
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

11 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 preparação 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 subseções 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 determinada 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 prestar 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 subseção 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 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:

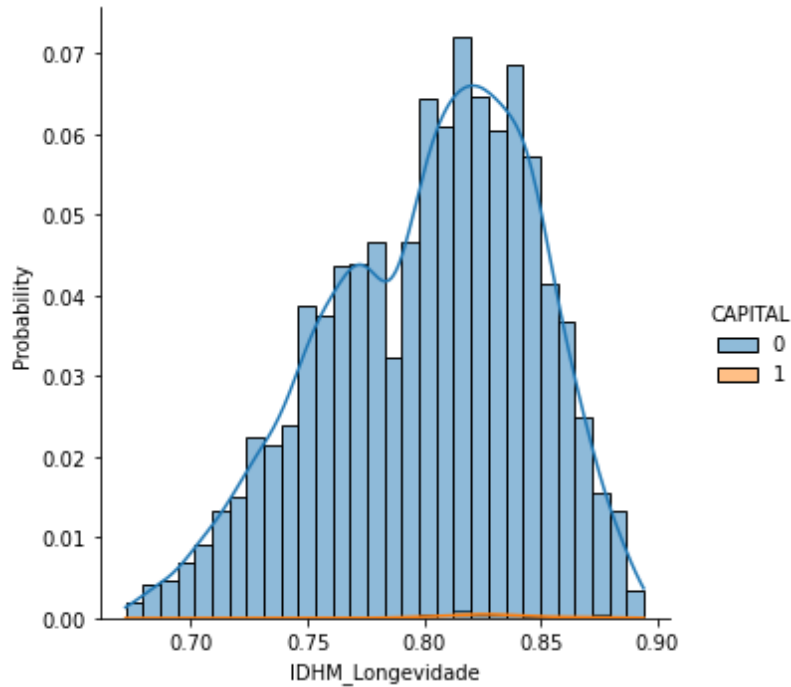


Figura 1: Histograma da Expectativa de Vida

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.

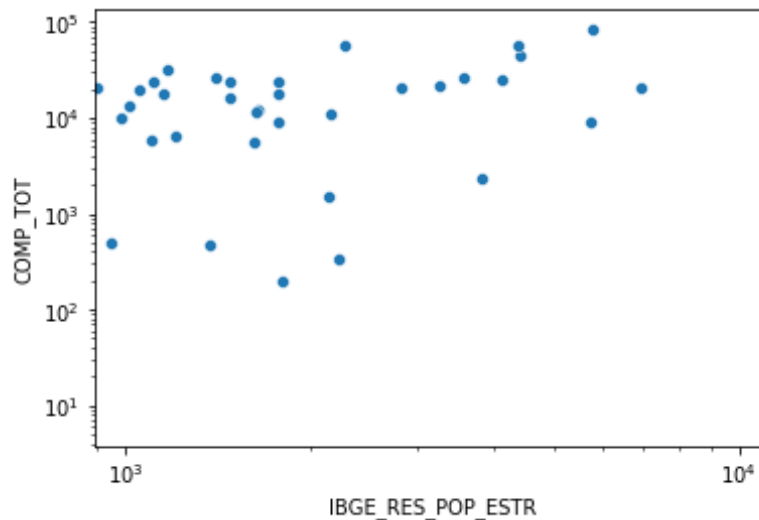


Figura 2: Gráfico de Pontos entre Quantidade de Companhias e Quantidade de Estrangeiros Residentes em cada cidade

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 subseção 3.7, a cidade de São Paulo foi removida da análise.

O gráfico abaixo apresenta a relação entre ambos atributos.

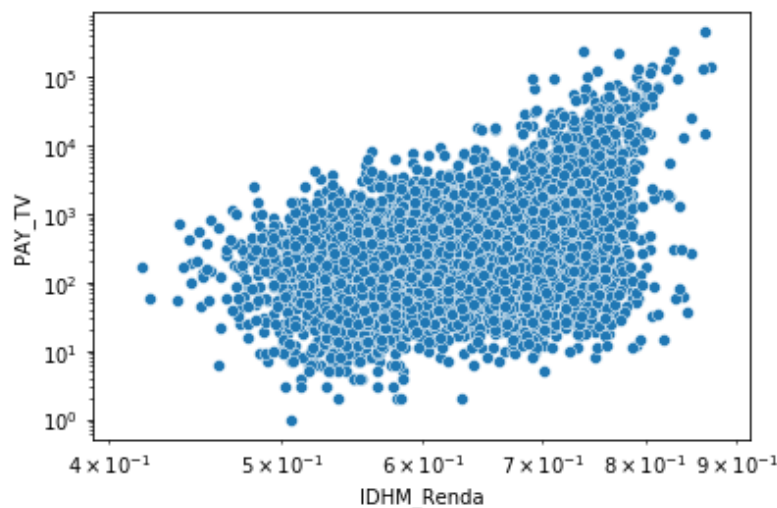


Figura 3: Gráfico de Pontos entre Quantidade de Assinaturas de TV a Cabo e Renda de cada Cidade

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 5×10^{-1} e 8×10^{-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 subseção 3.7, a cidade de São Paulo foi removida da análise.

O gráfico abaixo apresenta a relação entre ambos atributos.

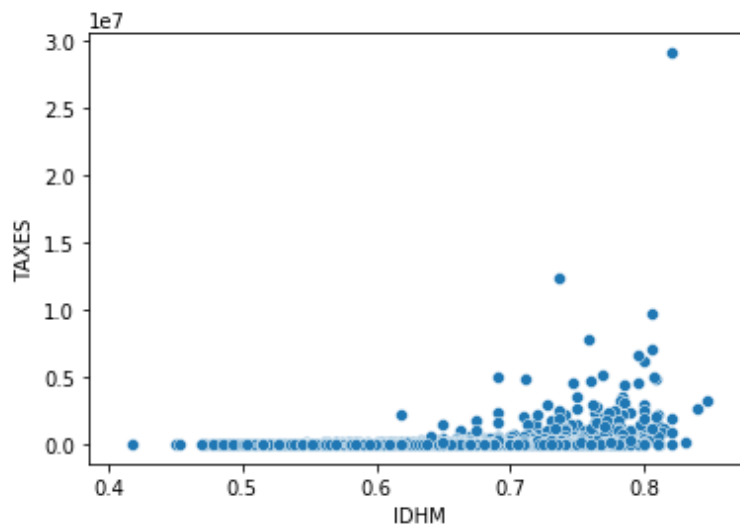


Figura 4: Gráfico de Pontos entre Carga Tributária e IDH

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.

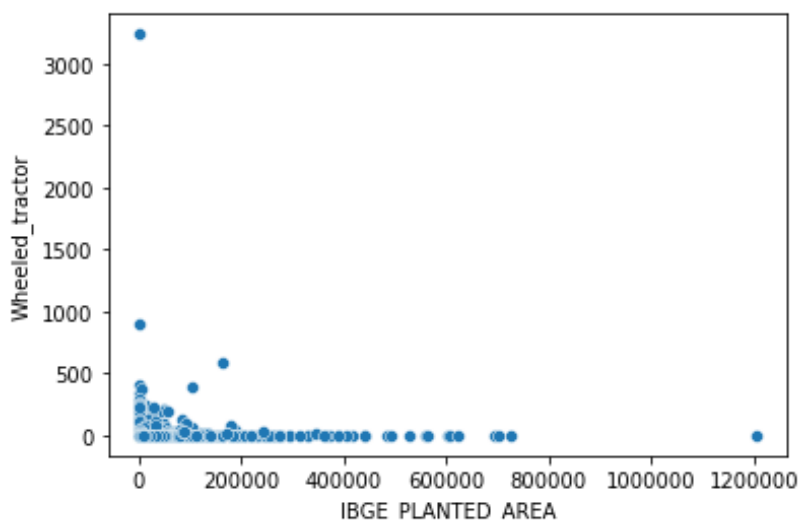


Figura 5: Gráfico de Pontos entre Quantidades de Rodas de Trator e Área Plantada

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 as 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 apresentadas 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.

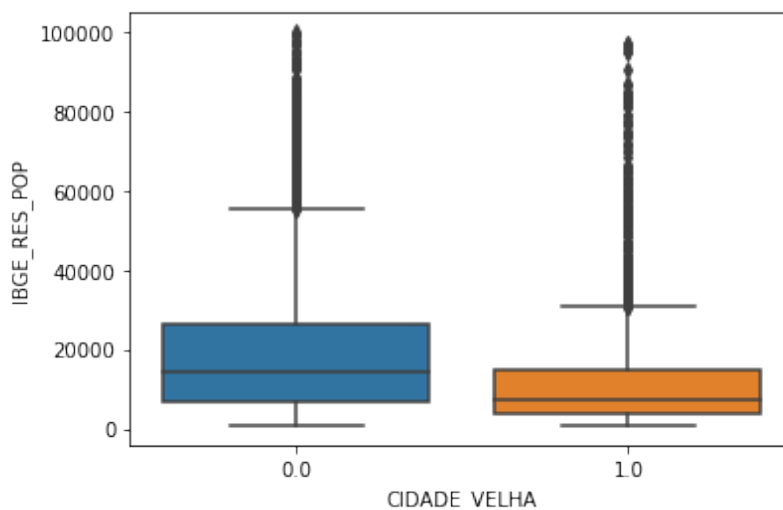


Figura 6: Boxplot população x cidade velha

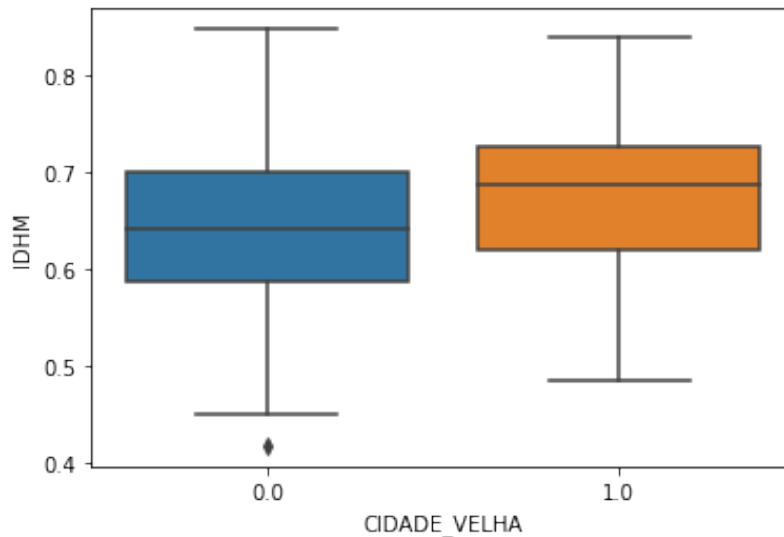


Figura 7: Boxplot população x IDH

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 hipótese de os dois conjuntos terem médias iguais não pode ser rejeitada, ou seja, são iguais