CptS 451 DBMS

Charles Nguyen

011606177

Project Milestone #2

Proposal for Analysis of Yelp Business Data

In order to determine whether a business is *popular* or *successful*, I plan to classify the businesses into three categories:

- all businesses (implicit)
- popular businesses
- successful businesses

The implicit category represents the entire set of business data as originally loaded into the database. The other two metrics are proper subsets of this data.

In order to be classified as *popular*, businesses are required to have high count in checkins. I believe this is a reliable metric for how well the business is performing. We can also further make sure that the business is online, as opposed to formerly popular and now closed since this type might not be relevant to our analysis.

On the other hand, businesses are determined to be *successful* only if they satisfy both the conditions that they should have high count in checkins and high value for mean/median rating. A successful business implies that it is popular. Therefore, it is a reasonable assumption that it should have high checkin volume. Additionally, a successful business implies consistent customer's satisfaction, which is reflected in high scores in rating/feedback. At the moment, I plan to rely on the star rating to extract rudimentary information. However, it is highly possible to use the review text for more granular analysis in the future.

In order to determine the validty of these two qualities, it may also be necessary to *compute the variances* for both checkin volumes and ratings.

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

For more granularity, it is possible to focus on the popularity and success metrics for the business in each state and city. This might be more relevant to my analysis when considering the local economic landscape. In these cases, we can use the sample variance:

$$s^2 = \frac{\sum (X - \bar{x})^2}{n - 1}$$

Due to time constraint and limitation on man power, I have decided to focus on using the data sets pertaining to business, review and checkin volumes in order to investigate whether some businesses are more popular and successful than other.

At the moment, the user data set does not seem to be immediately relevant to the analysis. However, that may change as the project moves further into completion, and/or there is a need to incorporate the user data to enrich the analysis.