Task 4 Mining Popular Dishes Report

Objective

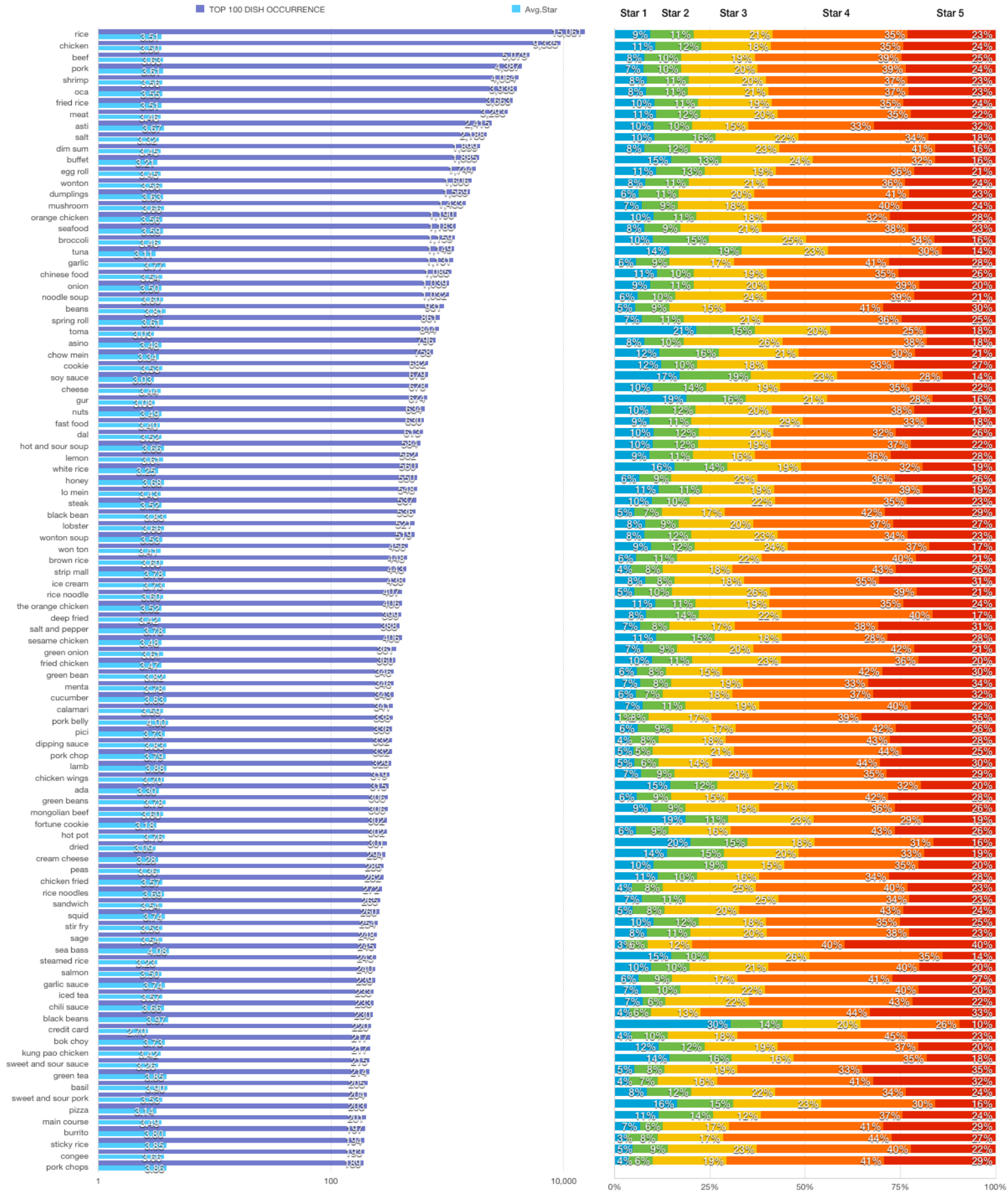
In this task, you will create a visualisation showing a ranking of the dishes for a Yelp cuisine of your choice. You may use the dish list we have provided, the list based on your annotations from Task 3 (or a subset of that list), or any other list for other cuisines. You might find it more interesting to work on a cuisine for which you can recognise many dishes than one with only a few dish names that you recognise.

Approaches

In the first place, I used the annotations provided based on my submission in task3. I did brute force search for all the labels in the original review and tips file to count the occurrence of each label. On top of that I sorted the label in descending order and get the top 100 dishes with most occurrence.

As mentioned in the instruction, simply go through the occurrence of each dish is not enough although occurrence could reflect the popularity of dish, therefore I used following statistics:

1. total occurrence of dish
2. the star distribution of each dish
3. average star of each dish



The occurrence indicates the times each dish is mentioned in review, however without identifying its source (from positive review or negative one) could be misleading since a dish could be mentioned many times for its distasteful. For

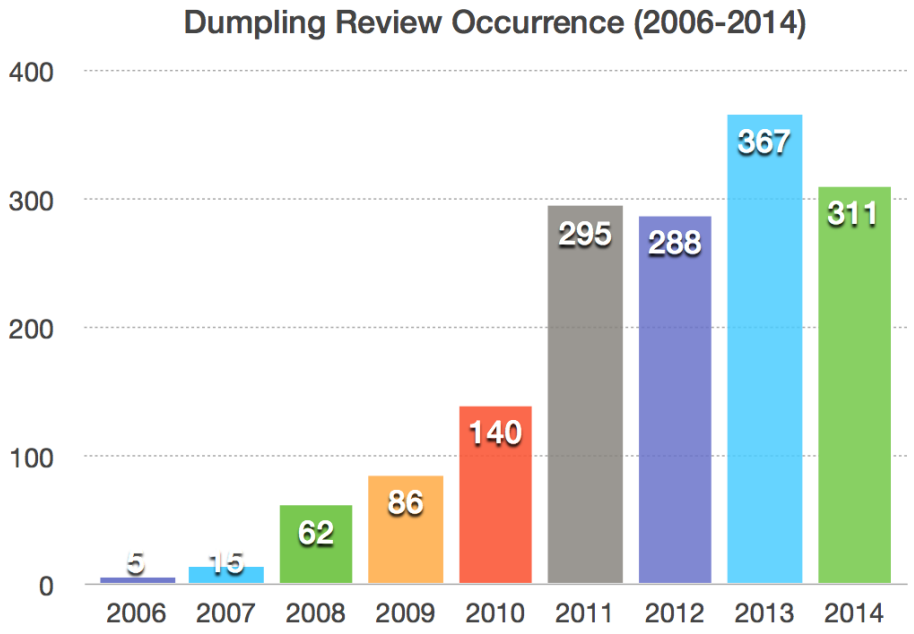
instance, “tuna” is mentioned 1149 times in reviews, but actually there are quite some negative feedbacks about it.

Therefore the average star could serve as another indicator to show the judgment of specific dish. Those dishes which has high average star should be of comparatively higher quality and popularity among all dishes.

Also it is interesting to see the star distribution for each dish, such kind of distribution is a good sign to find out the attitude of reviewers, for instance, a dish could be mentioned for many times with high average star, however the reality could be that many people held positive opinion while at the same time a certain group of people dislike it. The average star can be higher than average but the certain amount of negative review should also be considered.

Thus, customer could choose dishes according to above three criteria: occurrence, average star, and the star distribution for a comprehensive judgment.

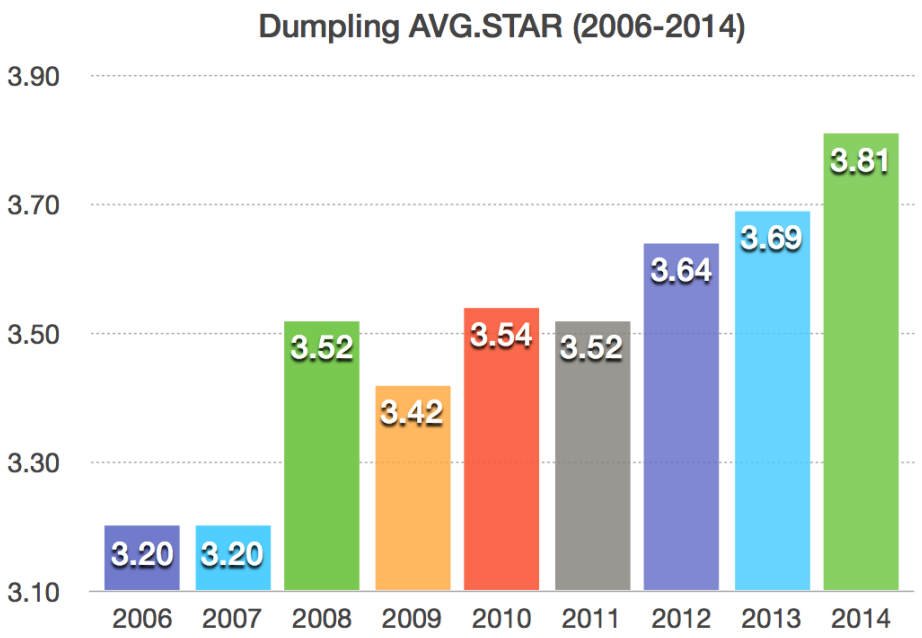
Specifically, I searched one dish “dumpling” for more details. I classified the reviews according to their date, then I collect all the reviews within each year to get the below graph.



From the graph we can find that dumpling related reviews are growing on yearly basis, which implies two possible reasons:

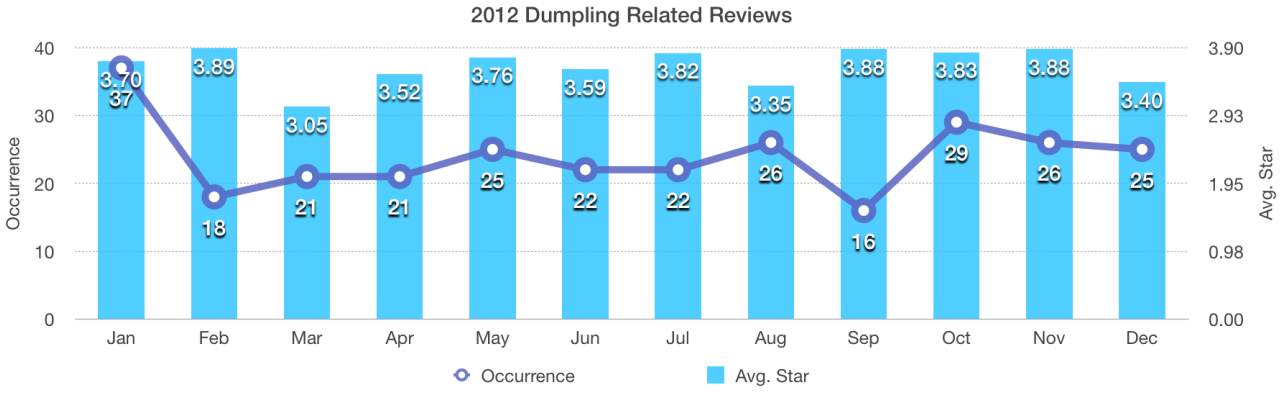
1. dumplings are gaining popularity among people, more and more people tried this dish and gave feedback
2. more and more people submit reviews on internet so that they can be found and used as reference for other people

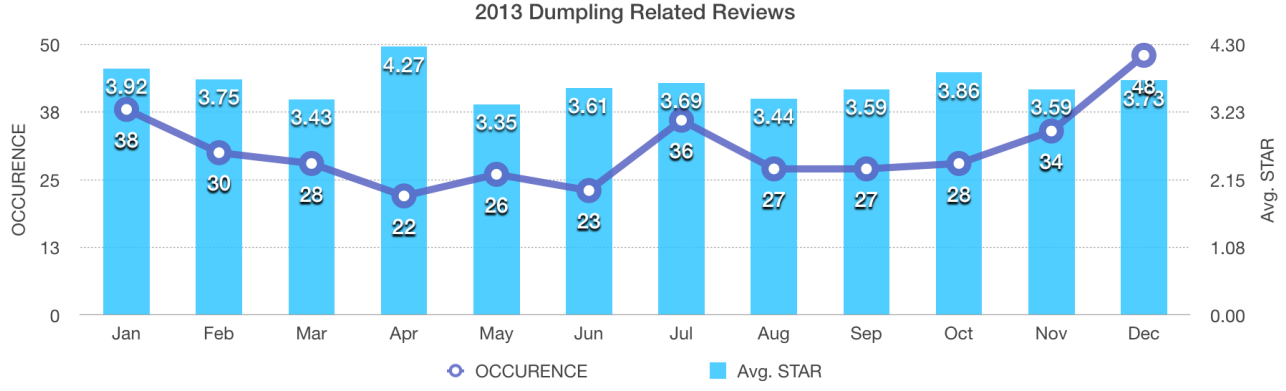
Another important factor is the average star, below are the average star people gave for dumpling during 2006 to 2014.

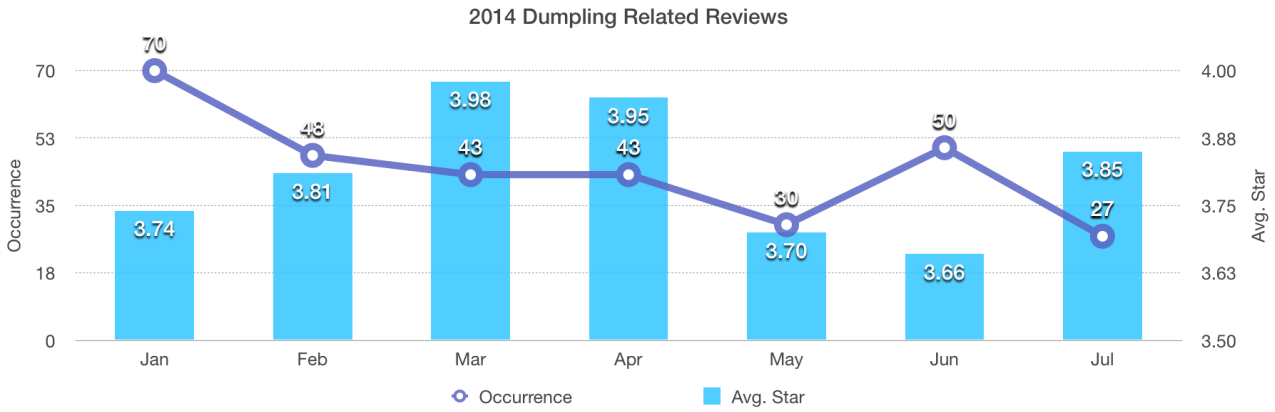


From above graph, we are sure that the assumption that dumpling is gaining popularity seems to be the case, according to the occurrence & average star, we can say that not only more and more people tried dumpling, but also gave positive feedback for this dish.

More specifically, during the year 2012, 2013 and 2014 (latest data from dataset), we can get the dumpling review occurrence and avg.star.







From above graphs, it is very clear that dumpling is gaining popularity from 2012 to 2014. Not only in terms of the monthly review counts but also the average star shows a rising trend.

Another interesting thing is that the review occurrence and avg.Star of second half year, especially winter (Nov, Dec, Jan), is higher than other seasons, it could be due to some Chinese related activities within the area. Another possible reason could be dumpling is popular during winter for its nutrition or other special effects.

Similarly, we can get such kind of statistics for other years, and analyse the trend regarding the popularity of dumpling.

Conclusion

From the review occurrence, star distribution we can analysis the overall reviewers' attitude toward specific dishes. From the review occurrence and average star we can infer the trend of dishes and the potential popularity for specific dish, which could serve as an indicator for future recommendation.