# Data Mining Project - Week 4 - Popular Dishes & Restaurant Recommendation

## Data Mining Specialization - Coursera / University of Illinois at Urbana-Champaign

- Author: Michael Onishi
- Date: November, 2019

## Introduction

The general goal of Tasks 4 and 5 is to leverage recognized dish names to further help people making dining decisions. Specifically, Task 4 is to mine popular dishes in a cuisine that are liked by people; this can be very useful for people who would be interested in trying a cuisine that they might not be familiar with. Task 5 is to recommend restaurants to people who would like to have a particular dish or a certain type of dishes. This is directly useful to help people choose where to dine.

For this task, I chose the Italian cuisine to explore. I used python with Pandas and Matplotlib for processing, filtering the text files and plotting.

The full notebook with some steps taken here, making it possible to reproduce this work, is available at:

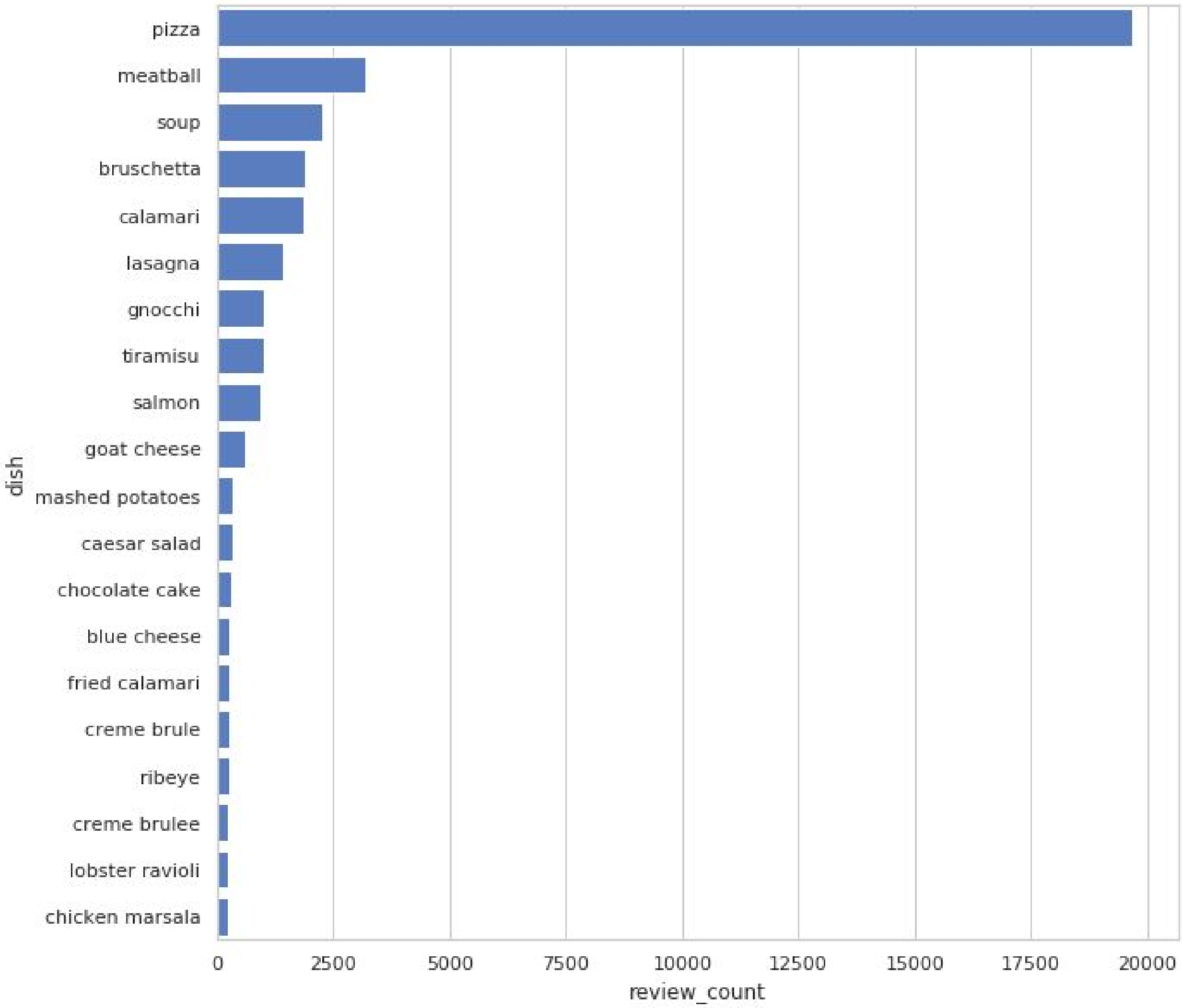
https://github.com/michaelonishi/coursera-data-mining-specialization/blob/master/c6-data-mining-project/task4-5/Data_Mining_Project_Week_4_Popular_Dishes_Restaurant_Recommendation.ipynb

## Task 4: Mining Popular Dishes

### In this task, you will create a visualization showing a ranking of the dishes for a Yelp cuisine of your choice.

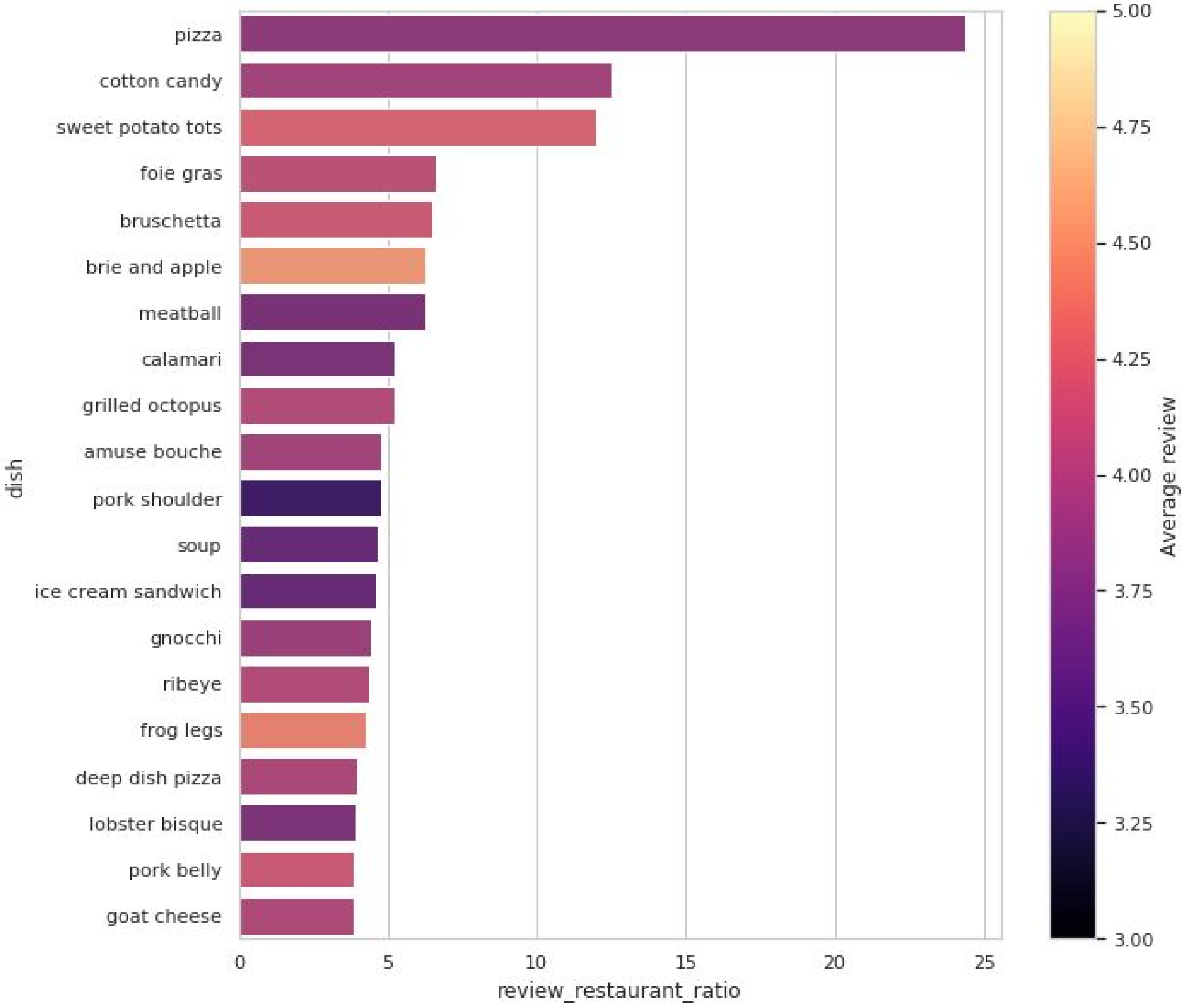
Here I used Italian dish names from task 3, compiled here:
https://raw.githubusercontent.com/michaelonishi/coursera-data-mining-specialization/master/c6-data-mining-project/task4-5/italian.txt

First, I plotted the simplest approach of counting the number of times each dish appeared on reviews, limiting to the top 20:



As we can see, it is not so helpful for recommendation, because the review mentions to a dish may be very high, but the average rating for it can be low. Also it lacks some kind of normalization, since the more restaurants serving the dish, the more reviews it is expected to have.

So the next approach was normalizing the number of dish reviews by the number of restaurants containing the dish review and showing the average review for each of the dishes with colors.

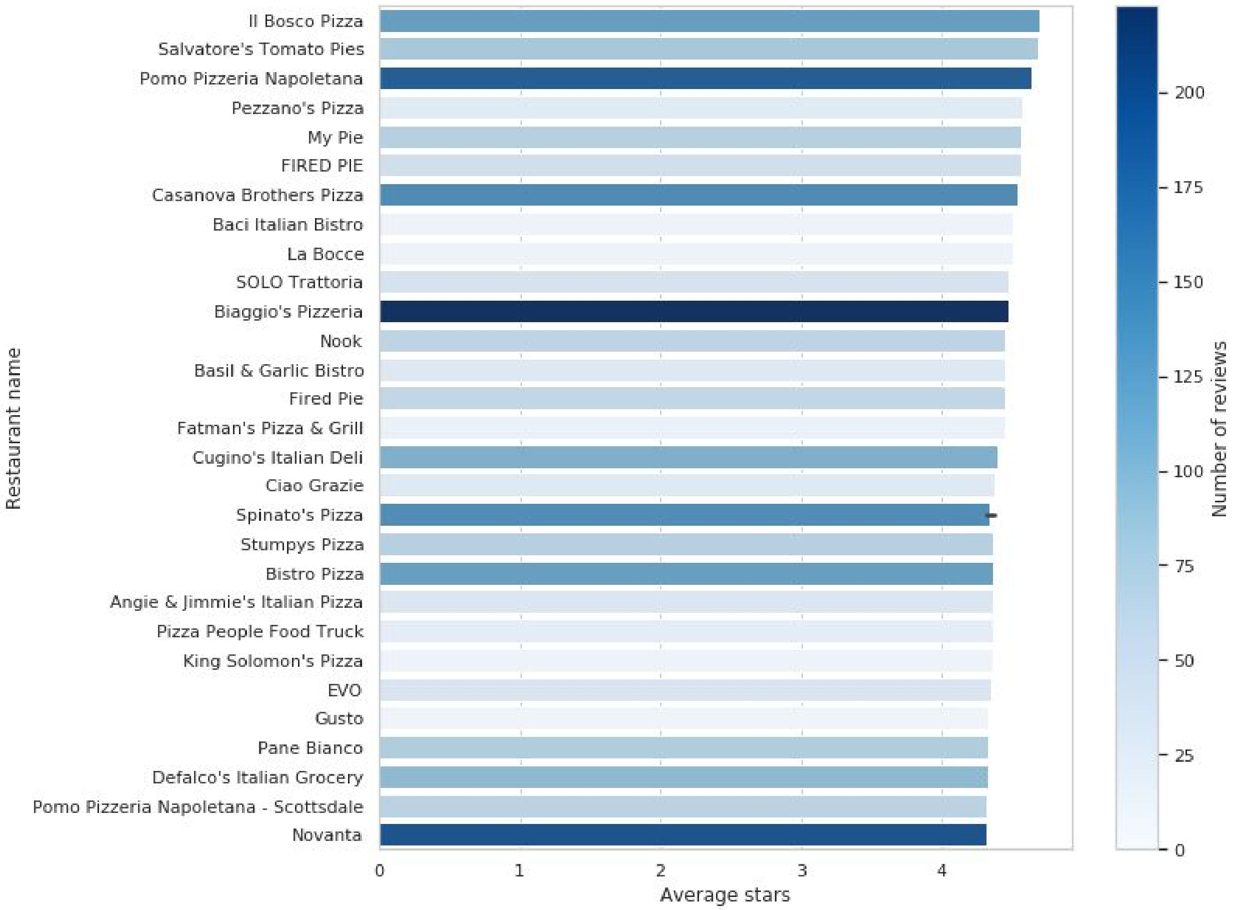


Now we can see that this visualization is much more interesting. Pizza continues to be on the top. Since it is one of the most popular italian dishes, it is expected. But its average review is not so good. I found very interesting that among that top 20 dishes, "brie and apple" and "frog legs" had the best average reviews. It is something I did not expect.

## Task 5: Restaurant Recommendation

In this task, your goal is to recommend good restaurants to those who would like to try one or more dishes in a cuisine.

Here I explored the pizza dish, since it is the most popular italian dish discovered. For ranking the best restaurants, I did the following: for each restaurant containing pizza reviews, I calculated the average review and review count (containing the word pizza). I discovered that some restaurants had only a few reviews, so it could not be consistent. So I only considered the restaurants with more than 10 reviews with pizza. Below I plotted a bar chart with the top 30 reviewed restaurants serving pizza with these filters applied. The colors represent the review count (darker colors represent more reviews).



Here we can see that in this list only a few recommended restaurants have a lot of reviews. But with a minimum of 10 reviews, I consider that all the restaurants in this list could easily be recommended to anyone seeking for a good pizza.

## Conclusion

These two tasks were very interesting because they were very related and could be easily adapted to some online recommender system with very good results even without applying any very sophisticated methods.