CAPSTONE PROJECT REPORT

Project : Mclaren's pub - IOWA

Author : LAMRAOUI Anass

Date : April 23, 2020

Table Of Content

I. Introduction

Many investors choose to invest into pubs, bars..., generally in liquor stores, mainly because the majority of the population of the United States spend their free time with their friends and loved ones consuming alcohol. But those investments cannot always be a success. It can involve different features such as the population, the rent, alcohol consumption. In fact, I found an investor who's ready to open a pub with the name '***Maclaren's Pub***' in the state of ***Iowa***.

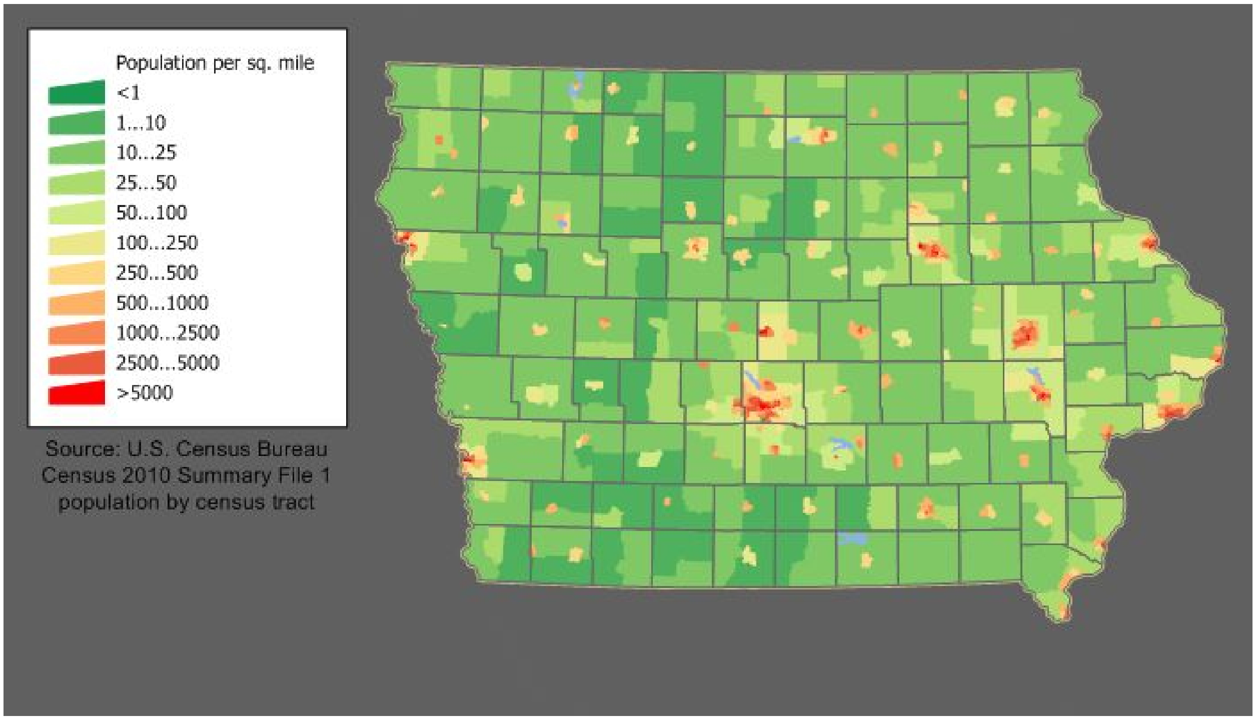
So how can we make sure that our investment doesn't turn into a big loss ? Where do we have to open our new pub ?

II. Project Description

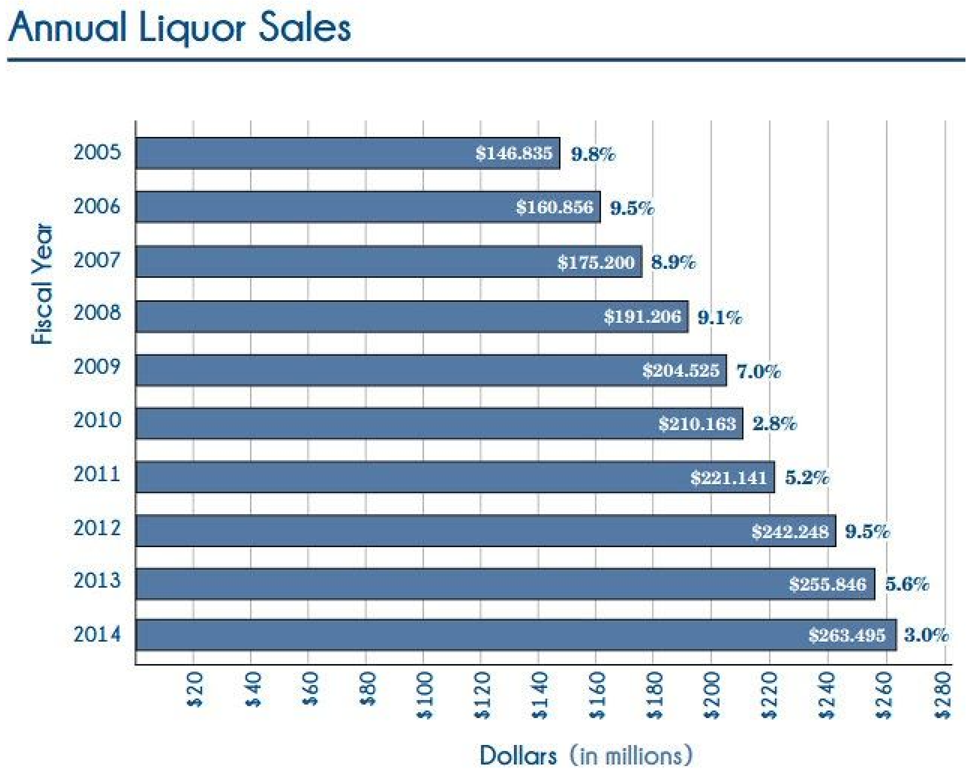
Our investor want to open '***Maclaren's Pub***' in one of the cities in the state of Iowa and make sure that this investment will make us a good profit.
Every town has a liquor store. However, the role as an important community center does not come cheap as the average startup costs for a pub can be quite high. The amount varies depending on many factors.
Finding the best location for a pub is a crucial part of planning and can help or hinder the business. The cost for the premises selected will depend on the area's popularity.

The state of Iowa is a state located in the Midwest region in the United States with 3 155 070 total population and ranks 26th in top states income in the U.S with an average of $58,570.

There are 99 counties in the U.S. state of Iowa and a population highly concentrated in Des Moines.



The Iowa Alcoholic Beverages Division 2014 Annual Report has shown that the drinking habits for Iowa Population increases every Year.



As we think about opening ***'Maclaren's Pub',*** what is the best place to choose to open our pub ? During this study we will explore liquor consumption over the year, population and also average rent value in the process of our evaluation to make the best decision.

Target Audience :

To solve this problem, as a data scientist my objective is to locate the best place with higher sales revenue, higher population and lowest rent value in order to get profitable business with lowest investment value.

III. Data Description

To solve this problem, as a data scientist my objective is to locate the best place with higher sales revenue, higher population and lowest rent value in order to get profitable business with lowest investment value.

1. Sales Data

The State of Iowa provides different datasets and visualizations about different categories such Communities, Commerce, Health… in the following website : https://data.iowa.gov/ .
In the Sales and Distribution category, specifically in the following website https://data.iowa.gov/Sales-Distribution/2019-Iowa-Liquor-Sales/38x4-vs5h , contains our first dataset. This filtered view contains the spirits purchase information of Iowa Class “E” liquor licensees by product and date of purchase for calendar year 2019. The dataset can be used to analyze total spirits sales in Iowa of individual products at the store level to get an idea on alcohol consumption and Sales. (.csv)

Column Name	Description
Invoice/Item Number	Concatenated invoice and line number associated with the ...
Date	Date of order
Store Number	Unique number assigned to the store who ordered the liqu...
Store Name	Name of store who ordered the liquor.
Address	Address of store who ordered the liquor.
City	City where the store who ordered the liquor is located
Zip Code	Zip code where the store who ordered the liquor is located
Store Location	Location of store who ordered the liquor. The Address, City...
County Number	Iowa county number for the county where store who order...
County	County where the store who ordered the liquor is located
Category	Category code associated with the liquor ordered
Category Name	Category of the liquor ordered.

Category Name	Category of the liquor ordered.
Vendor Number	The vendor number of the company for the brand of liquor...
Vendor Name	The vendor name of the company for the brand of liquor o...
Item Number	Item number for the individual liquor product ordered.
Item Description	Description of the individual liquor product ordered.
Pack	The number of bottles in a case for the liquor ordered
Bottle Volume (ml)	Volume of each liquor bottle ordered in milliliters.
State Bottle Cost	The amount that Alcoholic Beverages Division paid for eac...
State Bottle Retail	The amount the store paid for each bottle of liquor ordered
Bottles Sold	The number of bottles of liquor ordered by the store
Sale (Dollars)	Total cost of liquor order (number of bottles multiplied by t...
Volume Sold (Liters)	Total volume of liquor ordered in liters. (i.e. (Bottle Volume...
Volume Sold (Gallons)	Total volume of liquor ordered in gallons. (i.e. (Bottle Volu...

For more info, you can check the link to the datasets. This dataset contains a lot of information. I'll focus my study on Sales, Volume Sold for each store, city and county.

2. Population

The population dataset was taken also from data.iowa.gov. This dataset contains city population in Iowa from 2010 to 2018.

Column Name	Description	Type
FIPS	County FIP – A five-digit code for counties based on Federal...	Number #
County		Plain Text T
City		Plain Text T
Year	For population estimates the date of reference is always Jul...	Date & Time
Estimate		Number #
Primary Point	Primary latitude and longitude in decimal degrees for the c...	Point

FIPS	County	City	Year	Estimate	Primary Point
1966720	Keokuk	Richland	July 01, 2018	567	POINT (-91.9960251 41.1...
1924600	Scott	Eldridge	July 01, 2018	6,813	POINT (-90.5805427 41.6...
1906805	Benton	Blairstown	July 01, 2018	662	POINT (-92.0823291 41.9...
1912630	Dubuque	Centralia	July 01, 2018	136	POINT (-90.8365993 42.4...
1902350	Woodbury	Anthon	July 01, 2018	560	POINT (-95.8656328 42.3...
19031	Cedar	Balance of Cedar County	July 01, 2018	7,406	POINT (-91.1324125 41.7...
1965550	Fayette	Randalia	July 01, 2018	55	POINT (-91.8864465 42.8...
1980580	Cerro Gordo	Ventura	July 01, 2018	716	POINT (-93.4621042 43.1...
1962040	Mahaska	Pella (pt.)	July 01, 2018	2	POINT (-92.9177547 41.4...
1956055	Benton	Newhall	July 01, 2018	845	POINT (-91.9673411 41.9...
1944985	Jefferson	Libertyville	July 01, 2018	343	POINT (-92.0501957 40.9...
1963840	Pocahontas	Plover	July 01, 2018	70	POINT (-94.6222266 42.8...
1928020	Floyd	Floyd	July 01, 2018	318	POINT (-92.7402544 43.1...
1972345	Story	Sheldahl (pt.)	July 01, 2018	153	POINT (-93.6967164 41.8...

File source (.csv) is available in :

https://data.iowa.gov/Community-Demographics/City-Population-in-Iowa-by-County-and-Year/y8va-rhk9

3. Average Rent value

The average rent value is presented by the following Website :

https://www.rentdata.org/states/iowa/2019

This Website gives us the average rent and the population for every county in the state of Iowa in 2019. Using BeautifulSoup in python we will extract the data from the Website

County		0 BR	1 BR	2 BR	3 BR	4 BR	Est. Population
Adair County		$481	$502	$664	$877	$947	7,682
Adams County		$481	$517	$664	$960	$1,163	4,029
Allamakee County		$481	$506	$664	$865	$977	14,330
Appanoose County		$481	$502	$664	$840	$911	12,887
Audubon County		$481	$502	$664	$863	$898	6,119
Benton County	Metro	$522	$526	$683	$871	$966	26,076
Black Hawk County	Metro	$554	$662	$836	$1,087	$1,355	131,090
Boone County		$520	$603	$718	$980	$983	26,306
Bremer County	Metro	$496	$564	$740	$927	$1,198	24,276
Buchanan County		$501	$516	$682	$918	$922	20,958
Buena Vista County		$444	$509	$664	$914	$1,020	20,260
Butler County		$481	$502	$664	$848	$898	14,867

Using the three features we can cluster our data and conclude using Foursquare API the best places to fit for our new pub.

4. Data Preparation

After cleaning and assigning data types on each variable of the datasets, From the first dataset I started by adding a column that calculates the revenue of each operation made by every liquor store.
The revenue variable is equal to the Sales made minus the vendor price. The vendor price is the multiplication of the bottle sold and State Bottle Cost :

```
liq['Revenu']=liq['Sale (Dollars)']-liq['Bottles Sold']*liq['State Bottle Cost']
liq.head()
```

tore ame	Address	City	Zip Code	Store Location	County Number	County	...	Item Description	Pack	Bottle Volume (ml)	State Bottle Cost	State Bottle Retail	Bottles Sold	Sale (Dollars)	Volume Sold (Liters)	Volume Sold (Gallons)	Revenu
tfish rlie's	1630 East 16th St	Dubuque	52001.0	NaN	31.0	DUBUQUE	...	Crown Royal	12.0	1000.0	18.89	28.34	1.0	28.34	1.00	0.26	9.45
ntral City quor, Inc.	1460 2ND AVE	Des Moines	50314.0	POINT (-93.619787 41.60566)	77.0	POLK	...	Rumchata	12.0	375.0	7.00	10.50	2.0	21.00	0.75	0.19	7.00
im & #573 / SE DM	5830 SE 14th St	Des Moines	50315.0	NaN	77.0	POLK	...	Titos Handmade Vodka	12.0	750.0	9.64	14.46	6.0	86.76	4.50	1.18	28.92

The Revenue features represent our target variable that will help us take the decision to choose which city to open our Pub.

After having the revenu value on each operation. We should add the Population variable.
This dataset is csv file that contains each city population in the state of Iowa from 2010 to 2018. We cleaned our data to get the population on each city in the state of Iowa :

```
pop=pop[['City','Population']]
npop.head()
```

	City	Population
0	Orleans	589.0
1	Garwin	497.0
2	Mechanicsville	1133.0
3	Johnston	22040.0
4	Walnut	774.0

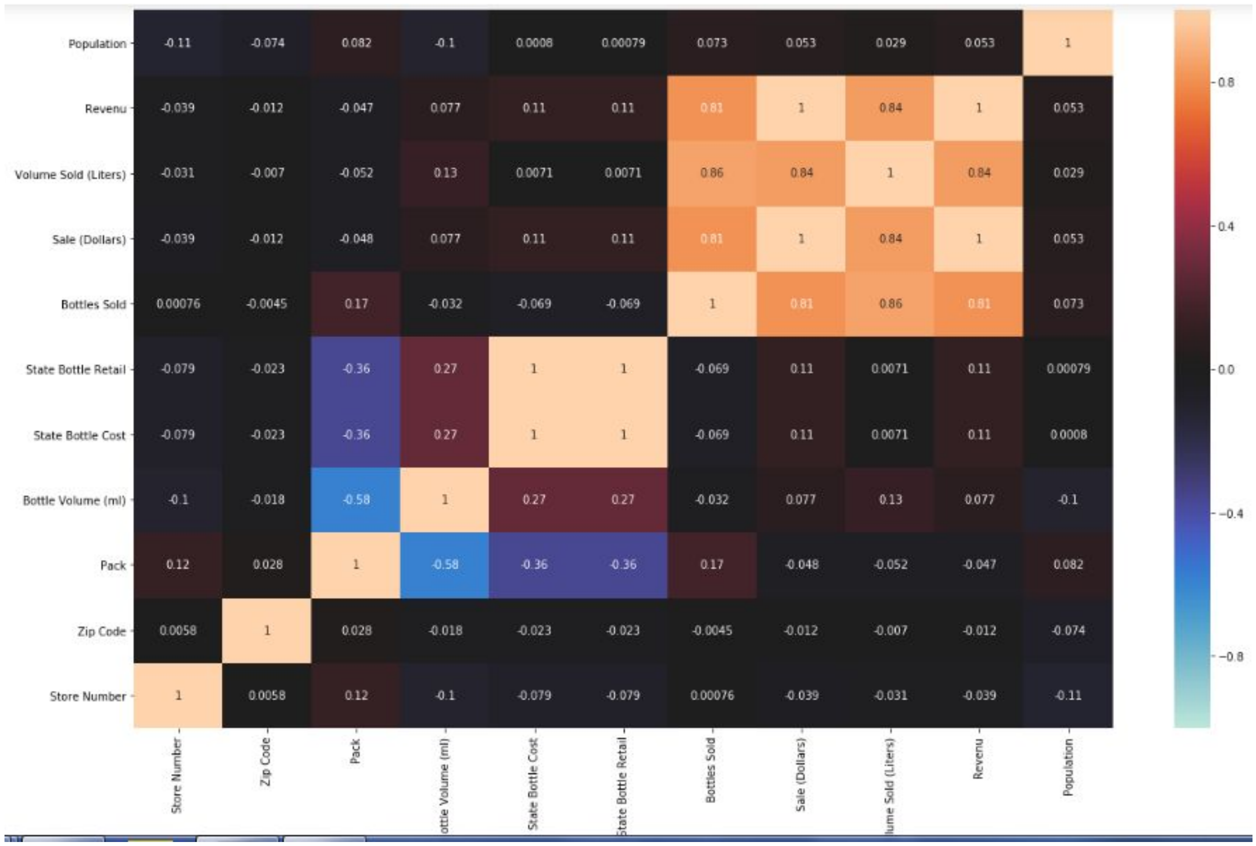
After filtering our data to get the average population in every city, let's add it to our first dataset. This column will be considered in the city analysis.

IV. EDA and Methodology

1. Exploratory Data Analysis I

Our dataset contains a lot of features, before eliminating different features. It is necessary to check the relationships between the variables to eliminate the loss of valuable data.

In our EDA, we started by using the function **describe()** on our dataset, and then study the correlation between the features. This EDA is resumed in our correlation Heat map :

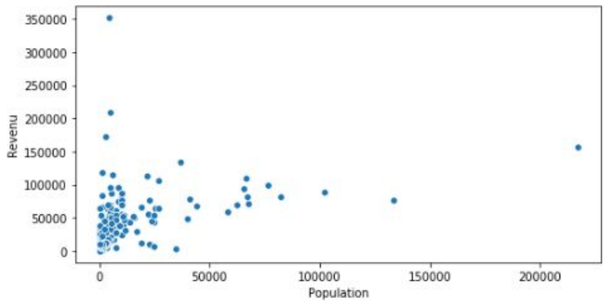


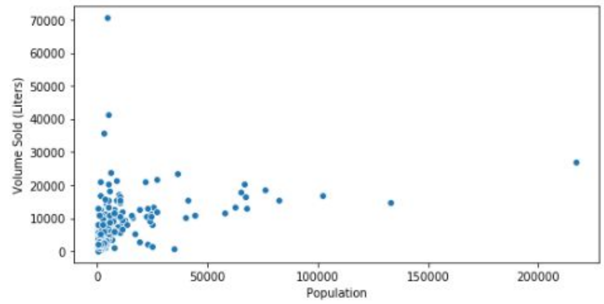
We see that there's a correlation between Revenue and Volume Sold (liters) and absolutely no correlation between the revenu and the population. Generally the correlation between other variables is poor.

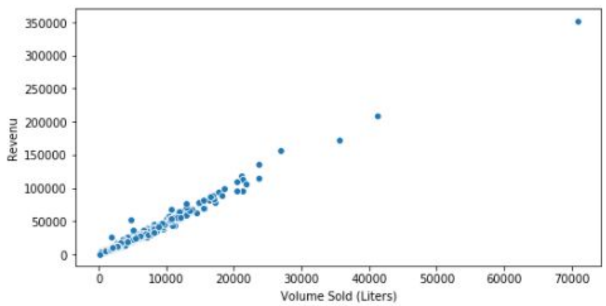
For this fact, we grouped our dataset on each store to have the total revenue made in 2019 for each store and then we grouped our data on each city, their county, population, average volume of liquor consumption and revenue made by the liquor stores present in each city.

	City	County	Total Population	Total Revenu	Volume Sold (Liters)
0	Ackley	HARDIN	1505.0	8360.060	1976.660
1	Adair	ADAIR	704.0	8055.770	1489.895
2	Adel	DALLAS	4954.0	39443.345	8110.635
3	Afton	UNION	821.0	9348.900	2323.350
4	Akron	PLYMOUTH	1473.0	9932.500	2302.590

After, focusing our analysis we plotted our dataset to see the data distribution and chose the modeling method.







First we see that there's a correlation between the Volume Sold and the revenue but we noticed that our data has a little bit of noise.

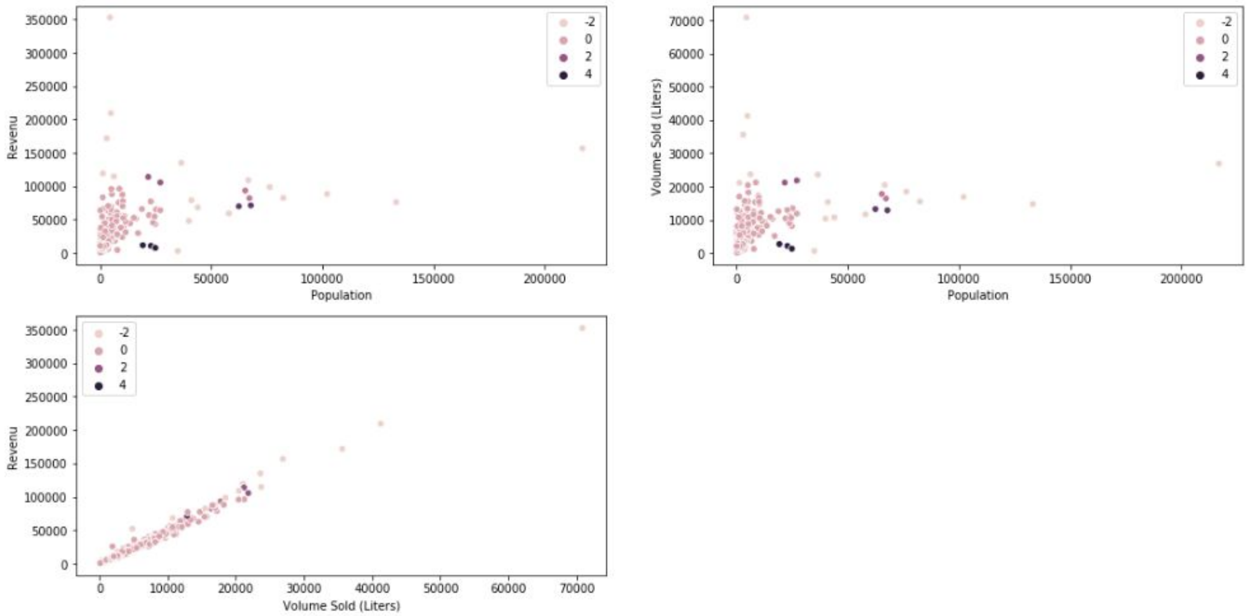
To eliminate the noise and get a bigger picture of our data we can use DBSCAN clustering method to eliminate the noise in our datasets. The noise in our dataset or outliers represent unlogical or useless data that could not help us to analyze our data.

First we start by scaling our data :

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(nobs[['Total Population','Volume Sold (Liters)','Total Revenu']])
X_scaled
```

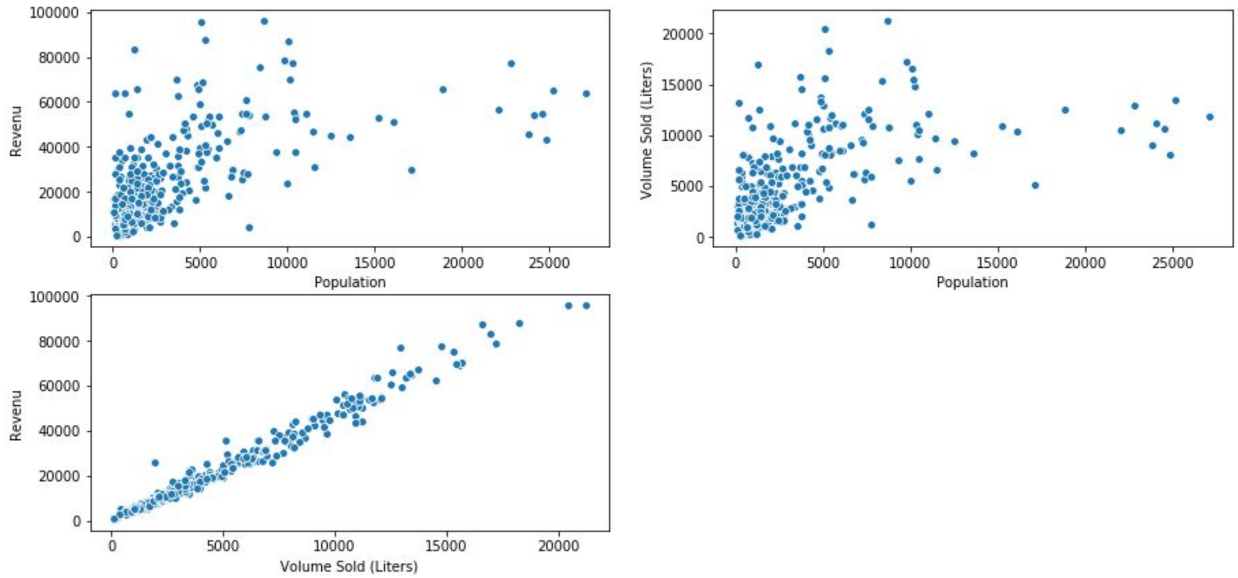
```
array([[-0.25571149, -0.56194031, -0.5697687 ],
       [-0.30366441, -0.64112912, -0.57955265],
       [-0.04923256,  0.43595847,  0.4296631 ],
       ...,
       [-0.26187772,  0.00265893, -0.08344475],
       [-0.26080013, -0.55044019, -0.5148511 ],
       [-0.34024274, -0.54341875, -0.49237623]])
```

Then we use the DBSCAN method to identify outliers as noises and plot the results to chose the cluster to study :



From our dataset we see that most of our data is focused in the cluster 0. The outliers are mostly errors and useless data such as points with higher population but with low liquor sale revenue.

Now we filter our data and plot it to choose the best machine learning algorithm for our study :



From the plots we see that the best machine learning algorithm that will help us predict the cities with highest revenue is K-means.

2. Modeling I

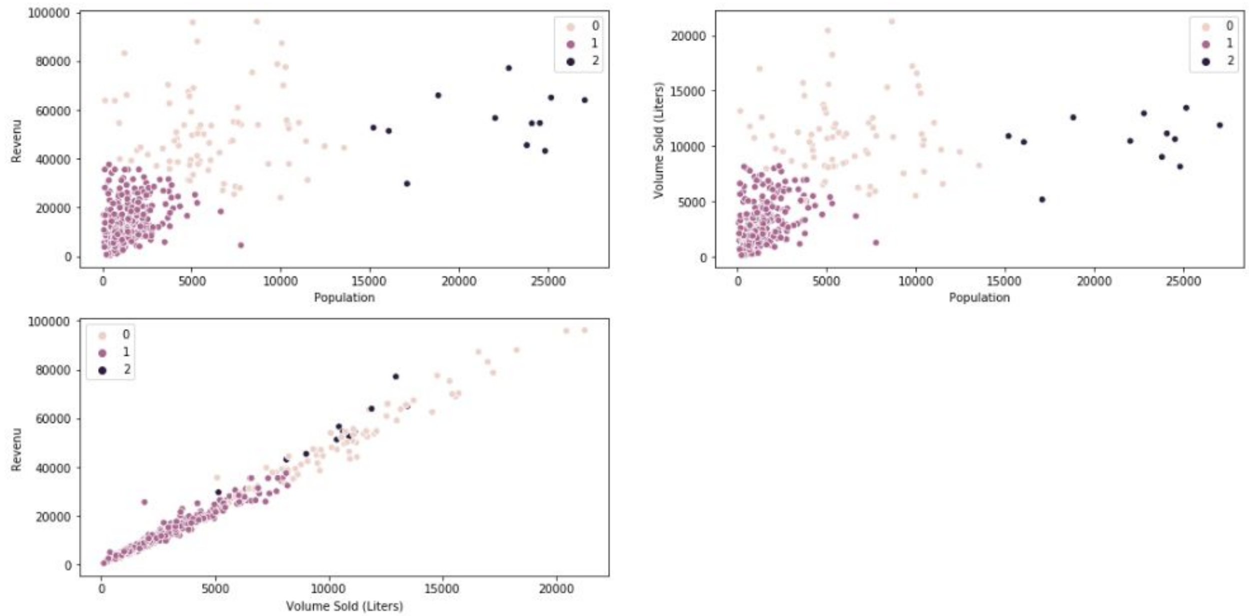
First we start by scaling our dataset for better result using the k-means algorithm :

```
scaler2 = StandardScaler()
X2_scaled = scaler2.fit_transform(nobs[['Total Population','Total Revenu']])
X2_scaled
```

```
: array([[-3.20509145e-01, -6.89867913e-01],
        [-5.11960765e-01, -7.05992674e-01],
        [ 5.03856192e-01,  9.57279714e-01],
        [-4.83995922e-01, -6.37467870e-01],
        [-3.28157650e-01, -6.06542073e-01],
        [-5.19848285e-01, -1.00097528e+00],
        [ 2.11539924e-01,  2.76294437e-01],
        [-5.62632105e-01, -8.10010076e-01],
        [-5.12438796e-01, -9.67930357e-01],
        [-5.02400135e-01, -5.80973975e-01],
        [ 6.17149660e-01,  1.65195051e+00],
        [-4.44080291e-01, -7.02504248e-01],
        [-2.20122528e-01, -9.01870450e-01],
        [-3.78111942e-01, -9.40158610e-01],
        [ 3.82378502e+00,  2.36039133e+00],
        [-5.74582893e-01, -2.31460805e-01],
        [ 6.36031905e-01,  1.71301170e+00],
```

Then we start by fitting our models with three clusters and then predict the cluster of each row and plot our data to see the best cluster for our study.

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=3,random_state=2).fit(X2_scaled)
c2 = kmeans.fit_predict(X2_scaled)
```



3. Results and interpretation I

From the plot we see three different clusters :

- **The first one k=0 interpretation :** Cities with low and medium population and high revenu on liquor consumption.
- **The second one k=1 interpretation :** Cities with low population and low revenu on liquor consumption.
- **The first one k=2 interpretation :** Cities with high population and medium revenu on liquor consumption.

The choice of the cluster will help us choose the best place to open our pub. From the third plot the best cluster to choose is k=0 Cities with low and medium population and the highest revenue on liquor consumption. Because most of the population of this cluster had the highest Revenu and Liquor Consumption.

But as we see our study intervals are pretty big, we should use another feature to judge the effectiveness of our model.

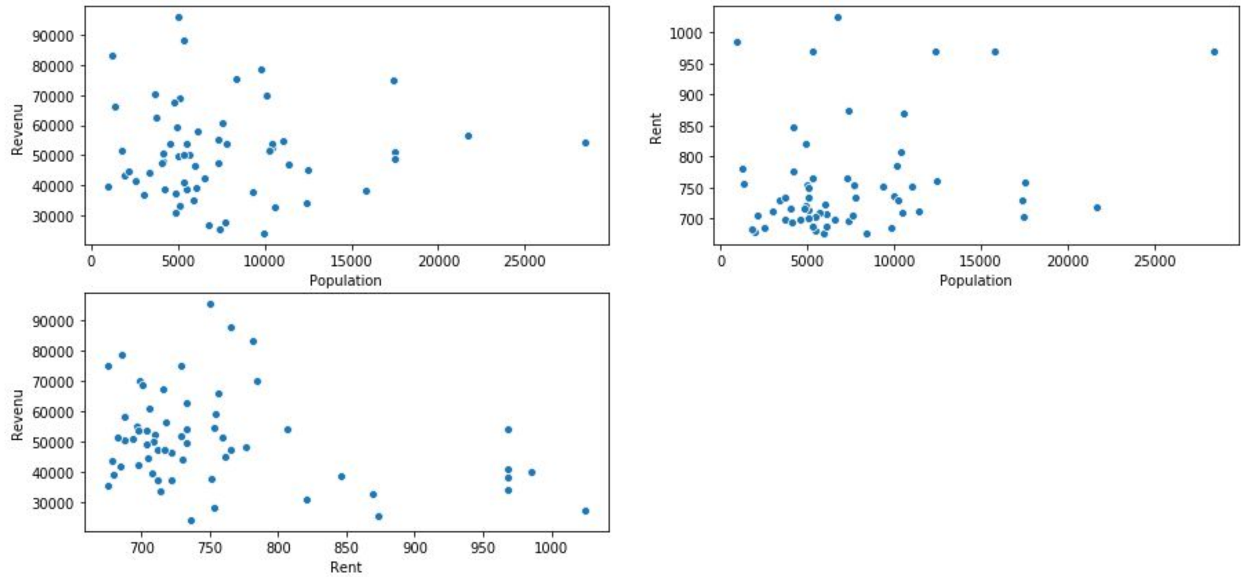
4. Exploratory Data Analysis II

Before thinking about choosing the city to invest in the pub we should think about the rent value. The cost for the premises selected depend generally on the area's popularity. With the average value on the rent we can focus our study into the cities with the lowest rent value.
The rent data : This dataset gives us an idea on the average rent in every county based on the population of the state of Iowa . With the dataset we could focus our area of study we started by cleaning and calculating the average rent for each city :

Now we group our first dataset in order to join it with the rent data :

	County	Total Population	Rent avg	Average Revenu
0	ALLAMAKEE	3683.0	698.6	70272.990000
1	APPANOOSE	5478.0	679.6	38918.207143
2	BENTON	5093.0	713.6	33403.435000
3	BOONE	12470.0	760.8	45074.831818
4	BREMER	10153.0	785.0	69963.692857

Since our data is ready to be treated. We plot the dataset to see the relationships and choose the algorithm to fit our data :



We can cluster our data to eliminate counties with the highest Rent value to focus our dataset. The best algorithm to filter our data is k-means.

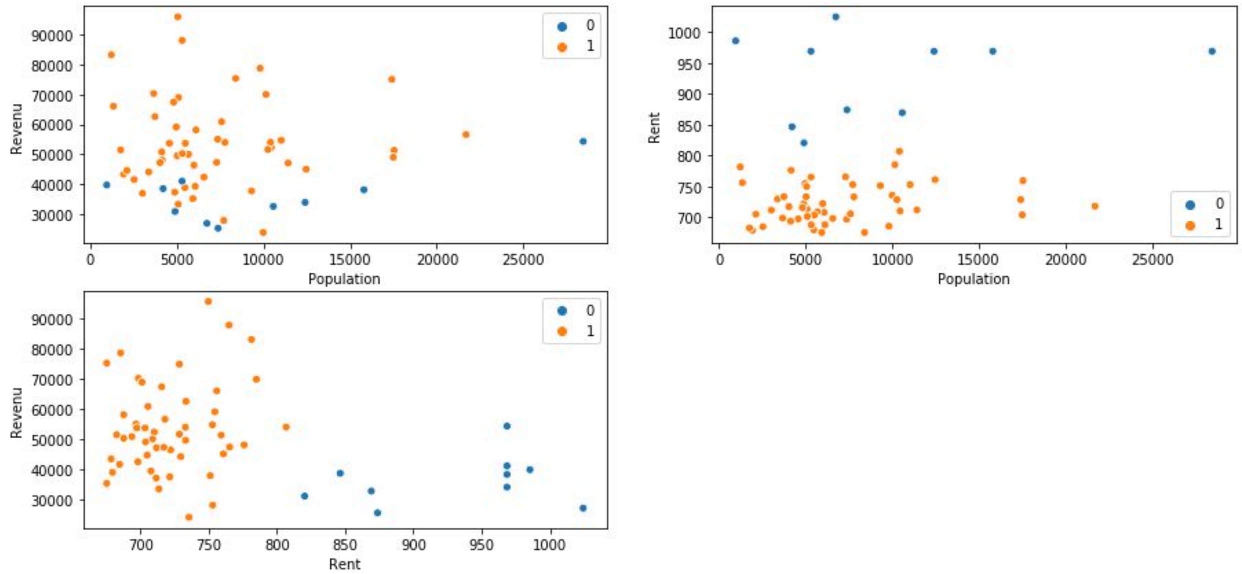
5. Modeling II

Before using the k-means cluster we should scale our dataset :

```
rscaler = StandardScaler()
Xr_scaled = rscaler.fit_transform(obsrent[['Total Population','Rent avg','Average Revenu']])
Xr_scaled
```

```
array([[-7.48997164e-01, -6.69120189e-01,  1.28636089e+00],
       [-3.96545635e-01, -8.89535804e-01, -7.67281902e-01],
       [-4.72141088e-01, -4.95107861e-01, -1.12848268e+00],
       [ 9.76346339e-01,  5.24509289e-02, -3.64041734e-01],
       [ 5.21399156e-01,  3.33190817e-01,  1.26610287e+00],
       [-2.79716298e-01, -5.62392628e-01, -7.37977254e-01],
       [ 5.81286463e-01, -5.34550656e-01,  1.10947696e-01],
       [-1.09064934e+00, -9.01136625e-01, -4.74080012e-01],
       [ 4.54443183e-01, -8.19930873e-01,  1.83932475e+00],
       [-1.80754977e-01, -6.73760518e-01, -5.36657358e-01],
       [-1.22857155e+00,  2.91427859e-01,  2.13196762e+00],
       [-5.10036991e-01, -4.02301287e-01, -8.66212638e-01],
       [-8.05742842e-01, -3.09494712e-01, -4.24673366e-01],
       [ 6.93796060e-01, -4.03556458e-02,  2.67735078e-01],
       [-4.27176520e-01,  1.01174381e-01,  2.44740467e+00],
       [ 1.78372514e-01, -9.38259255e-01,  1.61659654e+00],
       [ 9.64565229e-01,  2.46310170e+00, -1.08895281e+00],
       [-4.93936141e-01, -1.94741665e-02,  5.56351746e-01],
       [-4.76657180e-01, -7.51581113e-02,  2.96152869e+00],
```

Then we fit the model and predict the cluster for each County. To see the results of our machine learning algorithm we plot our data :



6. Results and interpretation II

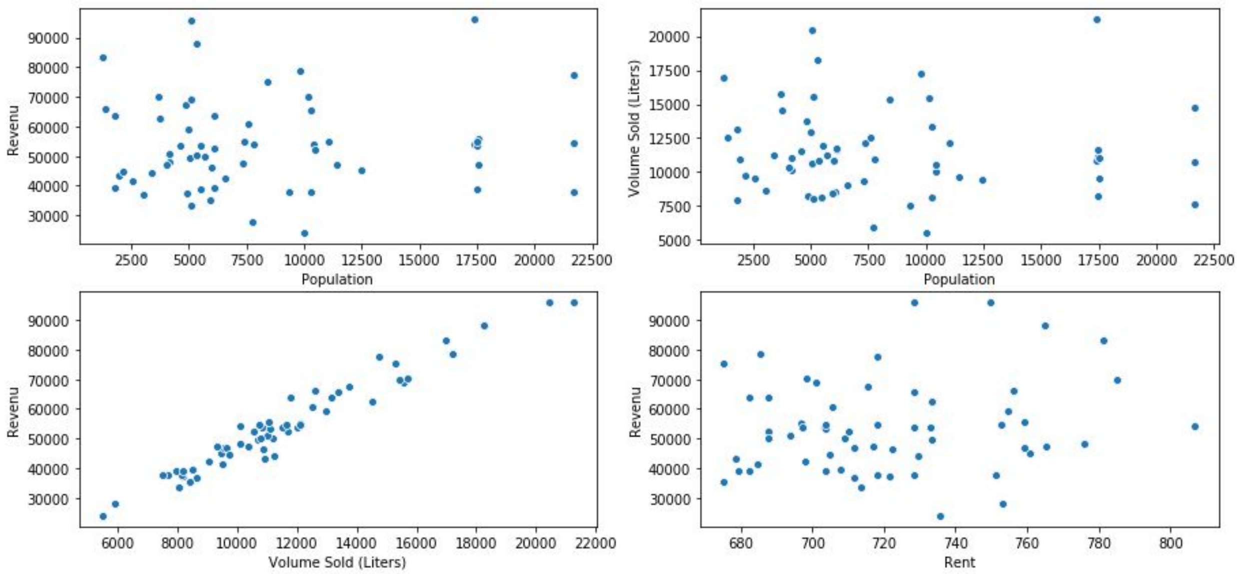
From the plot we see three different clusters :

- **The first one k=0 interpretation :** Counties with low rent value and high revenue on liquor consumption.
- **The second one k=1 interpretation :** Counties with high rent value and low revenu on liquor consumption.

The first cluster is generally the best one for our focus analysis. This choice will help us predict the city with lowest rent value, high revenue and diverse popularity in order to get the most profitable and optimal investment.

Now that we have chosen the cluster with the lowest rent values in each county, we should filter our city dataset on the counties with the cluster k=0.

We plot our result data :



With 60 City left in our dataset, we could use the Foursquare API to get the most common nightlife spots venues to predict the city with the highest revenu.

7. Exploratory Data Analysis III

Foursquare API is a database of more than 105 million places worldwide and an API that enables location data for Apple, Samsung, Microsoft, Tencent, Snapchat, Twitter, Uber, and others.
Using Foursquare data we will get the location and most common venues of each city. Using My developer Foursquare API credential we send a request to connect with the database and return the result in JSON file.
Foursquare API requests should have different input, such as the longitude and the latitude of the city. How can we get it ?
Using Geocoder package, we can get the location of each address and then we append it to our dataset :

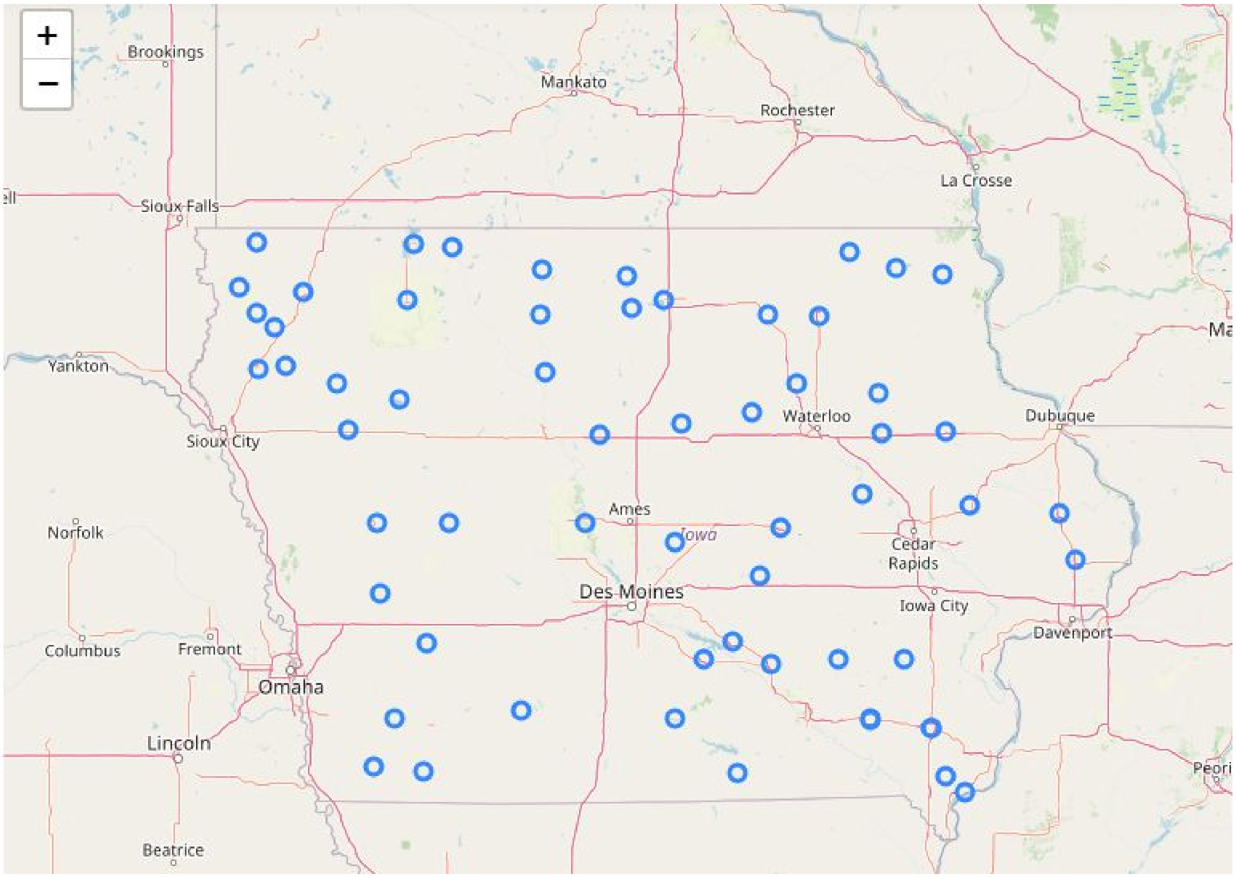
```python
from geopy.geocoders import Nominatim
import folium
```

```python
lat=[]
long=[]
i=0
y=0
for n in nobs['City']:
    g=Nominatim(user_agent="foursquire_agent",timeout=30)

    location=g.geocode('{}, Iowa'.format(n))
    if(location!=None):

            lat.append(location.latitude)
            long.append(location.longitude)
    else:
            lat.append(0)
            long.append(0)
nobs['Latitude']=lat
nobs['Longitude']=long
```

```python
nobs.head()
```

	City	County	Total Revenu	Volume Sold (Liters)	Total Population	Rent avg	Latitude	Longitude
0	Algona	KOSSUTH	52552.461667	11692.508333	6123.0	687.8	43.069966	-94.233019
1	Anamosa	JONES	53704.746000	11995.048000	5507.0	703.4	42.108337	-91.285159
2	Atlantic	CASS	42439.356250	9072.465000	6577.0	698.2	41.403601	-95.013878

Then we plot to see our data on the map and be sure to get the exact longitude and latitude of each city :



Now with the longitude and latitude of each city we send requests to the Foursquare API with the category Nightlife spots to collect data on bars, pubs, night clubs... Our Foursquare dataset is ready :

	City	City Latitude	City Longitude	id	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Algona	43.069966	-94.233019	4d66ed1f58155481542bde55	The Perky Parrot After Dark	43.069039	-94.236002	Cocktail Bar
1	Algona	43.069966	-94.233019	4bf5b5469abec9b69d8124e8	Billie Jo's Bar & Grill	43.068514	-94.237650	Bar
2	Algona	43.069966	-94.233019	4bac324df964a52082ea3ae3	Pep's	43.068985	-94.237120	Bar
3	Algona	43.069966	-94.233019	52685ff5498e9cba961df12f	Locker Room Bar & Grill	43.068753	-94.234676	Sports Bar
4	Anamosa	42.108337	-91.285159	4c018e80b58376b0145e443c	Tucker's Tavern	42.108258	-91.284374	Bar

Now we clean our data and get rid of errors on venues collection and focus our venues categories. Then we see how many venues for each city on a matrix.

	City	Bar	Beer Garden	Brewery	Cocktail Bar	Dive Bar	Lounge	Nightclub	Nightlife Spot	Pub	Sports Bar	Wine Bar
0	Algona	0	0	0	1	0	0	0	0	0	0	0
1	Algona	1	0	0	0	0	0	0	0	0	0	0
2	Algona	1	0	0	0	0	0	0	0	0	0	0
3	Algona	0	0	0	0	0	0	0	0	0	1	0
4	Anamosa	1	0	0	0	0	0	0	0	0	0	0

The best algorithm that could help us get the city with the top common venues we should k-means.

8. Modeling III

We should group rows by city and by taking the mean of the frequency of occurrence of each category :

	City	Bar	Beer Garden	Brewery	Cocktail Bar	Dive Bar	Lounge	Nightclub	Nightlife Spot	Pub	Sports Bar	Wine Bar
0	Algona	0.50	0.0	0.0	0.25	0.00	0.0	0.0	0.0	0.0	0.25	0.0
1	Anamosa	0.75	0.0	0.0	0.00	0.25	0.0	0.0	0.0	0.0	0.00	0.0
2	Atlantic	0.50	0.0	0.0	0.00	0.50	0.0	0.0	0.0	0.0	0.00	0.0
3	Centerville	0.50	0.0	0.0	0.00	0.00	0.0	0.0	0.5	0.0	0.00	0.0
4	Chariton	1.00	0.0	0.0	0.00	0.00	0.0	0.0	0.0	0.0	0.00	0.0

Let's create a new dataframe that displays the top 6 venues for each city.

	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
0	Algona	Bar	Sports Bar	Cocktail Bar	Wine Bar	Pub	Nightlife Spot
1	Anamosa	Bar	Dive Bar	Wine Bar	Sports Bar	Pub	Nightlife Spot
2	Atlantic	Dive Bar	Bar	Wine Bar	Sports Bar	Pub	Nightlife Spot
3	Centerville	Nightlife Spot	Bar	Wine Bar	Sports Bar	Pub	Nightclub
4	Chariton	Bar	Wine Bar	Sports Bar	Pub	Nightlife Spot	Nightclub
5	Clarinda	Bar	Wine Bar	Sports Bar	Pub	Nightlife Spot	Nightclub
6	Clear Lake	Bar	Brewery	Wine Bar	Sports Bar	Pub	Nightlife Spot
7	Creston	Bar	Nightlife Spot	Wine Bar	Sports Bar	Pub	Nightclub
8	Decorah	Bar	Dive Bar	Cocktail Bar	Beer Garden	Wine Bar	Sports Bar
9	Denison	Bar	Nightlife Spot	Wine Bar	Sports Bar	Pub	Nightclub

Now we use k-means algorithms with 4 categories and get our labels :

```python
kclusters = 4

# run k-means clustering
kmeansv = KMeans(n_clusters=kclusters, random_state=0).fit(iowa_grouped.drop('City', axis=1))

# check cluster labels generated for each row in the dataframe
kmeansv.labels_
```

```
array([0, 0, 0, 0, 3, 3, 0, 3, 0, 0, 3, 3, 3, 2, 3, 0, 0, 0, 0, 1, 2, 0,
       0, 2, 3, 3, 0, 3, 2, 0, 2, 2, 3, 0, 0, 0, 0, 1, 0, 0, 3, 2, 3, 3,
       0, 3, 3])
```

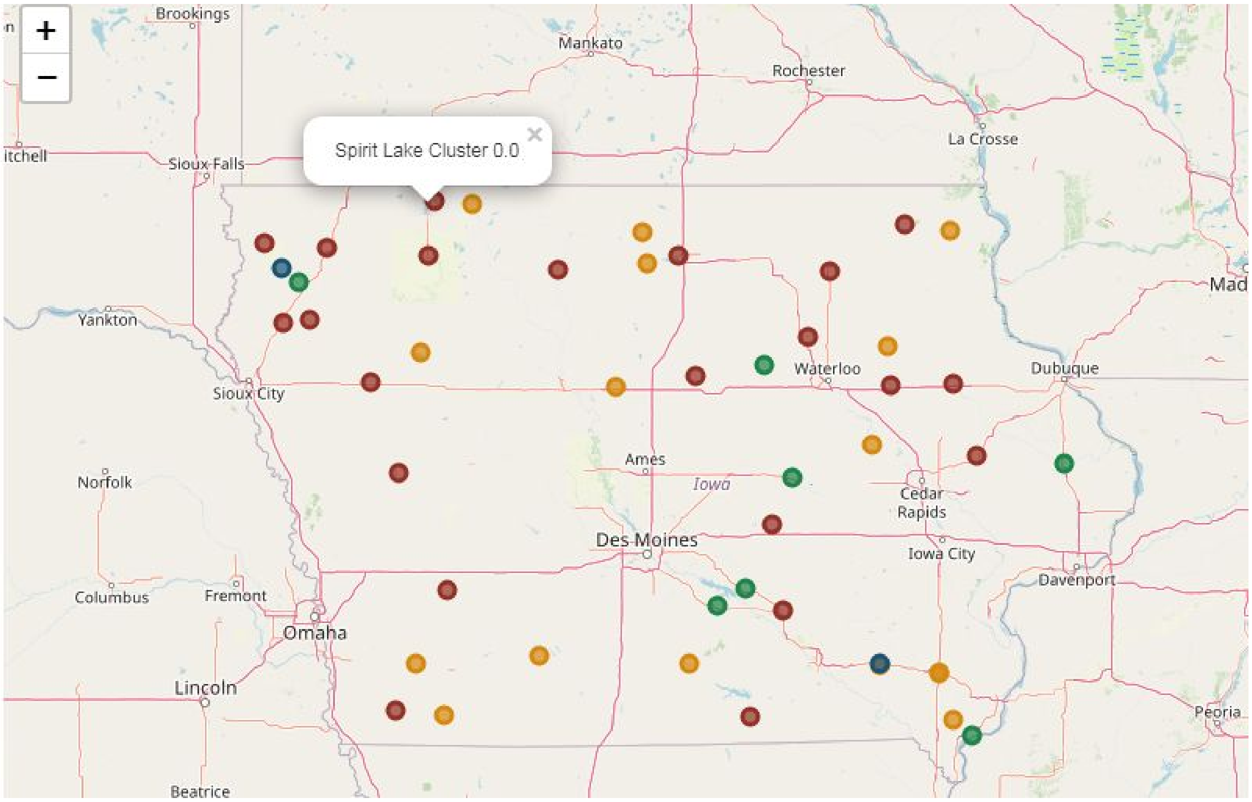
After getting the labels, let affect them to the data frame that displays the top 6 venues for each city , we get the following result and join it with our Foursquare dataset.

	City	County	Total Revenu	Volume Sold (Liters)	Total Population	Rent avg	Latitude	Longitude	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Con \
0	Algona	KOSSUTH	52552.461667	11692.508333	6123.0	687.8	43.069966	-94.233019	0.0	Bar	Sports Bar	Cocktail Bar	Wine Bar	Pub	Ni
1	Anamosa	JONES	53704.746000	11995.048000	5507.0	703.4	42.108337	-91.285159	0.0	Bar	Dive Bar	Wine Bar	Sports Bar	Pub	Ni
2	Atlantic	CASS	42439.356250	9072.465000	6577.0	698.2	41.403601	-95.013878	0.0	Dive Bar	Bar	Wine Bar	Sports Bar	Pub	Ni
3	Bancroft	KOSSUTH	63710.370000	11769.860000	6123.0	687.8	43.292739	-94.218019	NaN	NaN	NaN	NaN	NaN	NaN	
4	Boone	BOONE	45074.831818	9461.893636	12470.0	760.8	42.017180	-93.925411	NaN	NaN	NaN	NaN	NaN	NaN	

We clean the data with NaN values. These NaN are due to False values that we did get on venues categories after the collection of the Foursquare API.

9. Results and interpretation III

Finally we get the results by plotting our clusters on the map :

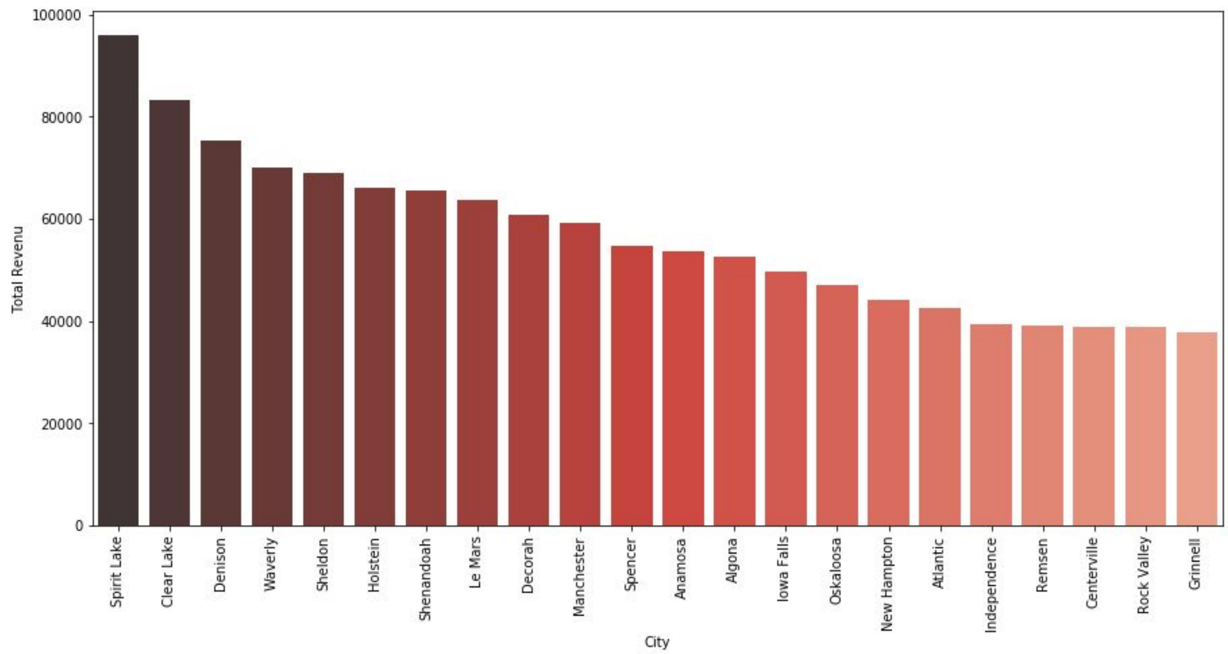


- Red spots : cluster k=0
- Blue spots : cluster k=1
- Green spots : cluster k=2
- Orange spots : cluster k=3

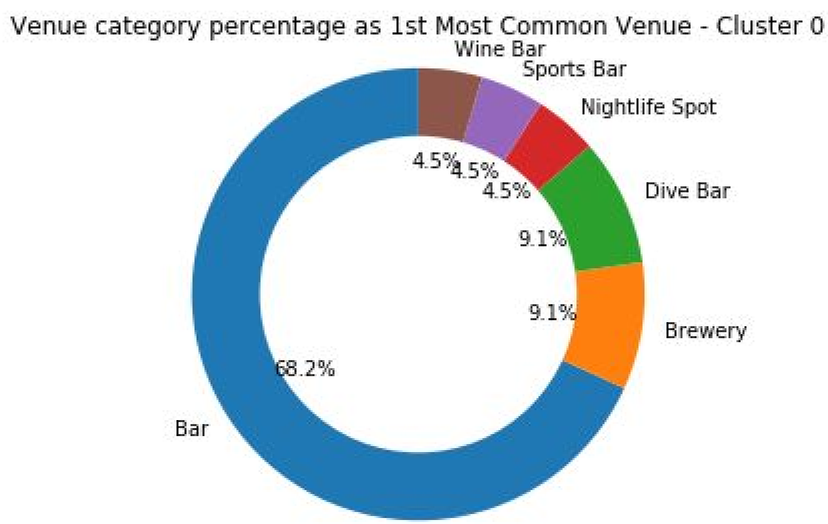
Let’s explore every cluster and see the difference between them by plotting the top cities with highest revenue and the most common venues for every cluster.

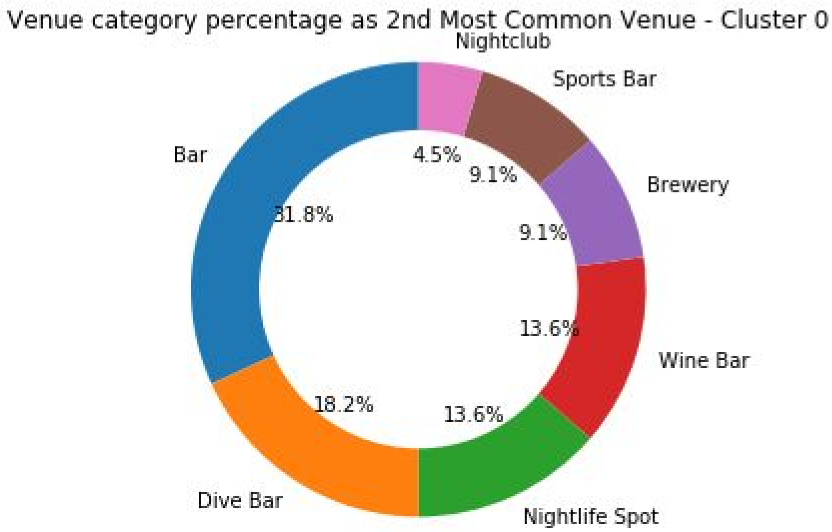
a. Cluster k=0

Let’s plot bar plot to show every city revenue in the cluster.



We notice that cities of the cluster k=0 have an average annual revenue between 90000 and 40000. Let's check the first and second common venues in the cluster :

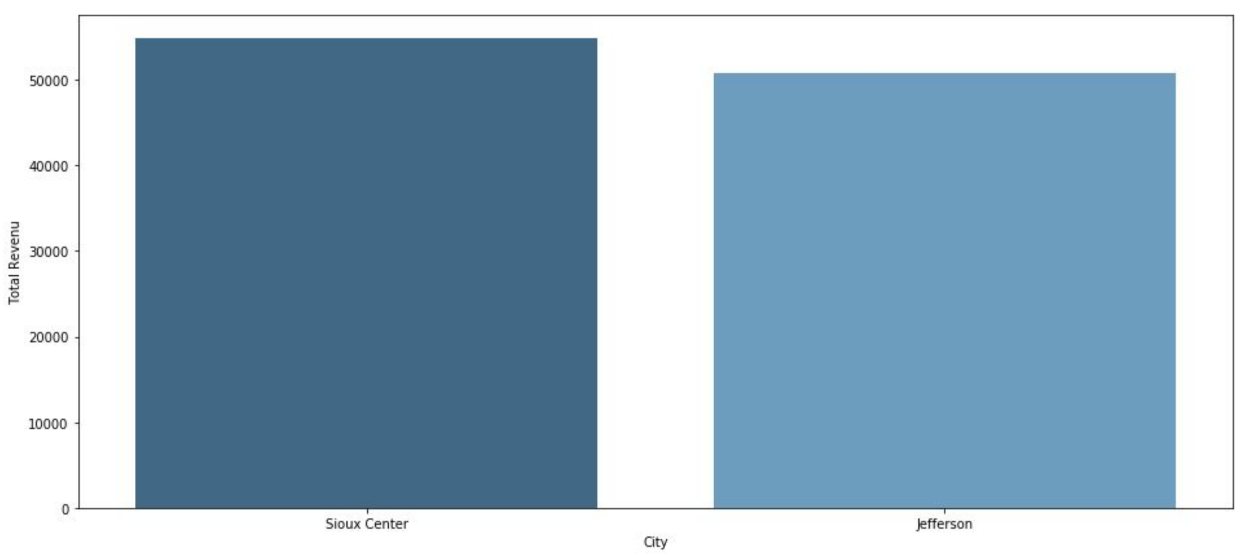




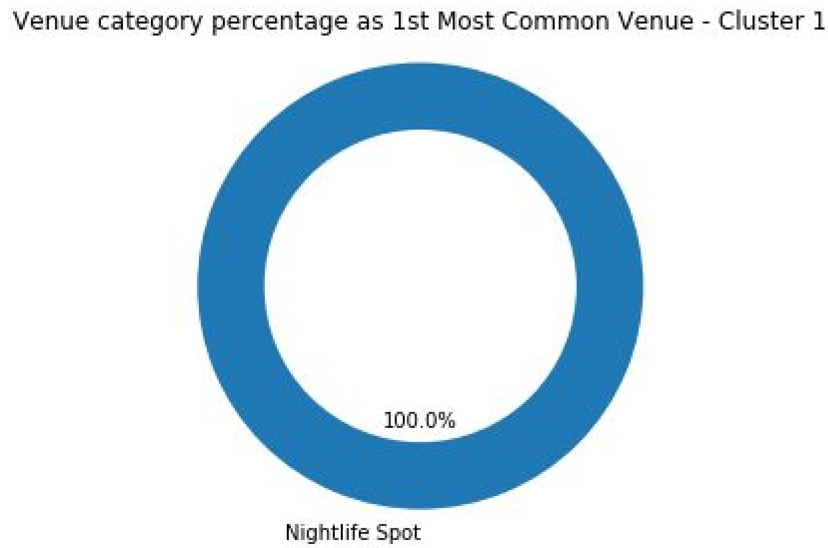
We see this cluster know a diversity in the most common venues categories controlled mostly with regular bars. This cluster is characterized with popularity for the nightlife spots, average rent value of 700$. This could be the bes cluster for our study

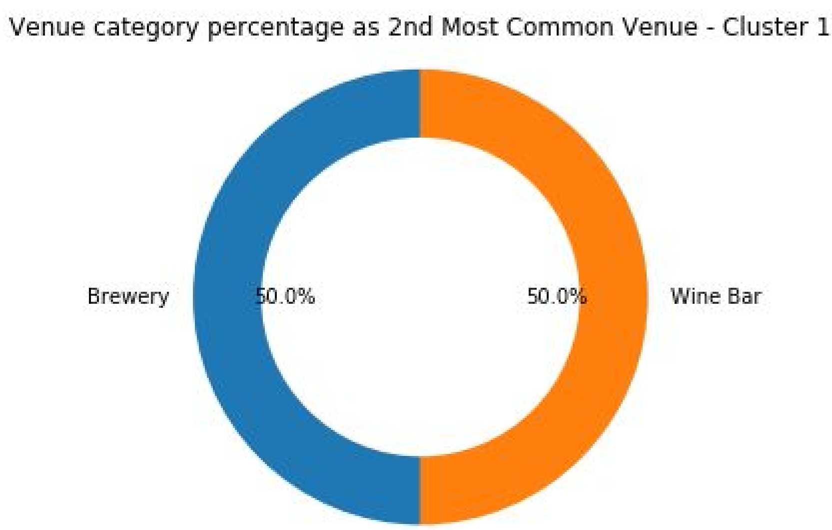
b. Cluster k=1

Let's start with the bar plot :



This cluster contains cities with annual revenue of 50000 and with less popularity. Let's check the most common venues :

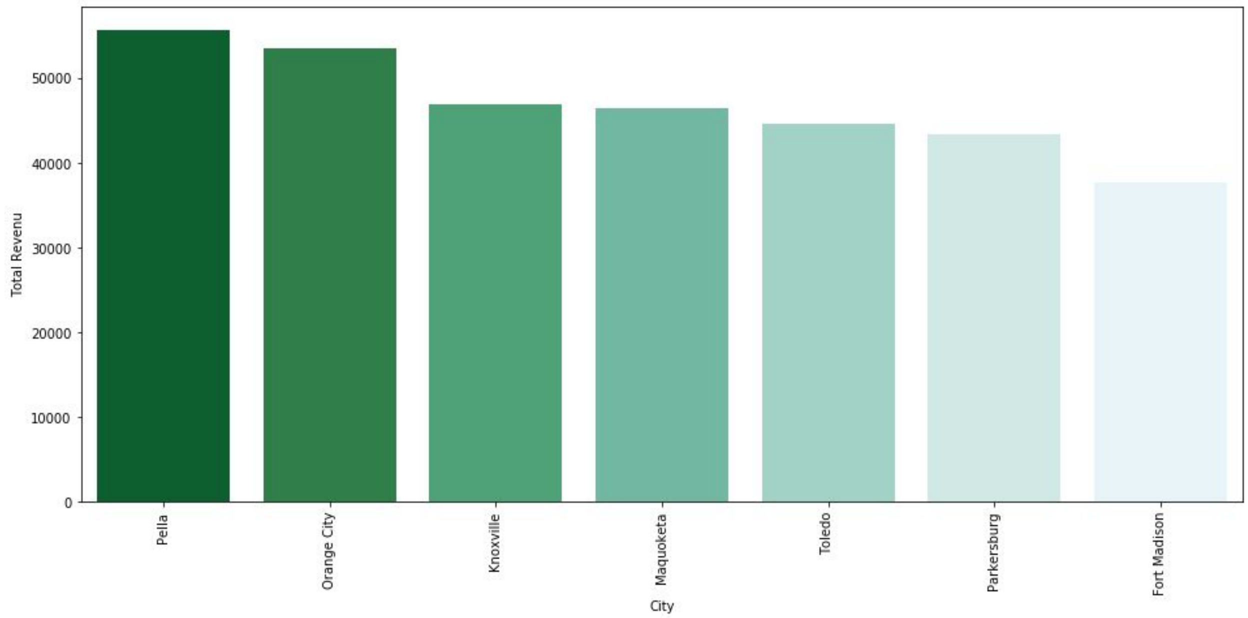




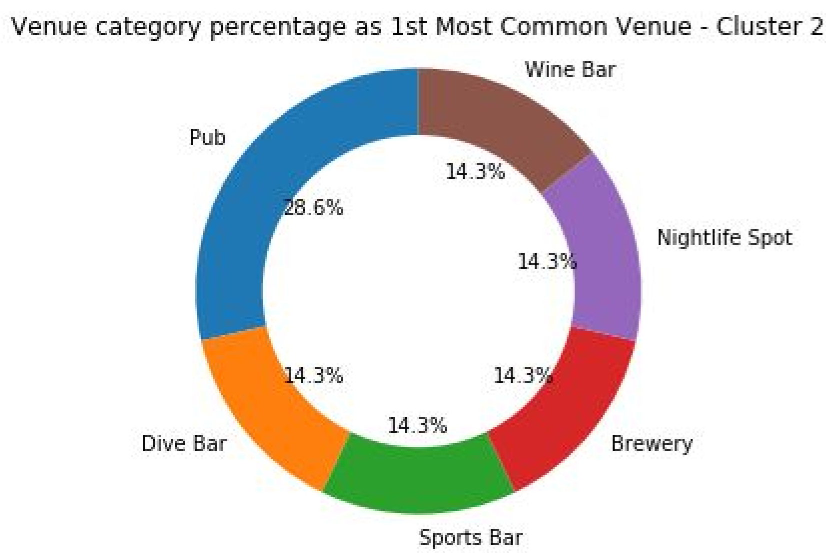
This cluster isn't good for our investment, it contains cities with less popularity and less attendances to nightlife spots.

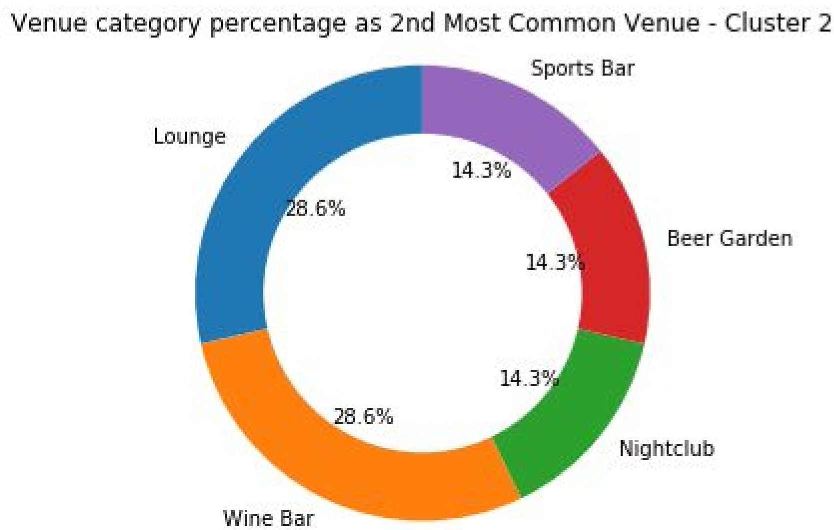
c. Cluster k=2

We start with the top cities with revenue in the cluster



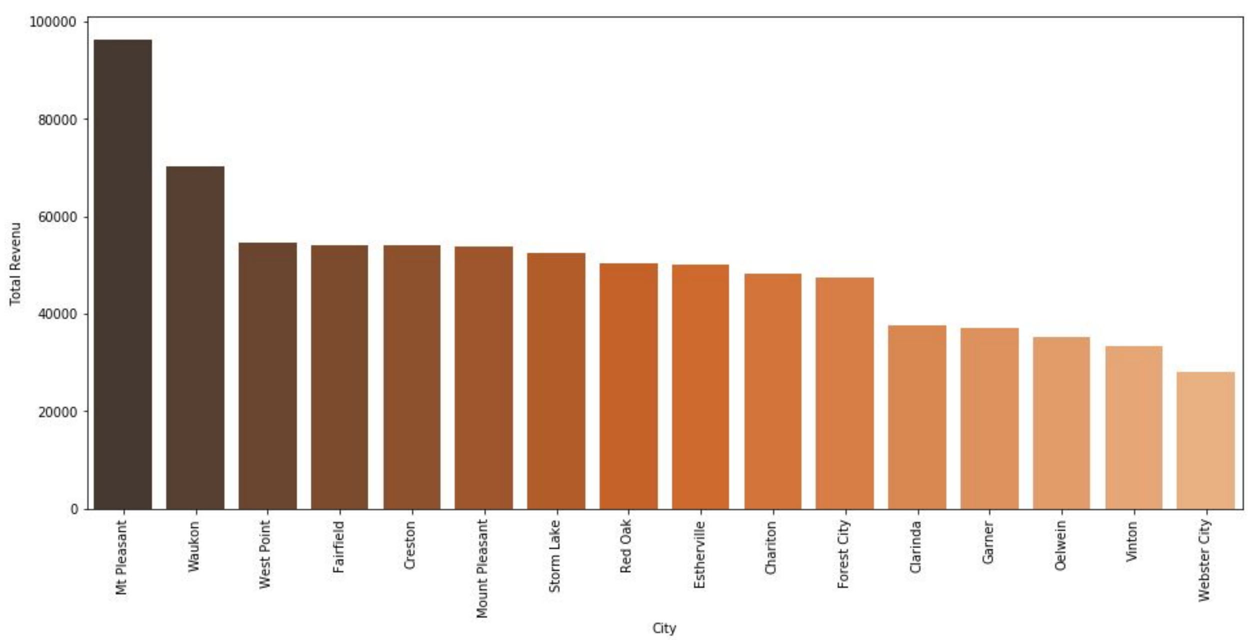
This cluster contains cities with an average annual revenue between 40000 and 50000. Let's see the popularity of the cities :



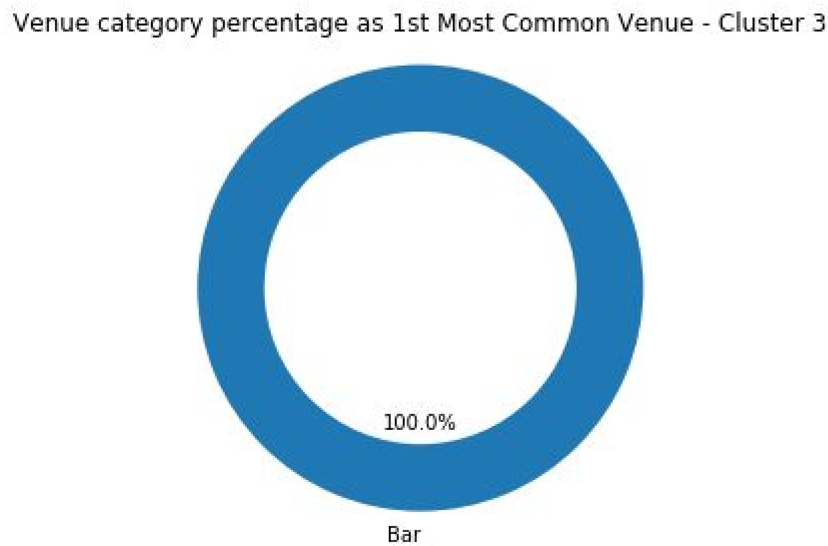


We notice that there's cities with higher attendance for pubs and a diversity of common venues categories. But the problem is that with these cities it can't be a good revenue with an average rent value of 710$ which impacts our revenue. This cluster isn't good for business but would have high attendance.

d. Cluster k=3



We notice that in this cluster there's higher average annual revenue on liquor consumption than the cluster 1 and 2. We should also see their popularity.



We see that these cities don't have a good popularity, showing one and category as a common venue category. Cities with lower popularity can't help us make our business profitable. The cluster 3 isn't good for us.

IV. Conclusion

From the plots we see that the cities in the first cluster k=0 has a higher profitability in terms of liquor consumption and different common venues on top of it, bars.

According to the 2016 American Community Survey, 5.6% of Iowa's population were of Hispanic or Latino origin (of any race): Mexican (4.3%), Puerto Rican (0.2%), Cuban (0.1%), and other Hispanic or Latino origin (1.0%). The five largest ancestry groups were: German (35.1%), Irish (13.5%), English (8.2%), American (5.8%), and Norwegian (5.0%).

From the cluster 0 the TOP 3 cities with high revenu on the liquor consumption are :

	City	County	Total Revenu	Volume Sold (Liters)	Total Population	Rent avg	Latitude	Longitude	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
51	Spirit Lake	DICKINSON	95849.260000	20437.322857	5070.000000	749.8	43.422184	-95.102217	0.0	Bar	Wine Bar	Pub	Sports Bar	Nightlife Spot
11	Clear Lake	CERRO GORD	83183.616667	16982.636667	1240.576139	781.4	43.138092	-93.379200	0.0	Bar	Brewery	Wine Bar	Sports Bar	Pub
16	Denison	CRAWFORD	75314.990000	15306.366667	8406.000000	675.4	42.017766	-95.355276	0.0	Bar	Nightlife Spot	Wine Bar	Sports Bar	Pub

Spirit Lake is the top city with the highest revenu with 95849$ estimated Population and a rent average of 750$ that could affect approximately 9% of average annual revenue.

Pubs are known as the 3rd Most Common Venue in the city. This city knows big popularity with the diversity of venues categories

Spirit Lake is the best place to open ***'Maclaren's Pub'***.