graph shows the explained variance ratios of each Principal component evaluated. Explained variance is the percentage of data explained by a principal component. As we can see, between Principal component 0 and 1, more than 75% of our data is being explained. So we can use the first 2 principal components to visualize the data.



As we can see clusters are very clearly separated in the data. Let's see some results from our clusters.

The number of recipes in each cluster

```
0    117516
1     44572
2     13092
3      2021
Name: nutr_cluster, dtype: int64
```

Let's summaize the data by the cluster and look at some properties

nutr_cluster	minutes	n_steps	n_ingredients	mean_rating	review_count	cal	totalFat	sugar	sodium	protein	satFat	carbs
3	31.60	6.66	7.81	4.28	5.30	61.59	2.63	15.36	39.02	5.72	1.94	2.41
1	41.44	9.42	8.75	4.34	5.09	247.92	16.85	72.16	9.40	10.59	22.80	10.13
0	42.10	9.38	9.20	4.37	4.83	340.36	26.63	17.01	22.45	35.29	31.95	8.65
2	30.08	6.86	6.57	4.39	4.24	135.67	3.24	85.32	4.27	3.91	4.85	8.71

As we can see there are no properties that define these clusters appropriately. Thus, clustering through Nutritional values didn't give us any good insights.

There is one more important field that defines a recipe, ingredients. Ingredients used in a recipe define both the nutritional values and cuisine of the item, thus playing an important role. We will now attempt to cluster based on ingredients to exploratorily search for insights.

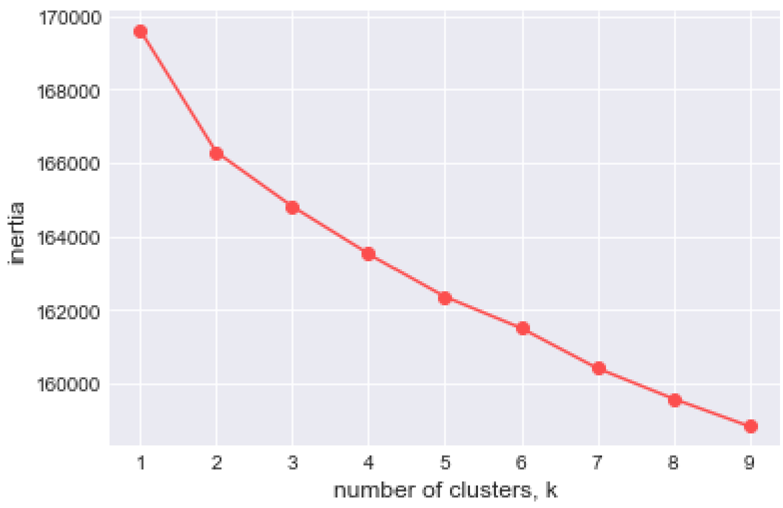
Clustering by ingredients

The pre-processing through TF-IDF vectorizer has already been explained in the previous section and we will be using the same preprocessing even for this process.

id	active_dry_yeast	allspice	almond_extract	almonds	american_cheese	apple	apple_cider	apple_cider_vinegar	apple_juice	apples	...	wc
137739	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	...	
31490	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	...	
112140	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	...	
59389	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	...	
5289	0.0	0.0	0.0	0.0	0.0	0.63	0.0	0.0	0.0	0.0	...	

5 rows × 500 columns

Plotting the elbow plot to find the optimal **K**.



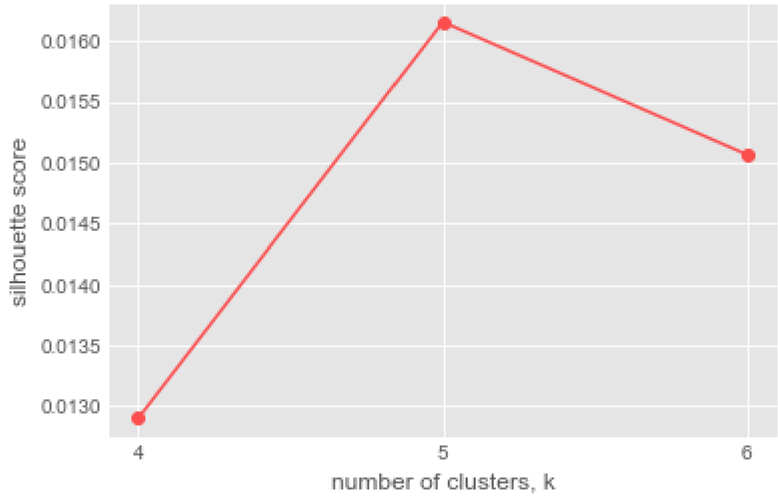

```
Process completed - 699.0452790260315 seconds elapsed.
```

From the above plot, we cannot discern a **K** value easily. Even though the graph is not straightforward, we have reasonable doubt between values 4,5 & 6.

In this case, we will need another method to find the optimal **K** from the values 4,5 & 6.

We will use the silhouette score.

- Silhouette method measures how similar a point is to it's own cluster compared to others.
- It is more likely a validation rather than a decision maker. Which is exactly what we want in this scenario.
- By using Euclidean distance as the metric, we will plot the graph for silhouette scores for the three values of K.




```
Process completed - 1973.9899640083313 seconds elapsed.
```

From the above graph, we can confidently say that the **5** is the most optimal value for **K**.

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
       n_clusters=5, n_init=10, n_jobs=None, precompute_distances='auto',
       random_state=0, tol=0.0001, verbose=0)
```

Let's look at the number of recipes in each cluster

```
0    80109
3    28961
4    25476
2    23507
1    19148
Name: ingr_cluster, dtype: int64
```

Now the top-ingredients in each of our clusters.

```
Crucial ingredients for each clusters:

Cluster 0:
salt onion mayonnaise garlic_cloves pepper sugar extra_virgin_olive_oil vegetable_oil garlic tomatoes lemon_
juice salt_and_pepper parmesan_cheese sour_cream black_pepper

Cluster 1:
water salt onion sugar butter pepper vegetable_oil oil cornstarch eggs flour garlic_cloves garlic soy_sauce
lemon_juice

Cluster 2:
sugar baking_powder eggs baking_soda flour salt vanilla butter egg milk cinnamon vanilla_extract brown_sugar
granulated_sugar unsalted_butter

Cluster 3:
butter milk salt eggs pepper onion flour parmesan_cheese cheddar_cheese salt_and_pepper egg sugar sour_cream
brown_sugar potatoes

Cluster 4:
olive_oil garlic_cloves salt onion garlic salt_and_pepper parmesan_cheese pepper tomatoes garlic_clove black
_pepper lemon_juice fresh_parsley balsamic_vinegar fresh_ground_black_pepper
```

Summarizing the data on Cluster number to look at some properties

	minutes	n_steps	n_ingredients	mean_rating	review_count	cal	totalFat	sugar	sodium	protein	satFat	carbs	nutr_cluster
ingr_cluster													
2	46.13	10.82	9.87	4.22	5.95	252.26	17.05	67.15	9.57	9.07	24.30	10.79	0.86
1	48.56	9.86	9.52	4.27	5.12	293.16	19.39	37.32	20.33	27.42	22.60	9.76	0.49
3	43.41	9.69	8.31	4.38	5.17	338.14	27.95	31.41	19.62	27.99	45.23	9.23	0.29
0	36.01	8.07	8.16	4.40	4.44	284.89	20.77	32.69	18.77	28.03	23.74	8.04	0.44
4	43.01	10.04	10.38	4.43	4.61	345.05	26.78	21.26	19.87	34.83	24.40	9.22	0.13

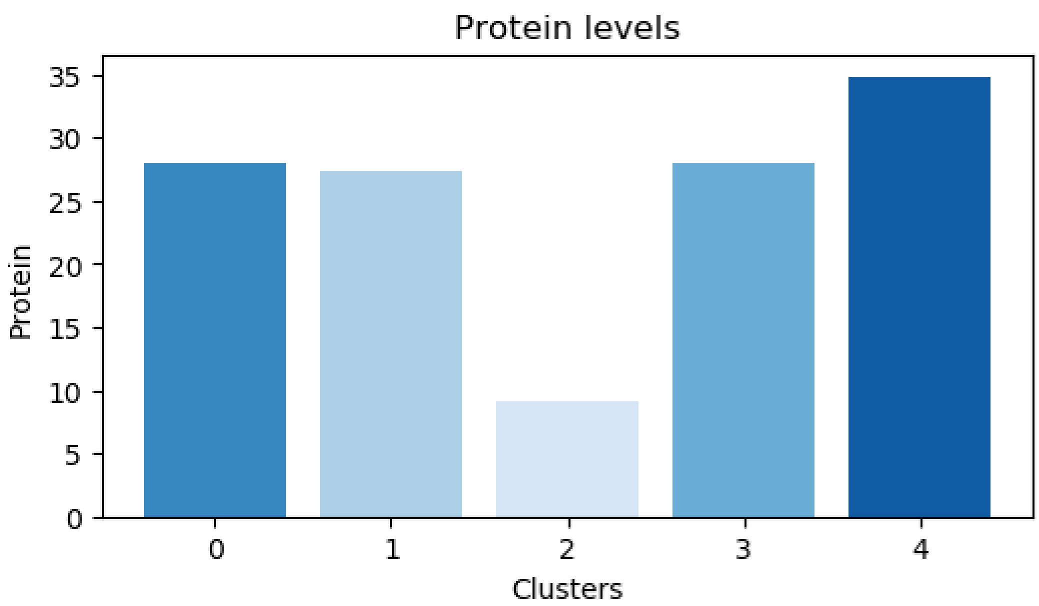
There are 2 features that define our clusters very well in terms of popularity.

- As we can see the cluster that has the highes average rating has the lowest sugar values and the highest protein values and
- the cluster with the least average rating has the highest sugar values and the least protein values.

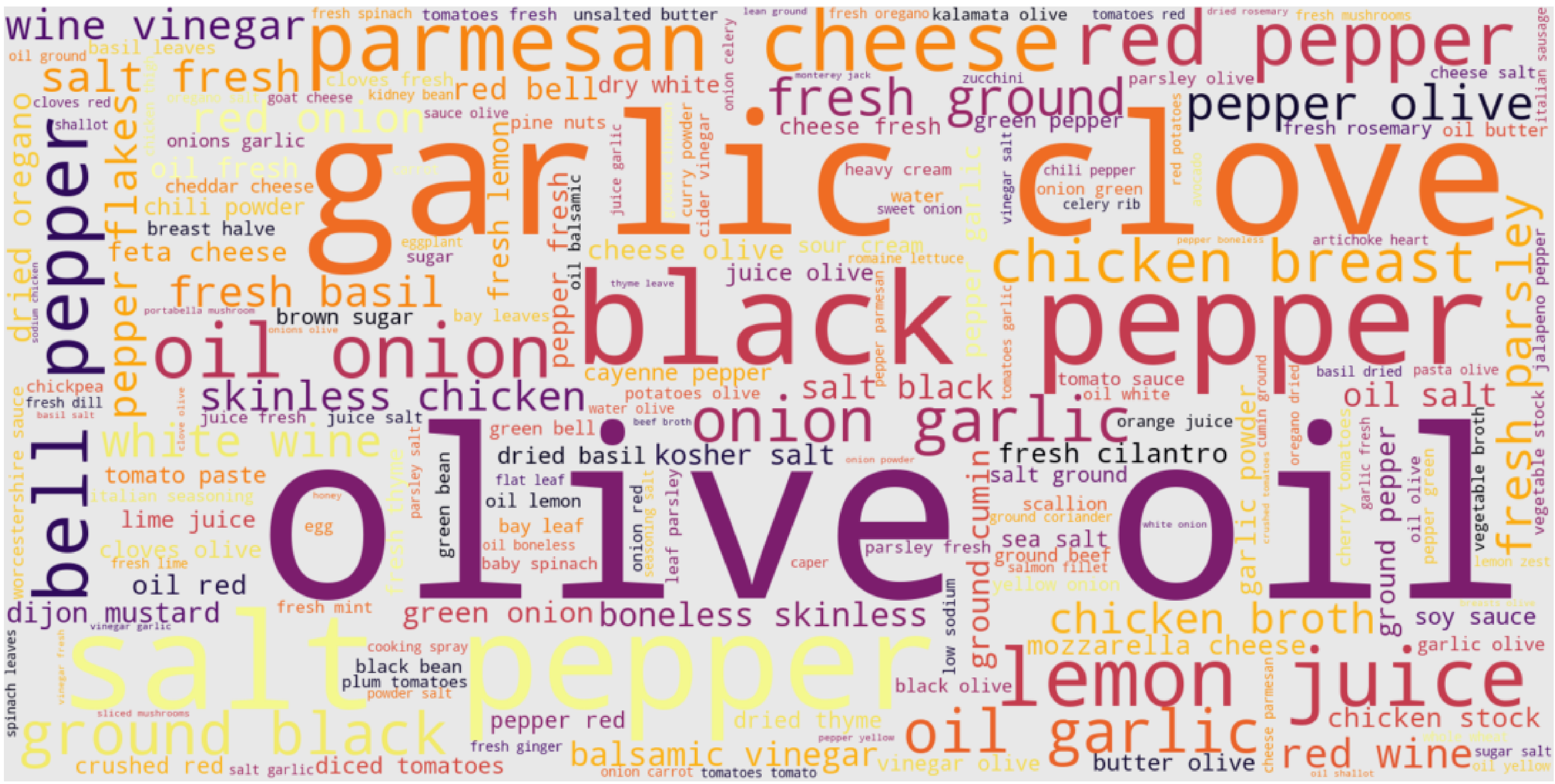
> Thus it shows us that recipes that contain less sugars and more proteins are more prone to beome popular than the sugary sweets and cakes with less protein value.

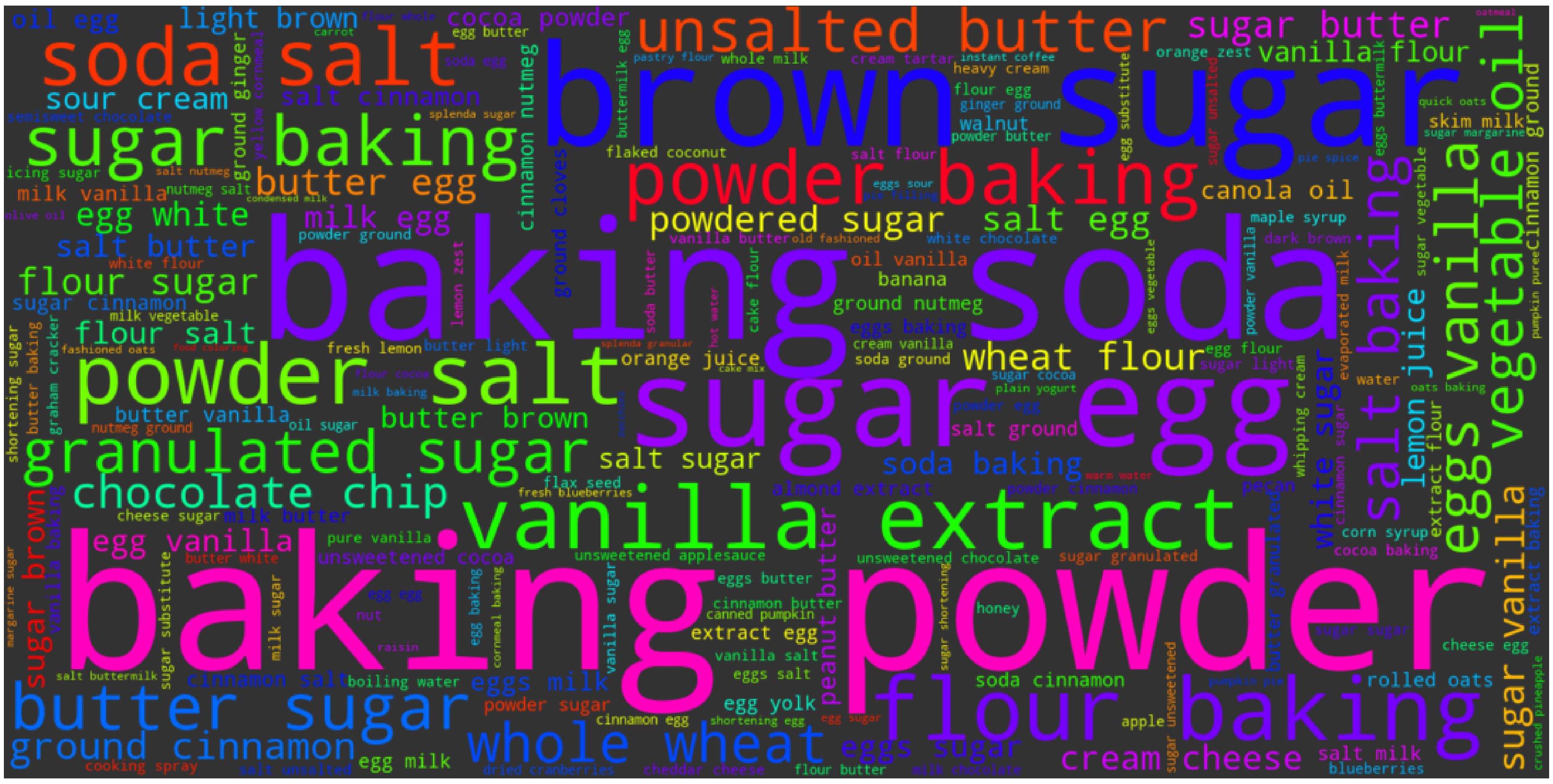
People on Food.com turned out to be healthy makers or eaters.

Please see below for bar-plots showing the sugar and protein levels of each cluster. Darker the color of bar, higher it's average rating.



Let's make word clouds for our most popular and least popular clusters and look at the ingredients precedence in them.





Regression

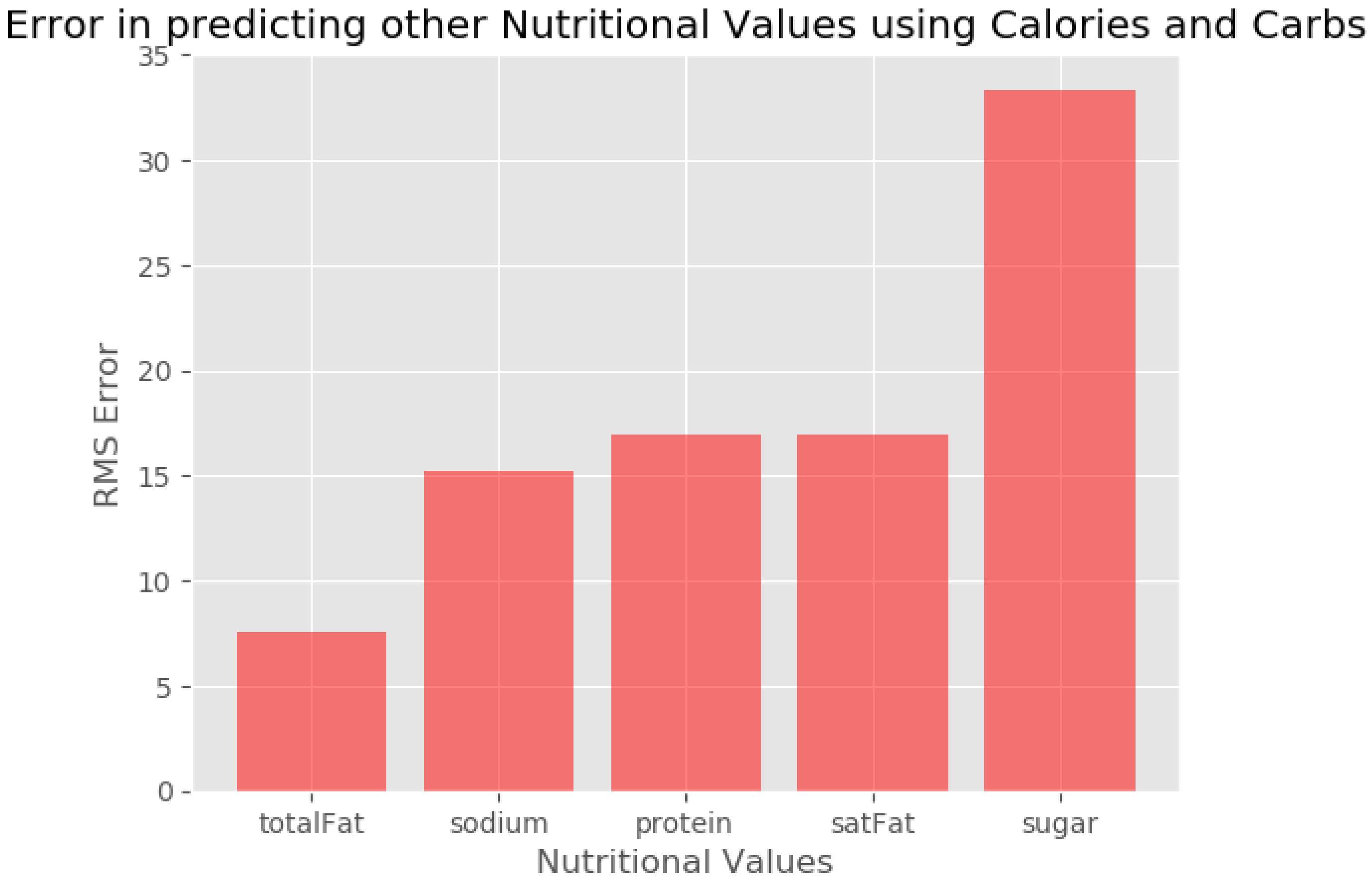
After classification and clustering, now we will see regression analysis. As we know regression analysis is done when we want to predict continious value. In our dataset we have nutritional values as 'cal', 'totalFat', 'sugar', 'sodium', 'protein', 'satFat', 'carbs' for each receipes. But we saw that for many of the receipes the nutritional values except calories and carbs are zero/missing. Hence we created a model using Gradient Boosting to predict the 'totalFat', 'sugar', 'sodium', 'protein', 'satFat' using values of 'cal','carbs'

On the similar lines are also predicting number of ingredients which will be used in a dish using the number of steps and time required to make the dish.

```
Some of the predicted values:
   totalFat_y_pred  totalFat_y_test  totalFat_abs_diff
0             8.04              8.0               0.04
1             8.04              8.0               0.04
2             2.93              3.0               0.07
3            12.92             13.0               0.08
4            12.92             13.0               0.08


Some of the predicted values:
   sugar_y_pred  sugar_y_test  sugar_abs_diff
0         16.77          17.0            0.23
1         16.77          17.0            0.23
2         16.77          17.0            0.23
3          1.28           1.0            0.28
4          1.28           1.0            0.28


Some of the predicted values:
   sodium_y_pred  sodium_y_test  sodium_abs_diff
0          19.02           19.0             0.02
1          13.92           14.0             0.08
2          19.34           19.0             0.34
3          19.34           19.0             0.34
4          11.74           11.0             0.74


Some of the predicted values:
   protein_y_pred  protein_y_test  protein_abs_diff
0            4.55             5.0              0.45
1            4.55             5.0              0.45
2            9.55            10.0              0.45
3            9.55            10.0              0.45
4            9.55             9.0              0.55


Some of the predicted values:
   satFat_y_pred  satFat_y_test  satFat_abs_diff
0          24.44           25.0             0.56
1           4.56            4.0             0.56
2          48.77           48.0             0.77
3          10.09           11.0             0.91
4           9.96            9.0             0.96



-----Predicting using ['cal', 'carbs']-----

  Predicted  with error  R2 Score
0  totalFat       21.39      0.86
1   protein       33.14      0.19
2    sodium       38.93      0.36
3    satFat       41.51      0.82
4     sugar      164.16      0.36


~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Some of the predicted values:
   n_ingredients_y_pred  n_ingredients_y_test  n_ingredients_abs_diff
0                  5.00                     5                2.59e-03
1                  5.00                     5                2.59e-03
2                  3.07                     3                7.32e-02
3                  9.11                     9                1.15e-01
4                  9.11                     9                1.15e-01



-----Predicting using ['n_steps', 'minutes']-----

       Predicted  with error  R2 Score
0  n_ingredients        3.79      0.01
```

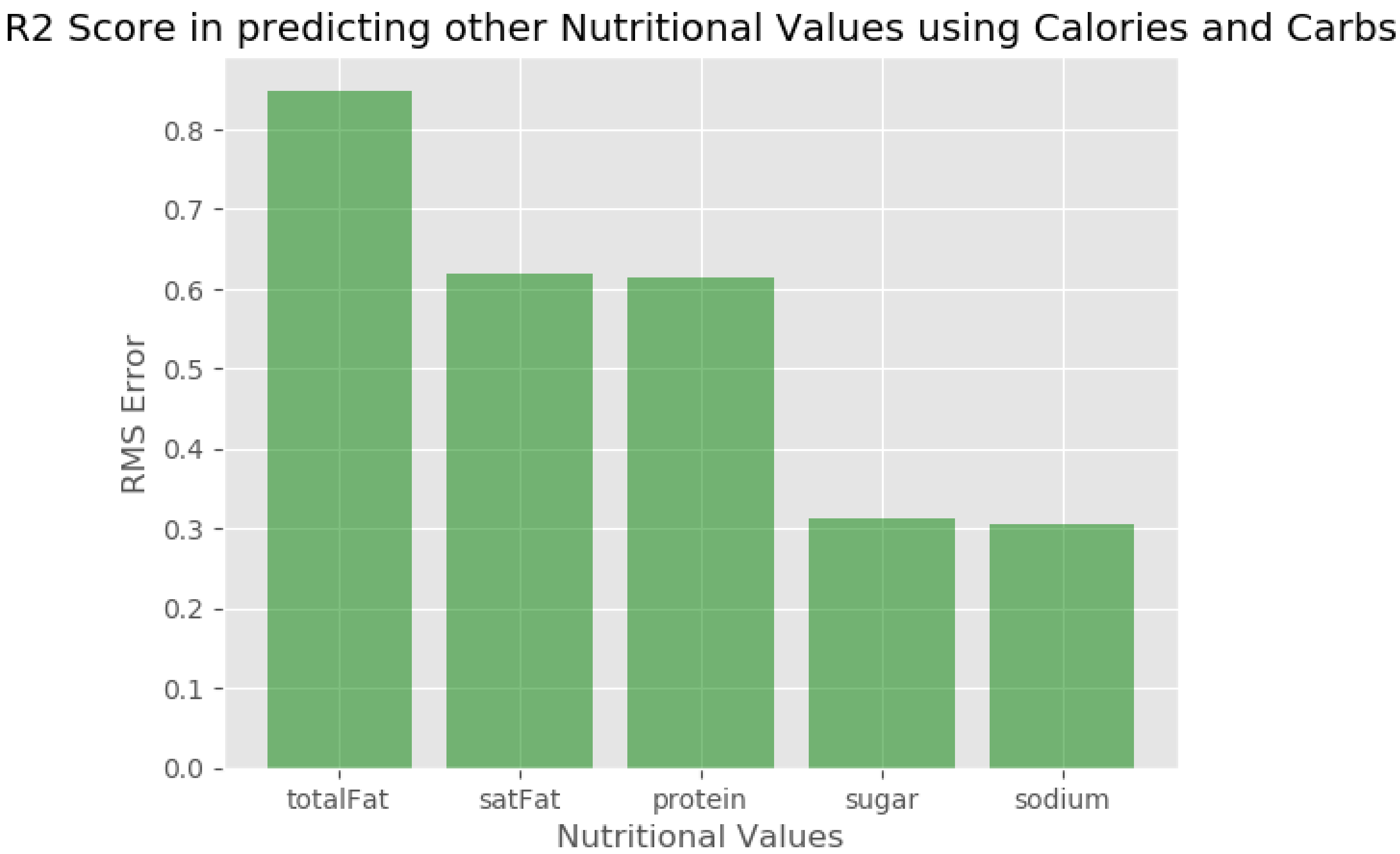
As we are predicting continuous value we can not measure accuracy but we can calculate the distance of actual and predicted value, which will be the error. One of the measure of error is Root Mean Square Error (RMSE).

So let's see what is the RMSE for different nutritional values predicted from Calories and Carbs.



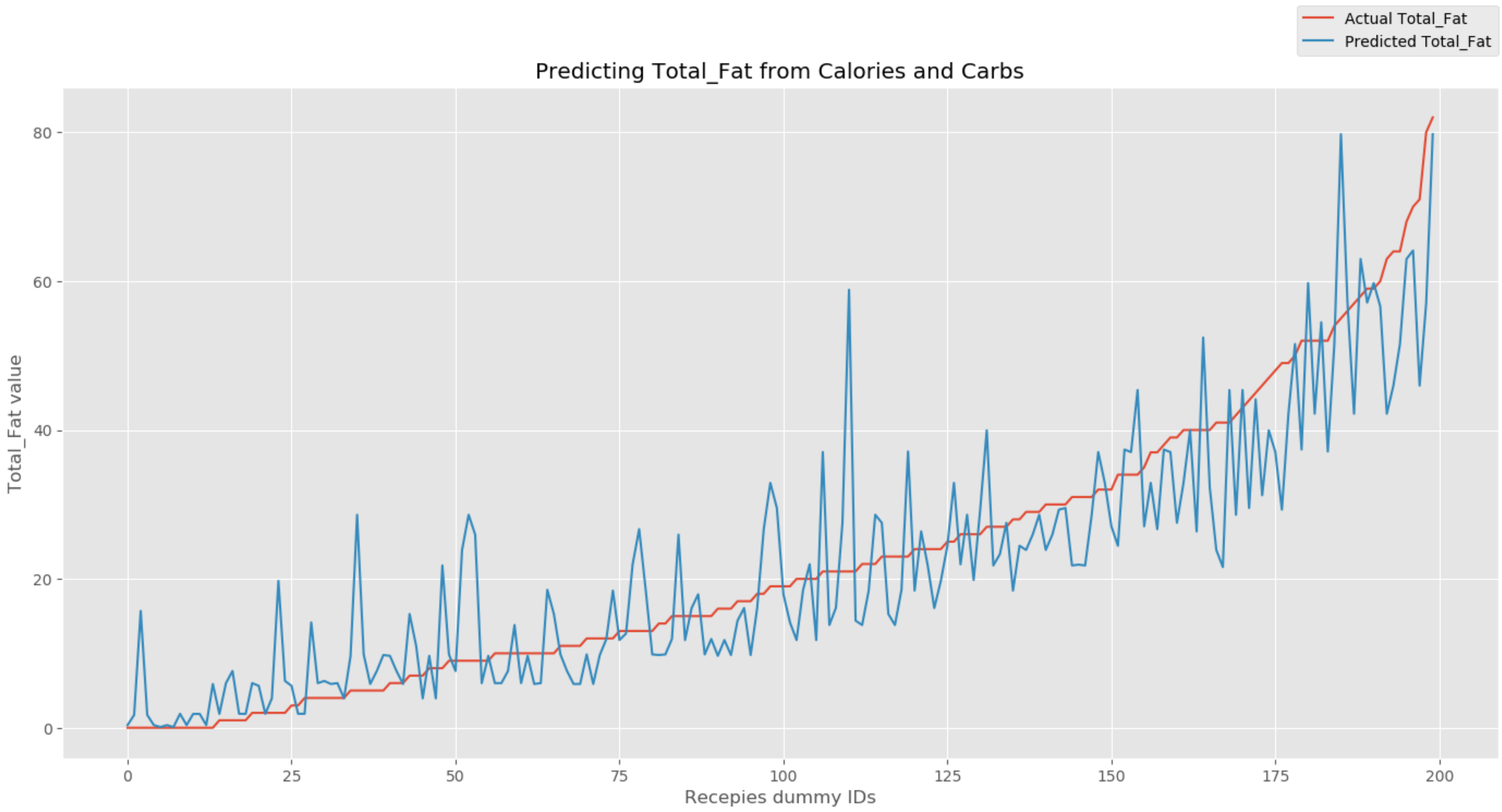
So as we can see we are able to predict TotalFat with least error and Sugar with highest error. This means we are able to predict TotalFat with most accuracy and Sugar with least accuracy.

Further let's see the R-squared (R2) Score, which represents how good the regression line fits the data. So below graph represents R2 Score for predicting values.



To vizulize how our actual and predicted values differs, let plot these values for smaple data. So we are taking 200 actual and predicted values for TotalFat to plot.

```
//anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:9: MatplotlibDeprecationWarning: Adding an axe
s using the same arguments as a previous axes currently reuses the earlier instance.  In a future version, a
new instance will always be created and returned.  Meanwhile, this warning can be suppressed, and the future
behavior ensured, by passing a unique label to each axes instance.
  if __name__ == '__main__':

<matplotlib.legend.Legend at 0x1c5bd9c828>
```



Apriori

As initially informed Food.com provides option to buy ingredients for a receipe in their portal. If we consider ingredients of a receipe are bought together, then we can consider them as items of a order/transaction. Some what like below:

	name	ingredients
id		
137739	arriba baked winter squash mexican style	['winter squash', 'mexican seasoning', 'mixed ...
31490	a bit different breakfast pizza	['prepared pizza crust', 'sausage patty', 'egg...
112140	all in the kitchen chili	['ground beef', 'yellow onions', 'diced tomato...
59389	alouette potatoes	['spreadable cheese with garlic and herbs', 'n...
44061	amish tomato ketchup for canning	['tomato juice', 'apple cider vinegar', 'sugar...
5289	apple a day milk shake	['milk', 'vanilla ice cream', 'frozen apple ju...
25274	aww marinated olives	['fennel seeds', 'green olives', 'ripe olives'...
67888	backyard style barbecued ribs	['pork spareribs', 'soy sauce', 'fresh garlic'...
70971	bananas 4 ice cream pie	['chocolate sandwich style cookies', 'chocolat...
75452	beat this banana bread	['sugar', 'unsalted butter', 'bananas', 'eggs'...

Can you guess what analysis we can do here to increase the items sale?

We can do Market Basket Analysis, which analyzes which items are frequently bought together and hence suggest items to buy based on the items on cart. By implementing this, user can get suggestion more items to add based on what s/he is buying at present.

We are using Apriori algorithm to implement Market Basket Analysis. So first like below are creating list of items/ingredients bought together.

```
Showing first to list of ingredients:

[['winter squash',
  'mexican seasoning',
  'mixed spice',
  'honey',
  'butter',
  'olive oil',
  'salt'],
 ['prepared pizza crust',
  'sausage patty',
  'eggs',
  'milk',
  'salt and pepper',
  'cheese']]
```

We get the below results once this list is passed to apriori model with the desired values for parameters of Support, Confidence and Lift.

```
Number of Rules:
1076


Example of a rule:
RelationRecord(items=frozenset({'cinnamon', 'allspice'}), support=0.007, ordered_statistics=[OrderedStatisti
c(items_base=frozenset({'allspice'}), items_add=frozenset({'cinnamon'}), confidence=0.7777777777777779, lift
=12.544802867383515)])



Listing 10 of the rules:

Rule: cinnamon --> allspice
Support: 0.007
Confidence: 0.7777777777777779
Lift: 12.544802867383515
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Rule: low-fat buttermilk --> baking powder
Support: 0.005
Confidence: 1.0
Lift: 13.157894736842106
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Rule: baking soda --> ground cloves
Support: 0.006
Confidence: 0.6
Lift: 8.108108108108109
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Rule: baking soda --> low-fat buttermilk
Support: 0.005
Confidence: 1.0
Lift: 13.513513513513514
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Rule: baking soda --> unsweetened cocoa
Support: 0.005
Confidence: 0.7142857142857143
Lift: 9.652509652509654
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Rule: onion --> bay leaves
Support: 0.005
Confidence: 0.625
Lift: 3.6982248520710055
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Rule: bread flour --> sugar
Support: 0.007
Confidence: 1.0
Lift: 5.2356020942408374
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Rule: butter --> egg yolks
Support: 0.006
Confidence: 0.75
Lift: 3.5885167464114835
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Rule: butter --> swiss cheese
Support: 0.005
Confidence: 1.0
Lift: 4.784688995215311
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Rule: onion --> celery
Support: 0.029
Confidence: 0.6304347826086957
Lift: 3.7303833290455364
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

Conlusion

To summarize the blog, let's see what all we did. We started with selecting interesting data. We choose data of Food.com from kaggle. To understand the data first we did the data analysis, where we saw different data files and there length, central tendency metric for various columns/features, the relation between the features and the outlier analysis. Once we analyzed the data and saw the issues, we worked upon to resolve them by handling outliers.

Then we started with classification where we introduced the new column as Cusine for our data set. Using the new column we analysed our data and came up with interesting analysis. Second, we performed clustering over the ingredients, for that we performed PCA, vectorization and K-means. Analysing cluster over rating and review provides us with meaningful insights. After that, we saw regression where by help of Carbs and Calories we predected other nutritional values using Gradient Boosting Regression. Last but not the least we performed Market Basket Analysis which can be profitable for the sales.

We hope this blog will be helpful, for any suggestions please email at any of the following: agupta33@student.gsu.edu sdasari3@student.gsu.edu sdawani1@student.gsu.edu

Credits and References

- http://www.ultravioletanalytics.com/blog/tf-idf-basics-with-pandas-scikit-learn (http://www.ultravioletanalytics.com/blog/tf-idf-basics-with-pandas-scikit-learn)
- https://medium.com/datadriveninvestor/an-introduction-to-grid-search-ff57adcc0998 (https://medium.com/datadriveninvestor/an-introduction-to-grid-search-ff57adcc0998)
- https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb (https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb)
- https://www.kaggle.com/etsc9287/food-com-eda-and-text-analysis (https://www.kaggle.com/etsc9287/food-com-eda-and-text-analysis)