Stat 133 Spring 2017 Final Project

Yelp Data: Exploring Seasonal Trends within User Reviewing Patterns

Yelp Me Pls -- Carmelia Muljadi, Madeline Wu, Brenda Zhang

I. Introduction

Yelp is a widely used platform that crowdsources business information and customer reviews. The product currently has 135 million monthly visitors and 120 million reviews, across 32 different countries. Yelp's primary user facing features are two-fold: 1) providing customers with access to a well-maintained database of business data and 2) allowing customers to interact with businesses and inform other users by rating their experiences via public reviews. Not only is this review information useful for customers, it is also extremely beneficial for businesses to know and analyze, as the reviewing behavior can essentially be viewed as user behavior data. With the increasing usage of data-driven business models and marketing plans, this Yelp data is an invaluable source of free, user data. Therefore, by analyzing user's reviewing patterns, we can utilize any potential trends and insights found to help businesses make better and more informed choices about how to change their business trajectory.

Yelp has historically put on its own dataset challenge in order to crowdsource this data analysis effort. With such an extensive dataset and so many avenues for exploration, they provide several starter ideas for data exploration and provide raw data files. One of the suggested topics of exploration was exploring "Seasonal Trends" -- determining what seasonal effects there are on businesses or if they exist at all. Our group thought that examining seasonal trends was the most interesting topic out of all the ones provided.

After taking a brief look at the data, we narrowed the scope of our exploration to cover the following main research questions:

1. **What business categories are most popular (are reviewed the most) during each season? Which categories are most affected by seasonal trends?**
2. **For a given business/category, how do ratings change with the seasons?**
3. **How does reviewing behavior change across seasons for businesses in the United States?**

Before proceeding with the data analysis, we had to define how we quantified business popularity, business success, and reviewing behavior. A business's popularity is represented by the number of reviews it received in the sample time period. A business's success is quantified by the average numerical star rating it received in the sample time period. We quantified reviewing behavior in a region by looking at the

number of review and/or rating entries. Although there was data related to check-ins, another way that users could interact with a business, we decided that we were unable to quantify and weigh check-ins in a standard manner when considering a business's popularity, success, and the aggregate user behavior.

Another important thing to note is the method by which we chose our categories of interest. When we assessed how ratings change with the seasons, we chose to examine a few specific categories by filtering our data set down to a subset of the categories -- food, beauty, shopping, and home. We chose these four categories based on the fact that they were the categories with the most respective review data points. We wanted to ensure that we had the largest data sets possible when examining specific categories for their seasonal trends, so we ended up with the four categories.

We hypothesized that we would find some seasonal trend amongst the business data. We were looking to identify whether there was a specific season or time period during the year where certain categories had increased popularity, through the form of higher review count. We thought this would be important to determine because of its practical, real life business applications. Knowing this information could help businesses to create more efficient staffing and stocking plans and they could better optimize to meet the demand.

II. Data Collection

Yelp provided the academic data set for free and we attained it via their website. The data was collected over the following time period, 2004-2015. It contained 4.1M reviews and 947K tips by 1M users for 144K businesses, 1.1M business attributes, aggregated check-ins over time for each of the 125K businesses, and 200,000 pictures from the included businesses. The data we decided to use included data on all the businesses, all the reviews, and all of the checkins. The latter two data sets contained multiple entries per business. The relevant data was available as three json files: yelp_academic_dataset_checkin.json (45 MB), yelp_academic_dataset_business.json (0.9 GB), and yelp_academic_dataset_review.json (2.9 GB).

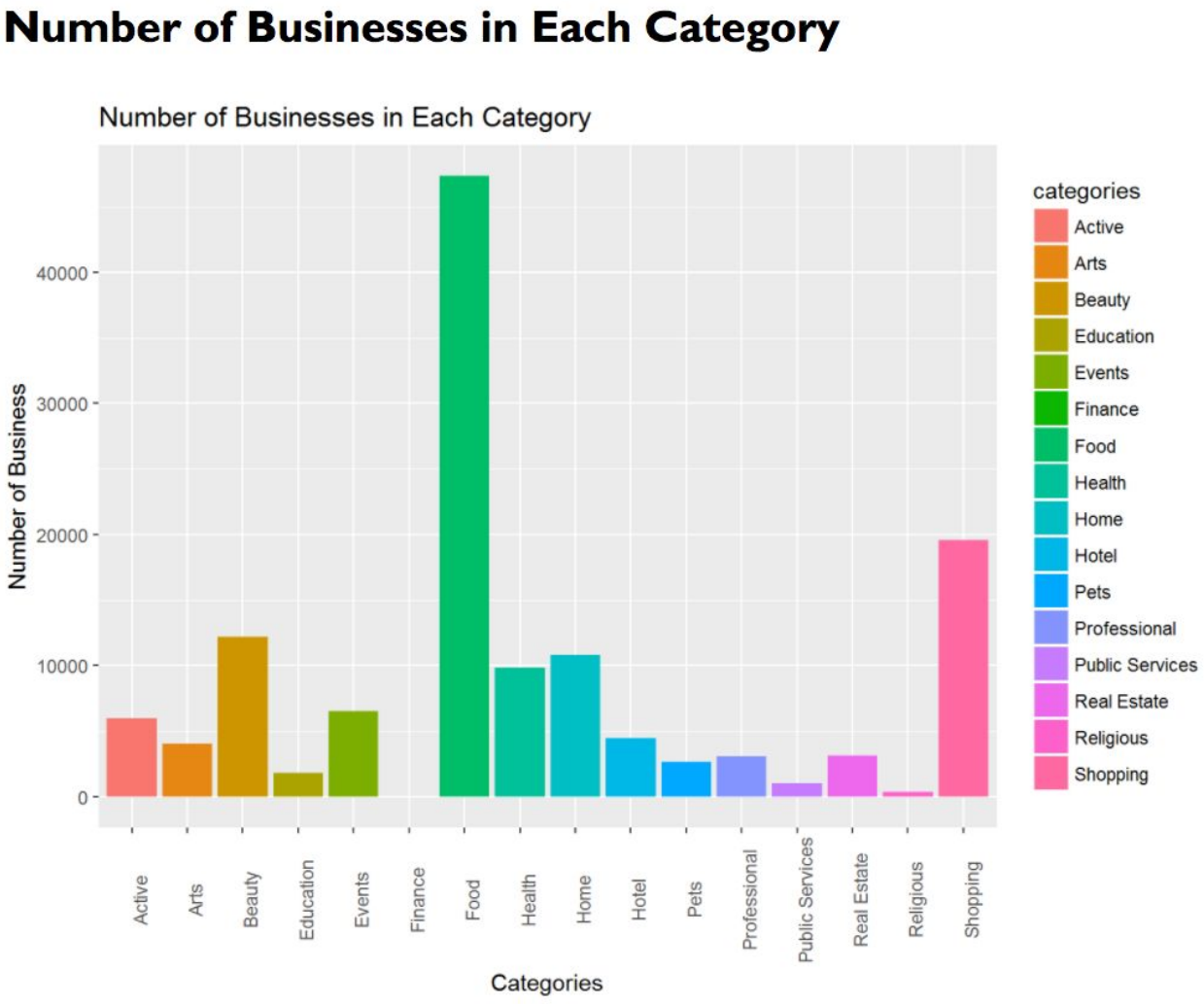
We decided to use the data on businesses and reviews because they were most relevant to addressing the questions we wanted to answer. We excluded the data related to check ins because it did not benefit our analysis. The relevant columns from the business table are business_id, categories, longitude, latitude, city, and review_count. The relevant columns from the review table are business_id, date and stars. The remainder of the categories were unnecessary and removed from the data table because they did not have relevance with the analysis we were doing.

Since the original data was in json format, we had to convert from json to csv, because csv files were what we learned to use in the course. Luckily, the Yelp dataset

challenge provided a link to a Python script that could convert from one format to the other. This allowed us to easily analyze the data, even though it was not initially in a form we were used to. The most rudimentary level of data we wanted to have involved rows that associated each rating with the appropriate business. We obtained this data by joining the business table with the review table, using the business_id as the key. This created the single table that we would use for the rest of our analysis because it combined all of the necessary data needed for our visualizations.

III. Analysis and Data Visualization

i. What business categories are most popular (are reviewed the most) during each season? Which categories are most affected by seasonal trends?

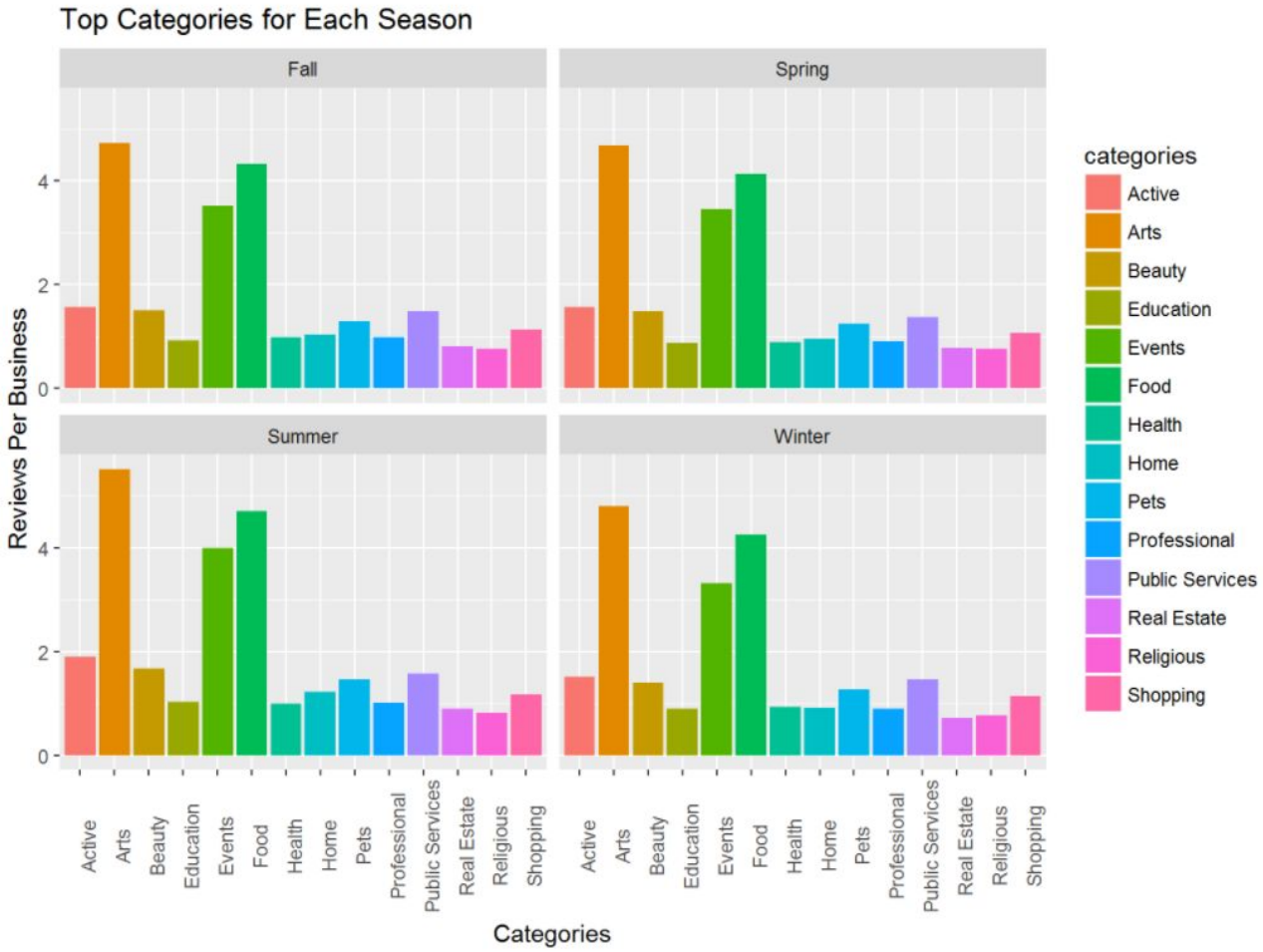


Before we could begin answering our research questions, we had familiarize ourselves with the data. There are 16 unique business categories that a business could

be placed into. Each business could be categorized by one or more categories. In order to assess how many businesses were in each category, we had to do a bit of data wrangling. This required us to use the *grepl* function to filter out rows that contained the category name within the categories column, for each of the 16 categories. This ensured that a given business was accurately represented in all the categories it fell into. We created separate tables for each category and used these resulting tables to calculate the number of businesses within each category. We combined all of this data in a table that contained the number of businesses for each category.

We visualized this data in a bar chart, shown above and found that the top four categories with the most businesses were food, shopping, beauty, and home. Thus, we decided to do our category-specific explorations on those categories, as they had sufficiently large enough data sets to make conclusions from.

Top Categories Per Season

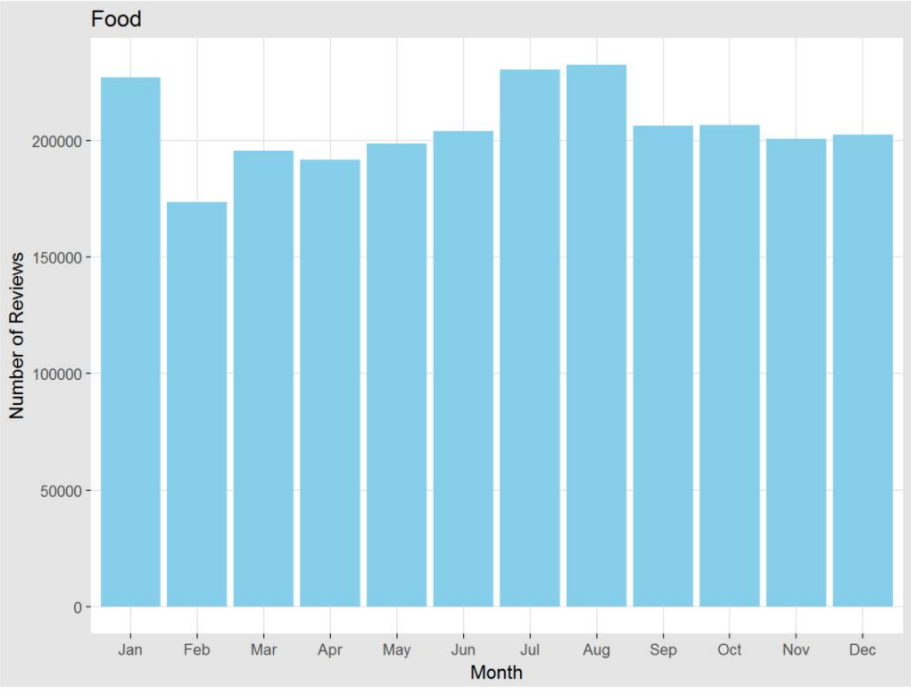


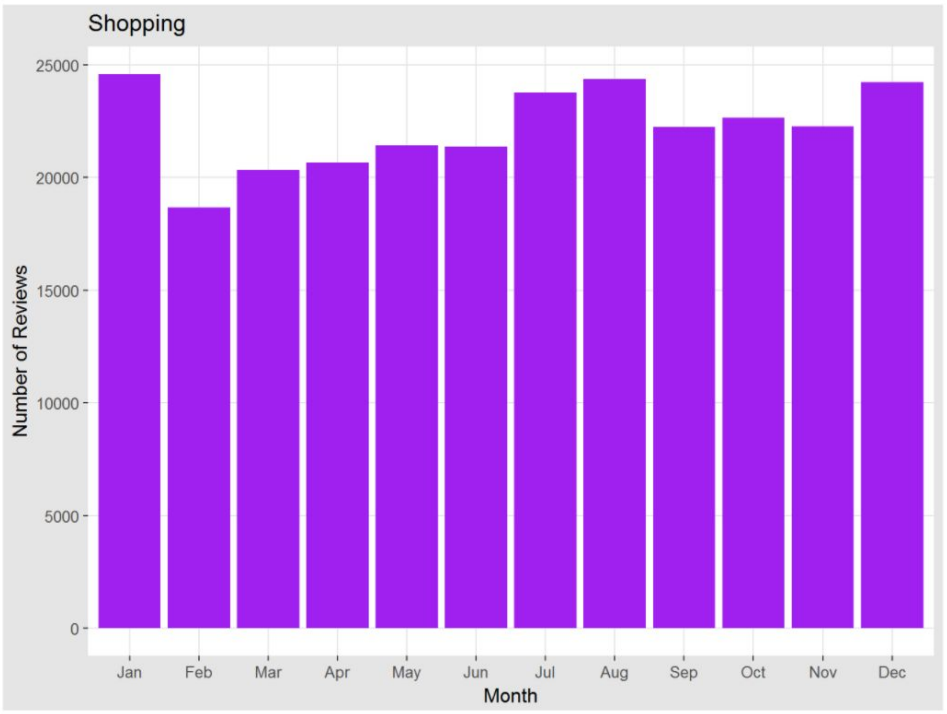
From the previous graph, we found the categories with the largest business data sets. We also wanted to see how categories performed differently among seasons. More specifically, which categories perform better in each season, in terms of sheer volume of reviews. We also wanted to know reviewing behaviour changes between seasons among the different categories. This graph displays the number of reviews per business for each category, faceted by the seasons.

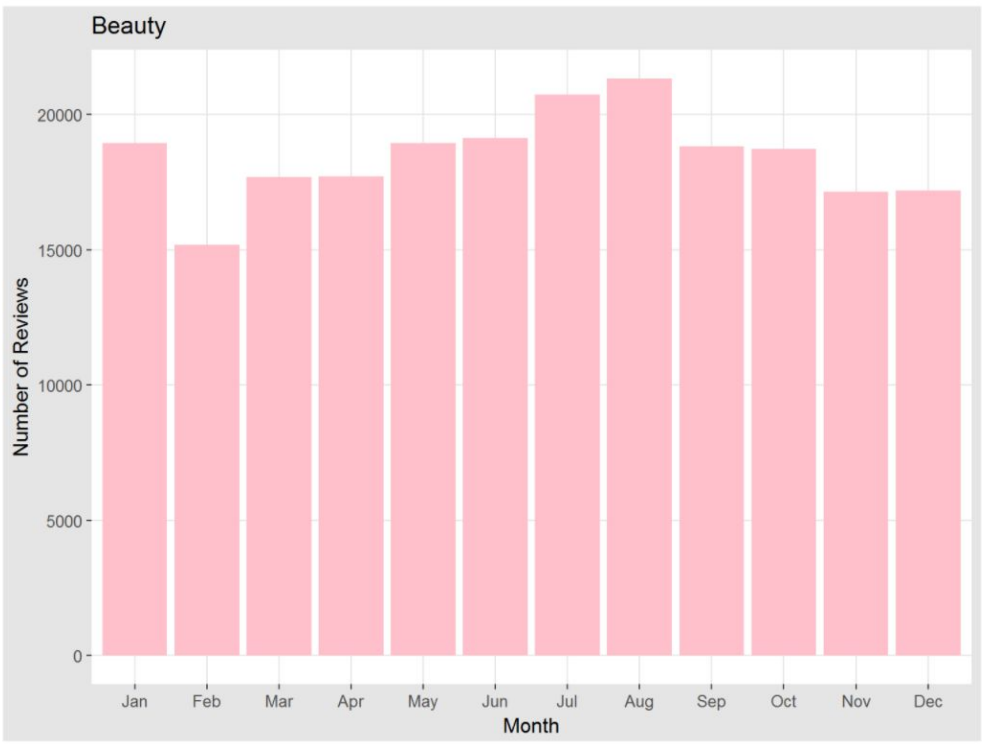
We started off with the cleaned data tables for each category. We used each of these data tables to attain the total number of businesses and as well of the variable we wanted to plot on our graph: total number of reviews per business. We initially thought about plotting the number of reviews. However, we decided to scale it by the number of the size of the category, by dividing by the number of businesses. Otherwise, our visualization would be skewed because the categories with a larger number of businesses would, naturally, have a larger number of reviewers. To find the reviews per business, we added up the total reviews for each category. We then divided this number with the total number of businesses in each category.

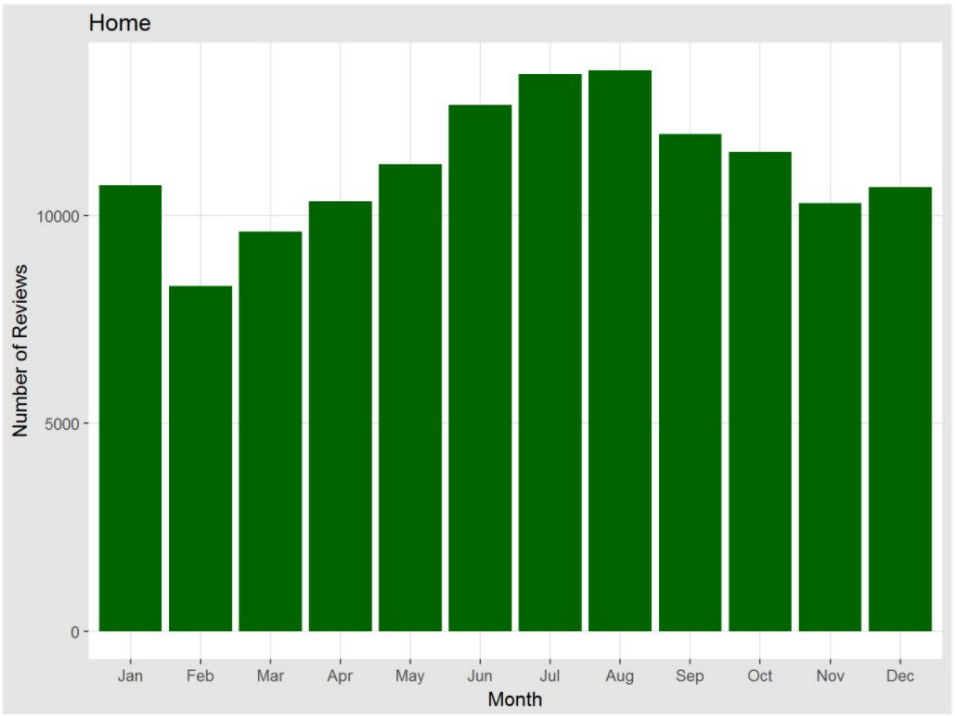
The Arts, Events and Food categories consistently have more reviews per business than other categories. The number of reviews per business does not change across the seasons, suggesting that choosing the same four categories for our analysis would produce consistent results. This did not demonstrate any seasonal trends in terms of reviewing behavior. This was not unexpected, because we were using the entire data set, so we did not anticipate to see large variation in overall customer reviewing behavior over time. We decided to look at the top categories in more detail.

Number of Reviews Over Time for Each Category




Shopping
25000
20000
15000
10000
5000
0
Number of Reviews
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month


Beauty
20000
15000
10000
5000
0
Number of Reviews
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month



These graphs plot the number of reviews over time, for each of the categories we have selected: food, home, beauty, and shopping. These plots were made by grouping the specific category's business data by month and summarizing by the total number of reviews within the category.

We used the original cleaned dataset and filtered out the following categories: food, beauty, shopping, and home. For each category, we grouped the data by month, and summarized by sum. We mapped each numeric month value to its string representation in a data table, and joined this with our data to make the data glyph-ready. To visualize, we used geom_col to create a bar chart and represent each month's number of reviews as a bar. We chose to visualize it in a bar because we could see trends across the year. We faceted the graph by category and assigned each category to a different color.

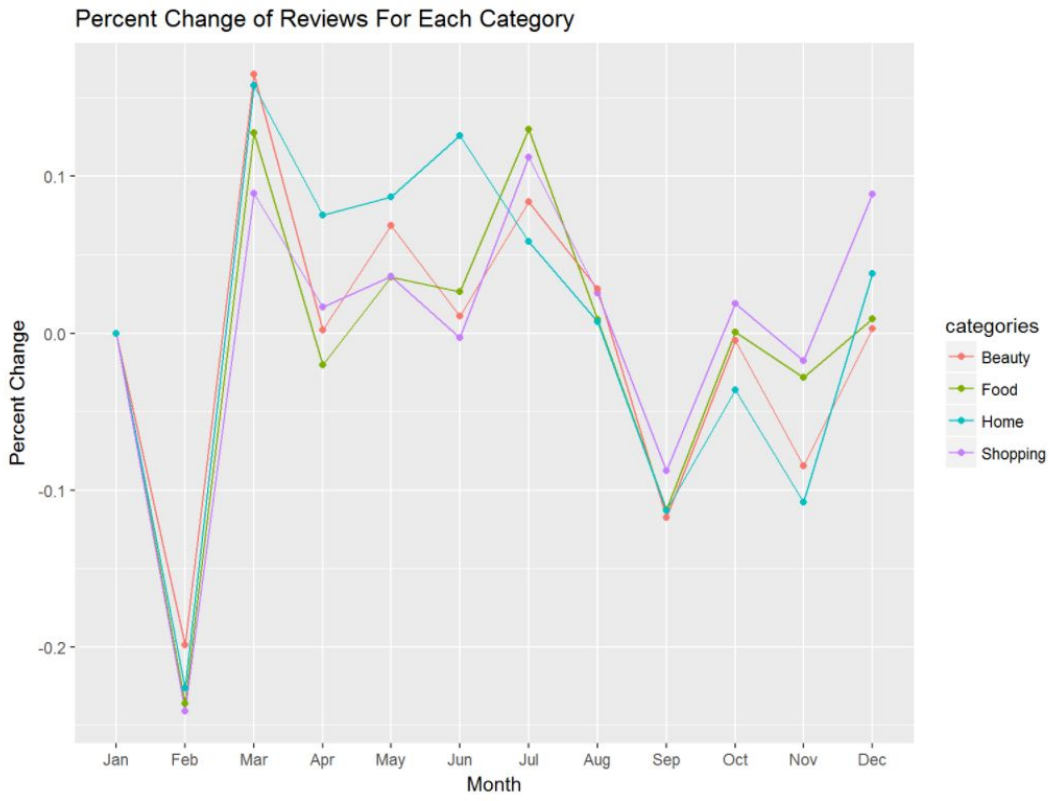
In this particularly analysis, the number of reviews is representative of a category's popularity and business success. If more users are reviewing businesses from a given category, then that means that they are engaging with the business and visiting the business more, which is a positive indication of popularity and success. We had several findings from these set of graphs. The first involves comparing the categories' trends to each other and then second involves analyzing the general trend itself.

All of the categories' bars seem to follow the same general trend, meaning that the popularity does not deviate across categories. Certain categories, at least the ones we chose to examine, are not more or less susceptible to changes in time. Category

does not seem to be a predictor for trend variance. Given this finding, we would not advise food businesses to do anything differently from shopping businesses, when accounting for time as a factor.

However, there was a significant finding in regards to the general trend. For all categories, we found that the months of January, July, and August had a particularly high number of reviewers. We hypothesize that this seasonal trend exists because of the summer months and holiday season. The number of reviews is correlated to number of customers visiting a business; a higher number of reviewers is preceded by and associated with more customers going to that business. By this explanation, the increased spike in reviews during the July and August could be due to the fact that tourism and vacationing is higher during these summer months. During this "summer vacation" time period, tourists are inclined to spend more and frequent businesses more often. January occurs after the holiday season, where many people are going to shopping businesses and eating at restaurants. The increased reviewing behavior could be a result of the increased business activity due to the holiday season.

Percent Change Per Month



This graph plots the percent change between each month and the previous month for the beauty, food, home, and shopping categories. We utilized the same data

we used in the previous data analyses. The only data wrangling involved was for the visualization. It required us to write a function to calculate the percent change and apply the function to all the rows in each of the category tables. Afterwards, we plotted the percent change against the month, making sure to use geom_line and geom_point to visualize the trend in an appropriate manner.
The percent change was calculated by the following standard percent change equation:

$$\% \; change = \frac{reviews_{currmonth} - reviews_{prevmonth}}{reviews_{prevmonth}}$$

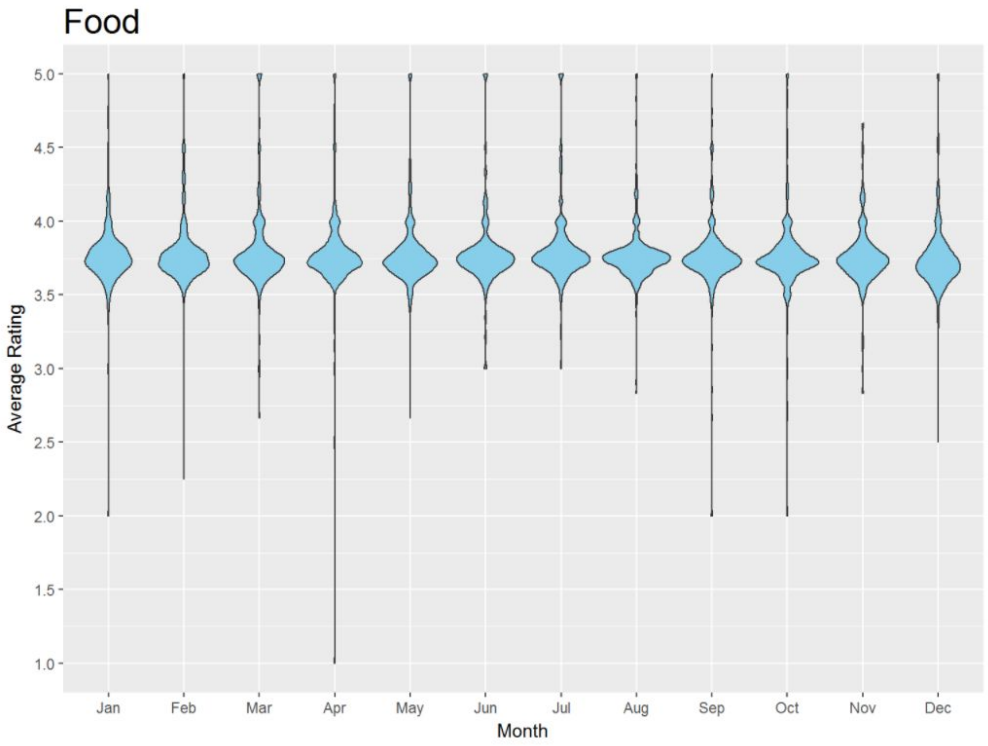
All the categories had the same general trend in the percent change in number of reviews. There is a decrease in reviews in February and a peak in March. There is almost no change from April to June except for the Home category. Reviewing behaviour peaks in the summer months, with all categories demonstrating an increase in reviews. There is also a decrease in reviews for all categories during the fall season, from August to November. The trends could be because businesses tend to promote themselves during the holidays - namely during the winter and summer months. This increased activity could explain the peak in reviews. This could also explain the dips in the months before summer, when there's a lull in shopping activity. The Home category had slightly greater fluctuations in percent change than the other categories. This could be because there aren't as many reviews for Home category compared to the other categories, so the percent change between months is larger.

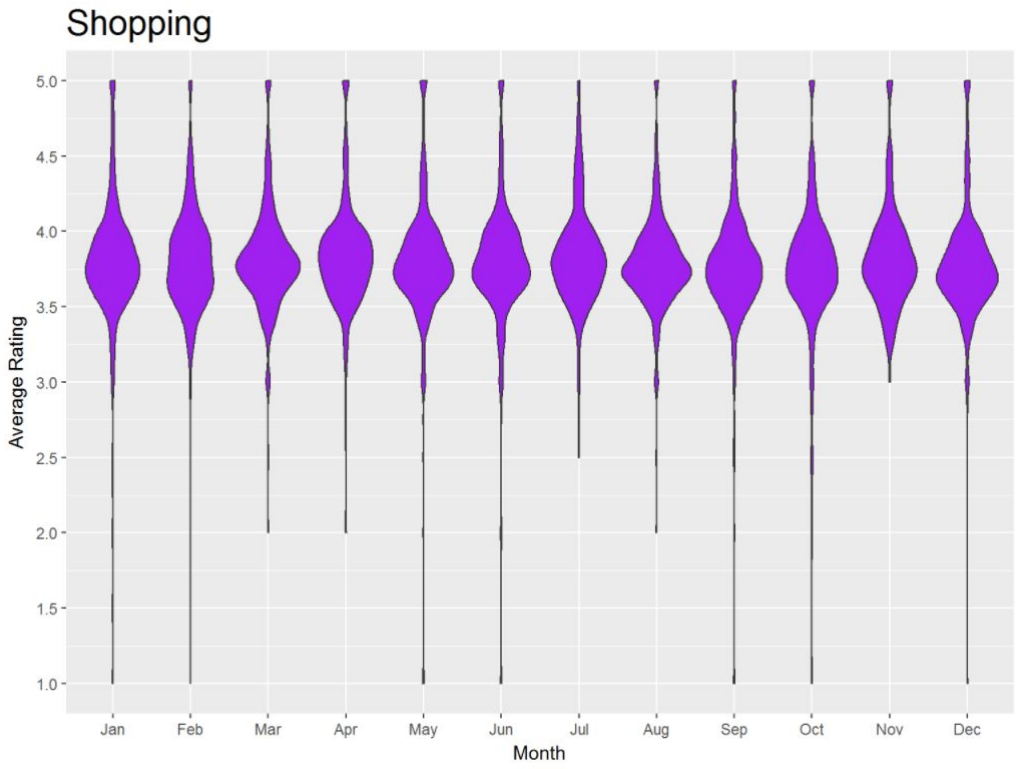
ii. How do the business's ratings change with the seasons?

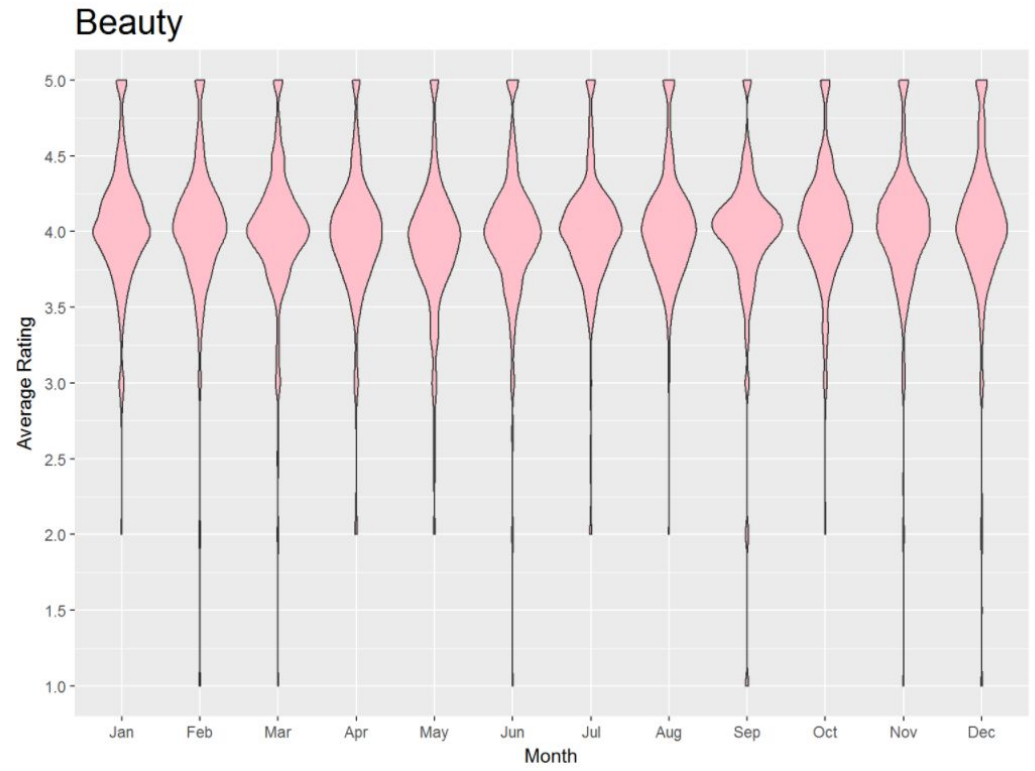
We made a violin plot that explore the how the average rating changes over time, for a given category. The four visualizations were made by grouping by the review's month and aggregating the rating values, thus providing a visual on the average rating value and the spread. We used the original cleaned dataset and filtered out the following categories: food, beauty, shopping, and home. For each category, we grouped the data by month, and summarized by the average rating. We mapped each numeric month value to its string representation to make the data glyph-ready.

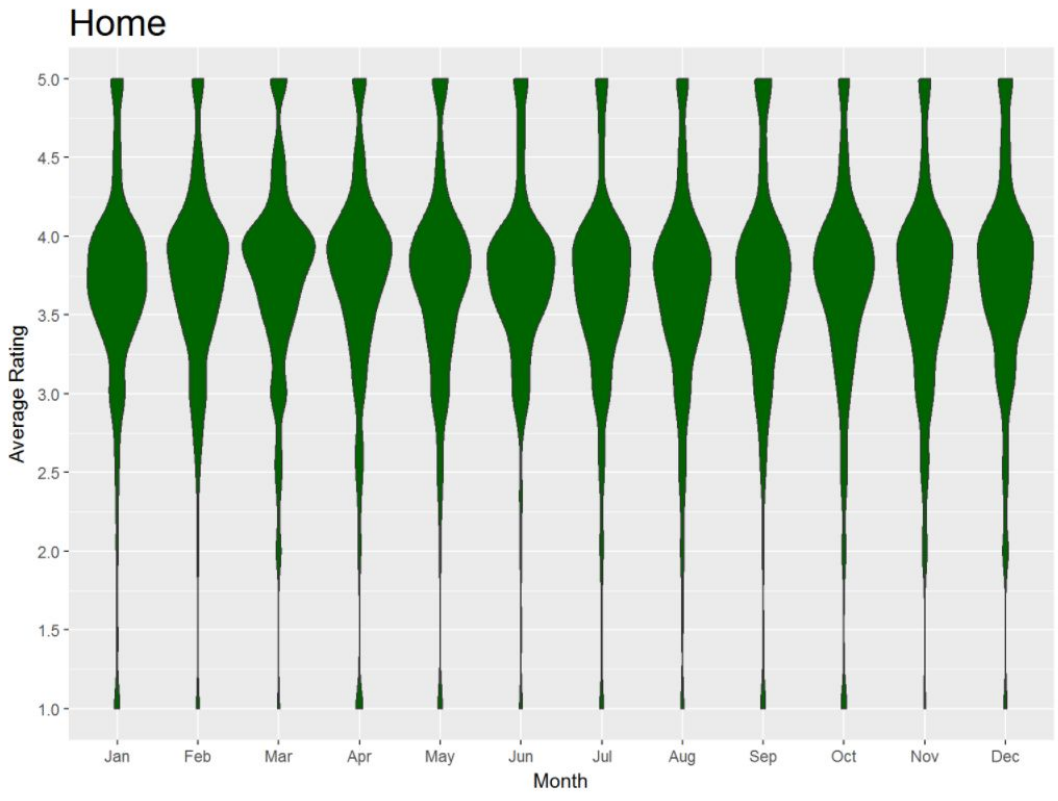
The statistic in question was the average rating for the category's business. A business's rating was another way to indicate how popular a business was, in addition to how many reviews it received over all. The rating quantifies the customer's satisfaction level with the business in a review. Therefore, the plot for each category visualized how much this satisfaction changed over the months. For this analysis, it is important to note that we were interested solely in the absolute change in rating, not the rating value itself. We are not comparing each category's popularity, but rather we are looking for whether popularity trends exist and if they do, comparing them across categories. This difference is subtle but crucial to mention as part of our analysis.

Average Rating Over Time for Each Category



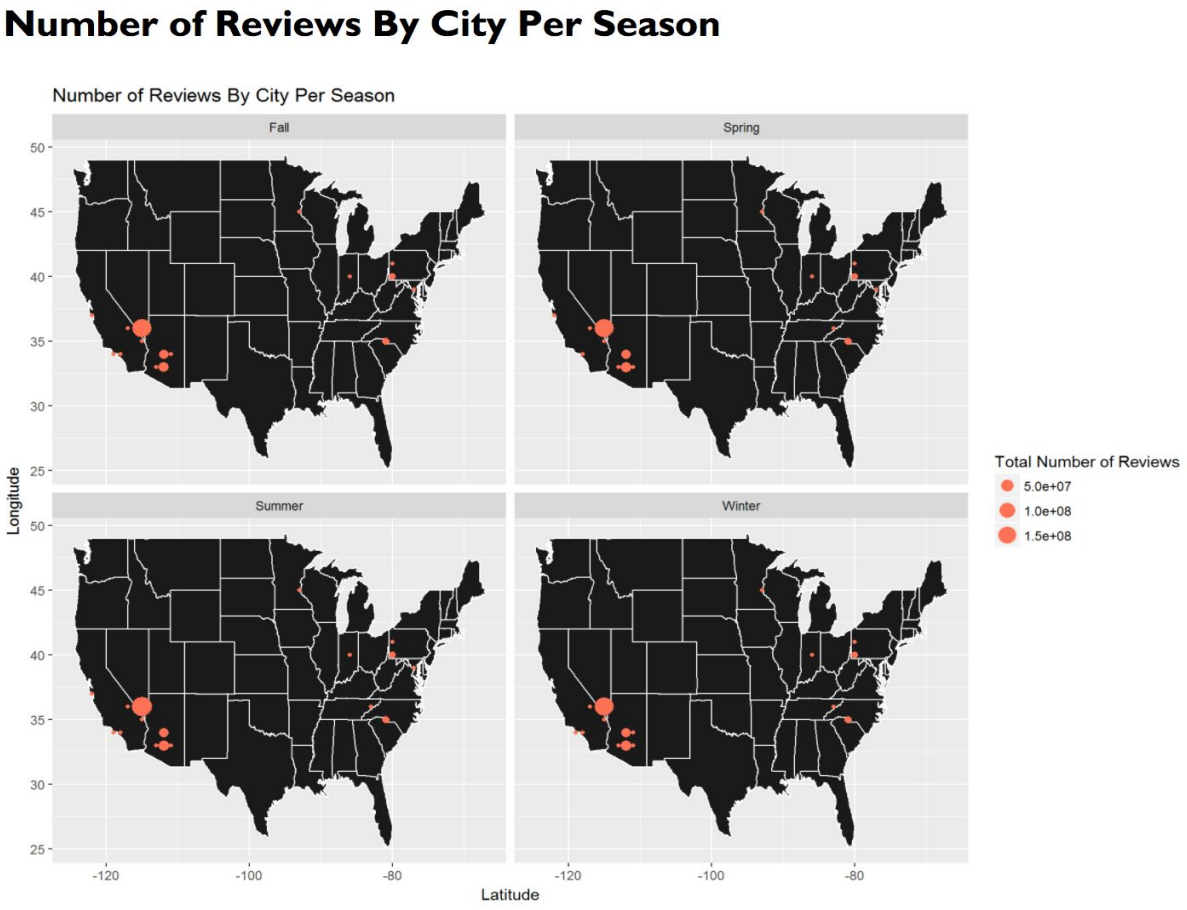



Beauty
Average Rating
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month


Home
Average Rating
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

For each category, there does not appear to be a difference in how the average rating changes across the months. The value that it is centered at does not change and neither does the spread. Across categories, there is no correlation between month and average rating, indicating that these variables are independent. This means that certain seasons do not cause customers to be any more or less satisfied with businesses. Originally, we had hoped to find some sort of trend in the data; however, in retrospect our hypothesis was based off of the wrong definition of popularity. In actuality, a certain month should not cause a customer's satisfaction levels to increase or decrease. It would make sense for a month to increase a customer's interest, but customer satisfaction and business likability are relatively less linked to time. For example, during holiday months in the winter, it is logical to predict that clothing stores to experience a greater number of reviews, but less intuitive to predict that people will like clothing stores more during these months. The latter is the conclusion that these four violin plots attempted but failed to support. Across the categories, there are differences in the rating values themselves, but again, we are not interested in these raw values.

<u>iii. Within the United States, does reviewing behavior change over the seasons?</u>



This graph displays total number of reviews for each city, faceted by season. This graph aims to show how the volume of reviews will change across seasons. We limited the scope of this analysis to a select few city regions: New York City, Las Vegas, Chicago, Pittsburgh, Charlotte, Phoenix, San Francisco, Los Angeles, Houston, and Boston.

We filtered our original cleaned dataset to only include cases from the select cities. An important data wrangling exercise we had to include was changing the coordinate data. Since we are visualizing the review by location, we want to group our cleaned data by location. Originally the latitude and longitude values have many significant digits, making it impossible to group and achieve the results we wanted. Since each business is in a unique location, grouping by the original latitude and longitude combinations would not yield a different data table. We decided to round each latitude and longitude value to the nearest whole number. This would categorize all businesses into its nearest latitude and longitude values, and the summarized table would be able to group over several rows for a given region. When we group, we are summarizing by the count.

To visualize, we plotted geom_points for each region, scaled by how many reviews there were in that region. The larger the point, the more reviews there were in that location. The data was intuitively plotted with the longitude on the x-axis and latitude on the y-axis. We then used geom_polygon to overlay a map and provide a geographical dimension to the visualization. The data were faceted by season. This resulted in a wrapped plot of four map visualizations for the four seasons.

As seen from the visualizations, the trend seems to be constant across the seasons. In general, there is not a large amount of variation in the number of reviews per city. Although we hoped to see some kind of difference in our statistic across the geographical area, the lack of trend can be explained. This means that overall, seasons don't significantly affect the frequency and volume at which customers review businesses overall. General reviewing behavior on Yelp is not a function of time of year or season. This is still useful information to know because this means that users behave similarly across the different locations and geographical factors do not have a strong correlation with reviewing behavior. This can be explained by the nature of the Yelp product itself. It is a standard product, regardless of what geographical location the user is in. The inherent product does not encourage more or less use depending on the user's location. Additionally, businesses operate in a relatively standardized manner across the country, so there does not seem to be anything in particular on the business end that would encourage a certain geographic region to have an abnormally large or small amount of reviews.

IV. Conclusion

Through this data analysis we hoped to determine whether there are general seasonal trends in Yelp business data. We were specifically interested in looking at which business categories' reviews were most associated with different seasons, how ratings changed across time, and how geographical location played a role in reviewing behavior.

After determining which categories had the largest datasets, we narrowed down our category-specific analysis to four categories: food, beauty, home, and shopping. We found that overall, there were no changes across categories and the correlation between time and all other variables of interest was close to 0. This was true for both reviewing behavior and rating values. However, we did find that there was a general trend in reviewing behavior and time, with more reviewing happening in the months of January, July, and August. Lastly, we were able to eliminate geographical location as a factor that affects reviewing behavior because we did not find an association between those two variables, when we looked at trends over time.

Many of the results we obtained did not demonstrate a significant correlation or trend. Therefore, this eliminates seasonality and temporal changes as a factor in determining user behavior. If businesses want to encourage more user behavior and increase their popularity, now they know that they should not depend on time of year to play a role in the factors that they choose to optimize on. This allows them to focus their analysis on other factors and methods of increasing their business and reaching customers.

However, we still have not explored the full scope of data in regards to seasonal trends. Our report provides a very generalized analysis on business data. If we had more time, we have many other areas of interests that we would like to continue digging into. We would have wanted to probe more specific research questions that were focused on more niche industries. For example, how ice cream shops, department stores, tax preparation companies are affected by seasons. Some intuitive hypotheses we had for these cases are that ice cream shops are more heavily reviewed in the spring and summer, while department stores have a spike in the winter months because of the holiday season, and tax preparation companies see a steady growth in popularity through winter and carrying through the spring, to align with tax season. These were all questions we were unable to answer with our current analysis because these kinds of businesses are more specific than the general home, food, beauty, etc. categories that our original data was categorized into. However, these are all important research questions to delve into, in order to help businesses make more informed decisions based on Yelp user data.

V. Appendix

R code

```
#import and read csv files into R

business <- read_csv("C:\\Users\\Brenda\\Documents\\UC Berkeley\\Sophomore\\Spring 2017\\STAT 133\\Final Project\\dataset-examples-master\\yelp_academic_dataset_business.csv")

review <- read_csv("C:\\Users\\Brenda\\Documents\\UC Berkeley\\Sophomore\\Spring 2017\\STAT 133\\Final Project\\dataset-examples-master\\yelp_academic_dataset_review.csv")
```

Initial Data Wrangling

```
# cleaning data to make business_clean and review_clean

business_clean <- business %>%
  select(business_id, city, stars, review_count, is_open, categories, latitude,longitude) %>%
  filter(is_open == 1) %>% # filtering out businesses that are no longer in business
  select(business_id, city, stars, review_count, categories, latitude, longitude) %>% # selecting relevant columns
  rename(avg_stars = stars)

review_clean <- review %>%
  select(business_id, date, stars) %>% # select relevant columns
  rename(review_stars = stars)

# Making separate business data tables for each of the 16 categories

arts <- business_clean %>% filter(grepl("Arts & Entertainment", categories))
nrow(arts)

active <- business_clean %>% filter(grepl("Active Life", categories))
nrow(active)

food <- business_clean %>% filter(grepl("Food|Restaurants", categories))
nrow(food)

beauty <- business_clean %>% filter(grepl("Beauty & Spas", categories))
nrow(beauty)
```

```
edu <- business_clean %>% filter(grepl("Education", categories))
nrow(edu)

event <- business_clean %>% filter(grepl("Event Planning & Services", categories))
nrow(event)

fin <- business_clean %>% filter(grepl("Financial Planning & Services", categories))
nrow(fin)

health <- business_clean %>% filter(grepl("Health & Medical", categories))
nrow(health)

home <- business_clean %>% filter(grepl("Home Services", categories))
nrow(home)

hotel <- business_clean %>% filter(grepl("Hotels & Travel", categories))
nrow(hotel)

pets <- business_clean %>% filter(grepl("Pets", categories))
nrow(pets)

prof <- business_clean %>% filter(grepl("Professional Services", categories))
nrow(prof)

publicserv <- business_clean %>% filter(grepl("Public Services & Government", categories))
nrow(publicserv)

realestate <- business_clean %>% filter(grepl("Real Estate", categories))
nrow(realestate)

relig <- business_clean %>% filter(grepl("Religious Organizations", categories))
nrow(relig)

shopping <- business_clean %>% filter(grepl("Shopping", categories))
nrow(shopping)

# Add an additional column with the category name for each table
food1 <- food %>% mutate (categories = "Food")
shopping1 <- shopping %>% mutate (categories = "Shopping")
beauty1 <- beauty %>% mutate (categories = "Beauty")
home1 <- home %>% mutate(categories = "Home")
arts1 <- arts %>% mutate(categories = "Arts")
```

```
active1 <- active %>% mutate(categories = "Active")
edu1<- edu %>% mutate(categories = "Education")
event1 <- event %>% mutate(categories = "Events")
fin1 <- fin %>% mutate(categories = "Finance")
health1 <- health %>% mutate(categories = "Health")
hotel1 <- hotel %>% mutate(categories = "Hotel")
pets1 <- pets %>% mutate(categories = "Pets")
prof1 <- prof %>% mutate(categories = "Professional")
publicserv1 <- publicserv %>% mutate(categories = "Public Services")
realestate1 <- realestate %>% mutate(categories = "Real Estate")
relig1 <- relig %>% mutate(categories = "Religious")

# Business table for food, shopping, beauty, and home businesses, with updated category
column
business_filtered <- rbind(food1,shopping1,beauty1,home1)

# Business table for all categories, with updated category column
business_all <- rbind(business_filtered, arts1, active1,edu1, event1,fin1,health1,hotel,
pets1,prof1,publicserv1,realestate1,relig1)

# Making a table mapping each month to season
seasons <- data.frame(season =
c("Winter","Winter","Winter","Spring","Spring","Spring","Summer","Summer","Summer","Fall","Fall","F
all"), month = c(12,1,2,3,4,5,6,7,8,9,10,11))
seasons

# Making table mapping each month number to a month name
months <- data.frame(month = seq(1,12), name = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul",
"Aug", "Sep", "Oct", "Nov","Dec"))
months

# Create fundamental data set for filtered categories, associating each review with its business
data <- business_filtered %>%
  left_join(review_clean, by="business_id") %>%
  select(business_id, city, categories, date, review_stars)

# Create fundamental data set for all categories, associating each review with its business

data_all <- business_all %>%
  left_join(review_clean, by="business_id") %>%
  select(business_id, city, categories, date, review_stars)
```

Answering research question #1

```
# Creating a two column data table of categories and their number of associated businesses

categ_count <- data.frame(
  categories = c("Arts", "Active", "Food", "Beauty", "Education", "Events", "Finance", "Health",
"Home", "Hotel", "Pets", "Professional", "Public Services", "Real Estate", "Religious", "Shopping"),
  count = c(nrow(arts), nrow(active), nrow(food), nrow(beauty), nrow(edu), nrow(event),
nrow(fin), nrow(health), nrow(home), nrow(hotel), nrow(pets), nrow(prof), nrow(publicserv),
nrow(realestate), nrow(relig), nrow(shopping))


# Create a visualization that plots the count for each category
categ_count_bar <- categ_count %>% ggplot(aes(x = categories,y = count)) + geom_col(aes(fill =
categories)) + labs(title = "Number of Businesses in Each Category", x = "Categories", y="Number
of Business") + theme(axis.text.x = element_text(angle = 90))

# Creates a table mapping each month to the number of reviews for each category

food_joined <- data %>%
  filter(categories == "Food") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month")
View(food_joined)

shopping_joined <-data %>%
  filter(categories == "Shopping") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month")

beauty_joined <- data %>%
  filter(categories == "Beauty") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month")

home_joined <- data %>%
  filter(categories == "Home") %>%
```

```
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month")

# Creating individual tables for each category: month, number of reviews, category

food_joined_all <- data_all %>%
  filter(categories == "Food") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Food")

shopping_joined_all <-data_all %>%
  filter(categories == "Shopping") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Shopping")

beauty_joined_all <- data_all %>%
  filter(categories == "Beauty") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Beauty")

home_joined_all <- data_all %>%
  filter(categories == "Home") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Home")

arts_joined_all <- data_all %>%
  filter(categories == "Arts") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
```

```
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Arts")

active_joined_all <- data_all %>%
  filter(categories == "Active") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Active")

edu_joined_all <- data_all %>%
  filter(categories == "Education") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Education")

event_joined_all <- data_all %>%
  filter(categories == "Events") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Events")

fin_joined_all <- data_all %>%
  filter(categories == "Finance") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Finance")

health_joined_all <- data_all %>%
  filter(categories == "Health") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Health")
```

```
hotel_joined_all <- data_all %>%
  filter(categories == "Hotel") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Hotel")

pets_joined_all <- data_all %>%
  filter(categories == "Pets") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Pets")

prof_joined_all <- data_all %>%
  filter(categories == "Professional") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Professional")

publicserv_joined_all <- data_all %>%
  filter(categories == "Public Services") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Public Services")

realestate_joined_all <- data_all %>%
  filter(categories == "Real Estate") %>%
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Real Estate")

relig_joined_all <- data_all %>%
  filter(categories == "Religious") %>%
```

```
  mutate(month = lubridate::month(date)) %>%
  group_by(month) %>%
  summarize(num_reviews = n()) %>%
  left_join(months, by="month") %>%
  mutate(categories = "Religious")

# Create aggregate table of reviews per business, grouped by category and month

top_categ <- rbind(food_joined_all,
shopping_joined_all,beauty_joined_all,home_joined_all,arts_joined_all,active_joined_all,
edu_joined_all, event_joined_all, fin_joined_all, health_joined_all, hotel_joined_all, pets_joined_all,
prof_joined_all, publicserv_joined_all, realestate_joined_all, relig_joined_all) %>%
  left_join(categ_count, by = "categories") %>% # create an intermediate table with month,
number of reviews, number of businesses, category
  mutate(reviews_per_business = num_reviews/count) %>% # create column with our quantifier
  left_join(seasons, by = "month") %>%
  group_by(categories, season) %>%
  summarise(reviews_per_business = mean(reviews_per_business))

# Function that calculates percent change
calc_per_change <- function(x) {
  change_vec <- c(0)
  for (i in 1:length(x)-1) {
    change <- (x[i+1]-x[i])/x[i]
    change_vec <-append(change_vec, change)
  }
  return(change_vec)
}

# Creates a table with added column for percent change for each category

food_joined_all_change <- food_joined_all %>%
  mutate(percent_change = calc_per_change(food_joined_all$num_reviews))

shopping_joined_all_change <- shopping_joined_all %>%
  mutate(percent_change = calc_per_change(shopping_joined_all$num_reviews))

home_joined_all_change <- home_joined_all %>%
  mutate(percent_change = calc_per_change(home_joined_all$num_reviews))

beauty_joined_all_change <- beauty_joined_all %>%
  mutate(percent_change = calc_per_change(beauty_joined_all$num_reviews))
```

```
# Creates aggregate percent change table for all selected categories

per_change <- rbind(food_joined_all_change, shopping_joined_all_change,
beauty_joined_all_change, home_joined_all_change)
```

Answering research question #2

```
# Creates a table mapping each month to the category's average resulting rating during that
month

food_joined_ratings <- data %>%
  filter(categories == "Food") %>%
  group_by(date) %>%
  summarize(ave_rating = mean(review_stars)) %>%
  mutate(month = lubridate::month(date)) %>%
  left_join(months, by="month")


shopping_joined_ratings <- data %>%
  filter(categories == "Shopping") %>%
  group_by(date) %>%
  summarize(ave_rating = mean(review_stars)) %>%
  mutate(month = lubridate::month(date)) %>%
  left_join(months, by="month")

beauty_joined_ratings <- data %>%
  filter(categories == "Beauty") %>%
  group_by(date) %>%
  summarize(ave_rating = mean(review_stars)) %>%
  mutate(month = lubridate::month(date)) %>%
  left_join(months, by="month")

home_joined_ratings <- data %>%
  filter(categories == "Home") %>%
  group_by(date) %>%
  summarize(ave_rating = mean(review_stars)) %>%
  mutate(month = lubridate::month(date)) %>%
  left_join(months, by="month")
```

Answering research question #3

```
# Create summary table for number of reviews for a given location in each season, limiting data
set to certain cities
```

```
geo <- data %>%
  mutate(latitude = round(latitude, digits = 0), longitude = round(longitude, digits = 0)) %>% # cleaning coordinates
  mutate(month = lubridate::month(date)) %>% # transforming month column into glyph-ready data
  left_join(seasons, by = "month") %>% # add a column associating a review with its season
  group_by(longitude, latitude, season) %>%
  mutate(tot_numrev = sum(review_count)) %>%
  select(city,tot_numrev, season, latitude, longitude) %>%
  unique() %>%
  filter(city == "New York City"|city == "Las Vegas"|city =="Chicago"|city =="Pittsburgh"|city =="Charlotte"|city =="Phoenix"|city =="San Francisco"|city =="Los Angeles"|city =="Houston"|city =="Boston") # select desired cities
```

Visualizing Number of Businesses in Each Category

```
categ_count_bar <- categ_count %>% ggplot(aes(x = categories,y = count)) + geom_col(aes(fill = categories)) + labs(title = "Number of Businesses in Each Category", x = "Categories", y="Number of Business") + theme(axis.text.x = element_text(angle = 90))

categ_count_bar
```

Visualizing Top Categories Per Season

```
top_categ_bar <- top_categ %>% ggplot(aes(x=categories, y = reviews_per_business)) + geom_col(aes(fill=categories)) + facet_wrap(~season) + labs(title = "Top Categories for Each Season", x = "Categories", y="Reviews Per Business") + theme(axis.text.x = element_text(angle = 90))

top_categ_bar
```

Visualizing Number of Reviews Over Time for Each Category

```
food_joined %>% ggplot(aes(x = factor(name, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov","Dec")), y=num_reviews)) + geom_col(fill = "sky blue") + labs(title = "Food",x = "Month", y= "Number of Reviews") + theme(plot.title = element_text(size = 25, color = "black")) + theme_igray()

shopping_joined %>% ggplot(aes(x = factor(name, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov","Dec")), y=num_reviews)) + geom_col(fill = "purple") + labs(title = "Shopping",x = "Month", y= "Number of Reviews") + theme(plot.title = element_text(size = 25, color = "black")) + theme_igray()
```

```
beauty_joined %>% ggplot(aes(x = factor(name, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov","Dec")), y=num_reviews)) + geom_col(fill = "pink") + labs(title = "Beauty",x = "Month", y= "Number of Reviews") + theme(plot.title = element_text(size = 25, color = "black")) + theme_igray()

home_joined %>% ggplot(aes(x = factor(name, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov","Dec")), y=num_reviews)) + geom_col(fill = "dark green") + labs(title = "Home",x = "Month", y= "Number of Reviews") + theme(plot.title = element_text(size = 25, color = "black")) + theme_igray()
```

Visualizing Percent Change Per Month

```
per_change_line <- per_change %>%
  ggplot(aes(x=factor(name, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov","Dec")), y=percent_change, col = categories, group=categories)) + geom_line() + geom_point() + labs(title = "Percent Change of Reviews For Each Category", x = "Month", y="Percent Change")

per_change_line
```

Visualizing Average Rating Over Time for Each Category

```
food_joined_ratings %>% ggplot(aes(x = factor(name, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov","Dec")), y = ave_rating)) + geom_violin(fill = "sky blue") + scale_y_continuous(breaks = seq(0,5,0.5)) + labs(title = "Food",x = "Month", y= "Average Rating") + theme(plot.title = element_text(size = 20, color = "black"))

shopping_joined_ratings %>% ggplot(aes(x = factor(name, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov","Dec")), y = ave_rating)) + geom_violin(fill = "purple")+ scale_y_continuous(breaks = seq(0,5,0.5)) + labs(title = "Shopping",x = "Month", y= "Average Rating") + theme(plot.title = element_text(size = 20, color = "black"))

beauty_joined_ratings %>% ggplot(aes(x  = factor(name, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov","Dec")), y = ave_rating)) + geom_violin(fill = "pink")+ scale_y_continuous(breaks = seq(0,5,0.5)) + labs(title = "Beauty",x = "Month", y= "Average Rating") + theme(plot.title = element_text(size = 20, color = "black"))

home_joined_ratings %>% ggplot(aes(x  = factor(name, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov","Dec")), y = ave_rating)) + geom_violin(fill = "dark green")+ scale_y_continuous(breaks = seq(0,5,0.5)) + labs(title = "Home",x = "Month", y= "Average Rating") + theme(plot.title = element_text(size = 20, color = "black"))
```

Visualizing Number of Reviews By City Per Season

```
map <-map_data("state")

p <- ggplot() + geom_polygon( data=map, aes(x=long, y=lat, group = group),colour="white",
fill="grey10" )
p1 <- p + geom_point( data=geo, aes(x=longitude, y=latitude, size = tot_numrev), color="coral1")
+ scale_size(name="Total Number of Reviews") + facet_wrap(~season) + labs(title = "Number of
Reviews By City Per Season", x="Latitude", y="Longitude")

p1
```