During the assignment 1, we have analyzed some useful statistical properties and drawn the histogram for the stars rating across the different three cities. Thus, we will continue to explore the question about what factors correlate strongly with high or low stars rating. Yelp dataset is very important for the business to increase the profit. Each addition of a star to a Yelp business review yielded a certain increase in revenue, with the large effect on independent business. Thus, understanding the stars rating behavior for individual business should be a priority for any ambitious business owner. We will mainly focus on the restaurant as the one particular business to analyze and the Las Vegas is regarded as the main area.

Based on some common senses and yelp dataset, we concluded some variables which affect the stars rating, including Wi-Fi, parking, take out, price range, noise level, location, good for and open hours. The type of Wi-Fi, parking, take out are Boolean, which can be noted by 1 if yes, 0 otherwise. And the longitude and altitude can be used to determine the location of the restaurant. Price range and noise level can be converted to the integer based on the real data. Good for is to describe the restaurant is good at cooking which categories’ foods, such as dessert, breakfast, lunch, brunch, late night and dinner. Converting categorical features to numerical features is a very important step before analyzing the data. Maybe we will add or delete additional data and features in the later process.

Since we need to explore whether the chosen features affect the stars rating or not, it is necessary to measure the covariance between each feature and stars rating. And then measuring the correlation coefficient between these two random variables with expected values and standard deviations, which is a scalar that can have values in the range [-1,1]. To measure how good the linear regression of the chosen factors with respect to the stars rating is, we can analyze the residual. The variance of residual can be used to measure the coefficient of determination (square of the correlation coefficient) which can be used to explain how much the linear model of a factor with respect to the variance of stars rating. The random forest is one of the most popular models to classify the data with the useful attributions. In order to simplify the exploration, stars ratings between 1 and 3, noted by 0, otherwise 1. After splitting the train and test dataset, we will build the random forest model based on the chosen factors and then obtain the feature importance generated from different factors.