IBM CAPSTONE PROJECT FOR DATA SCIENCE

Análise de agrupamento de lojas de roupas da cidade do Recife através de seus locais adjacentes

Nicólly Lira de Albuquerque

INTRODUÇÃO

Este relatório é referente às explicações e resultados do projeto final do Coursera Aplied Data Science Capstone.

O objetivo dessa análise é verificar se é possível categorizar em grupos lojas de roupas, na cidade de Recife - PE, através da similaridade de estabelecimentos das redondezas em um raio de 200 metros. A fim de encontrar um padrão de estabelecimentos comuns para cada grupo, auxiliando como um dos possíveis indicadores para fundamentar a melhor escolha de localidade para a abertura de outras lojas, que possuem similaridade com lojas de algum desses grupos que foram formados.

Para construir o problema, e devidamente respondê-lo, neste projeto utilizaremos com caráter obrigatório de especificações do Coursera, a API do Foursquare e o algoritmo de clusterização k-means.

Clusterização é a separação de dados em grupos de similaridades, realizada por algoritmos de aprendizagem de máquina não supervisionados. Onde, este agrupamento é baseado na distância da distribuição cartesiana de pontos relativos às categorias que foram escolhidas como parâmetros para a esta análise de similaridade.

O k-means é um algoritmo que agrupa os dados por meio de centróides, onde, é necessário se definir um número k de clusters, para cada cluster é atribuído um centróide e cada ponto da distribuição que pertence a um cluster está mais próximo do seu centróide que dos demais.

O Foursquare é um guia de cidades que permite os usuários cadastrarem estabelecimentos, fazer resenhas sobre eles e atribuir notas. Essas informações ficam disponíveis para consulta de qualquer usuário pelo site e aplicativo do Foursquare e pela API. O site é mais utilizado por usuários comuns para fazer buscas diversas sobre localidades, por possuir uma interface mais amigável e interativa para acessar esses dados. Mas esses mesmos dados podem ser buscados por programadores através da sua API, que foi a forma utilizada para construir este projeto.

METODOLOGIA

1. Da coleta de dados

Na realização deste estudo foram utilizadas duas fontes principais de dados para a coleta: a API do Foursquare e o site da prefeitura da cidade do Recife.

Da API do Foursquare foram extraídos dados de 50 lojas de roupas, num raio de 6.000 metros a partir das coordenadas do centro da cidade do Recife. Para esta primeira chamada na API foi utilizado o endpoint ‘venues’, passando além dos parâmetros de autenticação e localidade, o parâmetro de id da categoria lojas de roupas, fornecido pelo próprio Foursquare em sua página de apresentação e explicação de categorias.

Dessa chamada, os dados salvos foram salvos ID da loja, nome da loja, latitude e longitude.

Antes de partir para a segunda chamada no Foursquare, vamos acessar a API Nominatim e usar a latitude e longitude para resgatar a informação dos bairros aos quais essas lojas pertencem, pois, esse endpoint não nos retorna esta informação em seu arquivo json.

Na segunda chamada da API do Foursquare, utilizaremos o endpoint ‘explore’, para buscar, para cada loja de roupas encontrada acima, no máximo 200 estabelecimentos e facilidades num raio de 200 metros de distância. Armazenamos, desta vez, o nome desses estabelecimentos e a qual tipo/categoria de classificação do Foursquare eles pertencem.

Neste ponto, encontramos o nosso primeiro empecilho para o nosso objetivo de agrupar as lojas de roupas a partir da categoria dos estabelecimentos de seus arredores: o Foursquare possui um sistema de classificação hierárquica, onde existem classes, dentro de subclasses, dentro de sub-subclasses e assim sucessivamente. Diante disto, a API nos retorna a menor classe possível de categorização daquele estabelecimento, criando assim uma quantidade absurda de classes diferentes nos arredores das lojas, tornando impossível a identificação de um padrão.

Para resolver este problema, iremos recorrer mais uma vez a outra chamada na API do Foursquare, dessa vez com seu endpoint 'categories', para acessar uma ontologia disponível que determina determinar as categorias maiores de entendimento às quais cada subcategoria pertence. Criamos assim um dicionário ontológico das categorias do Foursquare que tornou possível a simples substituição das categorias menores por suas respectivas categorias maiores.

Agora usando o site da prefeitura do Recife, iremos realizar um webcrapping página por página, a partir de uma lista já criada com os nomes únicos de bairros das nossas lojas de roupas, e salvar para cada bairro o seu "Valor do Rendimento Nominal Médio Mensal dos Domicílios".

Esta métrica socioeconômica do IBGE se refere à renda média calculada com a soma de qualquer entrada de dinheiro no domicílio, não apenas os valores documentados formalmente. No site da prefeitura não existe informações se esta renda nominal é ou não per capta, ou seja, a soma da renda total dividida pelo número de moradores da residência, mas em uma breve pesquisa sobre métricas, pudemos confirmar que se trata sim de um valor per capto.

2. Do pré-processamento dos dados

Como já sabemos, iremos realizar o agrupamento a partir da similaridade de categorias dos estabelecimentos dos entornos da nossa amostra de lojas de roupas da cidade do Recife. Essas categorias são palavras que designam a qual classe de estabelecimento esses locais pertencem, das dez classes gerais fornecidas pelo Foursquare, sete foram encontradas neste caso.

Para contar quantas vezes cada uma dessas sete categorias aparece para uma respectiva loja de roupas, vamos isolar a coluna de categorias e representá-la de forma binária (one hot encoding), transformando os valores únicos da coluna em categorias.

De forma que para cada categoria existente para uma entrada, sua coluna receberá o valor 1 e para cada categoria inexistente, sua coluna receberá o valor 0.

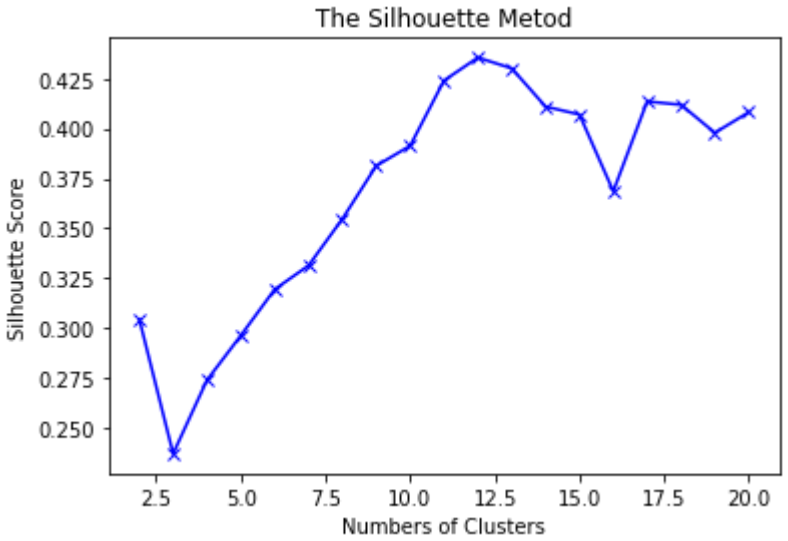
Feito isto, se reposiciona as informações de ID das lojas de roupas para suas respectivas linhas com colunas binárias, que representam cada estabelecimento próximo desse ID.
E por fim, agrupamos os IDs somando os valores das colunas binárias de categoria, encontrando assim a quantidade de categorias próximas a cada uma de nossas lojas de roupas.

O último passo da nossa fase de pré-processamento é o escalonamento, ou seja, iremos colocar nossos dados em uma mesma escala de valores, a fim de melhorar as métricas do algoritmo. Para este caso, iremos utilizar o escalonamento por Normalização, colocando nossos valores em uma escala entre -1 e 1.

3. Da clusterização

O k-means é o algoritmo de clusterização que iremos utilizar e como já mencionado acima, ele realiza o agrupamento por meio da distância de centróides. Porém, indicar como parâmetro o número errado de clusters, pode tornar nossa separação não supervisionada por similaridade de variância (clusterização) pouco precisa, seja por ter grupos demais que generalizam pouco, ou grupos de menos que não conseguem dar visibilidade a outras nuances que podem ser importantes na análise.

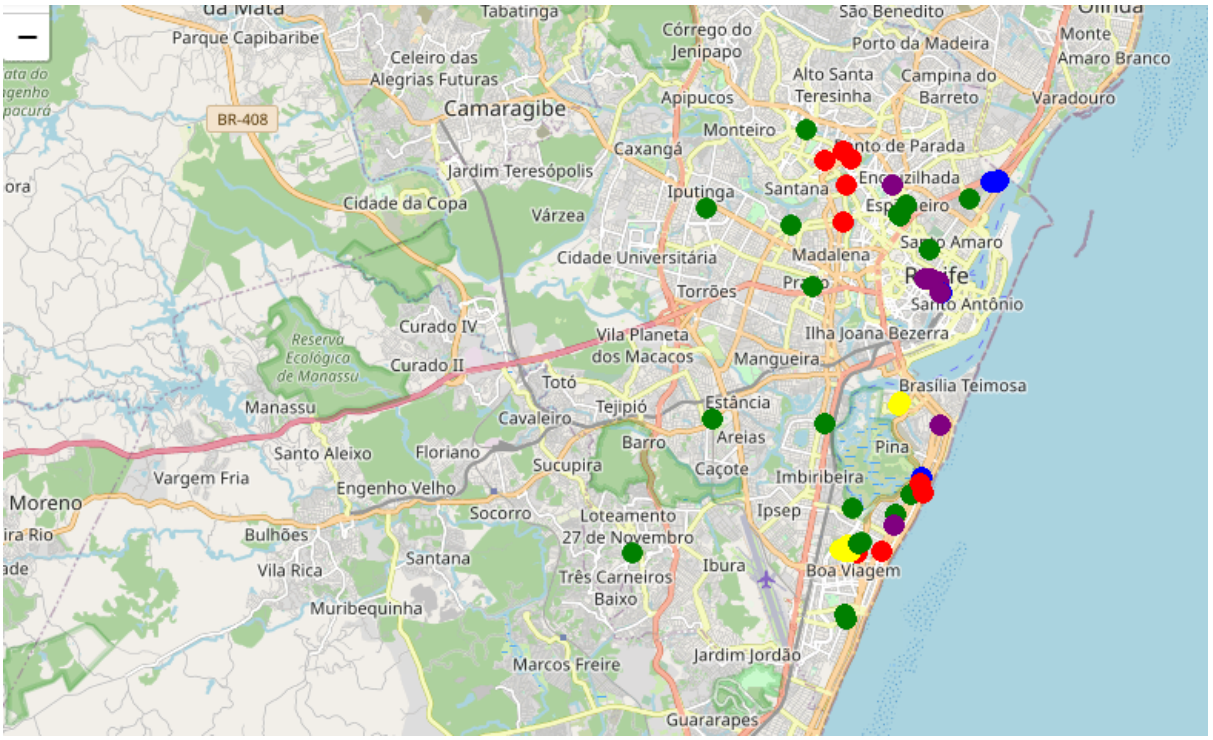
Para tentar resolver este problema, encontrando uma quantidade ótima de clusters, foram utilizados dois métodos Heurísticos: O método de elbow, que deu resultados muito inconclusivos, e o método da silhueta, que nos deu resultados um pouco menos nebulosos.



Apesar deste gráfico nos apontar mais ou menos dez ou onze clusters, este é uma quantidade muito grande de grupos em relação a quantidade dos nossos dados.

Desta forma, para evitar um overfit, assumiu-se um número 5 de clusters, porque é o melhor índice de silhueta antes do score subir indefinidamente e é o número de clusters que distribui mais igualitariamente as quantidade de lojas entre os clusters.

ENTENDIMENTO DOS CLUSTERS



Cluster 4 (Roxo): Possui oito lojas, onde a grande maioria está localizada na região central da cidade do Recife, no bairro da Boa Vista. Porém ainda possui duas lojas localizadas mais ao sul da cidade e apenas uma na zona norte no bairro dos Aflitos. A renda nominal média mensal por domicílio, dos bairros em que as lojas desse cluster estão localizadas, varia bastante, iniciando em mil e chegando a até sete mil reais.
Também é o cluster que possui o maior percentual de 'Nightlife Spot' em relação aos outros grupos.

	Category	amount	percentage
0	Food	116	52.25
1	Shop & Service	65	29.28
2	Nightlife Spot	22	9.91
3	Arts & Entertainment	11	4.95
4	Outdoors & Recreation	5	2.25
5	Travel & Transport	2	0.90
6	Professional & Other Places	1	0.45

Cluster 3 (Amarelo): Possui seis lojas de roupas, todas dentro de shoppings Centers, localizadas na zona sul do Recife, nos bairros de Boa Viagem e Pina, que possuem um valor do rendimento nominal médio mensal dos domicílios de sete e dois mil reais respectivamente. Outra similaridade entre a região é o fato de ambas serem regiões litorais, com praias inclusive vizinhas.
Esses shoppings também possuem um padrão entre si: eles são os maiores shoppings da cidade.
Com esse grupo, podemos supor que a disposição das lojas de roupas em shoppings dessa área da cidade, possuem um padrão de estabelecimentos nos arredores.

	Category	amount	percentage
0	Food	177	54.13
1	Shop & Service	123	37.61
2	Arts & Entertainment	23	7.03
3	Outdoors & Recreation	2	0.61
4	Nightlife Spot	2	0.61

Cluster 2 (Verde): Existem dezenove lojas classificadas como parte desse cluster. Com um espalhamento geográfico muito variado, sem a presença de um padrão identificável nesse aspecto. Porém o valor do rendimento nominal médio mensal dos domicílios possui uma menor variância entre si, com exceção do bairro da Boa Vista, essa métrica socioeconômica neste cluster varia entre mil e dois mil reais. Com isto, é

também o cluster com menor rendimento nominal médio mensal, dando indícios de que este é o padrão de arredores de lojas de roupas para bairros do Recife em que os moradores possuem, em média, um menor poder aquisitivo.

	Category	amount	percentage
0	Food	103	54.79
1	Shop & Service	50	26.60
2	Arts & Entertainment	10	5.32
3	Outdoors & Recreation	9	4.79
4	Nightlife Spot	9	4.79
5	Travel & Transport	7	3.72

Cluster 1 (Azul): Este grupo possui seis lojas, sendo sua maior concentração na região do centro do Recife, no bairro de Santo Amaro, dentro do shopping Tacaruna, com apenas duas lojas localizadas na região sul do Recife nos bairros do Pina e Boa Viagem.
O valor do rendimento nominal médio mensal dos domicílios desses dois últimos bairros é de dois e sete mil, respectivamente. O bairro de Santo Amaro fica em torno de mil, mas existem relativizações que podem ser feitas pelo fato de todas essas lojas estarem dentro de um shopping.
É importante ressaltar também, que apesar de serem lojas em um shopping, não foram agrupadas ao cluster 3. A hipótese mais provável é o indicativo de diferentes categorias de estabelecimentos disponíveis nos shoppings de acordo com a região, já que o cluster 3 possui uma porcentagem muito maior da categoria 'Arts & Entertainment' e levemente maior da categoria 'Outdoors & Recreation'. Mas uma outra hipótese é apenas uma disposição diferente das lojas de roupas entre esses shoppings.

	Category	amount	percentage
0	Food	117	50.87
1	Shop & Service	108	46.96
2	Nightlife Spot	2	0.87
3	Arts & Entertainment	2	0.87
4	Outdoors & Recreation	1	0.43

Cluster 0 (Vermelho) : Possui nove lojas, sendo todas completamente concentradas na zona sul e norte da cidade, não possuindo nenhuma loja no centro e em outras regiões.
Nesse cluster, além das duas primeiras categorias presentes em maior porcentagem em todos os clusters, se destaca os estabelecimentos de "Travel & Transport", seguido de um também bom índice de "Outdoors & Recreation".

Todos os bairros em que estão localizadas as lojas desse cluster também possuem um alto valor do rendimento nominal médio mensal dos domicílios que variam entre 4 a 11 mil reais, sendo desta forma um cluster de lojas em uma localidade de residentes de classe média e alta.

	Category	amount	percentage
0	Food	80	45.45
1	Shop & Service	65	36.93
2	Outdoors & Recreation	20	11.36
3	Travel & Transport	6	3.41
4	Nightlife Spot	5	2.84

DISCUSSÃO E CONSIDERAÇÕES FINAIS

Antes de nos concentrarmos nas discussões sobre os desdobramentos dos resultados desse estudo, eu gostaria de pontuar duas informações relevantes para a compreensão desse resultado como um todo.

O primeiro ponto é a dificuldade de encontrar dados referentes a bairros da cidade do Recife, as fontes de dados abertos encontradas possuíam dados mais gerais relacionados a cidade como um todo, com entradas sem divisões ou variáveis de identificação por bairro. Diante disso, realizei um webscrapping do site da prefeitura do Recife, mas ainda é muito distante de uma base de dados robusta.

O segundo e último é que infelizmente, o Foursquare foi gradativamente perdendo popularidade no Brasil, e nos dias atuais existem poucos registros de estabelecimentos ainda cadastrados e usuários alimentando a plataforma.

Deste modo, uma alternativa que pode responder a questão proposta neste trabalho, é a coleta de mais dados, tanto de lojas de roupas, quanto outras fontes de dados para fins comparação e cruzamento com os dados principais das lojas.

Dito dito, vamos nos debruçar nos resultados. Com uma breve observação das porcentagens de categorias por cluster, é possível determinar que todos eles possuem majoritariamente estabelecimentos da categoria "Food" e "Shop & Service". Indicando que todas as lojas de roupas,segundo dados e técnicas utilizadas nesse estudo, independente da região da cidade do Recife,

vão contar com uma significativa quantidade de estabelecimentos deste tipo ao redor, sendo mais de 50% do total de estabelecimentos próximos para todos os clusters.

É possível vislumbrar leves indícios de possíveis padrões em alguns dos clusters aqui agrupados, como exemplo, clusters que só aparecem em regiões mais valorizadas da cidade ou apenas dentro de shoppings, como foi o caso de alguns. Mas, com base na quantidade e qualidade de dados disponíveis aqui, para a aplicação desta metodologia em específica, os argumentos de similaridades que surgem ainda são muito frágeis. Podendo-se concluir que, com o algoritmo de clusterização k-means e os dados utilizados, não é possível identificar um padrão de arredores de lojas de roupas para cada cluster.

Das ideias possíveis para a otimização da resposta para este problema, uma delas é utilizar a informação de ticket médio dos estabelecimentos dos arredores das lojas de roupas, ao invés dos dados da prefeitura do Recife por bairros. Pois, a média não é um bom valor para generalizar a renda neste problema, apesar de à primeira vista parecer específico o suficiente, grande parte dos bairros não possuem uma distribuição de renda homogênea. Existindo blocos em maior vulnerabilidade econômica num mesmo bairro que possui outros blocos de com bons índices de desenvolvimento econômico.

REFERÊNCIAS

- Site da prefeitura do Recife: http://www2.recife.pe.gov.br/
  - Método Elbow e Silhueta: https://ichi.pro/pt/metodo-da-silhueta-melhor-do-que-o-metodo-do-cotovelo-para-encontrar-aglomerados-ideais-61080390822033
  - Sabores dos bairros da cidade de Nova York: https://towardsdatascience.com/exploring-the-taste-of-nyc-neighborhoods-1a51394049a4
  - Clusterização do metrô de São Paulo: https://medium.com/@felipe.testaa/unsupervised-machine-learning-kmeans-clustering-of-s%C3%A3o-paulo-subway-stations-using-foursquare-c5101727dd85