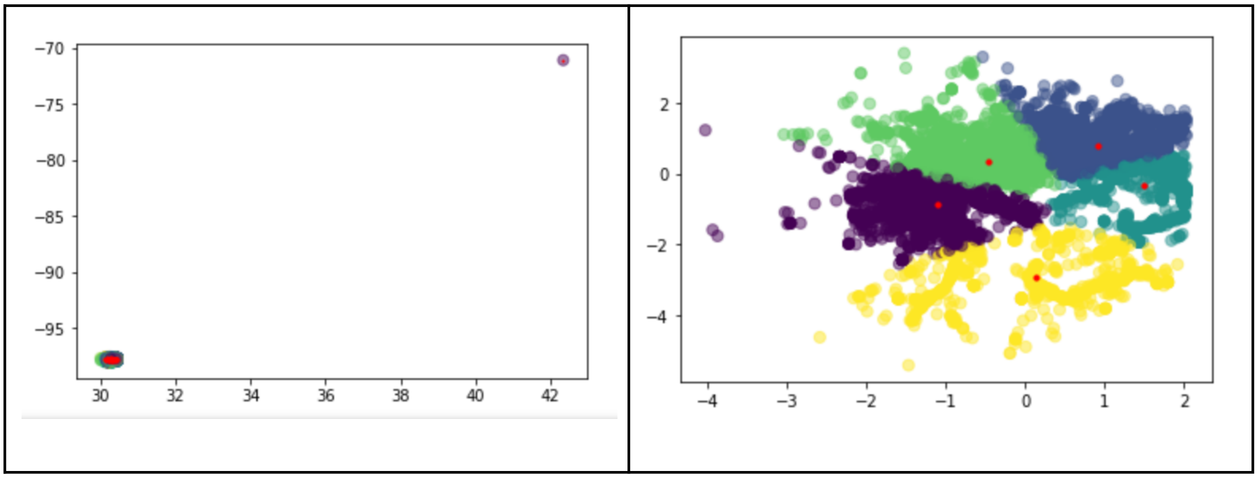
Project 2 Write-up

Vivian Gunawan

Task 1

The goal of this task is to compare the clustering of restaurants in Las Vegas based on different features. One clustering with longitude and latitude while the other with review texts. The first problem I ran into was the lack of businesses in Las Vegas due to the updated dataset. To solve this problem, I looked for the city with the most businesses, it happens to be Austin.

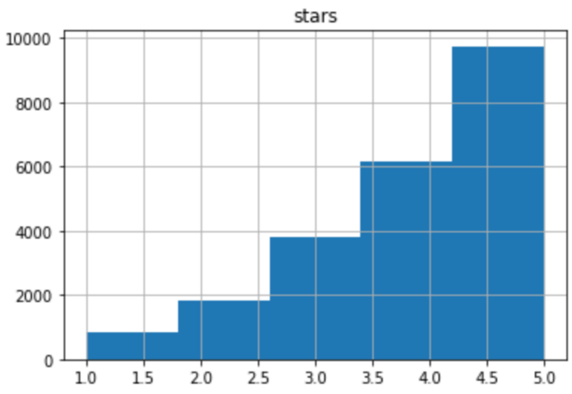
In this task, the clustering algorithm that I decided to use is the k-means clustering algorithm, it's simple and does the job. When clustering based on longitude and latitude, there happened to be an outlier in the data that caused a really odd clustering. When this outlier was removed, the clustering was a lot better. Below is the before and after plot of removing the outlier and normalizing the data.



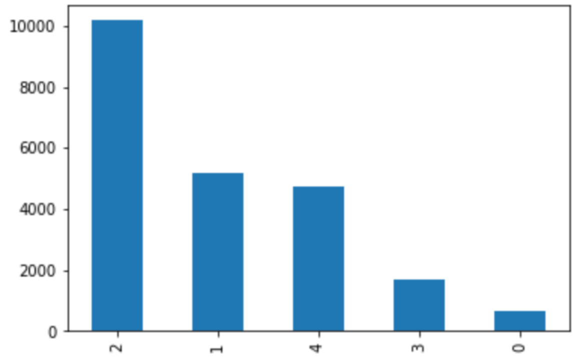
Classifying based on review text was not as straightforward, text data had to be converted into numerical data and there were multiple reviews for one business. The first thing I did was merged the review dataset down by concatenating all the reviews for each individual business. Then I proceed to preprocess these concatenated reviews, by removing non-english data. To convert these reviews to numerical data, I choose to use the Term Frequency-Inverse

Document measure. However I cannot just simply calculate this for each concatenated review, as the matrix would be too large and sparse. To deal with this, I used the same measure to first get the top 10 keywords for each business. Then I proceeded to use only these 10 keywords to get the measure among other businesses and clustered them. I was curious to see how my clustering did in comparison to the actual star rating of the businesses that customers gave, so below is a comparison of that. I think the clustering in terms of the distribution of businesses and their star rating is pretty close. However, the histogram and bar chart doesn't represent everything since it might not be the same business ids in the same cluster and rating group.

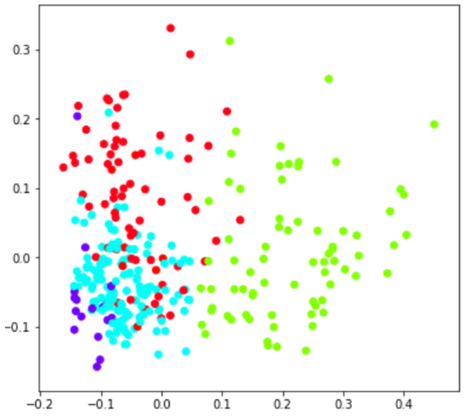
Actual star rating histograms



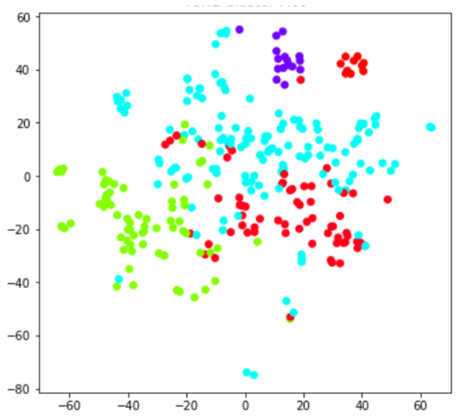
Clustering into 5 clusters bar chart



PCA 5 Cluster

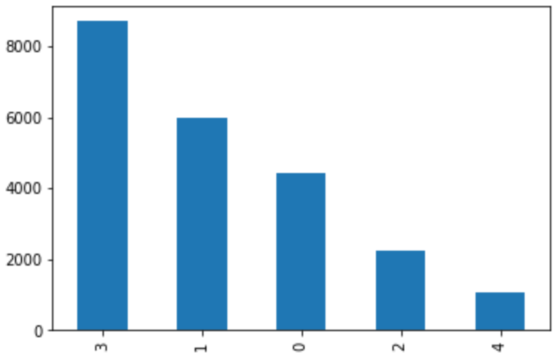


TSNE 5 Cluster

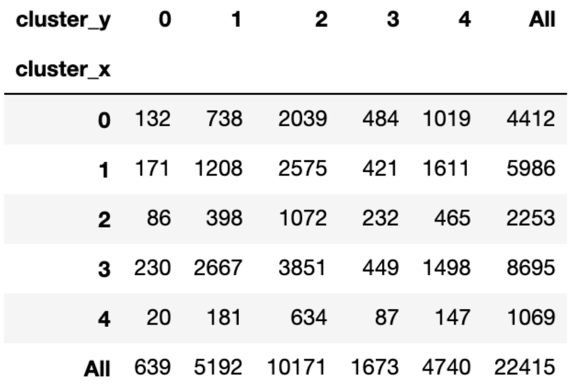


Below is how I compared the two clustering based on different features of the data. Based on the longitude and latitude clustering, most of the businesses are labeled as cluster 3 with 8695 of them. Based on the review text most of the business is labeled as cluster 2 with 10171 of them. However the number of businesses that are labeled 3 in the longitude and latitude cluster and lebled 2 in the review text cluster is only 3851. The way I interpret this is that restaurants with similar ratings are not necessarily close to each other in terms of location.

Clusterin into 5 cluster (longitude & latitude)



Crosstab between two different clustering



cluster_y	0	1	2	3	4	All
cluster_x						
0	132	738	2039	484	1019	4412
1	171	1208	2575	421	1611	5986
2	86	398	1072	232	465	2253
3	230	2667	3851	449	1498	8695
4	20	181	634	87	147	1069
All	639	5192	10171	1673	4740	22415

Task 2

For this task, we are supposed to classify a restaurant by looking at the scores of other restaurants that have the most similar review. This method fails, even with increasing the number of neighbours used during prediction. This could be due to the way TFIDF was used to vectorize the review or just because the reviews for similarly rated business are described very differently. When starting with k as the square root of the number of businesses, the accuracy achieved was only 52% with a 70:30 train test split.