Joann Lin

April 4, 2020

IBM Data Science Professional Certificate

Capstone Project

BATTLE OF THE NEIGHBORHOODS

Where to Live in America? A Comparison of Cost, Food, and Serial Killers

1. INTRODUCTION

The United States covers 3.797 million square miles across North America. As of 2020, there is an estimated 331.3 million people living within its 50 states. Each of the 50 states has its own unique characteristics. Some states, like New York, are home to booming metropolitan areas. Some hold beautiful vast national parks, like Arizona. Others still have the quiet rustic farmlands, like Oklahoma. Some states, such as California, contain a bit of everything.

Deciding where to move can be difficult. There are various factors to consider; jobs, housing, schools, etc. For this capstone project, I will dive into three:

- Cost of living
- Variety of food choices available
- The likelihood of coming face to face with a serial killer

Yes, those are three very different factors for one to consider. However, together, I believe they can make a more compelling case for states that do not receive as much attention.

My target audience is people looking for an inexpensive state to live, with a lot of food options, and do not want to face off with a serial killer!

2. DATA

For my analysis, I will be using three data sources:

- Foursquare
- developers.google.com
- worldpopulationreview.com

2.1 Foursquare

https://developer.foursquare.com/docs/

Utilizing the Foursquare API, I will obtain a samples different types of food venues available per state. I will then analyze and compare each state to determine the level of diversity. For example, if one state has 6 types of food venues in the sample of 10, they are more diverse than a state that only has 7 types of food venues in the sample of 20.
https://developer.foursquare.com/docs/build-with-foursquare/categories/

I will also use the Category Hierarchy provided by Foursquare to specify food-related venues (e.g. restaurants, fast food, etc).

2.2 developers.google.com

https://developers.google.com/public-data/docs/canonical/states_csv

I will use the latitudes and longitudes for each state provided here in conjunction with the Foursquare API.

2.2 worldpopulationreview.com (Cost of Living by State)

https://worldpopulationreview.com/states/cost-of-living-index-by-state/

There are various types of cost of living indexes that use different variables and metrics. Most indexes set 100 as the national average cost of living. This is a widely accepted average as a basis for comparison.

I will utilize the general cost of living index provided by worldpopulationreview.com in order to compare the 50 states and Washington D.C. Per their calculations, each state falls on a range between 86 and 193. Virginia, for example, is a 100.7, which is right about the national average.

This data also contains other more specific cost of living measurements, such as utilities and transportation. However, I will not be analyzing those variables in this project.

2.3 worldpopulationreview.com (Serial Killers by State)

https://worldpopulationreview.com/states/serial-killers-by-state/

This data contains the number of (known) serial killer victims in each state. It also contains the total population of each state (as of 2020). Based on the information provided, I will calculate the probability of encountering a serial killer in every state. For example, Virginia has 238 confirmed serial killer victims and a population of 8,626,207. Therefore, if you live in that state, you have a 0.0028% chance of a serial killer encounter.

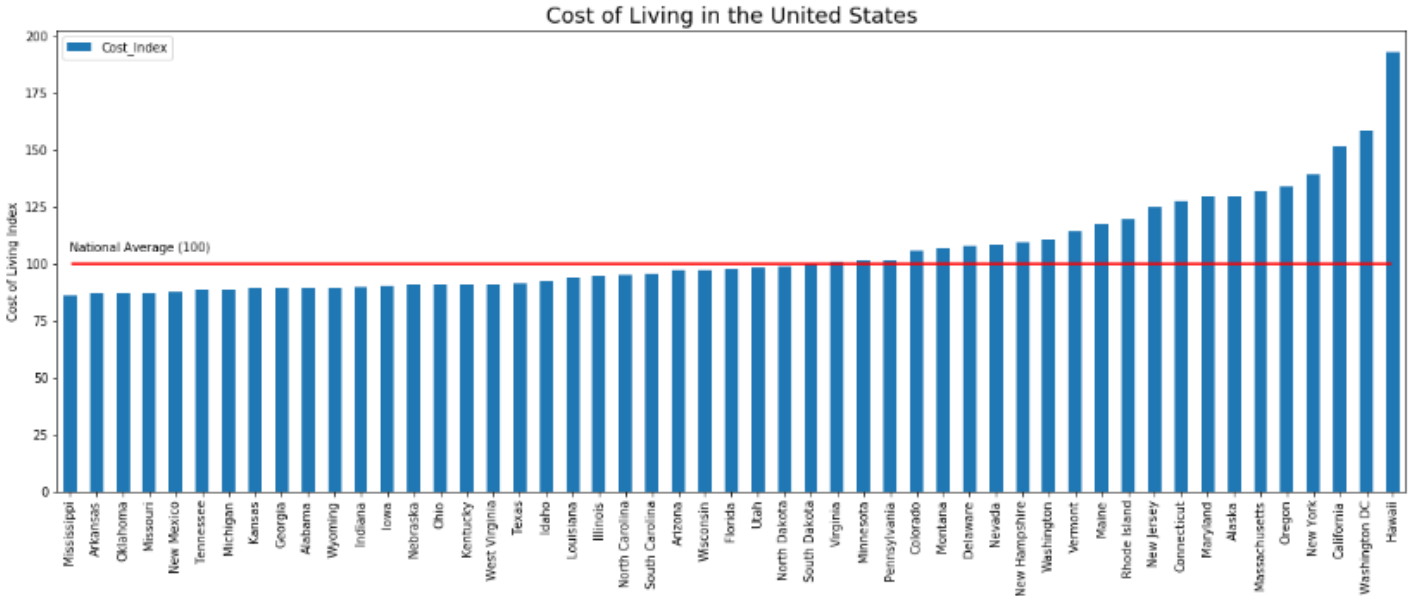
3. Methodolgy

3.1 Cost of Living

Here is a ranking of the 50 states and Washington DC, from lowest to highest cost of living.

	State	Cost_Index
0	Mississippi	86.1
1	Arkansas	86.9
2	Oklahoma	87.0
3	Missouri	87.1
4	New Mexico	87.5
5	Tennessee	88.7
6	Michigan	88.9
7	Kansas	89.0
8	Georgia	89.2
9	Alabama	89.3
10	Wyoming	89.3
11	Indiana	90.0
12	Iowa	90.1
13	Nebraska	90.8
14	Ohio	90.8
15	Kentucky	90.9
16	West Virginia	91.1
17	Texas	91.5
18	Idaho	92.3
19	Louisiana	93.9
20	Illinois	94.5
21	North Carolina	94.9
22	South Carolina	95.9
23	Arizona	97.0
24	Wisconsin	97.3
25	Florida	97.9
26	Utah	98.4
27	North Dakota	98.8
28	South Dakota	99.8
29	Virginia	100.7
30	Minnesota	101.6
31	Pennsylvania	101.7
32	Colorado	105.6
33	Montana	106.9
34	Delaware	108.1
35	Nevada	108.5
36	New Hampshire	109.7
37	Washington	110.7
38	Vermont	114.5
39	Maine	117.5
40	Rhode Island	119.4
41	New Jersey	125.1
42	Connecticut	127.7
43	Maryland	129.7
44	Alaska	129.9
45	Massachusetts	131.6
46	Oregon	134.2
47	New York	139.1
48	California	151.7
49	Washington DC	158.4
50	Hawaii	192.9

Here is a bar chart for a better visual comparison. The red line represents the national average of 100.



3.2 Food Variety

I will take a sample of 50 restaurants from each state and analyze the variety available.

First I will take the coordinates of all 50 states and Washington DC. Sample below.

	Latitude	Longitude	State
0	63.588753	-154.493062	Alaska
1	32.318231	-86.902298	Alabama
2	35.201050	-91.831833	Arkansas
3	34.048928	-111.093731	Arizona
4	36.778261	-119.417932	California

Now I will retrieve food venues information from Foursquare.

	State	State Latitude	State Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alaska	63.588753	-154.493062	Princess Theatre	63.588757	-154.493064	American Restaurant
1	Alaska	63.588753	-154.493062	Trident Grill	63.582062	-154.476957	Burger Joint
2	Alabama	32.318231	-86.902298	Charliey's Grilled Subs	32.318231	-86.902298	Fast Food Restaurant
3	Arkansas	35.201050	-91.831833	ﻟﻄ	35.201050	-91.831833	German Restaurant
4	Arizona	34.048928	-111.093731	Preston Main Mango Tree	34.048928	-111.093731	Food Court
5	Arizona	34.048928	-111.093731	Graziano's Pizzeria	34.048928	-111.093731	Pizza Place
6	Arizona	34.048928	-111.093731	Pete:s Fish In Chip	34.048927	-111.093735	Fast Food Restaurant
7	Arizona	34.048928	-111.093731	Mdonalds	34.048927	-111.093735	Burger Joint
8	Arizona	34.048928	-111.093731	Four Seasons Hotel " Restaurant Zafferano"	34.048928	-111.093731	Restaurant
9	Arizona	34.048928	-111.093731	Tortilla Lady	34.049077	-111.095808	Taco Place
10	Arizona	34.048928	-111.093731	Chester's Fried Chicken	34.056408	-111.089401	Fried Chicken Joint
11	California	36.778261	-119.417932	McDonald's	36.776593	-119.418501	Fast Food Restaurant
12	California	36.778261	-119.417932	Holy Aioli	36.778259	-119.417931	Food Truck
13	California	36.778261	-119.417932	Chester's Fried Chicken	36.778259	-119.417931	Fried Chicken Joint
14	California	36.778261	-119.417932	Bembos	36.778259	-119.417930	Burger Joint
15	California	36.778261	-119.417932	Pienburger Truck	36.777212	-119.419144	Food Truck
16	California	36.778261	-119.417932	Pescaderia DoÑa Elsa	36.787625	-119.420866	Seafood Restaurant
17	California	36.778261	-119.417932	Gaucho	36.785095	-119.426514	BBQ Joint
18	Colorado	39.550051	-105.782067	Hacienda Real	39.550054	-105.782065	Mexican Restaurant
19	Colorado	39.550051	-105.782067	Soda El Cevichito	39.549667	-105.781784	Seafood Restaurant

There are some general venue types (e.g. Food Court), which does not provide insight to the type of food it serves. Therefore, I will manually identify and remove those types.

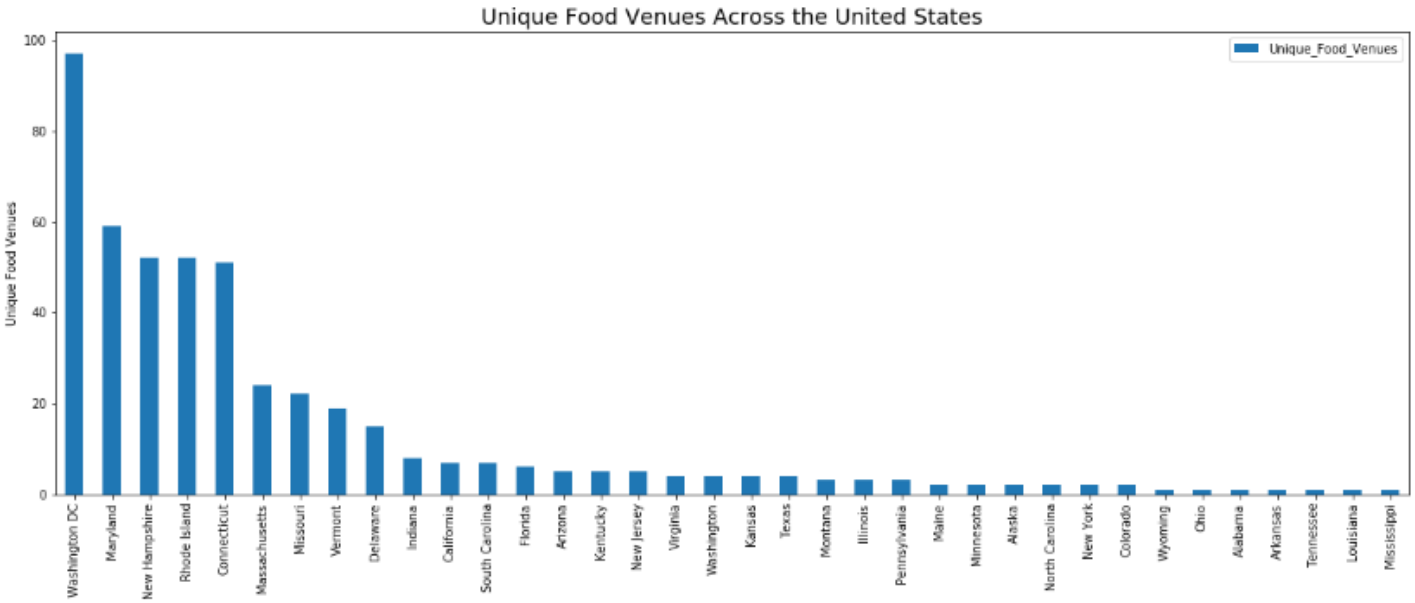
	State	State Latitude	State Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alaska	63.588753	-154.493062	Princess Theatre	63.588757	-154.493064	American Restaurant
1	Alaska	63.588753	-154.493062	Trident Grill	63.582062	-154.476957	Burger Joint
2	Alabama	32.318231	-86.902298	Charliey's Grilled Subs	32.318231	-86.902298	Fast Food Restaurant
3	Arkansas	35.201050	-91.831833	ظ	35.201050	-91.831833	German Restaurant
5	Arizona	34.048928	-111.093731	Graziano's Pizzeria	34.048928	-111.093731	Pizza Place
6	Arizona	34.048928	-111.093731	Pete:s Fish In Chip	34.048927	-111.093735	Fast Food Restaurant
7	Arizona	34.048928	-111.093731	Mdonalds	34.048927	-111.093735	Burger Joint
9	Arizona	34.048928	-111.093731	Tortilla Lady	34.049077	-111.095808	Taco Place
10	Arizona	34.048928	-111.093731	Chester's Fried Chicken	34.056408	-111.089401	Fried Chicken Joint
11	California	36.778261	-119.417932	McDonald's	36.776593	-119.418501	Fast Food Restaurant

I am left with the unique food venues in each state out of my sample.

Now I will count how many unique food venues there are in each state. Here is a sample.

	State	Unique_Food_Venues
0	Washington DC	97
1	Maryland	59
2	New Hampshire	52
3	Rhode Island	52
4	Connecticut	51

Here is the above data in a bar chart for visualization.



3.3 Serial Killers

Now I will examine the number of serial killer victims per state.

	State	Victims	Population	Probability
0	Washington DC	170	720687	0.00024
1	Alaska	51	734002	0.00007
2	Louisiana	300	4645184	0.00006
3	Kansas	153	2910357	0.00005
4	Indiana	341	6745354	0.00005
5	Missouri	311	6169270	0.00005
6	Washington	390	7797095	0.00005
7	Illinois	629	12659682	0.00005
8	Oklahoma	195	3954821	0.00005
9	Wyoming	27	567025	0.00005
10	California	1628	39937489	0.00004
11	Oregon	170	4301089	0.00004
12	Nebraska	76	1952570	0.00004
13	Florida	845	21992985	0.00004
14	Michigan	381	10045029	0.00004
15	Montana	41	1086759	0.00004
16	Ohio	433	11747694	0.00004
17	Connecticut	122	3563077	0.00003
18	Georgia	365	10736059	0.00003
19	Arkansas	103	3038999	0.00003
20	Nevada	105	3139658	0.00003
21	Alabama	164	4908621	0.00003
22	Pennsylvania	420	12820878	0.00003
23	New York	628	19440469	0.00003
24	South Carolina	162	5210095	0.00003
25	New Mexico	64	2096640	0.00003
26	Maryland	185	6083116	0.00003
27	Texas	893	29472295	0.00003
28	Kentucky	134	4499692	0.00003
29	Massachusetts	195	6976597	0.00003
30	Virginia	238	8626207	0.00003
31	Colorado	155	5845526	0.00003
32	Tennessee	179	6897576	0.00003
33	Mississippi	75	2989260	0.00003
34	North Carolina	266	10611862	0.00003
35	Rhode Island	25	1056161	0.00002
36	Utah	77	3282115	0.00002
37	New Jersey	207	8936574	0.00002
38	Maine	30	1345790	0.00002
39	Arizona	156	7378494	0.00002
40	West Virginia	37	1778070	0.00002
41	Idaho	37	1826156	0.00002
42	Wisconsin	118	5851754	0.00002
43	Vermont	11	628061	0.00002
44	Iowa	53	3179849	0.00002
45	North Dakota	11	761723	0.00001
46	Delaware	14	982895	0.00001
47	Minnesota	68	5700671	0.00001
48	South Dakota	7	903027	0.00001
49	New Hampshire	10	1371246	0.00001
50	Hawaii	10	1412687	0.00001

There are several numbers to consider. First, I will look at the highest victim count.

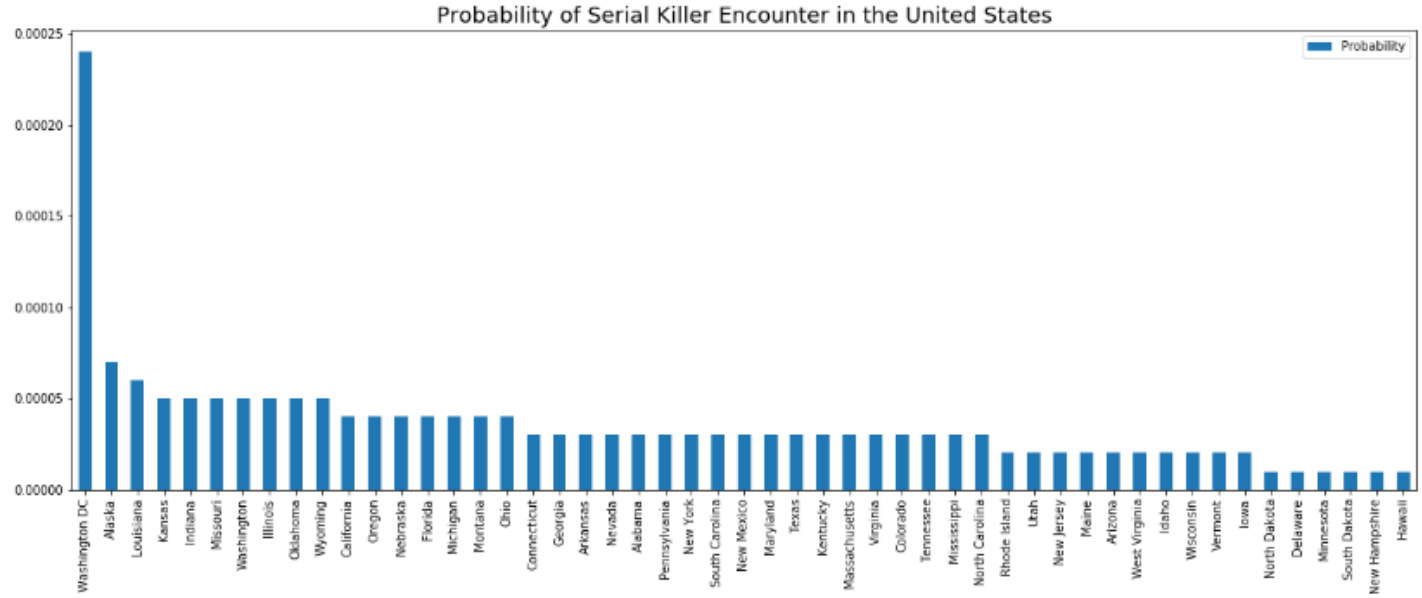
	State	Victims	Population	Probability
10	California	1628	39937489	0.00004
27	Texas	893	29472295	0.00003
13	Florida	845	21992985	0.00004
7	Illinois	629	12659682	0.00005
23	New York	628	19440469	0.00003

We can see the top 5 states with the highest victim count. However, those states do have a higher population, so it is expected that they would have more victims.

The probability in relation to the total population may be a better indicator.

	State	Victims	Population	Probability
0	Washington DC	170	720687	0.00024
1	Alaska	51	734002	0.00007
2	Louisiana	300	4645184	0.00006
3	Kansas	153	2910357	0.00005
4	Indiana	341	6745354	0.00005

Here is the probability in a bar chart for visualization.



4. Results

4.1 10 Highest and Lowest Results

From the three analyses done above, here are the 10 highest and lowest results of all three variables.

10 States with the LOWEST cost of living:

	State	Cost_Index
0	Mississippi	86.1
1	Arkansas	86.9
2	Oklahoma	87.0
3	Missouri	87.1
4	New Mexico	87.5
5	Tennessee	88.7
6	Michigan	88.9
7	Kansas	89.0
8	Georgia	89.2
9	Alabama	89.3

10 States (+ Washington DC) with the HIGHEST cost of living:

	State	Cost_Index
50	Hawaii	192.9
49	Washington DC	158.4
48	California	151.7
47	New York	139.1
46	Oregon	134.2
45	Massachusetts	131.6
44	Alaska	129.9
43	Maryland	129.7
42	Connecticut	127.7
41	New Jersey	125.1

We can see that Mississippi has the lowest cost of living at 86.1 compared to the national average of 100, while Hawaii by far has the highest at 192.9.

10 States (+ Washington DC) with the MOST number of unique food venues:

	State	Unique_Food_Venues
0	Washington DC	97
1	Maryland	59
2	New Hampshire	52
3	Rhode Island	52
4	Connecticut	51
5	Massachusetts	24
6	Missouri	22
7	Vermont	19
8	Delaware	15
9	Indiana	8

10 States (+ Washington DC) with the LEAST number of unique food venues:

	State	Unique_Food_Venues
35	Mississippi	1
29	Wyoming	1
30	Ohio	1
34	Louisiana	1
32	Arkansas	1
33	Tennessee	1
31	Alabama	1
23	Maine	2
24	Minnesota	2
27	New York	2

As you can see, there are definitely built in issues with the data. The smaller states (and Washington DC) have a significantly higher number of unique food venues. The bigger states have less. This is due to the radius value. I set the radius at 5000 in order to capture more area in the states.

For example, New York is a large state. However, the majority of the population lives in one small area of New York, and therefore, all the available restaurants are centered in that one small area.

Capturing such a large area unfortunately covers too much area for the smaller states, and therefore captures more unique food venues. Washington DC, being the smallest area, as it is just a city, expectedly shows the highest number of food venues.

We can also see that some states are also missing. This is because the 5000 radius did not pick up any food venues within the vicinity of the search.

Obviously, this is not a perfect indication of food diversity of each state. However, it will give us some indication of the availability of each state. Additionally, the other two variables also factor into the analysis. For the purpose of this project, I will keep the food venue variable as part of the anaylsis.

10 States with the LOWEST possibility of encountering a serial killer:

	State	Probability
50	Hawaii	0.00001
45	North Dakota	0.00001
46	Delaware	0.00001
49	New Hampshire	0.00001
48	South Dakota	0.00001
47	Minnesota	0.00001
44	Iowa	0.00002
42	Wisconsin	0.00002
41	Idaho	0.00002
40	West Virginia	0.00002

10 States with the HIGHEST possibility of encountering a serial killer:

	State	Probability
0	Washington DC	0.00024
1	Alaska	0.00007
2	Louisiana	0.00006
3	Kansas	0.00005
4	Indiana	0.00005
5	Missouri	0.00005
6	Washington	0.00005
7	Illinois	0.00005
8	Oklahoma	0.00005
9	Wyoming	0.00005

Now this gets interesting. People in Hawaii, North Dakota, Delaware, New Hampshire, South Dakota, and Minnesota have the lowest probability of encountering a serial killer at 0.00001%. On the flip side, people living in Washington DC have the highest probability of encountering a serial killer at 0.00024%. That is higher than the probability of being dealt a royal flush in poker (which is 0.000154%, according to Wikipedia). How scary!

4.2 All States (+ Washington DC) and All Results

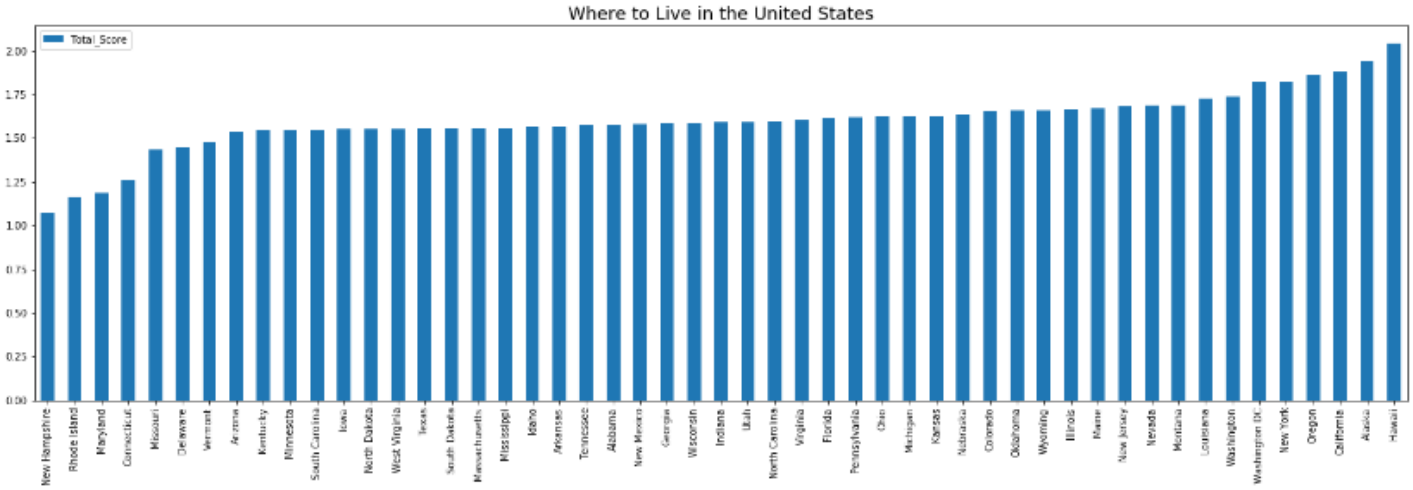
I combined all results below. I'm also replacing all NaN with "0". I will normalize the data in order to better analyze the correlation of the variables. I will also inverse the numbers for Unique_Food_Venues to match Cost_Index and Probability in that the lower the value, the better. Sample Below.

	State	Cost_Index	Unique_Food_Venues	Probability
0	Hawaii	1.000000	1.000000	0.041667
22	South Dakota	0.517367	1.000000	0.041667
20	Minnesota	0.526698	0.979381	0.041667
16	Delaware	0.560394	0.845361	0.041667
14	New Hampshire	0.568688	0.463918	0.041667
23	North Dakota	0.512182	1.000000	0.041667
26	Wisconsin	0.504406	1.000000	0.083333
27	Arizona	0.502851	0.948454	0.083333
32	Idaho	0.478486	1.000000	0.083333
11	Maine	0.609124	0.979381	0.083333

Now I will add values of the Cost_Index, Unique_Food_Venues, and Probability. This allows me to combine all three variables in order to compare each state.

State	Total_Score
New Hampshire	1.074273
Rhode Island	1.166224
Maryland	1.189122
Connecticut	1.261228
Missouri	1.433058
Delaware	1.447421
Vermont	1.481029
Arizona	1.534638
Kentucky	1.544682
Minnesota	1.547746
South Carolina	1.549984
Iowa	1.550415
North Dakota	1.553849
West Virginia	1.555599
Texas	1.558102
South Dakota	1.559033
Massachusetts	1.559796
Mississippi	1.561036
Idaho	1.561820
Arkansas	1.565183
Tennessee	1.574514
Alabama	1.577625
New Mexico	1.578603
Georgia	1.587416
Wisconsin	1.587740
Indiana	1.592422
Utah	1.593442
North Carolina	1.596346
Virginia	1.605795
Florida	1.612328
Pennsylvania	1.621288
Ohio	1.627068
Michigan	1.627527
Kansas	1.628475
Nebraska	1.637377
Colorado	1.651815
Oklahoma	1.659344
Wyoming	1.660958
Illinois	1.667297
Maine	1.671839
New Jersey	1.680309
Nevada	1.687468
Montana	1.689912
Louisiana	1.726471
Washington	1.740969
Washington DC	1.821151
New York	1.825480
Oregon	1.862364
California	1.880920
Alaska	1.944454
Hawaii	2.041667

Here is the bar chart of the final result.



5. Discussion

From the visualization of the total scores for each state, I can observe the following:

- New Hampshire, Rhode Island, Maryland, and Connecticut have the lowest scores. This means they have some combination of low cost of living, more food venues, and low possibility of encountering a serial killer.
- Hawaii, Alaska, California, Oregon, New York, and Washington DC have the highest scores. This means they have some combination of high cost of living, less food venues, and high possibility of encountering a serial killer. However, it doesn't mean all variables apply to all states. For example, Hawaii has the lowest probability of a serial killer encounter. However, the cost of living there is by far the highest, which weighed heavily on its score. Alaska and Washington DC, on the other hand, cover more of the variables.
- A majority of the 50 states are very average.
- The states with the lower scores (better) tend to be the smaller states, while the states with the higher scores (worse) tend to be the larger states (Hawaii and Washington DC are exceptions).

Based on my observations, here are my recommendations on where to live and where to avoid in the US:

- New Hampshire has the lowest score. It is cheaper and safer, relative to other states. This is a good state to consider moving to.
- Definitely do not move to Hawaii, as the cost of living there is insanely high. It is very safe though.
- I would also not move to Washington DC. The numbers for that city alone is so high, it was broken out into its own separate location. The actual state of Maryland itself scored very low and is a good place to consider moving to.
- Definitely avoid Alaska, as it hits all three down sides: expensive, minimal food options, and also a serial killer hot spot.

6. Conclusion

The United States is a large country with many options of where to live. Utilizing three variables (cost of living, food variety, and serial killer encounters), I built a ranking system which consider all three variables and provides a score for each state (and Washington DC).

I obtained cvs files from open sources (cost of living, US states coordinates, and serial killers) and utilized the Foursquare API to complete my analysis. I took the results and created bar charts for better visualization. Normalizing the data then allowed me to compare each state against another.

All of these steps lead me to the conclusion that New Hampshire, Rhode Island, Maryland, and Connecticut are states I recommend to consider moving to. I also recommend avoiding Hawaii, Alaska, California, Oregon, New York, and Washington DC.