

AC209a Milestone 5: Data Science with User Ratings and Reviews

Andrew Ross, Sophie Hilgard, Reiko Nishihara, Nick Hoernle

November 27, 2016

Proposal of Future Work

For Baseline Models in Milestone 4, we have implemented simple averaging as well as basic content-based recommendations and collaborative filtering models. Going forward, we hope to build more sophisticated models within these classes. In particular, we will explore creating a hybrid collaborative filtering/matrix factorization model or a model which more heavily weights recent ratings.

We also built a network-based recommendation system based roughly on the PageRank algorithm. This model finds the most ‘connected’ business, where path lengths between businesses are weighted inversely to how much a user liked both of the restaurants. To find the next recommendation for any given user, we search the graph built off of similar users and return the ‘closest’ business the user has not yet been to.

We have also explored a number of other models for predicting business success rather than focusing only on recommendation systems. In particular, we have spent time focusing on other aspects within the yelp challenge that ask you to predict business success based on location or investigate seasonal and otherwise temporal trends in business popularity. Location based models have focused largely on clustering methods, as neighborhood is a better predictor than absolute latitude and longitude values. Trend analysis will use regression models over time to analyze if businesses are cyclically popular at any given lag or to find points in time when a given business was comparatively ‘hot’ within its lifetime.

Lastly, we have been aiming to determine how location and seasonal trends alone impact the reviewers mindset (and consequently the reviews that they give to a business). To do this, we have used a time series analysis visualization (shown in figure 1) to highlight any repetitive trends that may be present in the data. For example the plot below shows the ratings that are above the mean in red and the ratings that are below the mean in blue. The average number of ratings per day of the week is shown in the right side vertical axis and the number of ratings on any one day of the year is shown in the axis above the plot. We clearly see that the reviews that are given on Sundays and Mondays tends to be more positive than the reviews given on the other days of the week.

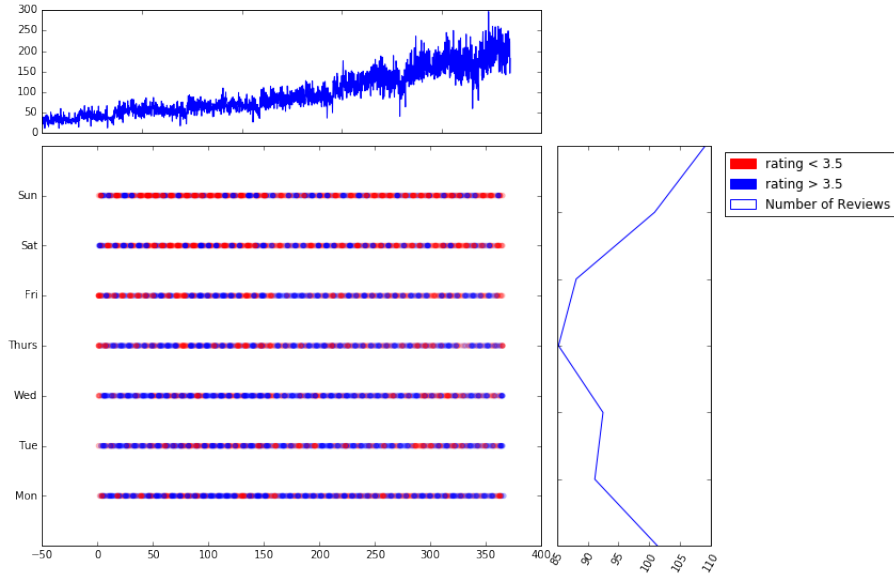


Figure 1: Plot ratings above and below the mean for days of the week vs days in a year for restaurants of categories bar and nightlife in Phoenix

A similar analysis can be applied to various sub locations within the greater Phoenix area. Figure 2 shows the locations of businesses that are tagged to three cities within the Phoenix area. We can compare the reviews that these locations are receiving over time, as shown in Figures 3, 4 and 5.

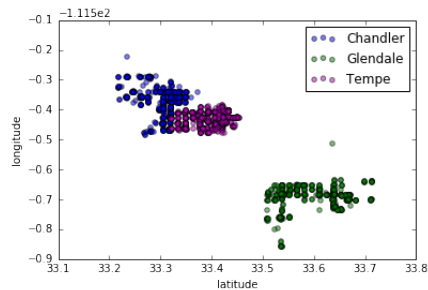


Figure 2: Plot of location of reviews tagged to three cities in Phoenix

For these different locations we see that to a greater or lesser extent, Sunday reviews receive the highest scores. We also see (focusing on the right hand line plot) that Chandler and Tempe receive roughly 5 and 10 reviews per day more than Glendale. The plot also uses the opacity of the dot to indicate numbers of reviews and there appears to be a trend that in the winter months (days of the year 200 through to 50), the Glendale area is particularly quiet (especially when comparing this to the other cities). Finally, we notice that Glendale appears to be a satellite city within Phoenix could be a possible reason for this.

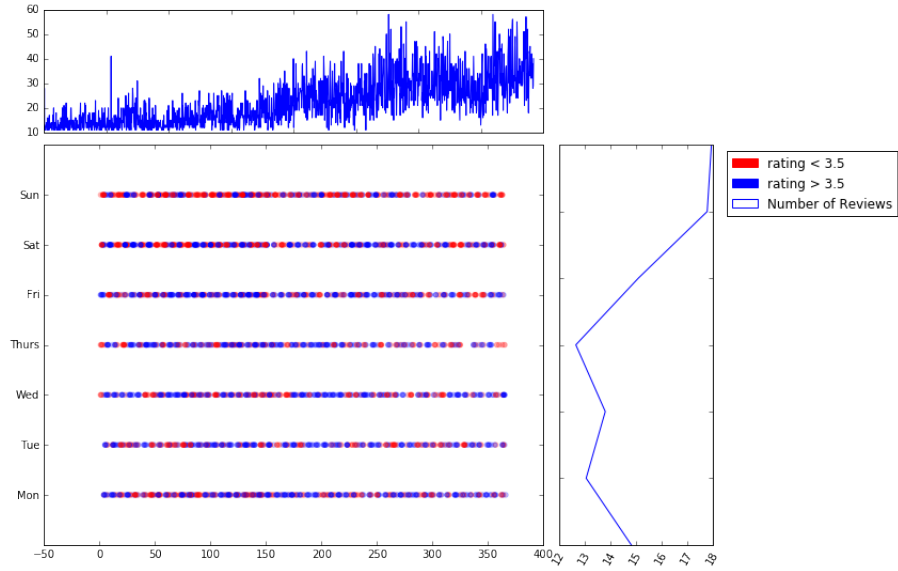


Figure 3: Plot of reviews above and below the mean for Restaurants in Chandler

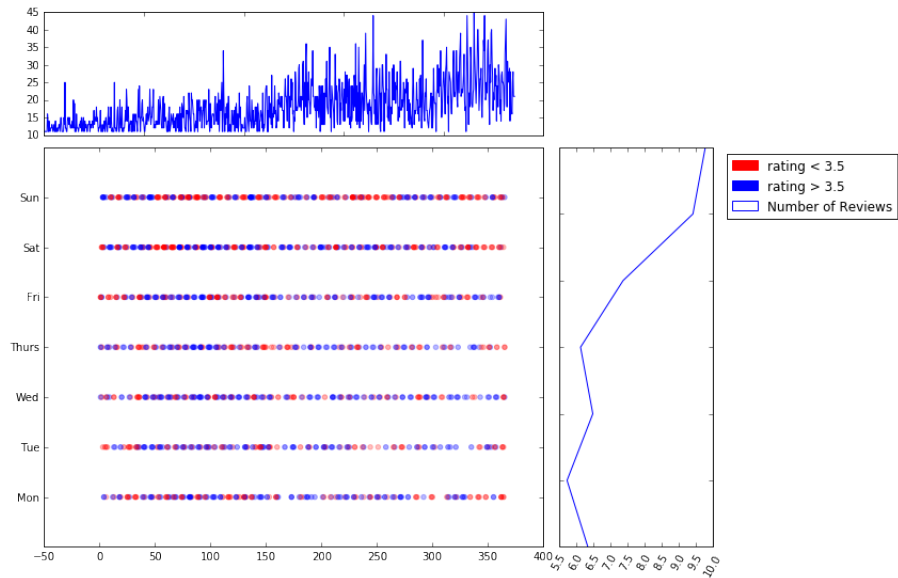


Figure 4: Plot of reviews above and below the mean for Restaurants in Glendale

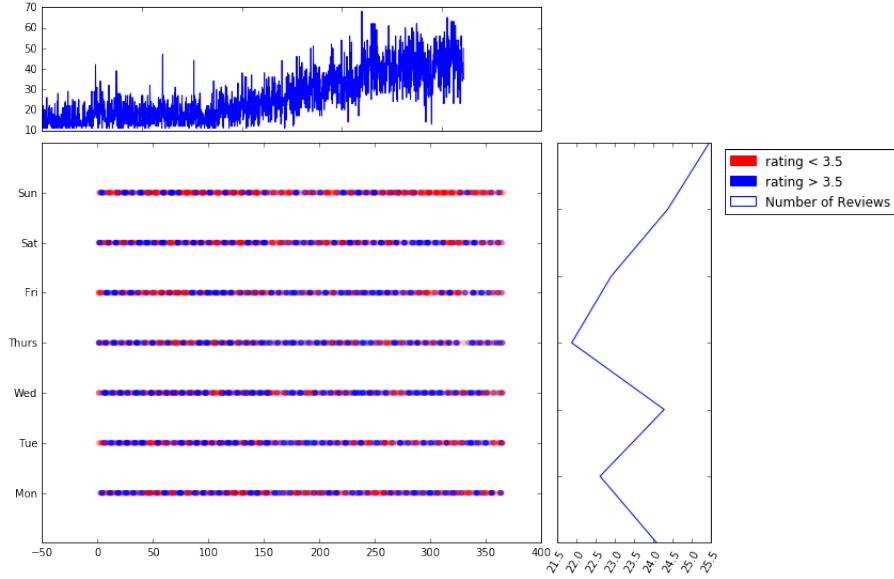


Figure 5: Plot of reviews above and below the mean for Restaurants in Tempe

We note that this location based study is likely to give us some insight into any seasonal and location based trends that are occurring within the restaurant businesses within the Phoenix area. Whether this information will only provide more qualitative information or if it will help us to refine our models by providing additional prior knowledge is to be explored further. Some additional source for future work is to focus on these trends and attempt to track specific categories of business as they become more or less popular.