



# Desafio Técnico Cientista de Dados

Jéssica P. S. Cardoso



## Questões

- O que faz os bairros do Rio serem considerados como alto, médio ou baixo potencial?
- Como podemos relacionar ou fazer uma associação com a cidade de São Paulo?
- Quais bairros de São Paulo apresentam uma maior aderência ao público alvo?
- E quais bairros de São Paulo podem ser bons em investir?



## O que possuímos?

- Quantidade de pessoas por faixa etária
  - Até 9 anos
  - de 10 a 14 anos
  - de 15 a 19 anos
  - de 20 a 24 anos
  - de 25 a 34 anos
  - de 35 a 49 anos
  - de 50 a 59 anos
  - 60 anos ou mais
- Renda média de cada bairro
- **Faturamento dos bairros do Rio**
- **Potencial dos bairros do Rio**
- Quantidade de domicílios de acordo com a classe social
- Nome do bairro
- Cidade onde está localizado o bairro
- Estado onde está localizado o bairro



## O que não temos?

- Faturamento dos bairros de São Paulo
- Potencial dos bairros do São Paulo
- Renda média por classe
- Tipo de estabelecimento alimentício  
(Restaurante, Lanchonete, Mercado etc)
- Quantidade de concorrentes na região
- Distância para os fornecedores
- Distância média entre o  
estabelecimento e seu público alvo



## Visão Geral

A maioria dos bairros da cidade do Rio de Janeiro, cerca de 62%, é categorizada com o potencial de consumo baixo.

Há 296 bairros registrados para o município de São Paulo.

	nome	cidade	estado	potencial
count	456	456	456	160
unique	447	2	2	3
top	Grajaú	São Paulo	SP	Baixo
freq	2	296	296	62

## Distribuição Rio de Janeiro & São Paulo

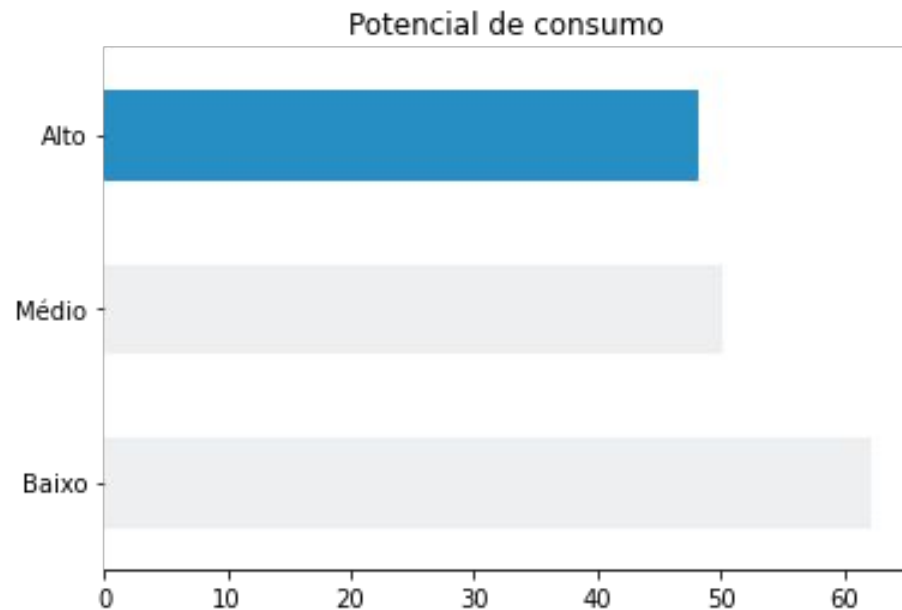
Vemos que há mais dados sobre os bairros de São Paulo do que do Rio.





# Potenciais

Dos 160 bairros do Rio, aproximadamente 50 são considerados como potencial alto.



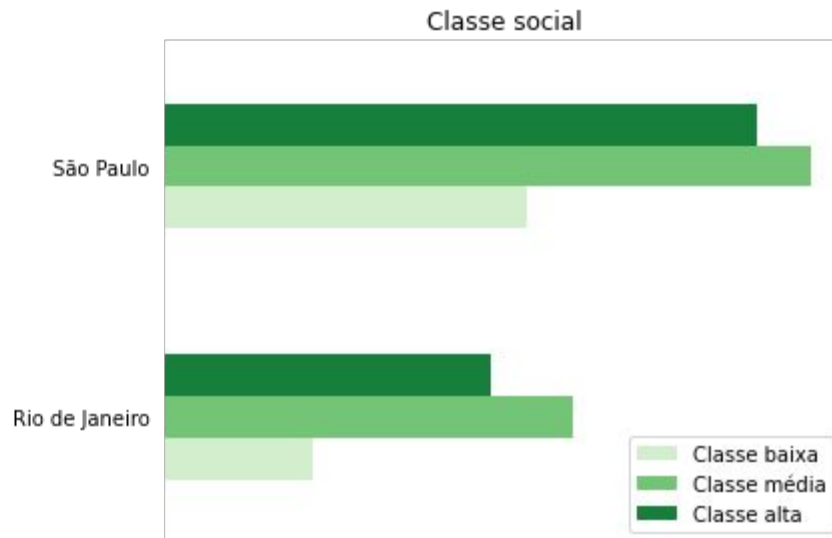
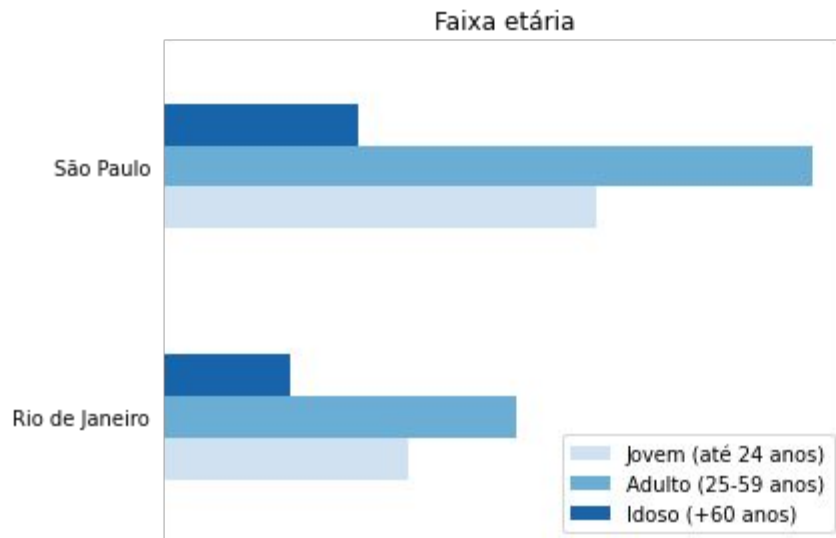


## Como podemos relacionar Rio e São Paulo?

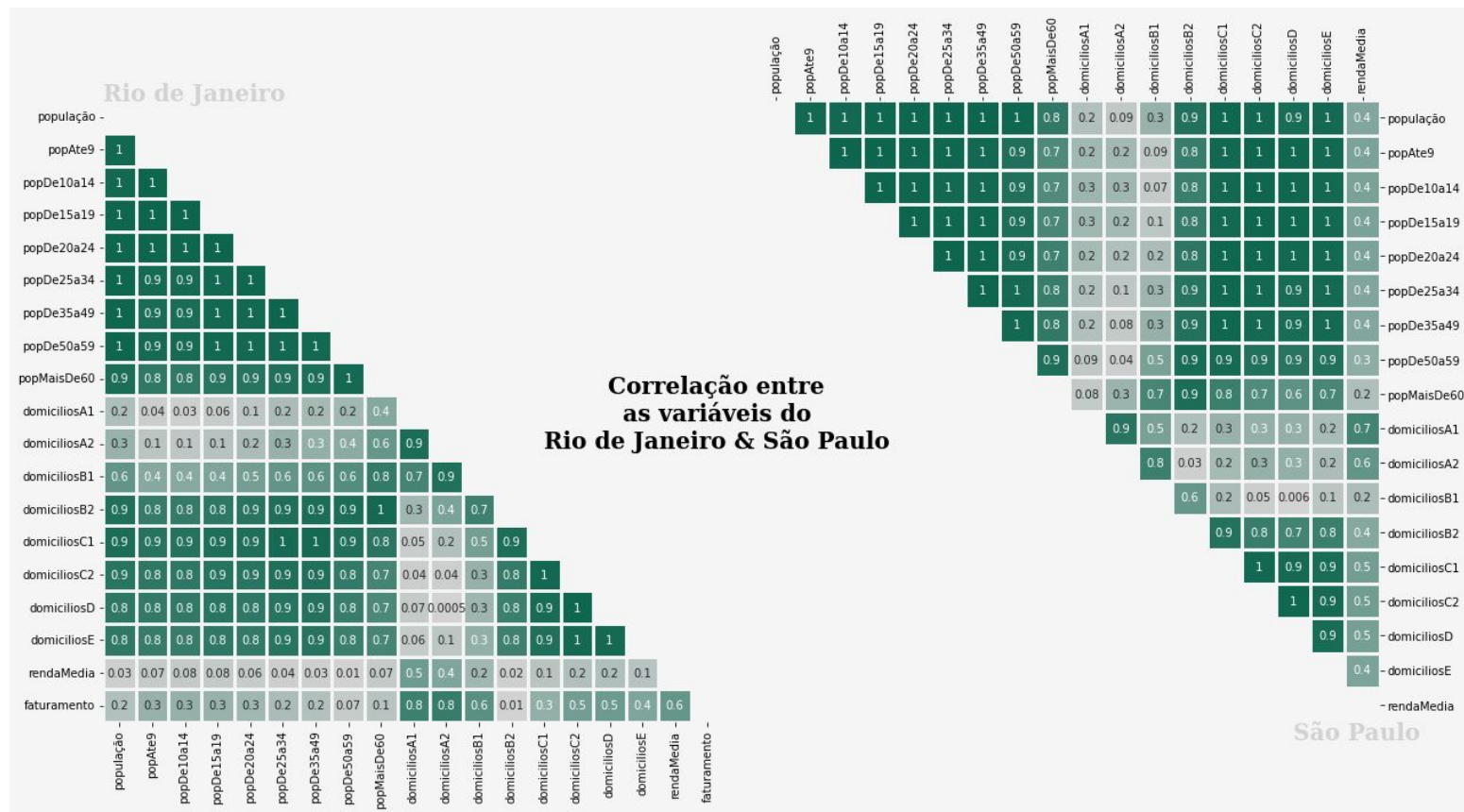
- Como estão distribuídas as suas populações?
- Quão correspondentes são as informações que possuímos de ambos?



# Distribuição população



Classe alta: A1, A2, B1 e B2; Classe média: C1 e C2; Classe baixa: D e E

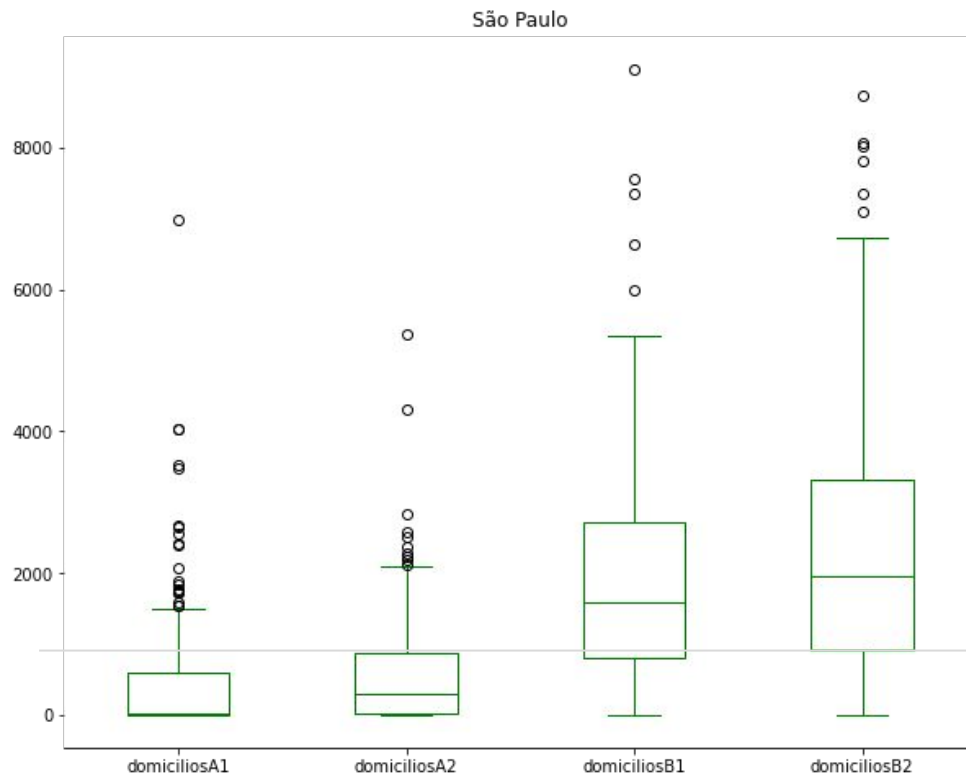


- Forte correlação entre as variáveis de faixa etária e população (tanto no Rio quanto São Paulo)
- Quanto mais verde, maior a correlação
- No Rio há uma correlação moderada entre a renda média com o faturamento

# Distribuição da população (São Paulo)

Público alvo: Domicílios A1-B2

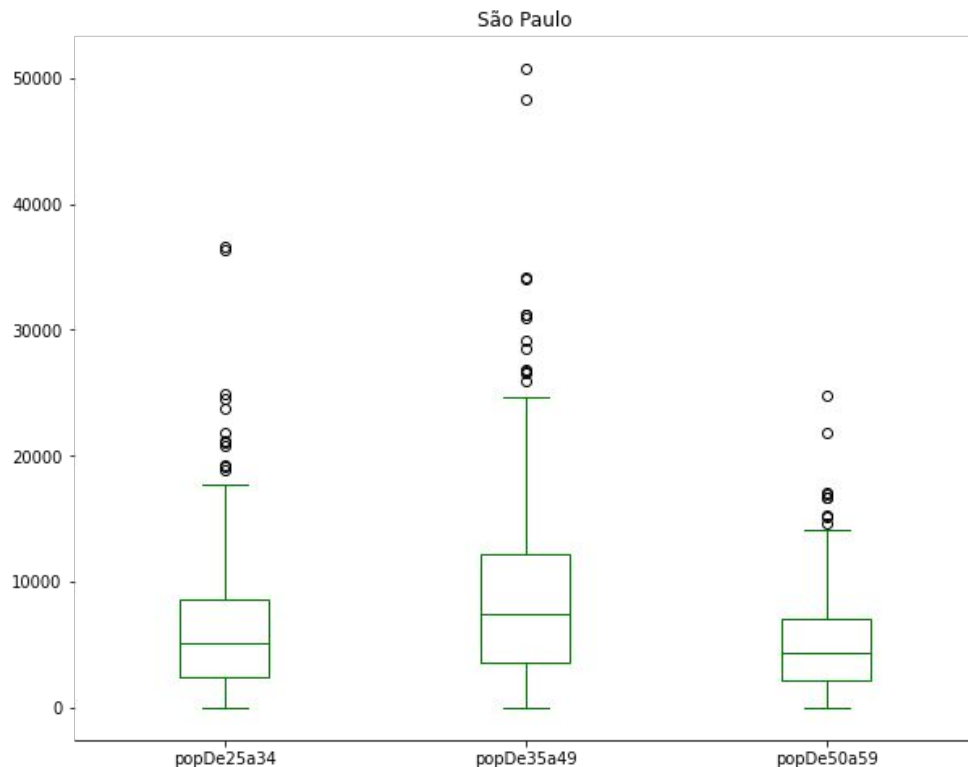
- Há uma menor quantidade de domicílios A1 e A2
- 25% dos bairros B1 e B2 tem aproximadamente 1.000 domicílios de cada
- Nota-se, que na classe A2, 75% dos bairros tem até 1.000 desses domicílios



# Distribuição da população (São Paulo)

Público alvo: Idade entre 25-50

- Há uma maior quantidade de bairros com população entre 35 e 49 anos
- A distribuição dos bairros aparenta possuir uma simetria






## Bairros de São Paulo ordenados pelo público alvo

Ao ordenar pelo número de domicílios A1-B2 e por faixa etária 25-59 obtemos os dados da tabela.

	nome	adulto	alta	população
318	Moema	34245.0	21456.0	60709.0
311	Mandaqui	56711.0	19710.0	108235.0
382	Sacomã	68975.0	17394.0	131007.0
359	Perdizes	26011.0	16520.0	47046.0
395	Saúde	30515.0	15445.0	55175.0
454	Vila Zatt	64041.0	14036.0	125864.0
436	Vila Mariana	20180.0	12940.0	38935.0
398	Tatuapé	26754.0	12604.0	51812.0
314	Marechal Deodoro	22756.0	12296.0	43550.0
403	Trianon	16816.0	12068.0	32963.0

- **Jovem** (Até 24 anos); **Adulto** (de 25 até 59); **Idoso**: 60+
- **Baixa** (D e E), **Média** (C1 e C2); **Alta** (A1, A2, B1 e B2)



# Bairros de São Paulo com renda média ausente

	codigo	nome	cidade	estado	população	popAte9	popDe10a14	popDe15a19
232	355030170.0	Eta Guaraú	São Paulo	SP	0.0	0.0	0.0	0.0
361	355030227.0	Pico Do Jaraguá	São Paulo	SP	0.0	0.0	0.0	0.0
376	355030167.0	Reserva Da Cantareira	São Paulo	SP	0.0	0.0	0.0	0.0

3 rows × 24 columns

- Todas as colunas estão com 0 cadastrado
- Eta Guaraú - Corresponde a uma estação de tratamento de água
- Pico de Jaguará - Corresponde a um parque estadual
- Excluídos da análise

# Distribuição dos bairros do Rio e São Paulo

Considerando a quantidade de domicílios (A1-B2) no eixo y e população entre 25-59 no eixo x.





# Análise Rio de Janeiro





## Bairros do Rio ordenados pelo faturamento (potencial alto)

	nome	população	rendaMedia	faturamento	potencial
9	Barra Da Tijuca	139761.0	18084.0	2915612.0	Alto
36	Copacabana	150524.0	7381.0	2384494.0	Alto
15	Botafogo	85229.0	8316.0	2211985.0	Alto
141	Tijuca	168267.0	7844.0	2157079.0	Alto
82	Leblon	47342.0	14738.0	2119774.0	Alto
50	Flamengo	51456.0	10619.0	1981817.0	Alto
68	Ipanema	43948.0	17188.0	1962438.0	Alto
80	Lagoa	21795.0	63887.0	1775547.0	Alto
81	Laranjeiras	46839.0	8980.0	1762798.0	Alto
93	Méier	51234.0	4671.0	1626856.0	Alto

# Bairros com potencial Alto & Baixo com maiores faturamentos

Investigaremos como esses bairros diferem e como podemos relacioná-los aos de São Paulo.

Méier			Maracanã
Laranjeiras			Anil
Lagoa			Ribeira
Ipanema			Vista Alegre
Flamengo			Zumbi
Leblon			Cocotá
Tijuca			Maria Da Graça
Botafogo			Praia Da Bandeira
Copacabana	Os bairros com maior rendimento da lista são considerados bairros de classe média alta ou alta.		São Francisco Xavier
Barra Da Tijuca			Campinho

# “Bairros de mesma região com faturamento similar”

---

- Análise no Rio de Janeiro

- O Rio de Janeiro contém 164 bairros
- Esses bairros estão contidos dentro de regiões administrativas (33).

# Potencial por região administrativas

- 20 regiões administrativas (*regiao\_adm*) ordenadas pelo potencial, na ordem, Alto, Médio e Baixo.
- Méier, Jacarepaguá, Botafogo, Lagoa e Madureira são os que possuem a maior parte dos bairros com potencial alto.
- Todos os bairros da região administrativa da Penha e Copacabana possuem potencial alto

		potencial		
		Baixo	Médio	Alto
rp	regiao_adm			
Méier	MEIER	5	4	7
Jacarepaguá	JACAREPAGUA	2	3	5
Zona Sul	BOTAFOGO	0	3	5
	LAGOA	1	1	5
Madureira	MADUREIRA	6	4	3
Barra da Tijuca	BARRA DA TIJUCA	3	2	3
Penha	PENHA	0	0	3
Tijuca	VILA ISABEL	1	0	3
Bangu	BANGU	1	1	2
Madureira	IRAJA	3	1	2
Ramos	RAMOS	1	1	2
Zona Sul	COPACABANA	0	0	2
Ilha do Governador	ILHA DO GOVERNADOR	8	6	1
Bangu	REALENGO	2	3	1
Campo Grande	CAMPO GRANDE	1	3	1
Tijuca	TIJUCA	0	2	1
Centro	CENTRO	0	0	1
	SANTA TEREZA	0	0	1
Inhaúma	INHAUMA	3	3	0
Pavuna	ANCHIETA	1	3	0



# Região administrativa por faturamento

- Méier
- Ilha do Governador
- Botafogo
- Lagoa
- Madureira
- Barra da Tijuca

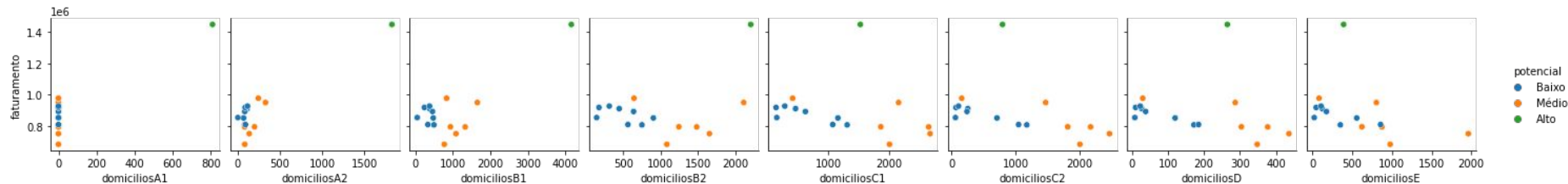
Ilha do Governador não aparecia no topo da tabela anterior como região de potencial alto.

		rendaMedia	faturamento
rp	regiao_adm		
Méier	MEIER	40796.800000	15500811.000000
Ilha do Governador	ILHA DO GOVERNADOR	40634.000000	13387629.000000
Zona Sul	BOTAFOGO	64711.000000	11491880.000000
	LAGOA	149391.000000	10811245.000000
Madureira	MADUREIRA	22955.000000	10379714.000000
Barra da Tijuca	BARRA DA TIJUCA	55670.000000	9207899.000000
Jacarepaguá	JACAREPAGUA	24015.000000	8763376.000000
Tijuca	VILA ISABEL	19673.333333	5526716.000000
Madureira	IRAJA	12682.000000	5314563.000000
Inhaúma	INHAUMA	11353.000000	4864190.000000

# Análise Ilha do Governador

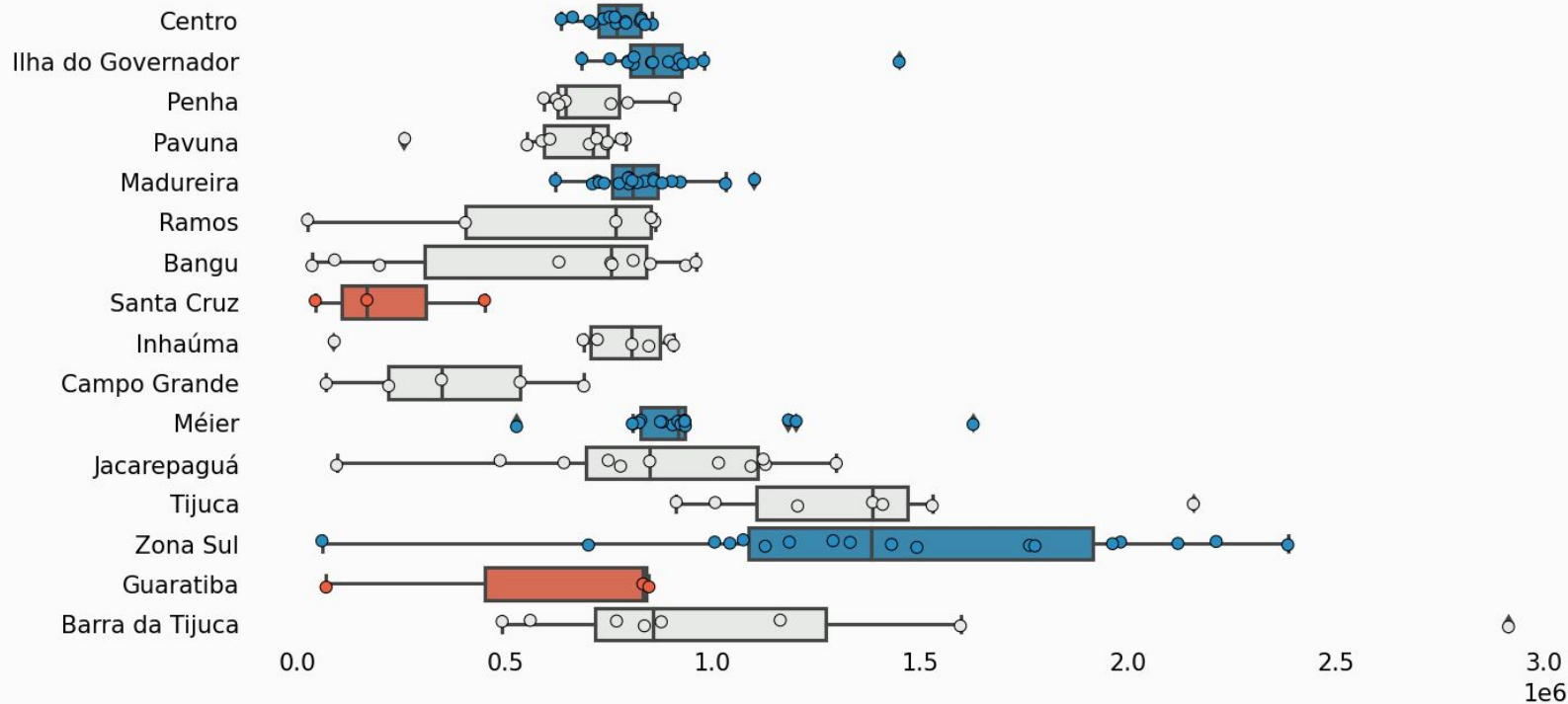
- Não apareceu entre a região com mais potenciais nível alto
- Foi uma das regiões do maior faturamento
- Faturamento levantado pelo Jardim Guanabara

	nome	população	faturamento	rendaMedia	potencial	regiao_adm	rp
13	Jardim Guanabara	33122.0	1448872.0	6499.0	Alto	ILHA DO GOVERNADOR	Ilha do Governador
1	Freguesia (Ilha Do Governador)	19984.0	796321.0	2125.0	Médio	ILHA DO GOVERNADOR	Ilha do Governador
3	Galeão	23620.0	684686.0	1740.0	Médio	ILHA DO GOVERNADOR	Ilha do Governador
4	Tauá	30403.0	752629.0	1704.0	Médio	ILHA DO GOVERNADOR	Ilha do Governador
5	Portuguesa	24529.0	950834.0	2511.0	Médio	ILHA DO GOVERNADOR	Ilha do Governador
6	Moneró	6381.0	978197.0	3669.0	Médio	ILHA DO GOVERNADOR	Ilha do Governador
10	Jardim Carioca	25549.0	795430.0	1943.0	Médio	ILHA DO GOVERNADOR	Ilha do Governador
2	Bancários	12864.0	808554.0	1962.0	Baixo	ILHA DO GOVERNADOR	Ilha do Governador
8	Cocotá	5012.0	912281.0	2927.0	Baixo	ILHA DO GOVERNADOR	Ilha do Governador
16	Praia Da Bandeira	6116.0	893640.0	2699.0	Baixo	ILHA DO GOVERNADOR	Ilha do Governador
18	Cacuaia	11325.0	852714.0	2022.0	Baixo	ILHA DO GOVERNADOR	Ilha do Governador
22	Pitangueiras	12088.0	810599.0	1688.0	Baixo	ILHA DO GOVERNADOR	Ilha do Governador
26	Zumbi	2072.0	919451.0	3791.0	Baixo	ILHA DO GOVERNADOR	Ilha do Governador
27	Ribeira	3629.0	928239.0	3420.0	Baixo	ILHA DO GOVERNADOR	Ilha do Governador
40	Cidade Universitária	1442.0	855182.0	1934.0	Baixo	ILHA DO GOVERNADOR	Ilha do Governador



## Variância do faturamento

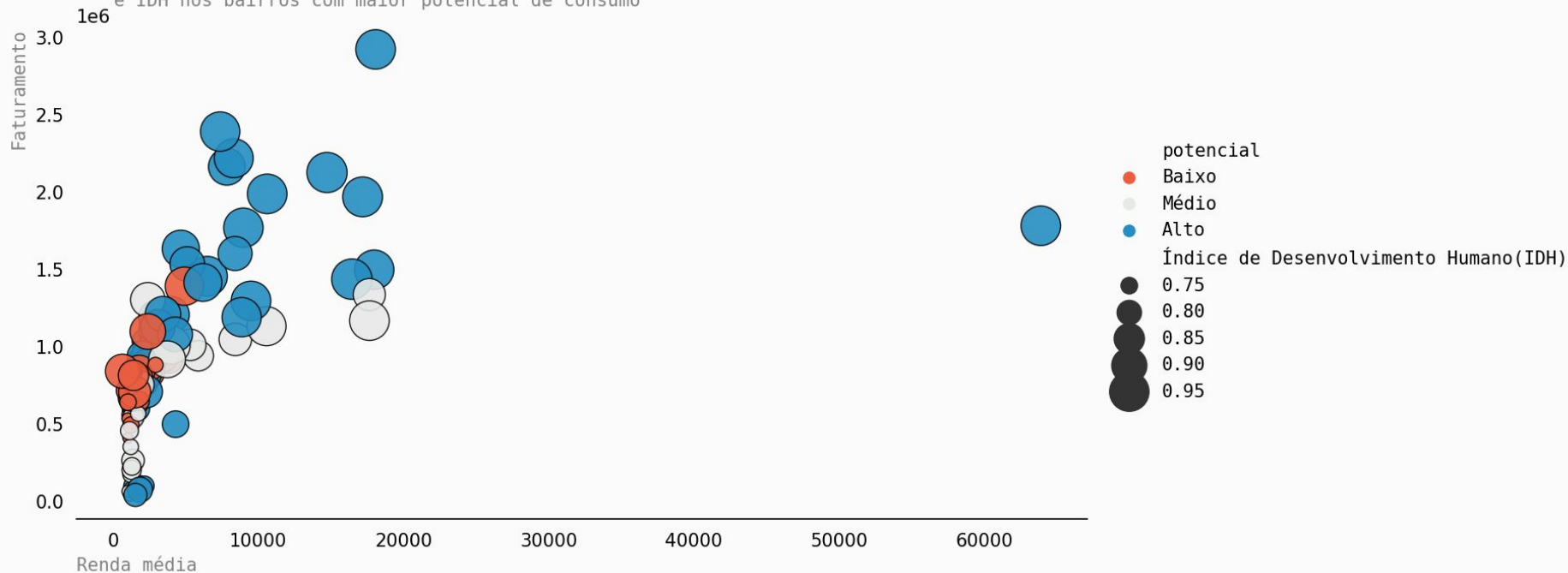
Há variância em todos os bairros, alguns variam mais do que outros, por exemplo, a Zona Sul que possui uma grande dispersão.



Faturamento por região de planejamento do Rio

## Renda média, Faturamento e IDH por bairro

Vemos um forte relacionamento entre as variáveis faturamento, renda média e IDH nos bairros com maior potencial de consumo





---

# Modelagem

# Predição de Faturamento

- Para etapa de previsão de faturamento fazemos uma busca por modelos de regressão
- Utilizamos parâmetros padrões para esses modelos
- O melhor modelo é selecionado para ser explorado posteriormente

```
"ElasticNet",  
"SGDRegressor",  
"SVM",  
"BayesianRidge",  
"KernelRidge",  
"LinearRegression",  
"XGBoost"
```

$$MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n}$$

$$RMSE(\hat{\theta}) = \sqrt{MSE(\hat{\theta})}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$



## Seleção de modelos

- Dividimos em teste/treino com partição 70/30
- Entre as métricas e modelos escolhidos, o de maior performance foi o XGBoost

	nome	mse	mae	rmse
<b>0</b>	ElasticNet	7.470713e+10	1.234394e+05	2.733261e+05
<b>1</b>	SGDRegressor	3.233579e+55	5.686457e+27	5.686457e+27
<b>2</b>	SVM	3.780502e+11	3.408051e+05	6.148579e+05
<b>3</b>	BayesianRidge	5.485924e+10	1.141340e+05	2.342205e+05
<b>4</b>	KernelRidge	8.798036e+10	1.363467e+05	2.966148e+05
<b>5</b>	LinearRegression	7.853243e+10	1.269443e+05	2.802364e+05
<b>6</b>	XGBoost	2.991091e+10	1.034727e+05	1.729477e+05



## Busca por hiperparâmetros

- Foi realizada uma busca de hiperparâmetros especificamente para o XGBoost
- A busca escolhida foi o grid search, que varia os valores dos parâmetros pré-definidos

```
parameters = {  
    "XGB__objective": ["reg:squarederror"],  
    "XGB__learning_rate": [0.1,0.01],  
    "XGB__booster": ["gbtree","gblinear"],  
    "XGB__max_depth": [1,3,5,7,10],  
    "XGB__min_child_weight": [1,3,5,7],  
    "XGB__colsample_bytree": [0.25,0.5,0.7],  
    "XGB__n_estimators": [30,50,100],  
    "XGB__reg_alpha": [0.2,0.5],  
    "XGB__reg_lambda": [2,5,7],  
    "XGB__gamma": [3,5,7],  
    "XGB__random_state": [123],
```

# Predições

- Predições no conjunto de teste do Rio de Janeiro
- Métricas finais:
  - mse: 61133322316.58545
  - mae: 122722.70572916667
  - rmse: 247251.53653028214
- O valor das métricas é alto pois a magnitude dos valores é alta

	nome	real	predicao
0	Pedra De Guaratiba	832018.0	8.012392e+05
1	Piedade	808082.0	8.422248e+05
2	Tijuca	2157079.0	1.534679e+06
3	Gardênia Azul	641865.0	7.316982e+05
4	Moneró	978197.0	9.236309e+05
5	Cidade De Deus	488021.0	5.620263e+05
6	Parque Anchieta	744303.0	8.037671e+05
7	Freguesia (Ilha Do Governador)	796321.0	8.416055e+05
8	Parada De Lucas	630075.0	6.835660e+05
9	Todos Os Santos	1200769.0	1.151904e+06
10	Caju	635348.0	6.821798e+05
11	Lins De Vasconcelos	932622.0	8.562518e+05
12	Botafogo	2211985.0	1.976184e+06
13	Inhaúma	688001.0	6.000611e+05
14	Cascadura	796395.0	7.321367e+05
15	Cidade Nova	837355.0	8.228039e+05
16	Santa Teresa	703465.0	7.936163e+05
17	Vargem Pequena	560631.0	6.546094e+05
18	Padre Miguel	629794.0	6.348360e+05
19	Brás De Pina	755073.0	7.468252e+05
20	Jardim Botânico	1491476.0	1.426256e+06
21	Cacuaia	852714.0	8.356323e+05
22	Benfica	662520.0	7.147164e+05
23	Barra Da Tijuca	2915612.0	1.940768e+06



## Predição de Potencial

- Tentamos prever as 3 classes de potencial: Alto, Médio, Baixo
- Realizamos uma busca de modelos
- Utilizamos parâmetros padrões para esses modelos
- O melhor modelo é selecionado para ser explorado posteriormente



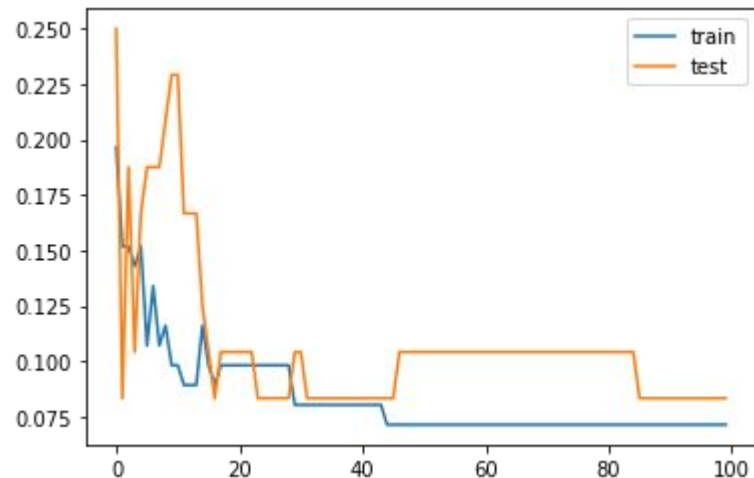
## Predição de Potencial

- Dentre os modelos experimentados, o XGBoost novamente foi o de melhor performance
- É posteriormente realizado um grid search nos parâmetros desse modelo

	model	precision	recall	f1-score
0	Nearest Neighbors	0.805159	0.791667	0.765049
1	RBF SVM	0.125434	0.354167	0.185256
2	Gaussian Process	0.097656	0.312500	0.148810
3	Decision Tree	0.837550	0.812500	0.811786
4	Random Forest	0.718915	0.729167	0.709398
5	Neural Net	0.624617	0.625000	0.624369
6	AdaBoost	0.521671	0.541667	0.500102
7	Naive Bayes	0.589471	0.604167	0.570896
8	QDA	0.795139	0.770833	0.771039
9	XGBoost	0.900000	0.895833	0.892580

## Predição de Potencial - Melhor XGBoost

- Depois de realizar o grid search e encontrar os melhores parâmetros, avaliamos o treinamento do modelo no conjunto de treino e teste
- Eixo y é a função objetivo sendo minimizada
- Eixo x é a quantidade de estimadores(árvores)





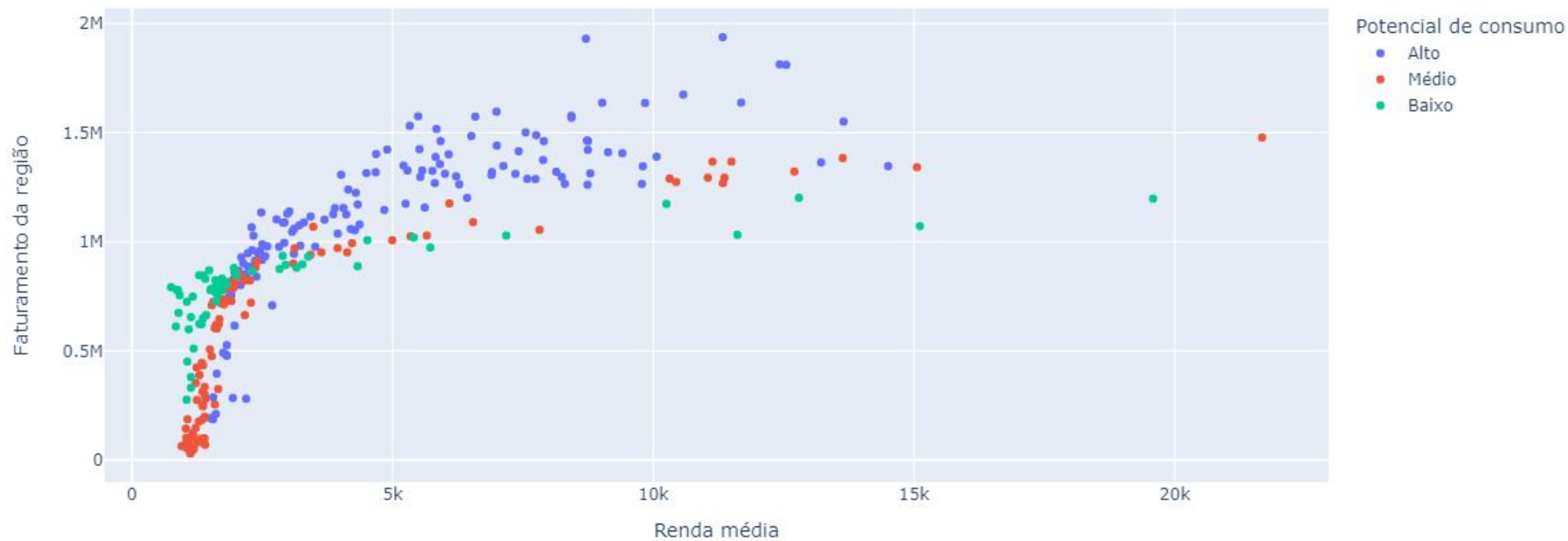
## Predição de Potencial - Melhor XGBoost

- Treino/Teste  
particionados  
em 70/30

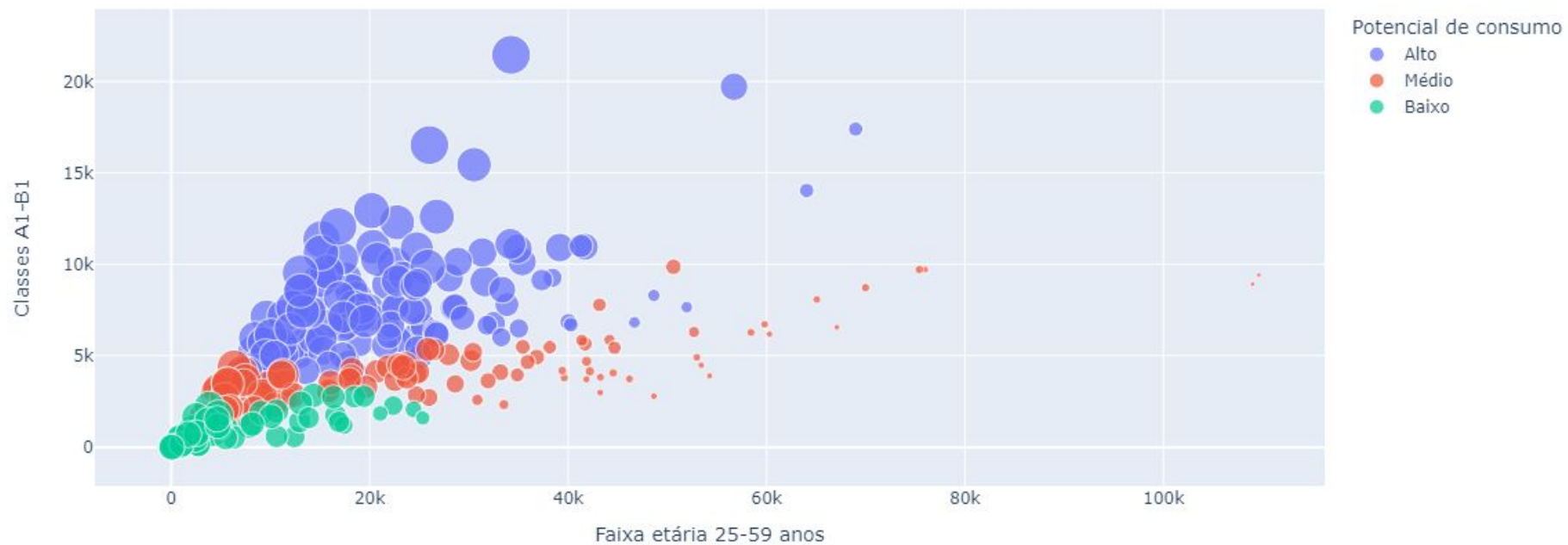
		precision	recall	f1-score	support
Treino	Alto	0.89	0.97	0.93	32
	Baixo	0.93	0.93	0.93	45
	Médio	0.97	0.89	0.93	35
	accuracy			0.93	112
	macro avg	0.93	0.93	0.93	112
	weighted avg	0.93	0.93	0.93	112

		precision	recall	f1-score	support
Teste	Alto	1.00	0.94	0.97	16
	Baixo	0.85	1.00	0.92	17
	Médio	0.92	0.80	0.86	15
	accuracy			0.92	48
	macro avg	0.92	0.91	0.91	48
	weighted avg	0.92	0.92	0.92	48

## Bairros de São Paulo com maior potencial por renda



## Bairros de São Paulo com maior potencial por público alvo

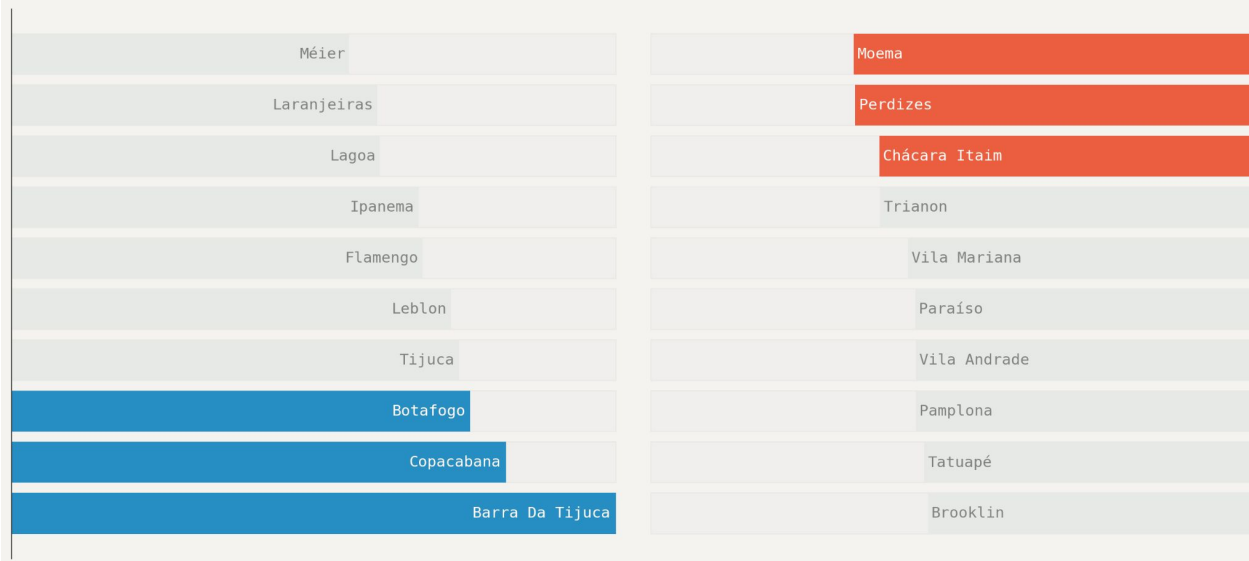




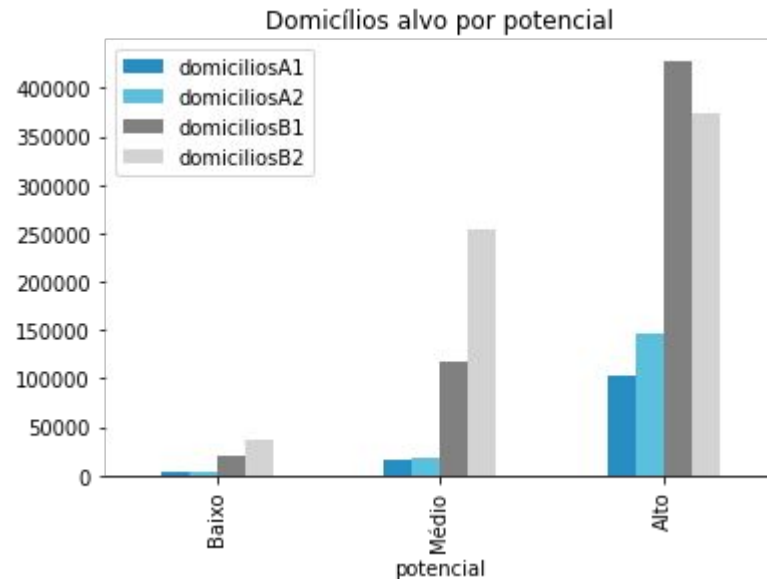
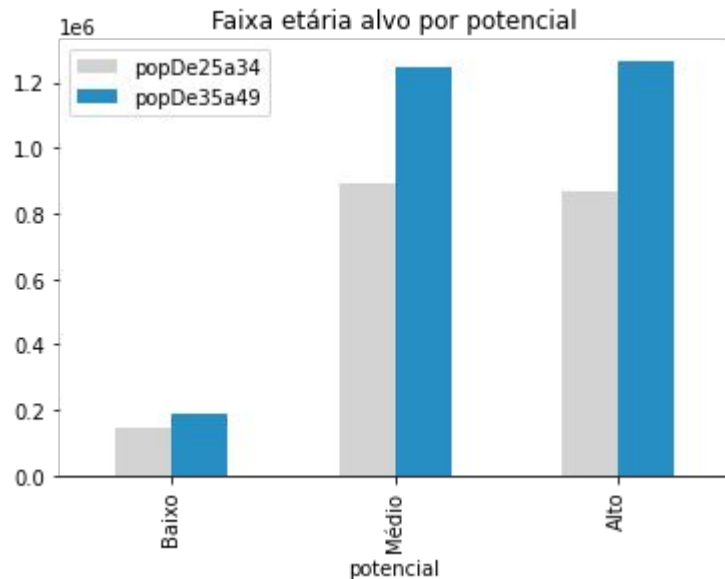
# Comparação entre cidades

## Bairros do Rio e São Paulo de potencial alto ordenados pelo faturamento

Bairros com maior faturamento considerados de potencial alto



# População alvo por potencial





## Bairros de acordo com o faturamento

	nome	rendaMedia	faturamento	potencial
157	Moema	11332.0	1938202.2	Alto
198	Perdizes	8707.0	1931360.6	Alto
50	Chácara Itaim	12424.0	1813645.0	Alto
240	Trianon	12550.0	1811308.0	Alto
273	Vila Mariana	10575.0	1675272.1	Alto
171	Paraíso	11686.0	1638488.9	Alto
246	Vila Andrade	9020.0	1637444.2	Alto
167	Pamplona	9843.0	1636636.1	Alto
235	Tatuapé	6994.0	1596681.6	Alto
30	Brooklin	8427.0	1577753.9	Alto



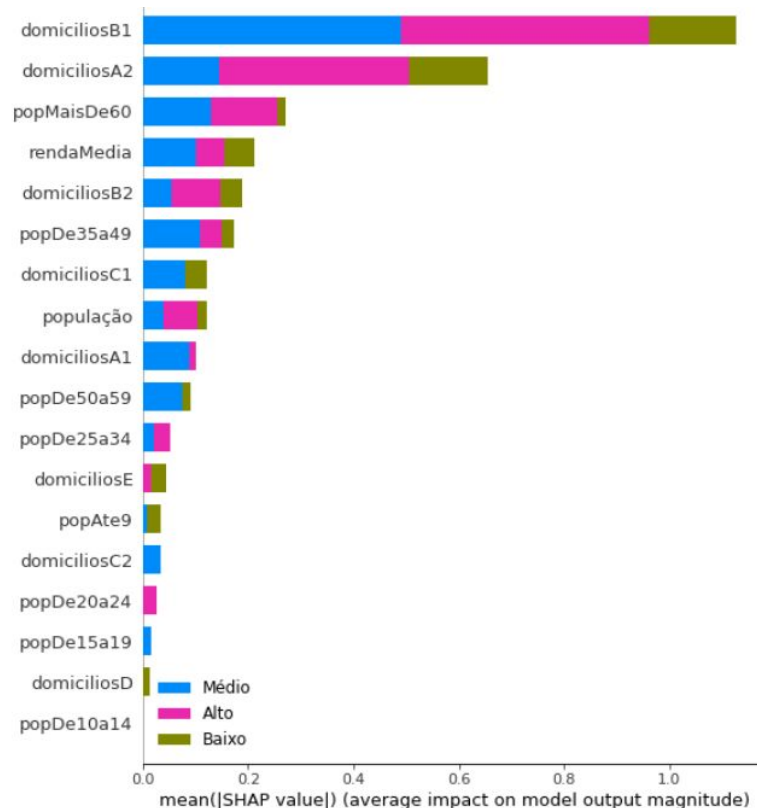
## Explicabilidade - Variáveis mais importantes

- Treino/Teste  
particionados em 70/30
- O XGBoost fornece as  
importâncias

```
[(0.0, 'popDe10a14'),  
(0.028203046, 'popAte9'),  
(0.036425527, 'domiciliosE'),  
(0.036655527, 'domiciliosD'),  
(0.039984833, 'domiciliosC2'),  
(0.040283673, 'popDe25a34'),  
(0.04118149, 'rendaMedia'),  
(0.05071579, 'popDe15a19'),  
(0.051711507, 'domiciliosC1'),  
(0.052092142, 'domiciliosA1'),  
(0.054709893, 'popDe35a49'),  
(0.06983957, 'população'),  
(0.07002102, 'domiciliosB2'),  
(0.071435034, 'domiciliosA2'),  
(0.08406548, 'popDe20a24'),  
(0.08732485, 'popDe50a59'),  
(0.09095586, 'popMaisDe60'),  
(0.09439472, 'domiciliosB1')]
```

## Explicabilidade - Shapley

- Para o método de shapley values, as variáveis de domicílio B1 e A1 são mais importantes para definir as classes Média e Alta.
- Já para prever a classe Baixa as variáveis domicílios E-D são importantes









## Informações adicionais

- Foram adicionadas informações do IDH dos bairros do Rio de Janeiro. Não foram observadas melhorias significativas na classificação
  - Os dados de IDH foram coletados nos anos 2000
  - Talvez seja necessário um tratamento desses dados e uma investigação maior
- Foram experimentadas abordagens não supervisionada como a clusterização.
  - A vantagem é descobrir bairros similares por suas características nos dados
  - Seria necessário, no entanto, um tempo maior para investigar os resultados.