Yelp Dataset Challenge - Seasonal Patterns

Christoph Fabianek

Sunday, November 22nd, 2015

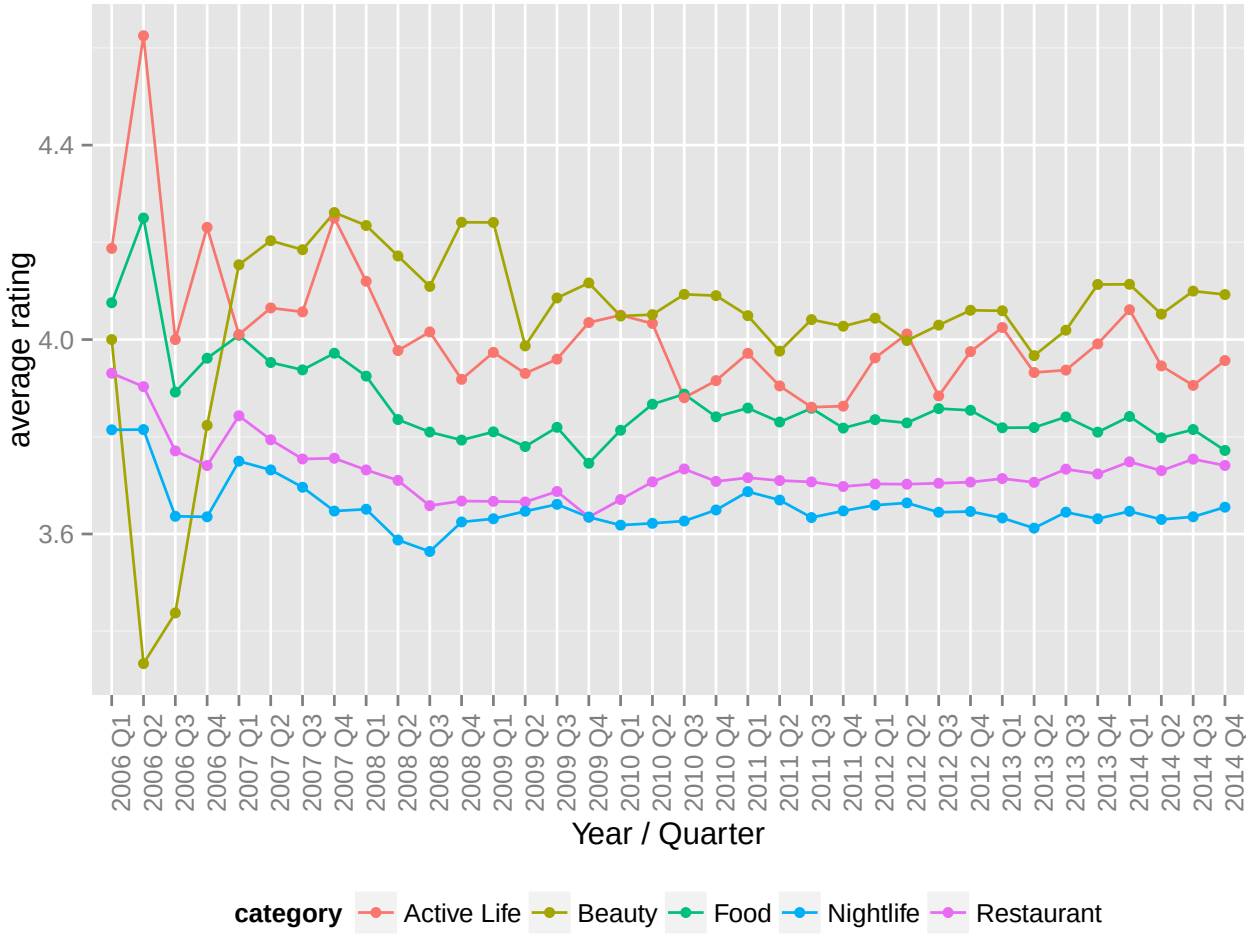
Introduction

This report addresses the following question: *Do seasonal review patterns exists in the Yelp Dataset?* I investigate businesses in different categories to search for patterns in the number of reviews and in the star rating. Identified patterns are afterwards tested for statistical significance using hypothesis testing.

This report was written for the *Coursera Data Science Specialization* and is based on data from the *Yelp Dataset Challenge*. The underlying source files are available on Github: https://github.com/fabianekc/DataScience_Capstone

Methods

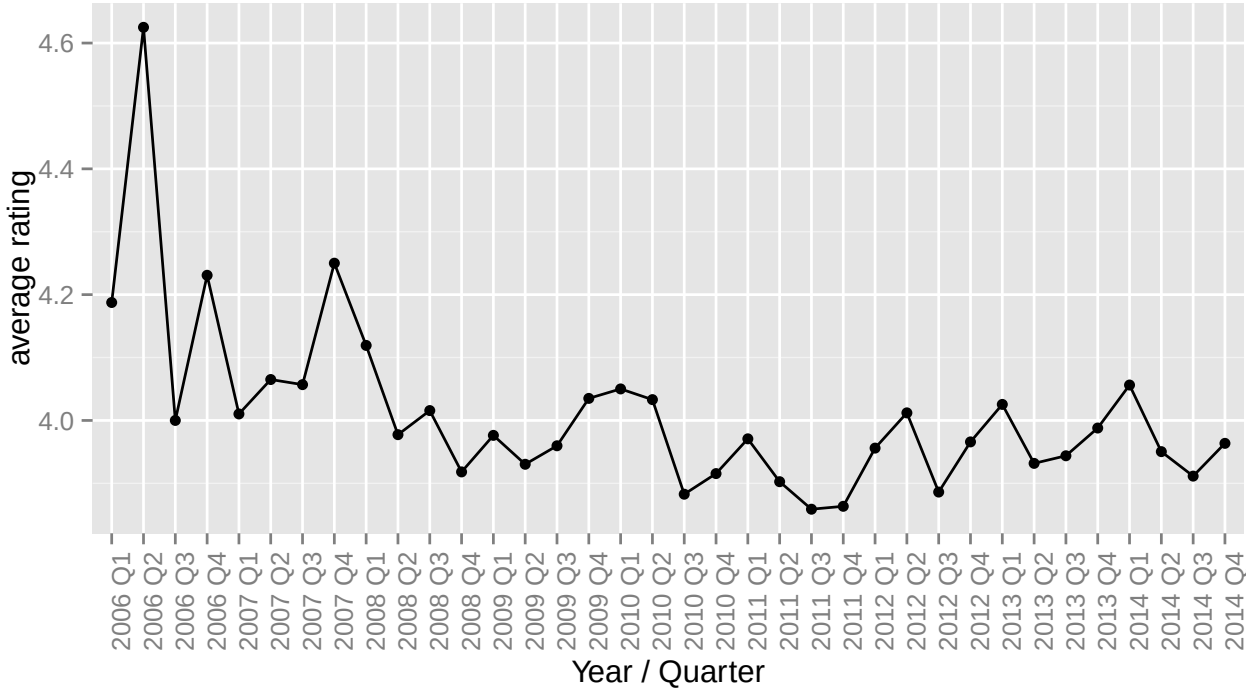
I started with some data exploration and the plot below shows the star rating for 6 categories with a high number of ratings.



From a first look we see that *Active Life* shows a lot of variation. For a further exploration I combined similiar categories to get a larger sample. The group **Sports** includes in the following anlysis these Yelp categories:

- Active Life
- Gyms
- Fitness & Instruction
- Pilates
- Yoga
- Weight Loss Centers

The following graph shows the average ratings per quarter for the group *Sports*.

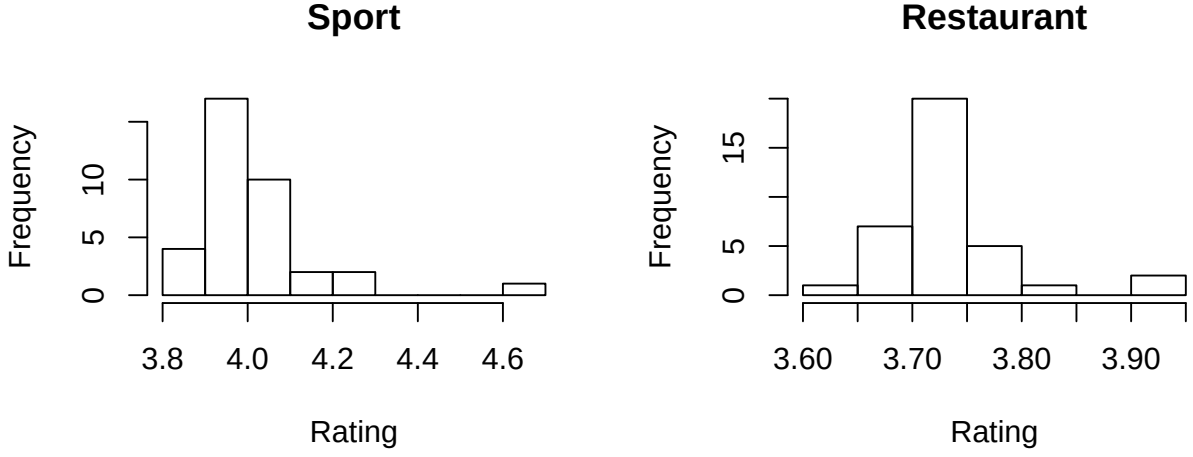


To address the question of a seasonal review pattern for sport categories I formulate the following null hypothesis:

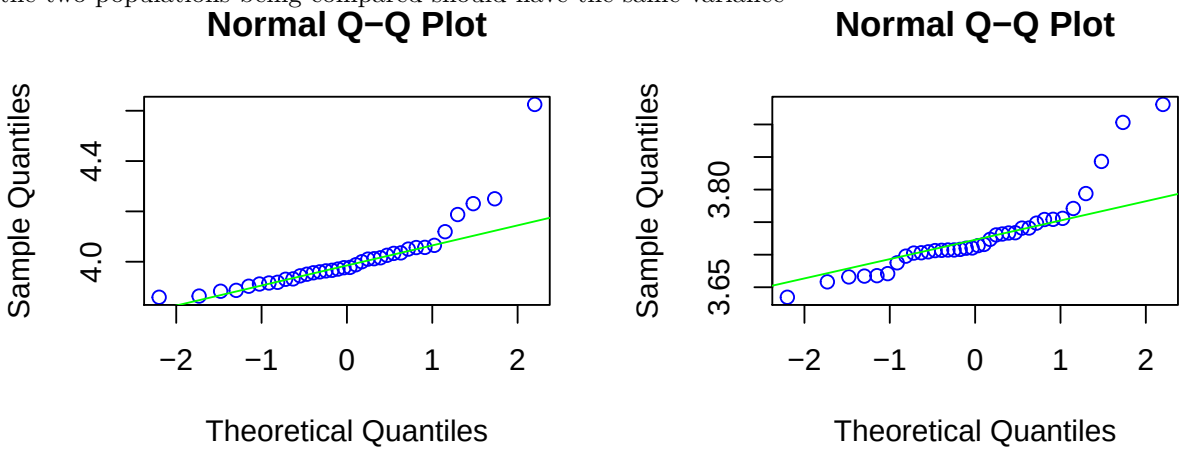
H_0: The season (time of the year) has no effect on sports activity ratings.

A T-Test is used to compare the average ratings for sport categories against average ratings for restaurants (restaurant ratings have the most steady average ratings as shown in the first figure). Before we perform the T-Test we need to check the following assumptions:

- the two populations being compared should follow a normal distribution



- the two populations being compared should have the same variance



- the data ushed should be sampled independently from the two populations being compared

Results

H_0: The season (time of the year) has no effect on sports activity ratings.

T-Test Result:

p.value	CI.low	CI.high
0	0.2318871	0.3342908

Interpretation: the confidence interval does not contain 0 and the p-value is less than 5% so H_0 can be rejected

Discussion

So I can answer the primary question of this report that the season (time of the year) has an effect on sports activity ratings. We have seen that ratings are usually higher at the begin of the year and this is maybe releated to New Year's resolutions and the positive spirit when starting a new activity.

Of course all of this is only true under the following assumptions:

- the experiment was performed with a random assignment of participants
- the sample ratings are representative for the overall population