Most Popular Yelp Cuisine by State

Kitu Komya

December 22, 2017

Introduction

In this extremely short group project, I chose to analyze which Yelp cuisines are most popular by the lower 48 states. This report contains the following sections: R Code, Methods, Analysis Findings, and Conclusions.

R Code

```
# load packages
library(ggplot2)
library(ggmap)
library(maps)
library(readr)
library(dplyr)

# load data
load(file = "Kitu/College/Senior Year/Fall Quarter/Stats 140SL/business.RData")

# look at categories
categories <- as.data.frame(table(business$categories.0))

# subset only certain cuisines
food <- c("Mexican", "Chinese", "Italian", "Indian", "Greek", "Thai", "Vietnamese")
business <- business[categories.0 %in% food, c(8, 11:13, 22:29)]

# import state abbreviations
state_abbreviations <- read_csv("~/Kitu/College/Senior Year/Geographic/state_abbreviations.csv")

# change abbreviation to full name
business$usstates <- ifelse(business$state %in% state_abbreviations$Abbreviation, state_abbreviations$S

# subset to only US states
business <- business[usstates != "", ]

# clean US states
business$usstates <- tolower(business$usstates)

# count how many total restaurants per cuisine exists
table(business$categories.0)

# create 2 way table for total count within each cuisine
table <- as.data.frame(table(business$usstates, business$categories.0))

# make proportion variable, because count is not appropriate (there are lots of Mexican restaurants)
table$prop <- 0
```

```
table$prop <- ifelse(table$Var2 == "Chinese", table$Freq/768, table$prop)
table$prop <- ifelse(table$Var2 == "Greek", table$Freq/138, table$prop)
table$prop <- ifelse(table$Var2 == "Indian", table$Freq/129, table$prop)
table$prop <- ifelse(table$Var2 == "Italian", table$Freq/697, table$prop)
table$prop <- ifelse(table$Var2 == "Mexican", table$Freq/1236, table$prop)
table$prop <- ifelse(table$Var2 == "Thai", table$Freq/186, table$prop)
table$prop <- ifelse(table$Var2 == "Vietnamese", table$Freq/115, table$prop)

# choose highest prop for each state
top <- table %>%
  group_by(Var1) %>%
  top_n(n = 1, wt = prop)

# left join to all_states dataframe
all_states <- map_data("state") # load US map data
names(top)[1] <- "region"
all_states <- left_join(all_states, top)
names(all_states)[7] <- "Cuisine"

# clean up dataframe
all_states <- all_states[ , -6]
all_states <- all_states[complete.cases(all_states), ]

# make map of lower 48 states
ggplot() + geom_polygon(data = all_states, aes(x = long, y = lat, group = group, fill = Cuisine), color
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_rect(fill = "white", colour = "white"),
        axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(), axis.text.x = element_blank(),
        axis.text.y = element_blank()) +
  ggtitle(label = "Most Popular Yelp Cuisine by State") + theme(plot.title = element_text(hjust = 0.5))
  labs(x = "", y = "")
```

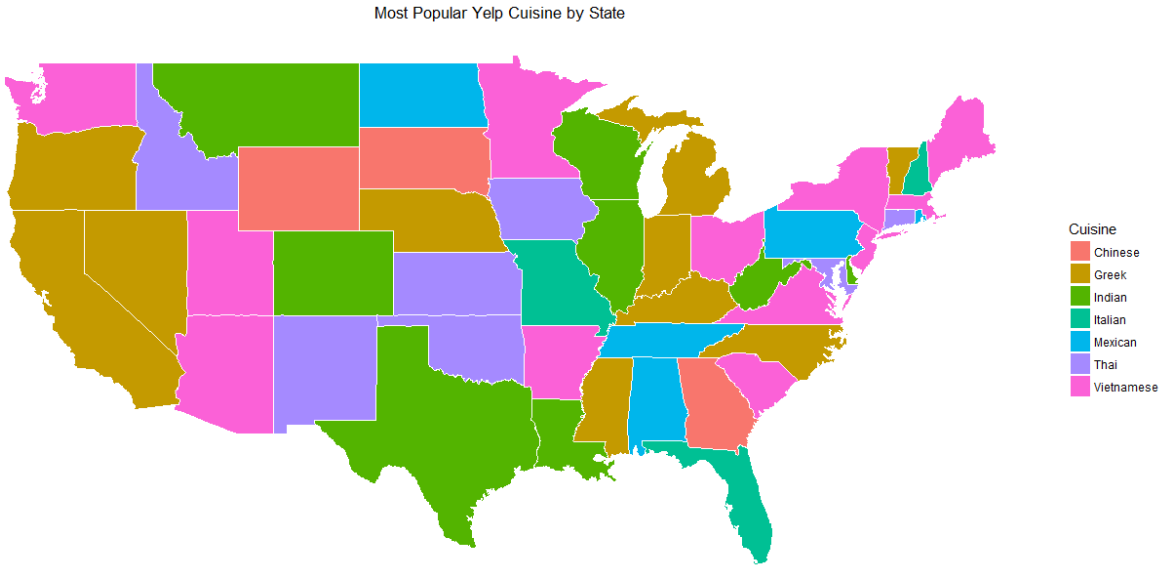
Method

My goal was to determine which cuisines are most popular in the lower 48 states of the USA. I did this by first examining the dataset and choosing some of the popular cuisines (hence reducing the total number of cuisines to a select few) to analyze, because there were too many cuisines otherwise. Then I calculated how many total restaurants of each cuisine is present on Yelp in the USA. From this, I found all the proportions of cuisine restaurants in each state, and chose the highest proportion.

For instance, I found that there are 697 Italian restaurants on Yelp in the USA and 1236 Mexican ones. I found out how many Mexican and Italian restaurants are in California, and used that number over the the total number (697 and 1236, respectively). Then, I chose the highest proportion cuisine for each state (in California, it was Greek).

The reason I used this method versus simple count of cuisine was simple: a simple count of all the cuisines does not standardize the total number of each cuisine restaurants in the USA. For instance, Mexican restaurants, in general, are found very frequently in the USA. If I did it by count, most, if not all, of the state's most popular Yelp cuisine would be Mexican. This method, however, does not take into account that Mexican restaurants are more popular in general anyway. Thus, this project essentially chose the cuisine most notably popular in each state, in comparison to the proportion of all other cuisines.

Analysis Findings



The West Coast typically enjoys Greek cuisines, while the other regions don't seem to share much of a pattern. This map is interesting to see, since it allows a user to compare her preferences to her state's or to other states'.

Conclusions

1. Cuisine popularity is pretty diverse across the states and shows little pattern across the US.

Future Analysis

2. Perhaps include all cuisines to use all available data and give a more realistic portrayal.
3. Include Hawaii and Alaska.
4. Color by zipcodes/local regions within states instead states.