Yelp Recommendation System

Kumari Nishu, Mohit Chander Gulla, Neelam Patodia

December 2019

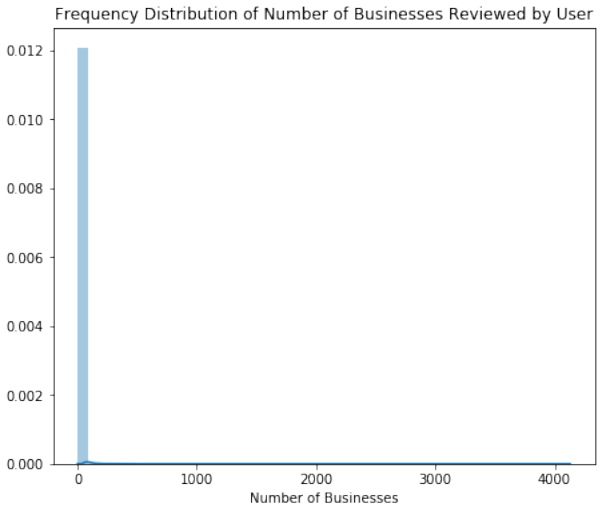
Table of Contents

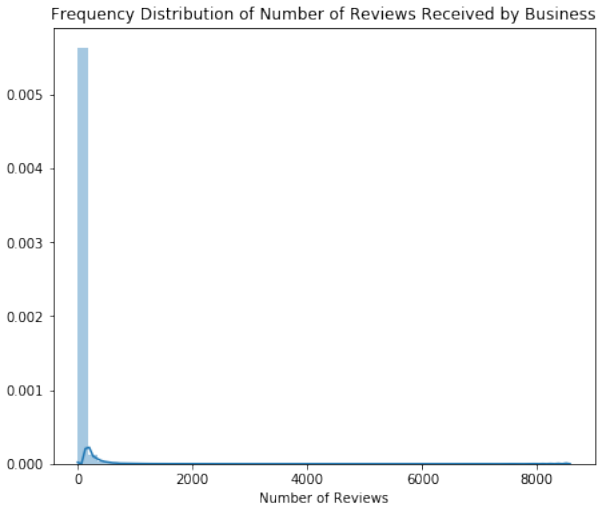
I. Project Scope

Our team's objective is to build a production-grade recommendation system using Yelp Dataset that provides curated business recommendations to users based on their explicit feedback and implicit features. This report goes in depth to define the scope of this project, to rationalize any assumptions that were taken in the process and arrive at a plan of action after analyzing results of all models that were explored.

II. Exploratory Data Analysis

In entirety, Yelp data comprises of more than 6.6M reviews for approximately 190K businesses. As we are prototyping, we will design and implement our models on a much smaller subset of data. This not just allows us iterate changes faster to improve our model, but also because it is possible to learn user preferences meaningfully on a smaller subset.

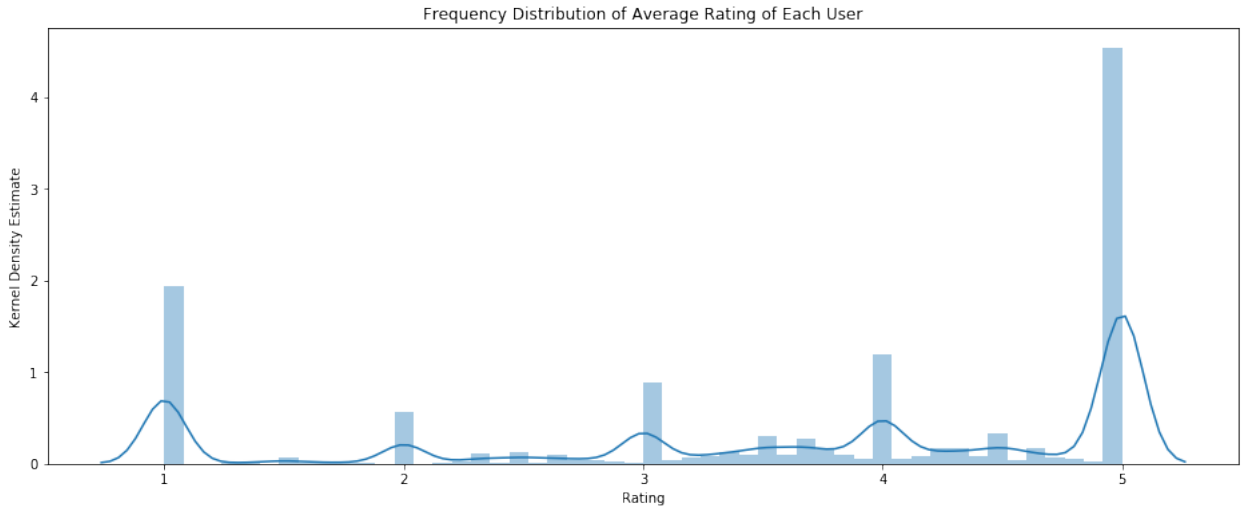




Looking at the frequency distribution of our data, both from user and business perspective, show us that majority of users have reviewed very few businesses and likewise majority of businesses have received very few reviews. The user-business interactions are not dense, and we would like to consider only active users and active businesses for our model in order to eliminate noise.

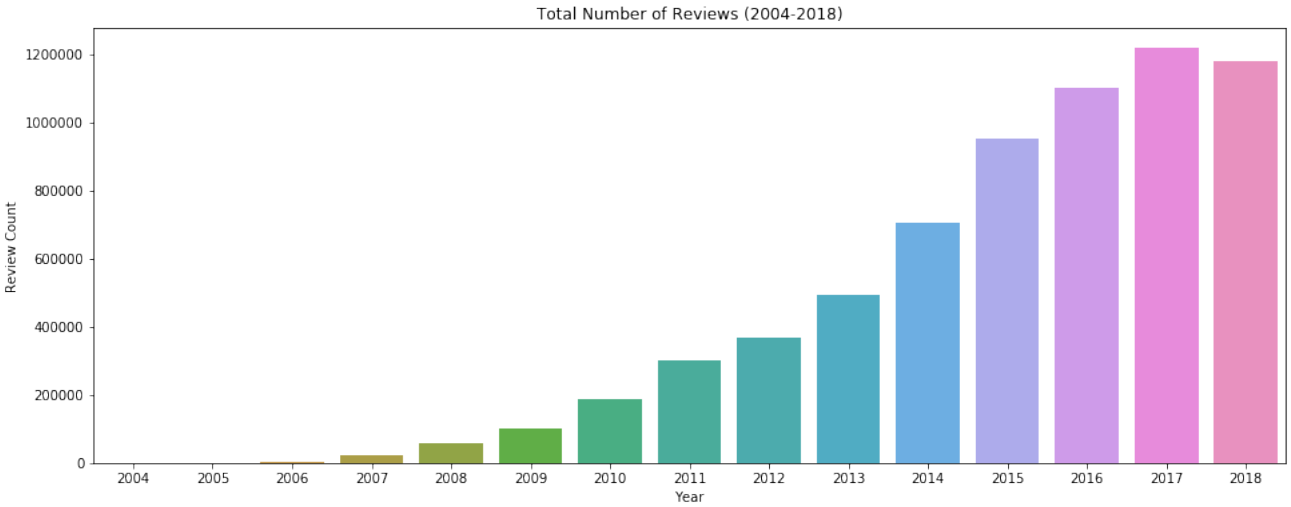
Total Number of Unique Users: **1,637,138**
Total Number of Unique Businesses: **192,606**

	Mean	Median
No. of Businesses Reviewed by User	4.08	1
No. of Reviews Received by Business	34.71	9

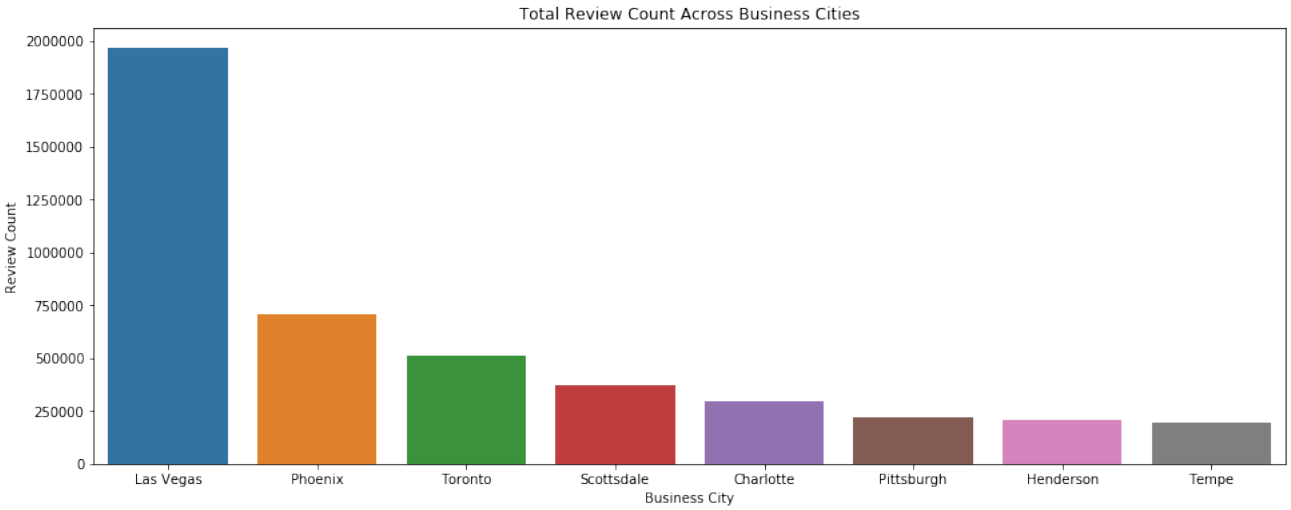


Mean of Users' Average Rating = **3.64** Median of Users' Average Rating = **4.0**

The above results were contradictory to our expectation. We assumed that users in general take the effort to provide a review mostly when they have had a negative experience. And even though there are significant number of users with their average rating across businesses equal to 1, the majority of users have their average rating greater than 4. As the perception of good and bad in terms of 1-5 scale differs from person to person, user bias is an important feature.



Over the last 14 years, the number of reviews that were posted on Yelp kept increasing before it dipped marginally in 2018. It is safe to assume that user preferences change over time and even the performance of a business can change considerably. As our end outcome is to provide users with recommendations, it should reflect the current state of a business, and align and learn from current taste of similar users. Moreover, as we are looking to train our model on a smaller subset of data, it is logical to consider recent data for analysis. We will subsequently add previous years data to analyze how well our model scales.



In term of total review count, the businesses situated in Las Vegas have received more reviews than businesses in any other city. Las Vegas covers 30.4% of total reviews in our data. As users' spending capacity, preferences and type of businesses vary from city to city, we will subset our data to only include reviews attributed to businesses from Las Vegas. At this stage, this project is not focused on building a location-aware recommendation system and in order to ensure that our business recommendations are relevant to users, in terms of proximity, we need to filter our data for Las Vegas.

III. Business Rules / Assumptions

IV. Sampling Strategy

V. Feature Engineering