Building a Recommendation Engine for Yelp Data

Yelp Data Set

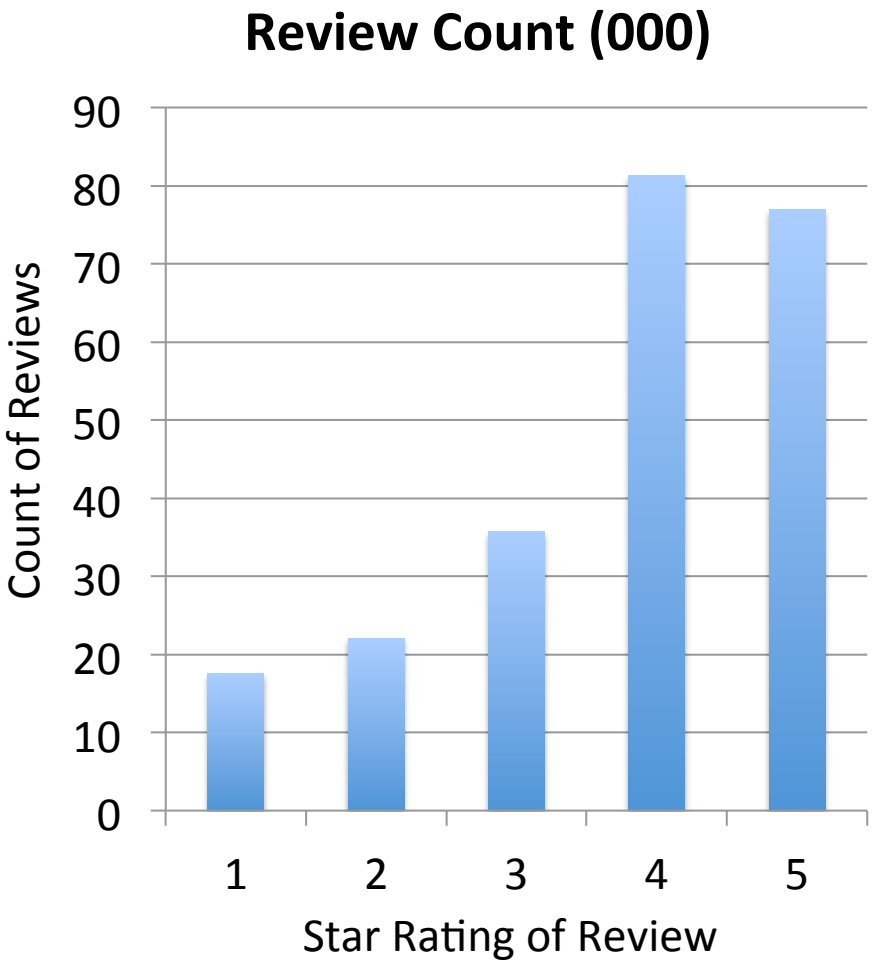
- Phoenix Metropolitan Area
  - 335k reviews
  - 16k businesses
  - 71k users
  - 112k business attributes
  - 11k check-in sets
  - 152k edge social graph
  - 114k tips

Data Used for Project

Data Munging and Exploration

- Cleaned up JSON files
  - Add missing delimiters
- Imported files into SQLite3 Database
  - Scripts for handling embedded dictionaries
  - Clean up missing values
- Used SQL Queries and Numpy to explore data

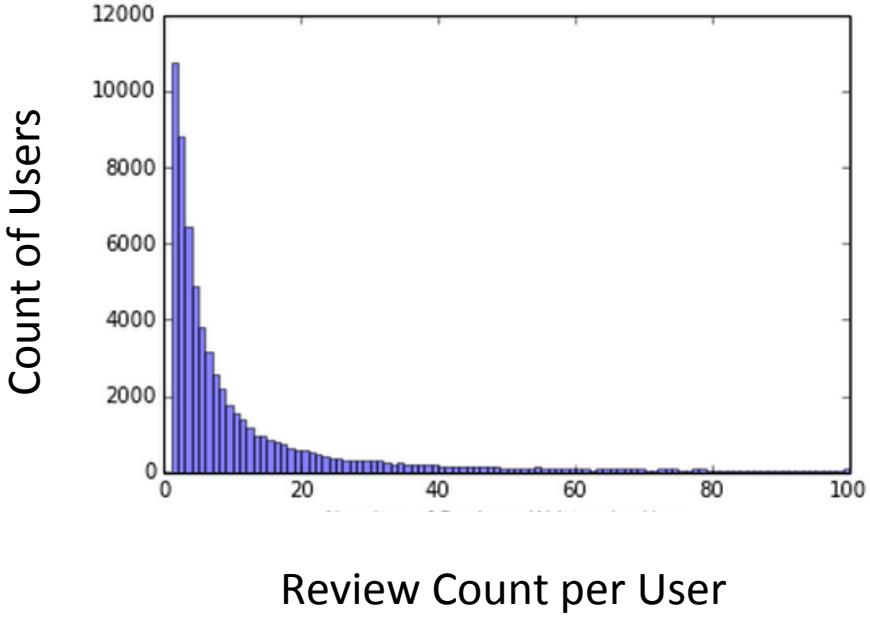
Distribution of Restaurant Reviews



Star Rating	Count of Reviews
1	17,576
2	22,085
3	35,718
4	81,363
5	76,976

Summary Statistics	
Mean	3.76
Standard Deviation	1.22

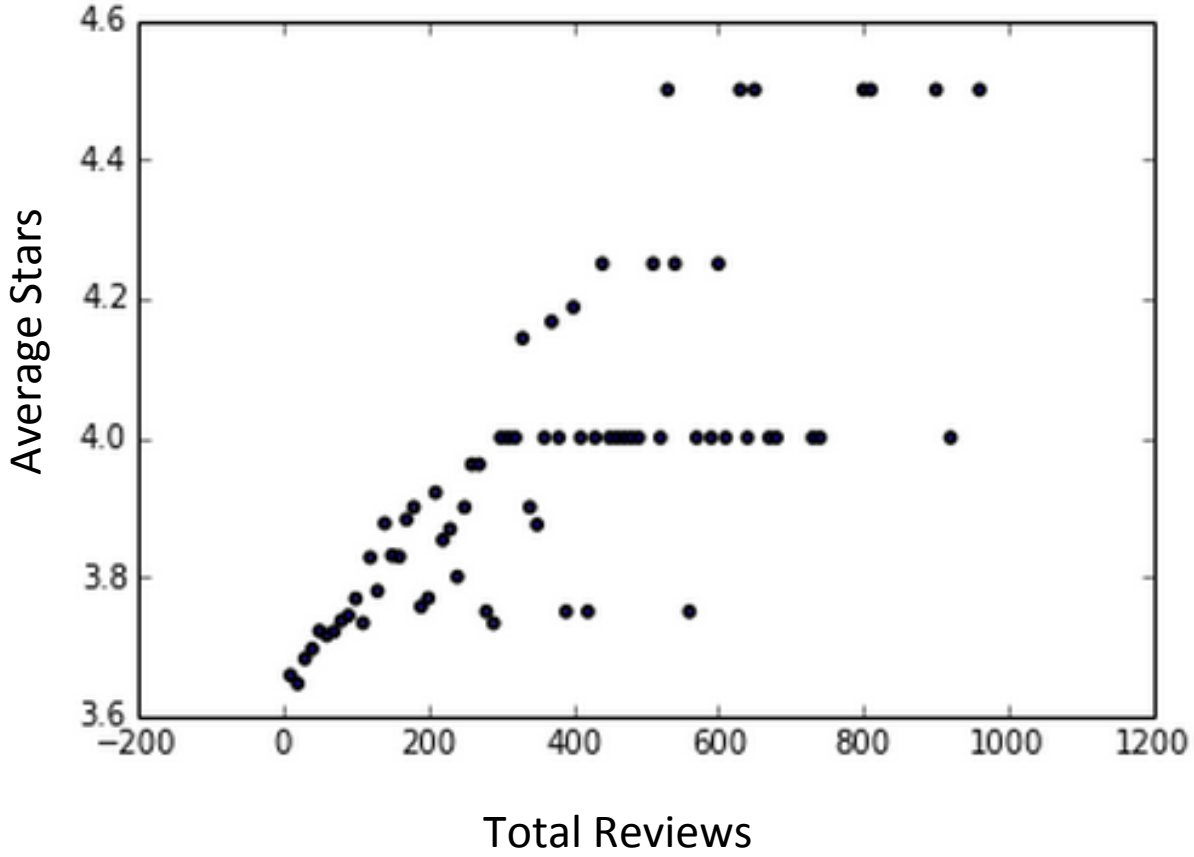
Distribution of Users



Total Reviews Written by User	Count of Reviews
1 -10	85k
11 - 20	42k
21 - 30	26k
31 -40	18k
41 - 50	14k
> 50	150k

38% of Reviews were Created by Users with < 20 total reviews

Distribution of Businesses

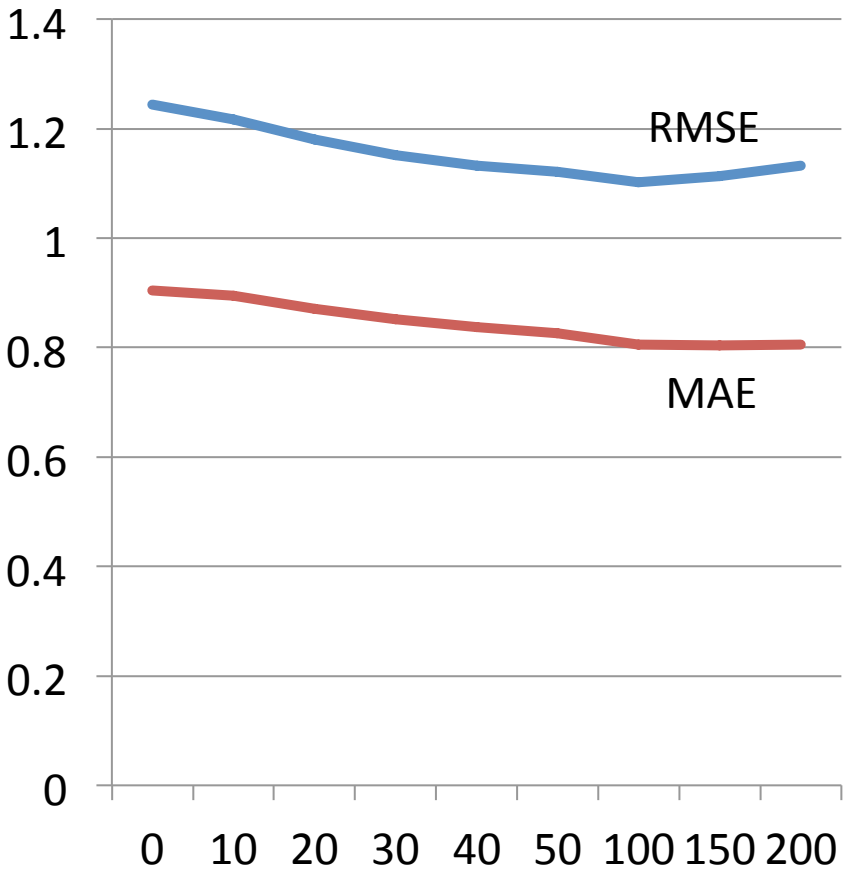


Business with higher total reviews have higher stars

Methodology

- RecSys Package
- Built sparse matrix based on:
  - Users
  - Restaurants
  - Ratings
- Calculated RMSE and MAE
  - Substituted multiple iterations of random sampling instead of cross-validation
  - Tested recommendations for similar restaurants

Measuring the Recommendation System



Users with Total Reviews	RMSE	MAE
> 0	1.244	0.904
> 10	1.217	0.895
> 20	1.181	0.871
> 30	1.152	0.851
> 40	1.133	0.837
> 50	1.122	0.827
> 100	1.102	0.805
> 150	1.114	0.804
> 200	1.132	0.806

Recommendation Systems Learnings

- Recommendations covered wide range of disparate categories (not intuitive at all)
- Filtering by categories helped to provide structure
- Difficult to measure practical accuracy without knowledge of Phoenix restaurants

Next Steps for Recommendations

- Test recommendation engine on friends from Phoenix
- Build ensemble recommendation system
  - Build more sophisticated filter for similar categories
  - Incorporate total reviews and average stars of restaurants
  - Scrape Zagat's scores from Phoenix

Other Things to DO

- Feature engineering
  - Average stars from only elite users
  - Weighted average stars based on how useful other users rated each other's reviews
  - Map reduce review text and build features based on count of key words
- NLP on review text
- k-Means Clusters on users and businesses
- Network analysis on friends