STAT 579 Final Project

Chelsey Legacy, Lindong Zhou, Evan Pete Walsh*

November 25, 2014

Abstract

Analysis of the Yelp Academic Dataset.

*Statistics graduate students at Iowa State University of Science and Technology

Contents

1 Introduction

Lindong

2 Data

The Yelp Academic Dataset[1] provides data enthusiants with the exciting opportunity to explore an incredible collection of information regarding the characteristics and quality of hundreds of businesses across the United States, Canada, and the UK. Specifically, the data includes details and reviews on 250 of the closest businesses to 30 large universities, including the Arizona State, UNLV, the University of Edinburgh, the University of Wisconsin, and the University of Waterloo, to name a few. The raw data is in json format and contains five different types of json objects: **Business**, **Review**, **User**, **Check-in**, and **Tip**.

Each **Review** object represents an individual user-based review of a particular business. The unique encrypted business ID is given along with the date of the review, the number of stars (out of 5) that were awarded, the number and type of votes that the review received, and an optional text description provided by the user. **User** objects are unique to every person that has an active Yelp account. Each user has a name, a unique encrypted user ID, the number of votes they have cast, the average number of stars they have given, and the date they signed up for Yelp, among other things. A **Check-in** object represents the count and time of all the registered check-ins for a particular business, and **Tip** objects represent a tip given by user for a particular business. Tips include the user's ID, the business's ID, the date, and the message that the user gave. While these objects all provide a rich source of information, for the scope of this paper we will only be examing the **Business** objects.

Business objects are unique to business ID's, and include the following information:

- the name of the business,
- the name of neighborhood in which the business is located,
- the city in which the business is located,
- the full address of the business,
- the exact latidude and longitude coordinates of the business,
- the average number of stars awarded to the business,
- the number of reviews received by the business,
- whether or not the business is still open,
- the hours that the business is open,
- the categories that the business falls under,

[1] *https://www.yelp.com/academic_dataset*

- and a number of different attributes which mostly concern restaurants and bars, such as whether or not smoking is allowed and the price range of the food.

To work with the data, we converted the set of all **Business** objects to a csv file in which the columns are variable names representing each aspect of a **Business** object, and each row corresponds to a unique business. Each different type of attribute was converted to its own character, numerical, or logical variable depending on what was appropriate. For example, the attribute **smoking** was converted to a logical variable where the value is "true" if the smoking is allowed at the establishment, and "false" if it is not allowed, while the attribute **price range** was converted to a numerical variable which ranges from 1 to 4.

3 Restaurants

3.1 Breakfast

3.2 Lunch

3.3 Dinner

4 Bars

Chelsey

5 Hotels

Lindong

6 Fitness

Lindong

7 Conclusion

Chelsey