The analysis process starts with examining the relation between population and review count. It's essential to know if these two variables are correlated. If the result shows they are unrelated then I would have to find other variables to determine the accurate range of review count that indicate a four stars and above food business is actually good. During story telling of the dataset, the scatter plot shows there is a slightly positive correlation between population and review count. To further 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.

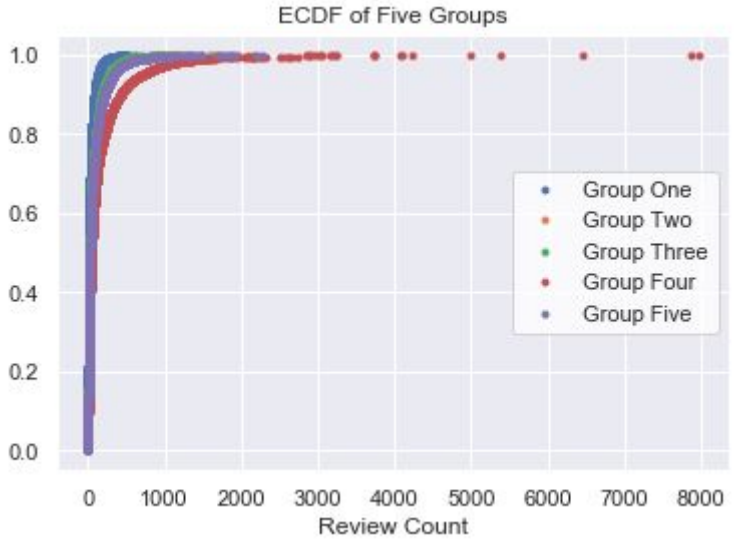
The Pearson correlation between population and review count is 0.098716. It indicates there's a positive correlation between the two variables. The p-value from the permutation test is very close to 0, which is small enough to reject the null hypothesis. Therefore, we can conclude that statistically there's a positive correlation between population and review count.

Stars, review count and population are three important variables in my model. I decided to test if they are normal. I used normaltest from Scipy stats package to perform the testing. The results show that only review count and population are normal.

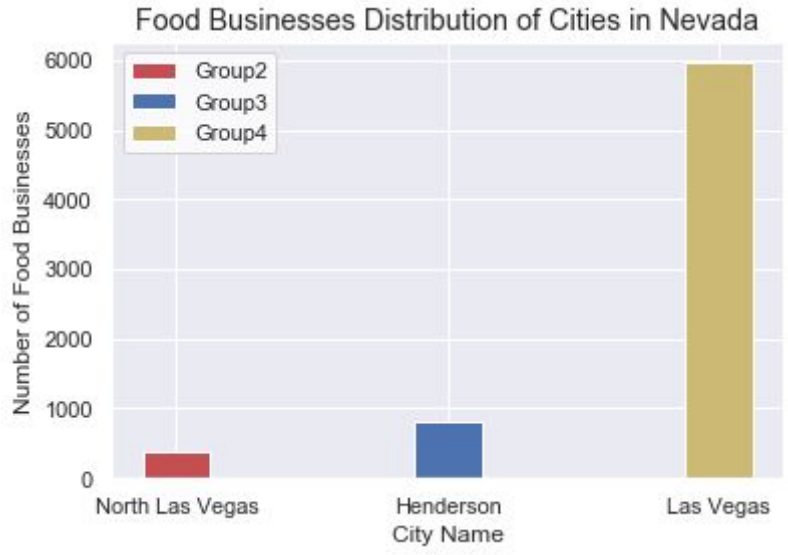
Since part of the design of this model is using different size group to predict if the food business is good, I explored more on population by splitting the dataset into five groups, small to large respectively. I used percentile to split the dataset instead of dividing by five. I calculated the mean and standard deviation of the review count for all five groups. The results show a peak at Group Four. This implies there could be extreme values exist. I then calculated the median of their review count. Interestingly, I discovered the Group Four also has a higher number than the other four 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

Before exploring Group Four data further, I plotted ECDF graphs for all five groups to check their similarity on review count. The graph shows they have identical distribution.



Next, 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. Based on city, I got the result below.



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.