Final Report

1. Problem Statement:

Yelp is commonly used in our life when we want to find or checkout if a place is good to go for food or drink. As a heavy Yelp user, it occurred to my friends and me that stars rating doesn't truly reflect whether a restaurant is good or not. We also need to consider the numbers of reviews, the average number of reviews and the contents of reviews to determine if it is a good restaurant to try out. However, it would take us more than few minutes, even an hour, to find good places and decide where to go. When we don't have enough time, we often just pick a 4-5 stars rating restaurant, and often turns out not as good as we expect.

2. Description of Dataset:

2.1. General Description:

2.1.1. **Yelp dataset from Kaggle**:

A dataset provided by Yelp with businesses data that are in and outside of US. The dataset contains 188593 entries and 15 columns.

2.1.2. **Population data from Government Census:**

A datast contains 81546 entries and 20 columns. It includes estimated population from 2010 to 2017.

2.1.3. **Zipcode data from an online database:**

A dataset maintained by non-government but sources are mostly from government sites. It contains 81831 entries and 20 columns. The entries are not only in US but also in the territories and military bases outside of US.

2.2. Data Wrangling Process:

2.2.1. **Initial Data Handling:**

All three dataset are imported as DataFrame for initial examination. Next, I extracted the columns that are needed and renamed them for easier understanding and future merging purpose for each dataset. Afterwards, I used drop_duplicates() function to remove redundant rows. Last, I removed all the records that are not in US from Yelp and zip code data. Besides these records, I also removed decommissioned zip codes from the zip code data.

2.2.2. Main Process:

2.2.2.1. **Yelp Data:**

I constructed a list with categories that I think are food related, like "Food", "Bars", "Bakeries", "Bakeries", etc. I then used the list to filter out non food business related records. A helper function is_empty was created to help find empty string in all string columns. Next, I merged Yelp data with zip code data to find missing or incorrect state name. In addition, latitude and

longitude columns are used to help fill in missing city and state names; correct city, state and zip code value; and verify records. Before merging with census data, I unified many cities name for better merging result. For example, any city name with Mt. or Mt are changed to Mount and N or N. are changed to North.

2.2.2.2. Population Data:

A lookup table consists of states full name and their abbreviation was created to help merging. I removed words from the census data, like “Borough”, “Town” and “City”, and leave the area name only to match as many records. This word-removal process was gradually done multiple times to achieve the best merging result.

2.2.3. Final Process:

Some manual validations and corrections were required after merging all three datasets. During the process, I used a website that can convert latitude and longitude into actual address as a helper. The data correction are mainly done in two ways. One is using business_id where only single record or only few other ones with the same correct value, like business_id '110iMPMPEEjFlf8HKVq84g' and '1jdE-PeiQHvL8165vebWrw' both has incorrect city name that are supposed to be “Charlotte”; or the city name exists but in another state, like “Pittsburg” exists in California but with state is PA the city name is actually “Pittsburgh”. Besides city name, this method is also applied to change incorrect zip code and state name. The other way is changing city name in the case where it is misspelled and doesn’t exist in another state; or in the case where a zip code has multiple area names in the census dataset, thus, I changed the city name to match. Last, I removed the ones that I was unable to find population number from the result dataset.

2.3. Summary:

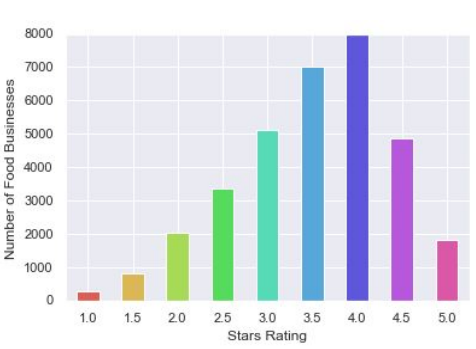
After examining original datasets and data wrangling, the result dataset contains 33289 entries and 10 columns with only food related businesses.

3. Exploratory Data Analysis:

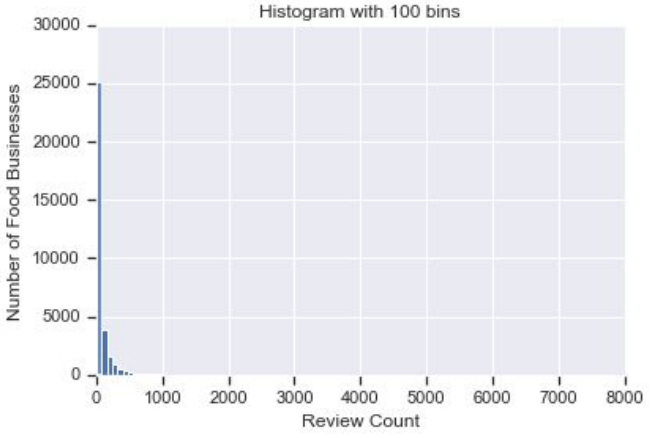
3.1. Initial Exploration:

After the food_business dataset is loaded, I started with .info() and .describe() to observe the general information on the data. There are 3 numeric columns and 6 string columns. For my project, stars rating, review count, and population are important variables. I decided to visualize the distribution on these three variables.

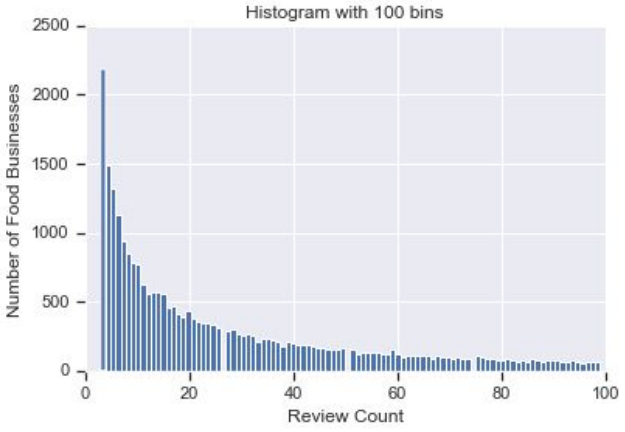
From the distribution, percentiles, and mean of the stars rating, we can see most ratings falls between 3.0 to 4.0 stars.



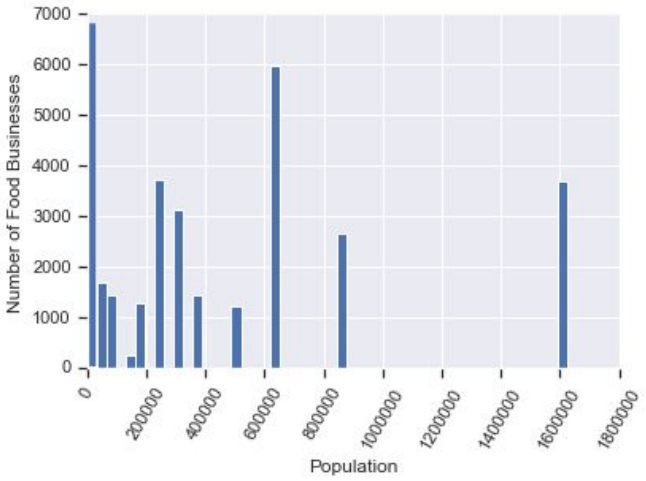
The numeric and visualized general review count distribution both showing there are many outliers and most of the counts fall under 100. The result graph is shown below.



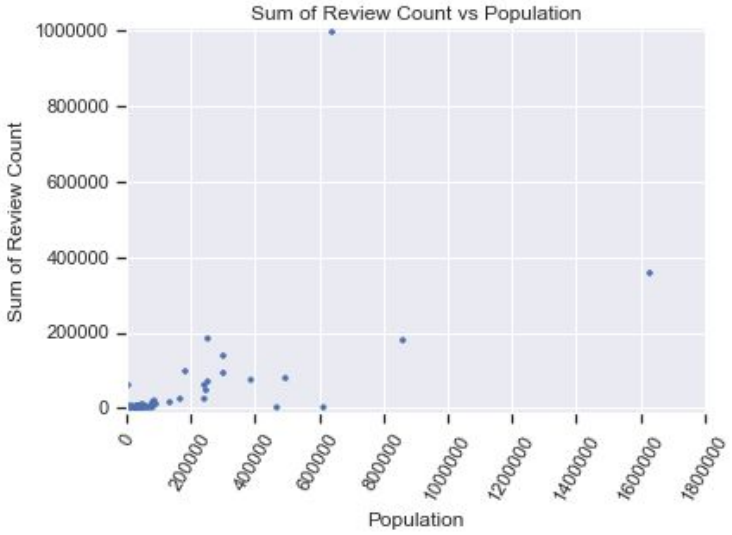
Below is a “zoom in” version of the review counts that are under 100. The result shows most review counts fall under 40 counts.



The distribution of population overthrows my initial assumption that more population means more food businesses. The graph shows no correlation between the number of food businesses and population.



Since I plan on using population to determine the review counts range, I plot population and review count to see if there's relation between the two variables. I plotted 4 graphs, all with population as x axis and review count as y axis. At first, the graphs don't seem to show any correlation between the two, until I plot the sum of review counts with population. The graph is roughly showing more population does end up getting more review counts in total.



After researching the relation between population and review counts, I suspect that whether the city is a tourist place could also be a key factor. Tourists don't count in the population but they do leave reviews. Thus, I observed the relation between review count and states by plotting the distribution of state. I found there is only one food business in KY, MD and TX in the dataset, which may cause my dataset to be biased. I also listed out the cities in top 5 and bottom 5 review counts and population respectively. Comparing the results, I found it interesting that there are 4 cities in both top 5 list, but only one in both bottom 5 list.

3.2. Deeper Exploration:

The scatter plot of review count and population doesn't show a strong relation

between the two variables. Thus, I computed their Pearson correlation value, and got 0.098716, which indicating there exists a positive correlation. To confirm it wasn't happened by chance, I applied bootstrap permutation test on the Pearson correlation coefficient of the two variables. The hypothesis is set as the following,

H_0: There's no correlation between population and review count.

H_A: There is a correlation between population and review count.

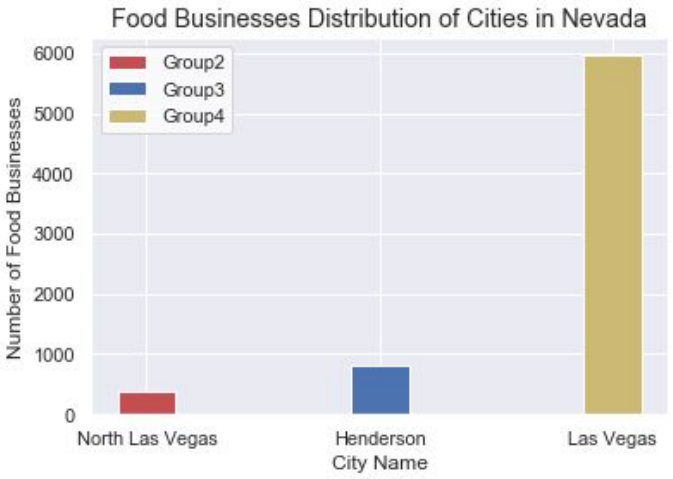
The p-value from the permutation test is very close to 0 that it's small enough to reject the null hypothesis. Therefore, we can conclude that statistically there's a positive correlation between population and review count.

As mentioned earlier, stars rating, review count and population are important variables in this project. I used normaltest from Scipy stats package to test if they are normal. The results show that only review count and population are normal.

I also explored further on population by splitting the dataset into five groups, small to large respectively, since I plan on using population to determine the optimal review counts range that predict whether a food business is good. I calculated mean, standard deviation and median of the five groups. Below is the results of each group,

	# of Records	Mean	Standard Deviation	Median
Group One	6757	35.74	58.36	16
Group Two	7668	75.79	134.88	28
Group Three	6551	70.03	118.09	27
Group Four	5966	167.53	376.34	47
Group Five	6347	84.29	156.47	29

From the result above, we can see Group Four has a significantly higher number than the other groups in mean, standard deviation and median. To find what might cause this, I researched the state and city distribution in the Group Four. Specifically targeting the data whose review count is larger than the mean. All of them are in Las Vegas, NV. Intrigued by the finding, I looked into all NV data in the whole dataset.

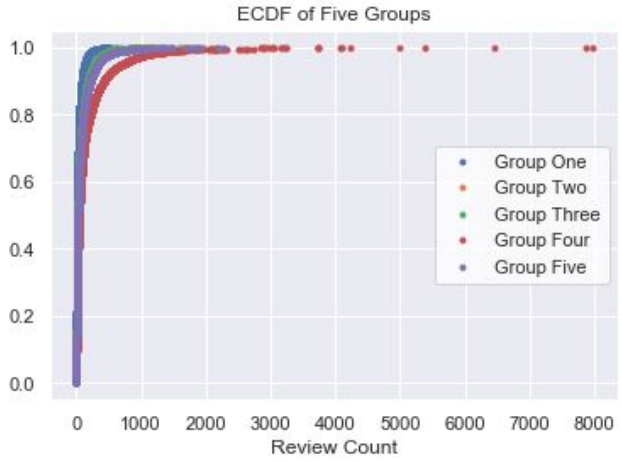


There are only three cities from NV in the whole dataset, which are North Las Vegas, Henderson, and Las Vegas. The population are 376, 808, and 5965.

3.3. Conclusion:

There seems no correlation between number of food businesses and population from the distribution of population graph, which might due to lack of spaces to open businesses. High population doesn't necessarily mean bigger land. However, from the permutation test result, we can see there's a positive relationship between review count and population.

I also plotted ECDF graphs for all five groups to check their similarity on review count. The graph shows they have identical distribution.



4. Results and In-Depth Analysis using Machine Learning:

4.1. In-Depth Anaylsis:

As mentioned in the previous section, I realized touristy cities, like Las Vegas in this case, contain higher review counts with moderate population. This affects the exploratory result of the relations between review counts and population. The 5,965 records that in Las Vegas are removed to improve the accuracy of the model. The Pearson correlation coefficient was about 0.099, and 0.12 after the removal. I randomly selected 250 records from rest of the dataset, which has a size of 27,324.

Since a food businesses is good or not is quite subjective, there's no definite answer to this model. Due to this reason, I decided to use clustering, and make the conclusion based on results. I applied K-Means clustering with divisive hierarchical clustering approach, first by population and next by review count. find_best_k function was created to help find the best number of cluster that generate the best model. The range of k to test on is from 2 to 8. Silhouette score is used to help identify the best clustering model. I also tried Agglomerative clustering. However, the dendrogram was visually difficult to observe the relations. Moreover, I only need two hierarchies. Therefore, I did not proceed with this approach.

After couple runs of the codes, I found the best number of clusters is either 5 or 7 depends on the 250 random samples. The second clustering mostly fall on 2 sometimes 3 or 4. In order to get more clusters to observe on review count's distribution, I used 3 instead of 2 if the best K is 2. The skewness score of review count column is around 5.65, which means the review count data is highly skewed to the right. This indicates most of the review counts are at lower numbers. I organized statistic results into a dataframe to compared. In addition, I ranked population range and review count range of each cluster from small to big. As the statistics show, the smaller the review count range is, the more estimated good food businesses it contains.

		population_rank	review_cnt_rank	estimated_good_cnt
cluster_label	**cluster_label_2**			
0	**0.0**	2	1	4.0
	1.0	2	3	0.0
	2.0	2	2	1.0
1	**0.0**	7	2	1.0
	1.0	7	1	8.0
	2.0	7	3	1.0
2	**0.0**	6	1	9.0
	1.0	6	2	2.0
	2.0	6	3	0.0

3	**0.0**	1	2	5.0
	1.0	1	1	32.0
	2.0	1	3	1.0
4	**0.0**	4	2	3.0
	1.0	4	3	1.0
	2.0	4	1	3.0
5	**0.0**	5	2	2.0
	1.0	5	3	0.0
	2.0	5	1	3.0
6	**0.0**	3	2	4.0
	1.0	3	1	13.0
	2.0	3	3	2.0

Exploring more on how likely the clustering could determine the good businesses, I decided to sample another 250 records. I divided the dataset without Las Vegas into five groups using percentile at 20, 40, 60, 80 and 100. 50 records are randomly selected from each group. The initial assumption of good food businesses are defined as four stars and above in certain review count range based on population of the area. To find that review count range, I first applied logarithm to reduce the skewness of the review count data. Next, I computed the upper bound and lower bound by adding or subtracting one standard deviation from the mean, then “converted” the range back by applying exponential. The food businesses with four stars and above in this review count range are labeled as good. This process was repeated on the 50 samples in all five groups. The 50 samples from each group are combined after labeling as the second sample dataset, which has the same size of 250 as the first sample dataset used in clustering. I used the same clustering approach on the second sample dataset, and created a similar dataframe as above. From the dataframe, we can see statistically labeled good businesses are less than the ones from clustering. However, some actually has the same counts in both columns.

		population_rank	review_cnt_rank	good_count	estimated_good_cnt
cluster_label	**cluster_label_2**				
0	**0.0**	8	1	6.0	11.0
	1.0	8	3	0.0	1.0
	2.0	8	2	0.0	7.0
1	**0.0**	3	1	3.0	4.0
	1.0	3	3	0.0	3.0
	2.0	3	2	3.0	4.0
2	**0.0**	7	1	5.0	6.0
	1.0	7	2	1.0	4.0
	2.0	7	3	0.0	0.0
3	**0.0**	1	1	27.0	32.0
	1.0	1	2	1.0	6.0
	2.0	1	3	0.0	2.0
4	**0.0**	6	2	0.0	3.0
	1.0	6	3	0.0	1.0
	2.0	6	1	5.0	7.0
5	**0.0**	5	1	0.0	0.0
	1.0	5	3	0.0	1.0
	2.0	5	2	0.0	2.0
6	**0.0**	4	1	3.0	5.0

	1.0	4	3	0.0	0.0
	2.0	4	2	3.0	3.0
7	**0.0**	2	1	5.0	5.0
	1.0	2	3	0.0	1.0
	2.0	2	2	0.0	1.0

4.2. Results:

Everyone has a different standard of deciding if a food business is good. Some thinks four stars and above while others think only five stars are good. Because the standard is very subjective, the result of this model only tells if a food business is highly possible to be good compared to others in the area. The result shows the cluster groups with the smallest review count and the second smallest review count both can be reference for determining good food places. However, each has its pro and con. The groups with the smallest review count range contain more estimated good businesses but also contain the minimum review count. A food place with five stars but only single digit of review count may not be actually good. The groups with the second smallest review count range contain less estimated good businesses than the former one but the minimum review count is excluded.

Due to the removal of Las Vegas data, the result of the model may not applied in touristy cities. It's also important to keep in mind that this model aim to help people quickly choose a food business when visiting an unfamiliar area or want to try something new. The food businesses can be restaurants, bakeries, coffee shops, etc., that has a higher chance to be good compared to others in the same category. Other factors should also take into consideration, like reviews, for higher accuracy.

4.3. Future Work:

NLP can be applied on review details to analyze customers' sentiments or eliminates fake reviews to increase accuracy. The attributes may also be taken into consideration. For example, a place is hot and new and with four stars and above, even though with very low review count, it could be good.

Besides increasing the accuracy, the model can incorporate users' information to give more customized recommendations.