Google Data Analytics Professional Capstone

Case Study : Should Your City Have More Restaurants with Veg Options?

DISCLAIMER The vegan/vegatarian restaurants dataset used for this report has many potential sources of error. The dataset appears to be last updated in 2016, many incomplete values, and I couldn't validate if the dataset was a full aggregate of vegan/vegetarian restaurants. While the data was cleaned for the incomplete values, the other potential sources of error are still there. Correspondingly, any conclusion from this report should be viewed with skepticism.

1. Introduction

Restaurants with vegan and vegetarian options are an integral option for people who are vegan, vegetarian, flexitarian, pescatarian, and for those who simply want to try something different. While we know there are many restaurants out there, it is hard to know where the options are lacking for certain cities and states. By analyzing data on these restaurant locations and census information, I hope to understand which cities are ripe for more options.

2. Scenario

As a data analyst friend of someone who is open to moving anywhere with the intention of opening a vegan/vegetarian restaurant, I intend to supply them with the best city locations. I will refer to vegan/vegetarian as veg throughout this report for brevity. My friend also suspects that the density of these restaurants is likely contributed by the population's voting patterns. So they would like to ensure that this is not left off as a lurking variable.

Data Sources:

Thanks to these amazing public datasets this report was made possible! County Presidential Election Results:
MIT Election Data and Science Lab, 2018, "County Presidential Election Returns 2000-2020", https://doi.org/10.7910/DVN/VOQCHQ, Harvard Dataverse, V12; countypres_2000-2020.tab [fileName], UNF:6:KNR0/XNVzJC+RnAqIx5Z1Q== [fileUNF]

City Census Data: https://www.census.gov/data/tables/time-series/demo/popest/2020s-total-cities-and-towns.html#v2023

State Census Data: https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html

State Presidential Election Results: https://www.fec.gov/documents/4228/federalelections2020.xlsx

Vegan/Vegetarian Restaurants: https://www.kaggle.com/datasets/datafiniti/vegetarian-vegan-restaurants/versions/1?resource=download

3. Process

Tools: Since none of the datasets are very large, I decided to use Google Sheets to Combine and Clean

First step was to import all data sources into Google Sheets. Once the data was imported. all rows with empty city and province information were filtered out. Consequently any restaurant with a value of "TRUE" in the isCLOSED field were filtered out. To pare down to pertinent information I used the following query to get just city and state information:

```
=Query(US_Restaurants_List_Raw!A1:AC18156, "SELECT C,Z
```

Next was a couple of formulas to select distinct cities, and count the number of restaurants in each city:

```
=UNIQUE(A5:A17538)

=COUNTIF(A$5:A$17538,E4)
```

With this done, next was to combine this with city population data:

```
#Change city name to allow matches

=LEFT(B2,FIND("city",B2,1)-2)

#Find the matches and add to existing list
=VLOOKUP(E4,City_Populations!C$2:D$324,2,false))
```

This list was then filtered down to populations of over 300,000. Then, a value for population/Restaurant was calculated for each city.

```
Population per Restaurant = Census Population/# of restaurants
```

These steps were repeated for state population data. The next step was to investigate the presidential election results by state and city. To do this first the sheet with county presidential election results needed to be adjusted:

```
# Filter to the correct year of 2020

# Match county to city

=VLOOKUP(C3, 'City-County'!A$1:E$311121,2,FALSE)

# Create unique list

=UNIQUE(G3:G18069)

# Sum votes for the county for Democrat

=sumifs($E$3:$E$18069,$D$3:$D$18069,"DEMOCRAT",$G$3:$G$18069,$J3)

# Sum votes for the county for Republican

=SUMIFS($E$3:$E$18069,$D$3:$D$18069,"REPUBLICAN",$G$3:$G$18069,$J3)
```

With the knowledge of the Population per Restaurant for both the city and the state, the # of restaurants expected for the city to meet the state average can be found by the following:

```
Num of restaurants to meet state average = (People per Restaurant_city / People per Restaurant_state) *

#Normalize for population per 100,000

Num of restaurants to meet state average / City_Population * 100000
```

This leaves values for the number of restaurants to meet state average per 100,000 people in the city.

Next, a rough Liberalism score was calculated through this formula:

```
Liberalism Score = Democrat votes/(Republican Votes + Democrat Votes)
```

This process was then repeated for the states, though this process was much easier due to the format of the data.

These were graphed to see the strength of the relationship between the two. Using this relationship data was another way to calculate the number of extra veg restaurants a city needs. This was simply done by graphing city liberalism score by population/restaurant by city. Using this corresponding line of best fit calculation, and the same calculations done previously a number of veg restaurants needed per city normalized per 100,000 people was calculated.

To practice my skills in BigQuery and R, I ran some databaset processing through them as well. In R, I first loaded all of the libraries I would potentially need:

```
#load libraries
library(tidyverse) #calculations
library(lubridate) #dates
library(hms) #time
library(data.table) #exporting data frame
library(dplyr)
library(ggplot2)
```

Load csv file with restaurant data:

```
vegdata <- read.csv("C:\\Users\\djenn\\OneDrive\\Documents\\Case Study\\vegdataset.csv")
```

Take a quick look at the dataset

```
tibble::glimpse(vegdata)
```

Get column names and total number of restaurants

```
colnames(vegdata)
```

```
nrow(vegdata)
```

Further information for clarity:

```
#Get total number of restaurants by City
vegdata %>%
  select(city) %>%
  count(city)

#Get total number of restaurants by State
vegdata %>%
  select(province) %>%
  count(province)

#See how many restaurants are closed
vegdata %>%
  count(isClosed)

#Filter out closed restaurants
vegdata %>%
  filter(isClosed = TRUE )
```

Generate map plot based on coordinates:

```
ggplot(vegdata, aes(x = long, y = lat)) +
  geom_point() +
  coord_cartesian(xlim = c(-125, -65), ylim =c(24, 50)) +
  labs(x = "Longitude", y = "Latitude") +
  ggtitle("Location of Restaurants offering Veg Options") +
  theme(plot.title = element_text(face = "bold",
                                  margin = margin(10, 0, 10, 0),
                                  size = 16))
```

Then with BigQuery, joining the city census data with the city restaurant data. Filtering out cities with 300k people and under.

```
SELECT DISTINCT vegdata.city, vegdata.province, census_city.population, census_state.population
FROM vegdata
INNER JOIN vegdata ON vegdata.city = census_city.city
INNER JOIN census_state ON vegdata.province = census_state.state
WHERE census_city.city < 300000
AND vegdata.city is NOT NULL
AND vegdata.province is NOT NULL;
```

Counting the number of restaurants in the dataset that are not null:

```
SELECT COUNT(address)
FROM vegdata
WHERE address is NOT NULL
```

Then counting the number of unique cities in the restaurant database:

```
SELECT COUNT(DISTINCT city) AS [Count]
FROM vegdata
WHERE city is NOT NULL
```

Separating and cleaning up the county political results:

```
SELECT county, party, votes
FROM political_county
WHERE party = 'DEMOCRAT'
AND year = 2020
ORDER BY county

SELECT county, party, votes
FROM political_county
WHERE party = 'REPUBLICAN'
AND year = 2020
ORDER BY county
```

4. Analysis and Share

The data analysis has been stored and prepared for a further analysis. Visualizations come from Google Sheets, R Studio, and Tableau. The objective of the analysis was to determine which cities would be best suited for a new vegan/vegetarian restaurant and to see if there is an existing political bias that could act as a lurking variable.

To first visualize this data geographically maps were made in Tableau and R Studio showing them:

Now following onto the data processing done previously, the political analysis by state corresponding to vegan restaurant density resulted in the following:

The corresponding R squared value for this was 0.26 which indicates a moderate-weak correlation. I will posit that further analysis to determine whether the data can be deemed normal was not done.

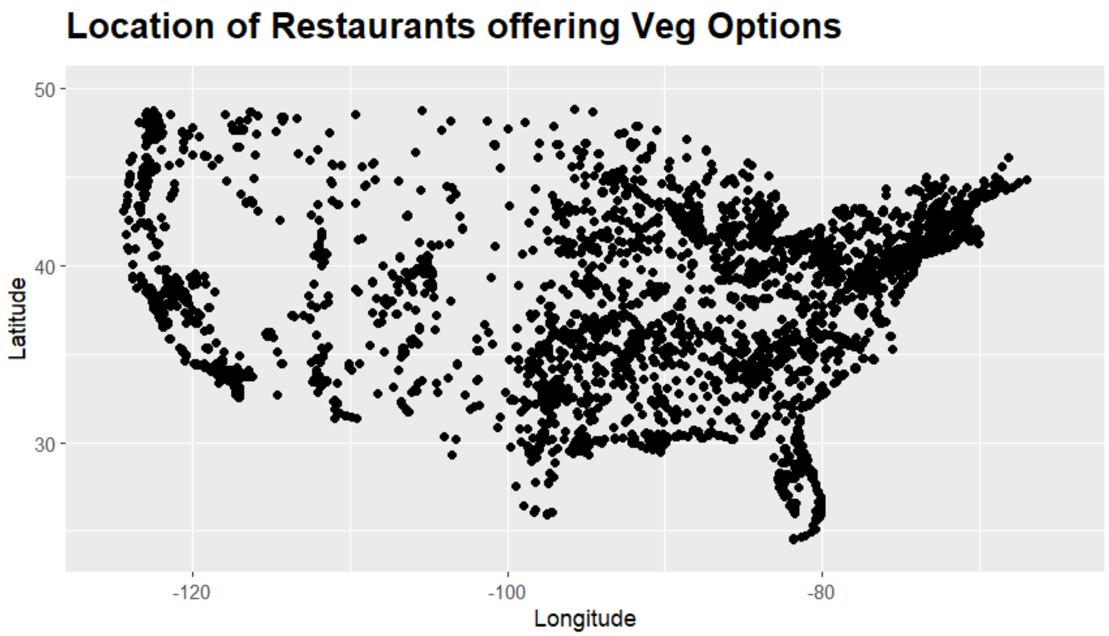
After this, the political analysis by city corresponding to vegan restaurant density resulted in the following:

The corresponding R squared value for this 0.21 which is also a moderate-weak correlation. However, using this as explained in the process stage, the variation of the city's from this line of best fit was done.

With this information in hand, it can be shown which cities veg restaurant density are the lowest in comparison to their expected amount. As explained in the Process step, the first is the corresponding calculations from the veg restaurant density by city variance to the expected values by liberalism score:

The other method explained in the process step for determining this top ten was the veg restaurant density by city variance to the veg restaurant density by state. That top ten resulted in:

Taking the average of the values from these two charts resulted in the combined chart.



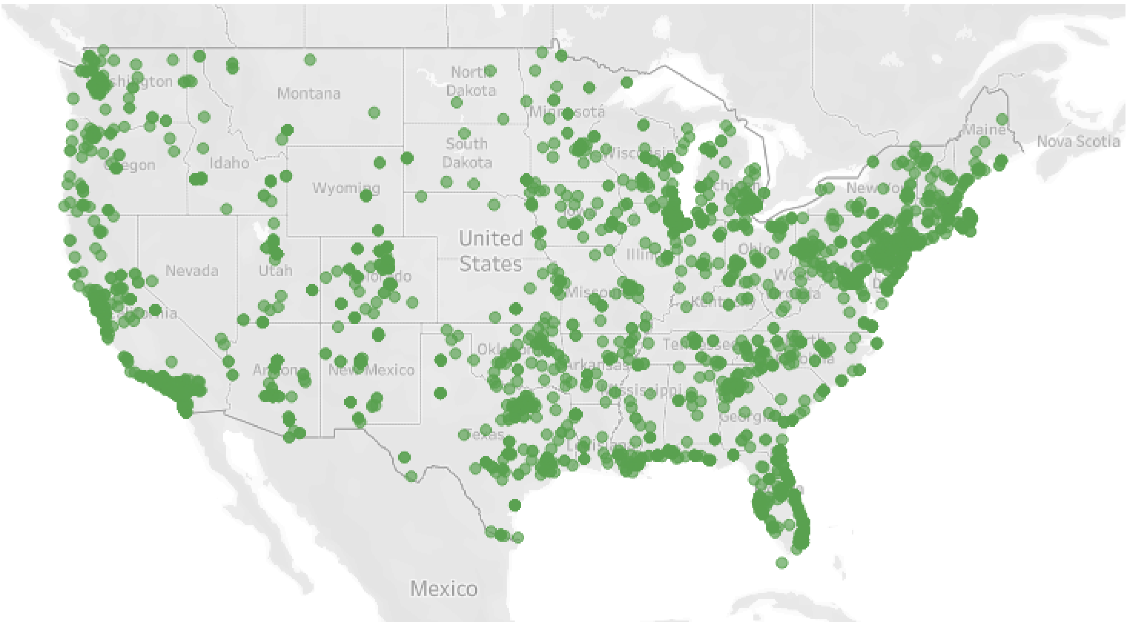
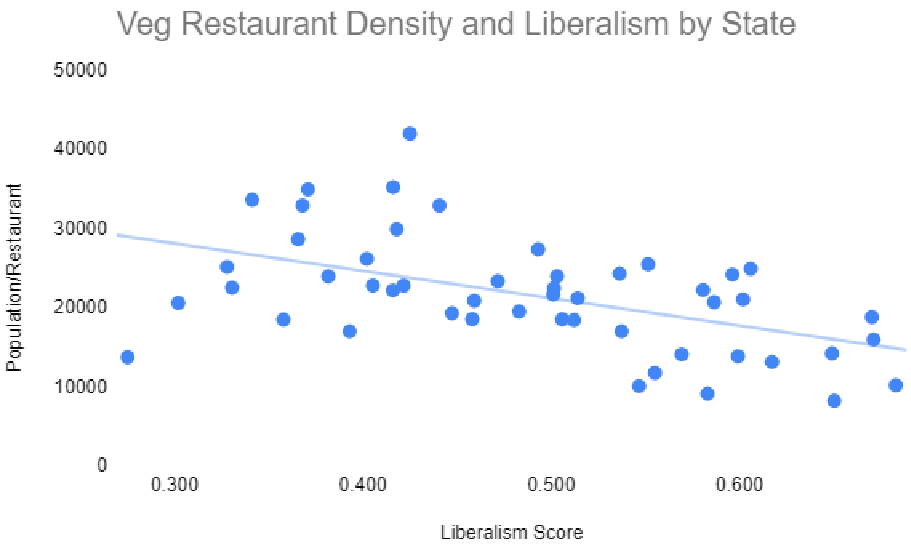
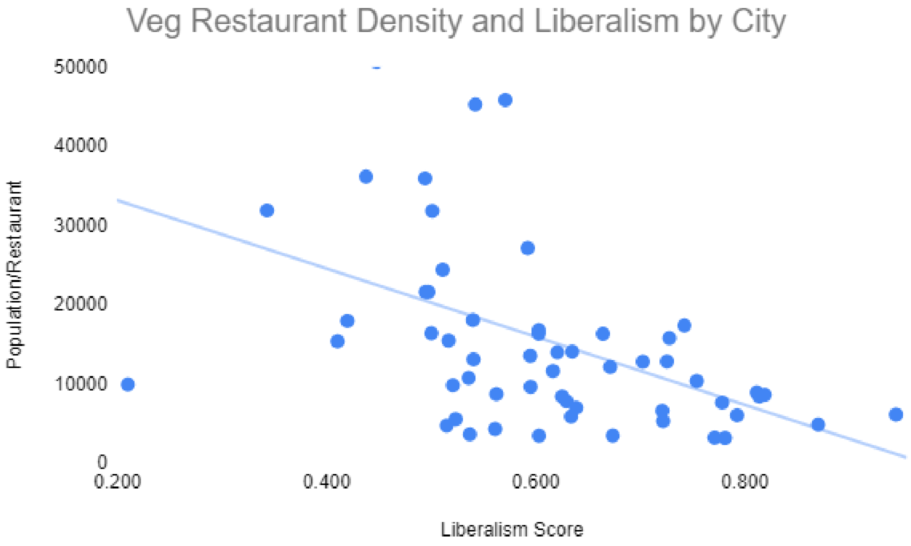
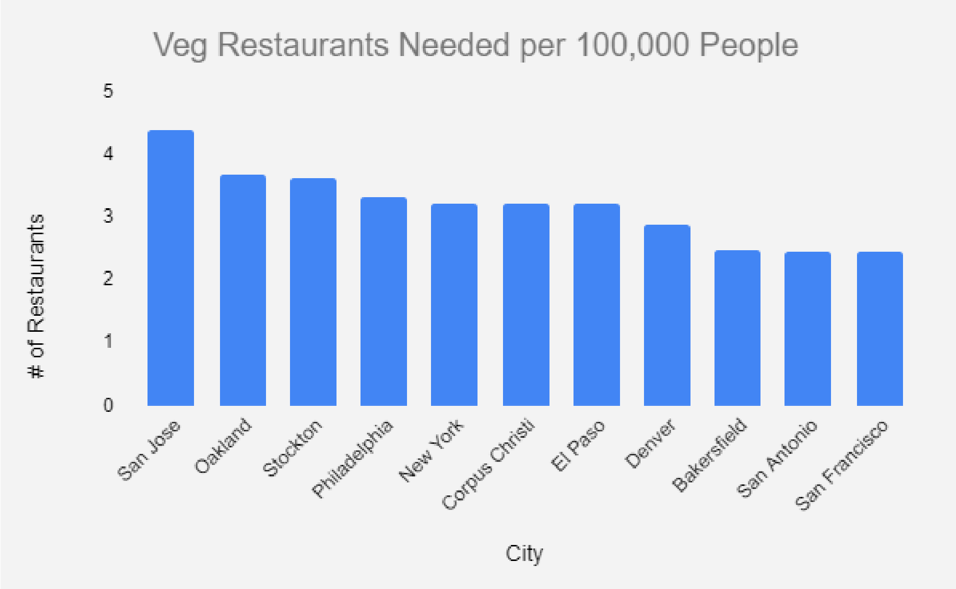
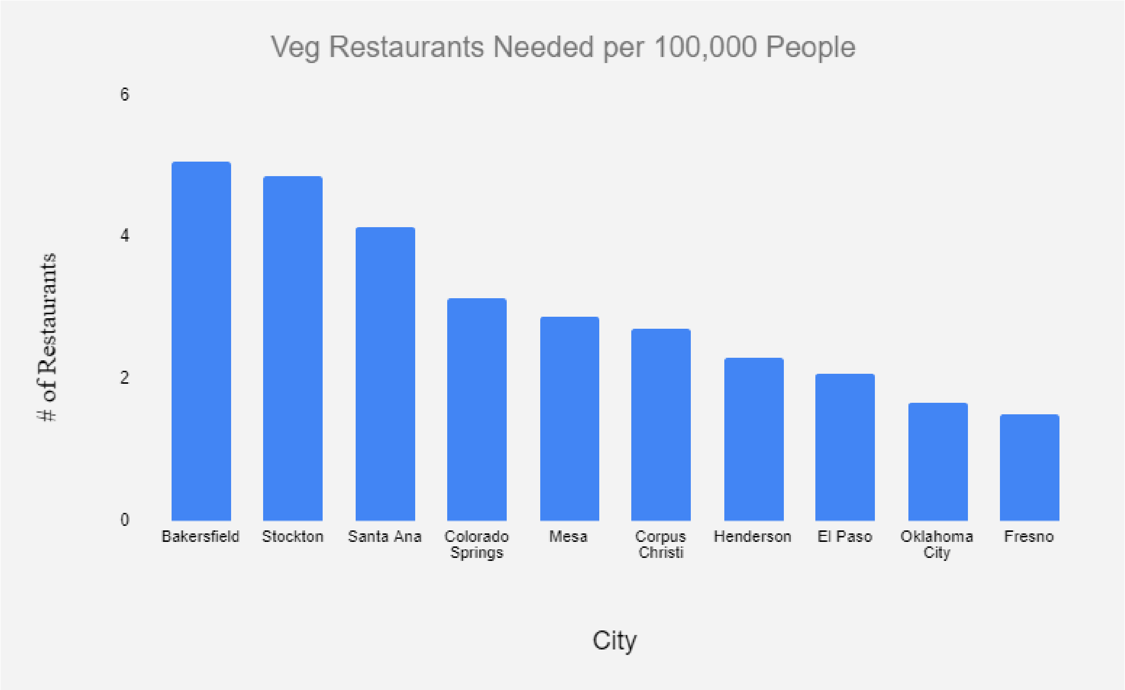


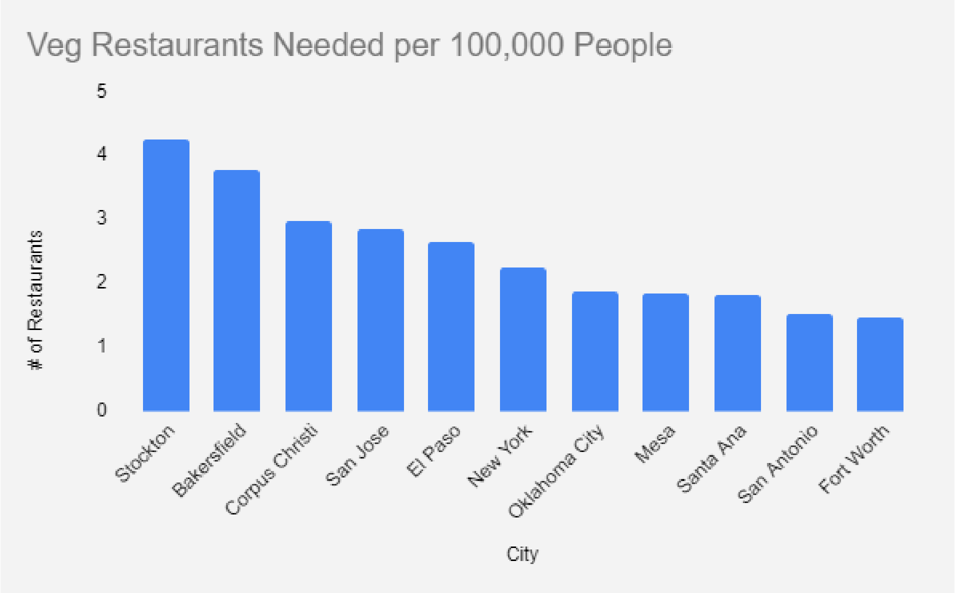
Figure 1: Tableau Version


Veg Restaurant Density and Liberalism by State
50000
40000
30000
20000
10000
0
Population/Restaurant
0.300
0.400
0.500
0.600
Liberalism Score


Veg Restaurant Density and Liberalism by City
50000
40000
30000
20000
10000
0
Population/Restaurant
0.200
0.400
0.600
0.800
Liberalism Score


Veg Restaurants Needed per 100,000 People
# of Restaurants
0
1
2
3
4
5
San Jose
Oakland
Stockton
Philadelphia
New York
Corpus Christi
El Paso
Denver
Bakersfield
San Antonio
San Francisco
City


Veg Restaurants Needed per 100,000 People
# of Restaurants
0
2
4
6
Bakersfield
Stockton
Santa Ana
Colorado Springs
Mesa
Corpus Christi
Henderson
El Paso
Oklahoma City
Fresno
City



6. Act

After analyzing the information on the veg restaurant density by city, there are definitely cities with the opportunity to open more veg restaurants. Using the values shown through the analysis phase I would recommend the following cities as the best cities for opening a veg restaurant:

- Stockton
- Bakersfield
- Corpus Christi
- San Jose
- El Paso

Those five cities had the highest combined average of deviation from expected based upon liberalism scores and state averages. While I recommend these cities, to anyone else out there, please open a veg restaurant anywhere you can and want to!