STAT 605 Project Proposal

Xiangyu Wang xwang2439

Zijin Wang zwang2548

Jingshan Huang jhuang456

Yicen Liu liu943

Yuan Cao cao234

1. Introduction

1.1. Description of data

URL of data: `https://www.kaggle.com/yelp-dataset/yelp-dataset`.

Our data is Yelp dataset obtained from Kaggle. Yelp is founded in 2004 and it develops, hosts, and markets website and mobile app, which publish user-sourced reviews and ratings about businesses. This dataset is a subset of the businesses, reviews, and user data from Yelp. It contains 8,021,122 reviews, 209,393 businesses, 1,968,703 users and 1,320,761 tips from users and is stored in 4 separate files, respectively. Also, business attributes like hours, parking, availability, and ambience are included in business data.

We will only process the data of reviews and users as the data of tips and businesses are relatively small (about 150-250MB). However, we will also use these two small datasets during project as they contain more information related to other datasets so that we can do more analysis instead of only focusing on a single file.

1.2. Description of variables available

The review data contains 9 variables, the important ones include the ID of user and business, rating, review date and content of user, and the number of votes received for the review (whether it is useful, funny or cool to other users). The unique ID of user and business can help us map more information in user or business data.

The user data is a integration of user profile and his/her Yelp usage behavior, containing 22 variables. However, we will only consider some for the relevance to our problem. For example, user ID will lead to a mapping to review file just as mentioned above, the number of fans, received votes and average rating of user will be a reference to classify users into different groups and perform test for inference about whether there are some differences between different groups.

2. Statistical questions

2.1. Reviews

As described above the review data has two numerical variables which are stars and date. We are curious about if people have different usage habits when time goes by. What's more, we want to know if people's criteria are the same in different years. Basically, our first analysis is to test whether the average stars people give are the same between year 2016 and 2017. If people have the same judgement criteria, then we can use much longer time period data to form a more accurate analysis on a specific category. We think that t-test will preform a reasonable result.

2.2. Users

Our user dataset has a specific category ”fans”. As we all know, for these days celebrities can make a huge impact on the business market. Naturally, we come out a question which is whether people who have a large amount of fans tend to give more(or less) stars compared with people who only have a few fans. Visually, we will firstly form a bar-plot of number of fans to select those “celebrities” and amateurs. Then we could use t-test to measure the difference of average-stars that those two kinds of users give. The results will directly answer our question.

3. Code snippet that reads the data

Since our raw data are all formed as json files (each line is a json object), it is easy to read our data line by line with shell pipeline script (*code: cat review_01 | while read line; do...; done*).

4. Statistical methods

4.1. T-test

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. Since t-test is very useful when we want to compare the average values of the two data sets, we plan to use it to compare the average stars of hotels from different groups.

4.2. ANOVA

Since there are many factorial variables in our data set, we plan to use ANOVA to analyse the their effects on the hotel stars.

4.3. Linear regression

We plan to use linear regression to fit a model and figure out how the different variables are going to affect the stars of hotels.

5. Computation

First, in data processing, CHTC or slurm with implement of shell scripts is suitable for our data. We will use it to split large file into small ones and extract useful information with regular expression.

Depends on the size of what we get from the first step, we may consider to use Python, R or shell to draw some plots and do statistical computations.