Capstone Project – Expanding to New Areas

Jordan Arthur

December 29, 2019

Introduction/Business Problem

A marketing company specializing in producing and mailing direct mail marketing pieces for their clients, independent small businesses like hair salons and restaurants in Los Angeles county California, is looking to expand into Ventura and Santa Barbara counties. With years of experience and by tracking the redemption of coupon codes, the leadership at the marketing company has a good understanding of what neighborhoods are best suited to receive the direct mail marketing pieces.

The problem is a lack of similar knowledge of neighborhoods in Ventura and Santa Barbara counties. Targeting the correct neighborhoods is important since unlike email marketing that has a very low cost, direct physical mail pieces are relatively expensive to produce and deliver.

The deliverables for this project are:

1. Find census block level areas in Santa Barbara and Ventura that are most like the know target areas in Los Angeles. This is who will be mailed the marketing material by the marketing company on behalf of their small business clients.

2. Find businesses by category that are within 5 kilometers of the target areas. These are the marketing company's potential clients.

Data

This project required two different data sets. The first is demographic data for neighborhoods that will be the target of the direct mailing pieces. The second is location data for businesses that will be offered the services of the marketing company.

A data file was provided by the marketing company that contains Los Angeles Census Block level data for the following datapoints. All values are show as a percent to total of the available data for each census block. In addition, each census block is marked as Target = True or False.

- Males 50 years and older
- Females 50 years and older
- Education level of a bachelor's degree or higher
- Household income of $100 or more
- No children under 18 in the household

Data Sources

The demographic data for the neighborhoods was retrieved using the US Census data API (https://www.census.gov/data/developers/data-sets.html)

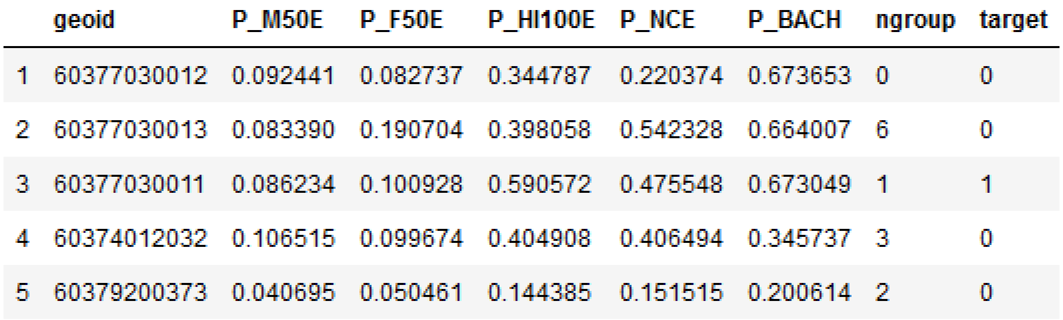
Geolocation data came from TIGER/Line Shape files (https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html)

The foursquare.com API was used to gather business location data. (https://developer.foursquare.com/docs/api)

Methodology

Matching the provided data file

After viewing the provided Los Angeles data file and the data available from the US Census API it became clear that some data preparation and aggregation would be necessary in order to create similar datasets for Ventura and Santa Barbara.



	geoid	P_M50E	P_F50E	P_HI100E	P_NCE	P_BACH	ngroup	target
1	60377030012	0.092441	0.082737	0.344787	0.220374	0.673653	0	0
2	60377030013	0.083390	0.190704	0.398058	0.542328	0.664007	6	0
3	60377030011	0.086234	0.100928	0.590572	0.475548	0.673049	1	1
4	60374012032	0.106515	0.099674	0.404908	0.406494	0.345737	3	0
5	60379200373	0.040695	0.050461	0.144385	0.151515	0.200614	2	0

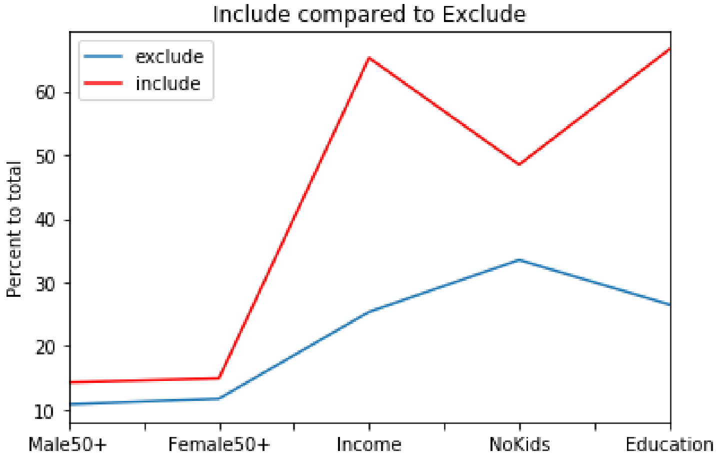
The following definitions were provided with the Los Angeles dataset.

geoid = The US Census Block FIPS number
P_M50E = Males 50 years and older
P_F50E = Females 50 years and older
P_HI100E = Household income of $100 or more
P_NCE = No children under 18 in the household
P_BACH = Education level of a bachelor’s degree or higher
target = Target area 1 for true, 0 for false

To recreate these columns the following mapping was performed. The census values are integers representing the count of the households or individuals in the census block. For our purposes we need the percent to total for that datapoint so we will also need to retrieve the totals.

Description	Column Name	Supporting Data from US Census
SEX / AGE		
Male 50 and older	P_M50E	B01001_016E + B01001_017E + B01001_018E + B01001_019E + B01001_020E + B01001_021E +B01001_022E + B01001_023E + B01001_024E + B01001_025E
Female 50 and older	P_F50E	B01001_040E + B01001_041E + B01001_042E + B01001_043E + B01001_044E + B01001_045E + B01001_046E + B01001_047E + B01001_048E + B01001_049E
HOUSEHOLD INCOME		
100k or more	P_HI100E	B19001_014E + B19001_015E + B19001_016E + B19001_017E
FAMILY TYPE		
No related persons under 18	P_NCE	B11004_007E
EDUCATIONAL ATTAINMENT		
Bachelor's degree or higher	P_BACH	B15003_022E + B15003_023E + B15003_024E + B15003_025E

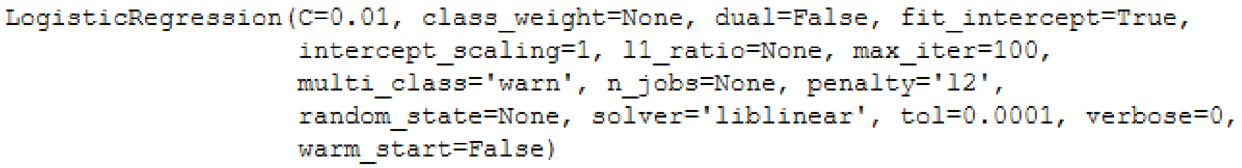
With the census data for Ventura and Santa Barbara gathered we now need to determine which census block are most like the records marked as Target = 1 in the Los Angeles file. After some manipulation and column remaining, the difference between the Target = 1 (include) and Target=0 (exclude) records can be seen in the following plot. The difference appears to be clear, the include group is older, has a higher average income, fewer households have kids at home, and they have more education.



Machine Learning

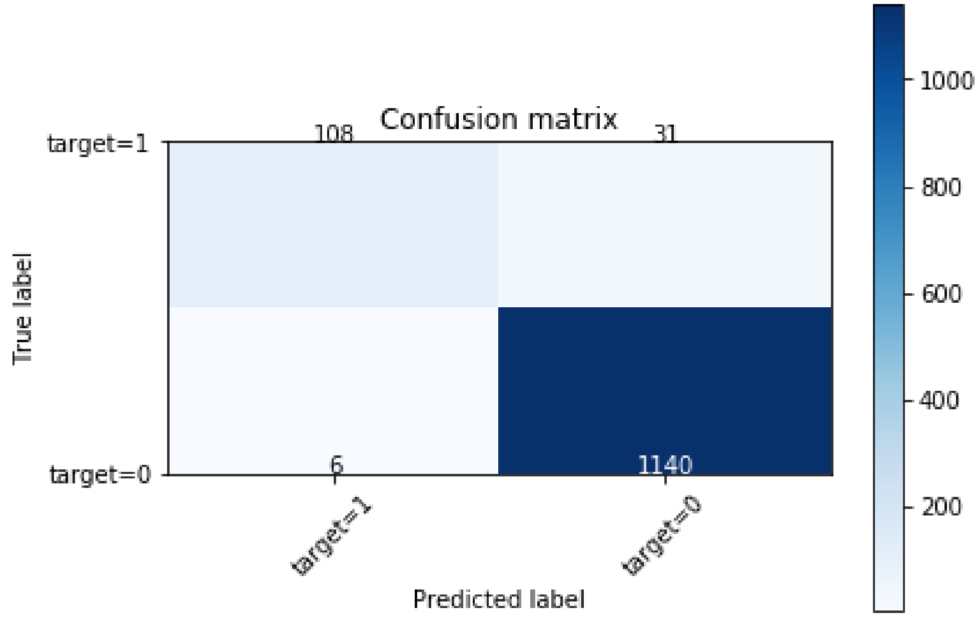
Although the difference between the two groups appears obvious, determining which census blocks in our new Ventura/Santa Barbara dataset most closely resemble the target=1 group across all five datapoints will not be easy. Unless of course we employ some machine learning in the form of Logistic Regression. The next step in our project will be to create a model that we can then run our new dataset through and determine the “target=1” census blocks.

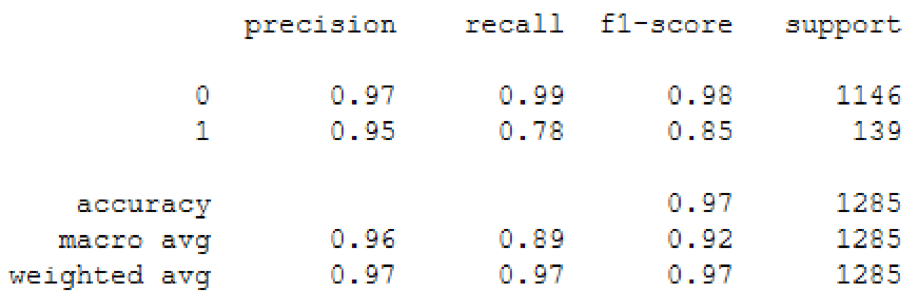
After preparing, scaling, and splitting the 6425 records in the Los Angeles dataset into Test and Train groups the following Logistic Regression process was run.



```
LogisticRegression(C=0.01, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='warn', n_jobs=None, penalty='l2',
                   random_state=None, solver='liblinear', tol=0.0001, verbose=0,
                   warm_start=False)
```

The outcome of this process can be seen in this Confusion matrix and classification report.

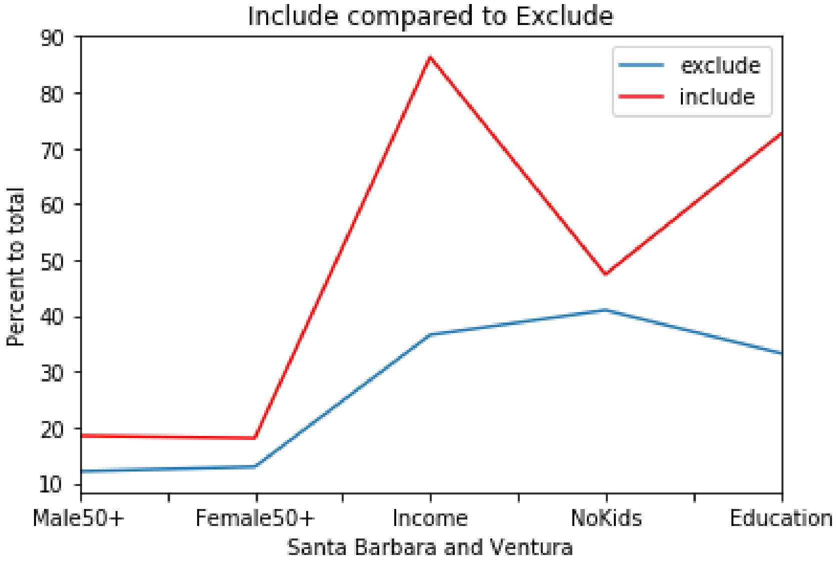


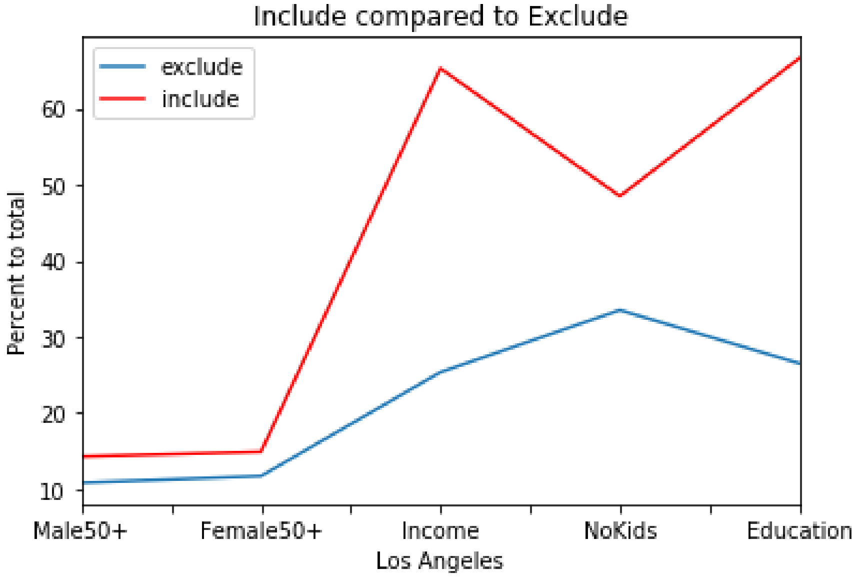


```
              precision    recall  f1-score   support

           0       0.97      0.99      0.98      1146
           1       0.95      0.78      0.85       139

    accuracy                           0.97      1285
   macro avg       0.96      0.89      0.92      1285
weighted avg       0.97      0.97      0.97      1285
```

Satisfied that this model will work our new dataset can now be appended with target values.

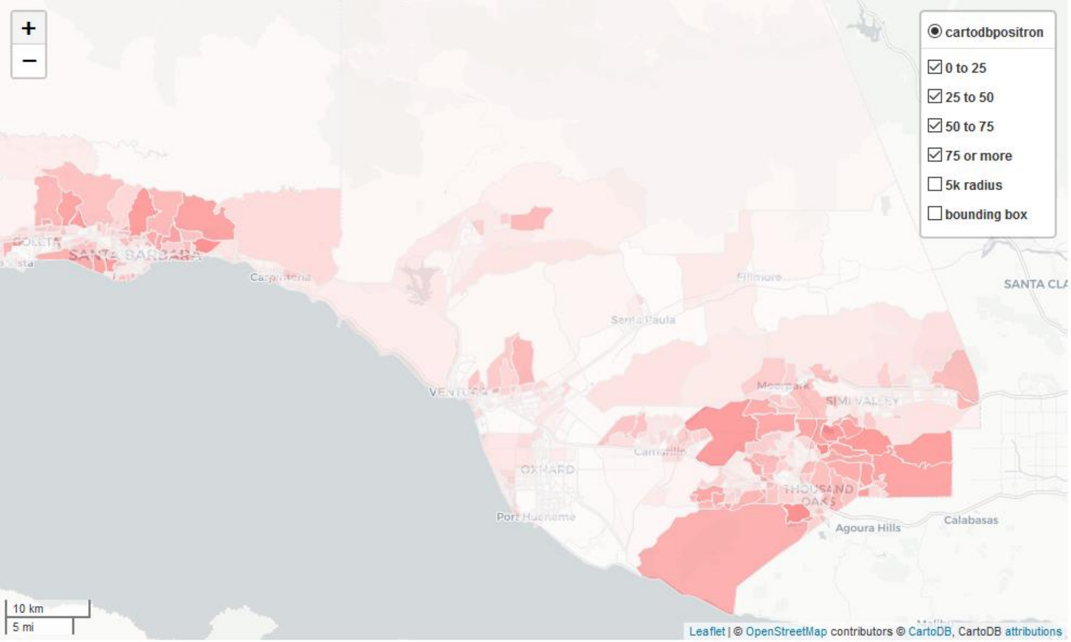
After running the new dataset through the model to determine the target value we can plot the values and compare the results to the original Los Angeles plot. Both plots show a similar relationship between the include and exclude target values.





Mapping the Data

With the Ventura and Santa Barbara dataset updated with a probable Target=1 value the data can be displayed on map using the folium library. Census Blocks have been divided into four layers based on the target value, 0 to .25, .25 to .50, .50 to .75, and .75 or more.



● cartodbpositron

☑ 0 to 25

☑ 25 to 50

☑ 50 to 75

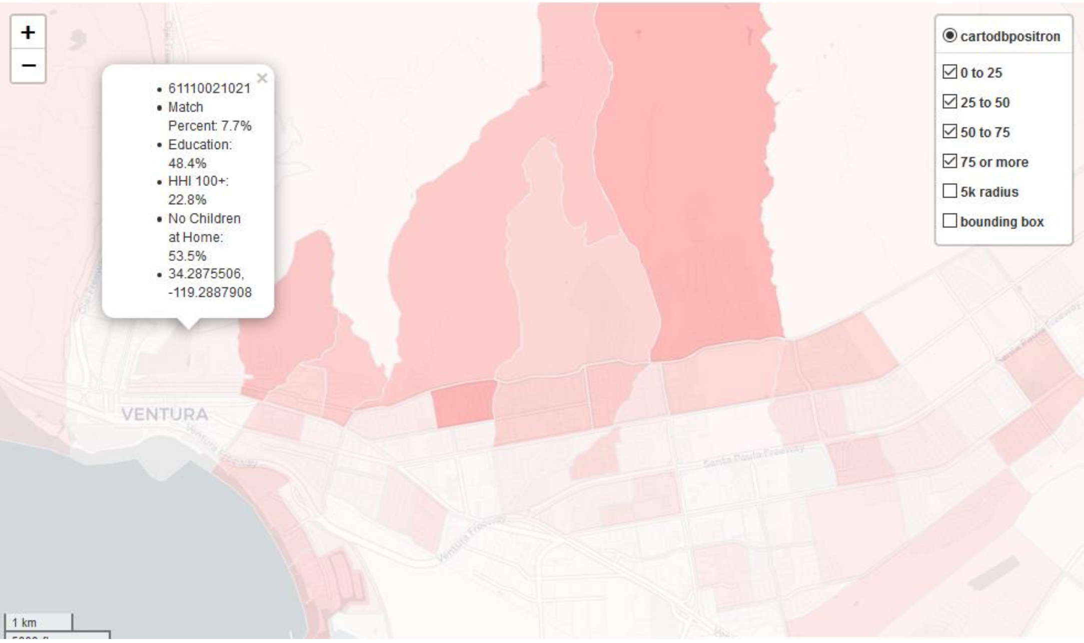
☑ 75 or more

☐ 5k radius

☐ bounding box

Leaflet | © OpenStreetMap contributors © CartoDB, CartoDB attributions

Detail view map



- 61110021021
- Match Percent: 7.7%
- Education: 48.4%
- HHI 100+: 22.8%
- No Children at Home: 53.5%
- 34.2875506, -119.2887908

● cartodbpositron

☑ 0 to 25

☑ 25 to 50

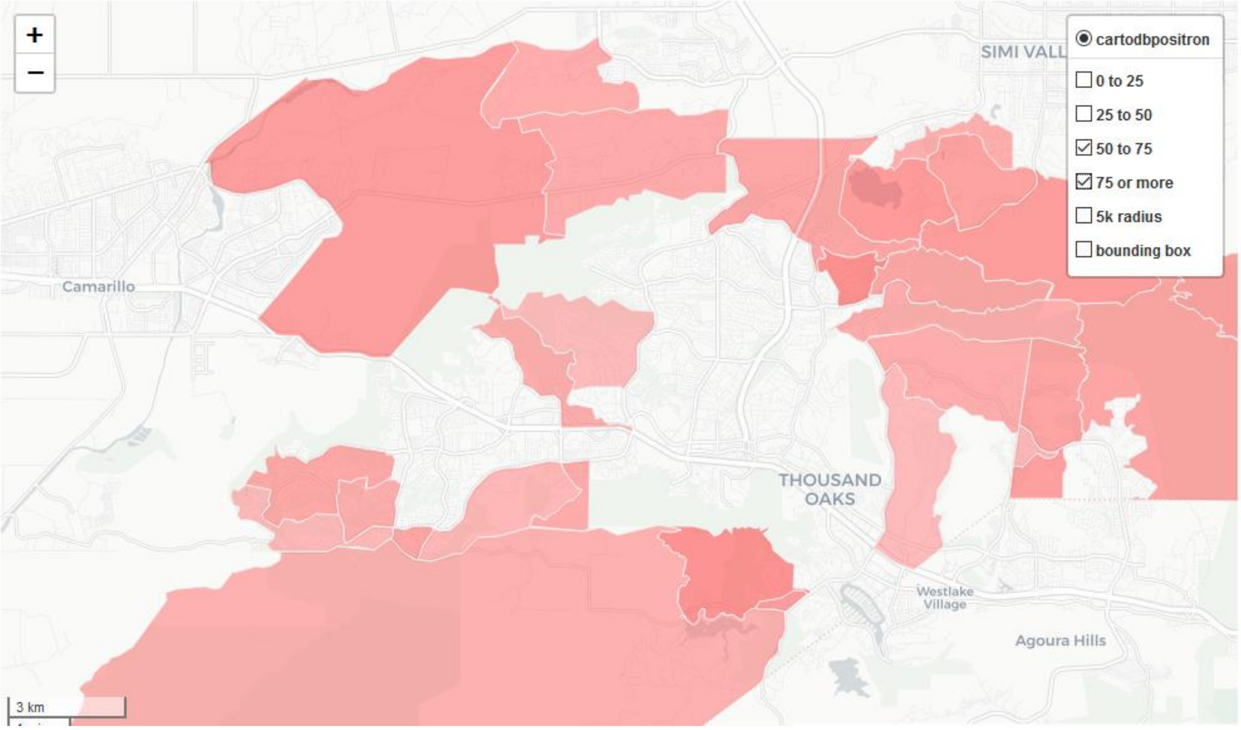
☑ 50 to 75

☑ 75 or more

☐ 5k radius

☐ bounding box

Detail map with some layers removed

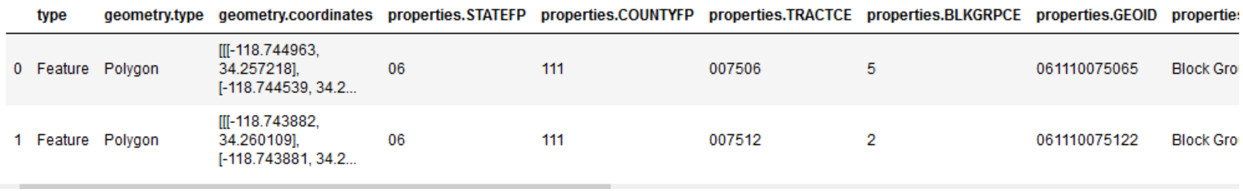


● cartodbpositron
□ 0 to 25
□ 25 to 50
☑ 50 to 75
☑ 75 or more
□ 5k radius
□ bounding box

TIGER/Line Shape files

In order to make the maps show above we first must have geographic border data that can be joined to the data we pulled with the US Census API. The Tiger/Line Shape files that contain these borders are readily available but are very large. For this project I downloaded the shape file and processed it offline to extract the data for Los Angeles, Ventura and Santa Barbara into json files that can be easily imported into the Jupyter notebook.

The processed shape files look like this when imported into a dataframe. You can see the GEOID column that will be used to join to the US Census files.



	type	geometry.type	geometry.coordinates	properties.STATEFP	properties.COUNTYFP	properties.TRACTCE	properties.BLKGRPCE	properties.GEOID	propertie:
0	Feature	Polygon	[[[-118.744963, 34.257218], [-118.744539, 34.2...	06	111	007506	5	061110075065	Block Gro
1	Feature	Polygon	[[[-118.743882, 34.260109], [-118.743881, 34.2...	06	111	007512	2	061110075122	Block Gro

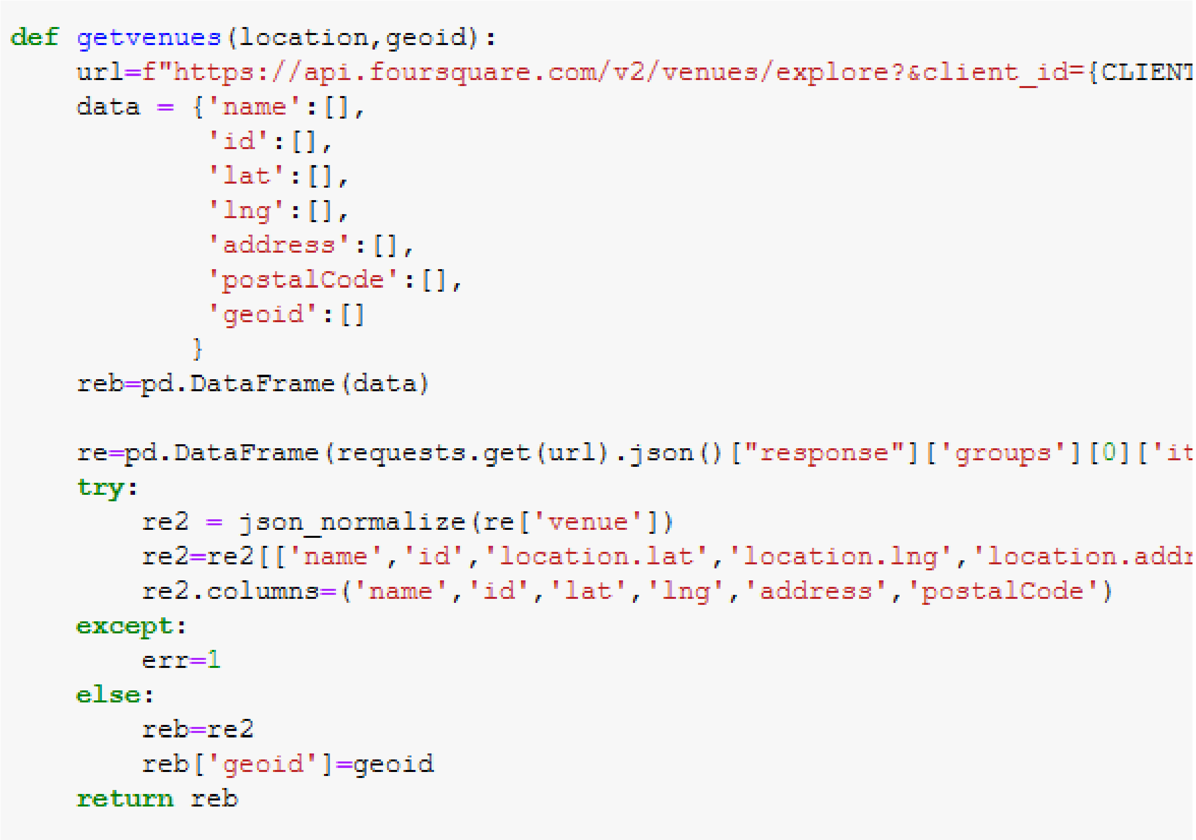
Sample json file

https://jordan-arthur.github.io/datascicapstone/counties_111.json

Foursquare API data

After joining the US Census data to the json shape file we now have all the information we need to use the Foursquare API to find businesses near are target census blocks and we know the census blocks to target for any particular business that hires the marketing company to send direct mail marketing pieces. The foursquare API is fast, simple to use and data rich. Adding the business venue information from Foursquare to this project was as easy as looking up the category id https://developer.foursquare.com/docs/resources/categories and passing it along with the location information to the API.

And processing the results is also simple

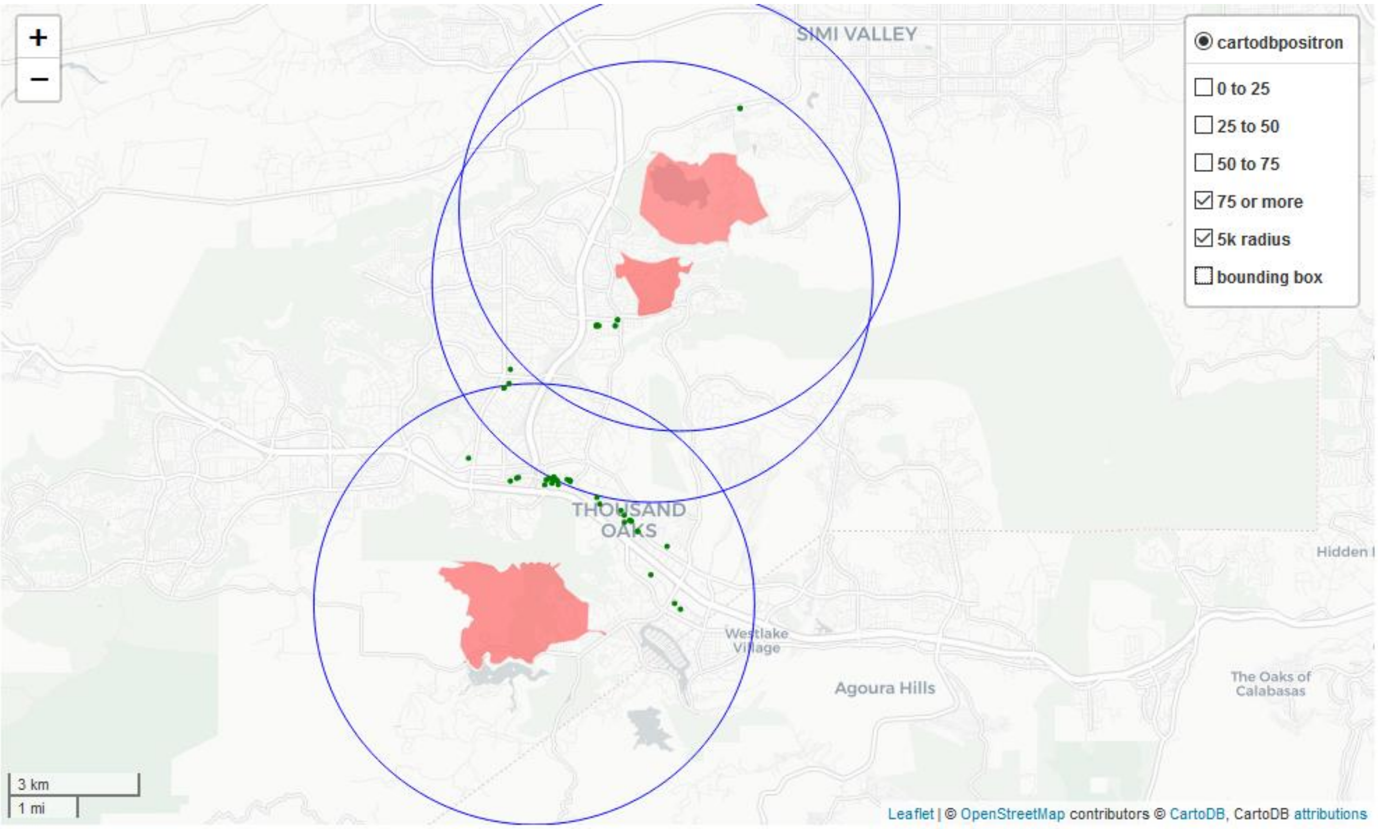


```
def getvenues(location,geoid):
    url=f"https://api.foursquare.com/v2/venues/explore?&client_id={CLIENT
    data = {'name':[],
            'id':[],
            'lat':[],
            'lng':[],
            'address':[],
            'postalCode':[],
            'geoid':[]
           }
    reb=pd.DataFrame(data)

    re=pd.DataFrame(requests.get(url).json()["response"]['groups'][0]['it
    try:
        re2 = json_normalize(re['venue'])
        re2=re2[['name','id','location.lat','location.lng','location.addr
        re2.columns=('name','id','lat','lng','address','postalCode')
    except:
        err=1
    else:
        reb=re2
        reb['geoid']=geoid
    return reb
```

Results

This Ventura county map shows red census blocks that have a probability percent of 75% or more of matching the target=1 value in our Los Angeles county data that was used in or machine learning model. The blue circles represent the 5-kilometer business search area that we used in our call to the Foursquare API. The green markers are the hair salon businesses returned from Foursquare.

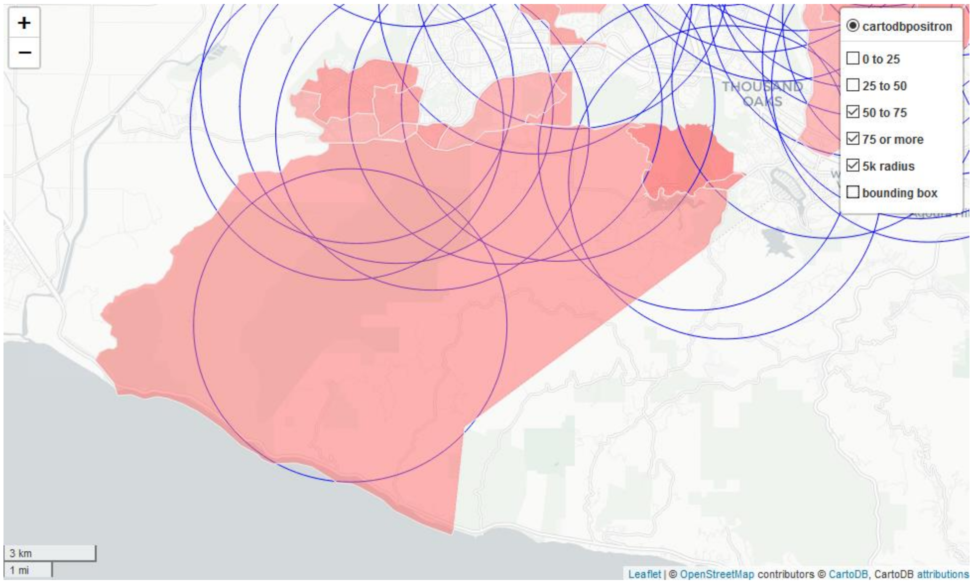
live version of this map here



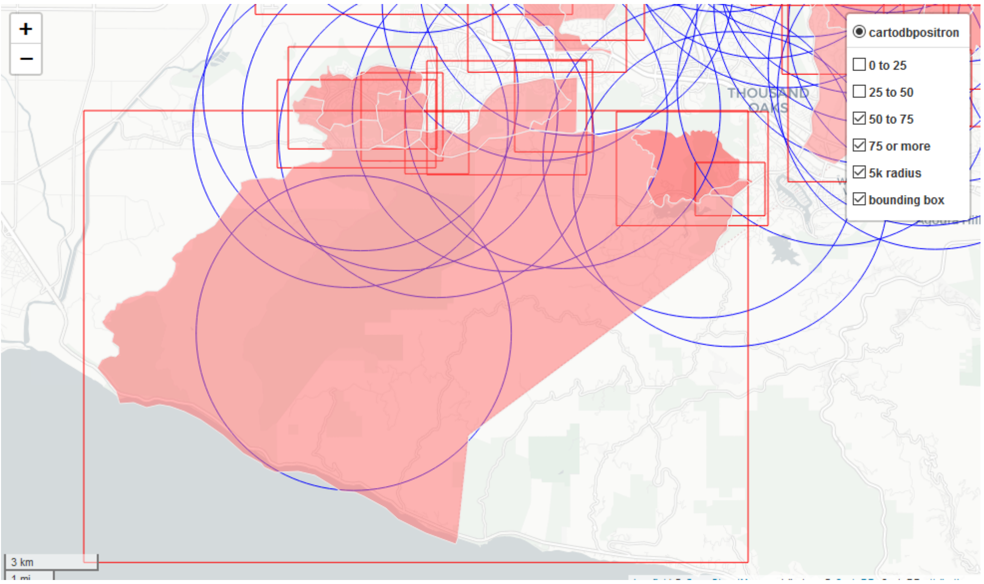
● cartodbpositron
□ 0 to 25
□ 25 to 50
□ 50 to 75
☑ 75 or more
☑ 5k radius
□ bounding box

Discussion

One issue with the requirements for this project was that businesses within 5 kilometers of the target census blocks center where to be identified. A few rural census blocks are larger than 5 kilometers, so the radius of the search circle excluded portions of the census block. This is illustrated here.



To mitigate this issue a second venue search is performed using a bounding box around the census block.



Conclusion

The abundance and quality of the data and tools available to accomplish a project like this one is amazing. That said the marketing company in this report may get better results in targeting customers by paying for personalized data about consumers. By buying quality data they would not have to limit their mailings to small geographic areas like census blocks but could target the same number of customers, but more qualified customer, spread out over a larger area and get a better engagement rate.