Análise de dados São Paulo / Rio de Janeiro

- Análise dos bairros do Rio de Janeiro

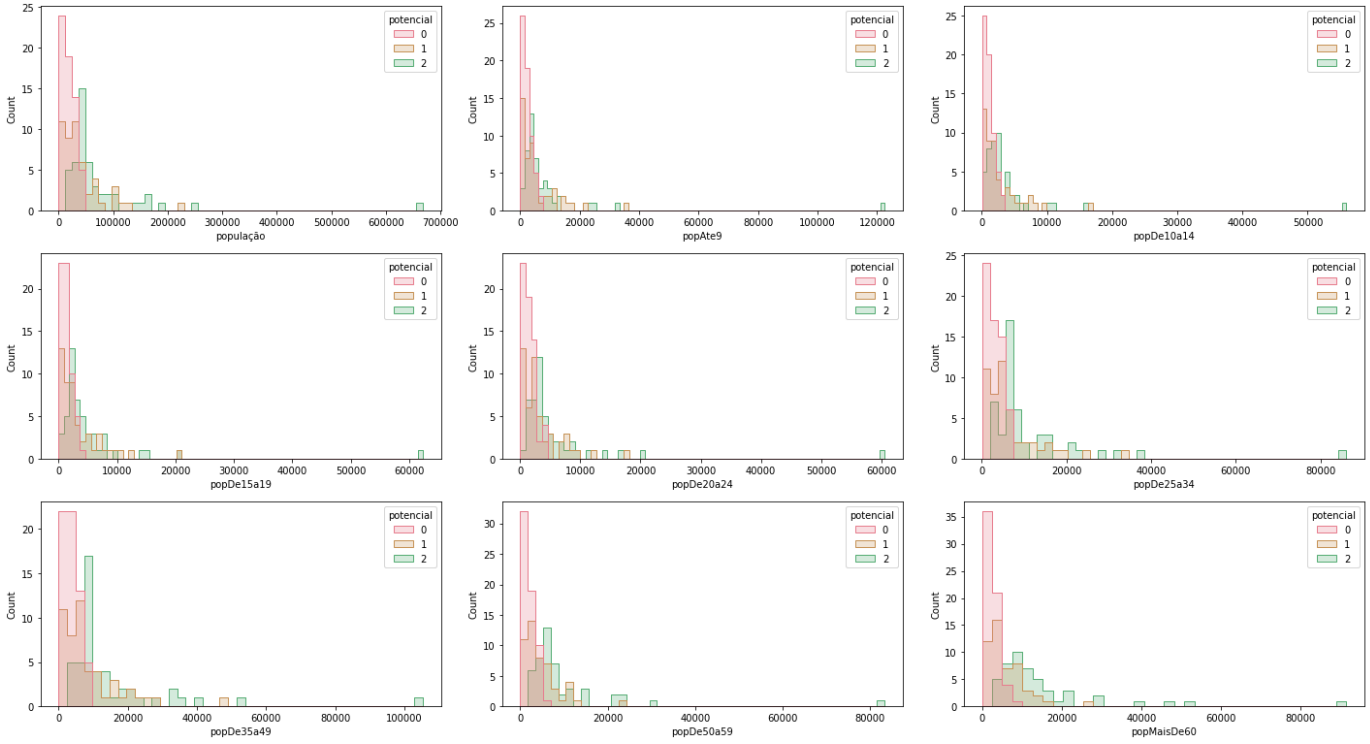
No dado do Rio de Janeiro, em seis bairros, não havia a informação *rendaMedia*. Visando preencher os valores faltantes, olhei as seguintes estatísticas desta coluna:

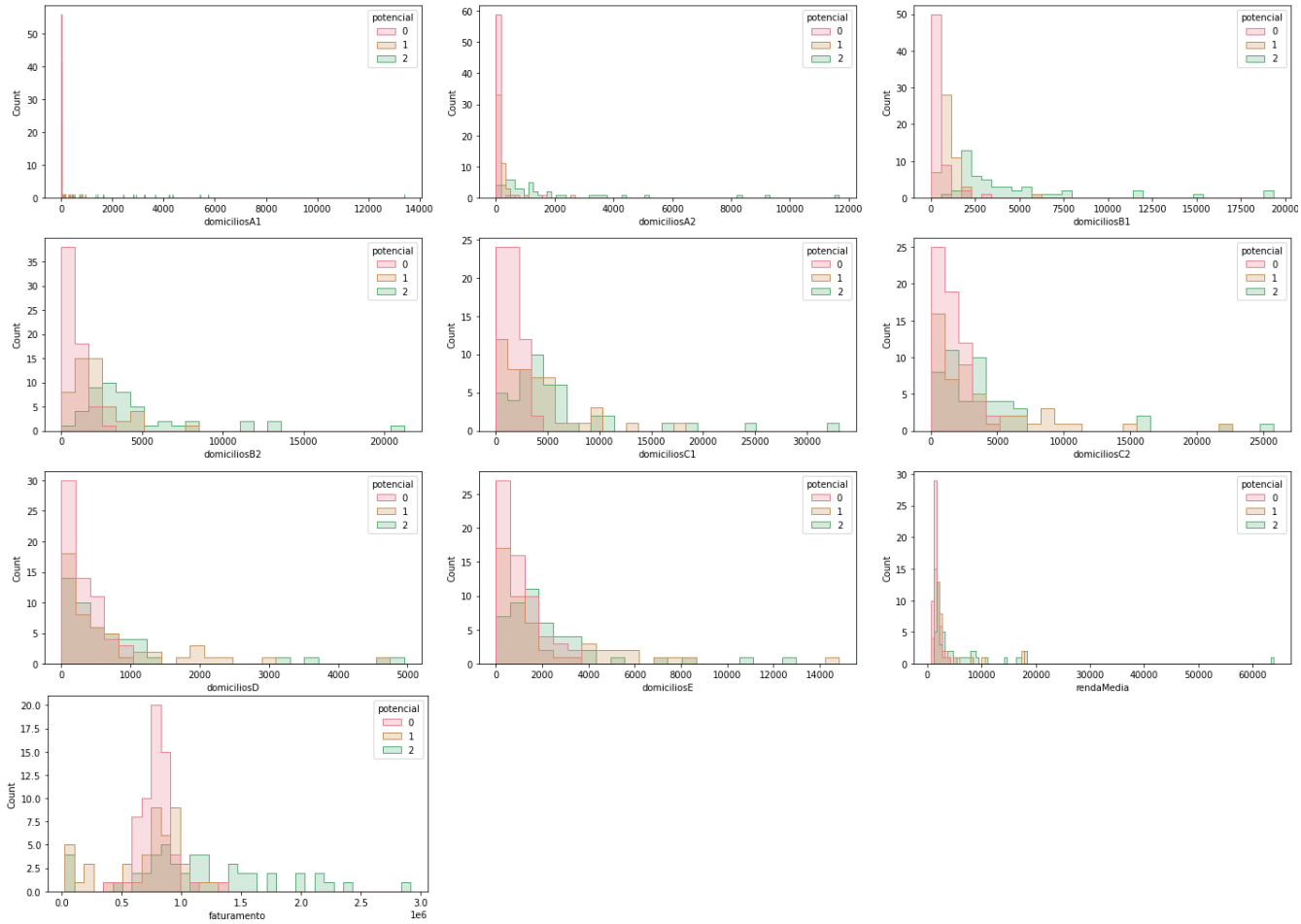
```
count      154.000000
mean      3608.071429
std       6091.865305
min        654.000000
25%       1486.000000
50%       1915.500000
75%       2954.500000
max      63887.000000
```

Legenda: *count* → Quantidade de registros; *mean* → média dos valores; *std* → desvio padrão; *min* → menor valor; *max* → maior valor. O 25%, 50% e 75% correspondem ao primeiro, segundo e terceiro quartil.

Por conta da média ser muito alta, a mediana (50%) correspondia a um valor mais realista para o preenchimento dos campos faltantes.

Em seguida, retirei as colunas *codigo*, *nome*, *cidade* e *estado*, pois não agregavam em nada a análise. Depois de retiradas, transformei os valores da coluna *potencial*, de texto (‘Baixo’, ‘Médio’ e ‘Alto’) para valor numérico (0, 1 e 2, respectivamente), e então gerei os gráficos (histogramas) abaixo para comparar os diferentes potenciais:





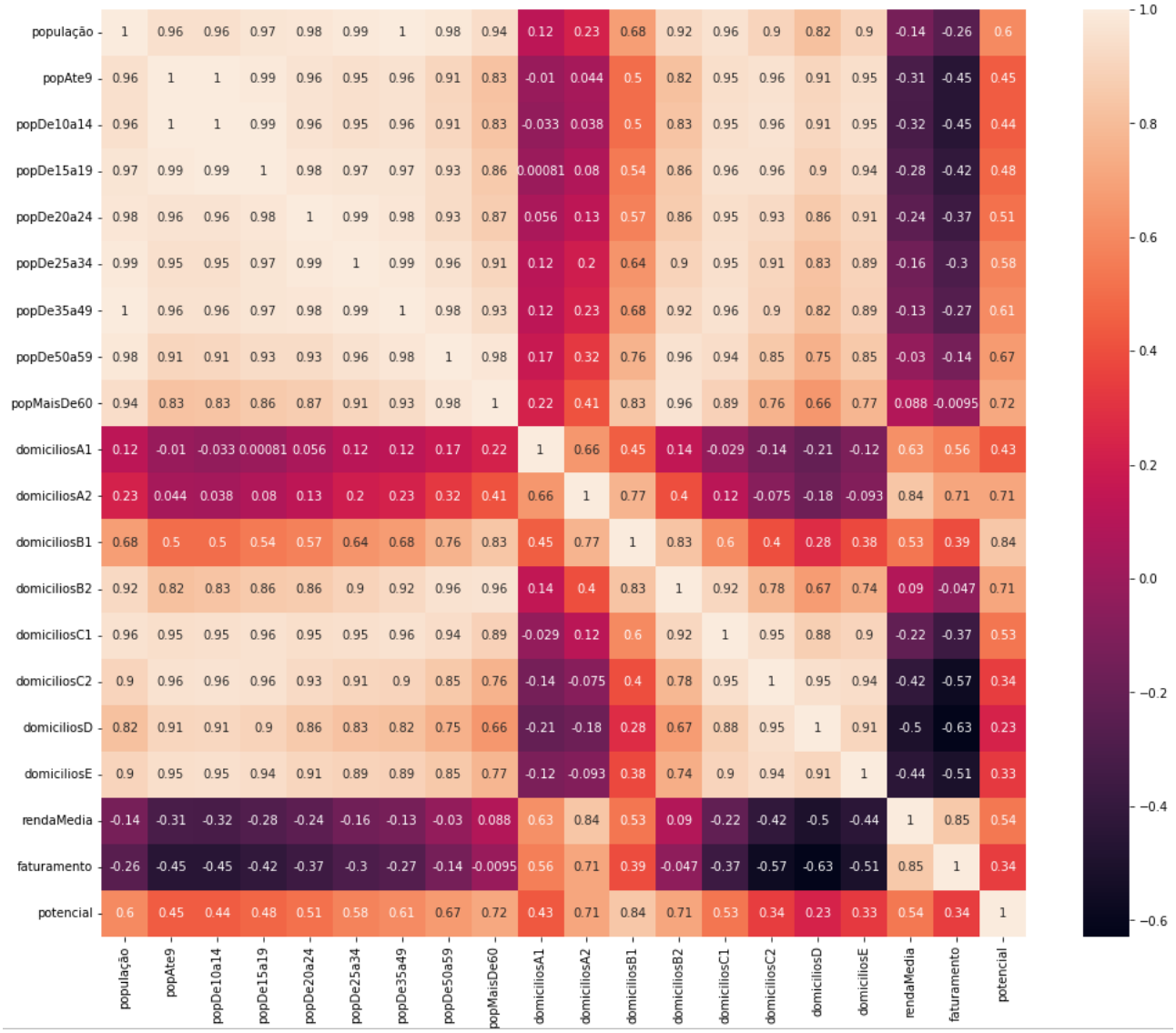
Legenda: O eixo Y indica quantidade, o eixo X indica valores particionados de uma coluna específica. O potencial Baixo (0) é representado pela cor vermelha, o Médio (1) é representado pela cor amarela e o Alto (2), pela cor verde. Vale ressaltar que no gráfico faturamento, o eixo X está em milhões.

Dando uma rápida olhada, já é possível perceber que não há uma fronteira definida entre os tipos de potencial, e há valores discrepantes, no qual serão ignorados nesta análise. Nos primeiros gráficos, de população, é possível ver que para o potencial Baixo, os bairros são pouco populosos, já para o Médio e Alto, existe um equilíbrio maior. Mas ainda, para alguns gráficos é perceptível uma leve tendência dos bairros com potencial Alto para a direita do Médio, indicando ser um pouco mais populoso.

Analisando agora os gráficos referentes a domicílios, é notável também que os bairros de potencial Baixo também estão mais à esquerda, uma possível consequência de menor população. Nos gráficos de domicílios A1, A2, B1 e B2, há uma presença notável do Alto mais à direita, correspondendo com o publico alvo da empresa. Em C1, os potenciais Médio e Alto se equilibram um pouco, e nos potenciais C2, D e E, há uma maior presença de potencial Médio.

No gráfico de renda média, é perceptível uma maior presença do potencial Alto mais à direita, indicando uma maior renda. No gráfico de faturamento, é indicado um maior faturamento em geral para o potencial Alto, mas em relação aos potenciais Baixo e Médio, não há uma diferença clara.

Visando analisar melhor o faturamento e detalhar outras correlações, gerei um gráfico de correlação:



Legenda: Cada quadrado representa uma correlação entre duas colunas, indicadas pelo eixo X e Y. A cor está associada com um índice de correlação (ver barra a direita do gráfico). Partindo da interpretação onde r é o índice indicado em cada quadrado, temos r=0.3 → correlação fraca, r=0.5 → correlação moderada e r=0.7 → correlação forte e r=1 → correlação perfeita. Além disso quando o número for negativo, a correlação é inversa, ou seja, quando um valor tender a aumentar, o outro tende a diminuir.

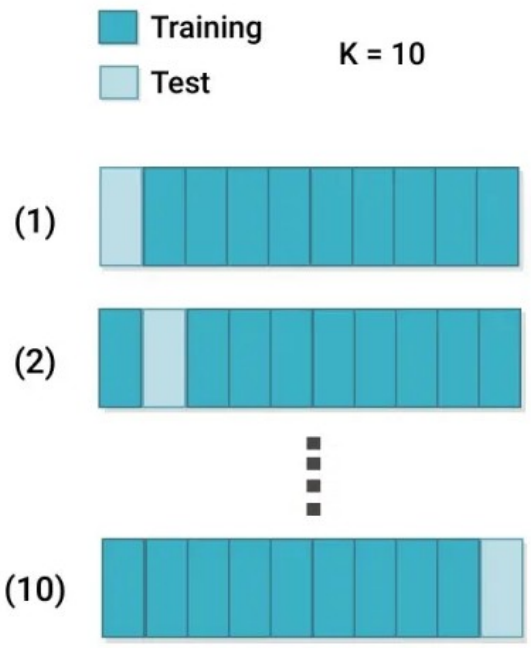
Neste gráfico é possível ver algumas das conclusões dos gráficos anteriores, como impacto da população no potencial (0.6) e o impacto do tipo de residência no potencial, sendo destaque a presença de domicílios A2, B1 e B2 no potencial. Curiosamente, domicílios A1 teve um impacto mais baixo do que domicílios C1. Também é possível verificar melhor o impacto crescente da população no potencial, mostrando que a presença de crianças/jovens em uma família, reduz o interesse na empresa. Além disso, o faturamento não se mostrou estar muito atrelado com o potencial.

Analisando agora o faturamento, é bem claro que a quantidade de pessoas em um bairro tem impacto negativo no faturamento, porém esse impacto negativo reduz com a idade. Em relação ao tipo de domicílio, o faturamento é positivo para domicílios A1, A2 e B1, mostrando que bairros de média/baixa e baixa renda contribuem negativamente para o faturamento, fato comprovado pela correlação entre faturamento e renda média (0.85).

- Treinamento de modelo para estimar o faturamento de um bairro (objetivo 1)

Os atributos utilizados para este treinamento foram todos os apresentados nas análises, retirando, logicamente, os objetivos das predições (faturamento e potencial).

Após normalizar o dado, fiz uma busca por hiperparâmetros utilizando os algoritmos Random Forest, Ada Boost, Gradient Boosting e a Regressão Linear. Utilizei validação cruzada com 10-folds. A imagem abaixo ilustra o processo:



Legenda: cada retângulo grande representa o dado completo. Cada retângulo pequeno representa uma fração do dado completo.

Fiz a média de algumas métricas e escolhi o melhor modelo baseado no MAE (Mean Absolute Error). Esta métrica indica o quanto o modelo errou em termos absolutos. O modelo que teve melhor desempenho foi o Gradient Boosting.

- Treinamento de modelo para estimar o potencial de um bairro (objetivo 2)

Utilizando o dado normalizado, fiz uma busca por hiperparâmetros utilizando os algoritmos Random Forest, Ada Boost e Gradient Boosting. Utilizei validação cruzada com 10-folds. O modelo que teve melhor desempenho, com base no F1-score, também foi o Gradient Boosting.

- Predizendo o dado de São Paulo

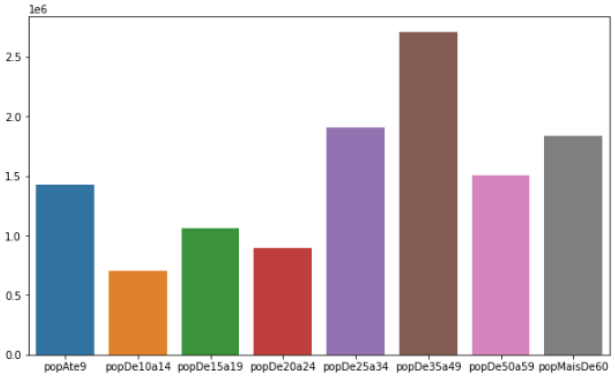
No dado de São Paulo havia três bairros com informações completamente zeradas, então decidi retirá-las. Após isso, criei um arquivo apenas com os bairros de São Paulo, “final.csv”, com as colunas *faturamento* e *potencial* preenchidas a partir das predições dos dois modelos. Concluindo então os objetivos 1 e 2.

- Segmentação dos bairros de São Paulo (objetivo 3)

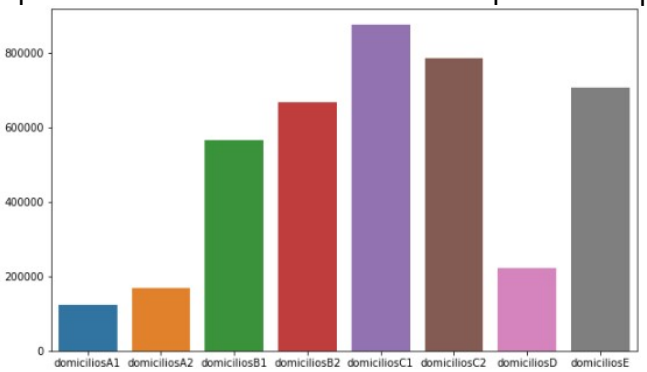
Na etapa de segmentação, utilizei apenas as colunas referentes à faixa etária e aos tipos de domicílios.

Para segmentar, usei o algoritmo K-Means. Como primeiro passo, apliquei o método do cotovelo e encontrei 4 clusters como sendo a divisão ideal. Antes de mostrá-los, os gráficos a seguir contribuíram para uma melhor interpretação dela. São eles um gráfico mostrando a

distribuição etária dos bairros da cidade e a distribuição dos tipos de domicílios.



Legenda: O eixo X corresponde às faixas etárias e o eixo Y representa a população de cada um.



Legenda: O eixo X corresponde os tipos de domicílios e o eixo Y representa a quantidade de cada um.

Observa-se que as barras em ambos os gráficos não são iguais, sendo maior a faixa 25-39 no gráfico de população e domicílios de renda C1 no gráfico de domicílios. Com isso em mente,

A tabela abaixo mostra os centroides de cada atributo para os clusters, este centroide é interpretado como valor médio.

	popAte9	popDe10a14	popDe15a19	popDe20a24	popDe25a34	popDe35a49	popDe50a59	popMaisDe60	domiciliosA1	domiciliosA2	domiciliosB1	domiciliosB2	domiciliosC1	domiciliosC2	domiciliosD	domiciliosE
0	1443.025210	671.302521	1059.008403	971.630252	2184.478992	3125.092437	1822.831933	2597.033613	366.285714	393.655462	975.890756	812.428571	885.991597	691.478992	173.571429	735.420168
1	10930.326531	5476.428571	8097.306122	6429.632653	12998.959184	18349.142857	9529.163265	10143.775510	22.326531	217.224490	2015.612245	4090.020408	6410.102041	6301.346939	1910.775510	5372.510204
2	4137.734513	1981.769912	3124.053097	2832.840708	6381.530973	9162.707965	5515.787611	7530.557522	691.150442	951.097345	2821.185841	2582.752212	2786.000000	2189.362832	544.646018	2163.327434
3	21240.416667	10621.750000	15449.750000	11937.916667	24031.166667	33197.500000	16406.583333	14719.500000	0.666667	212.583333	2674.250000	6423.166667	11750.333333	12336.416667	3957.166667	9389.916667

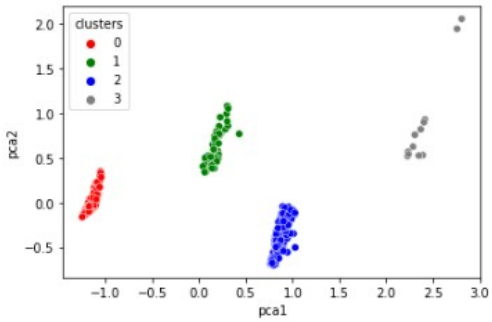
No cluster 0 há menos população que os outros e há um equilíbrio de tipos de renda com uma leve predominância para classe média (B1 e B2).

No cluster 1 há uma grande população para todas as faixas etárias e há comparativamente poucos domicílios de renda mais alta (A1 e A2).

No cluster 2, os bairros possuem proporcionalmente uma maior população com 25+ anos em relação às outras faixas etárias e aos outros clusters. Além disso, possui domicílios de com maior renda (A1, A2, B1 e B2). Portanto é onde ocorreria a maior presença de público-alvo.

No cluster 3 temos bairros populosos, mais de 10k para cada faixa etária. Havendo grande predominância presença de população de renda mais baixa (C1, C2, D e E) e comparativamente com pouquíssimas residências de classe alta (A1 e A2).

O gráfico seguinte apliquei o algoritmo de redução de dimensionalidade, o PCA, utilizei para conferir se a divisão dos clusters estava realmente consistente, o que se mostrou verdade, pois eles estão bem separados.



Legenda: Cada círculo representa um bairro e cada cor corresponde a um cluster diferente.

Para concluir, comparei os clusters com resultados das predições.

cluster	potencial	quantidade
0	0	40
	1	37
	2	42
1	1	34
	2	18
2	0	6
	1	18
	2	86
3	1	10
	2	2

cluster	faturamento
0	1018900
1	384530
2	1100024
3	113941

De acordo com a análise da clusterização, o cluster mais adequado a se construir uma franquia da empresa é o 2, e em segundo lugar, o 0. Estas tabelas validam as análises apresentadas e pode-se dizer que também validam os dados preditos de potencial e faturamento. Ao final, foi gerado um arquivo com a identificação dos clusters, "final_clusterizado.csv". Isso conclui o objetivo 3.

- Dados externos

Procurei um arquivo que mostre os polígonos dos bairros, porém não encontrei. Além de contribuir para uma boa visualização, contribuiria nos resultados caso apresentasse maior proximidade a bairros similares. Com isto conferido, poderia ser usado, caso exista, dado de sensus block ou sensus tracts, e então aplicar o algoritmo de CCA para fazer o agrupamento entre setores similares, pois nem sempre a divisão administrativa é a melhor opção para análise social.