Project Proposal

Dataset: Yelp challenge data—business, hours, category.
(The business data contains 156639 observations of 12 variables. The hours data contains 734421 observations of 2 variables. The category data contains590290 observation of 2 variables)

Purpose: There are many Beauty and Spas businesses in United states. The purpose of my project is try to find out what factors will have impacts on the rating of the Spas shops. I would like to show these shops in the map to figure out the distribution about the shops. Will the location impact the rating of the shops? Are there have any difference between shops that have longer opening hours and have shorter opening hours? If the Beauty shops provide many different items or products such as nails spa, hair salon, medical spas and so on, will the shops get higher rating? What about the influence from the number of the reviews that those shops have?
There might have other factors that will affect the ratings, but the questions above may be the main questions that I want to figure out in my project.

Steps:

1 Read the data and clean the data by R.
2 Exploratory data analysis and find the variables.
3 Build a mixed effect model and interpret the results.
4 Model checking and the limitations.
5 Draw a conclusion.