
WHAT IS
THE BEST
RECIPE FOR YOU?
Ofir Morhaim
Maxim Svishech
Data science Project

What are we looking for?

Investigating the site Food.com, we would want to see a connection between recipe parameters and its rating.
We will test if preparation time\calories\number of ingredients influence the rating.
We would also check the most popular ingredients and the most common food categories.

Our research assumption is that time\calories\number of ingredients affect the rating of the recipe

Ofir Morhaim

Maxim Svishech

Main process steps of the project

Ofir Morhaim

Maxim Svishech

Web Crawling

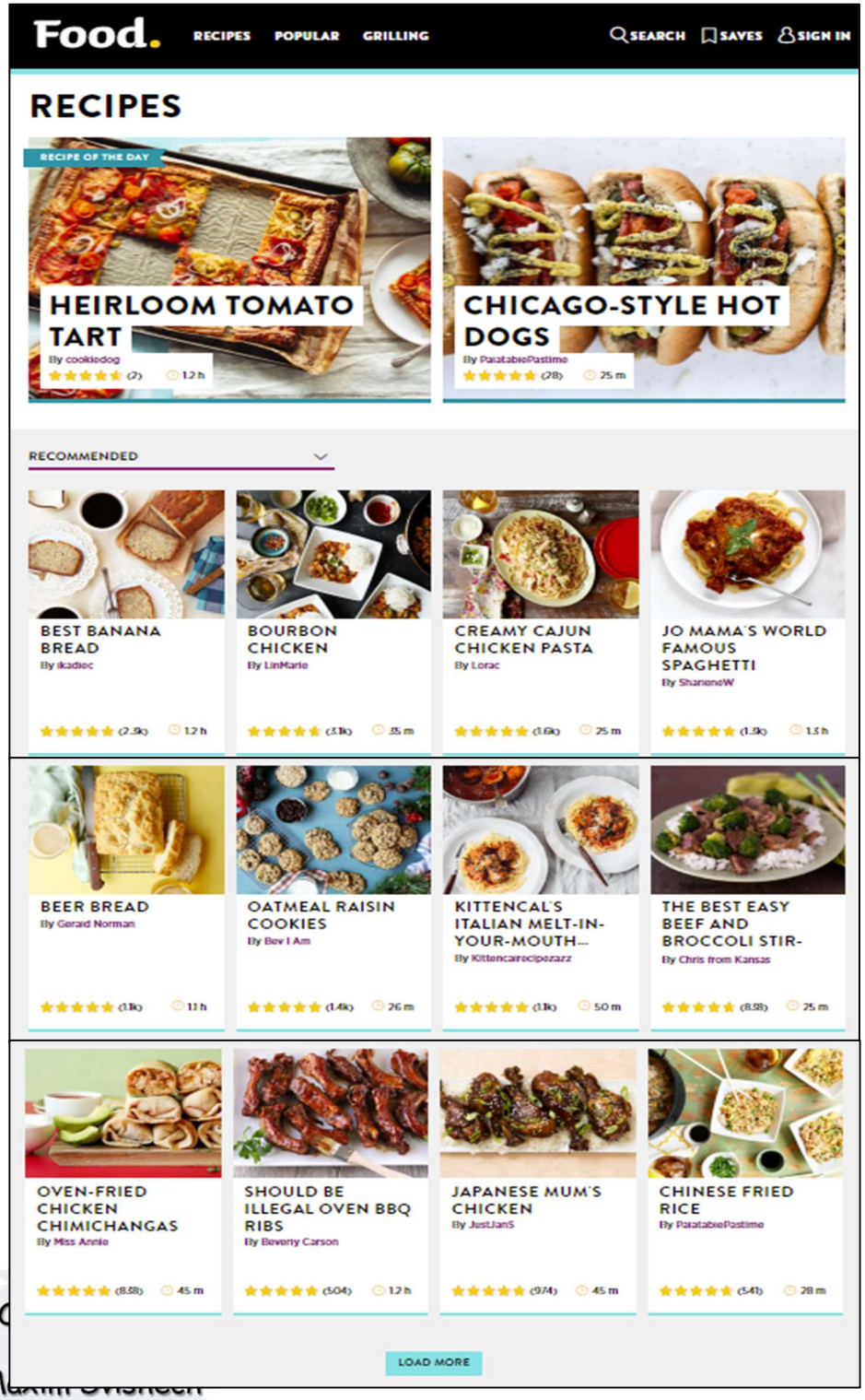
At this step we started with data extraction using Crawling.

We took all the data from a recipe site Food.com .

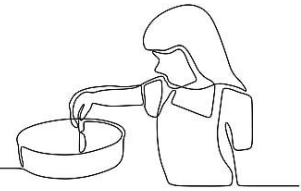
At this step, our main tools were: **BeautifulSoup and Selenium**

Ofir Morhaim

Maxim Svishech



Web Crawling



In this picture, you can see the website page from which we extracted the data, we sorted the recipes by recommended.

This page is an example, in order to pull data, we will have to go through all the pages of the recipes (although on the site everything is visible on one page).

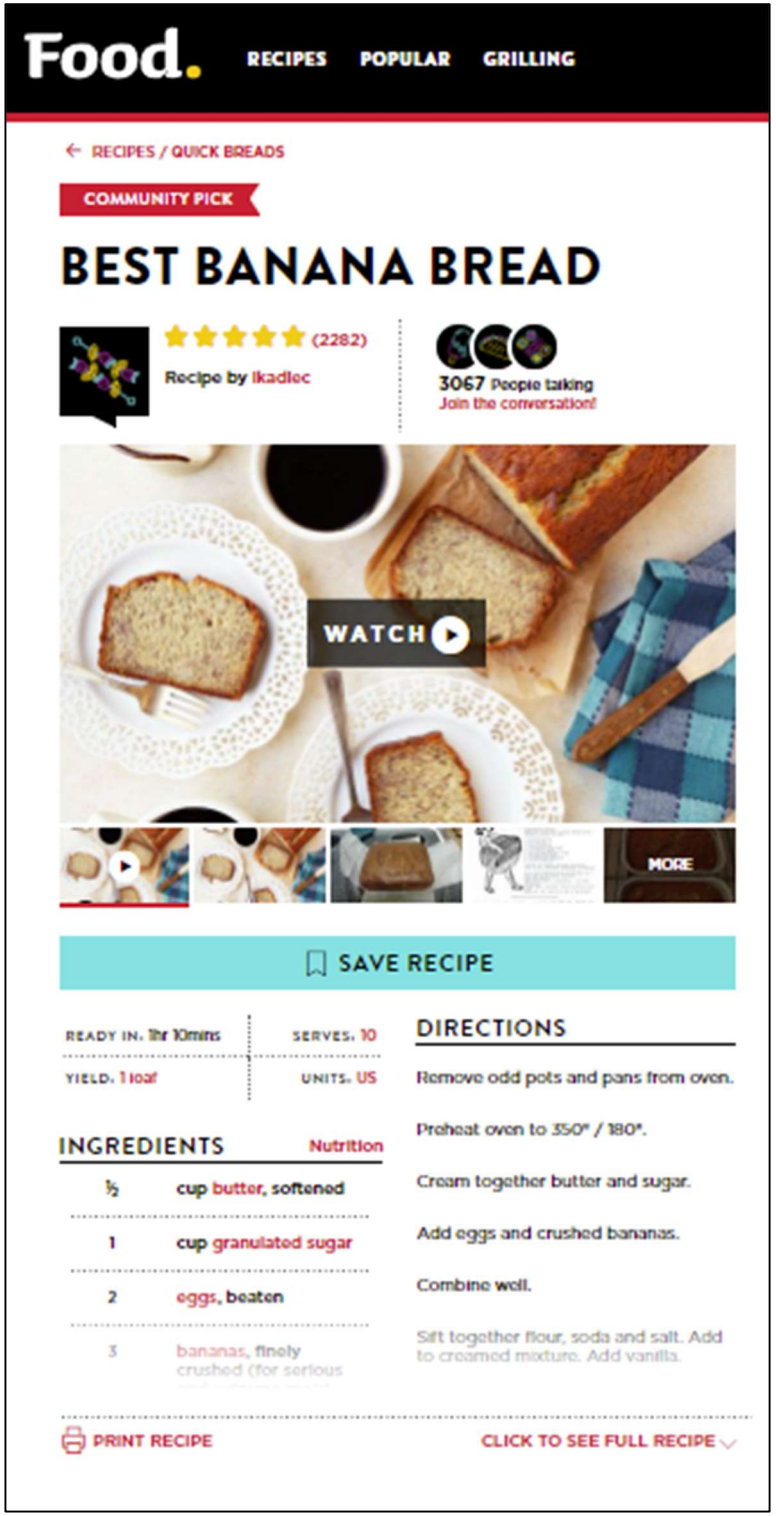
For example, the recipes found on page 3: "https://www.food.com/recipe?ref=nav&pn=3"

Web Crawling

Within each page, we go through all the recipes that exist within it.

From this page we begin to pull the data:

* Name of the dish
* Preparation time
* Several ingredients
* List of ingredients
* Rating
* Number of rated
* Recipe category
* Calories



Ofir Morhaim

Maxim Svishech

Web Crawling

The DataFrame after extracting all the data.

	Name	Rating	Rank number	Preparation time in mins	Ingredients number	Ingredients	Category
0	Best Banana Bread	width:96.2%;	2282	1hr 10mins	8	[butter, granulated sugar, eggs, bananas, all-...	Quick Breads
1	Bourbon Chicken	width:91.8%;	3069	35mins	10	[boneless chicken breasts, olive oil, garlic c...	Chicken Breast
2	Creamy Cajun Chicken Pasta	width:96.8%;	1591	25mins	12	[boneless skinless chicken breast halves, ling...	Chicken Breast
3	Jo Mama's World Famous Spaghetti	width:94.60000000000001%;	1340	1hr 20mins	15	[Italian sausage, onion, garlic cloves, diced ...	Spaghetti
4	Best Ever Banana Cake With Cream Cheese Frosting	width:96%;	1414	1hr 15mins	15	[bananas, lemon juice, flour, baking soda, sal...	Dessert
5	To Die for Crock Pot Roast	width:91.60000000000001%;	1695	9hrs 5mins	2	[beef roast, water]	One Dish Meal
6	Beer Bread	width:95.8%;	1055	1hr 3mins	6	[flour, baking powder, salt, sugar, beer, butter]	Quick Breads
7	Oatmeal Raisin Cookies	width:96.2%;	1418	26mins	11	[all-purpose flour, baking soda, baking powder...	Drop Cookies
8	Kittencal's Italian Melt-In-Your-Mouth Meatballs	width:97.8%;	1073	50mins	9	[ground beef, egg, parmesan cheese, garlic, sa...	Meat
9	The Best Easy Beef and Broccoli Stir-Fry	width:93%;	838	25mins	12	[cornstarch, water, water, garlic powder, bone...	One Dish Meal
10	Yes, Virginia There is a Great Meatloaf	width:93%;	1389	1hr 20mins	10	[ground beef, egg, sweet onion, table salt, bl...	Meatloaf
11	Beth's Melt in Your Mouth Barbecue Ribs (Oven)	width:99%;	747	3hrs 15mins	5	[light brown sugar, hickory smoke salt, paprik...	Pork
12	Olive Garden Copycat Zuppa Toscana	width:97.6%;	721	1hr	12	[Italian sausage, russet baking potatoes, Acce...	Pork
13	Soft Snickerdoodle Cookies	width:95.19999999999999%;	775	25mins	10	[butter, sugar, eggs, flour, cream of tartar, ...	Drop Cookies
14	Sour Cream Banana Bread	width:97.39999999999999%;	631	1hr 15mins	8	[margarine, sugar, eggs, vanilla, flour, bakin...	Quick Breads
15	Delicious Chicken Pot Pie	width:97.39999999999999%;	767	1hr 45mins	11	[potato, onion, celery, carrot, margarine, all...	Savory Pies
16	Crock-Pot Chicken With Black Beans & Cream Cheese	width:89.2%;	1659	4hrs 3mins	5	[boneless chicken breasts, black beans, corn, ...	One Dish Meal
17	Simply Sour Cream Chicken Enchiladas	width:93%;	808	45mins	9	[chicken breast, onion, flour tortillas, monte...	Chicken Breast
						[unbleached all-purpose flour,	

Ofir Morhaim

Maxim Svishech

Scrubbing Data

At this step, we clean, arrange and make the table we created more readable and easier to perform operations on.

The actions we perform:

* Handling missing data
* Duplicate handling
* Noise cleaning
* Create additional tables

Scrubbing Data

First, we edited the table for more readable and thus also more useful.

* Deleting duplicates, deleting rows with missing values

* The "preparation time" string became minutes (int)

* The "Ingredients List" string has become more readable

* "Rating" string turned into a number between 1-100 (float)

Before

	Name	Rating	Rank number	Preparation time in mins	Ingredients number	Ingredients	Category
0	Best Banana Bread	width:96.2%;	2282	1hr 10mins	8	[butter, granulated sugar, eggs, bananas, all-...	Quick Breads
1	Bourbon Chicken	width:91.8%;	3069	35mins	10	[boneless chicken breasts, olive oil, garlic c...	Chicken Breast
2	Creamy Cajun Chicken Pasta	width:96.8%;	1591	25mins	12	[boneless skinless chicken breast halves, ling...	Chicken Breast
3	Jo Mama's World Famous Spaghetti	width:94.60000000000001%;	1340	1hr 20mins	15	[Italian sausage, onion, garlic cloves, diced ...	Spaghetti

After

	Name	Rating	Rank number	Preparation time in mins	Ingredients number	Ingredients	Category
0	Best Banana Bread	96.2	2282	70	8	butter,granulated sugar,eggs,bananas,all-purpo...	Quick Breads
1	Bourbon Chicken	91.8	3069	35	10	boneless chicken breasts,olive oil,garlic clov...	Chicken Breast
2	Creamy Cajun Chicken Pasta	96.8	1591	25	12	boneless skinless chicken breast halves,lingui...	Chicken Breast
3	Jo Mama's World Famous Spaghetti	94.6	1340	80	15	Italian sausage,onion,garlic cloves,diced toma...	Spaghetti

Ofir Morh

Maxim Svishech

Scrubbing Data

After, we created tables for "preparation time" and "category" which replace the existing value with number.

In a category table, each category has a number.

In the time table for each range of minutes, there is a category value, for example from 33 minutes to 64 the value will be 2.

Then, we will replace the original values of "Category" with the numeric value that belongs to it.

And we will add a column with the "time group" number (we will not replace the original value because we still want the exact time for each recipe)

Time Table

	Start Range	End Range	Number
0	0	32	1
1	33	64	2
2	65	128	3
3	129	256	4
4	257	512	5
5	513	1024	6
6	1025	2048	7

Category Table

	Category	Numbers
0	Quick Breads	16
1	Chicken Breast	6
2	Chicken Breast	6
3	Spaghetti	19
4	Dessert	8
5	One Dish Meal	13
6	Quick Breads	16
7	Drop Cookies	9
8	Meat	11
9	One Dish Meal	13
10	Meatloaf	12
11	Pork	14
12	Pork	14
13	Drop Cookies	9
14	Quick Breads	16
15	Savory Pies	18
16	One Dish Meal	13
17	Chicken Breast	6
18	Breads	3
19	Potato	15
20	Breakfast	4
21	Bar Cookie	1

Ofir Morhaim

Maxim Svishech

Scrubbing Data

Create a new table that contains all the dishes (each in a row), and columns of the ingredient names of all the dishes (and delete double columns).

For a recipe that contains the ingredient, the value will be 1, otherwise 0

	Name	butter	granulated sugar	eggs	bananas	all-purpose flour	baking soda	Preparation time in mins_categorial	salt	vanilla	...	kidney beans	dry mustard
0	Best Banana Bread	1	1	1	1	1	1	0	1	1	...	0	0
1	Bourbon Chicken	0	0	0	0	0	0	0	0	0	...	0	0
2	Creamy Cajun Chicken Pasta	1	0	0	0	0	0	0	1	0	...	0	0
3	Jo Mama's World Famous Spaghetti	0	0	0	0	0	0	0	1	0	...	0	0
4	Best Ever Banana Cake With Cream Cheese Frosting	1	0	1	1	0	1	0	1	1	...	0	0
5	To Die for Crock Pot Roast	0	0	0	0	0	0	0	0	0	...	0	0
6	Beer Bread	1	0	0	0	0	0	0	1	0	...	0	0
7	Oatmeal Raisin Cookies	0	0	1	0	1	1	0	0	1	...	0	0
8	Kittencal's Italian Melt-In-Your-Mouth Meatballs	0	0	0	0	0	0	0	1	0	...	0	0

Ofir Morhaim

Maxim Svishech

Scrubbing Data

The DataFrame after Scrubbing Data.

	Name	Rating	Rank number	Preparation time in mins	Ingredients number	Ingredients	Category	Preparation time in mins_categorial
0	Best Banana Bread	96.2	2282	70	8	butter,granulated sugar,eggs,bananas,all-purpo...	16	3.0
1	Bourbon Chicken	91.8	3069	35	10	boneless chicken breasts,olive oil,garlic clov...	6	2.0
2	Creamy Cajun Chicken Pasta	96.8	1591	25	12	boneless skinless chicken breast halves,lingui...	6	1.0
3	Jo Mama's World Famous Spaghetti	94.6	1340	80	15	Italian sausage,onion,garlic cloves,diced toma...	19	3.0
4	Best Ever Banana Cake With Cream Cheese Frosting	96	1414	75	15	bananas,lemon juice,flour,baking soda,salt,but...	8	3.0
5	To Die for Crock Pot Roast	91.6	1695	545	2	beef roast,water	13	6.0
6	Beer Bread	95.8	1055	63	6	flour,baking powder,salt,sugar,beer,butter	16	2.0
7	Oatmeal Raisin Cookies	96.2	1418	26	11	all-purpose flour,baking soda,baking powder,ko...	9	1.0
8	Kittencal's Italian Melt-In-Your-Mouth Meatballs	97.8	1073	50	9	ground beef,egg,parmesan cheese,garlic,salt,fr...	11	2.0
9	The Best Easy Beef and Broccoli Stir-Fry	93	838	25	12	cornstarch,water,water,garlic powder,boneless ...	13	1.0
10	Yes, Virginia There is a Great Meatloaf	93	1389	80	10	ground beef,egg,sweet onion,table salt,black p...	12	3.0
11	Beth's Melt in Your Mouth Barbecue Ribs (Oven)	99	747	195	5	light brown sugar,hickory smoke salt,paprika,g...	14	4.0
12	Olive Garden Copycat Zuppa Toscana	97.6	721	60	12	Italian sausage,russet baking potatoes,Accent ...	14	2.0
13	Soft Snickerdoodle Cookies	95.2	775	25	10	butter,sugar,eggs,flour,cream of tartar,baking...	9	1.0
14	Sour Cream Banana Bread	97.4	631	75	8	margarine,sugar,eggs,vanilla,flour,baking soda...	16	3.0
15	Delicious Chicken Pot Pie	97.4	767	105	11	potato,onion,celery,carrot,margarine,all-purpo...	18	3.0
16	Crock-Pot Chicken With Black Beans & Cream Cheese	89.2	1659	243	5	boneless chicken breasts,black beans,corn,sals...	13	4.0
17	Simply Sour Cream Chicken Enchiladas	93	808	45	9	chicken breast,onion,flour tortillas,monterey ...	6	2.0
18	Southern Buttermilk Biscuits	94.4	810	22	6	unbleached all-purpose flour,baking soda,bakin...	3	1.0
19	Scalloped Potatoes	94.8	854	75	9	potatoes,butter,flour,milk,salt,cayenne pepper...	15	3.0
20	Pancakes	93.8	1150	10	7	egg,milk,butter,flour,sugar,baking powder,salt	4	1.0

Ofir Morhaim

Maxim Svishech

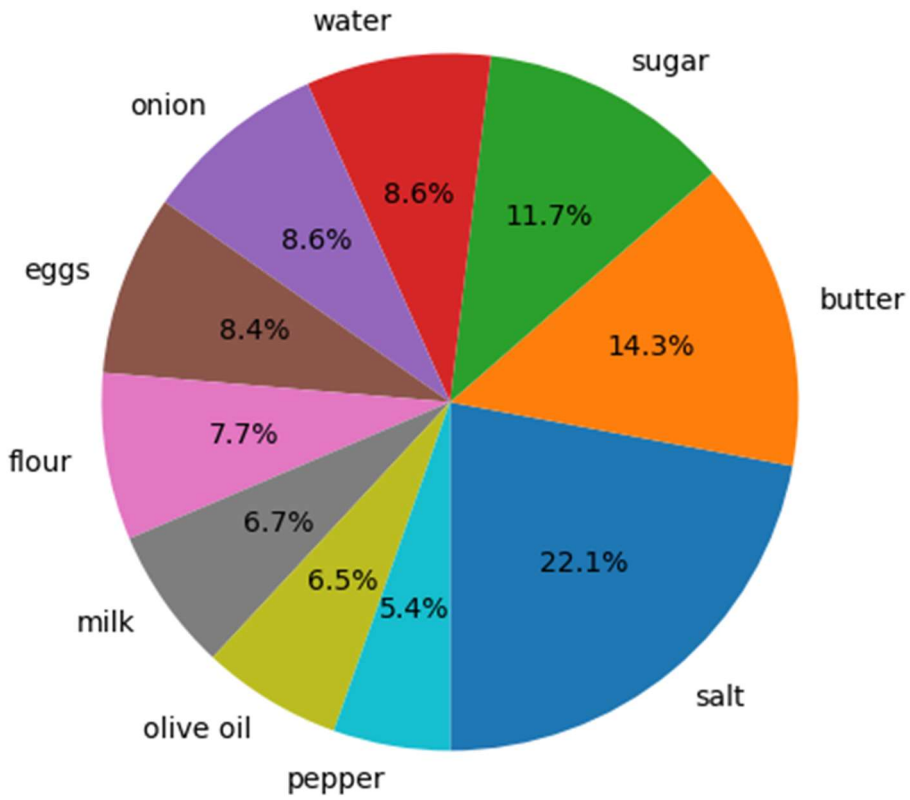
Exploring Data

In this section, we will present graphs and tables derived from the data in the main table

Ofir Morhaim

Maxim Svishech

Percentage the 10 most common ingredients



	Name	Number
6	salt	4739
0	butter	3080
41	sugar	2516
16	water	1843
19	onion	1838
2	eggs	1802
39	flour	1663
52	milk	1435
9	olive oil	1397
140	pepper	1163

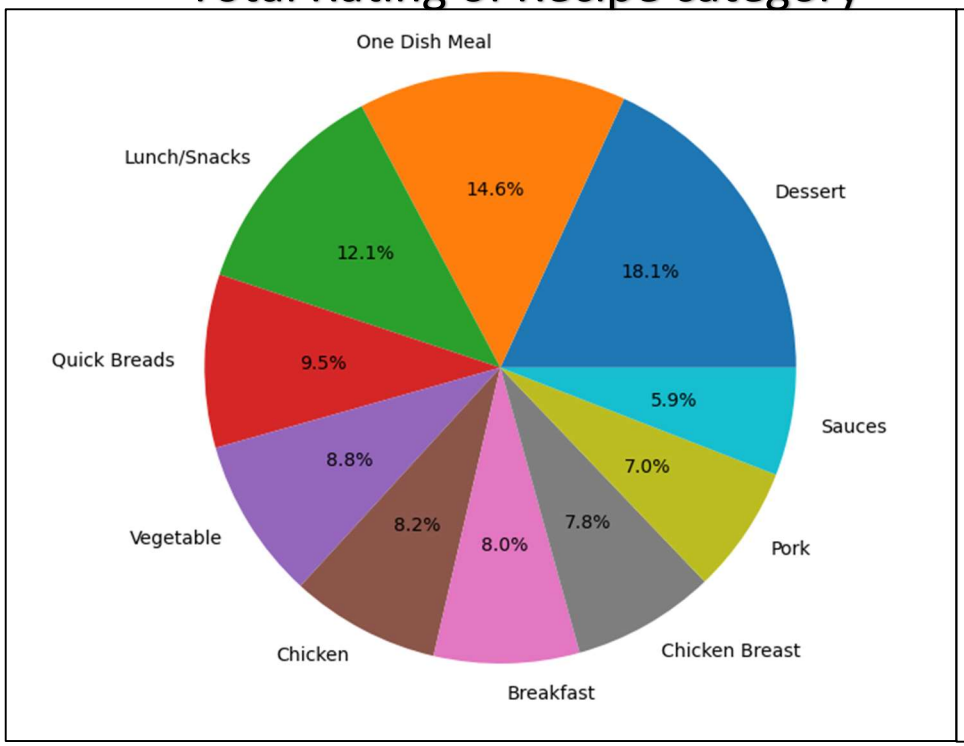
Ofir Morhaim

Maxim Svishech

Top 10 of:

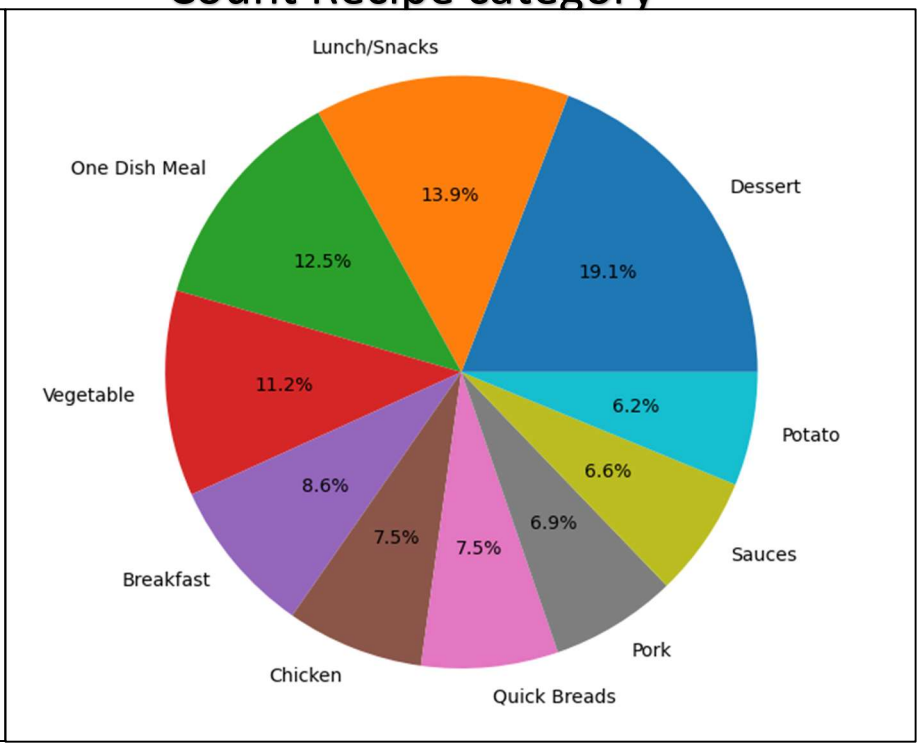
Exploring Data

Total Rating of Recipe category



Category	Value	Counts	Total Rating
Dessert	52	1024	54133
One Dish Meal	114	671	43693
Lunch/Snacks	95	745	36221
Quick Breads	140	400	28267
Vegetable	187	601	26347
Chicken	32	401	24394
Breakfast	18	458	23745
Chicken Breast	33	311	23324
Pork	132	370	20793
Sauces	146	354	17598

Count Recipe category

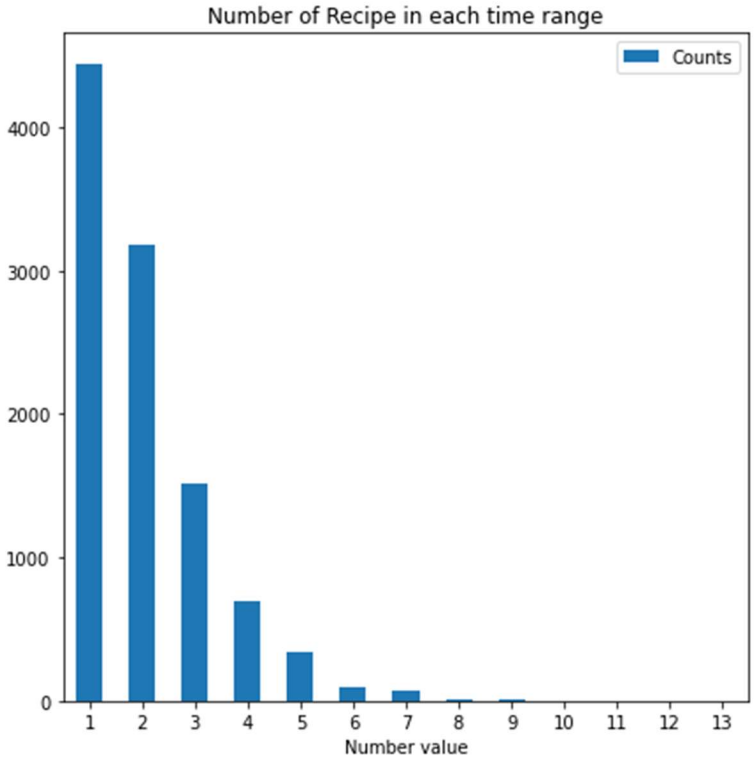


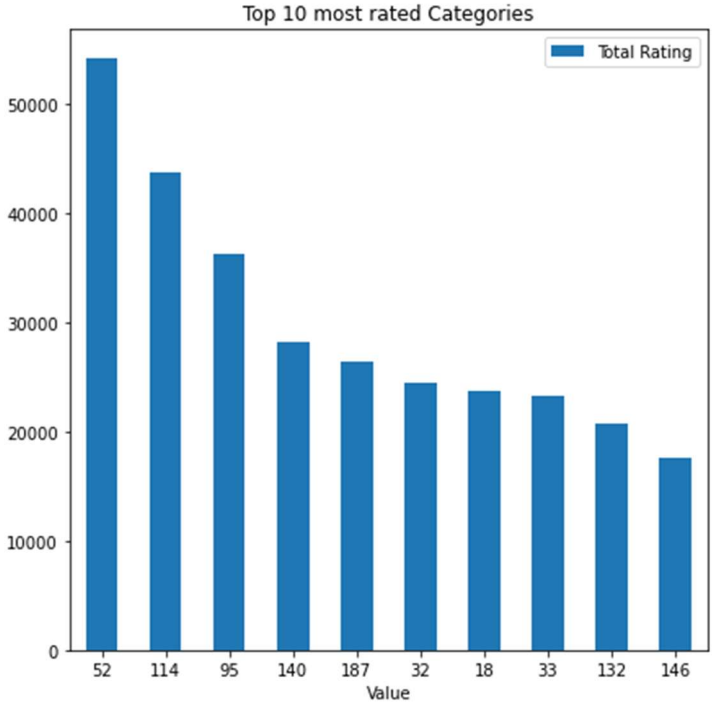
Category	Value	Counts	Total Rating
Dessert	52	1024	54133
Lunch/Snacks	95	745	36221
One Dish Meal	114	671	43693
Vegetable	187	601	26347
Breakfast	18	458	23745
Chicken	32	401	24394
Quick Breads	140	400	28267
Pork	132	370	20793
Sauces	146	354	17598
Potato	134	332	16248

Ofir Morhaim

Maxim Svishech

Exploring Data





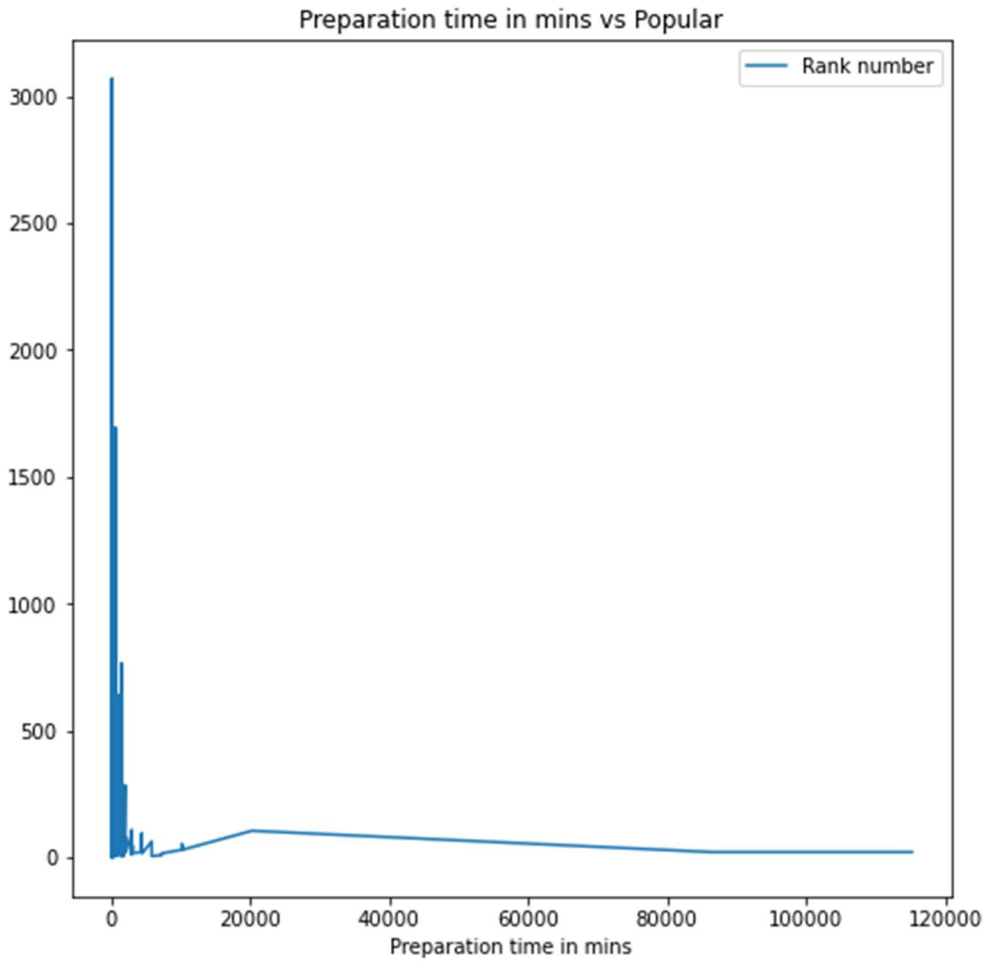
	Number value	Start time Range	End time Range
0	1	1	32
1	2	33	64
2	3	65	128
3	4	129	256
4	5	257	512
5	6	513	1024
6	7	1025	2048
7	8	2049	4096
8	9	4097	8192
9	10	8193	16384
10	11	16385	32768
11	12	32769	65536
12	13	65537	131072

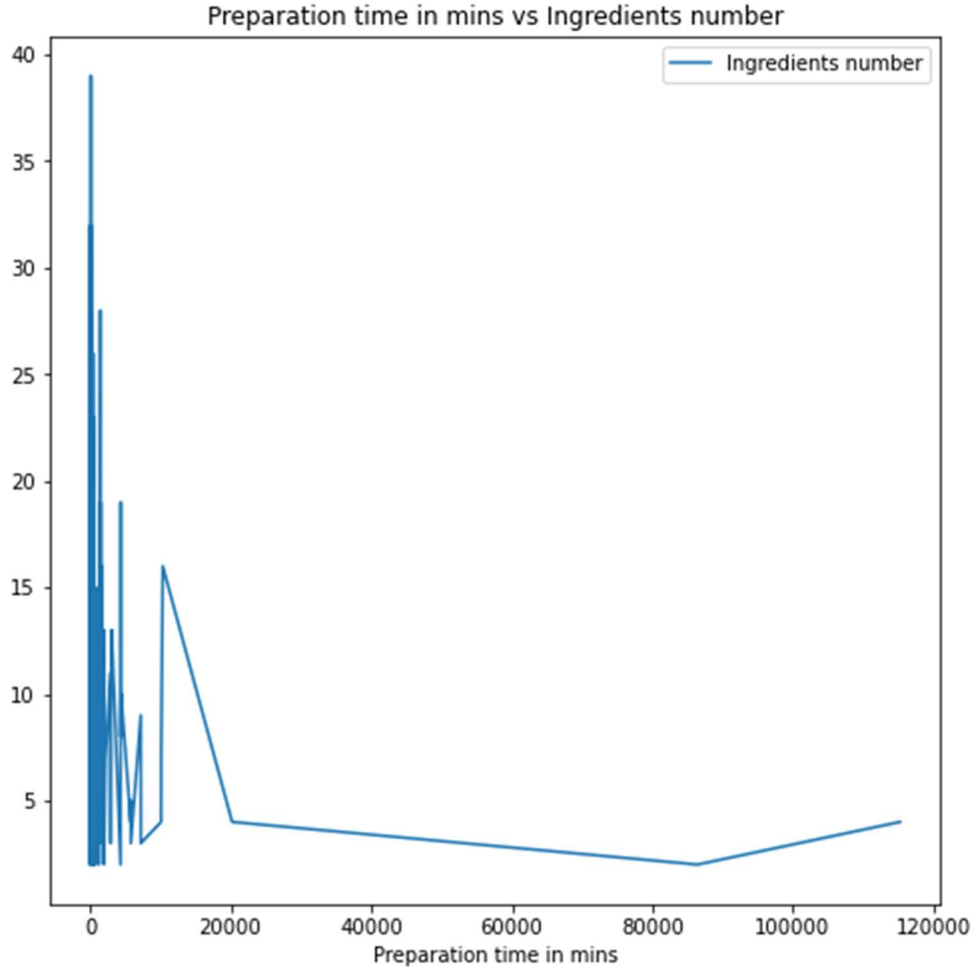
	Category	Value	Total Rating
52	Dessert	52	54133
114	One Dish Meal	114	43693
95	Lunch/Snacks	95	36221
140	Quick Breads	140	28267
187	Vegetable	187	26347
32	Chicken	32	24394
18	Breakfast	18	23745
33	Chicken Breast	33	23324
132	Pork	132	20793
146	Sauces	146	17598

Ofir Morhaim

Maxim Svishech

Exploring Data





Ofir Morhaim

Maxim Svishech

Modelling Data

In this part, we will create a machine learning algorithm for predicting the data we want and display the results.

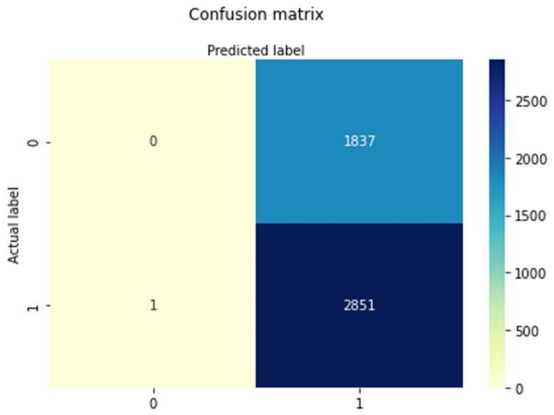
Modelling Data

```
feature_cols = ['Preparation time in mins', 'Ingredients number','Calories per serving']
X = df_CleanTable[feature_cols] # Features
y = df_CleanTable.Rating_categorial

X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.6,random_state=1)

logreg = LogisticRegression()

# fit the model with data
logreg.fit(X_train,y_train)
y_pred=logreg.predict(X_test)



cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
cnf_matrix


class_names=[0,1] # name  of classes
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
plt.yticks(tick_marks, class_names)
# create heatmap
sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu" ,fmt='g')
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)
plt.ylabel('Actual label')
plt.xlabel('Predicted label')


print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
print("Precision:", metrics.accuracy_score(y_test, y_pred))
print("Recall:",metrics.recall_score(y_test, y_pred))

df = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
df
```

```
Accuracy: 0.6080187673277885
Precision: 0.6080187673277885
Recall: 0.9996493688639552
```

	Actual	Predicted
4939	1	1
5113	1	1
7565	0	1
39	0	1
299	1	1
...	...	...
764	0	1
6549	0	1
4155	1	1
5239	0	1
7268	1	1

4689 rows × 2 columns

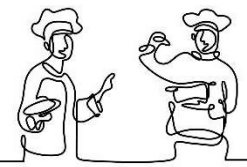


After the EDA part, we understand that most of the recipes are high rated (above 80 percent).

Based on the parameters we have (preparation time\calories\number of ingredients) we want to use machine learning to teach the algorithm to predict if a dish is excellent (above 90 percent) or not and to predict the final rating

Ofir Morhaim

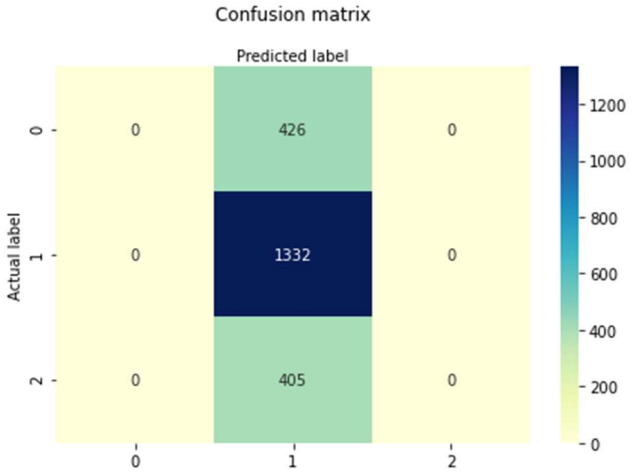
Maxim Svishech



Modelling Data

Naïve bayes

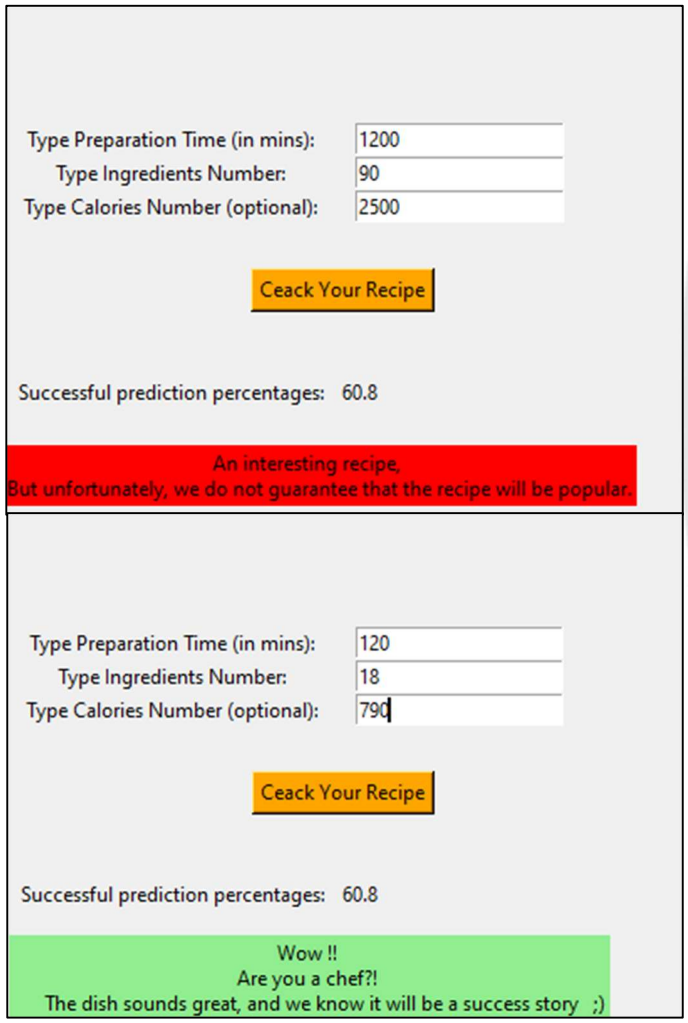
```
X = df_CleanTable["Rating"] # Features
y = df_CleanTable.Calories_categorial

x_train, x_test, y_train, y_test = train_test_split(X, y, test_size= 0.25, random_state=1)

x_train= x_train.values.reshape(-1, 1)
y_train= y_train.values.reshape(-1, 1)
x_test = x_test.values.reshape(-1, 1)

classifier = GaussianNB()
classifier.fit(x_train, y_train)
y_pred = classifier.predict(x_test)

ac = accuracy_score(y_test,y_pred)
cm = confusion_matrix(y_test, y_pred)

print("Accuracy:",ac)

class_names=[0,1] # name  of classes
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
plt.yticks(tick_marks, class_names)
# create heatmap
sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu" ,fmt='g')
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)
plt.ylabel('Actual label')
plt.xlabel('Predicted label')
```



Ofir Morhaim

Maxim Svishech

Bonus



Window that predicts if a dish is successful based on the preparation time\ingredient number\calories (optional)

As you can see the prediction model is not good enough, the accuracy percentage is not good enough and it is only predicting good dishes and not bad dished.

Ofir Morhaim

Maxim Svishech

Summary

As you can see with all the data ,the graphs and the machine learning module There is not really a parameter that makes a recipe more popular than other recipe.

You can get good recipe rating with many or few ingredients, the preparation time is not a parameter too and not the calories.

Ofir Morhaim

Maxim Svishech

Summary

As you can see with all the data ,the graphs and the machine learning module There is not really a parameter that makes a recipe more popular than other recipe.

You can get good recipe rating with many or few ingredients, the preparation time is not a parameter too and not the calories.