New Dental Office in Miami

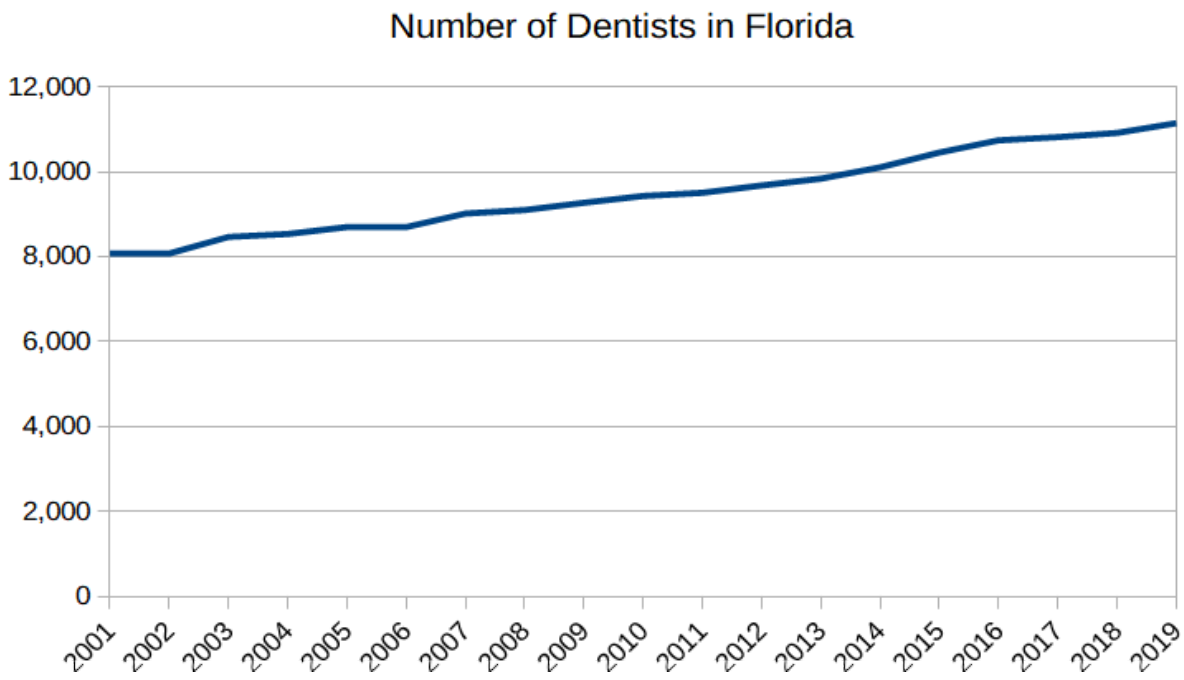
Applied Data Science Capstone Report

Pawel Kalinowski

06/09/2020

INTRODUCTION

According to American Dental Association each year the number of practicing dentists in Florida grows by average by 170.



Most new dentists start as Associates in the established practices, but finally they want to open their own offices.

Here comes the question and the business problem -- what is the best place to open the new dental practice?

I have examined in this exercise a naïve approach of comparing population numbers in a given neighborhood to density of existing dental practices. Places with least dental practice density per population will be considered as candidate spots.

This approach was further expanded by comparing additional data as wealth i.e. median household income.

I am narrowing the scope of the exercise to the Miami metropolitan area, with focus on Miami-Dade county.

HYPOTHESIS

The following hypothesis was considered: number of Dental Offices in a given area is in (linear) relation to this area population and median household income.

The hypothesis was scrutinized using multi linear regression.

The resulting model should present information about the expected number of dental offices in the area. Comparing this value with the actual data we should be able to find areas where new offices should be established.

Finally, the result will name the best place(s) to open a Dental Office in the Miami area.

DATA

The exercise will base on the following data sets:

1. Foursquare (the business type `Dentist's Office`) - location and popularity

2. Geographical / geocoding data zip code boundaries] (https://gis-mdc.opendata.arcgis.com/datasets/zip-code), [KML file with boundaries] (https://opendata.arcgis.com/datasets/fee863cb3da0417fa8b5aaf6b671f8a7_0.kml?outSR=%7B%22latestWkid%22%3A3857%2C%22wkid%22%3A102100%7D)

3. Demographics [by zip code](https://worldpopulationreview.com/zips/florida/)

4. Zip Code Characteristics: Mean and Median Household Income (https://www.psc.isr.umich.edu/dis/census/Features/tract2zip/MedianZIP-3.xlsx)

PROCEDURE

1. Geographical data in form of **KLM** file was downloaded from **arcgis.com** site. The file was converted to **.topojson** format using MyGeoData on-line converter. **Zip_Code.topojson** file is natively recognized by **folium** library to create choropleth maps. The KML file was also converted to Zip_Code.tsv using bash tools. The file has a following structure:

ZIP_CODE1<tab>point1.1<tab>point1.2<tab>...<tab>point1.N

ZIP_CODE2<tab>point2.1<tab>point2.2<tab>...<tab>point2.M

...

where *point* is a pair of (latitude, longitude) coordinates. The file **Zip_Code.tsv** is then loaded into a dictionary **zip_borders**.

2. Population by ZIP code was downloaded in **.csv** format from **worldpopulationreview.com** as file **population_by_zip.csv**. This data was directly imported into Pandas DataFrame, then converted to integers:

zip	city	...	pop
33012	Hialeah	...	75666
33024	Hollywood	...	75306
33023	Hollywood	...	73671
33311	Fort Lauderdale	...	73034
33025	Hollywood	...	71763

3. Household Median Income was downloaded from **umich.edu**. The **XLSX** file was manually converted to **MedianZIP-3.csv** and loaded to **zip_income** DataFrame:

Zip	Median	Mean	Pop
1001	56,663	66,688	16,445
1002	49,853	75,063	28,069

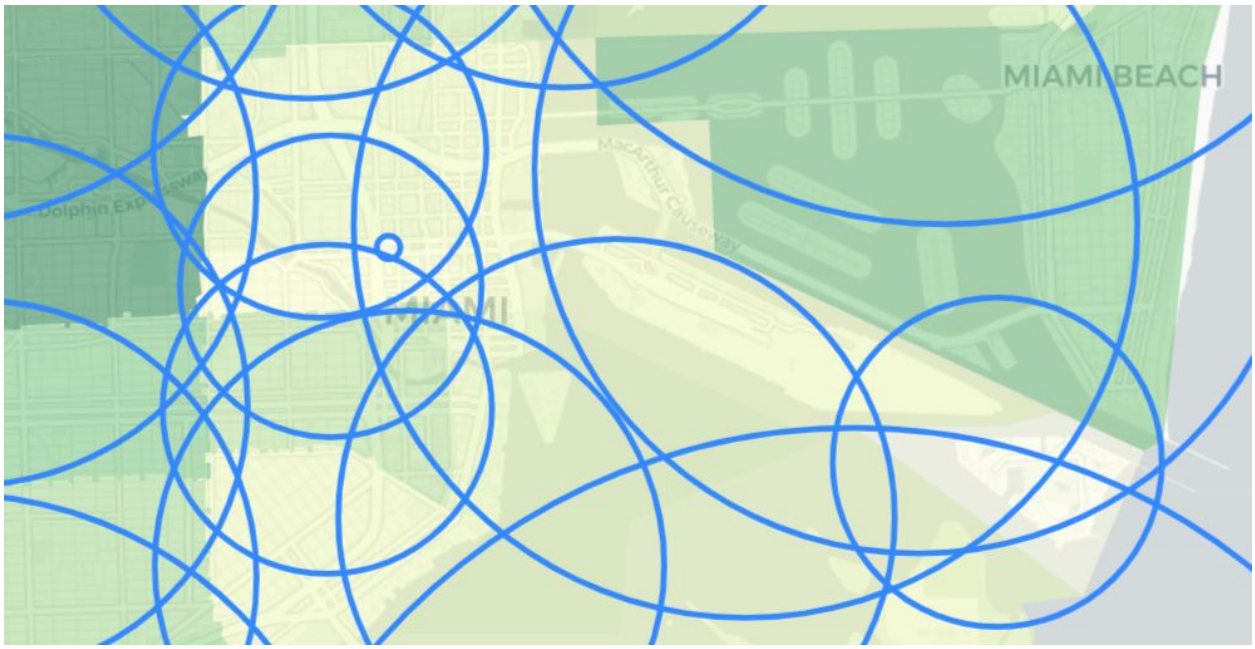
1003	28,462	35,121	8,491
1005	75,423	82,442	4,798
1007	79,076	85,802	12,962

4. The most important set of data comes from **Foursquare.** The right endpoint and API call for this exercise was **venues/search**. It requires **latitude, longitude** and **radius** for the queried area with additional filtering possible. Here we used filtering by category. As we wanted to get a list of all Dental Offices in the Greater Miami area, we might call the **API** just once with a large enough radius to cover it all. Unfortunately, the hard limit for returned results in **Foursquare API** is **50**, so we turned to querying each ZIP code individually.

   The function **get_dentists(lat, lng, r)** returns DataFrame with all returned offices in a given radius.

   Now we need to run the above function over all ZIP areas. The dictionary **zip_borders** contains coordinates of points making a border of a ZIP area. So, we need to find a **smallest enclosing circle** for each set of points.

   The enclosed open source library https://www.nayuki.io/page/smallest-enclosing-circle fits this purpose. The original library was enhanced with **make_circle_metric** function that returns an approximate radius in meters, as required by **Foursquare API**.
   Below a portion of the map with smallest enclosing circles overlaid over few Miami ZIP code areas.

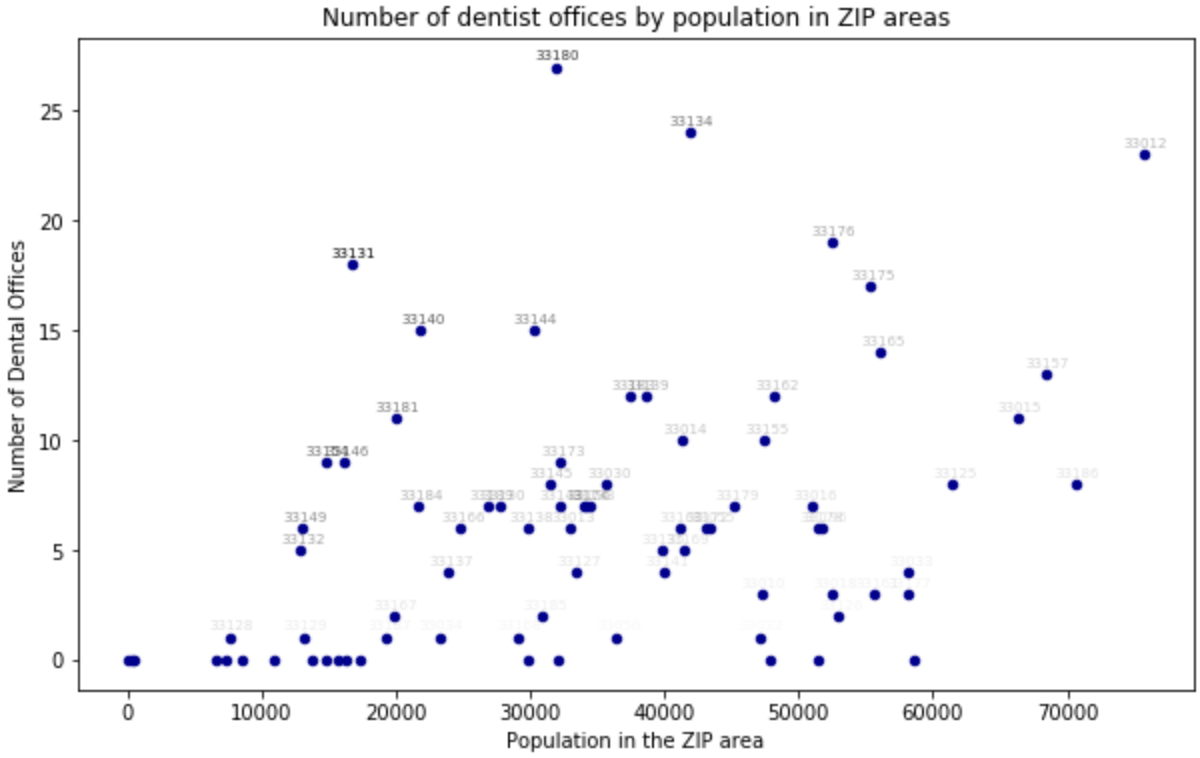


MODEL

DataFrame with all collected data was prepared. Excerpt:

zip	population	dentists	income
33035	15658	0	49203
33010	47231	3	26222
33154	14790	9	65690
33037	10955	0	46276
33140	21807	15	86562

Scatterplot of population by number of dental offices does suggest some kind of relation between these two:



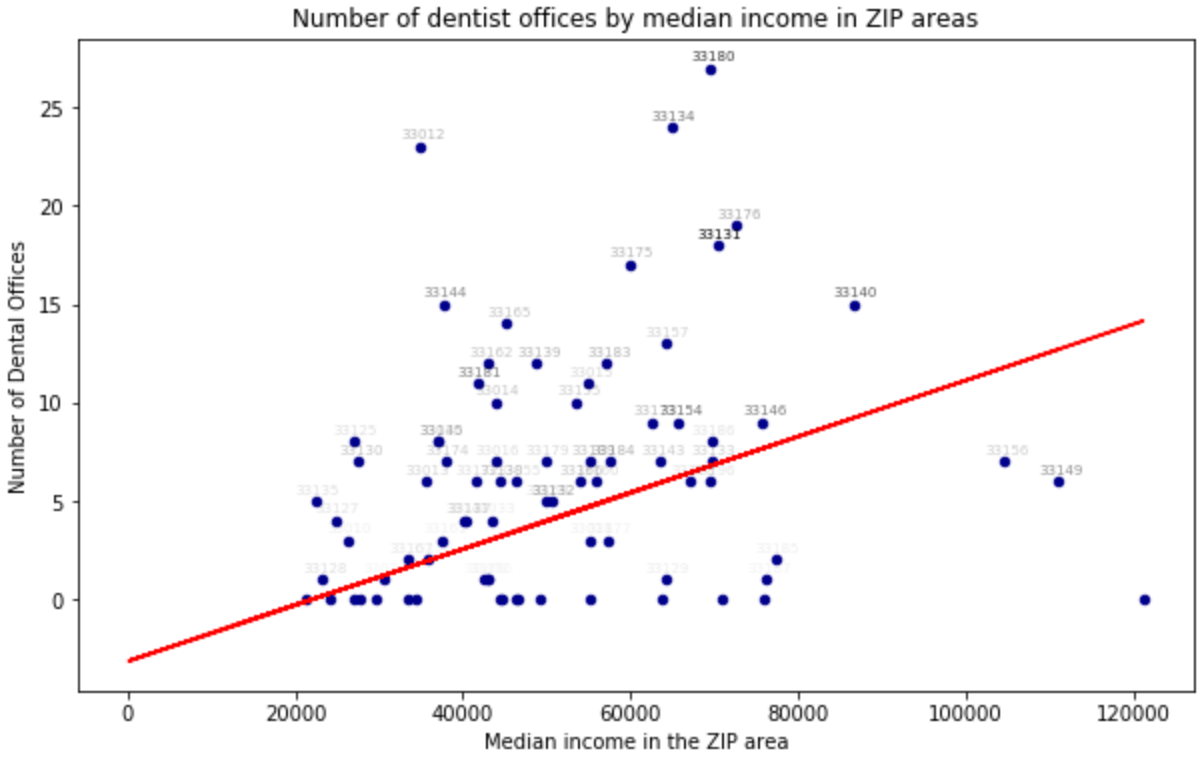
The linear regression has been calculated using **linear_model** class from **sklearn** library:

```
from sklearn import linear_model
regr = linear_model.LinearRegression()
x = np.asanyarray(zip_data[['pop','income']])
y = np.asanyarray(zip_data[['dentists']])
regr.fit (x, y)
# The coefficients
print ('Coefficients: ', regr.coef_)
y_hat= regr.predict(zip_data[['pop', 'income']])
x = np.asanyarray(zip_data[['pop', 'income']])
y = np.asanyarray(zip_data[['dentists']])
print("Residual sum of squares: %.2f"
      % np.mean((y_hat - y) ** 2))
# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(x, y))
```

```
Coefficients:  [[1.42718244e-04 8.73925591e-05]]
```

Residual sum of squares: 28.27
Variance score: 0.22

With such results several attempts to clean-up data have been implemented, including narrowing the dataset to non-empty areas, excluding obvious outliers. Unfortunately none of these attempts improved the Variance Score, therefore decision has been made to continue with the above model.



The above graph shows the regression line overlaid on scatterplot of income by number of Dental Offices.

Subtracting $\hat{y}$ (predicted number of Dental Offices) from the actual number of offices with the following code

```
result = np.rint(y-y_hat).astype(int)
```

and combining it with the zip_data dataframe gives a following array where the **result** column shows the difference between the real and expected number of offices. “-3” means that the model predicts 3 more dental offices.

zip	pop	dentists	income	ratio	result
33035	15658	0	49203	0.000000	-3
33010	47231	3	26222	0.635176	-3
33154	14790	9	65690	6.085193	4
33037	10955	0	46276	0.000000	-2
33140	21807	15	86562	6.878525	7

The following code extracts the top 5 prospective areas:

```
Prospective = zip_data.sort_values(by=['result','income','pop'])[0:5].index
```

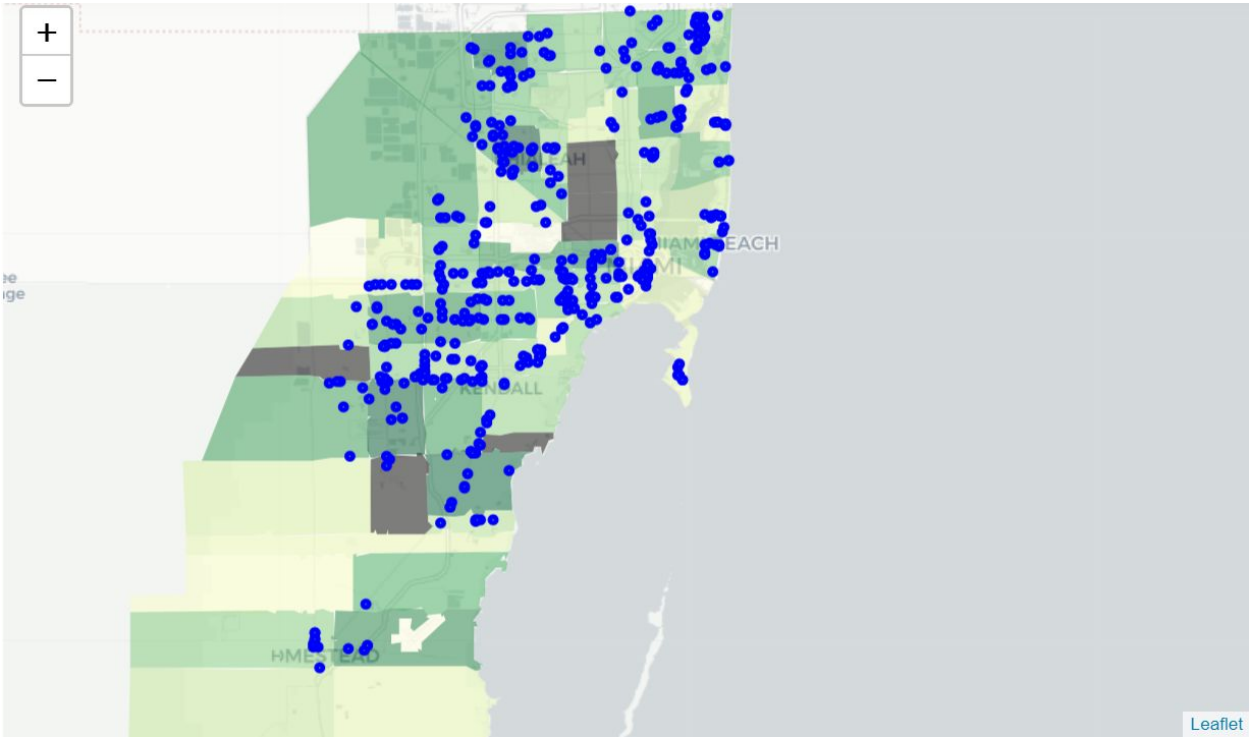
zip	pop	dentists	income	ratio	result
33193	51378	0	55287	0.000000	-9
33158	6641	0	121193	0.000000	-8
33142	58574	0	24024	0.000000	-7
33177	58129	3	57246	0.516094	-7
33147	47834	0	29739	0.000000	-6

List of prospective ZIP codes selected:

```
[33193, 33158, 33142, 33177, 33147]
```

VISUALIZATION

The selected areas are marked dark on the choropleth map:



Blue dots denote the existing Dental Offices.

CONCLUSION

The used model is too simple to accurately predict the number of Dental Offices per area, but still may have a value for a quick assessment of the neighborhoods.

Accuracy of the model has a room for improvement, like additional data on ZIP areas, e.g.

- Commercial real estate rent rates and availability
- Criminal activity
- Age profile of the population

Capture more information on Dental Practices for further exploration:

- Number of Doctors
- Number of Patients
- Review score
- Financial data