Final Project

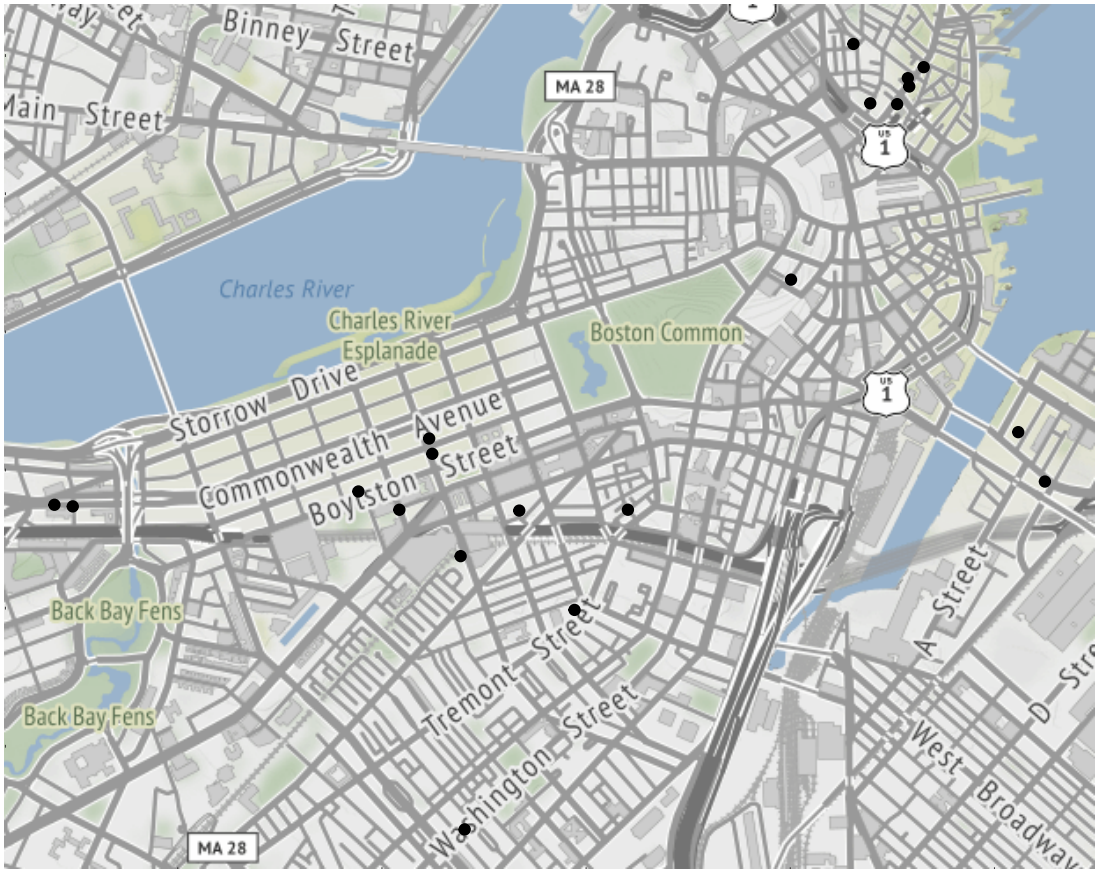
He Guo

11/27/2019

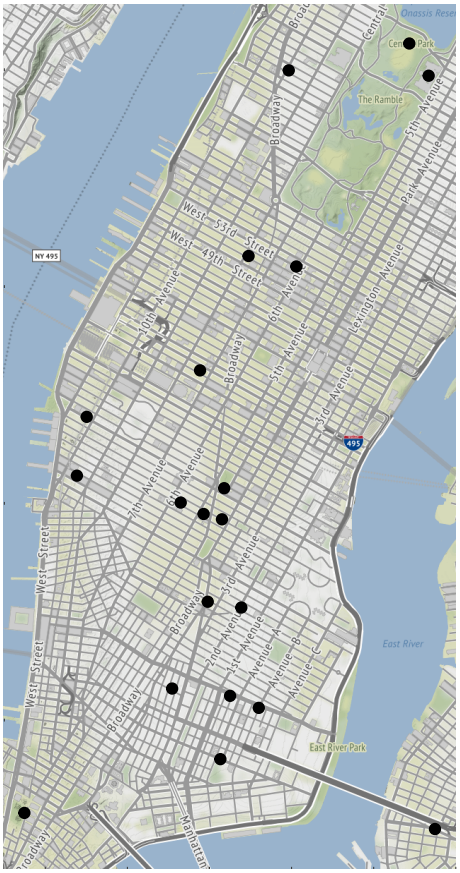
Yelp API

The current study gets business data from Yelp API. The study uses longitude and latitude data from Yelp API to plot the map of business in Boston.

```
## Using zoom = 14...
## Source : http://tile.stamen.com/terrain/14/4956/6059.png
## Source : http://tile.stamen.com/terrain/14/4957/6059.png
## Source : http://tile.stamen.com/terrain/14/4958/6059.png
## Source : http://tile.stamen.com/terrain/14/4956/6060.png
## Source : http://tile.stamen.com/terrain/14/4957/6060.png
## Source : http://tile.stamen.com/terrain/14/4958/6060.png
## Source : http://tile.stamen.com/terrain/14/4956/6061.png
## Source : http://tile.stamen.com/terrain/14/4957/6061.png
## Source : http://tile.stamen.com/terrain/14/4958/6061.png
```



The current study get business data from Yelp API. The study use longitude and latitude data from Yelp API to plot the map of business in New York.



Cluster Analysis for business dataset

This study want to apply cluster analysis to get the sub group of yelp business data set. The study want to see each business in yelp business dataset in the same subgroup to be similar, and businesses in yelp business dataset from different subgroup to be different.

Data Preparation

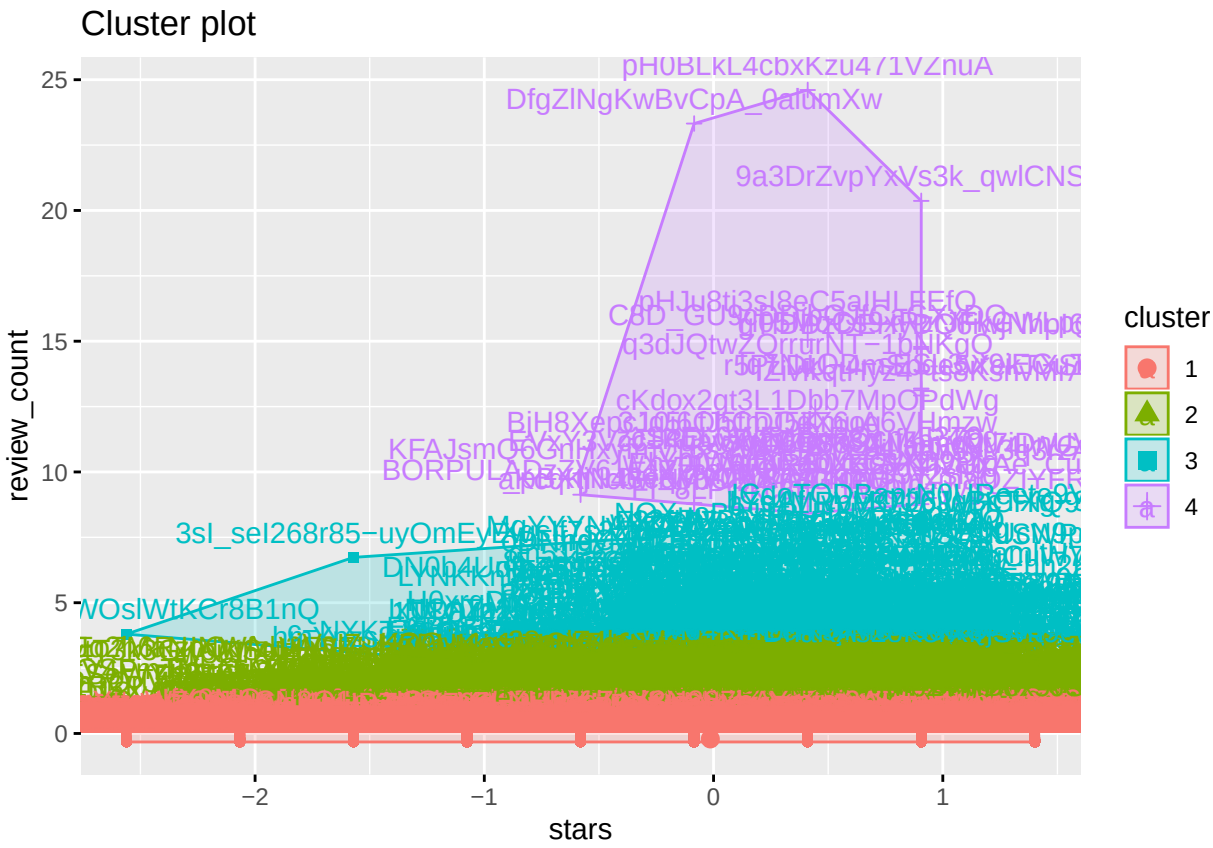
This study random select 10000 observations from yelp business data set , and get the numeric information need to be used in furture analysis. Next the study removes the missing value. The study picks the stars of business and review count of business as interested variable and apply cluster analysis to those variable. The cluster analysis is based on the stars of business and review count of business.

Computing k-means clustering

This study want to use cluster analysis splits observation into four clusters, and generate 25 initial configurations based on review count and stars.

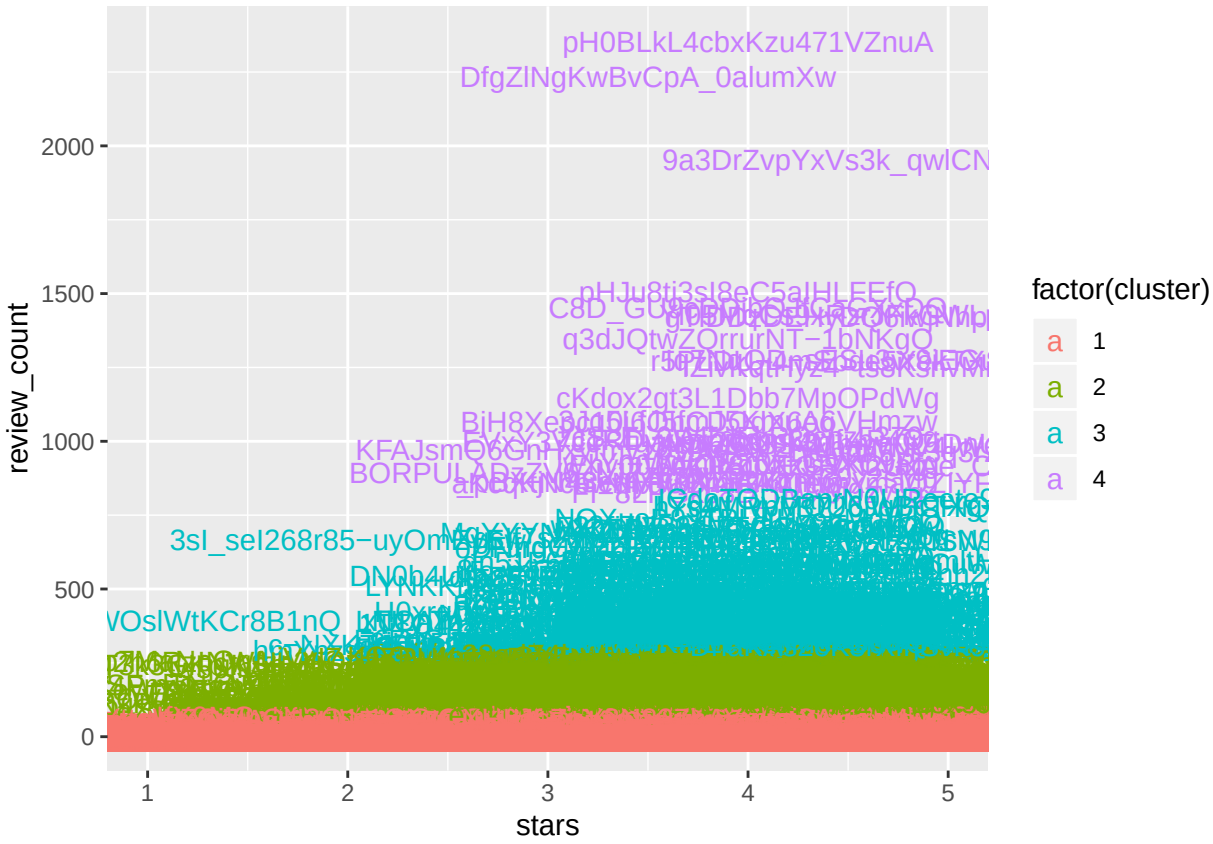
Using fviz_cluster

This methods could visualized the cluster the study did.



Using fviz_cluster

This study also want to plot cluster with the business_id as their label on it. Based on the plot the study can know that the business from cluster group 1 have zero review count. Business from cluster group 2 have more review count and higher stars than business cluster group 1. Business from cluster group 3 have more review count and higher stars than business cluster group 2. Business from cluster group 4 have more review count and higher stars than business cluster group 3.



Using fviz_cluster

This study gets the cluster after apply cluster analysis, then put the culster into yelp.business data frame. The study want to draw the map based on the cluster. The study use qmplot to draw the loaction mqp for four cluster data frame, which are yelp.business.C1, yelp.business.C2, yelp.business.C3, yelp.business.C4.

Cluster 1

Business in Cluster type 1 data frame are located in 14 states. The sate cotains largest number of business in cluster group 1 is AZ.

```
## Warning in kableExtra::kable_styling(., bootstrap_options = c("striped", :
## Please specify format in kable. kableExtra can customize either HTML or
## LaTeX outputs. See https://haozhu233.github.io/kableExtra/ for details.
```

state	n
AZ	2586
ON	1634
NV	1498
OH	743
NC	719
PA	546

```
## Using zoom = 5...

## Source : http://tile.stamen.com/terrain/5/5/10.png

## Source : http://tile.stamen.com/terrain/5/6/10.png
```

```
## Source : http://tile.stamen.com/terrain/5/7/10.png
## Source : http://tile.stamen.com/terrain/5/8/10.png
## Source : http://tile.stamen.com/terrain/5/9/10.png
## Source : http://tile.stamen.com/terrain/5/5/11.png
## Source : http://tile.stamen.com/terrain/5/6/11.png
## Source : http://tile.stamen.com/terrain/5/7/11.png
## Source : http://tile.stamen.com/terrain/5/8/11.png
## Source : http://tile.stamen.com/terrain/5/9/11.png
## Source : http://tile.stamen.com/terrain/5/5/12.png
## Source : http://tile.stamen.com/terrain/5/6/12.png
## Source : http://tile.stamen.com/terrain/5/7/12.png
## Source : http://tile.stamen.com/terrain/5/8/12.png
## Source : http://tile.stamen.com/terrain/5/9/12.png
```

Cluster 2

Business in Cluster group 2 data frame are located in 4 states. The sate cotains largest number of business in cluster group 2 is NV.

state	n
AZ	276
NV	200
ON	118
NC	58
PA	37
OH	36

```
## Using zoom = 5...
```

Cluster 3

Business in Cluster group 3 data frame are located in 10 states. The sate cotains largest number of business in cluster group 3 is NV.

state	n
NV	74
AZ	50
ON	10
OH	9
PA	6
NC	4

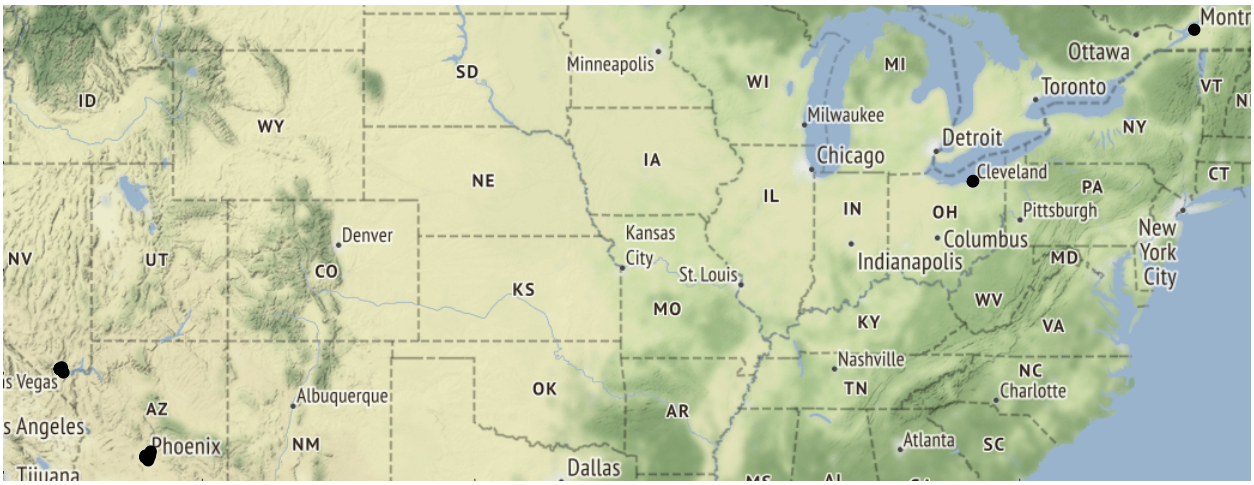
```
## Using zoom = 5...
```

Cluster 4

Business in Cluster group 4 data frame are located in 11 states. The sate cotains largest number of business in cluster group 4 is AZ.

state	n
NV	18
AZ	9
OH	2
QC	1

```
## Using zoom = 5...
```



Discussion

Based on the cluster analysis, the study find out AZ have more business in cluster group 1 and cluster group 4. It means that the number of business with low review count in AZ is the largest, and the number of business with with high stars and large review count in AZ is the largest. However, we can see that the number of business in AZ is the largest. Maybe, this is the reason business in AZ take a large propoation in cluster group 1 and cluster group 4.

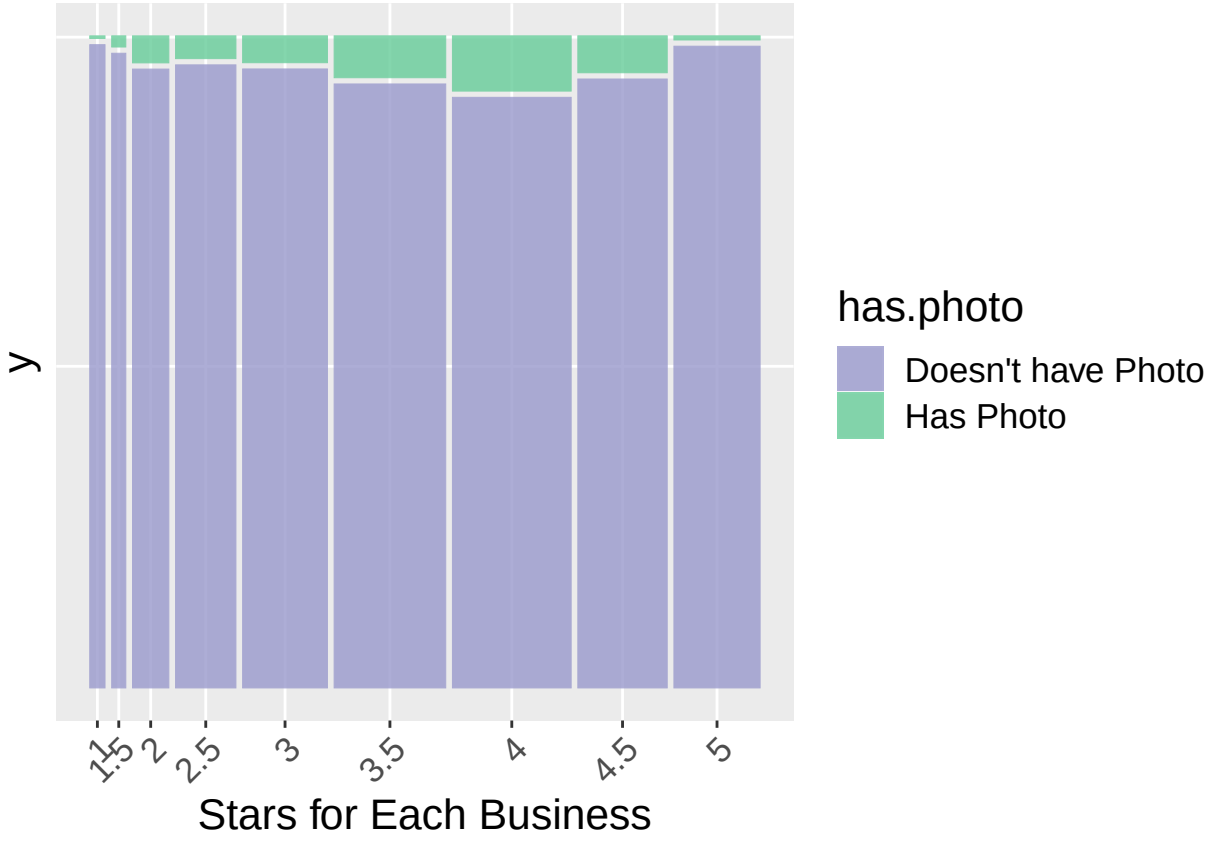
state	n
AZ	2921
NV	1790
ON	1762
OH	790
NC	781
PA	589

Stars and Photo

The study are interesting the distribution of business stars with photo and without photo. The study take a sample data frame from photo jason file in Yelp data set. This study use left_join to combine the photo data frame and business data frame. Left joint will keep the all row of left data frame. It means that if a business_id is not contained in photo data set, The row of this business id will return NA for its photo_id columns. The study trades NA as this business has no photo on Yelp.

According to the following plot, a lot of business do not have photo. The study take the sample of photo data file and business file. It may cauese that a lot of business do not have photo. Based what we have for now, when business have more photo, the stars for thoes business will aries to 3.5 to 4.5. There is a big surprise that only a few business, which have 5 starts, have photo.

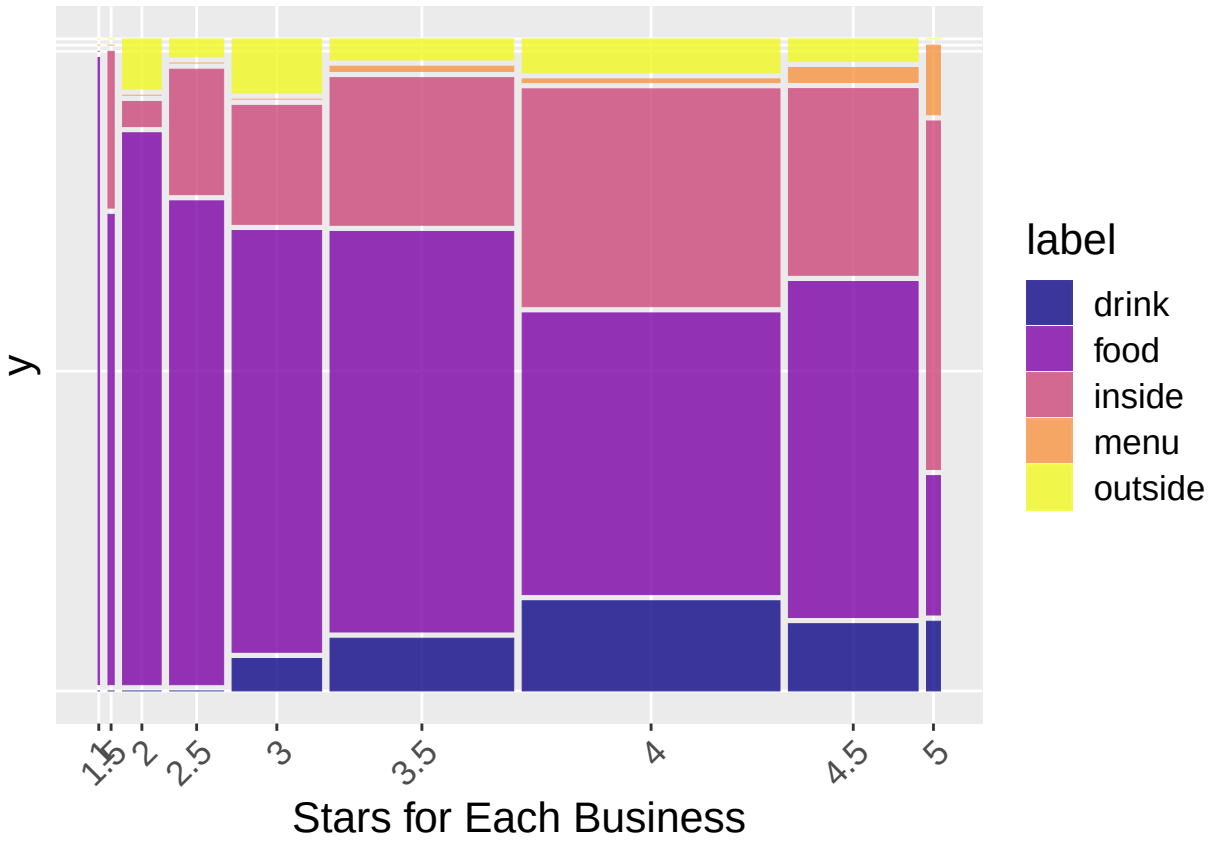
The plots used to analyze the distribution of stars for business with photo and without photo are mosaic plots. The mosaic plot is a graphical representation of the 2 kinds of frequency table. The mosaic plot is been divide by rectangles, then vertical length is the proportion of has.photo variable on y axis in each level of stars variable on



axis.

According to the following figure, business with more than 4 stars has more picture witl outside business, inside business, and menu. Business, which has star picture with inside label, take the largest propotion. Then, the study conclude that including more more picture witl outside business, inside business, and menu could help business to receive more higher stars.

The plots used to analyze the the distribution of stars for business with different photo is mosaic plots.The mosaic plot is a graphical representation of the 2 kinds of frequency table. The mosaic plot is been divide by rectangles, then vertical length is the proportion of photo label variable on y axis in each level of stars variable on axis .


y
1.5
2
2.5
3
3.5
4
4.5
5
Stars for Each Business
label
drink
food
inside
menu
outside