STAT 511 Final Project (Initial Sketch)

Jim Curro, Eric Hare, Alex Shum

Apr. 15, 2013

Introduction

This is a sketch of our results thus far. We have made progress on our two biggest goals with the dataset. First, we've conducted an EDA on the data and discovered some interesting things, which is summarised in section two. Second, we've used a Term Document Matrix to condense the text of the reviews into a manageable format. We also have fit an initial SVM to the data in order to predict which reviews have at least one useful vote. These results are summarised in section three.

Exploratory Data Analysis

Figure X displays the number of useful votes against the number of cool votes in blue, and the number of useful votes against the number of funny votes in red. It can be seen that while both cool and funny votes are great predictors of useful votes, the slope is larger for funny. In other words, we would expect that the reviews with a set number of funny votes would tend to have more useful votes than those with the same set number of cool votes.

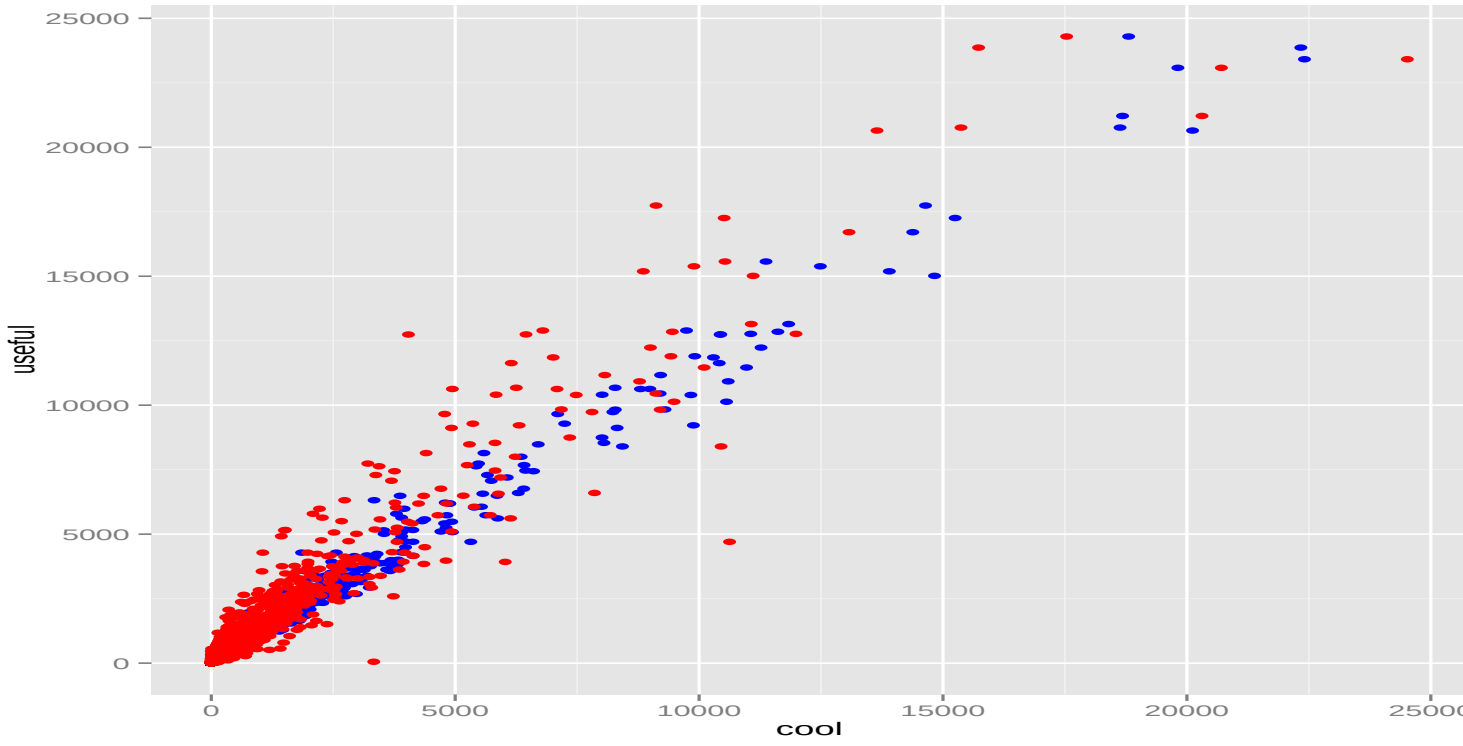
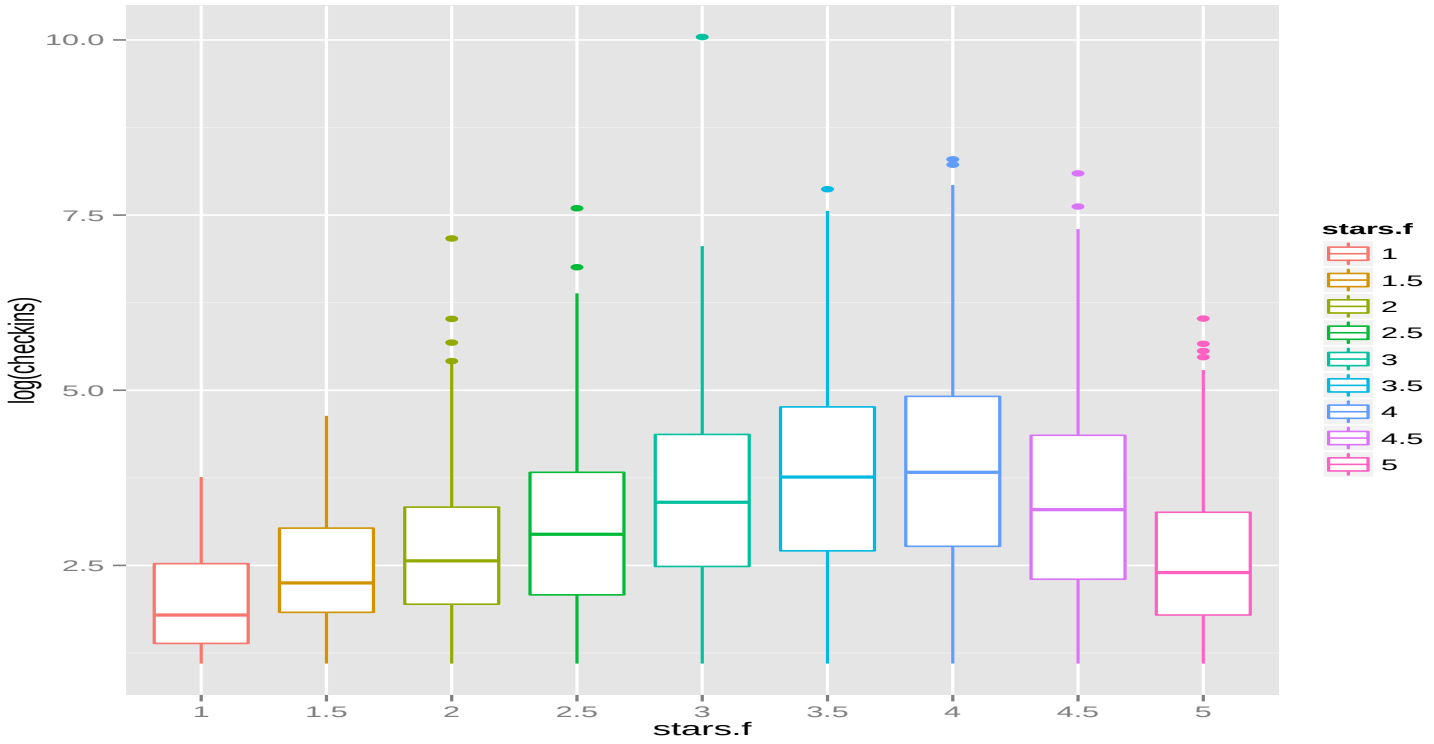


Table X shows the top 10 cities by number of checkins in the database. It also shows the total number of reviews for all businesses in that city, as well as the percentage of reviews over checkins. It can be seen that for the biggest cities in the database, there is a very uniform number of about .27 reviews per checkin.

city	reviews	checkins	percentage
Phoenix	91453.00	310640.00	0.29
Scottsdale	49042.00	202946.00	0.24
Tempe	27021.00	93446.00	0.29
Chandler	13944.00	48067.00	0.29
Mesa	9786.00	34328.00	0.29
Glendale	7017.00	26306.00	0.27
Gilbert	5803.00	22723.00	0.26
Peoria	2599.00	11361.00	0.23
Surprise	1274.00	5109.00	0.25
Avondale	1243.00	5053.00	0.25

Figure X shows boxplots of the number of checkins to each business by the average star rating of that business. It can be observed that businesses with about a four star rating tend to have the most checkins, but five star rating businesses actually have fewer. This may be either due to the fact that businesses with such a high rating have fewer total reviews, or they are more expensive and less likely to have a high number of customers.



TDM