UNIVERSITY OF TARTU
Institute of Computer Science
Software Engineering (Enterprise System)

Analysis of Recipes by Ratings, Ingredients and Nutritional benefits

Data Mining Project

Victor Aluko

Madhushree Signh

Antony Orenge

Bilal Abdullah

Supervisor(s): Jaak Vilo

Tartu 2017

Introduction

In today's world, there are millions of recipes that makes food yummy! But, do you know what is in your food? Do you know the ingredients of the dish that you are consuming? Does questions like healthy and nutritious but tasty food often come to your minds?

So, to help you to live a healthier life datamining helped us to let you know all the nutrition and calories present in your food!

Objective

The main objective by this analysis is to have a graphical analysis of the correlation between ratings, calories and nutritional features. Another interesting objective was to find the nutritional value present in non-vegetarian and vegetarian food.

Data Preparation

The first task was to analyzed the data and understand the dataset. Based on this understanding we can then assess the level of cleanup required for the dataset.

We identified the columns which are actually ingredients and found that some columns are specifications about how the recipes are prepared. So we noted the actual ingredient columns.

Secondly, we developed a script in *python* to remove non-ingredient columns. The script reduced the number of columns by 60%. Next, we developed another python script to sort through the over 300 columns and only identify the ingredients for each recipe. This gave us the ratings, nutritional benefits and ingredients of each recipe.

Thirdly, we also did some filtering using *Excel* to sort the non-vegetarian and the vegetarian recipes along with their ratings. In the analysis of nutrients in vegetarian and non-vegetarian recipes, we selected equal number of samples from both subsets.

The source code used for cleaning up the dataset can be found here:

https://github.com/valuko/data_cleaner

Finally, after these initial data cleanup, we divided the task into 4 parts and each team member worked on one task. For each task, we developed R-scripts for visualizing the data. The URL to the repository containing the R scripts used for visualizing the dataset is below:
https://github.com/valuko/dm_project

Methods

1. Correlation
2. K-means clustering
3. Confidence and Lift

Results: (Questions)

1. Which ingredients are the most in demand? - Predict recipe rating by ingredients
2. Which ingredients occur most of all in the highest rating meals? - Cluster
3. What is the distribution of the most frequent ingredients and how do they affect the ratings of meals? Do these ingredients also give the highest calorie meals? -
4. Wish to know the nutritional features based on your diet? Which is more nutritious, vegetarian or non-vegetarian recipes? - Nutritional features of vegetarian and non-vegetarian recipes in 5 star ratings

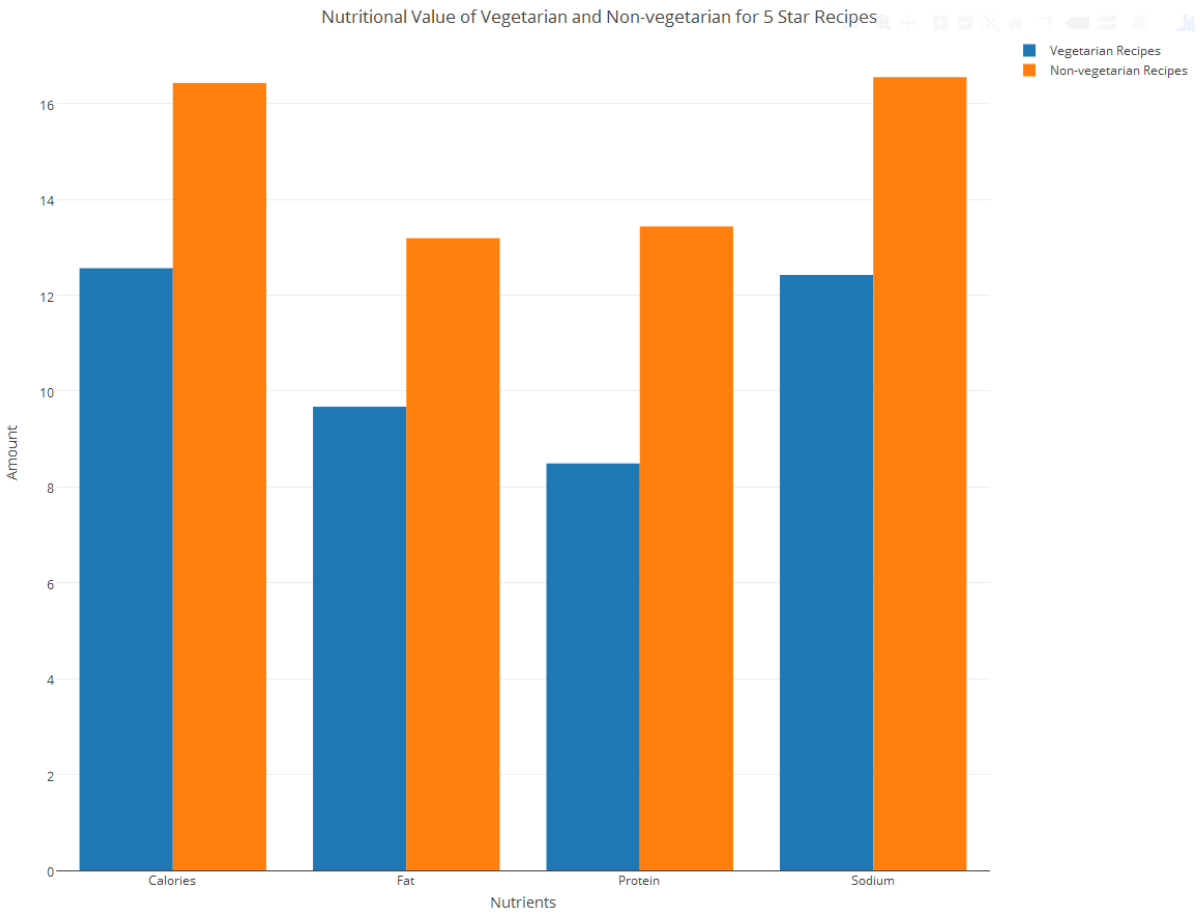


Fig. 4.1

The cumulative nutritional value of non-vegetarian recipes was observed to be almost always higher than those of vegetarian recipes. This is to be expected, but does mean that non-vegetarian diets are better? The next plot answers that question.

Fig. 4.2

Comparing Fat and Calories, we observed that while non-vegetarian recipes had a higher amount of Fat and Calories, those in vegetarian recipes are more evenly spread as shown in the plot above. In other words, more vegetarian recipes contain a healthy amount of fat and calories. That is in contrast to the vegetarian recipes where the nutrients are concentrated in a smaller number of recipes.

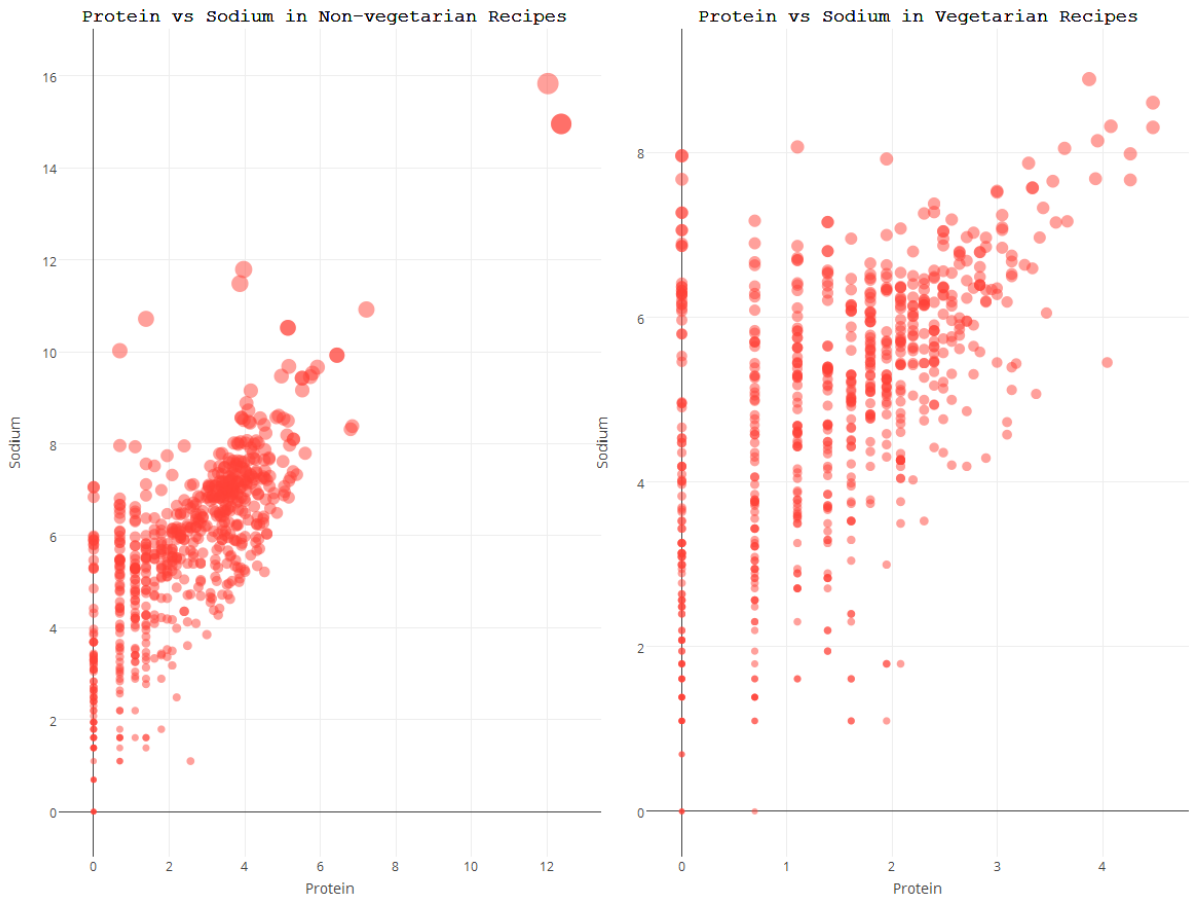
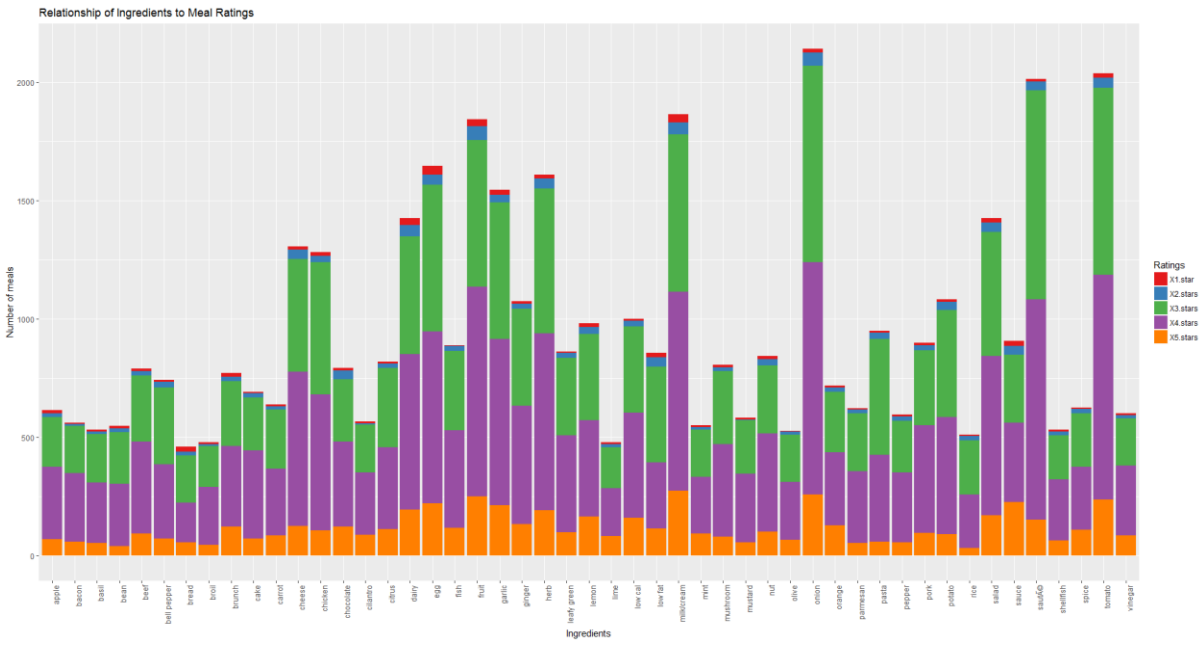


Fig. 4.3

The same pattern is observed here, but now, the spread of Protein and Sodium among a larger number of vegetarian recipes is even more pronounced. A larger number of vegetarian recipes contain a healthy dose of Protein and Sodium while in non-vegetarian recipes they are concentrated in smaller number of recipes.

In conclusion, while eating non-vegetarian dishes might sometimes provide you with a higher amount of nutrients, vegetarian dishes are more balanced and contain just the right amount of nutrients for healthy eating.



The visualization shows the plot of the top 50 ingredients and their effects on the ratings. It shows that Onions and Milk/cream impact more on the 5 stars and 4 stars' recipes while ingredients like sauce impact more on the 3 stars' recipes.

Conclusion

We analyzed the dataset and came up with details regarding the benefits of the ingredients in relation to the recipes. The analysis also involved predicting the ratings of an athlete based on recipe ratings by regression modelling. In addition, we displayed how the nutritional benefits for vegetarian and non-vegetarian recipes are distributed. Based on this dataset, we see that for an athlete, vegetarian meals give more nutritional benefits over non-vegetarian meals in terms of sodium, calories and proteins.

References

https://www.kaggle.com/hugodarwood/epirecipes

http://www.kdnuggets.com/

Source Codes

https://github.com/valuko/dm_project

https://github.com/valuko/data_cleaner