Final Paper

Introduction

introduction here

Using Yelp Data

The yelp data set (https://www.yelp.com/dataset_challenge) consists of information about Yelp users, reviews they wrote and business information.

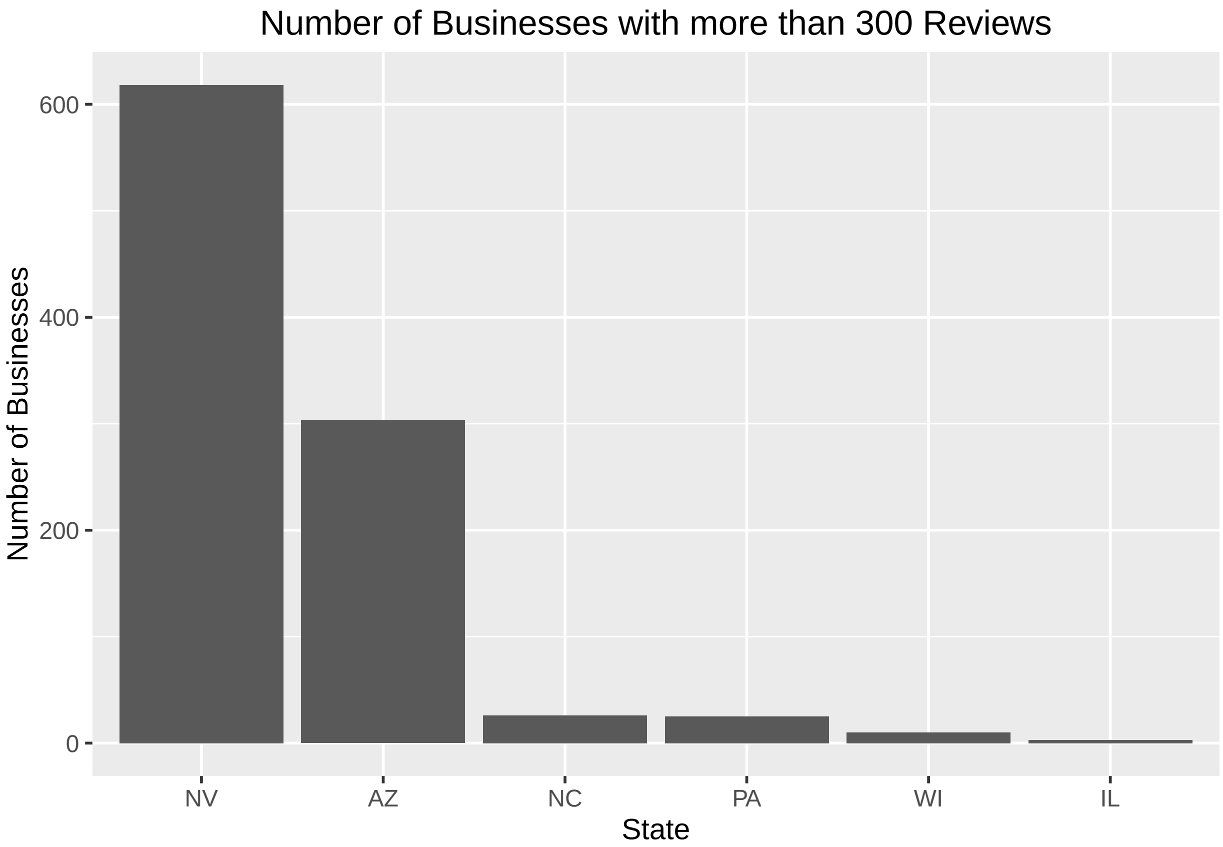
Business Clusters in the US

The business data set provided 77,445 rows of information such as review counts, ratings and locations of local businesses. After carefully looking at the data set, we have decided to filter out the more significant data by limiting it to businesses having more than 300 reviews in the US.

We have first tried to plot this information with points on a map. Points with different sizes stacked up on each other and unexpectedly, the size and color of the points did not tell much information. Even though 900+ points were plotted, there were six clusters of points and there was no point of having so many individual points.

Instead, we ended up with a bar graph, which clearly shows which state has the most number of businesses with over 300 reviews. There were 6 distinct regions in total. Interestingly, the graph shows that the majority of the businesses is located in two states: about 62% of the filtered businesses were clustered in Nevada area and another 31% was in Arizona.

We have hypothesized earlier that location will influence a business' success. The plot shows



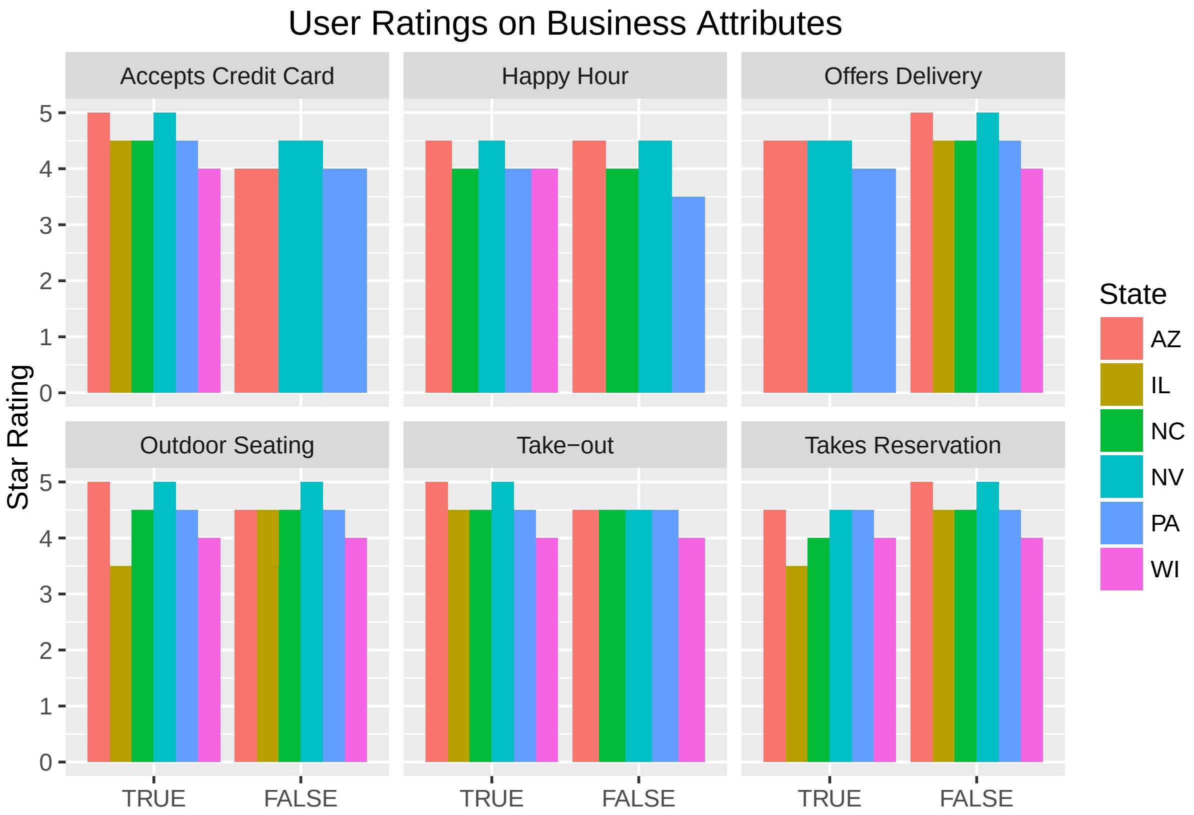
Business Attributes

Looking more into the data set, we wanted to know if certain attributes in specific regions make businesses stand out. The business data set is now cleaned and gathered in order to make each row contain one business attribute.

We have first tried to include only two representative plots. However, subsetting too much of the data set made us lose many valuable information., It could also potentially lead to overgeneralization.

We wanted to see if any of the attributes of the businesses makes them successful. Six major attributes were chosen among many. There were slight differences in some categories. For example, businesses that accept credit cards seem to have higher ratings compared to those that don't. There were no major findings from the plots, however.

Except for slight differences in average ratings, there was no exceptional difference shown among the choices. The user review ratings were evenly spread out no matter the specific attribute exists or not. The result points out that business attributes do not have huge influences over user ratings.



Users Behavior Changes When Travelling

If business attributes do not impact business' success, then what would?

From analyzing over 2 million reviews, we wanted to find out what would affect people's behavior of rating businesses. Since the data set given by Yelp was academic, this did not include every single review by every user in data. We have focused on active users that have written more than 50 reviews in total and have 10 reviews in the data provided at the same time. This seemed enough to analyze the user trends.

First, we went with facet_wrap in plot after selecting top 15 influencers with most reviews. However, since facet focuses rather on individual data than overall trends, it was hard to generalize the findings. Some individual preferences were really different from others.

Another approach we took was to plot everyone's ratings into the histograms. From the filtered user data set from above, one familiar region was chosen per user. This was based on where the user wrote most reviews. Two average star ratings were calculated for a user: one for reviews in familiar state and another from reviews when travelling unfamiliar places. The results are shown below in density curves.

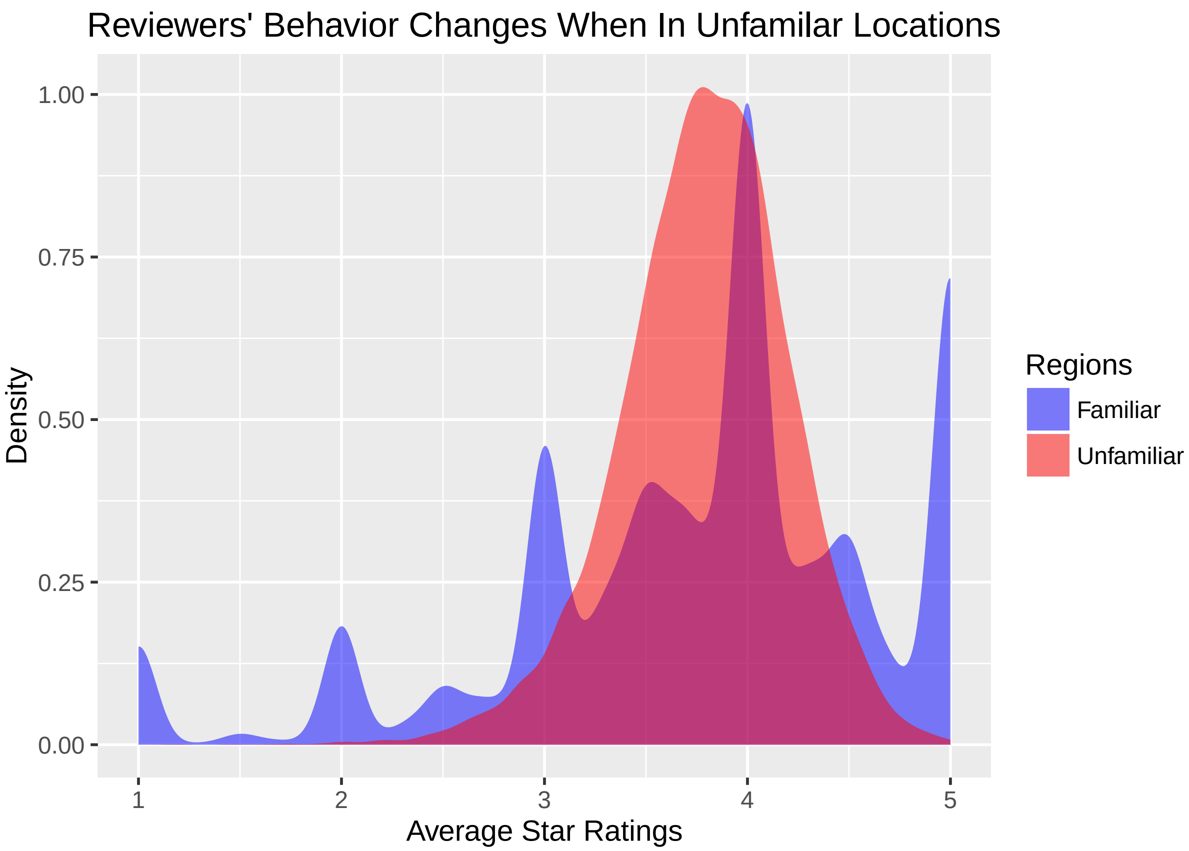
At initial observation, the density curve for unfamiliar region is smooth and the majority of users tend to give star ratings in between 3 to 4.5 out of 5. Number of people giving 1 or 2 stars while traveling is barely seen. On the other hand, the ratings in familiar places were rather widely distributed. There were good amount of people giving star ratings from 1 through 5, even though the peak of the curve is around 4.

Interestingly, the mean value of each place turned out to be 3.718 for unfamiliar and 3.762 for familiar places. However, the standard deviation for the familiar place plot is 0.41 and unfamiliar one is 1.37 so that star ratings for unfamiliar places looked higher at first.

```
## Joining by: "user_id"
```

```
## Joining by: "business_id"
## Joining by: c("user_id", "n")
```

```
## Joining by: c("user_id", "state")
```


Reviewers' Behavior Changes When In Unfamilar Locations
Density
Average Star Ratings
Regions
Familiar
Unfamiliar