Midterm project pre-screening

Introduction

The love of people for delicious food has never decreased. Yelp, as an informative platform, gives us opportunity to view, learn, search and comment for restaurants or other business easily. As it becomes more and more popular as an application, the reviews and comments that customers leave on Yelp have larger and larger impact on other customers' choices. When we are looking at the list of restaurants, the first thing we will check is the "stars", scaled from 0 to 5, that one restaurant has. This does beg my initial question: what factors can influence the star number of a restaurant?

Data Choosing

I use the data called "business" from the SQL database for Yelp Data Challenge. It has over 156 thousand rows and 12 variables, which contains basic information for each business, such as its name, neighborhood, city and stars it has.

Thoughts and Plans

My first thought is to try to fit into a multilevel linear model and treat each state as every group. My dependent variable is stars, and my predictors are the state a business is located, the amount of reviews it received and whether it is still open, and try to find their correlations with the star it receives.

From the summary, I found that the number of reviews seems not a very good predictor, or in other word this predictor is not relevant with stars. So, my model needs more improvements, which is what I am intended to do in the future. Also, Before I fit another model, I will do some data cleaning, and then do some EDA, in order to find a better way to fit the model.