Real Estate Developer Opportunity Finder

IBM Data Science Professional Certificate

Capstone Project, November 2019

Introduction

Business Problem

We are a consulting firm specialized on advising real-estate developers on potential business opportunities in cities around the US and Canada.

Clients come to us for advice where to start new developments e.g. a new mall or a new office building in a certain city or a particular region. Based on our market expertise, business intelligence, and internal databases we seek interesting areas or neighborhoods which we then rank and propose to our clients. The client will receive a landscape report which gives a market overview and highlights most promising areas.

After having learned about Foursquare we decided to start a project to evaluate to what extend we can leverage data from Foursquare Places for our analyses. As a first step we set out to prepare an analysis for mall developments in NYC and/or one other major US city. After this project being completed and evaluated we will decide if we take this further or not.

Analytical Approach

In order to identify „Mall“ business opportunities we will

- create a grid over the target city or use a given segmentation as postal code areas or boroughs
- find the existing malls in a given section
- try to find public demographic data for these sections such as population, age, income, etc
- combine and analyze these information sources with Foursquare data in order to derive a ranking

Data Requirements

Public Geo-Demografic Data

The analysis will depend on the availability of demographic data for the city or region to be evaluated.

Data will have to be available for subareas of the city defined by e.g.

- Postal Code Areas,
- Boroughs
- Neighborhoods
- or another grid of coordinates laid across the city

For each subarea I will try to find the following data

- geo-coordinates (mandatory)
- population (mandatory)
- average income or income distribution (optional)
- gender distribution (optional)
- - average age or age distribution (optional)

Foursquare Data

For each subarea we will determine the number of already existing malls in that region.

In oder to do so I will use the Foursquare Venues API

https://developer.foursquare.com/docs/api/venues/

and there the API endpoints

- search and/or explore

The relevant subarea will be targeted using the parameters

- ll (geo location)

- ne/sw (north-east and south-west corner a rectangular area (available for ‚search‘ endpoint only))

depending available geo-coordinates for the demographic data.

Malls will be identified by their Foursquare category

https://developer.foursquare.com/docs/resources/categories

categoryId: 4bf58dd8d48988d1fd941735

In addition we may consider shopping streets categoryId: 5744ccdfe4b0c0459246b4dc

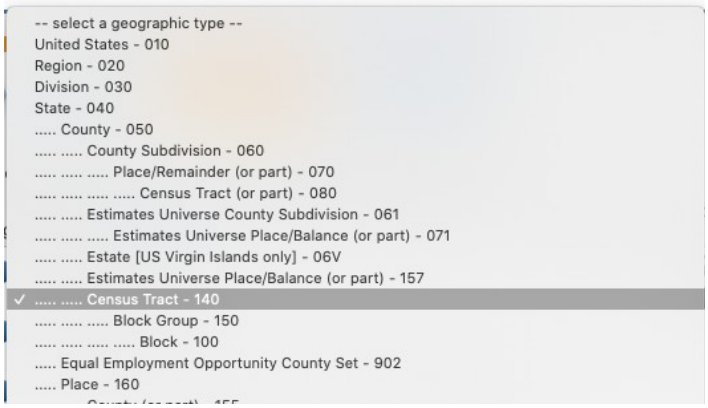
Data Acquisition

Demoscopice Data

After some research I identified results of the American Community Survey as a suitable dataset, available on https://factfinder.census.gov/. There go to “Download Center”.

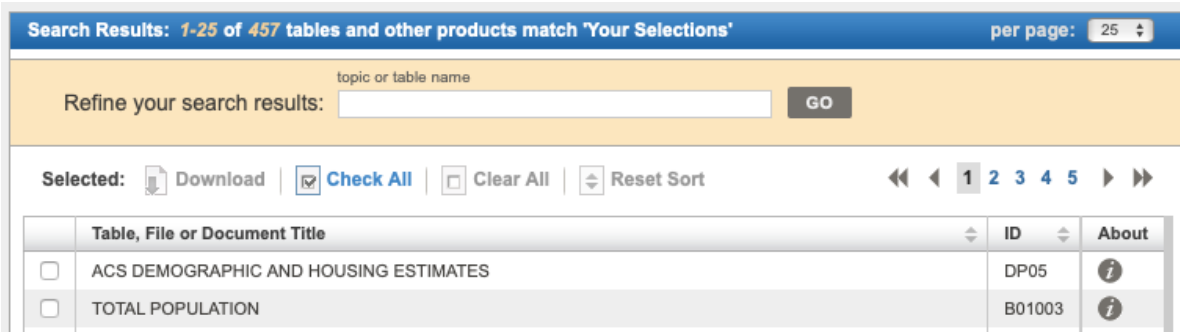
The data are based on the 2010 US census and are propagated into the future by prognostic models. I chose the ACS 2017 5YR dataset for our analysis.

Data is made available in different geographical resolutions e.g. State, County, Place.



For the study at hand I use “Census Tracts” as underlying geographical unit. Census tracts are small, relatively permanent statistical subdivisions of a county. For further discussion see https://www2.census.gov/geo/pdfs/education/CensusTracts.pdf

For each geographical sub-division a large variety of data is available. Selection is possible using the search field.



For our analysis I used the POPULATION data set. The following screenshot shows the file structure.

	GEO.id	GEO.id2	GEO.display-label	HD01_VD01	HD02_VD01
0	Id	Id2	Geography	Estimate; Total	Margin of Error; Total
1	1400000US24510010100	24510010100	Census Tract 101, Baltimore city, Maryland	3201	255
2	1400000US24510010200	24510010200	Census Tract 102, Baltimore city, Maryland	3145	228
3	1400000US24510010300	24510010300	Census Tract 103, Baltimore city, Maryland	2552	241
4	1400000US24510010400	24510010400	Census Tract 104, Baltimore city, Maryland	2573	268

After removing column GEO.id and row “0” and some renaming I used the data in the following table format.

	GEOID	GEO.display-label	Population	Error
1	24510010100	Census Tract 101, Baltimore city, Maryland	3201	255
2	24510010200	Census Tract 102, Baltimore city, Maryland	3145	228

Geographic Data

The census data themselves do not contain any geo information. In order to relate census data to the actual location of the Census Tract a so called “gazetteer” file is provided on census.gov see https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.2017.html.

For a record layout of the gazetteer file see https://www.census.gov/programs-surveys/geography/technical-documentation/records-layout/gaz-record-layouts.html.

Please note that the ALAND (land area) and AWATER (water area) areas are provided in square meters.

	USPS	GEOID	ALAND	AWATER	ALAND_SQMI	AWATER_SQMI	INTPTLAT	INTPTLONG
0	AL	1001020100	9817812	28435	3.791	0.011	32.481959	-86.491338
1	AL	1001020200	3325679	5670	1.284	0.002	32.475758	-86.472468

For further analysis the two data sets were merged.

Foursquare Data

In order to identify the best locations for new mall developments I need the location of existing malls since I assume that one should develop in areas which are under-supplied with malls.

I use the search endpoint of the Foursquare API and search for categoryIds 4bf58dd8d48988d1fd941735, and 5744ccdfe4b0c0459246b4dc' which represent malls and shopping streets.

ULR: 'https://api.foursquare.com/v2/venues/search?&

client_id=MU4OZMN3032OXPGG5JSJSFKLQ0F0JJ4JDITTP4B4FG3J5RB2&
client_secret=KXMHI0MF40B3G2A3DCGNGEHHGSJ2BPF2LQNCPE0KCNNKDAB1&
v=20180605&
ll=39.299511200000005,76.609125&
radius=20000&
limit=100&
categoryId=4bf58dd8d48988d1fd941735,5744ccdfe4b0c0459246b4dc'

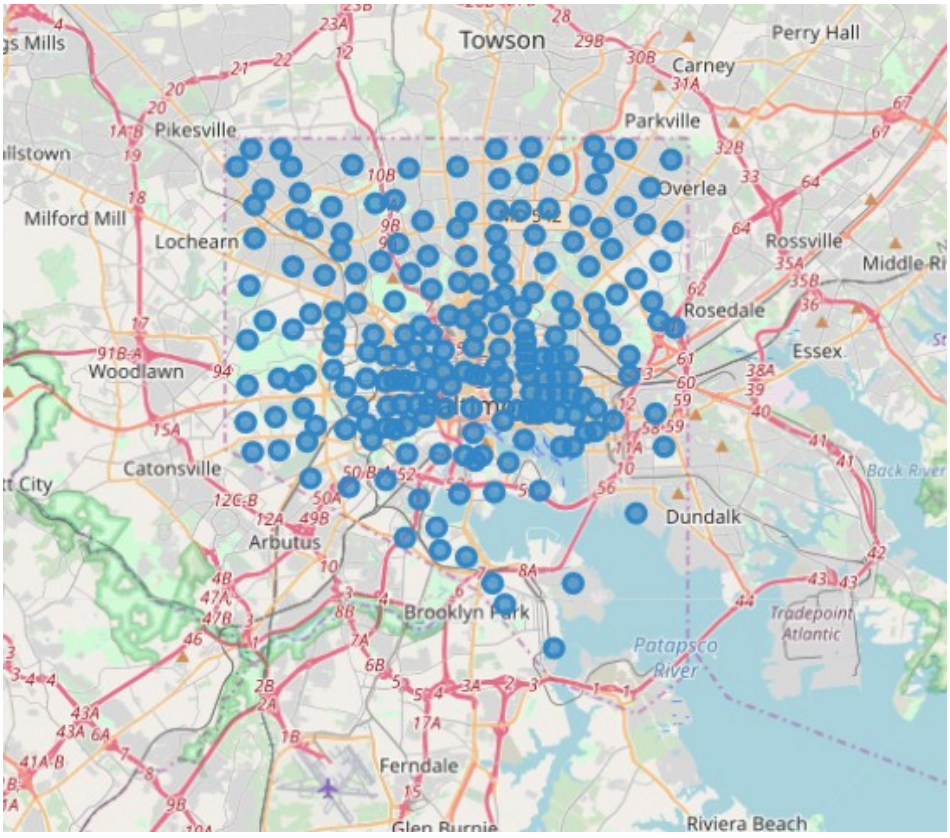
Exploratory Data Analysis

In a first step a simple Dataframe.describe() was used to get a first impression of the data

	GEOID	Population	Error	ALAND	AWATER	ALAND_SQMI	AWATER_SQMI	INTPTLAT	INTPTLONG
count	2.000000e+02	200.000000	200.000000	2.000000e+02	2.000000e+02	200.000000	200.000000	200.000000	200.000000
mean	2.451019e+10	3098.980000	382.800000	1.048218e+06	1.438381e+05	0.404740	0.055530	39.310767	-76.618382
std	8.195549e+04	1381.106508	171.904905	1.165135e+06	9.071620e+05	0.449868	0.350212	0.031201	0.044089
min	2.451001e+10	0.000000	12.000000	1.552440e+05	0.000000e+00	0.060000	0.000000	39.217624	-76.706489
25%	2.451012e+10	2048.000000	249.250000	3.879285e+05	0.000000e+00	0.150000	0.000000	39.289274	-76.648776
50%	2.451020e+10	2905.500000	354.500000	7.858755e+05	0.000000e+00	0.303500	0.000000	39.307377	-76.615444
75%	2.451026e+10	4049.500000	491.000000	1.208222e+06	0.000000e+00	0.466500	0.000000	39.334280	-76.585213
max	2.451028e+10	7144.000000	999.000000	1.000742e+07	9.589221e+06	3.864000	3.702000	39.369990	-76.535045

There a 200 census tracts related to Baltimore city with an average population of 3100 people. The average land area is 1.05 square kilometers, the northernmost tract has a latitude of 39.370, the southernmost 39.218. The difference of 0.152 degree corresponds to roughly 16.9 km (1 degree of latitude or longitude corresponds to around 111 km).

In order to get a better feeling for the data at hand the following plot shows the Census Tracts across a map of Baltimore.



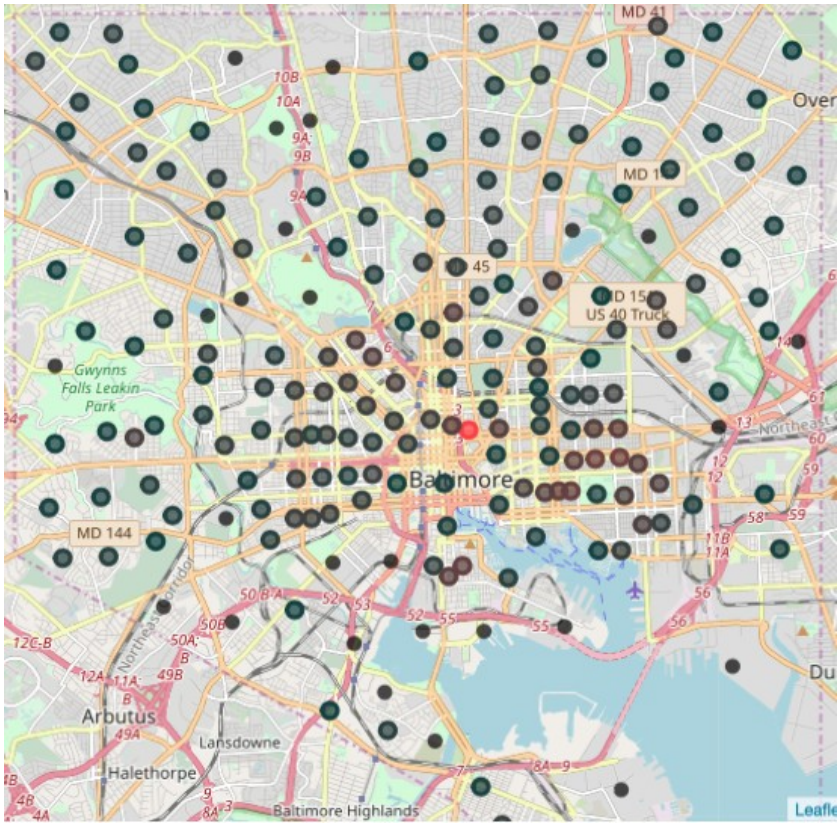
Since the census tracts are not evenly spaced and do not have the same size I created three additional columns:

ATOTAL = ALAND + AWATER

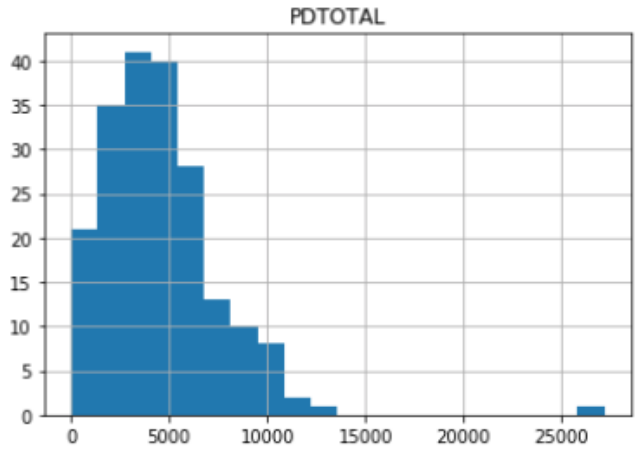
PDTOTAL = Population / ATOTAL (Population density)

PDLAND = Population / ALAND (Population density per land area)

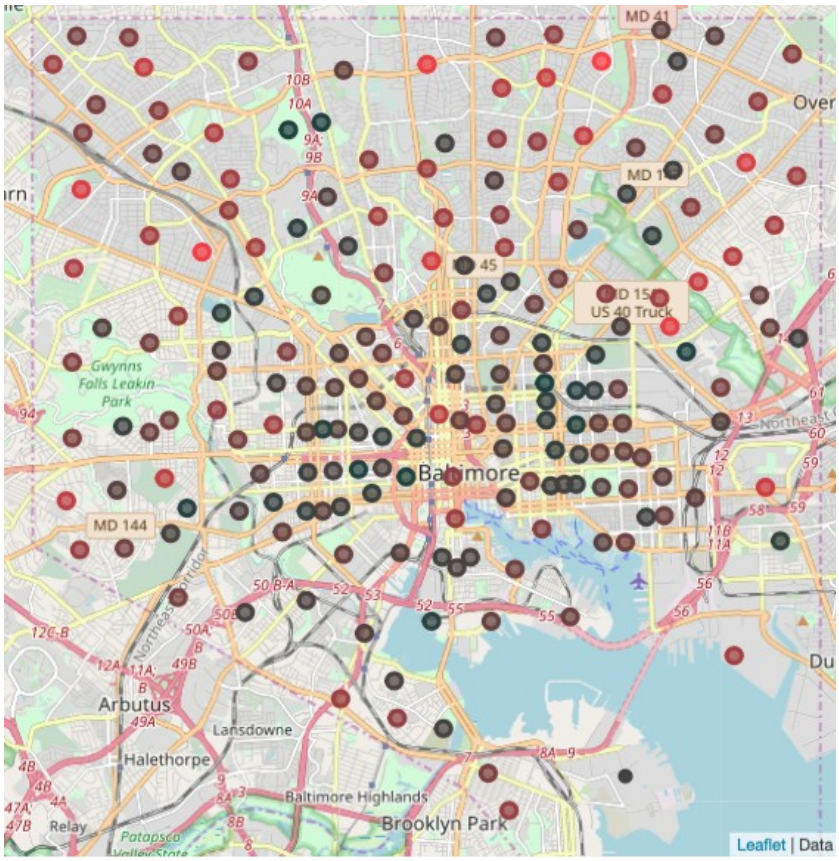
The following plot shows the Census Tracts with color coding the population density (people per square kilometer, black: low, red: high).



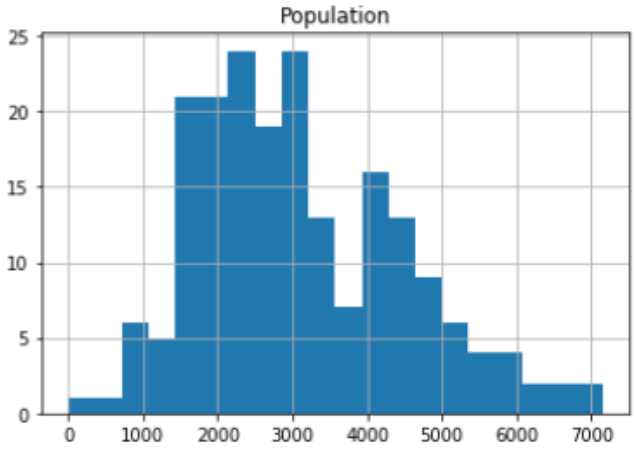
While the Census Tracts seem to have a relatively constant population density there is one outlier in the center. In order to get a better understanding of the distribution here is a histogram of population densities.The histogram clearly shows the outlier at beyond 25k in the city center.



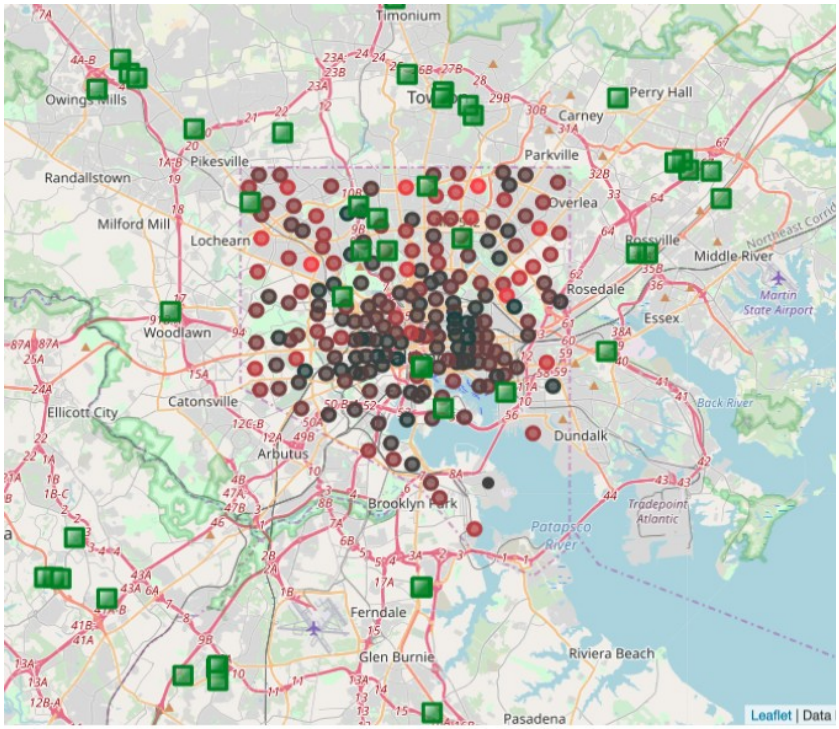
For the original business problem the actual population as show in the plot below is a better measure of attractiveness for a mall developer than the population density.



The histogram now shows a distribution without outliers



In the next step I add the mall data retrieved from Foursquare. The following image shows the map of Baltimore with all malls within a radius of 20 km.



The data set is now complete and I can proceed to modelling.

Modelling

Our goal is to identifiy attractive locations for new mall developments. Quality criteria are:

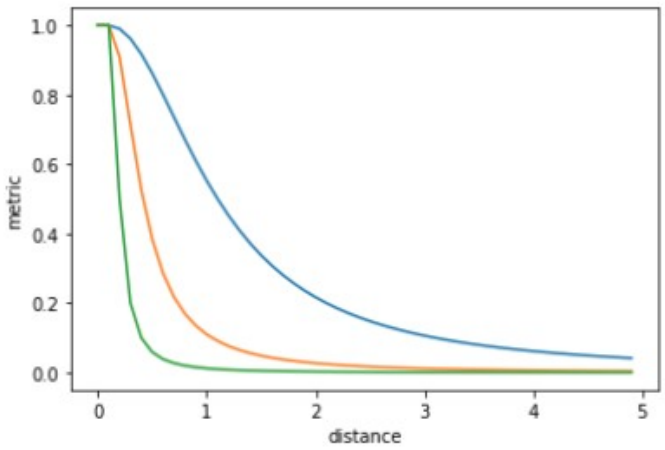
- population close to the mall
- distance to other malls

The population is given for each Census Tract, the distance measure remains to be created. Using the geo-coordinates and the fact that one degree of latitude or longitude corresponds to 111 km I determine the Euclidian distance between Census Tract and mall.

In oder to create a metric for the attractiveness of the location I start by calculating the Nearness between Census Tract and existing mall.

```
def nearness(s, walking_distance, laziness):
    if s <=walking_distance:
        return 1
    else:
        return 1 / (1 + laziness*((s-walking_distance)/walking_distance)**2)
```

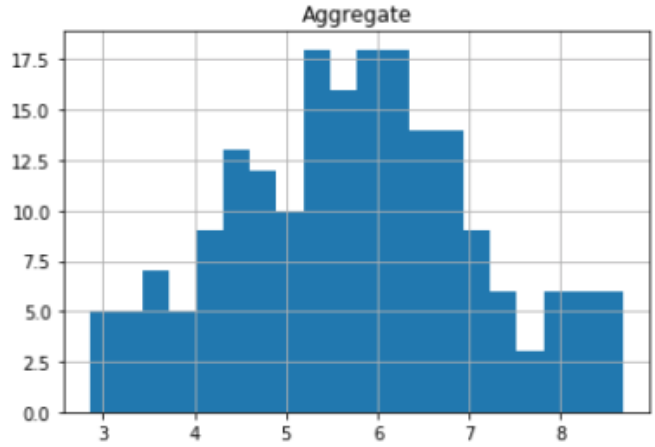
The following plot shows the metric for slope 1 (green), 0.1 (orange), and 0.01 (blue)



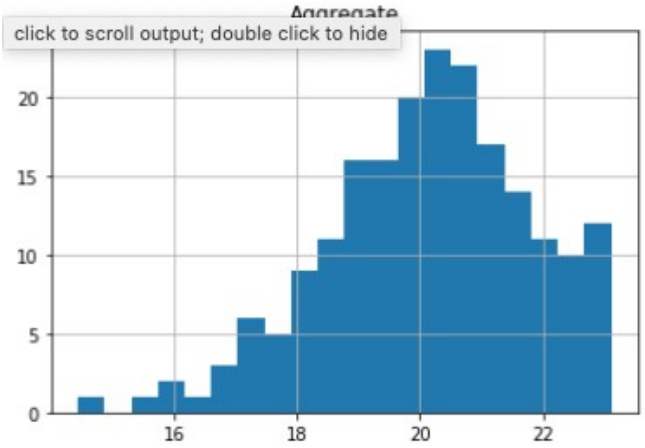
In the next step I calculate the Nearness for each Census Tract the in respect to each mall. Next again for each Census Tract I add-up all those metrics to create an “Aggregate Nearness”.

Intuition: People are expected to go to a mall in walking distance. This is assumed to have a weight 1. How far they go beyond that depends on how lazy they are. Low Laziness increases the weight for larger distances.

The following image shows the histogram depicting the Census Tract result for a slope of 0.1

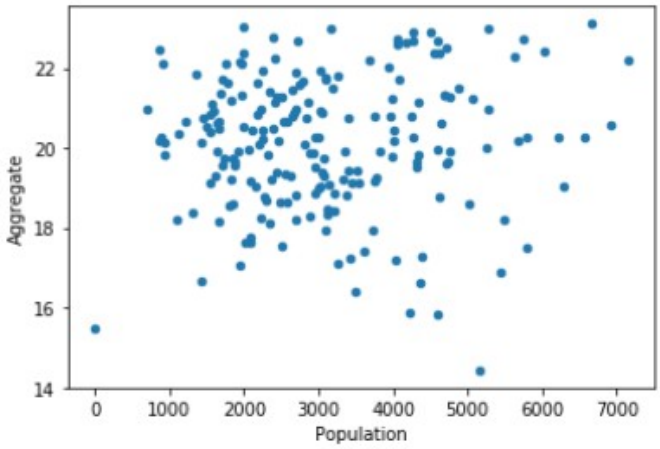


and a slope of 0.01.



Remember that high values of the aggregate Nearness metric mean that there are many malls nearby. Low Laziness values therefore provide better discrimination than hight Laziness values, because they take malls further away into account as well.

In order to calculate the actual score of the location I have to include the population as well. The following chart is a scatter plot of aggregate Nearness metric vs Population

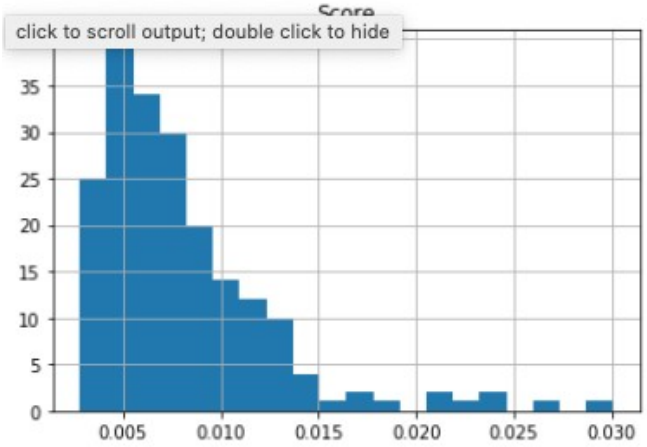


There is a certain but weak correlation between the variables which means that areas with high population have a higher number of malls nearby. However even areas with small population have a high coverage of malls.

In the lower right quadrant there is a certain number of Census Tracts which have a large population but not many malls in the immediate vicinity. Since the actual correlation between Aggregate Distance Metric and Population is below 10% I refrain from doing any regression.

Starting with the actual aggregate Nearness metric which indicates a good location when the value is small I calculate the final score as aggregate metric divided by the Population in the Census Tract. Please remember that the score indicates a good location when it is small.

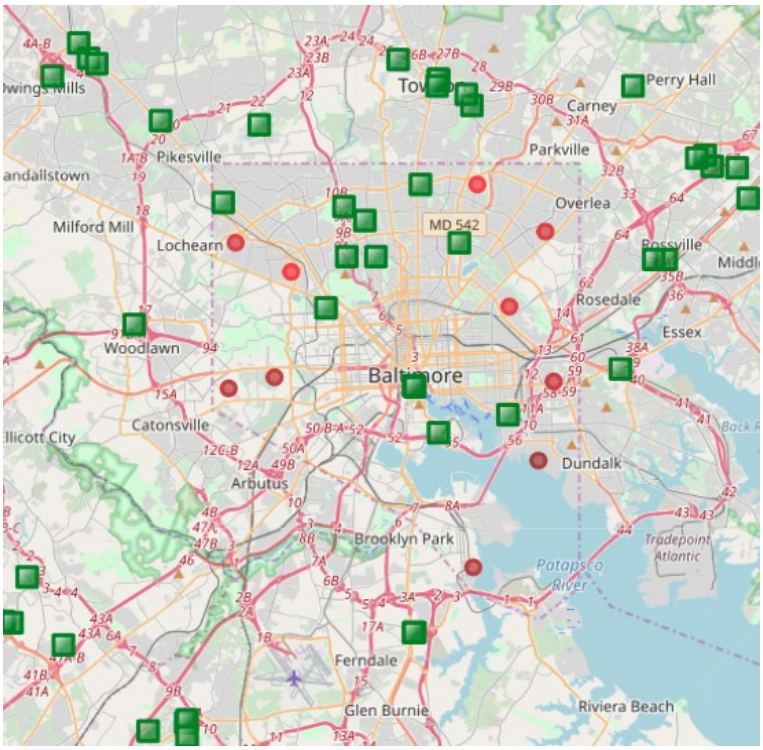
The following histogram shows the distribution of Scores



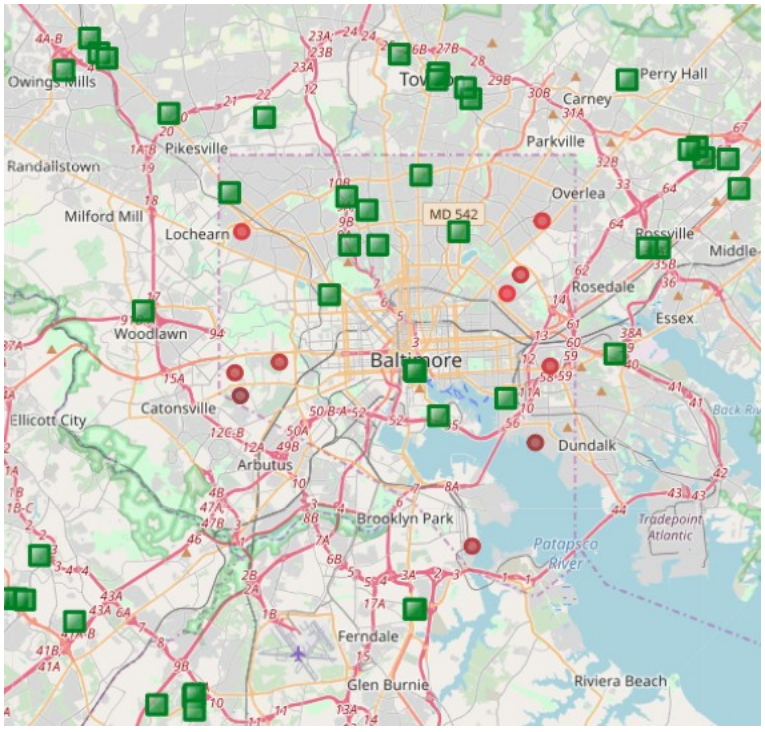
The distribution is compressed at the region of lower Scores which corresponds to the fact that even when the is a low population the Aggregate Nearness metric remains high.

However, the score discriminates to some extent and I now use it to identify the 10 Census Tracts with the lowest (the lower the better) Score.

The result (walking distance 1 km, laziness: 0.01) is shown in the plot below.The next chart shows



the results of the same analysis for walking distance: 1 km and laziness: 0.1.



Result and Discussion

Results for Laziness: 0.01

GEO ID	Score	Neighborhood
24510250500	0.002792	Hawkins Point
24510151100	0.002974	Ashburton
24510260501	0.003023	Bayview

Results for Laziness: 0.1

GEO ID	Score	Neighborhood
24510280403	0.000553	Ten Hills
24510250500	0.000603	Hawkins Point
24510250101	0.000656	Beachfield

The assignment Census Tract – Neighborhood has been done using the map provided be the Baltimore city planning authority: https://planning.baltimorecity.gov/sites/default/files/Neighborhood%20Statistical%20Areas%20with%20Census%20Tracts.pdf

Conclusion and Next Steps

The goal was to develop a model which allows a company to identify areas where building malls is an attractive business opportunity.

I created a model using US census data and data from the geo-data provider Foursquare.

The model produces results which are in line with intuition from visual inspection. However, before the model is actually deployed I suggest to perform on-site visits or inspection through Google Earth. I addition the following steps should be taken.

The results show some dependence on model parameters which needs further inspection. In the current model the dependency on Walking Distance and Laziness needs to be further investigated.

Currently the model neglects the population density in adjacent Census tracts. This should be added.

Furthermore we need to include additional census data such as income, gender distribution, and actual space available into our analysis. This should improve model quality significantly.