# Vegan and Vegetarian Restaurant Model

**Authors:** Jorit Studer and Larissa Eisele
**Module:** R-Bootcamp

Submitted on February 25th, 2022

Supervisor: Matteo Tanadini and Claude Renaux

Lucerne University of Applied Sciences and Arts

# Table of Contents

**1 Introduction** **1**

**2 Methodology** **1**
2.1 Importing Data . . . 1
2.2 Data Processing . . . 2
2.3 Missing Values . . . 3

**3 Data Visualization** **3**
3.1 Summary Statistics . . . 3
3.2 Histogram . . . 4
3.3 Boxplots . . . 7
3.4 Scatterplot . . . 9

**4 Modeling** **10**
4.1 Model . . . 11
4.2 Interpretation . . . 11
4.3 Most significant predictors . . . 12
4.4 Residuals . . . 12

**5 Chapter of Choice: Maps, Wordcloud and Bookdown** **14**
5.1 Maps . . . 14
5.2 Wordcloud . . . 17
5.3 Bookdown . . . 18

**6 Reflection and Conclusion on the Project** **19**

**7 Conclusion for ourselves** **19**

**References** **20**

# 1 Introduction

Inspired by Veganuary, we discovered a dataset about vegetarian and vegan restaurants in the U.S. (Datafiniti, 2018). The dataset was found through a dataset search on Google, hosted on Kaggel. The reason for selecting this dataset was that many useful parameters were already available. The description of the dataset states that it contains data from 18,0000 restaurants. Unfortunately, we only later noticed that only 211 individual restaurants with 10,000 menu items are included in the dataset, as it serves as an advertisement for the Datafiniti Business Database. The entire dataset is only available for purchase, and after consulting with the lecturer (Matteo Tanadini), it was agreed that the study should continue with less meaningful results.

The purpose of this paper is to examine the relationship between the number of vegetarian or vegan restaurants, population and median household income for the U.S. zip code (Michigan Population Studies Center, 2020), and the frequency of Google searches for vegetarian and vegan restaurants in that state (Google Trends, n.d.a, n.d.b). With the small sample size of 211 restaurants and 10000 menu items, this study cannot significantly predict the number of vegetarian restaurants in a zip code. Nevertheless, this paper offers some insights into vegetarian and vegan supply in the U.S., as well as the demand and whether demographics, income, and Google searches match up.

For the following work, different R packages have been used. The respective packages and their installation can be found in the original R Markdown file or the README.

# 2 Methodology

The first step was to research the relevant data. The data was imported into R and first visualizations were created to get an insight into the data. The R programming language was used for the calculations in the following document. For simplicity, not all of the code is included. However, all code can be found in the original R Markdown file.

In the following document, different calculation and presentation methods are used. The different methods are intended to reflect both the teaching content from the R-Bootcamp and the knowledge that the authors have gained during the learning process itself.

**Model**

$$\text{restaurants} = \beta_0 + \beta_1 * \text{residents} + \beta_2 * \text{income} + \beta_3 * \text{searches}$$

The formula above illustrates our original concept for the model. The model aims to explain the supply of vegetarian and vegan restaurants in the U.S. based on demand in different regions. Demand is interpreted by the number of searches for “vegan restaurants” and “vegetarian restaurants.” In addition, the model is designed to explain a possible relationship between supply and demand in terms of mean household income and population size.

## 2.1 Importing Data

The following work is based on three different data sources. We are gathering data from Datafiniti (2018) through Kaggel (Restaurant Data), Michigan Population Studies Center (2020) (Mean household income U.S.) and from Google Trends (n.d.a) and Google Trends (n.d.b) (Google Search Data).

```
#setwd("/Users/larissaeisele/switchdrive/R-Bootcamp/project-R_Bootcamp")
restaurants <-read.csv("./Data/veg_restaurants_US.csv")
income_pop_by_zip <- read_excel("./Data/MeanZIP-3.xlsx")
google_searches_vegetarian <- read_excel("./Data/searches_vegetarian_res.xlsx")
google_searches_vegan <- read_excel("./Data/searches_vegan_res.xlsx")
```

## 2.2 Data Processing

### 2.2.1 Restaurant Data

As a first step, we took a closer look into the restaurant data and decided which parameter was needed. With the following code, we are selecting the relevant parameter. Further, we had to change the type of zip code to numeric and create a new row containing the cuisines served in the restaurant as a list.

```
restaurants <- restaurants %>% select(id, dateAdded, cuisines, latitude,
                                      longitude, menus.category, name,
                                      province, postalCode)
restaurants$postalCode <- as.numeric(restaurants$postalCode)
restaurants <- restaurants %>% mutate(cuisines_split = strsplit(cuisines,","))
```

While looking closer into the restaurant data, we found a restaurant without a zip code. We looked up the respective restaurant and found the right zip code for the restaurant. We cleaned the missing data from the restaurant and added the zip code to the dataset with the following code.

```
restaurants[restaurants$id == "AV1Tp59p3D1zeR_xE2DL",]$postalCode <- 94596
anyNA(restaurants$postalCode) # There are no missing values anymore in postalCode (zip code)
```

```
## [1] FALSE
```

### 2.2.2 Population Size and Mean Household Income Data

To enrich our restaurant data, we merged it with the mean household income and population size by zip code. This was done using a left join with the key being our zip code.

**Joining the Data**

```
joined_data <- left_join(restaurants, income_pop_by_zip %>%
                           select(Zip, Mean, Pop), by = c("postalCode" = "Zip"))
```

**Checking for missing data**

```
sum(is.na(joined_data$Mean)) # all postal codes have a mean household income
```

```
## [1] 0
```

```
sum(is.na(joined_data$Pop)) # all postal codes have a population
```

```
## [1] 0
```

### 2.2.3 Google Search Data

During the project, it occurred to us that it would be interesting to compare demand with the supply of vegetarian and vegan restaurants. However, no data relating to vegetarian or vegan diets in the U.S. population was found. Therefore, we decided to visualize demand using Google search data, and used the search query for “vegetarian restaurants” and “vegan restaurants” through the website of Google Trends with a time period of 5 years.

```
# Vegetarian Restaurant Search Results
vegetarian_temp <- google_searches_vegetarian %>% group_by(province) %>%
  summarise(total_searches_vegetarian = sum(search_vegetarian_res, na.rm = TRUE))
joined_data <- left_join(joined_data, vegetarian_temp %>%
                           select(province, total_searches_vegetarian), by = "province")
# Vegan Restaurant Search Results
vegan_temp <- google_searches_vegan %>% group_by(province) %>%
  summarise(total_searches_vegan = sum(search_vegan_res, na.rm = TRUE))
```

```
joined_data <- left_join(joined_data, vegan_temp %>%
                           select(province, total_searches_vegan), by = "province")
```

### 2.3 Missing Values

We checked the search data for missing values. Unfortunately, data was not available for all states in the U.S. Google did not provide any information on whether the data is not collected for certain states or whether there are no searches for these states. We consider the latter unlikely given the time period selected. The missing information was treated as NA in the further analysis.

```
sum(is.na(joined_data$total_searches_vegetarian))
```

```
## [1] 42
```

```
sum(is.na(joined_data$total_searches_vegan))
```

```
## [1] 42
```

```
# 42 provinces have no google search hits
```

## 3 Data Visualization

### 3.1 Summary Statistics

As a first step, we explored our data sources in more detail. The following summary statistics on median household income and population by zip code (Michigan Population Studies Center, 2020) provided several insights. Regarding median household income, the following table indicates that the data is not adjusted, as 7 values are missing. However, since the missing values did not affect our further analysis, we decided to proceed with them. The summary statistics also reveal that the difference between incomes is very large. One might assume that this is related to different living situations. For example, the dataset could include very low-income students and households with two high-income earners. The summary statistics on population by zip code suggest that there appears to be only one person registered in a zip code. We are aware that this is most likely an error in the dataset, but nevertheless decided to use the dataset for our analysis. Another interesting finding is that the U.S. population by zip code has an average population of 9193, which is equivalent to a small town in Switzerland.

```
# yearly mean household income whole U.S.
summary(income_pop_by_zip %>% select(Mean, Pop))
```

```
##       Mean               Pop
##  Min.   :    53.6   Min.   :     1
##  1st Qu.: 48593.2   1st Qu.:   736
##  Median : 56949.6   Median :  2756
##  Mean   : 63452.2   Mean   :  9193
##  3rd Qu.: 70341.2   3rd Qu.: 12513
##  Max.   :361842.3   Max.   :113916
##  NA's   :7
```

```
#summarize household income in U.S.
joined_data %>% group_by(province) %>%
  summarize(average_income = sum(Mean) / length(id), std_dev=sd(Mean,na.rm=TRUE))
```

```
## # A tibble: 25 x 3
##    province average_income std_dev
##    <chr>             <dbl>   <dbl>
##  1 AZ               58325.  18706.
```

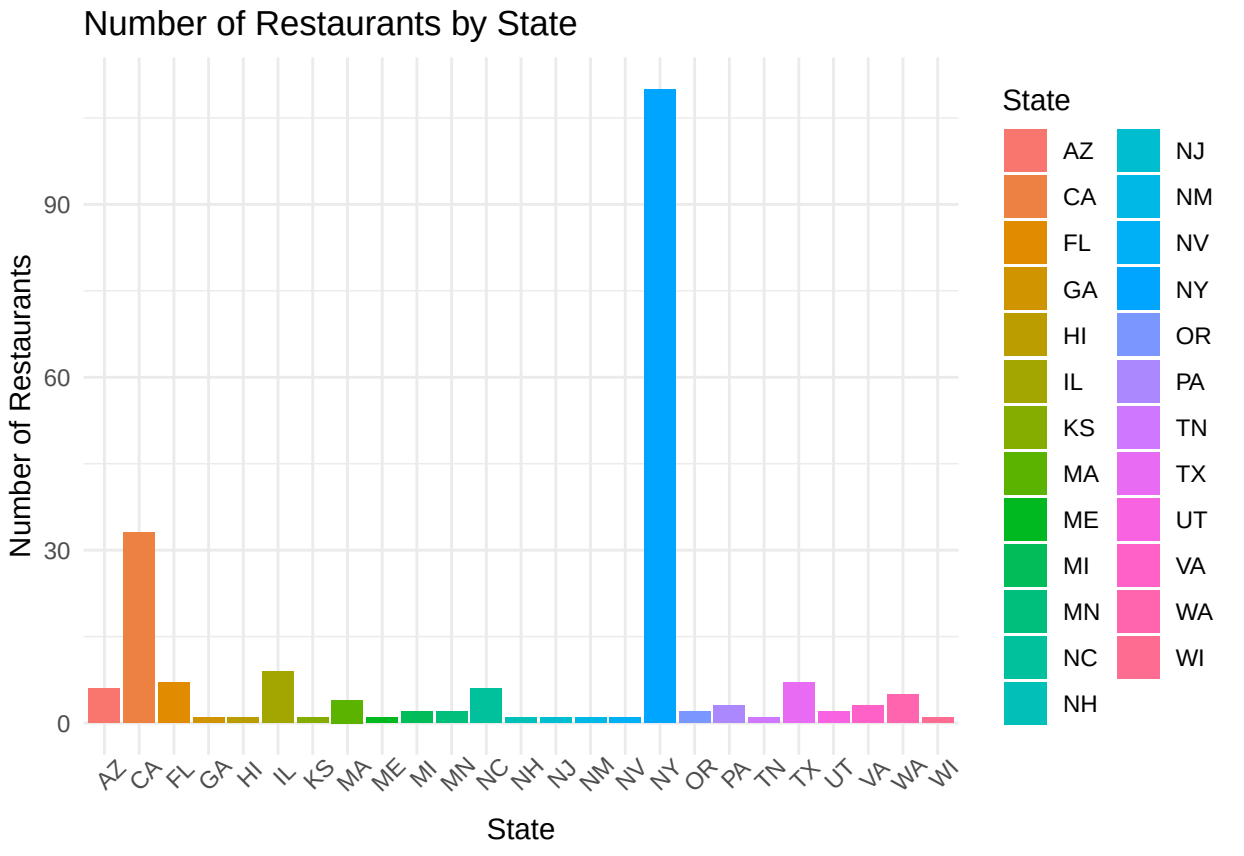
```
##  2 CA               95888.  33155.
##  3 FL               96091.  12324.
##  4 GA               35938.      0
##  5 HI               83599.      0
##  6 IL              111035.  26749.
##  7 KS               64736.      0
##  8 MA              183076.  83678.
##  9 ME               52719.  12109.
## 10 MI               58059.   3889.
## # ... with 15 more rows
```

## 3.2 Histogram

### 3.2.1 Vegetarian and Vegan Restaurant by State

The histograms below present the number of vegetarian and vegan restaurants by state to give a sense of the regional distribution of the data. It is evident that New York has the most vegetarian and vegan restaurants in our dataset, while California is the state with the second most vegetarian and vegan restaurants. In addition, we can see that the number of restaurants in other states seem to be fairly evenly distributed.

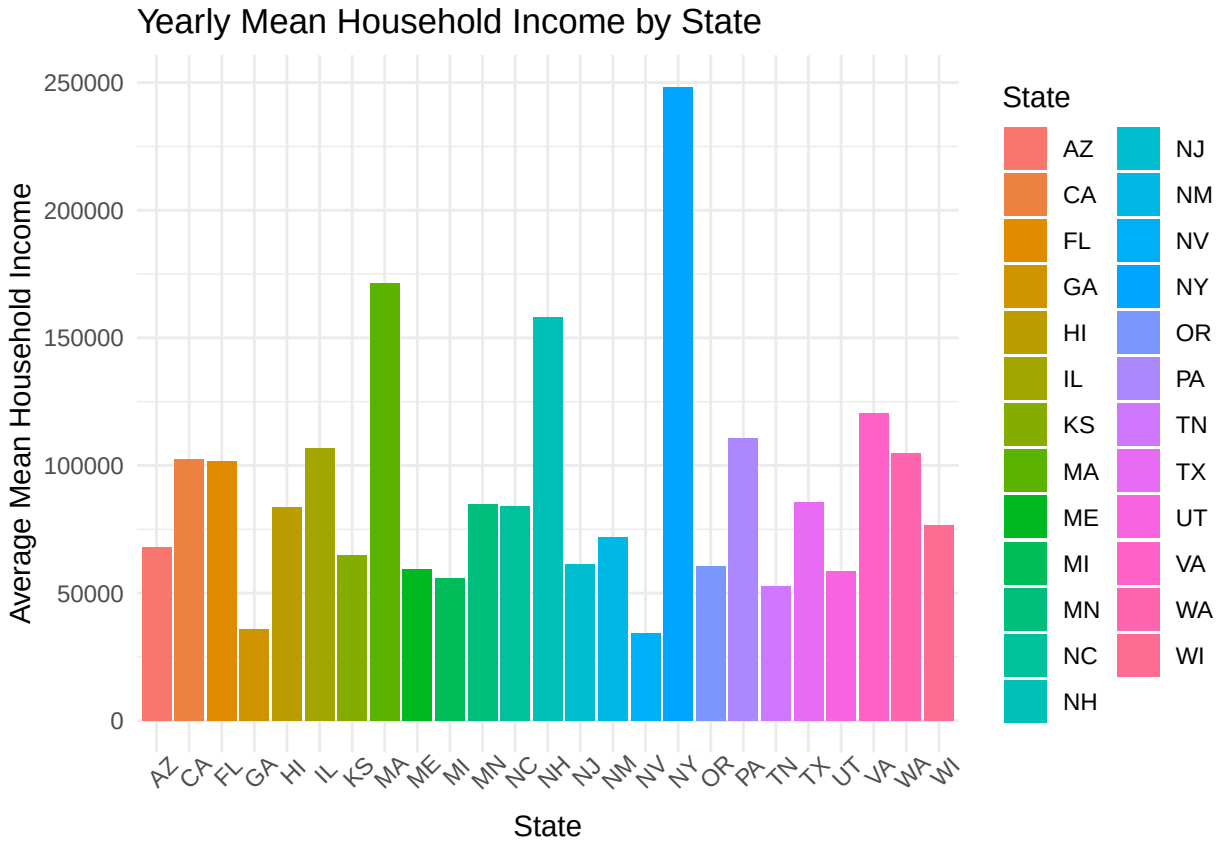
```
#Vegetarian and Vegan Restaurants by State
number_of_res <- joined_data %>% group_by(province) %>%
  summarise(Number_Of_Restaurants = n_distinct(id))
ggplot(number_of_res, aes(x=province, y=Number_Of_Restaurants, fill=province)) +
  geom_bar(stat = "identity") +
  ggtitle("Number of Restaurants by State") +
  xlab("State") + ylab("Number of Restaurants") + labs(fill='State') +
  theme_minimal() + theme(axis.text.x = element_text(angle = 45))
```



### 3.2.2 Mean Houshold Income by State

The following histogram shows median household income by state. Based on our dataset, New York is the state with the highest median household income, followed by Massachusetts and New Hampshire. The other states have a fairly balanced median household incomes. Nevada and Georgia have the lowest median household income.

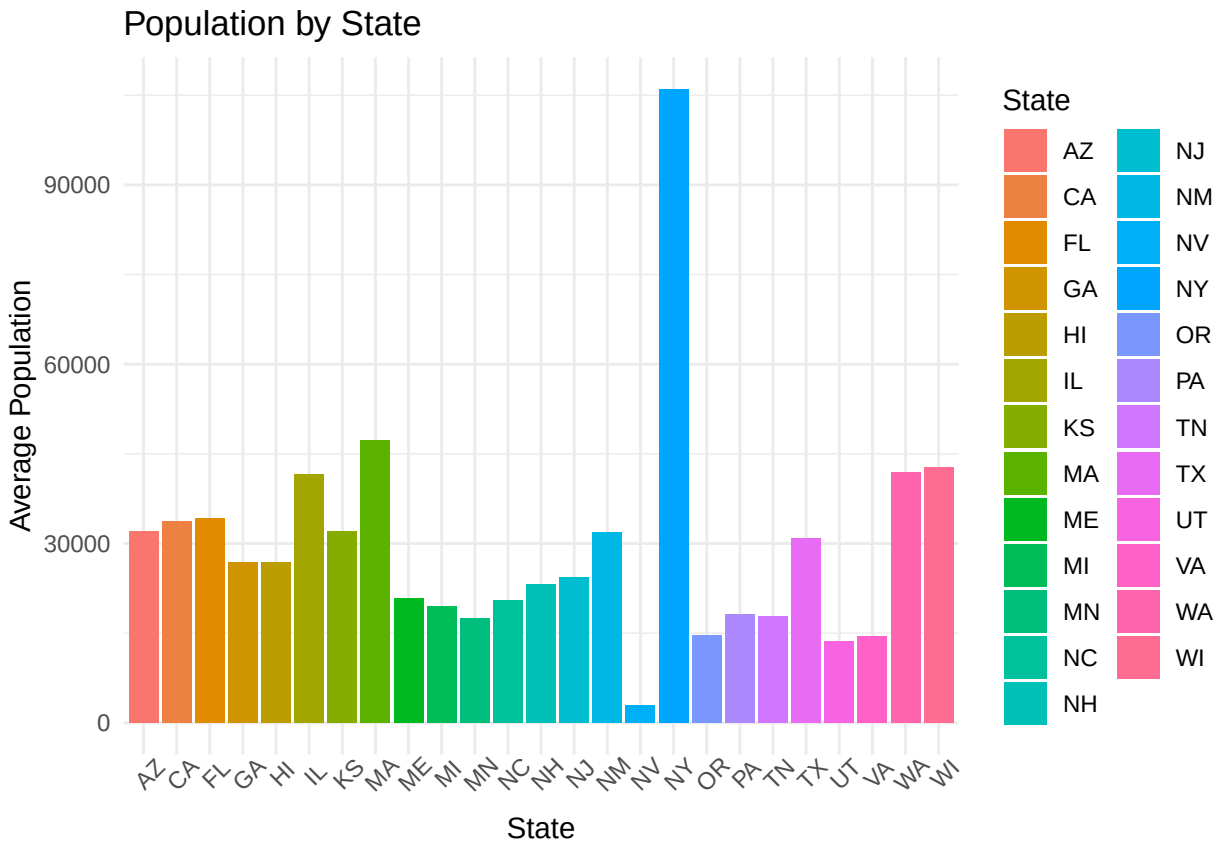
```
# Mean Household income by State
temp <- joined_data %>% distinct(id, province, Mean) %>% group_by(province) %>%
  summarise(Avg_Mean_Province = (sum(Mean) / n_distinct(Mean)))
ggplot(data=temp, aes(x=province, y=Avg_Mean_Province, fill=province)) +
  geom_bar(stat="identity") +
  ggtitle("Yearly Mean Household Income by State") +
  xlab("State") + ylab("Average Mean Household Income") + labs(fill='State') +
  theme_minimal() + theme(axis.text.x = element_text(angle = 45))
```



### 3.2.3 Population by State

The following histogram shows the population per state. New York has the highest population, which is probably due to the high population density in the city of New York. The state of Nevada has by far the lowest population density, which can be explained by the environmental and climatic conditions in this region.

```
# Population by State
temp <- joined_data %>% distinct(id, province, Pop) %>% group_by(province) %>%
  summarise(Avg_Pop = (sum(Pop) / n_distinct(Pop)))
ggplot(data=temp, aes(x=province, y=Avg_Pop, fill=province)) +
  geom_bar(stat="identity") +
  ggtitle("Population by State") +
  xlab("State") + ylab("Average Population") + labs(fill='State') +
  theme_minimal() + theme(axis.text.x = element_text(angle = 45))
```

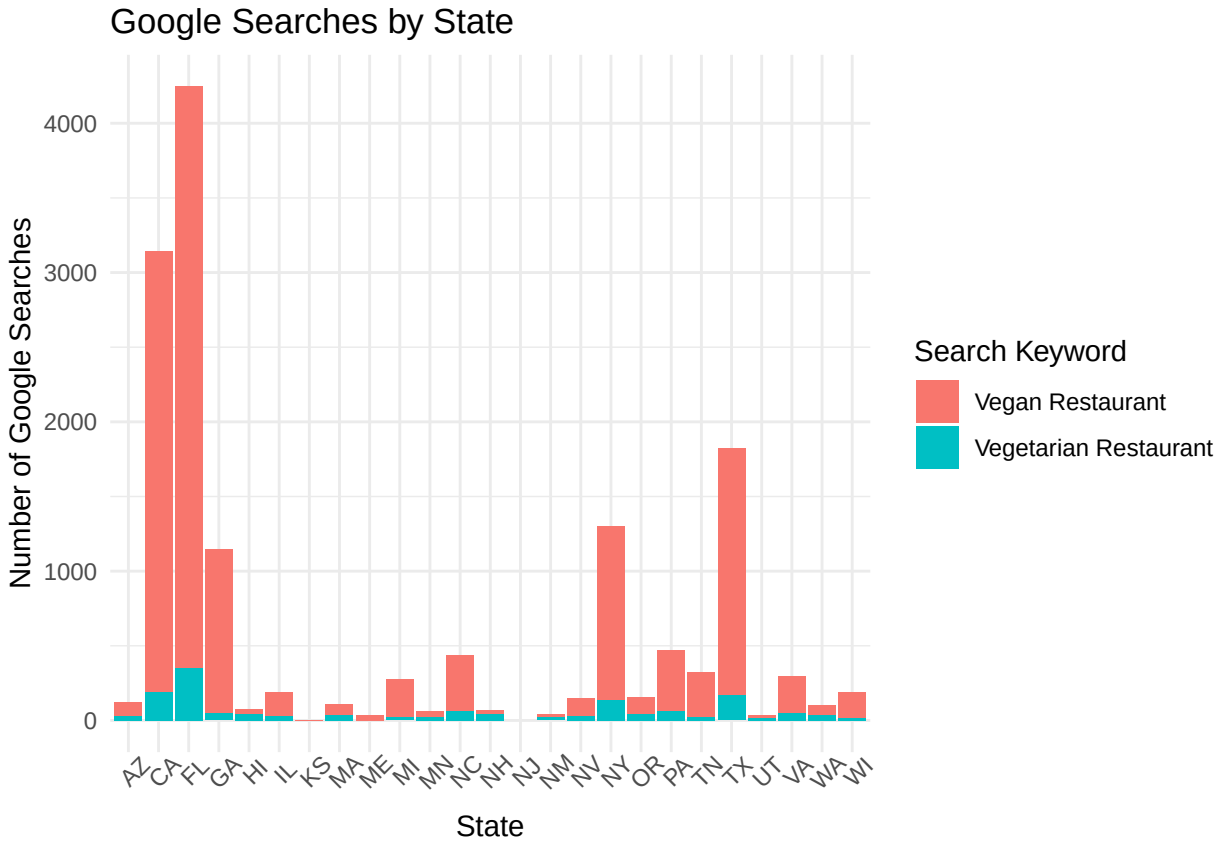



### 3.2.4 Vegetarian and Vegan Searches by State

Based on Google search data, Florida has the highest demand for vegetarian and vegan restaurants. The data does not match our findings from the restaurant data, according to which the supply of vegetarian and vegan restaurants in Florida is not that high. Comparing the searches for vegetarian or vegan restaurants, the graph shows that the demand for vegan restaurants seems to be higher. This may also be attributed to the fact that it is less common for a regular restaurant to serve vegan dishes, while most restaurants offer at least a few vegetarian dishes.

```
# province / type (vegan, veg) / value
data <- joined_data %>% group_by(province) %>%
  distinct(province, total_searches_vegan, total_searches_vegetarian) %>%
  pivot_longer(cols = c(`total_searches_vegetarian`, `total_searches_vegan`),
               names_to = "type",
```

```
                values_to = "value")
# Google Searches Stacked
ggplot(data, aes(x=province, y=value, fill=type)) +
    geom_bar(position="stack", stat="identity") +
    scale_fill_discrete(name = "Search Keyword",
                        labels = c("Vegan Restaurant",
                                   "Vegetarian Restaurant")) +
    ggtitle("Google Searches by State") +
    xlab("State") + ylab("Number of Google Searches") +
    theme_minimal() + theme(axis.text.x = element_text(angle = 45))
```

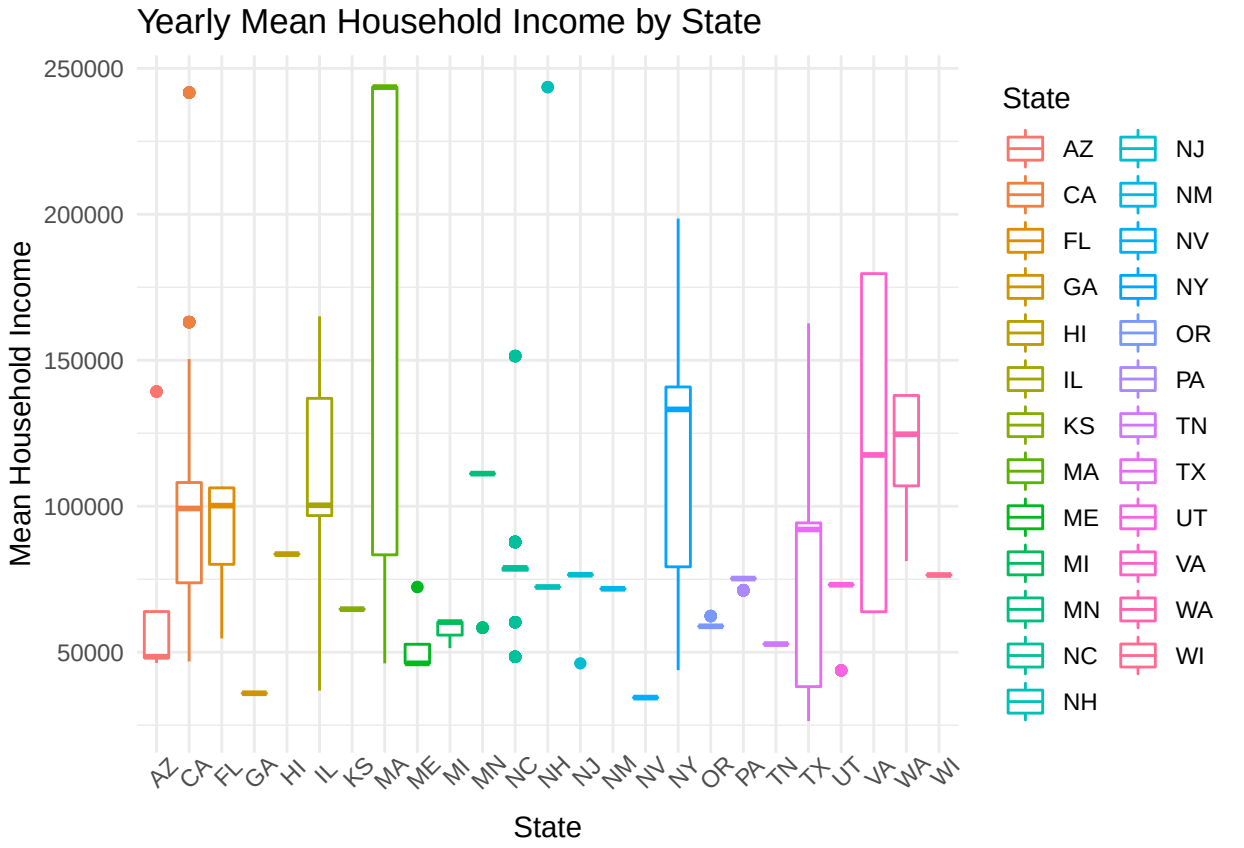



## 3.3 Boxplots

### 3.3.1 Household Income

A boxplot can be used to graphically display the distribution of at least one ordinally scaled value. The boxplot below shows median household income by state. The median value for Massachusetts appears to be quite high. California and New Hampshire have some outliers with very high median household incomes. There are a few groups of states that have nearly the same median.

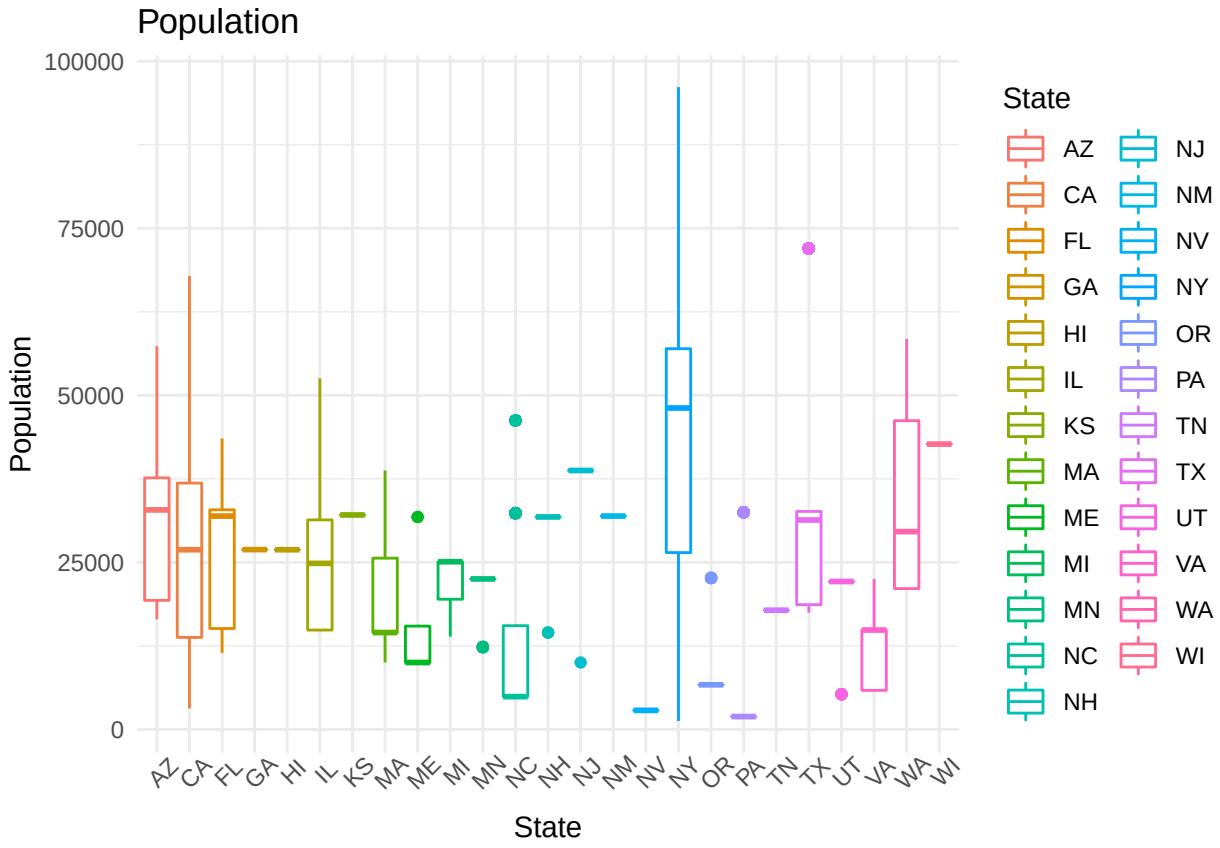
```
#Boxplot Household Income
ggplot(data=joined_data, aes(x=province, y=Mean, color=province)) +
  geom_boxplot(alpha=1) +
  ggtitle("Yearly Mean Household Income by State") +
  xlab("State") + ylab("Mean Household Income") + labs(color='State') +
  theme_minimal() + theme(axis.text.x = element_text(angle = 45))
```




### 3.3.2 Population

The following illustrates the distribution of the population by state, again the spread for New York and California are the largest in this dataset.

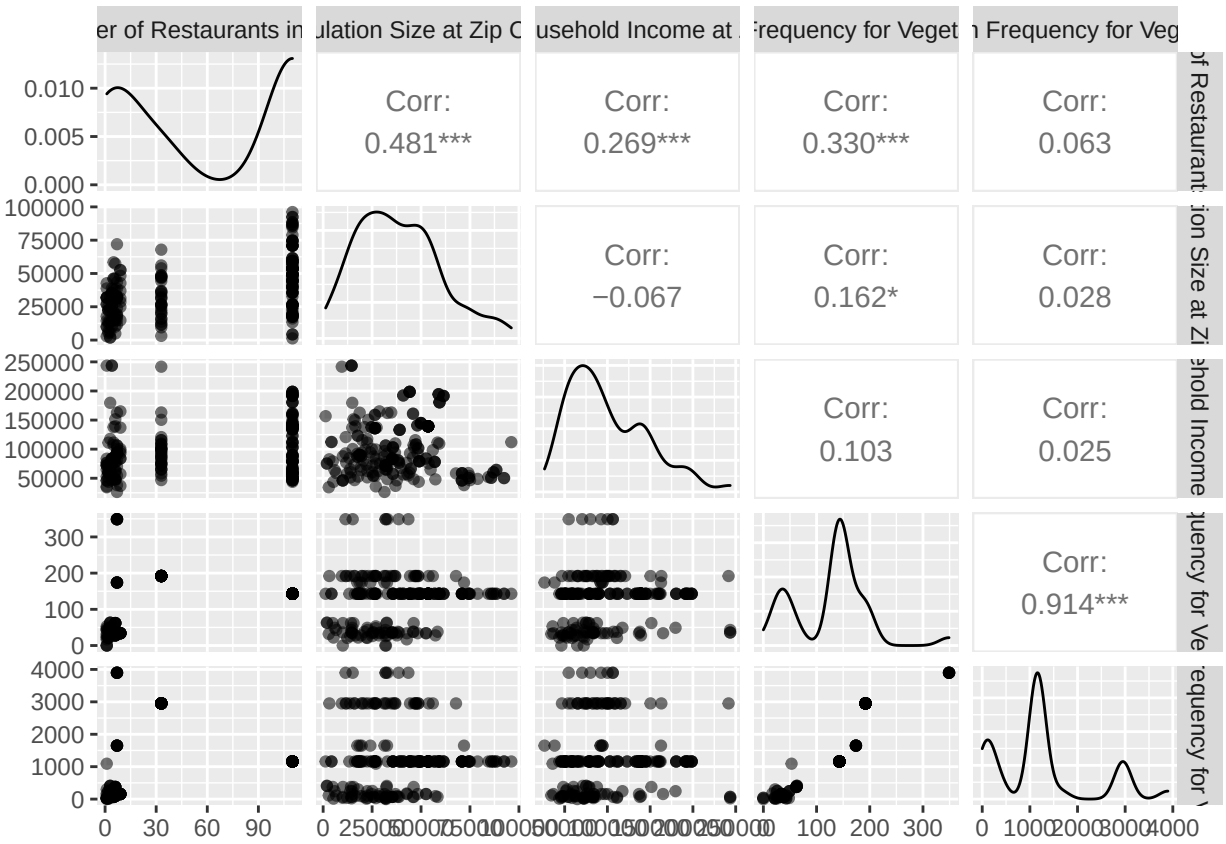
```
#Boxplot Population
ggplot(data=joined_data, aes(x=province, y=Pop, color=province)) +
  geom_boxplot(alpha=1) +
  ggtitle("Population") + xlab("State") + ylab("Population") + labs(color='State') +
  theme_minimal() + theme(axis.text.x = element_text(angle = 45))
```




### 3.4 Scatterplot

With scatterplots a relationship between two variables can be visualized. The following scatterplot shows the relationship between the demand of vegan and vegetarian restaurants and the amount of offers by region.

```
number_of_res <- joined_data %>% group_by(province) %>%
  summarise(Number_Of_Restaurants_In_Province = n_distinct(id))
unique_res_data <- joined_data %>% distinct(id,
                                            Pop,
                                            Mean,
                                            total_searches_vegetarian,
                                            total_searches_vegan, province)
unique_res_data <- left_join(unique_res_data, number_of_res %>%
                           select(province, Number_Of_Restaurants_In_Province),
                           by = "province"
                           )
colnames(unique_res_data) <- c("ID",
                               "State",
                               "Mean Household Income at Zip Code",
                               "Population Size at Zip Code",
                               "Google Search Frequency for Vegetarian Restaurant",
                               "Google Search Frequency for Vegan Restaurant",
                               "Number of Restaurants in State")
```




# 4 Modeling

```
fit <- lm(`Number of Restaurants in State` ~ `Population Size at Zip Code` +
            `Mean Household Income at Zip Code` +
            `Google Search Frequency for Vegetarian Restaurant` +
            `Google Search Frequency for Vegan Restaurant`,
          data=unique_res_data)
summary(fit)
```

```
##
## Call:
## lm(formula = `Number of Restaurants in State` ~ `Population Size at Zip Code` +
##     `Mean Household Income at Zip Code` + `Google Search Frequency for Vegetarian Restaurant` +
##     `Google Search Frequency for Vegan Restaurant`, data = unique_res_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -108.469  -16.242    7.669   21.748   52.087
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                      -3.459e+01  7.589e+00
## `Population Size at Zip Code`                     7.341e-04  1.150e-04
## `Mean Household Income at Zip Code`               1.898e-04  4.984e-05
## `Google Search Frequency for Vegetarian Restaurant`  9.301e-01  8.733e-02
## `Google Search Frequency for Vegan Restaurant`    -5.598e-02  5.885e-03
```

```
##                                                  t value Pr(>|t|)
## (Intercept)                                       -4.558 8.66e-06 ***
## `Population Size at Zip Code`                      6.382 1.06e-09 ***
## `Mean Household Income at Zip Code`                3.808 0.000183 ***
## `Google Search Frequency for Vegetarian Restaurant`  10.651  < 2e-16 ***
## `Google Search Frequency for Vegan Restaurant`      -9.513  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.22 on 215 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.5554, Adjusted R-squared:  0.5471
## F-statistic: 67.14 on 4 and 215 DF,  p-value: < 2.2e-16
```

## 4.1 Model

$$\begin{aligned}\text{Number of Restaurants in a State} = &-34.59\\ &+ 0.0007341 * \text{Population}\\ &+ 0.0001898 * \text{Mean Household Income}\\ &+ 0.9301 * \text{Google Search Frequency for Vegetarian Restaurant}\\ &- 0.05598 * \text{Google Search Frequency for Vegan Restaurant}\end{aligned}$$

## 4.2 Interpretation

- The R-squared seems to indicate that ~56% of the variability of the data around its mean is explained by the model. The F-statistics p-value is below the significance level of 0.05% and thus significant, hence we reject the null hypothesis. There is at least one $\beta$ significantly different from 0. Thus at least one variable contributes significantly to the model. Moreover based on the p-value's $\beta_1$, $\beta_2$, $\beta_3$ and $\beta_4$ all seem to significantly contribute to the model.
- With no population, no income, and no Google searches there are negative ~35 restaurants in a state.
- For each additional resident in U.S. state the amount of vegetarian/vegan restaurants increases by 0.0007341. Therefore, there is approximately one restaurant for every 1'400 residents at a specific zip code in a given state.
- For each $ in Mean Household Income the number of restaurants increases by 0.0001898. Thus, for approximately $5300 in Mean Household Income the number of restaurants increases by one for a given state.
- For every Google search with the term 'Vegetarian Restaurant' the number of restaurants increases by 0.9301.
- For every Google search with the term 'Vegan Restaurant' the number of restaurants decreases by 0.05598.

### 4.3 Most significant predictors

```
reg_test <- unique_res_data %>% select(`Number of Restaurants in State`,
                                       `Population Size at Zip Code`,
                                       `Mean Household Income at Zip Code`,
                                       `Google Search Frequency for Vegetarian Restaurant`,
                                       `Google Search Frequency for Vegan Restaurant`)
reg <- regsubsets(`Number of Restaurants in State` ~ .,
                  data = reg_test,
                  method = "forward",
                  nvmax = 6)
summary(reg)$which
```

```
##   (Intercept) `Population Size at Zip Code` `Mean Household Income at Zip Code`
## 1        TRUE                          TRUE                                FALSE
## 2        TRUE                          TRUE                                 TRUE
## 3        TRUE                          TRUE                                 TRUE
## 4        TRUE                          TRUE                                 TRUE
##   `Google Search Frequency for Vegetarian Restaurant`
## 1                                               FALSE
## 2                                               FALSE
## 3                                                TRUE
## 4                                                TRUE
##   `Google Search Frequency for Vegan Restaurant`
## 1                                          FALSE
## 2                                          FALSE
## 3                                          FALSE
## 4                                           TRUE
```

Based on forward algorithm the predictors **population and mean household income** are most often included in the model (i.e most significant).

### 4.4 Residuals

```
par(mar = c(4, 4, .1, .1))
model.diag.metrics <- augment(fit)
# Population
ggplot(model.diag.metrics, aes(`Number of Restaurants in State`,
                               `Population Size at Zip Code`)) +
  geom_point() +
  stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = `Number of Restaurants in State`,
                   yend = .fitted), color = "red", size = 0.3)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
# Income
ggplot(model.diag.metrics, aes(`Number of Restaurants in State`,
                               `Mean Household Income at Zip Code`)) +
  geom_point() +
  stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = `Number of Restaurants in State`,
                   yend = .fitted), color = "red", size = 0.3)
```

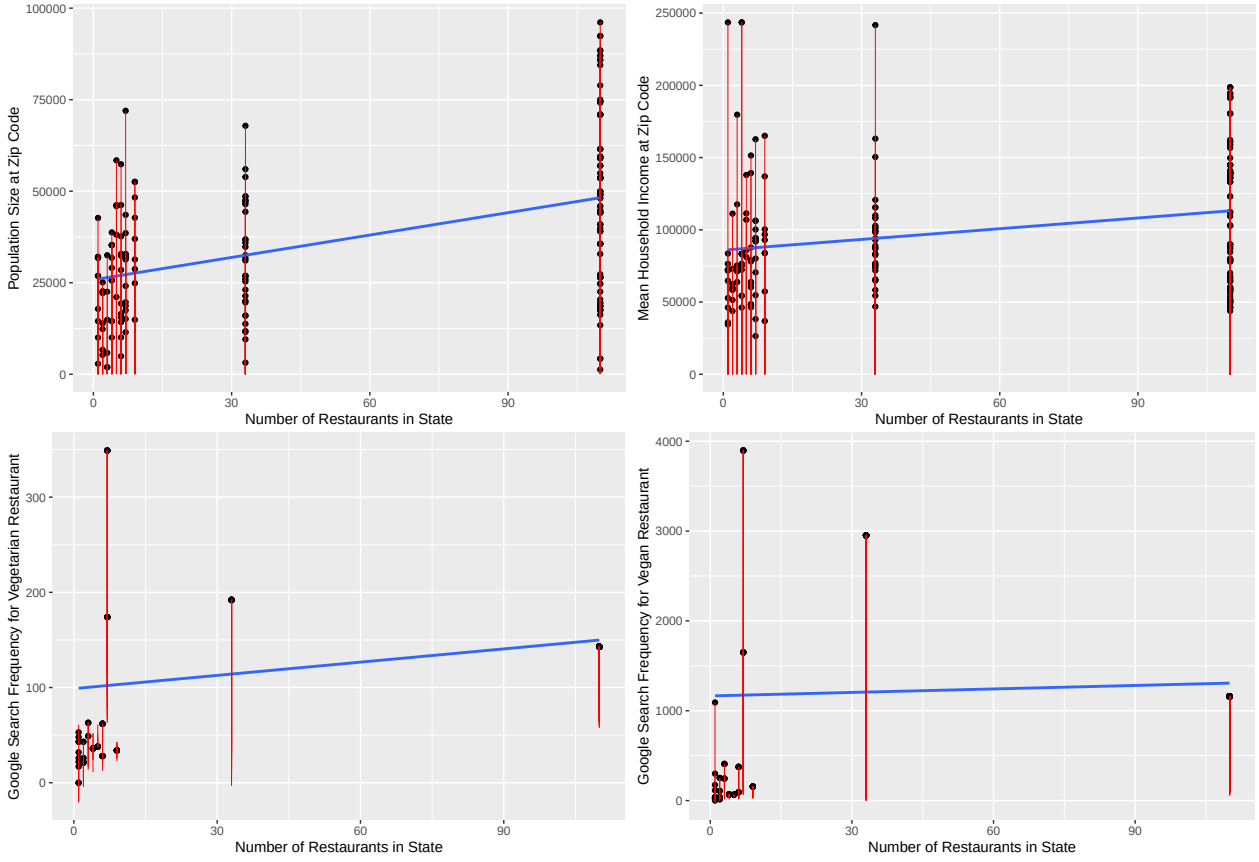
```
## `geom_smooth()` using formula 'y ~ x'
```

```
# Google Search Frequency for Vegetarian Restaurant
ggplot(model.diag.metrics, aes(`Number of Restaurants in State`,
                               `Google Search Frequency for Vegetarian Restaurant`)) +
  geom_point() +
  stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = `Number of Restaurants in State`,
                   yend = .fitted), color = "red", size = 0.3)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
# Google Search Frequency for Vegan Restaurant
ggplot(model.diag.metrics, aes(`Number of Restaurants in State`,
                               `Google Search Frequency for Vegan Restaurant`)) +
  geom_point() +
  stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = `Number of Restaurants in State`,
                   yend = .fitted), color = "red", size = 0.3)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




### Interpretation

There does not appear to be a clear pattern in the residual plots, however New York at 110 and California at 33 are clearly outliers in this data set. This is most likely due to the fact that the data set contains only a portion of the full data set.

# 5 Chapter of Choice: Maps, Wordcloud and Bookdown

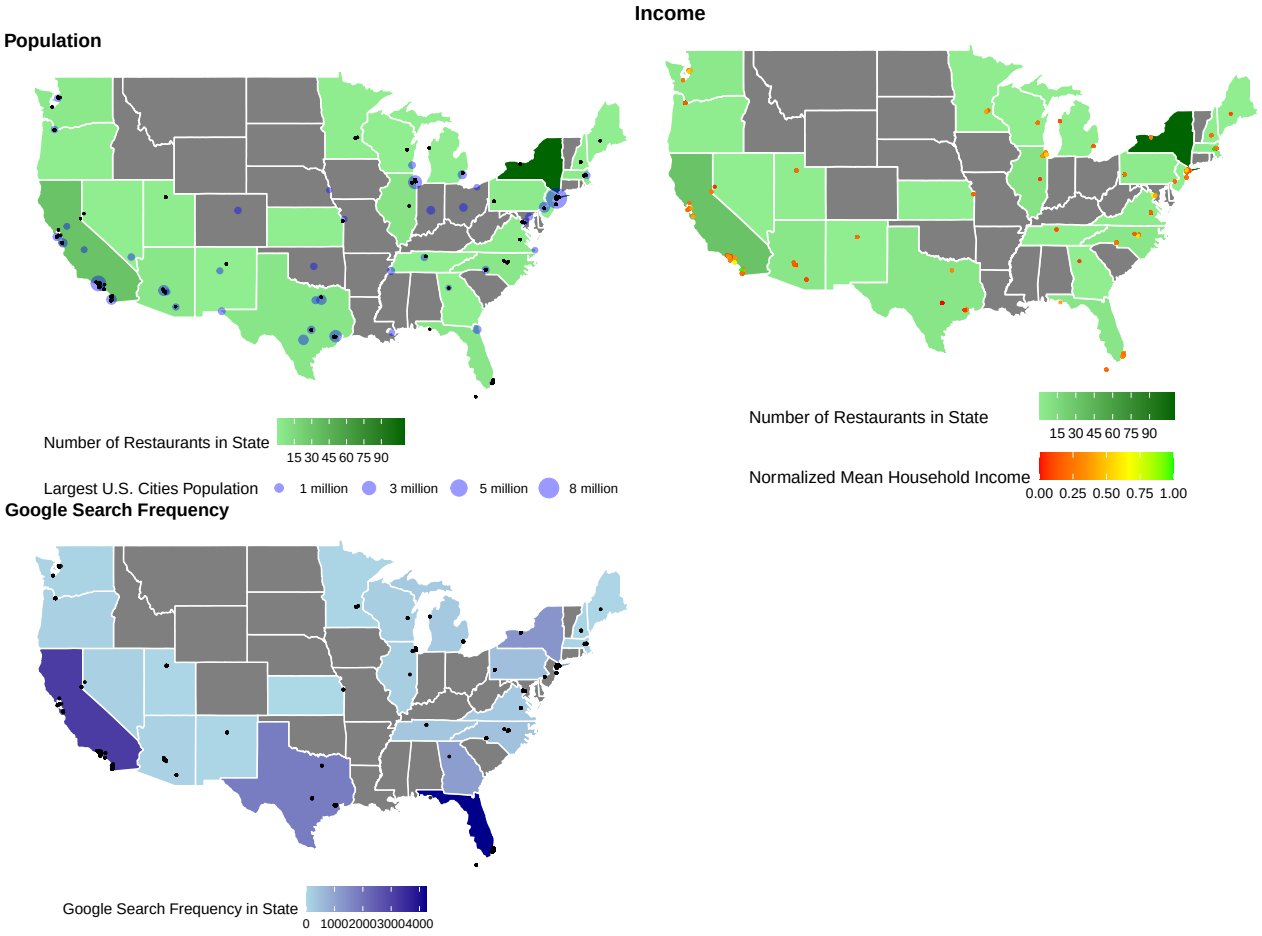
## 5.1 Maps

Since we were examining location data, the obvious choice was to visualize it as a map. The following code was used to visualize three different maps to show the relationship between population size, number of restaurants, and Google search frequency.

```
temp <- read.csv(
  curl("https://raw.githubusercontent.com/cphalpert/census-regions/master/us%20census%20bureau%20regions
)
states_map <- map_data("state")
states_map <- left_join(states_map, temp %>% mutate(State = tolower(State)) %>%
                          select(State, State.Code), by = c("region" = "State"))
google_search_frequency <- joined_data %>% group_by(province) %>%
  distinct(province, total_searches_vegan, total_searches_vegetarian) %>%
  summarise(google_search_frequency = total_searches_vegan + total_searches_vegetarian)

# Map Data
states_map_google_search_frequency <- left_join(states_map, google_search_frequency,
  by = c("State.Code" = "province"))
states_map_number_restaurants_state <- left_join(states_map, joined_data %>%
  group_by(province) %>%
  summarise(Number_Of_Restaurants = n_distinct(id)), by = c("State.Code" = "province"))
joined_data$MeanNormalized <- (joined_data$Mean - min(joined_data$Mean)) /
  (max(joined_data$Mean) - min(joined_data$Mean))

# Population
ggplot() +
  # Number of Restaurants by state
  geom_polygon(data = states_map_number_restaurants_state,
               aes(long, lat, group = group,
                   fill=Number_Of_Restaurants),
                   color = "white") +
  scale_fill_gradientn("Number of Restaurants in State",
                       colors = c('lightgreen', 'darkgreen'),
                       breaks = seq(from = 0, to = 100, by = 15)) +

  # mapping 44 largest cities in the U.S. since 45 is in hawaii
  geom_point(data=us.cities %>% arrange(pop) %>% tail(44),
             aes(x=long, y=lat, size = pop/1000000),
                 color = "blue", alpha = 0.4) +
  labs(size="Largest U.S. Cities Population") +
       scale_size(breaks=c(1, 3, 5, 8),
                  labels=c("1 million","3 million","5 million","8 million"),
                  guide="legend") +

  # mapping the veg. restaurants excluding hawaii (96753)
  geom_point(data=joined_data %>% filter(!str_detect(postalCode, "96753")),
                                        aes(x=longitude, y=latitude),
                                        size=0.3) +
  theme_void() + coord_map() +
  theme(plot.title = element_text(face = "bold"),
        legend.position="bottom",
        legend.box="vertical",
```

```
        legend.margin=margin(),
        legend.box.just = "left") +
  labs(title = "Population")

# Income
ggplot() +
  # Number of Restaurants by state
  geom_polygon(data = states_map_number_restaurants_state,
               aes(long, lat, group = group, fill=Number_Of_Restaurants),
               color = "white") +
  scale_fill_gradientn("Number of Restaurants in State         ",
                       colors = c('lightgreen', 'darkgreen'),
                       breaks = seq(from = 0, to = 100, by = 15)) +

  # mapping the veg. restaurants excluding hawaii (96753)
  geom_point(data=joined_data %>% filter(!str_detect(postalCode, "96753")),
                                        aes(x=longitude, y=latitude,
                                            color = MeanNormalized),
                                        size=0.5, alpha = 0.4) +
  scale_color_gradientn("Normalized Mean Household Income",
                        colors = c('red', 'darkorange', 'yellow', 'green'),
                        breaks = seq(from = 0, to = 1, by = 0.25)) +
  theme_void() + coord_map() +
  theme(plot.title = element_text(face = "bold"),
        legend.position="bottom",
        legend.box="vertical",
        legend.margin=margin(),
        legend.box.just = "left") +
  labs(title = "Income") +
  guides(fill = guide_colourbar(order = 1))

par(mar = c(0, 5, 0, 0))
# Google Search Frequency
ggplot() +
  # Number of Google Searches by State
  geom_polygon(data = states_map_google_search_frequency,
               aes(long, lat, group = group, fill=google_search_frequency),
               color = "white") +
  scale_fill_gradientn("Google Search Frequency in State",
                       colors = c('lightblue', 'darkblue'),
                       breaks = seq(from = 0, to = 5000, by = 1000)) +

  # mapping the veg. restaurants excluding hawaii (96753)
  geom_point(data=joined_data %>% filter(!str_detect(postalCode, "96753")),
                                        aes(x=longitude, y=latitude),
                                        size=0.5,
                                        alpha = 0.4,
                                        color = "black") +
  theme_void() + coord_map() +
  theme(plot.title = element_text(face = "bold"),
        legend.position="bottom",
        legend.box="vertical",
        legend.margin=margin(0,125),
```

```
        legend.box.just = "left") +
  labs(title = "Google Search Frequency")
```




### 5.1.1 Population by numbers of restaurants

Map one shows the population size in relation to the number of restaurants. It can be clearly seen that New York and California have quite high populations and are also home to the highest number of vegetarian and vegan restaurants. While in the state of New York the highest population is counted in the city of New York, in California there is more than a high population spread all over the state. In further analysis, it might be interesting to see where the most restaurants are located in each state and if this correlates with population size.

### 5.1.2 Mean household income by numbers of restaurants

Map two "Income" shows the mean household income in relation to the number of restaurants. It can be seen that based on our data set, there is not much correlation between the number of restaurants and the average household income.

### 5.1.3 Google search frequency by numbers of restaurants

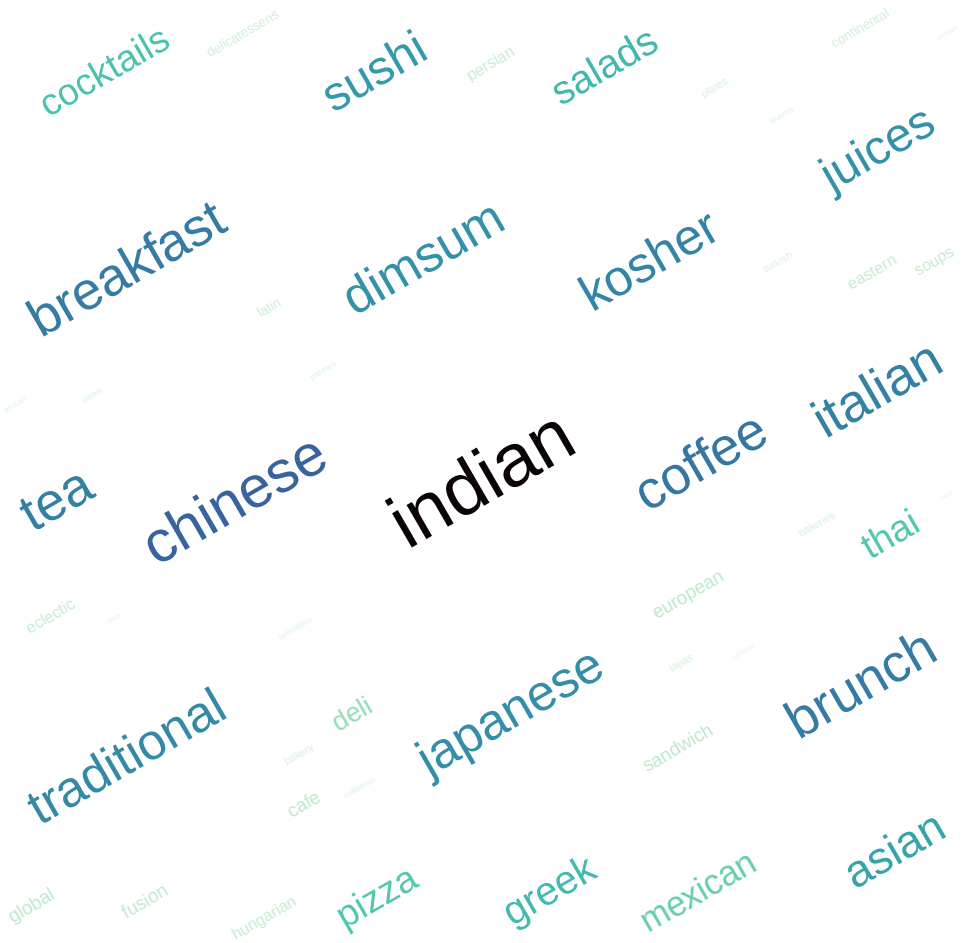
Map three "Google Search Frequency" shows the demand for vegetarian and vegan restaurants. Based on our dataset, it can be interpreted that the demand is quite high in Florida and California, followed by Texas and New York. If we disregard the low quality of the data, one could recommend restaurant owners to open a vegetarian or vegan restaurant in the mentioned states.

## 5.2 Wordcloud

Wordclouds were used to visually arrange words based on their frequency. Using wordcloud2 and ggworcloud, we were able to visualize an interactive and static wordcloud version with the most frequently mentioned cuisine types in our restaurant dataset. Indian seemed to be the most commonly offered cuisine in terms of vegetarian and vegan offerings.

```
# devtools::install_github("lchiffon/wordcloud2") fixed rmarkdown
df <- as.data.frame(do.call(rbind,
                            lapply(joined_data$cuisines_split, as.data.frame)))
colnames(df) <- c("cuisines")

# remove leading whitespaces
df$cuisines <- trimws(df$cuisines)

# additional fixes
df$cuisines[df$cuisines == 'Dim Sum'] <- 'DimSum'
df$cuisines[df$cuisines == 'Hot Dogs'] <- 'HotDogs'
df$cuisines[df$cuisines == 'Fast Food'] <- 'FastFood'
df$cuisines[df$cuisines == 'Deli Food'] <- 'Deli'
df$cuisines[df$cuisines == 'Ice Cream and Desserts'] <- 'IceCream and Desserts'

# remove nonsense words
df <- df %>% filter(!(str_detect(cuisines, "Restaurant|Friendly|Vegetarian|Vegan|Health|Bar|Gluten|Bist

# seperate words by slash; the word ' and '; whitespace
filtered <- df %>% separate_rows(cuisines, sep = "/") %>%
  separate_rows(cuisines, sep = " and ") %>%
  separate_rows(cuisines, sep = " ")
filtered <- filtered %>%
  filter(!(str_detect(cuisines, "New|Modern|Small|Shops|Lunch|Pacific|American")))

v <- sort(rowSums(
  as.matrix(TermDocumentMatrix(
    Corpus((VectorSource(filtered$cuisines))))
        )),decreasing=TRUE)
word_frequency_df <- data.frame(word = names(v),freq=v)

if(knitr::is_html_output()) {
  wordcloud_colors <- colorRampPalette(brewer.pal(11,"BrBG"))
  wordcloud2(word_frequency_df, size = 1, minRotation = pi/6, maxRotation = pi/6,
             rotateRatio = 1, color=rev(wordcloud_colors(32)[seq(1,32,1)]))
}

if (knitr::is_latex_output()) {
  ggplot(word_frequency_df, aes(label = word, size = freq,
                                color = freq, angle = 30)) +
  geom_text_wordcloud(shape = "square", rm_outside = TRUE, max_steps = 1,
                      grid_size = .1, eccentricity = .1) +
  scale_size_area(max_size = 10) +
  scale_color_viridis(option = "mako", direction = -1) + theme_void()
}
```

```
## Some words could not fit on page. They have been removed.
```




### 5.3 Bookdown

This paper was optimized using the Bookdown package:

- a custom title page done in Latex
- a table of contents with roman lettering
- custom header and footer notations
- references in APA7 using BibTex
- optimized exports for HTML (interactive) and PDF (static)

# 6 Reflection and Conclusion on the Project

At the beginning of the project, we very quickly came across the restaurant data from (Datafiniti, 2018). At first glance, the data met all the necessary parameters for a comprehensive analysis. Unfortunately, as mentioned earlier, it was only later that we realized that the dataset did not contain the full amount of data as stated in the description. This taught us that when using secondary data, the data must be very well validated.

However, we think that overall the subject was worth analyzing. We assumed that with better data quality or with the full data set, more interpretations could have been made. For the project now, we still found some interesting insights into vegetarian and vegan restaurant supply and demand.

If further analyzed with more accurate data, the project would lead to insights for potential restaurant owners to decide where to open vegetarian and vegan restaurants. Another group of interest could be the Vegetarian and Vegan Association to provide insight to the population.

# 7 Conclusion for ourselves

The R-Bootcamp week was very valuable for both of us and the content of the course covered a lot of important content for our further studies and professional life. It was a very steep learning curve for us to work independently on a project and to deepen and expand on the basics we were taught. We are both positively surprised by our learning progress.

# References

Datafiniti. (2018). *Vegetarian and vegan restaurants*. https://www.kaggle.com/datafiniti/vegetarian-vegan-restaurants

Google Trends. (n.d.a). *Vegan restaurant*. https://trends.google.com/trends/explore?geo=US&q=vegan%20restaurant

Google Trends. (n.d.b). *Vegetarian restaurant*. https://trends.google.com/trends/explore?geo=US&q=vegetarian%20restaurant

Michigan Population Studies Center. (2020). *Zip code characteristics: Mean and median household income*. https://www.psc.isr.umich.edu/dis/census/Features/tract2zip/