O BRASIL EM DADOS - ANÁLISE GERAL DAS CIDADES DO BRASIL

INTRODUÇÃO À CIÊNCIA DOS DADOS - CCF 425

André Elias
Matrícula 3013
Aluno de Ciência da Computação
Universidade Federal de Viçosa - Florestal, MG
andre.elias@ufv.br

Thomas Chang
Matrícula 3052
Aluno de Ciência da Computação
Universidade Federal de Viçosa - Florestal, MG
thomas.chang@ufv.br

Victor Hugo
Matrícula 3510
Aluno de Ciência da Computação
Universidade Federal de Viçosa - Florestal, MG
victor.h.santos@ufv.br

27 de outubro de 2021

RESUMO

Este relatório tem como objetivo apresentar os resultados obtidos através do trabalho O Brasil em dados - Análise Geral das Cidades do Brasil. Desenvolvido pelos autores ao cursar a disciplina Introdução à Ciência de Dados - CCF 425, durante o primeiro semestre de 2021.

1 Introdução

O objetivo do trabalho prático é aplicar os conteúdos aprendidos em sala de aula em um projeto real, com dados reais disponíveis publicamente que possuem relação com o Brasil. E foi dividido em quatro etapas práticas.

Na etapa de *Escolha de Dados e Planejamento* o grupo escolheu um conjunto de dados com todas as 5570 cidades brasileiras, que possuem até 80 atributos diferentes. Podendo ser acessado aqui. Elaborou-se uma lista de 20 questões a serem respondidas com o trabalho.

Após a escolha do conjunto de dados, o grupo realizou a etapa de *Preparação dos Dados*, com o objetivo de entender os atributos e objetos dos dados, identificar possíveis ruídos ou informações ausentes,formatar valores, com o objetivo de realizar uma análise mais precisa. A preparção dos dados será apresentada mais afundo posteriormente.

A etapa seguinte foi *Análise exploratória e extração de conhecimento*, onde se gerou estatísticas descritivas, gráficos e tabelas para conhecer os dados. E poder relacionar os diferentes objetos e atributos, visando responder parte dos questionamentos. É importante ressaltar que essa parte foi mais focada em responder as hipóteses iniciais. As visualizações serão aprimoradas para a apresentação final. A próxima etapa é a *Análise Preditiva*, que ainda será realizada ao decorrer do semestre.

2 Preparação dos Dados

Esta etapa foi realizada após a escolha do conjunto de dados, e seus objetivos são entender os atributos e objetos dos dados, identificar possíveis ruídos ou informações ausentes, formatar valores, com o intuito de realizar uma análise mais precisa.

Primeiramente os atributos 'REGIAO_TUR' e 'CATEGORIA_TUR' receberam a string 'Nenhum' em campos nulos, que representam, justamente, a ausência de turismo naquela determinada região. Pela string 'Nenhum', foi mais fácil e dinâmico de tratar dados futuramente em alguma possível análise.

Em seguida, o atributo 'MUN_EXPENDIT' em campos nulos recebeu a mediana de todos os outros valores da coluna. Inicialmente foi pensado em colocar 0 em tais campos, porém o valor 0 poderia enviezar os dados, pelo campo se tratar de um valor em reais e por 0 ser um valor extremo. Colocar a média também pode não ser uma ideia boa, pois o desvio padrão está relativamente alto, então o grupo decidiu colocar a mediana. Porque a mediana dos outros valores, a chance daquele determinado valor variar muito a análise é pequena.

Analogamente a ideia aplicada nos atributos 'REGIAO_TUR' e 'CATEGORIA_TUR', os atributos 'HOTELS', 'BEDS', 'UBER', 'MAC' e 'WAL-MART' possuem o valor nulo, que indica ausência daquele atributo naquela cidade, então, colocamos o valor 0 em tais campos.

Quanto às instituições bancárias e quantidade de ativos, para as instituições, o valor 0 foi colocado pois indica ausência e para os ativos ('Pr_Assets' e 'Pu_Assets') colocou-se a mediana pelo mesmo motivo de ter sido colocado a mediana em 'MUN_EXPENDIT', por se tratar de dinheiro e o valor 0 ser capaz de enviezar a análise.

A maioria dos campos nulos foram preenchidos e, portanto, os restantes, que é uma quantidade extremamente pequena, será removida a linha da tabela que os contenha, para evitar futuros problemas. O grupo acredita que a remoção de poucas cidades não atrapalhará na análise, pela quantidade restante de cidades. Sendo assim, ficamos com 5368 cidades com todos os dados preenchidos.

3 Análise exploratória e extração de conhecimento

Com os dados já preparados, o grupo buscou explorar e extrair conhecimento para tirar algumas conclusões e possivelmente responder algumas perguntas que foram propostas e serão discutidas posteriormente.

Nas subsessões a seguir, têm-se cada uma das perguntas e as conclusões tiradas. No notebook apresentado no **github** é possível ver a codificação feita para obter tais conclusões.

3.1 Quais cidades possuem maior IDH?

Uma coluna presente no conjunto de dados utilizado é a 'IDHM', que informa o valor do IDH de cada cidade e possibilita encontrar as cidades com os maiores IDH. A tabela a seguir apresenta tal resultado.

CITY	IDH
Florianópolis	0.85
Santos	0.84
Joaçaba	0.83
Jundiaí	0.82
Brasília	0.82
Valinhos	0.82
Vinhedo	0.82
Araraquara	0.81
Nova Lima	0.81
Ilha Solteira	0.81

Tabela 1: 10 Cidades do Brasil com maior IDH

3.2 Existe alguma relação entre IDH e algum outro índice (renda, educação, etc)? O que leva um bom IDH?

Inicialmente foi calculado a correlação entre o atributo 'IDHM', que, como explicado anteriormente, representa o IDH de uma determinda cidade, e os demais atributos.

Com os resultados obtidos, foi possível notar que os atributos a seguir apresentaram um valor consideravelmente alto para a correlação:

- IDH de educação;
- IDH de renda;
- Expectativa de Vida e

- PIB

Além disso, pode-se observar que a **latitude** também é um atributo interessante de se prestas atenção para essa questão, pois os países do Sul, segundo a latitude, tendem a ter um IDH maior.

3.3 Cidades que possuem maior IDH também possuem mais residentes?

Calculando a correlação entre os atributos necessários ('IDHM' e 'ESTIMATED_POP') obteve-se um valor consideravelmente baixo. Porém, considerando um bom IDH como um valor acima de **0.72**, que representa o 3º quartil (75%) dos dados em relação ao IDH e separando as cidades com bom IDH das demais, após calcular a média da população para ambos grupos de cidades (IDH bom e IDH ruim), podemos perceber que a média das cidades com IDH bom é **consideravelmente maior**.

3.4 Quais as melhores cidades para se viver? Elas são capitais?

Antes de buscar responder o questionamento, é importante definir **o que é uma boa cidade para se viver**. Por consentimento do grupo, foi definido que o **IDH** é uma boa medida, pois o IDH engloba outros índices bem importantes para o ser humano, em geral.

A subsessão 3.1 apresenta as cidades com os maiores IDH, porém, somente 2 delas são capitais, são elas: **Florianópolis e Brasília**. Temos então que somente 20% das 10 cidades com maior IDH do Brasil são capitais.

As 10 capitais com melhor IDH são:

CITY	IDH
Florianópolis	0.85
Brasília	0.82
São Paulo	0.81
Goiânia	0.80
Palmas	0.79
Cuiabá	0.79
Campo Grande	0.78
São Luís	0.77
João Pessoa	0.76
Salvador	0.76
Boa Vista	0.75
Teresina	0.75

Tabela 2: 10 Capitais do Brasil com maior IDH

3.5 Podemos relacionar diretamente o número de McDonald's com o IDH?

Calculando a correlação entre os atributos ('IDHM' e 'MAC'), temos um valor consideravelmente baixo (0.08), o que nega o questionamento feito. Tal resposta pode ser explicada pelo fato da rede McDonald's ser uma franquia muito famosa e estar presente em diversas cidades, até mesmo aquelas com baixo IDH.

3.6 Podemos afirmar com confiança que cidades que possuem maior expectativa de vida também são capitais?

Plotando o histograma com o eixo x sendo a expectativa de vida e e o eixo y sendo a probabilidade e traçando a curva de densidade para cidades que são capitais (1) e também para cidades que não são capitais (0). Obtemos:

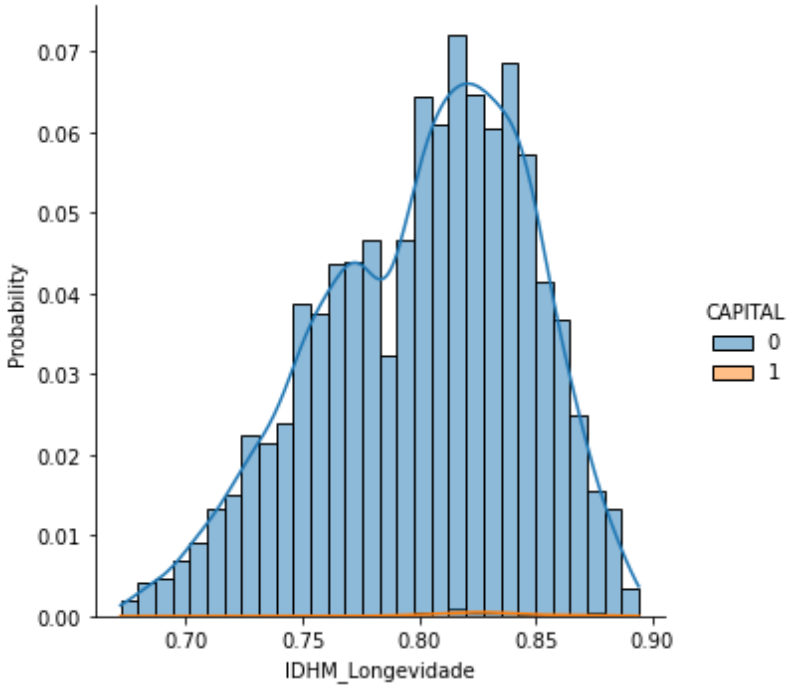

Figura 1: Histograma da Expectativa de Vida

Com o gráfico obtido, podemos afirmar que a confiança para afirmar que cidades que possuem maior expectativa de vida também são capitais é extremamente baixa, com uma probabilidade próxima de 0.

3.7 Cidades que tem mais estrangeiros são as que tem mais oportunidades de emprego?

As oportunidades de emprego de uma cidade, foram definidas, pelo grupo, como sendo proporcional à quantidade de companhias que aquela determinada cidade possui. Além dessa definição, o grupo retirou a cidade de São Paulo de tal análise, pois tal cidade estava se comportando como um outlier e assim enviesando os dados. A visualização dos resultados considerando a cidade de São Paulo está no notebook, porém, para este relatório, será apresentado somente o resultado desconsiderando tal cidade.

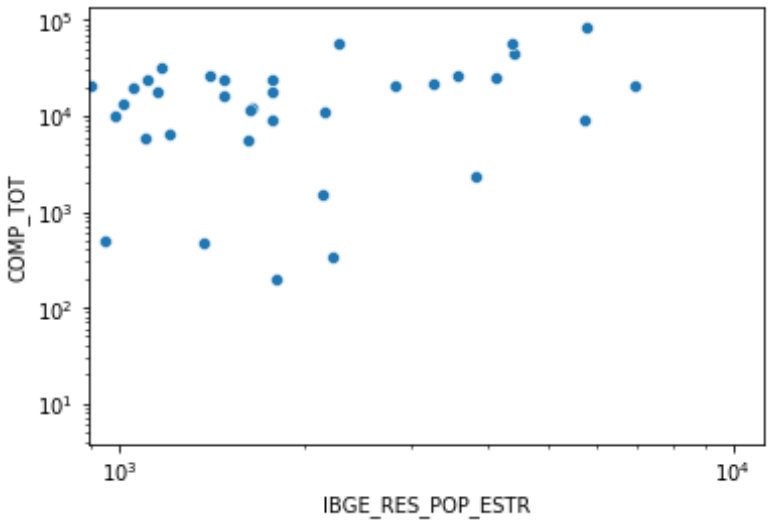

Figura 2: Gráfico de Pontos entre Quantidade de Companhias e Quantidade de Estrangeiros Residentes em cada cidade

O gráfico obtido nos mostra visualmente que os atributos não se relacionam de forma linear. Existem muitas cidades com **muitas companhias e poucos estrangeiros**, porém as **cidades com muitos estrangeiros são cidades com muitas companhias**. Não podemos dizer que as cidades que possuem mais estrangeiros são as que tem **mais** oportunidades de emprego, mas podemos dizer que as **cidades com muitos estrangeiros são cidades com boas oportunidades**.

3.8 Existe alguma relação entre cidades que possuem maior renda e assinam TV a cabo?

Nesta análise, assim como na subsessão 3.7, a cidade de São Paulo foi removida da análise.

O gráfico abaixo apresenta a relação entre ambos atributos.

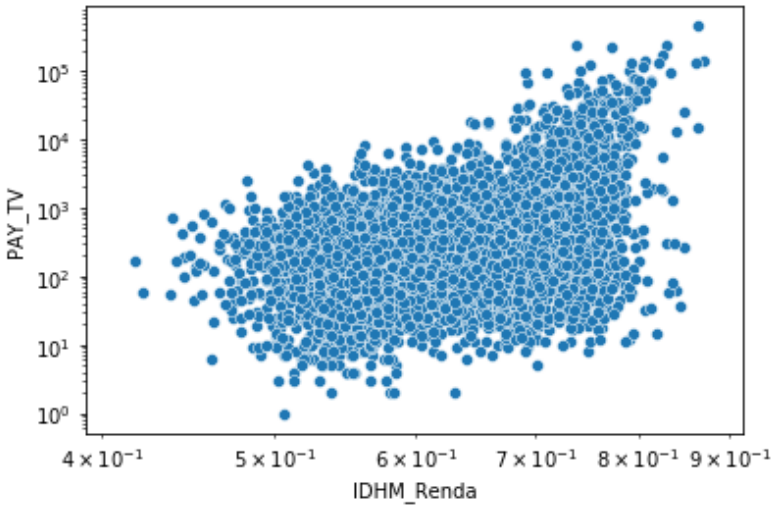

Figura 3: Gráfico de Pontos entre Quantidade de Assinaturas de TV a Cabo e Renda de cada Cidade

Com o gráfico obtido, podemos observar que os valores estão concentrados entre os valores de 10^1 e 10^5 para quantidade de assinantes de TV a Cabo e 5 x 10^{-1} e 8 x 10^{-1} para o IDH de renda. Além disso, a correlação entre ambos atributos é de **0.21**, logo, não podemos afirmar com significativa confiança a relação entre cidades que possuem maior renda e assinam TV a Cabo.

3.9 Cidades com maior carga tributária são as que possuem maior IDH?

Nesta análise, assim como na subsessão 3.7, a cidade de São Paulo foi removida da análise.

O gráfico abaixo apresenta a relação entre ambos atributos.

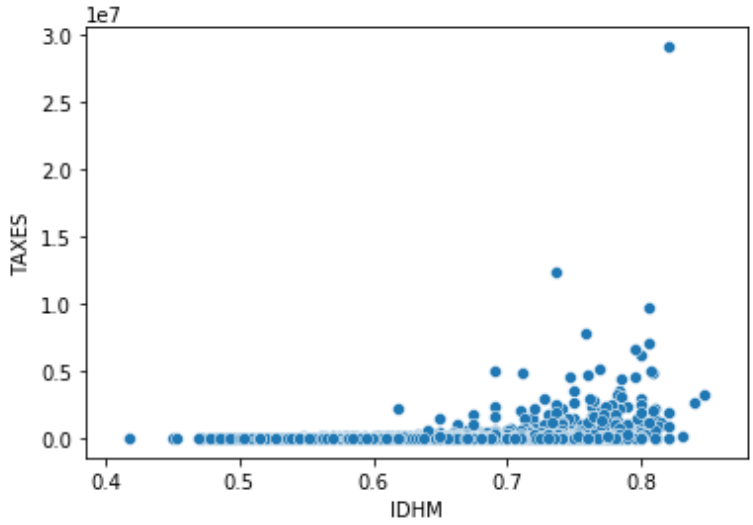

Figura 4: Gráfico de Pontos entre Carga Tributária e IDH

O gráfico plotado nos mostra uma certa concentração dos pontos ao lado direito do gráfico. Isso nos diz que as cidades com maior IDH são cidades que tendem a cobrar uma maior carga tributária.

3.10 As cidades que possuem mais hectares de plantação são também as que mais possuem número total de tratores de rodas?

Ao elaborar as perguntas deste projeto, o grupo supôs que a quantidade de rodas de trator estava proporcionalmente relacionada à quantidade de área plantada, porém os dados mostraram exatamente o oposto, como conta na figura 5.

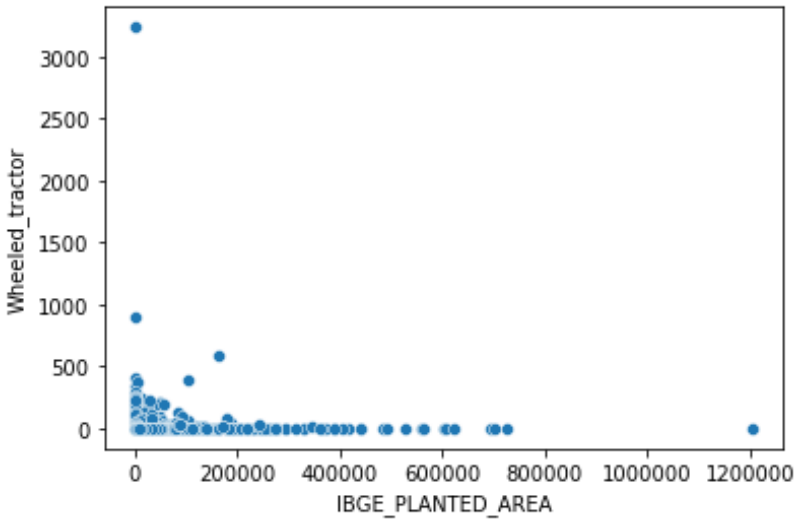

Figura 5: Gráfico de Pontos entre Quantidades de Rodas de Trator e Área Plantada

Logo, pode-se dizer que as cidades que possuem mais hectares de plantação não são as que possuem maior número de rodas de trator.

3.11 Quais a 3 cidades que mais pagam impostos?

CITY	TAXES
São Paulo	117125386.74
Brasília	29145585.42
Manaus	12433232.22

Tabela 3: Cidades que mais pagam impostos

A tabela 3 apresenta as 3 cidades que mais pagam impostos, seguidas de seus respectivos valores.

3.12 Quais são as maiores cidades que não são capitais? Existem cidades que não são capitais e são maiores do que alguma capital?

O tamanho de uma cidade foi definido, pelo grupo, como sendo proporcional à quantidade de habitantes e as maiores cidades que não são capitais são apresentas na tabela 4.

CITY	IBGE_RES_POP
Campinas	1080113.00
São Gonçalo	999728.00
Duque De Caxias	855048.00
Nova Iguaçu	796257.00
São Bernardo Do Campo	765463.00

Tabela 4: Maiores cidades que não são capitais

Sobre o questionamento de existir ou não cidades que não são capitais e são maiores que capitais, conhecendo o Brasil, pode-se afirmar que sim. O Brasil contém cidades não capitais com grande expressão em diversos fatores, como por exemplo Campinas. Porém também foram manipulados os dados de forma a apresentarem tal resultado.

CITY	IBGE_RES_POP
Florianópolis	421240.00
Macapá	398204.00
Rio Branco	336038.00
Boa Vista	284313.00
Palmas	228332.00

Tabela 5: População de algumas capitais

Analisando a tabela 4 e 5, temos Campinas (não capital), por exemplo, sendo maior que Florianópolis (capital).

3.13 Quais estados que mais produzem safra (agricultura), elas possuem muita mão de obra estrangeira?

A tabela 6 mostra os estados que mais produzem safra, o valor agregado bruto das safras e a quantidade de residentes em cada estado.

Os resultados obtidos nos dizem que **PR, RS e SP** são dominantes no quesito produção de safra. Além disso, **SP** é um estado que possui uma notável concentração estrangeira em relação aos outros.

STATE	GVA_AGROPEC	IBGE_RES_POP_ESTR
PR	31038394.40	19538.00
RS	30512205.76	17384.00
SP	30364219.18	180395.00
MG	27798564.49	10149.00
MT	22018684.51	3089.00
GO	16777610.12	5377.00
PA	14912877.30	3566.00
BA	13457196.93	8875.00
MS	13396656.03	8620.00
SC	12697019.73	10317.00

Tabela 6: Produção de Safra e Residentes Estrangeiros por Estado

3.14 Quais as cidades mais altas do brasil?

A tabela 7 apresenta as cidades mais altas do Brasil seguidas de sua respectiva altitude.

CITY	ALT
Divisa Nova	874579.00
São Miguel Arcanjo	665758.00
Carmésia	572655.00
Nova Xavantina	271009.00
Olho D'Água Grande	134461.00
Governador Archer	132852.00
Campos Do Jordão	1639.15
Senador Amaral	1495.64
Bom Repouso	1378.71
Urupema	1345.42

Tabela 7: Cidades mais altas do Brasil

3.15 Existe alguma diferença significativa entre cidades com populações mais jovens e mais velhas?

Considerou-se cidades mais velhas como as que tem mais de 12% de população com mais de 60 anos. Visando uma melhor visualização do boxplot aplicou-se também um filtro para as cidades com menos de 100 mil habitantes.

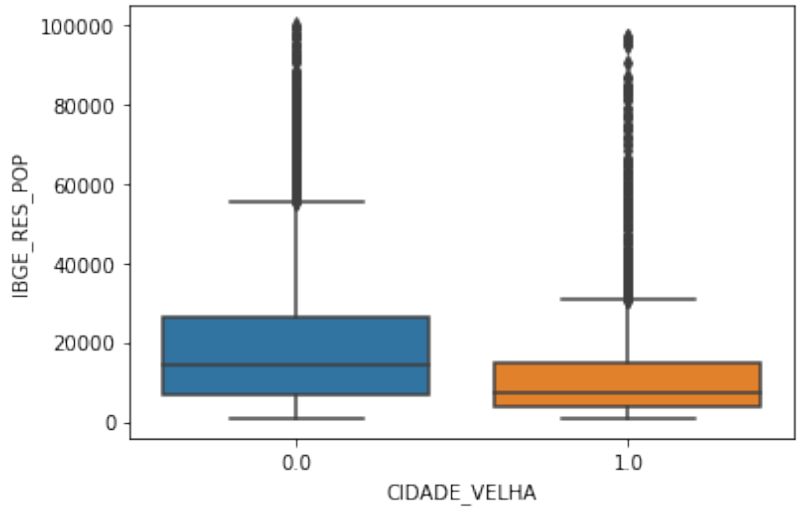

Figura 6: Boxplot população x cidade velha

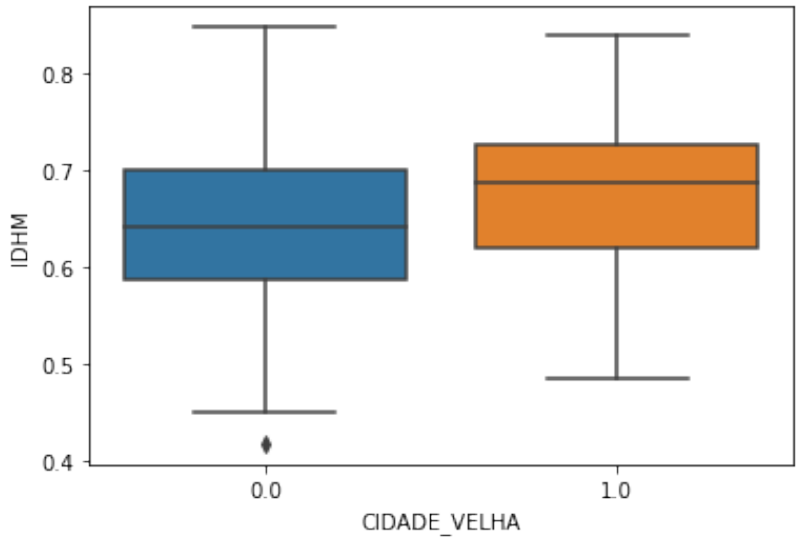

Figura 7: Boxplot população x IDH

Com isso podemos dizer que cidades com população mais velha possuem menos habitantes e também IDH maior, que cidades com população mais jovens.

3.16 Qual cidade com a maior taxa de veículos/habitantes? Esse dado pode ter alguma relação com outro dado disponibilizado?

A cidade com maior taxa de veículos/habitantes é Bom Jesus Do Norte-ES, com a taxa de 1.13 veículos/habitantes. Na tabela abaixo está presente as correlações, enquanto mais próximo de 1 maior a correlação crescente e enquanto mais próximo de -1 maior a correlação negativa.

Atributo	Correlação
IDHM Ranking 2010	-0.84
IDHM	0.84
IDHM_Renda	0.85
IDHM_Longevidade	0.73
IDHM_Educacao	0.75
LONG	-0.41
LAT	-0.72
GDP_CAPITA	0.45
Pu_Bank	0.37
IDH_BOM	0.46
veiculos/habitantes	1.00

Tabela 8: Correlação entre veículos/habitantes e outros atributos

3.17 Qual a cidade com menos habitantes que possui Uber?

A cidade com menos habitantes que possui uber é Joaçaba-SC, com 27020 habitantes.

3.18 Qual a quantidade de cidades que possuem mais motos do que carro?

A quantidade de cidades que possuem mais motos do que carro são 2364.

3.19 Podemos afirmar com confiança que cidades que possuem uber são cidades com uma área maior?

O intervalo de confiança cidades com Uber: (1593.905262040819, 3136.1287730469003);

Já o intervalo de confiança cidades sem Uber: (1380.7936210629734, 1688.1820508060785);

Logo, a hipotese de os dois conjuntos terem médias iguais não pode ser rejeitada, ou seja, são iguais.

4 Aprendizado de Máquina

Nesta 4ª etapa, o grupo ficou responsável por aplicar algoritmo(s) de aprendizagem de máquina para classificar os dados e/ou tentar prever algum acontecimento desconhecido.

Essa etapa, então, foi separada em duas subetapas melhores descritas nas subseções 4.1 e 4.2.

4.1 Clusterização

A Clusterização ou Agrupamento é um aprendizado de máquina considerado não-supervisionado, ou seja, os dados não possuem rótulos. Seu objetivo é agrupar os dados em grupos coerentes e distintos entre si.

Nessa etapa, o conjunto de dados passou por uma redução de dimensionalidade. Alguns atributos consideradas irrelevantes foram removidos e neste novo data frame, foram deixadas somente atributos considerados importantes no quesito desenvolvimento. Alguns atributos deixados foram:

PIB/Capita
Quantidade de Residentes
Despesas Municipais

Além disso, assim como anteriormente, a cidade de São Paulo foi removida desta etapa por apresentar valores extremos e prejudicar o tratamento dos dados.

O algoritmo utilizado para efetuar o agrupamento foi o **KMeans** que é referência, além de eficiente, para resolver esse tipo de problema.

O parâmetro *n_clusters* recebeu valor 4, ou seja, a execução do algoritmo agrupará os dados em 4 grupos distintos.

O grupo julgou os principais grupos de cidades como sendo:

Metrópoles,
Grandes Cidades,
Cidades Metropolitanas e
Cidades Comuns

Antes de prosseguir para os resultados, é necessário entender como cada grupo de cidade é, de modo geral.

As metrópoles são cidades de grande desenvolvimento que organizam, em torno de si, uma rede composta por cidades a ela dependentes. Como por exemplo São Paulo, Rio de Janeiro, BH, etc.

Grandes cidades são as cidades que apresentam alto desenvolvimento e não estão em regiões metropolitanas do país, diferente do grupo 3, que apresenta as cidades metropolitanas, e por fim, o grupo 4 conta com as cidades interiores e outras cidades de baixo desenvolvimento de modo geral.

Para cada um dos grupos obtidos, temos as seguintes quantidades de cidades:

- Metrópoles: 5
- Grandes Cidades: 24
- Cidades Metropolitanas: 125
- Cidades Comuns: 3773 cidades

Dentre os atributos restantes no data frame, alguns apresentaram relações interessantes de serem observadas, a fim de obter conclusões significativas.

As figuras a seguir apresentam algumas das principais relações obtidas.

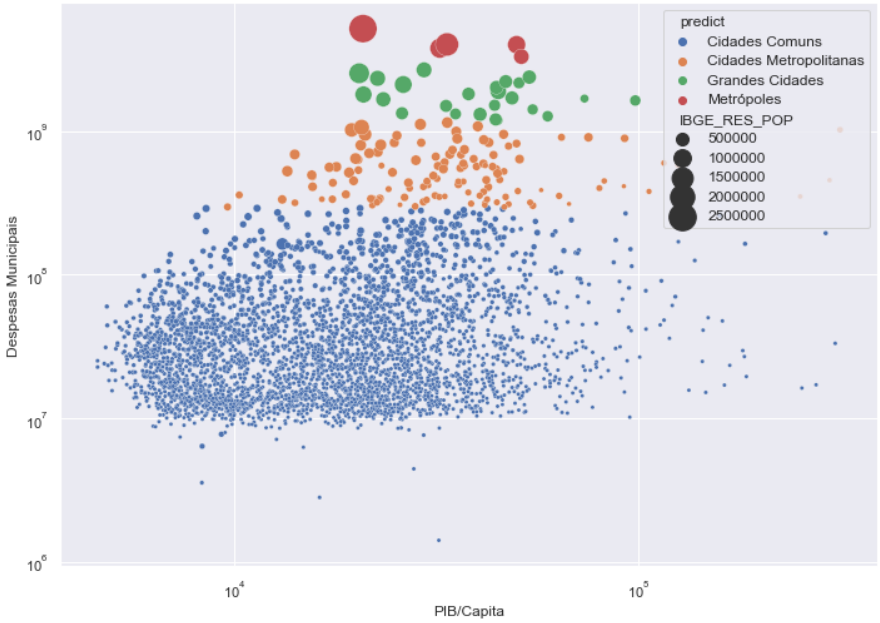

Figura 8: Gráfico PIB/Capita x Despesas Municipais

O gráfico de PIB/Capita x Despesas Municipais, permite observar que as cidades ficaram muito bem agrupadas, uma vez que é nítida a separação dos dados por meio das cores, melhores descritas na legenda do gráfico.

Vale ressaltar que os pontos no gráfico estão de acordo com a quantidade de residentes de cada cidade, por isso os pontos das metrópoles, em vermelho, por exemplo, estão significativamente maiores.

Pode-se extrair que quanto maior a cidade, maior suas despesas municipais, porém as despesas não resultam necessariamente em PIB/Capita, ou seja, ambas as características não possuem uma relação significativa.

Figura 9: Gráfico População Residente x Despesas Municipais

No gráfico da figura 9, que apresenta a relação entre População Residente x Despesas Municipais é possível notar uma clara tendência de aumento do número de habitantes de acordo com o aumento do valor das despesas municipais.

Essa observação é bem coerente, pois mais habitantes geram mais custos ao município, além, disso, é possível notar como os grupos aparecem de forma bem linear de acordo com os 2 atributos.

Por fim, pode-se também concluir que quanto maior uma cidade (metrópoles, por exemplo), mais altos são os valores, como é possível notar no gráfico.

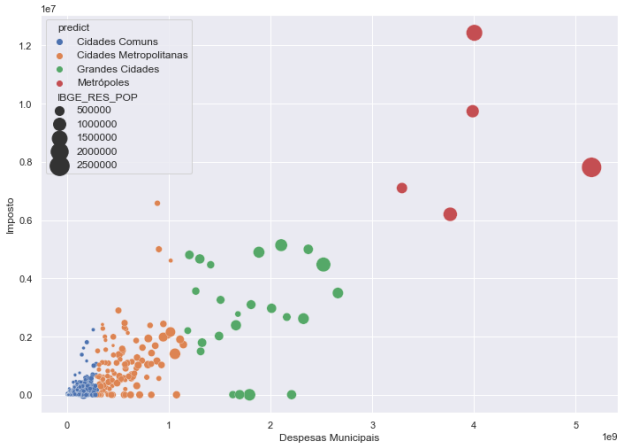

Figura 10: Gráfico Despesas Municipais x Impostos

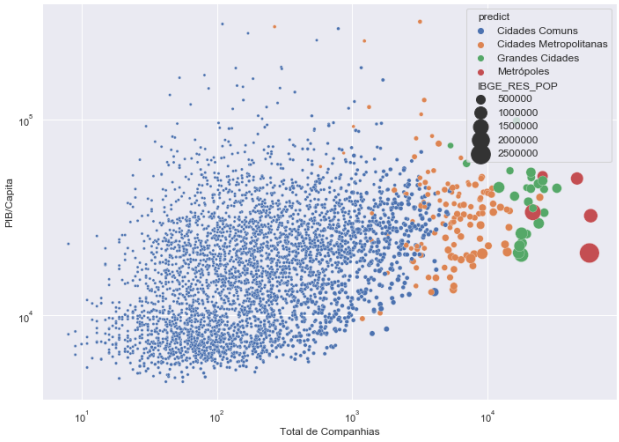

Figura 11: Gráfico Total Companhias x PIB/Capita

A figura 10 apresenta a relação entre os gastos municipais e impostos cobrados, o resultado obtido mostrou forte evidência da alta correlação entre os atributos (correlação = 0.86), ou seja, ambos atributos relacionam significativamente entre si. Então, conforme a cidade recebe mais impostos, maior os gastos municipais desta. E claro, quanto maior o desenvolvimento da cidade, maior são seus valores.

Um questionamento levantado, durante o desenvolvimento deste trabalho, pelo grupo foi e que não foi apresentado na seção 3 é: O número de empresas em uma cidade tem relação com o PIB/Capita? É possível notar no gráfico da figura 11 que a relação não apresenta um valor significativo de correlação. As cidades com maior PIB /Capita são as 'Cidades Comuns', com pontos em azul e 'Cidades Metropolitanas', como é possível observar na parte superior do gráfico, enquanto as metrópoles, apresentam muitas companhias e PIB/Capita não tão alto. A reflexão que fica é: será que cidades mais desenvolvidas tem maior concentração de produção de riquezas? O gráfico evidencia isso, ressaltando um dos diversos problemas de segregação social do Brasil.

Figura 12: Gráfico Número Residências Urbanas x Número de Companhias

Por fim, uma relação interessante de se observar também é presente entre os atributos que indicam o número de residências urbanas e número de companhias. Tem-se que o número de residências urbanas nas cidades acompanha o número de companhias locais, apresentando uma relação praticamente linear entre ambos atributos. Ou seja, quanto mais residências, mais empresas. Uma afirmação de fácil entendimento considerando o intenso e presente êxodo rural no país.

Sabe-se que o Brasil conta com cidades com alto desenvolvimento, como as metrópoles, que oferecem uma grande diversidade de oportunidades, atraindo assim mais pessoas do campo.

4.2 Predição

Nesta etapa selecionamos 3 atributos categóricos o 'CATEGORIA_TUR', o 'RURAL_URBAN' e 'GVA_MAIN', os quais permitem separar os objetos em grupos. E removemos os seguintes atributos do dataset 'CITY', 'CATEGORIA_TUR', 'STATE', 'REGIAO_TUR', 'RURAL_URBAN' e 'GVA_MAIN'.

Atributo	**Quantidade**
Nenhum	1557
D	1350
E	473
C	387
B	128
A	33

Tabela 9: Quantidade de valores de cada categoria do atributo 'CATEGORIA_TUR'

Atributo	**Quantidade**
Rural Adjacente	2121
Urbano	1080
Intermediário Adjacente	496
Rural Remoto	193
Intermediário Remoto	38
A	33

Tabela 10: Quantidade de valores de cada categoria do atributo 'RURAL_URBAN'

Atributo	**Quantidade**
Administração, defesa, educação e saúde públicas e seguridade social	1728
Demais serviços	1136
Agricultura, inclusive apoio à agricultura e a pós colheita	584
Indústrias de transformação	210
Pecuária, inclusive apoio à pecuária	129
Eletricidade e gás, água, esgoto, atividades de gestão de resíduos e descontaminação	63
Comércio e reparação de veículos automotores e motocicletas	36
Indústrias extrativas	23
Produção florestal, pesca e aquicultura	12
Construção	7

Tabela 11: Quantidade de valores de cada categoria do atributo 'GVA_MAIN'

Sendo assim, tentou-se prever cada uma das categorias, com o test size = 0.35, dos atributos e observar a acurácia, o recall, f-1 score e suporte.

Accurary(Acuracia):

Accuracy = 44.945454545454545

Precision (Precisão) e Recall(Revocação ou sencibilidade):

	precision	recall	f1-score	support
A	0.40	0.33	0.36	12
B	0.55	0.43	0.48	53
C	0.45	0.42	0.44	139
D	0.47	0.51	0.49	477
E	0.19	0.16	0.17	159
Nenhum	0.49	0.50	0.49	535
accuracy			0.45	1375
macro avg	0.42	0.39	0.41	1375
weighted avg	0.44	0.45	0.45	1375

Tabela 12: Resultado do KNN predict para o atributo 'CATEGORIA_TUR'

Accurary(Acuracia):

Accuracy = 69.89090909090909

Precision (Precisão) e Recall(Revocação ou sencibilidade):

	precision	recall	f1-score	support
Intermediário Adjacente	0.33	0.25	0.28	186
Intermediário Remoto	0.00	0.00	0.00	12
Rural Adjacente	0.73	0.90	0.81	742
Rural Remoto	0.20	0.02	0.03	60
Urbano	0.79	0.66	0.72	375
accuracy			0.70	1375
macro avg	0.41	0.36	0.37	1375
weighted avg	0.66	0.70	0.67	1375
weighted avg	0.44	0.45	0.45	1375

Tabela 13: Resultado do KNN predict para o atributo 'RURAL_URBAN'

Accurary(Acuracia):

Accuracy = 60.07272727272727

Precision (Precisão) e Recall(Revocação ou sencibilidade):

	precision	recall	f1-score	support
Administração, defesa, educação e saúde públicas e seguridade social	0.62	0.84	0.71	603
Agricultura, inclusive apoio à agricultura e a pós colheita	0.34	0.20	0.25	209
Comércio e reparação de veículos automotores e motocicletas	0.00	0.00	0.00	9
Construção	0.00	0.00	0.00	3
Demais serviços	0.66	0.70	0.68	394
Eletricidade e gás, água, esgoto, atividades de gestão de resíduos e descontaminação	0.00	0.00	0.00	22
Indústrias de transformação	0.09	0.01	0.02	76
Indústrias extrativas	1.00	0.20	0.33	5
Pecuária, inclusive apoio à pecuária	0.00	0.00	0.00	48
Produção florestal, pesca e aquicultura	0.00	0.00	0.00	6
accuracy			0.60	1375
macro avg	0.27	0.20	0.20	1375
weighted avg	0.52	0.60	0.55	1375

Tabela 14: Resultado do KNN predict para o atributo 'GVA_MAIN'

5 Conclusão

Dessa forma, podemos concluir que o trabalho foi bastante desafiador, principalmente por ter sido divido em 5 etapas distintas, e também divertido de ser executado, uma vez que pudemos aplicar na pratica grande parte do conteúdo aprendido durante a disciplina. Além disso, com o trabalho pudemos conhecer melhor o nosso país e alguns padrões de cidades existentes no território nacional.

Contudo, foi uma experiência bastante produtiva e positiva para o desenvolvimento do grupo na disciplina, uma vez que nos mostra, por meio da prática, o uso das técnicas de tratamento de dados. Sendo importante ressaltar a evolução na organização de desenvolvimento de projetos e desenvolvimento de códigos eficientes e de qualidade.