Classifications of cities in US

Table of contents

- Introduction: Business Problem
- Data
- Methodology
- Analysis
- Results and Discussion
- Conclusion

Introduction

Description & Discussion of the Background

United States is one of the top three countries with the highest population in the world, based on the 2020 census the total population was **331 million** people, all the living in a land of **9.1 million square kilometers** that represents **.027 square kilometer** per person. Moreover, every year the country graduates **1.2 million** students from a bachelor's degree, this represents a high demand on new employment as well as for the decision of where to live. Is well know that one of the important decision makers is to have places to visit and have time to relax or a variety of things to do, not living aside that the income has an enormous factor to make a verdict.

when I analyze this situation I identified the need to have an overview of the cities that has the highest income as well as the highest land spaces to live, providing and insight for those who are looking to have a house with enough space to raise kids and have a big place to live, further, this project includes the top venues of those top cities in order to provide more information to take a better decision.

Data

Data Description

Golden Oak Research Group LLC, "U.S. Income Database Kaggle". Publication: 5, August 2017. Accessed, 29, June 2021.

I Will use Income database taken from the Golden oak project which has the information about the cities with the below columns:

Household & Geographic Statistics:

- Mean Household Income (double)
- Median Household Income (double)
- Standard Deviation of Household Income (double)
- Number of Households (double)
- Square area of land at location (double)
- Square area of water at location (double)

Geographic Location:

- Longitude (double)
- Latitude (double)
- State Name (character)
- State abbreviated (character)
- State_Code (character)
- County Name (character)
- City Name (character)
- Name of city, town, village or CPD (character)
- Primary, Defines if the location is a track and block group.
- Zip Code (character)
- Area Code (character)

Foursquare API Data:

To gain more information about the cities that are identified as more relevant based on initial analysis, we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Venue
2. Name of the venue e.g. the name of a store or restaurant
3. Venue Latitude
4. Venue Longitude
5. Venue Category

Map of US states

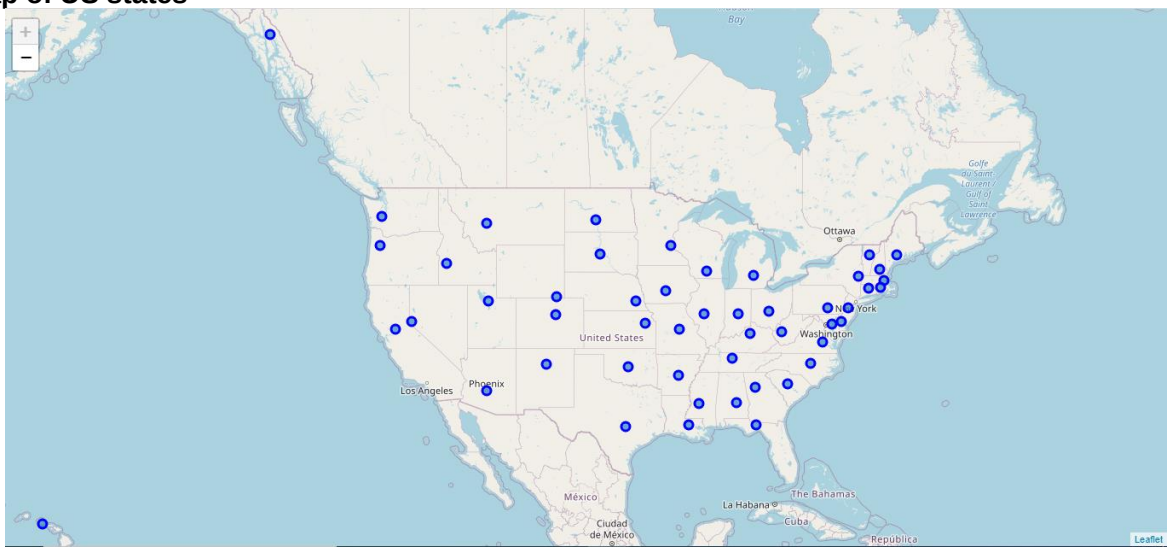


Figure 1

Methodology

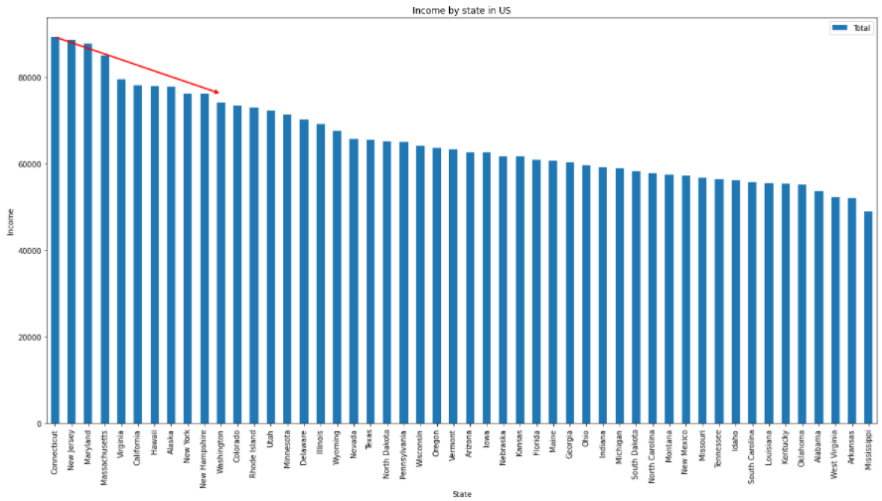
Creating data table and data pre-processing

As mentioned on the data section, for this analysis I used the data collected on the project “U.S. Income Database Kaggle” with the main components *city, state, land size, income, latitude, and longitude* as well as additional information that won't be used on this project. Additionally, I used the information from the census portal to collect the capital of the states and with this, to identify on a map the center point of each state.

Once I had our dataset with the necessary information, I first make a graph to visualize the average income per state and highlight which states will have more demand in terms of graduate students looking for a job.

	State_Name	Total
6	Connecticut	89227.219718
29	New Jersey	88657.644144
19	Maryland	87689.604096
20	Massachusetts	84878.683582
45	Virginia	79401.740127
4	California	78126.737805
10	Hawaii	77859.586957
1	Alaska	77670.209524
31	New York	76201.220833
28	New Hampshire	76113.503817

Table 1



Graph 1

Analysis

First insights and Visual Maps

First of all we can see by using the bar graph, the top 10 states in relation to the highest income, this provide the first sign that those state should have the top cities in the country, but in this first step of analysis we can provide now recommendations on which states you should focus your job hunting.

	State_Name	Total	Capital	latitude	longitude
0	Connecticut	89227.219718	Hartford 	41.764046	-72.682198
1	New Jersey	88657.644144	Trenton	40.220596	-74.769913
2	Maryland	87689.604096	Annapolis	38.978764	-76.490936
3	Massachusetts	84878.683582	Boston	42.358162	-71.063698
4	Virginia	79401.740127	Richmond	37.538857	-77.433640
5	California	78126.737805	Sacramento	38.576668	-121.493629
6	Hawaii	77859.586957	Honolulu	21.307442	-157.857376
7	Alaska	77670.209524	Juneau	58.301598	-134.420212
8	New York	76201.220833	Albany	42.652843	-73.757874
9	New Hampshire	76113.503817	Concord	43.206898	-71.537994

Table 2

Now that we have the states, we create a map to visualize the region where is concentrated the highest income, this help us to highlight that most of the states are located on the east side of the country. Another insight is that none of the states located on the center, south, or north of the country, are considered with high income.

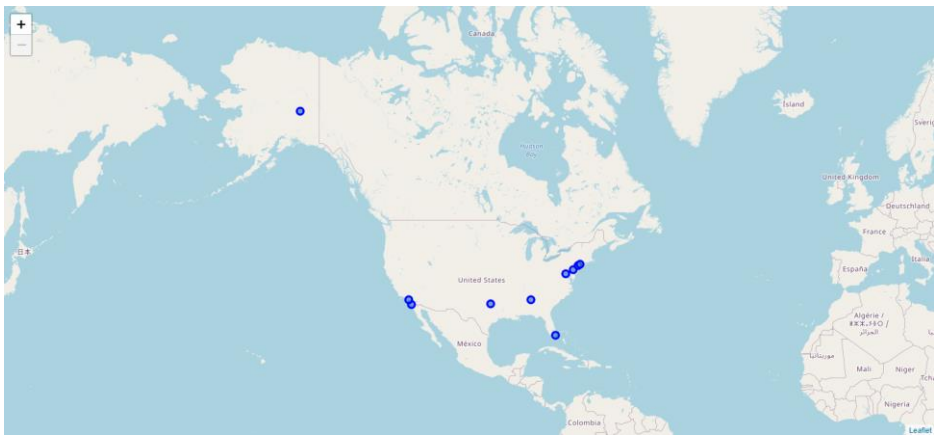


Figure 2

On the second step of the analysis, we focus on the cities to see if the behavior is similar as with the state results and to do that, we will take the top 10 highest income cities.

	State_Name	City	ALand	Lat	Lon	Mean	Capital
560	Alaska	Delta Junction	18298887	64.152838	-145.906385	242857	Juneau
3341	California	San Diego	1961071	32.737719	-117.197744	242857	Sacramento
56	Alabama	Odenville	27893577	33.691576	-86.503766	242857	Montgomery
24213	Pennsylvania	West Chester	604077	39.933520	-75.530306	242857	Harrisburg
18080	New Jersey	Short Hills	9094477	40.739665	-74.342521	216503	Trenton
19990	New York	Bronxville	1805685	40.938765	-73.823654	209392	Albany
5974	Florida	Miami Beach	1420761	25.756717	-80.140066	207128	Tallahassee
27162	Texas	Dallas	2883569	32.858013	-96.799359	206380	Austin
30017	Virginia	Leesburg	8876960	39.096570	-77.595294	205835	Richmond
2910	California	Huntington Beach	3060683	33.679397	-118.020316	203910	Sacramento

Table 3

We now have the top 10 cities with the highest income, and we will proceed to visualize on a map where are those cities located.

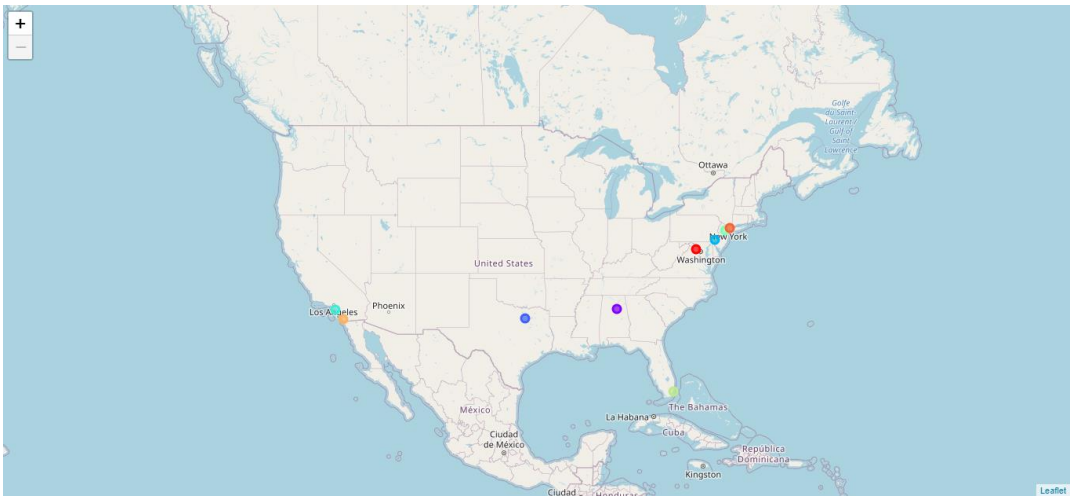


Figure 3

The result of the map confirms that the states that has in average the highest income in general are the same with the highest cities, with a mayor focus on the east side of the country.

Let's utilizing the Foursquare API to explore the boroughs and segment them.

First, we will need to create the function that will help us to collect the venue information that will show us additional data to take more complete decision.

In this step we will use the top 10 list to collect the information that Foursquare has about them.

```
1 Delta Junction
2 San Diego
3 Odenville
```

```
4 West Chester
5 Short Hills
6 Bronxville
7 Miami Beach
8 Dallas
9 Leesburg
10 Huntington Beach
```

Let's see what where the initial results of the consult.

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	San Diego	32.737719	-117.197744	Southwest Airlines	32.732094	-117.196897	Airport Service
1	San Diego	32.737719	-117.197744	Barnett Avenue Adult Superstore	32.745927	-117.197552	Video Store
2	San Diego	32.737719	-117.197744	MCRD Fitness Center	32.743753	-117.198761	Gym / Fitness Center
3	San Diego	32.737719	-117.197744	MCRD San Diego Museum	32.742742	-117.194319	Museum
4	San Diego	32.737719	-117.197744	Stone Brewing	32.732268	-117.203878	Beer Bar

Table 4

City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Bronxville	43	43	43	43	43	43
Dallas	45	45	45	45	45	45
Huntington Beach	9	9	9	9	9	9
Leesburg	4	4	4	4	4	4
Miami Beach	15	15	15	15	15	15
Odenville	1	1	1	1	1	1
San Diego	57	57	57	57	57	57
Short Hills	7	7	7	7	7	7
West Chester	5	5	5	5	5	5

Table 5

We now can categorize the venues to visualize more in detail what each city offers.

The result doesn't mean that inquiry run all the possible results in boroughs. It depends on given Latitude and Longitude information and here is we just run single Latitude and Longitude pair for each borough. We can increase the possibilities with Neighborhood information with more Latitude and Longitude information.

Let's find out how many unique categories can be curated from all the returned venues.

```
There are 93 unique categories.
```

Using one hot encoding

For this part of the analysis, we will used one hot encoding to normalize the data and be able to clustery.

Next, let's group rows by borough and by taking the mean of the frequency of occurrence of each category

We analyze the top 5 venues by city with the goal to identify the frequency that the category appears on the top 5 and with this make the clusters.

```
----Bronxville----
                venue  freq
0                Pool  0.05
1                 Bar  0.05
2         Coffee Shop  0.05
3  Italian Restaurant  0.05
4  Mexican Restaurant  0.05
```

```
----Dallas----
                  venue  freq
0                Bakery  0.09
1   American Restaurant  0.09
2         Sandwich Place  0.04
3  Gym / Fitness Center  0.04
4           Cupcake Shop  0.04
```

```
----Huntington Beach----
            venue  freq
0       Surf Spot  0.33
1      Playground  0.22
2     Golf Course  0.22
3    Tennis Court  0.11
4  Scenic Lookout  0.11
```

```
----Leesburg----
           venue  freq
0   Home Service  0.25
1           Park  0.25
2      Dive Spot  0.25
3           Pool  0.25
4  Movie Theater  0.00
```

```
----Miami Beach----
                venue  freq
0  Italian Restaurant  0.13
1       Boat or Ferry  0.13
2          Restaurant  0.07
3     Harbor / Marina  0.07
4              Resort  0.07


----Odenville----
                venue  freq
0  Electronics Store   1.0
1            Airport   0.0
2      Movie Theater   0.0
3         Playground   0.0
4        Pizza Place   0.0


----San Diego----
                venue  freq
0     Airport Service  0.11
1         Coffee Shop  0.07
2      Airport Lounge  0.07
3  Chinese Restaurant  0.04
4          Bagel Shop  0.04


----Short Hills----
          venue  freq
0  Home Service  0.14
1   Golf Course  0.14
2        School  0.14
3          Pool  0.14
4        Bakery  0.14


----West Chester----
            venue  freq
0  Scenic Lookout   0.2
1            Park   0.2
2     Music Venue   0.2
3  Farmers Market   0.2
4      Smoke Shop   0.2
```

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bronxville	Bar	Pool	Pharmacy	Italian Restaurant	Coffee Shop	Restaurant	Mexican Restaurant	Mediterranean Restaurant	Playground	Park
1	Dallas	American Restaurant	Bakery	Sandwich Place	Gym / Fitness Center	Men's Store	Seafood Restaurant	Shoe Store	Cupcake Shop	Playground	Coffee Shop
2	Huntington Beach	Surf Spot	Playground	Golf Course	Tennis Court	Scenic Lookout	Department Store	Chinese Restaurant	Cocktail Bar	Coffee Shop	Convenience Store
3	Leesburg	Home Service	Park	Dive Spot	Pool	Airport Terminal	Electronics Store	Cosmetics Shop	Airport Service	Cupcake Shop	Cycle Studio
4	Miami Beach	Boat or Ferry	Italian Restaurant	Beach	Cocktail Bar	Restaurant	Resort	Harbor / Marina	Gym / Fitness Center	Sushi Restaurant	Tennis Court

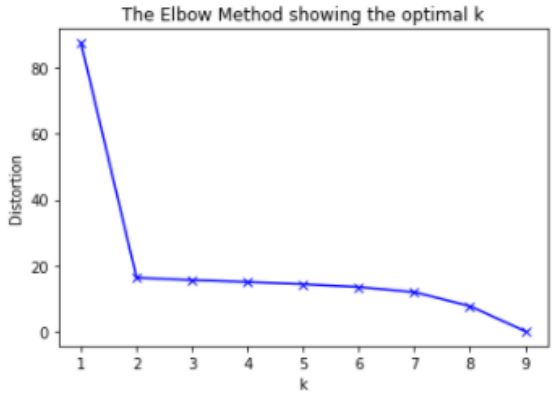
Table 6

On the table above we can see that there is a lot of categories on this dataset, this will cause to clustered in many groups.

Cluster of Boroughs

K-Means algorithm is one of the most common cluster methods of unsupervised learning. I will use K-Means algorithm for my study in this project.

I used the elbow method to identify the best degree for the K-Means and ensured the optimum result.



Graph 2

	State_Name	City	ALand	Lat	Lon	Mean	Capital	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
560	Alaska	Delta Junction	18298887	64.152838	-145.906385	242857	Juneau	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3341	California	San Diego	1961071	32.737719	-117.197744	242857	Sacramento	8.0	Airport Service	Airport Lounge	Coffee Shop	Chinese Restaurant	Bar	Ice Cream Shop	Pizza Place	Fast Food Restaurant	Café
56	Alabama	Odenville	27893577	33.691576	-86.503766	242857	Montgomery	1.0	Electronics Store	Wine Bar	Coffee Shop	Convenience Store	Cosmetics Shop	Cupcake Shop	Cycle Studio	Deli / Bodega	Department Store
24213	Pennsylvania	West Chester	604077	39.933520	-75.530306	242857	Harrisburg	3.0	Smoke Shop	Park	Music Venue	Scenic Lookout	Farmers Market	Discount Store	Coffee Shop	Convenience Store	Cosmetics Shop
18080	New Jersey	Short Hills	9094477	40.739665	-74.342521	216503	Trenton	5.0	Home Service	Bakery	Golf Course	Pool	School	Baseball Field	Hockey Rink	Asian Restaurant	Convenience Store
19990	New York	Bronxville	1805685	40.938765	-73.823654	209392	Albany	7.0	Bar	Pool	Pharmacy	Italian Restaurant	Coffee Shop	Restaurant	Mexican Restaurant	Mediterranean Restaurant	Playground
5974	Florida	Miami Beach	1420761	25.756717	-80.140066	207128	Tallahassee	6.0	Boat or Ferry	Italian Restaurant	Beach	Cocktail Bar	Restaurant	Resort	Harbor / Marina	Gym / Fitness Center	Sushi Restaurant
27162	Texas	Dallas	2883569	32.858013	-96.799359	206380	Austin	2.0	American Restaurant	Bakery	Sandwich Place	Gym / Fitness Center	Men's Store	Seafood Restaurant	Shoe Store	Cupcake Shop	Playground
30017	Virginia	Leesburg	8876960	39.096570	-77.595294	205835	Richmond	0.0	Home Service	Park	Dive Spot	Pool	Airport Terminal	Electronics Store	Cosmetics Shop	Airport Service	Cupcake Shop
2910	California	Huntington Beach	3060683	33.679397	-118.020316	203910	Sacramento	4.0	Surf Spot	Playground	Golf Course	Tennis Court	Scenic Lookout	Department Store	Chinese Restaurant	Cocktail Bar	Coffee Shop

Table 7

The next step on the analysis was to clean the clusters and categories that doesn't have data and work with useful ones.

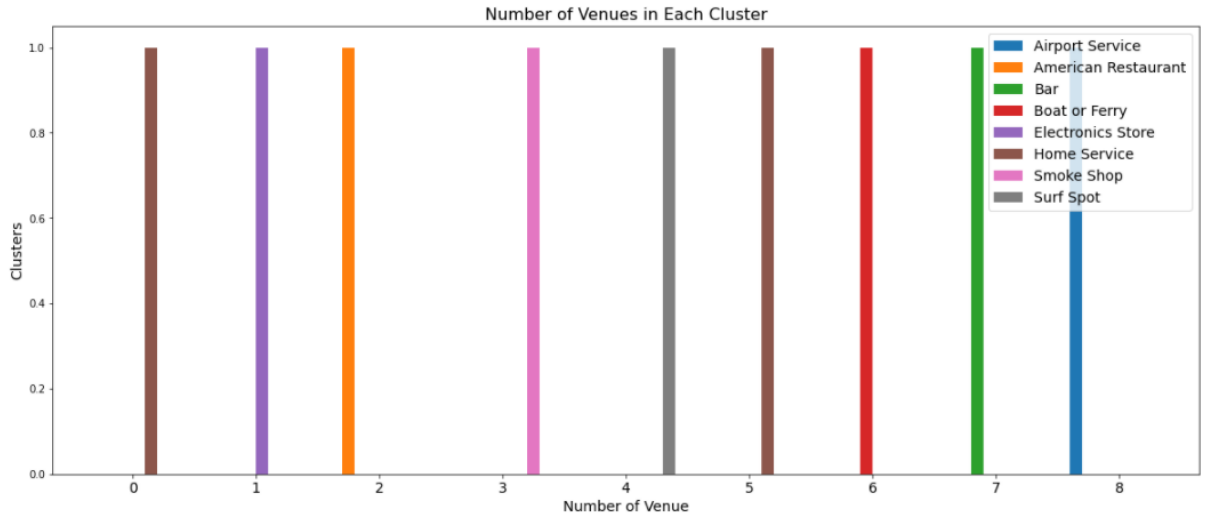
	State_Name	City	ALand	Lat	Lon	Mean	Capital	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	California	San Diego	1961071	32.737719	-117.197744	242857	Sacramento	8	Airport Service	Airport Lounge	Coffee Shop	Chinese Restaurant	Bar	Ice Cream Shop	Pizza Place	Fast Food Restaurant	Café	Mexican Restaurant
1	Alabama	Odenville	27893577	33.691576	-86.503766	242857	Montgomery	1	Electronics Store	Wine Bar	Coffee Shop	Convenience Store	Cosmetics Shop	Cupcake Shop	Cycle Studio	Deli / Bodega	Department Store	Discount Store
2	Pennsylvania	West Chester	604077	39.933520	-75.530306	242857	Harrisburg	3	Smoke Shop	Park	Music Venue	Scenic Lookout	Farmers Market	Discount Store	Coffee Shop	Convenience Store	Cosmetics Shop	Cupcake Shop
3	New Jersey	Short Hills	9094477	40.739665	-74.342521	216503	Trenton	5	Home Service	Bakery	Golf Course	Pool	School	Baseball Field	Hockey Rink	Asian Restaurant	Convenience Store	Cupcake Shop
4	New York	Bronxville	1805685	40.938765	-73.823654	209392	Albany	7	Bar	Pool	Pharmacy	Italian Restaurant	Coffee Shop	Restaurant	Mexican Restaurant	Mediterranean Restaurant	Playground	Park

Table 8

We can also estimate the number of **1st Most Common Venue** in each cluster. Thus, we can create a bar chart which may help us to find proper label names for each cluster.

1st Most Common Venue	Airport Service	American Restaurant	Bar	Boat or Ferry	Electronics Store	Home Service	Smoke Shop	Surf Spot
0	0	0	0	0	0	1	0	0
1	0	0	0	0	1	0	0	0
2	0	1	0	0	0	0	0	0
3	0	0	0	0	0	0	1	0
4	0	0	0	0	0	0	0	1
5	0	0	0	0	0	1	0	0
6	0	0	0	1	0	0	0	0
7	0	0	1	0	0	0	0	0
8	1	0	0	0	0	0	0	0

Table 9



Graph 3

To conclude with the analysis, we can do a map to see how was distributed the clusters and additionally we present each cluster venues bases con the label that was used to identify the group.

Results

Main tables with results:

Cluster 1

	ALand	Capital	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
7	8876960	Richmond	0	Home Service	Park	Dive Spot	Pool	Airport Terminal	Electronics Store	Cosmetics Shop	Airport Service	Cupcake Shop	Cycle Studio

Cluster 2

	ALand	Capital	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	27893577	Montgomery	1	Electronics Store	Wine Bar	Coffee Shop	Convenience Store	Cosmetics Shop	Cupcake Shop	Cycle Studio	Deli / Bodega	Department Store	Discount Store

Cluster 4

	ALand	Capital	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	604077	Harrisburg	3	Smoke Shop	Park	Music Venue	Scenic Lookout	Farmers Market	Discount Store	Coffee Shop	Convenience Store	Cosmetics Shop	Cupcake Shop

Cluster 5

	ALand	Capital	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
8	3060683	Sacramento	4	Surf Spot	Playground	Golf Course	Tennis Court	Scenic Lookout	Department Store	Chinese Restaurant	Cocktail Bar	Coffee Shop	Convenience Store

Cluster 6

	ALand	Capital	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	9094477	Trenton	5	Home Service	Bakery	Golf Course	Pool	School	Baseball Field	Hockey Rink	Asian Restaurant	Convenience Store	Cupcake Shop

Cluster 7

	ALand	Capital	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
5	1420761	Tallahassee	6	Boat or Ferry	Italian Restaurant	Beach	Cocktail Bar	Restaurant	Resort	Harbor / Marina	Gym / Fitness Center	Sushi Restaurant	Tennis Court

Discussion

We have been able to see how the income information in the United States in the first instance can give us an important reference that allows us to help recent graduates to make a decision about their future as to where is the best place to live taking into consideration In this area, it is clear that to make more conclusive recommendations it must be necessary to involve other variables that were not considered in this dataset, such as security, mobility or even labor supply and demand. Foursquare provides you with relatively simple information but to be able to relate to the locality and help recommend, in the same way, a place to live, you must consider other variables.

Conclusion

From this analysis I can conclude that there are 3 main cities to be considered as the best option to find a job, those are **Delta Junction, San Diego and Odenville**, and important hint from this data is that they are most focus on a certain region, the distribution on the country brings an opportunity to everyone to be part of those communities. Another topic to highlight is that some of the most important states on the country has at least 1 city on the top 10 like **New York, Texas, California, and Florida**, this represents the impact that has on the economy for the country and how they can capture the most talented persons due to the capacity to pay higher salaries.

References:

- [1] US Household Income Statistics - Golden Oak Research Group
- [2] Forsquare API