Opening a new coffee shop

Introduction

In this project I will be describing how a hypothetical investor who is looking to enter restaurant industry in Austin. I am going to propose a model that will find a desirable location for his new business using Machine Learning and Foursquare API. I picked Austin because it is one of the fastest growing city (+22% of population within the last decade) in the United States and I am familiar with its neighborhoods.

Assume the investor wants to open a restaurant business. For this project I chose a Coffee shop. I am going to omit most intricacies of the business plan except for location. Some of the main location properties are:

- Human traffic density
- Demographics
- Income
- Local laws
- Competition

There are services (https://www.safegraph.com/) that provide almost real time information on how many people are in a certain area, I contacted them and asked for sample data to use for my project, however the declined. Their services start at $10K/year.

For the project I will use population density data that is available for free. I will focus on competition, population density and demographics in certain neighborhoods. The data on population by zipcode can be obtained from here:

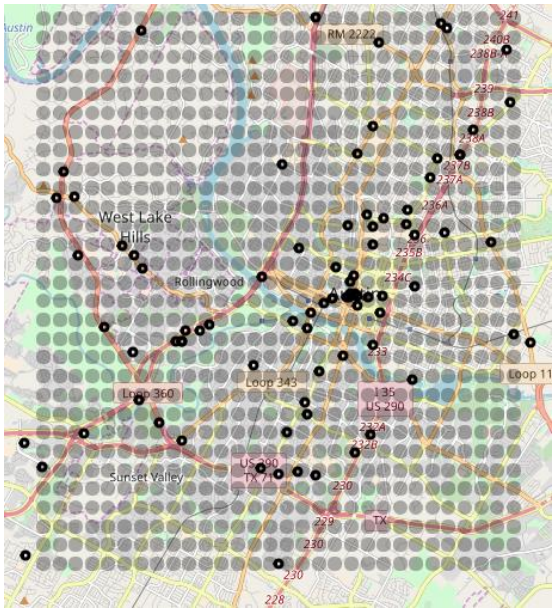
http://zipatlas.com/us/tx/austin/zip-code-comparison/population-density.htm

Information on median age and median income by zipcode can be queried from this website:

http://zipwho.com/

Data

To find all the venues that match the criteria of “coffee” I divided the area of interest into a rectangular grid (nodes are indicated by gray semi-opaque circles). The distance between adjacent circles is equal to 500m.



FourSquare ‘coffee’ search query with radius of 1 km was executed for each node and the following parameters were saved to the data frame: id, name, latitude, longitude. The search returned 86 unique venues that are shown on the map as black dots.

id	name	location.lat	location.lng
4aa51b71f964a520674720e3	Starbucks	30.207837	-97.815372
584978d8e03e5771646253d8	Summer Moon Coffee Trailer	30.205375	-97.806272
579b791e498e9bd835c09ce6	Machine Head Coffee	30.210956	-97.771058
4f32165119836c91c7b47a90	Ag Horn Coffee Service Company	30.215841	-97.763750
58c549ff13c2236f0d0ffd8d	Ruta Maya Coffee	30.203389	-97.704302

Each node was assigned a zipcode based on how close it is to the center of zipcode area. This is a very crude approximation, but I could not find a free service that would provide me with this information. The easiest way to obtain this info is by using Google API reverse geolocation. For each zipcode, median age and income data was queried from zipwho.com.

The following data frame was compiled for analysis. It consists of 900 rows where each row represents a node and contains the following information: geo coordinates, zip code, population density, median income, median age and distance to the closest competing venue in meters. 'f' is the optimization function that will be discussed in the methodology section.

	lat	lng	zip	density	income	age	minD	f
0	30.2041	-97.8332	78749	2795	68244	32.2	1960.59	3102.44
1	30.2041	-97.8285	78749	2795	68244	32.2	1517.25	3102.27
2	30.2041	-97.8239	78745	4063	43458	31.3	1028.88	3464.65
3	30.2041	-97.8192	78745	4063	43458	31.3	591.809	2142.61
4	30.2041	-97.8146	78745	4063	43458	31.3	420.658	1264.05

Methodology

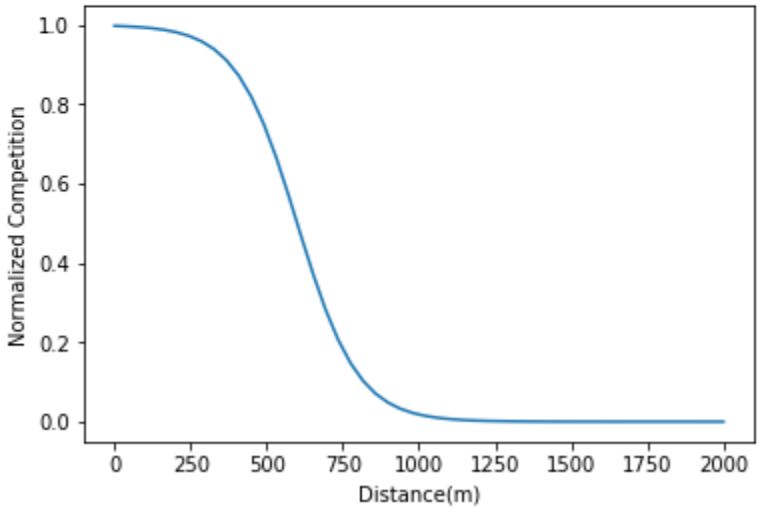
In order to predict the number of customers that visit our café based of predictors we acquired in the data section, I am going to ask a few questions.

Question one: How does the competition in the vicinity affect the bubble tea sales?

First we need to define the scale of distance. I will assume anything that is further than a 10 minute walk is far enough for the customer not to consider the competition. Average walking speed of an adult is 1.4 m/s, which translates into roughly 900 meters per 10 minutes. So our function should return 0 at 900 meters and above and approximately 1 within 200 meters (~3 minute walk). Sigmoid (*Competition_F*) will do well:

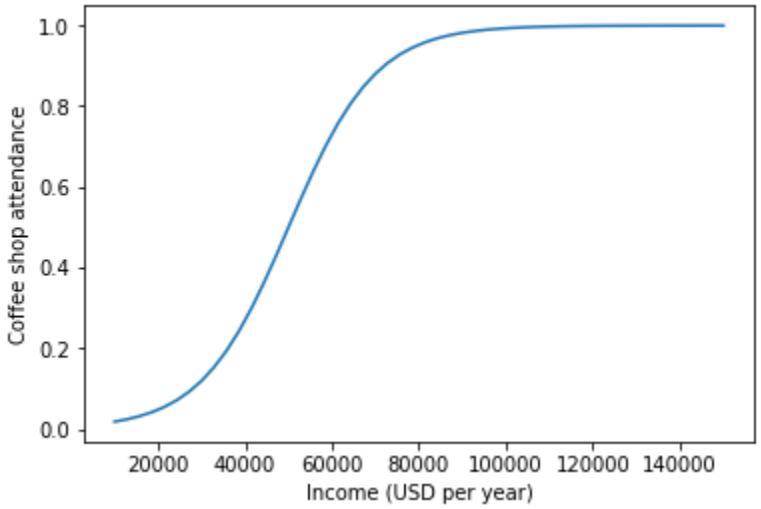
```
x=np.linspace(0,2000,50)
y=1/(1+np.exp((x-600)/100))
plt.plot(x,y)
plt.xlabel('Distance(m)')
plt.ylabel('Normalized Competition')
```

```
Text(0,0.5,'Normalized Competition')
```



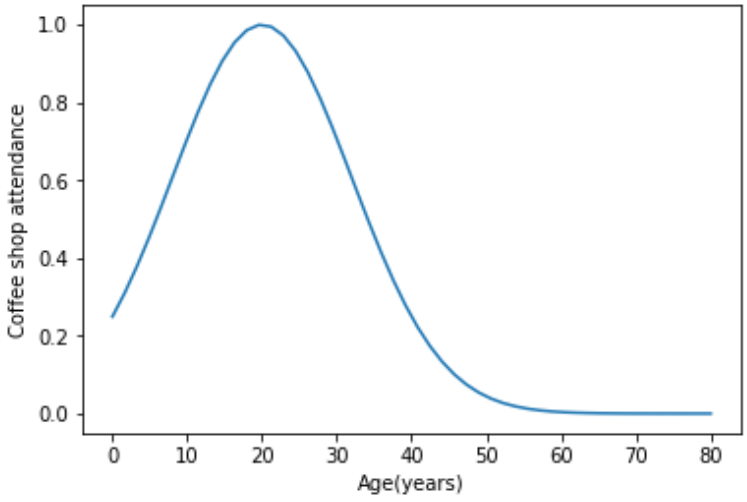
Question two: How does income in the area affect the sales?

People who make less than $20000 a year are unlikely to spend 5 dollars on a fancy coffee, while those who make $ 100000 will probably not overindulge themselves in caffeine drinks. So it is natural to assume that sales saturate at around $60000 median income. Once again sigmoid (*Income_F*) will do the trick:



Question three: How does the median age in the area affect the sales?

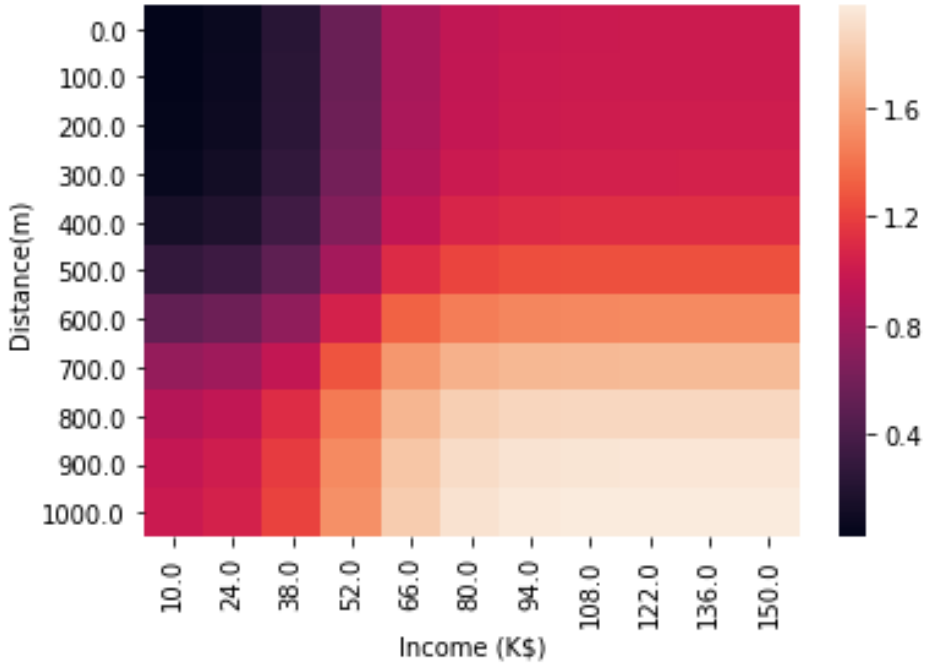
Unfortunately I couldn't find any information on coffee shop demographics, so I will make some assumptions from what I see. I mostly see people age 20-40 at coffee shops around the world, so there we go. (*Age_F*)



Finally, I constructed the optimization function as follows:

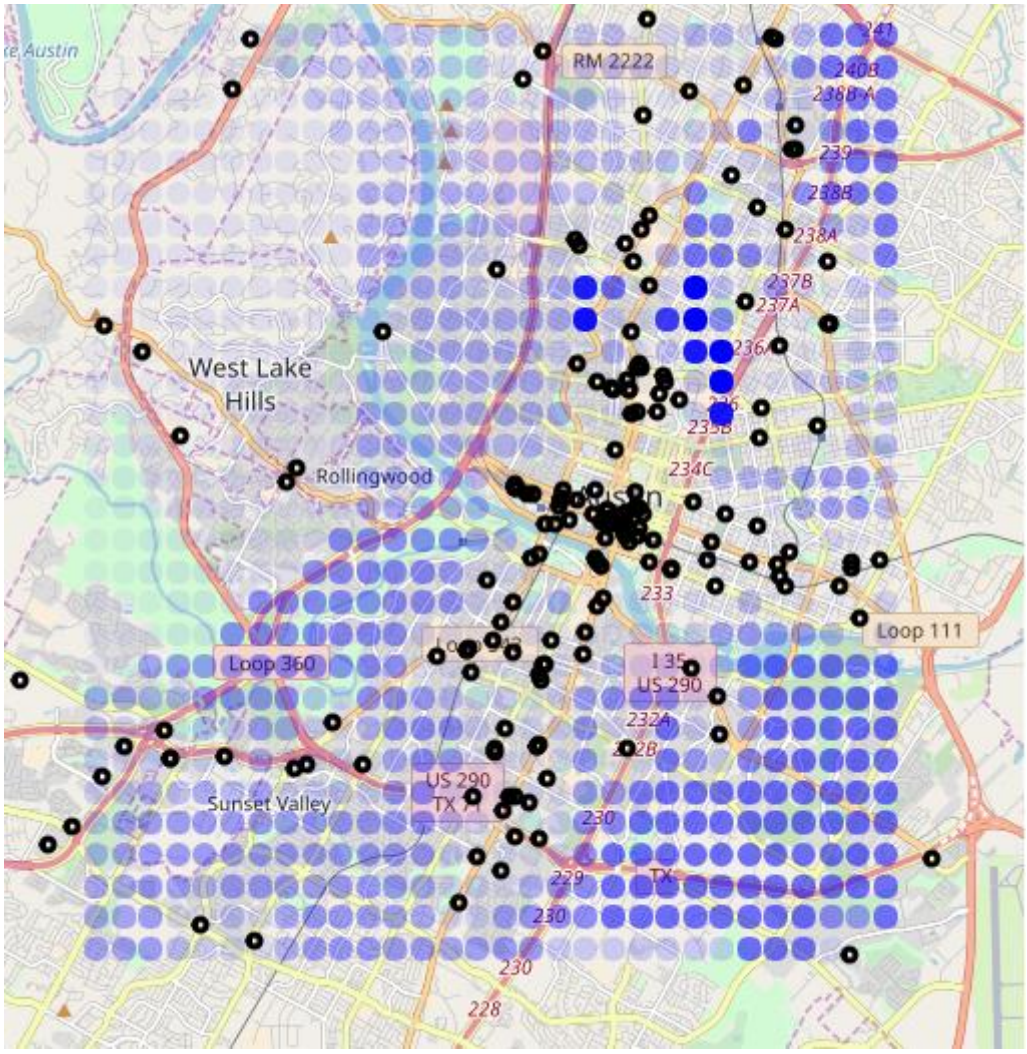
$$F=Density*Age\_F*(1+Income\_F-Competition\_F)$$

The number of customers should be proportional to the total population flow within a certain demographic in the area. Thus the density and age functions come in as multipliers. The main part of the function **(1+Income_F-Competition_F)** can be described by the heat map. Function range is [0,2].



Results

The optimization function was calculated for each node and its value was color coded onto the map. Dark circles represent areas where the value is high, while semi-transparent circles represent areas where F is low.



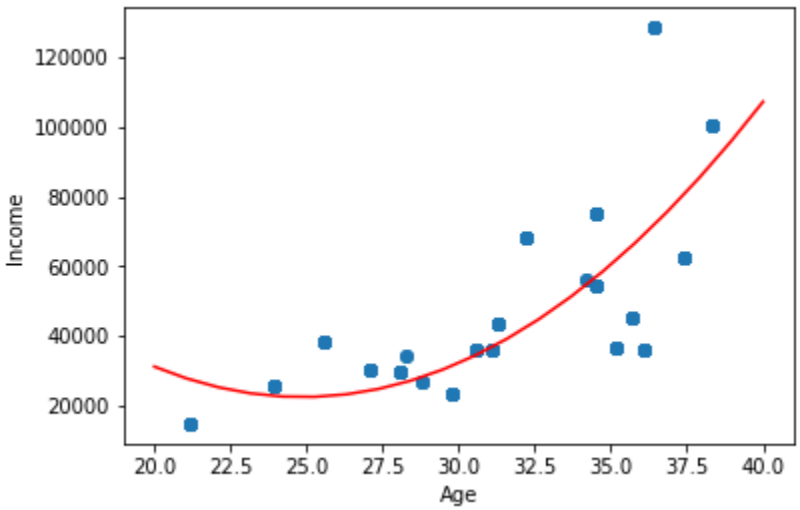
The results seem reasonable on a grand scale. The West Lake Hills area mostly consists of private residential houses with low population density and a lot of vegetation. The brightest spots on the map show North Campus area that is populated with students of UT Austin however it is devoid of coffee shop within the area from 26th to 43rd streets. This is the best location for a new coffee shop.

Discussion

In the discussion section I will focus on ways this model can be improved. I will start with the geography. Some areas that were identified as desirable locations for new a coffee shop are actually inaccessible by car, or simply are bodies of water. This is something that can be fixed by introducing new predictors, such as car and or population traffic information. There are heat maps that show real time concentration of smartphone users (SnapChat uses such maps that are generated from their own app).

Another useful piece of information would be purchase history. Banks collect data about credit card purchase trends for their customers. Convenience stores like 7/11 have data on product purchases.

As the number of predictors increases we will need to start thinking about computational limits. It is possible that some of the predictors are correlated and it makes sense to sacrifice some of them to meet computational limits. For example, it seems natural that age and education is correlated with income. Since I didn't have any information on education, I plotted income as a function as age and tried to fit it with a 2nd degree polynomial.



Conclusion

In this project I proposed an optimization model for picking a location for a new coffee shop in the city of Austin. I identified the parameters that are used as predictors and explained why they are important. I collected the data from free sources and provided references to them. I constructed an optimization function that returns feasibility of opening a new coffee shop at a certain location based on its geo coordinates. Finally, I talked about how this model can me improved and what paid data providers can be used for improvement.