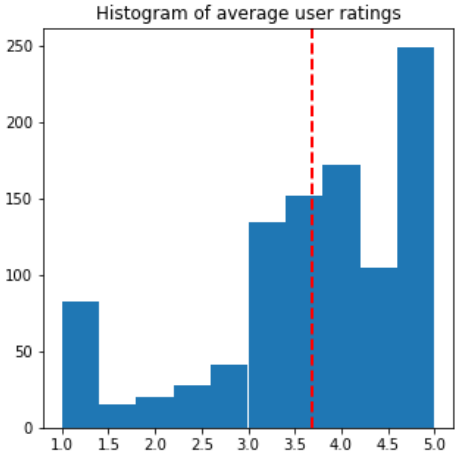
CS 109a Final Project: Recommendations

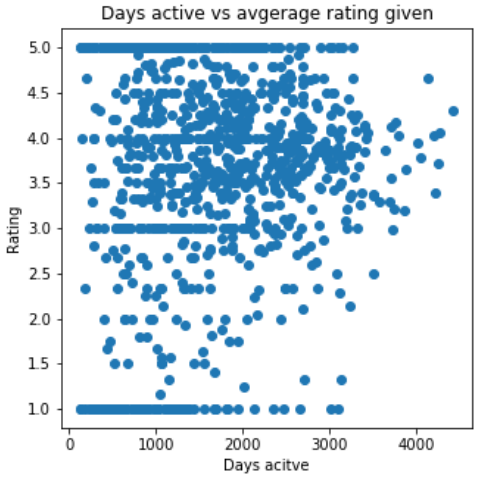
Group #20:
Maciej Holubiec, Paul Von Chamier, Jimena Romero

EDA:

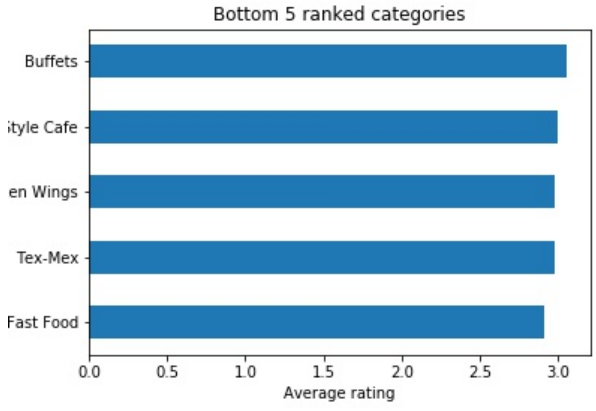
Data decription: We are dealing with 5 different json files, out of which 3 are useful: business.json, review.json and user.json. We are using all 3 for preliminary EDA. To build our models, we will need to transform review database to create a user_id-business_id matrix (filled with ratings). With our baseline model, we will fill the missing data with averages or predictions from a multiple regression. With the latent matrix model, we will try to predict using matrix factorization backed up by alternating least squares reguralization.

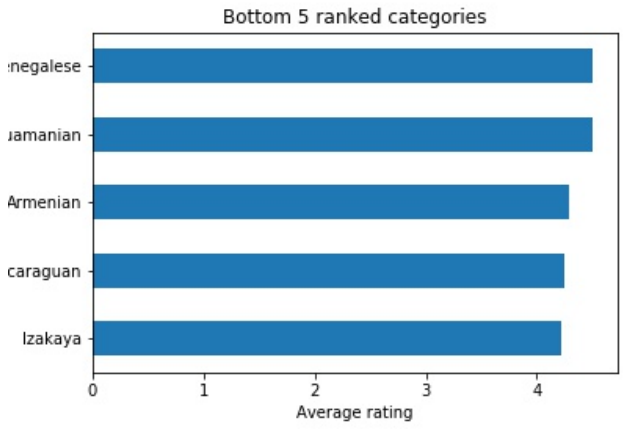
User analysis: As shown on plots below (sample of size 1000 from user.json), users give a rating above average (~3.6) and 5 is the most popular rating. Moreover, there is no clear linear relationship between how long the user has been active and the ratings they give. Moreover, we see ratings forming horizontal lines at 1,2,3,4,5. Further analysis (not included in this document) shows that those are the users that gave very few rankings.



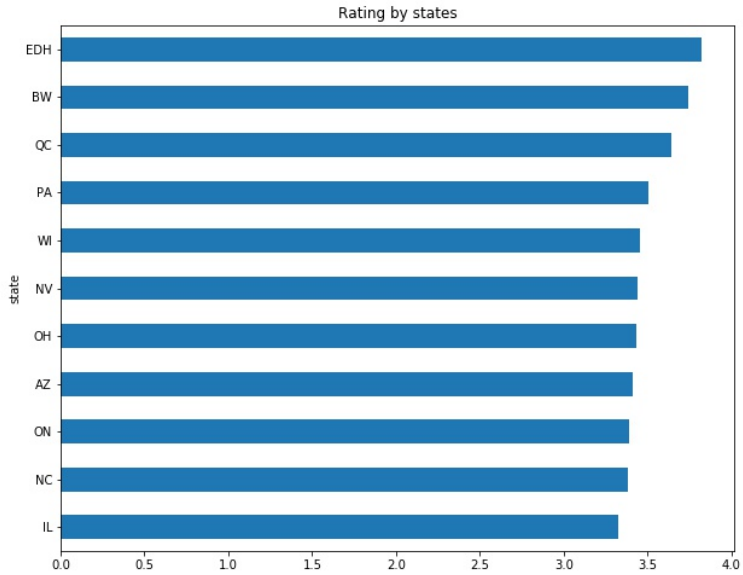


Categories analysis: The barh plots below (from business.json) show top 5 highest rated categories and and bottom 5 lowest rated categories. We see that highest rated categories include *Senegalese* whereas the lowest rated categories include *Fast Foods*. Further analysis (not included here) shows that highest scored cuisines usually have very few instances (all under 10), which makes it easier to maintain high ranking, whereas the lowest scored cuisines have a lot of instances (in thousands).

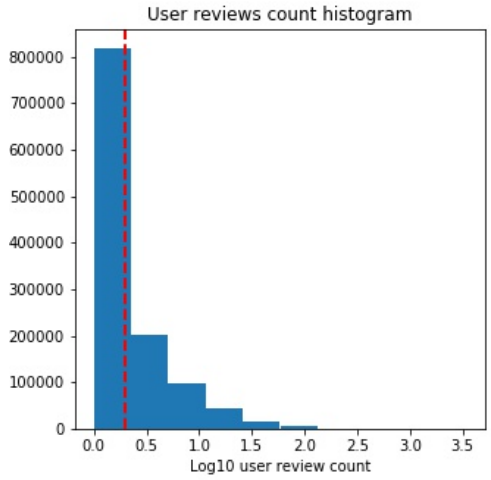


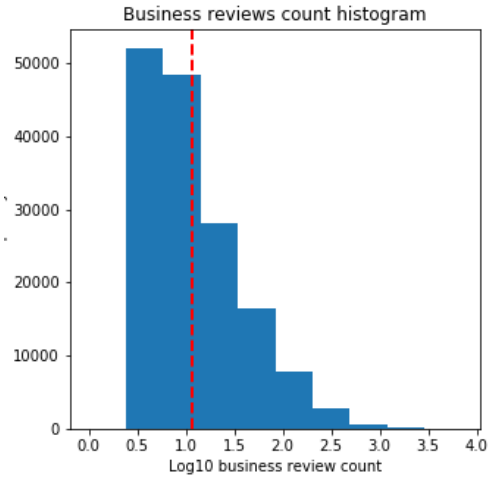


Ratings by geography analysis: Preliminary analysis showed that regions with highest average rankings ha very few restaurants. Therefore, we only look at regions with more than 500 ratings (shown below). The highest rated region is EDH, followed by BW, QC, PA, WI.



Ratings count analysis: Plots below show the histograms of logged user and business review counts. Both are heavily skewed, which shows that majority of the restaurants in the system receive a few ratings and majority of the users five only a few ratings. This means we have a lot of missing data, but thankfully, the dataset is big enough to fill it all in using our 3 approaches. To quantify this a bit, we see that use average is around 10^0.4=2.5 ratings and the business average is around 10^1.1=12.5 ratings.





REFINED PROJECT QUESTION:

Our project question stays the same: we want to build a matrix of ratings using business_id and user_id data from review.json and fill the missing data using the matrix factorization model. This will allow us to extract the top predictions for a given user. The only issue might come from the fact that on average, users rate only a few places and places are rated by only a few users. To fix this, we might only select the most frequented restaurants (choosing based on most frequented categories).