Capstone Project - The Battle of Neighborhoods (Week 2):

Setting up a new Restaurant in California

By Arunkumar Ramachandran

2/2/2021

Table of Contents

Table of Figure:

1. Introduction:

This project's main aim was to determine a suitable area in California to set up a restaurant business. The project starts off with reading the dataset, preprocessing/ cleaning the dataset, using foursquare credentials and finally using geo encoders to plot the map.

2. Dataset:

The dataset was taken from https://share.cocalc.com/share/e9d2f604-5c15-48c1-8c69-4d560cf9a933/PythonDataScienceHandbook/notebooks/data/california_cities.csv?viewer=share

The dataset can be shown as follows:

Load Dataset:

```
# Import zip codes, lat/longs and cities/neighborhoods of california
import types
import pandas as pd
from botocore.client import Config
import ibm_boto3

def __iter__(self): return 0

# @hidden_cell
# The following code accesses a file in your IBM Cloud Object Storage. It includes your credentials.
# You might want to remove those credentials before you share the notebook.
client_9a44eb9d5a79464eba6780c0d5d7902a = ibm_boto3.client(service_name='s3',
    ibm_api_key_id='HPilVCAY9DvqjSrgVeZQzkx3eMG7GZ5JKQTx-SaBsxKE',
    ibm_auth_endpoint="https://iam.cloud.ibm.com/oidc/token",
    config=Config(signature_version='oauth'),
    endpoint_url='https://s3-api.us-geo.objectstorage.service.networklayer.com')

body = client_9a44eb9d5a79464eba6780c0d5d7902a.get_object(Bucket='capstoneprojectthebattleofneighbo-donotdelete-pr-ummtwu3m8ncwlv',Key='california_cities.csv')['Body']
# add missing __iter__ method, so pandas accepts body as file-like object
if not hasattr(body, "__iter__"): body.__iter__ = types.MethodType( __iter__, body )

df_data_1 = pd.read_csv(body)
df_data_1.head()
```

12]:

	Unnamed: 0	city	latd	longd	elevation_m	elevation_ft	population_total	area_total_sq_mi	area_land_sq_mi	area_water_sq_mi	area_total_km2	area_land_km2	area_water_km2	area_water_percent
0	0	Adelanto	34.576111	-117.432778	875.0	2871.0	31765	56.027	56.009	0.018	145.107	145.062	0.046	0.03
1	1	AgouraHills	34.153333	-118.761667	281.0	922.0	20330	7.822	7.793	0.029	20.260	20.184	0.076	0.37
2	2	Alameda	37.756111	-122.274444	NaN	33.0	75467	22.960	10.611	12.349	59.465	27.482	31.983	53.79
3	3	Albany	37.886944	-122.297778	NaN	43.0	18969	5.465	1.788	3.677	14.155	4.632	9.524	67.28
4	4	Alhambra	34.081944	-118.135000	150.0	492.0	83089	7.632	7.631	0.001	19.766	19.763	0.003	0.01

Figure 1 Load Dataset

3. Project Description:

The primary focus of this project was to ensure that the area selected in California was suitable for business or not. Hence some key factors were taken into consideration:

- Area Population
- Longitude
- Latitude
- The area in sq.m

The project focuses on the struggles dealt with setting up a restaurant. Many factors come into picture when it comes to starting a restaurant and a lot of time goes with investment.

4. Data Wrangling:

The dataset was loaded from the URL mentioned above. Some of the critical aspects done were:

4.1. Locate all cities in California

Initially, all the cities are loaded from the dataset and saved in pandas dataframe.

4.2. Foursquare credentials

The foursquare credentials are used and the id details can be found in the foursquare developer app that we create.

4.3. Locate all venues in California

We locate all the venues in California and list down all the restaurants in California The primary forcus is on the cuisine of the restaurants.

```
In [85]: venues_cali.groupby('city').count()
```

Out[85]:

city	city Latitude	city Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
AgouraHills	19	19	19	19	19	19
Albany	59	59	59	59	59	59
Alhambra	14	14	14	14	14	14
AlisoViejo	66	66	66	66	66	66
Alturas	10	10	10	10	10	10
AmadorCity	7	7	7	7	7	7
AmericanCanyon	32	32	32	32	32	32
Anaheim	40	40	40	40	40	40
Anderson	11	11	11	11	11	11
AngelsCamp	10	10	10	10	10	10
Antioch	15	15	15	15	15	15
AppleValley	1	1	1	1	1	1
Arcadia	5	5	5	5	5	5
Arcata	39	39	39	39	39	39
ArroyoGrande	12	12	12	12	12	12
Artesia	33	33	33	33	33	33
Arvin	13	13	13	13	13	13

Figure 2 Venues in California

4.4. Selected all the restaurant's cuisine and located them in California

Based on the cuisine, we see the suitable places in california

4.5. One hot coding

For categorical variables where no such ordinal relationship exists, the integer encoding is not enough. In fact, using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories). In this case, a one-hot encoding can be applied to the integer representation. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value.

One Hot Coding:

```
In [89]: # one hot encoding
TO_new_onehot = pd.get_dummies(TO_new[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
TO_new_onehot['city'] = TO_new['city']


# move neighborhood column to the first column
fixed_columns = [TO_new_onehot.columns[-1]] + list(TO_new_onehot.columns[:-1])
TO_new_onehot = TO_new_onehot[fixed_columns]

TO_new_onehot.head()
```

Out[89]:

	city	Afghan Restaurant	Airport Food Court	American Restaurant	Arepa Restaurant	Asian Restaurant	BBQ Joint	Bar	Belgian Restaurant	Bistro	Brazilian Restaurant	Breakfast Spot	Burger Joint	Burrito Place	Café	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Coffee Shop	Comfort Food Restaurant	Creperie	Cuban Restaurant	Deli / Bodega
0	AgouraHills	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	Avalon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2	BaldwinPark	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
3	Belvedere	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	Burbank	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 3 One hot Coding

5. Methodology:

The algorithm selected was K means. K means can be defined as follows:
"K-means clustering is an iterative clustering algorithm where the number of clusters K is predetermined, and the algorithm iteratively assigns each data point to one of the K clusters based on the feature similarity." k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.

The problem is computationally difficult (NP-hard); however, efficient heuristic algorithms converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both k-means and Gaussian mixture modeling. They both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

The unsupervised k-means algorithm has a loose relationship to the k-nearest neighbor classifier, a popular supervised machine learning technique for classification that is often confused with k-means due to the name. Applying the 1-nearest neighbor classifier to the cluster centers obtained by k-means classifies new data into the existing clusters.

The k value selected was 5. The clusters were then run, and the data was segmented into clusters, and the labels were generated. The geographic center was also determined.

6. Results:

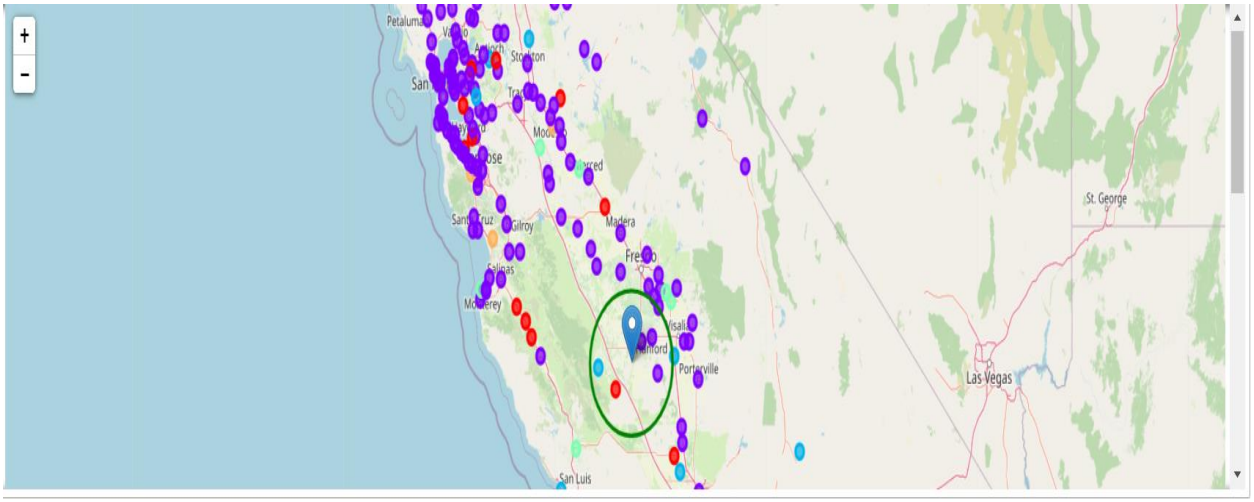

Figure 4 Area Suitable for restaurant setup

The green circle on the map shows the area where a restaurant can be set up.

6.1. Examine Clusters:

Five clusters were examined and depicted in the code.

7. Discussion:

A lot of critical factors need to be considered while setting up a restaurant. The area and the population in that area is one of them. The preference of people living in the area is also another critical factor to be considered.

A k- means were also used to locate all the nearby points into a cluster. This is the reason why using K means felt like an ideal choice to me.

Locating the map area to set up a restaurant was one of the primary aims, and I was successful.

8. Conclusion:

The area marked on the map is by far the best suitable place to start a restaurant. An improvement I can make to this project is checking the crime rate and other restaurant ratings to check for any competition in the business.