Title

A study on exploring the impact of geographic location (east or west regions of USA) to the Yelp review score of a restaurant.

Introduction

This is a report on studying the Yelp data set to explore whether people live in the east and west regions of United States having different trends to review their restaurants. Those who want to receive higher review score in Yelp when they open a new restaurant may refer to this.

There are definitely much more other factors impacting the review score of a restaurant, like the exact location in town, the price, whether having wifi, the type of the restaurants and etcs. Well, in this report we just focus on the impact from the broad geographic location, namely the east and west regions. The techinques to answer a more narrowed question than this will be similar to this report.

Basic running environment of this report is as below.

```
## [1] "R version 3.2.2 (2015-08-14)"
```

```
## [1] "OS X 10.11.1 (El Capitan)"
```

Methods and Data

Download and Read Data

The Yelp Dataset is over 500MB and 1.7GB after unzip. All files are in json format and some brief introduction of the dataset can be found from Yelp website. The data that will be used in the report are *business* and *review* datasets.

There are various methods to read joson files. In this report I choose the *stream_in* method from *josonlite* package. Below code will do the job.

```
if(require(jsonlite) == FALSE) install.packages("jsonlite")
setwd("./Chapter10.Data_Science_Capstone/SourceData/yelp_dataset_challenge_academic_dataset/")
business <- stream_in(file("yelp_academic_dataset_business.json"))
review <- stream_in(file("yelp_academic_dataset_review.json"))
```

After parsing the json data, there is a need to "flat" the data as there are nested data frames. I used *flattern* method to flat the data and also saved the data into *".rds"* format so next time it'll be much faster to just read the saved data.

```
businessf <- flattern(business)
reviewf <- flattern(reviewf)
saveRDS(businessf, "./SourceData/businessf.rds")
saveRDS(reviewf, "./SourceData/reviewf.rds")
```

Determine the Region of the Restaurant

There are *city* and *state* column in the *business* dataset. However, after a little exploration I found it was not so accurate to determine the location of the business. And luckily there is *longtitude* and *latitude* of the business.

```
names(business)
```

```
##  [1] "business_id"   "full_address"  "hours"         "open"
##  [5] "categories"    "city"          "review_count"  "name"
##  [9] "neighborhoods" "longitude"     "state"         "stars"
## [13] "latitude"      "attributes"    "type"
```

Filter Restaurant Businesses in USA

The *Maps* package has great features on geography. The country of the businesses can be determined by below code chunk.

```
library(maps)
businessf$Country <- map.where(database = "world", businessf$longitude, businessf$latitude)
unique(businessf$Country)
```

```
## [1] "USA"              "Canada"           "UK:Great Britain"
## [4] "Germany"
```

There are 4 countries, they are USA, Canada, UK:Great Britain, Germany. We only focus on USA restaurants. When we talk about "restaurant", sometimes a bar or pub may be referred to. To make it clear, I only choose those businesses whose *category* contains "restaurant" only. Eventually, *businessUSRst* will be the data frame of the USA restaurants I'm interested in.

```
businessfUS <- businessf[which(businessf$Country == "USA"), ]
businessfUSRst <- businessfUS[grepl("[Rr]estaurant", businessfUS$categories), ]
```

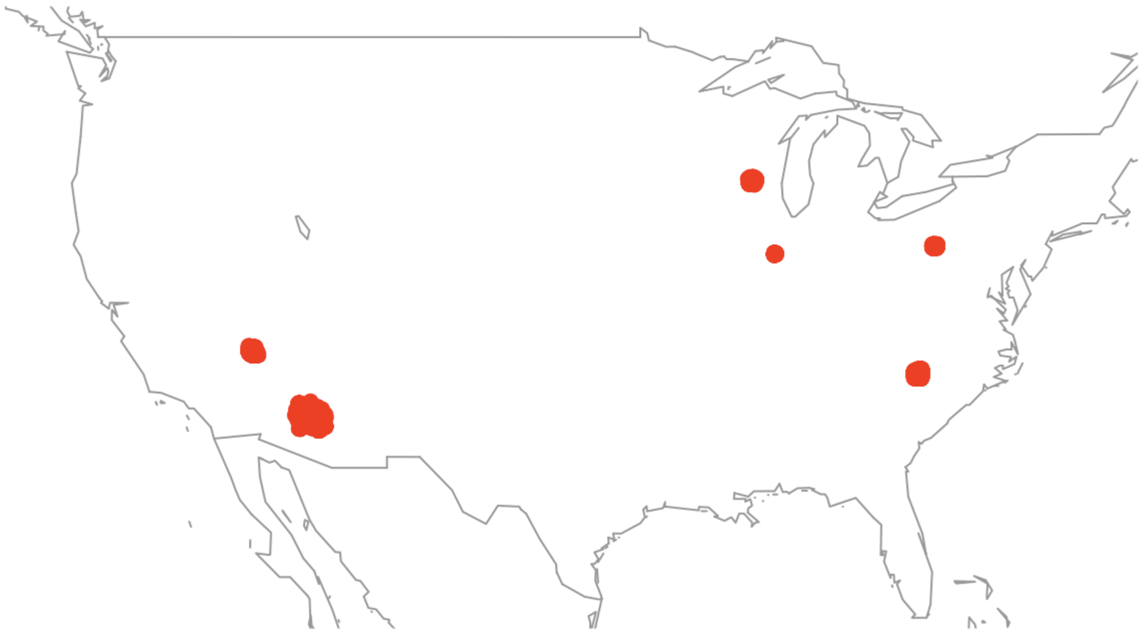
Now let's take a look at the range of longtitude and latitude so that we can plot all the business on the map properly.

```
rbind(range(businessfUSRst$longitude), range(businessfUSRst$latitude))
```

```
##             [,1]      [,2]
## [1,] -115.35190 -79.85026
## [2,]   32.87664  43.25227
```

The above figures will be used to plot below map.

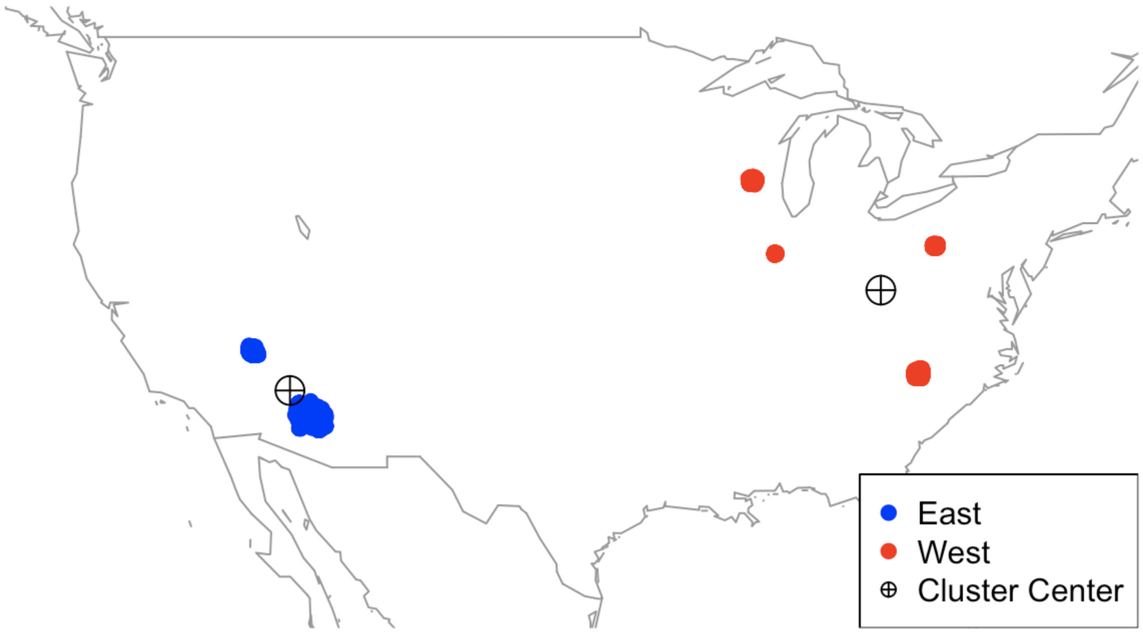
```
map("world", ylim=c(25,50), xlim=c(-130,-70), col="gray60")
points(businessfUSRst$longitude, businessfUSRst$latitude, pch = 19, col = 2)
```



Determine East and West Regions

It can be quite objective when we determine whether one business is in the east or west region of USA. However, in this report, I think it's quite clear by the map plot in previous section. I used *kmeans* algorithm to cluster the businesses into two clusters.

```
kcluster <- kmeans(cbind(businessfUSRst$longitude, businessfUSRst$latitude), 2)
businessfUSRstArea <- businessfUSRst
businessfUSRstArea$Area <- factor(kcluster$cluster, labels = c("East", "West"))
businessfUSRstEast <- subset(businessfUSRstArea, Area == "East")
businessfUSRstWest <- subset(businessfUSRstArea, Area == "West")
map("world", ylim=c(25,50), xlim=c(-130,-70), col="gray60")
points(businessfUSRstEast$longitude, businessfUSRstEast$latitude, pch = 19, col = "blue")
points(businessfUSRstWest$longitude, businessfUSRstWest$latitude, pch = 19, col = "red")
points(kcluster$centers, pch = 10, col = "black", cex = 2)
legend("bottomright", pch = c(19, 19, 10) , col = c("blue", "red", "black"), legend = c("East", "West", "Cluster
 Center"))
```



Methodology to Answer the Question

By here, the answer to the question will be simplified to just perform a statistical inference on the review scores from the two regions. We know the distribution of the review scores and in this report I set the higher average review score meaning higher chance for a new restaurant to receive higher review socre.

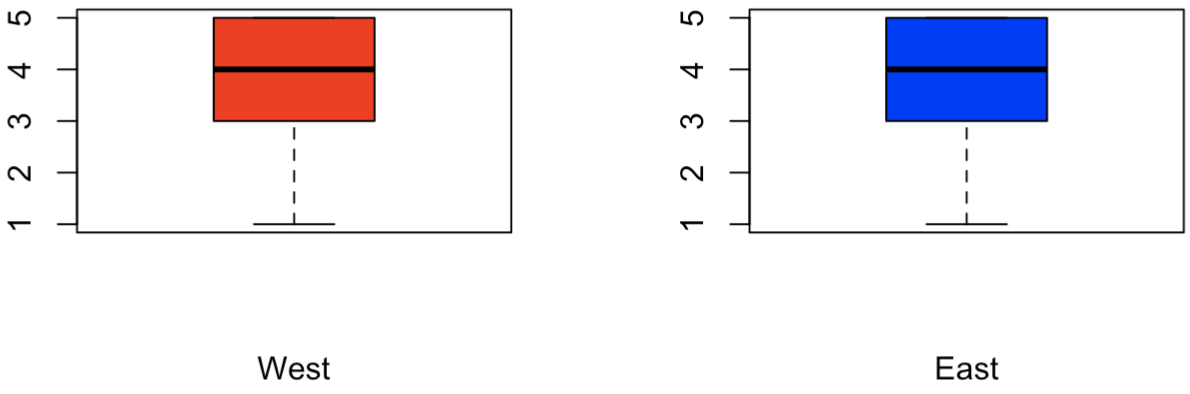
Results

Now we know which region the business is from. Next step will be linking it with the *review* dataset. For each of the business review we determine the region by mapping its *business_id* to the dataset created earlier. So the new column *Area* in the *review* dataset will be with vaule of either “East”, “West” or “NA”.

```
businessIndex <- match(reviewf$business_id, businessfUSRstArea$business_id)
reviewfArea <- reviewf
reviewfArea$Area <- businessfUSRstArea[businessIndex,]$Area
```

The final two sets of data will be the review score for the restaurants in the east and west of USA, namely *eastStars* and *westStars* in this report. Let’s take a look by plotting them.

```
eastStars <- subset(reviewfArea, Area == "East")$stars
westStars <- subset(reviewfArea, Area == "West")$stars
par(mfrow = c(1,2))
boxplot(westStars, col = "red", xlab = "West")
boxplot(eastStars, col = "blue", xlab = "East")
```



Quit surprising that despite the difference on consuming power from the two regions, the pattern how people review their restaurants look the same. And let’s do a more regurious t test. Null hypothesis is that the average review scores are the same.

```
t.test(eastStars, westStars)
```

```
##
##  Welch Two Sample t-test
##
## data:  eastStars and westStars
## t = 8.9442, df = 224640, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.02390519 0.03732222
## sample estimates:
## mean of x mean of y
##  3.721167  3.690553
```

Now here comes the funny part. The p value is so small that we'll reject the null hypothesis to conclude the review scores are different. But how much difference? It's just 0.03 in the rating system of 5. I prefer treating them as the same.

Discussion

So all in all, my conclusion is the chance of getting higher review score is the same for a restaurant in east and west of United States. If one wants to improve the review score, definitely needs to consider about other factors, like pricing strategy, whether having wifi and etcs. This will be not be covered in this report as mentioned earlier.

There is one more point I want raise up in this report. Statistically the restaurants in the west receive higher review score with strong supporting evidence due to the very small p value. However, 0.03 is really negligible in the rating system 5, at least from my point of view.