IFSP - Instituto Federal de Educação, Ciência e Tecnologia
Câmpus São Paulo

Leonardo Naoki Narita SP3022498

Análise Exploratória de Dados (EDA)
Estudo de Caso

São Paulo - SP - Brasil

27 de Janeiro de 2022

IFSP - Instituto Federal de Educação, Ciência e Tecnologia
Câmpus São Paulo

Leonardo Naoki Narita SP3022498

Análise Exploratória de Dados (EDA)
Estudo de Caso

Trabalho desenvolvido no curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo como requisito parcial para a conclusão da disciplina de Programação Funcional.

Professor: Guilherme Werneck de Oliveira

IFSP - Instituto Federal de Educação, Ciência e Tecnologia
Câmpus São Paulo

Tecnologia em Análise e Desenvolvimento de Sistemas

PFUEL - Programação Funcional

São Paulo - SP - Brasil

27 de Janeiro de 2022

Sumário

1 Introdução

Nesse capítulo, são abordados conceitos fundamentais que embasam o desenvolvimento do projeto.

1.1 Definição de EDA

A Análise Exploratória de Dados (do inglês, "Exploratory Data Analysis", conhecida também pela sigla "EDA") é a aplicação de um conjunto de técnicas que visam analisar uma população ou amostra de dados (conjunto de dados dentro de um mesmo contexto).

A EDA visa identificar conclusões significativas dentro dessa massa de dados, podendo ser padrões ou outliers (inconsistências de dados).

1.2 Ciclo de Vida da EDA

Dado que há uma massa de dados importada, deve-se aplicar técnicas de consistência de dados, tratando os dados não existentes, tornando os dados manipuláveis para cálculos, entre outros.

Após esse processo, o ciclo de vida da EDA baseia-se em três passos:

1.2.1 Gerar Questionamentos

O primeiro passo é gerar questionamentos sobre os dados que estão dispostos, criando variáveis e agregações.

Para gerar questionamentos, uma questão é fundamental a ser indagada: "Que tipo de variação ocorre com as variáveis?", onde a variável é um atributo mensurável com um valor atribuído, e a variação é a mudança dos valores da variável.

1.2.2 Modelas Dados

O segundo passo é organizar os dados de modo que seja visualmente fácil de analisar para que, assim possa encontrar conclusões.

Nesse momento, são buscados modos de relacionar variáveis em modos visuais (gráficos, tabelas) a fim de conseguir buscar respostas aos questionamentos realizados anteriormente.

Não há um processo exato há ser seguido para organizar os dados, tendo em vista que cada caso é um caso. Contudo, há métodos amplamente utilizados que podem auxiliar nesse processo. Entre os métodos existentes, podem ser utilizados removação de outliers e/ou substituição de dado (aplicando medidas de posição central, tais como média, moda, mediana, máximo, desvio padrão, mínimo, entre outros).

1.2.3 Buscar Conclusões

O terceiro e último passo é analisar as conclusões obtidas, podendo chegar a novos questionamentos e outros pontos de vistas sobre os mesmos dados.

Uma vez que tendo as conclusões necessárias, é possível realizar aplicação de estatísticas ou técnicas de inteligência artificial.

2 Apresentação do Conjunto de Dados

Nesse capítulo, é definida a massa de dados a ser importado no projeto, bem como suas variáveis e detalhes.

2.1 Escolha dos Dados

O conjunto de dados escolhidos para o projeto é o "Brazilian Cites", onde há 5573 cidades brasileiras, disponível em: https://www.kaggle.com/crisparada/brazilian-cities.

2.2 Dicionário de Dados

Nessa seção, é definido os principais atributos que serão trabalhados e analisados pela EDA.

Os principais atributos das cidades do Brasil, dispostas na planilha, são:

Quadro 1 – Definição dos Principais Atributos

Dicionário de Dados - Cidades do Brasil		
Atributo	**Descrição**	**Valores**
CITY	Nome da cidade	
STATE	Nome do estado	
CAPITAL	Indica se a cidade é a capital do estado	1 (SIM) ou 0 (NÃO)
IBGE_RES_POP	População residente na cidade	
IBGE_RES_POP_BRAS	População brasileira residente na cidade	
IBGE_RES_POP_ESTR	População estrangeira residente na cidade	
IDHM	Índice de Desenvolvimento Humano (IDH)	
IDHM_Renda	Índice de renda pelo IDH	
IDHM_Longevidade	Índice de longevidade pelo IDH	
IDHM_Educacao	Índice de educação pelo IDH	

Fonte: Kaggle

Na coluna "Valores", as células não preenchidas representam variáveis contínuas (conjunto de valores abertos, ou seja, possuindo uma quantidade de variáveis que são variáveis). Enquanto, nessa mesma colunas, as células preenchidas representam variáveis categóricas (com um conjunto de dados fechados e pré-determinados).

3 Desenvolvimento da EDA

Nesse capítulo, é aplicado os três passos do ciclo de uma EDA em um contexto prático aplicando o conjunto de dados "Brazilian Cities".

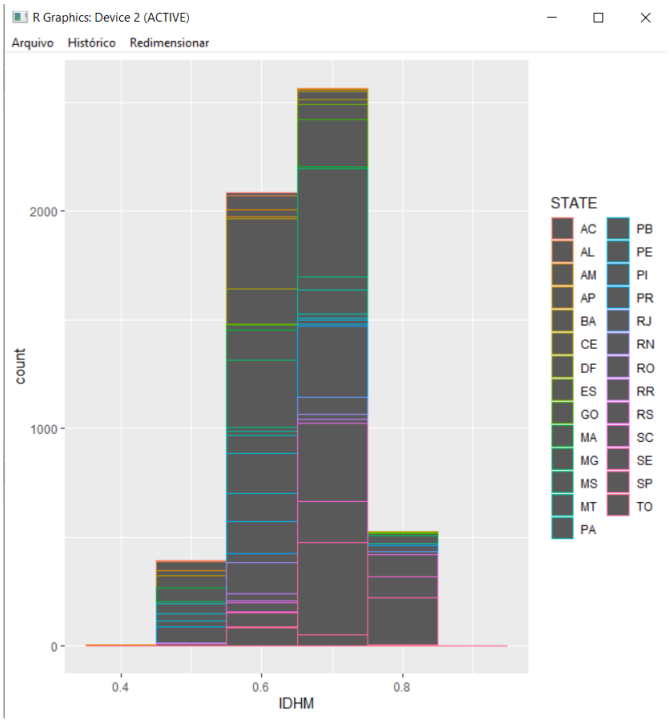
3.1 Contextualização

O IDH (Índice de Desenvolvimento Humano) é um índice que avalia a qualidade de vida em um determinado local, sendo mensurado a partir de 3 fatores: Longevidade, Educação e Renda. No Brasil, o órgão responsável por avaliar o IDH é o IBGE (Instituto Brasileiro de Geografia e Estatística).

Sendo um parâmetro importante, o IDH é capaz de influenciar tomadas de decisões, tais como quais setores aplicar investimento público pelos políticos, se é vantajoso mudar-se para morar nesse local, e se é viável a iniciativa privada investir nesse local para atuar.

Vide na figura Figura 1 a distribuição do IDH nas cidades do Brasil.

Figura 1 – Distribuição do IDH nas cidades brasileiras



Fonte: O autor

Visualizando esse gráfico, é perceber que não há cidades brasileiras com IDH menor do que 0.4, nem com IDH maior do que 0.8.

3.2 Tratamento de Dados

Para tornar os dados mais consistentes, foi necessário filtrar os dados cujo IDH não seja NA (Não Aplicável).

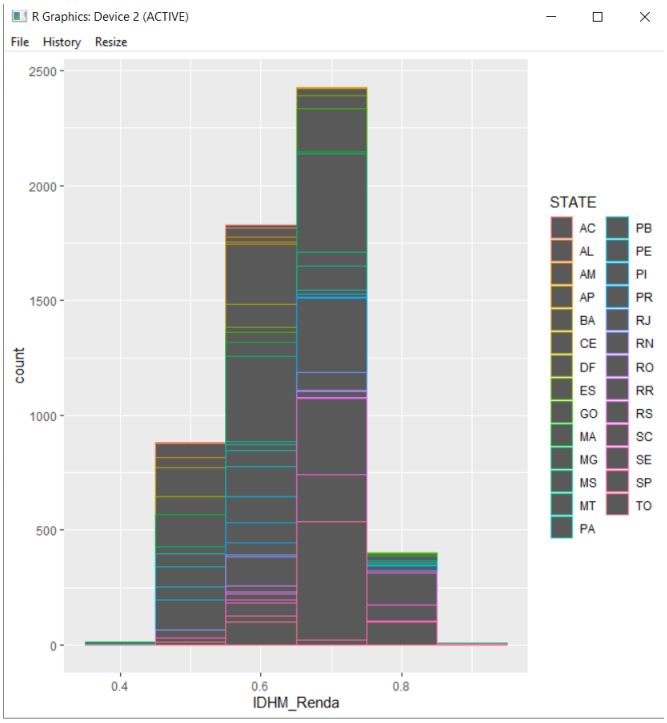
3.3 Problematização

Com isso, surgem-se duas questões:

1. "Qual é a razão da discrepância entre os dados de IDH?";
2. "Por que não existe IDH maior que 1?".

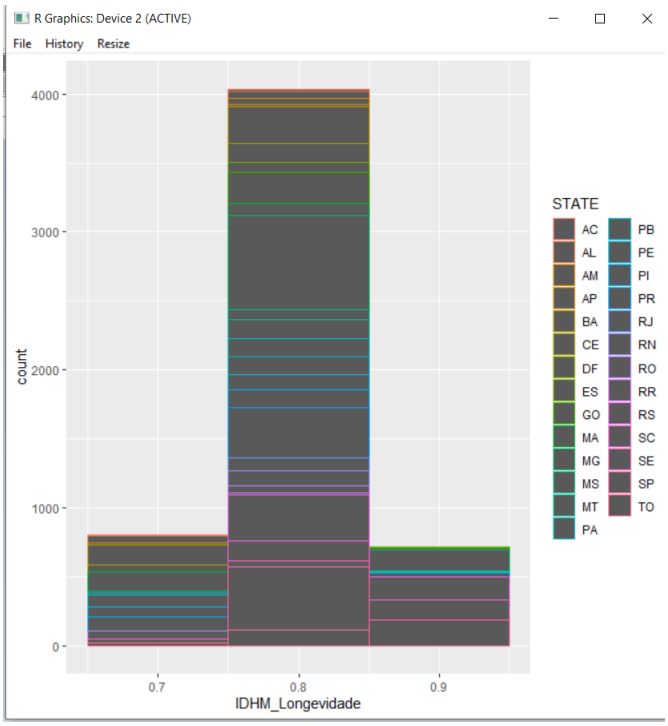
Para responder essas questões, foram feitas análises da variação de cada critério do IDH. Na Figura 2, é analisada a variação de renda. Na Figura 3, é analisada a variação de longevidade. E, por fim, na figura Figura 4, é analisada a variação de educação.

Figura 2 – Distribuição do IDH de Renda nas cidades brasileiras



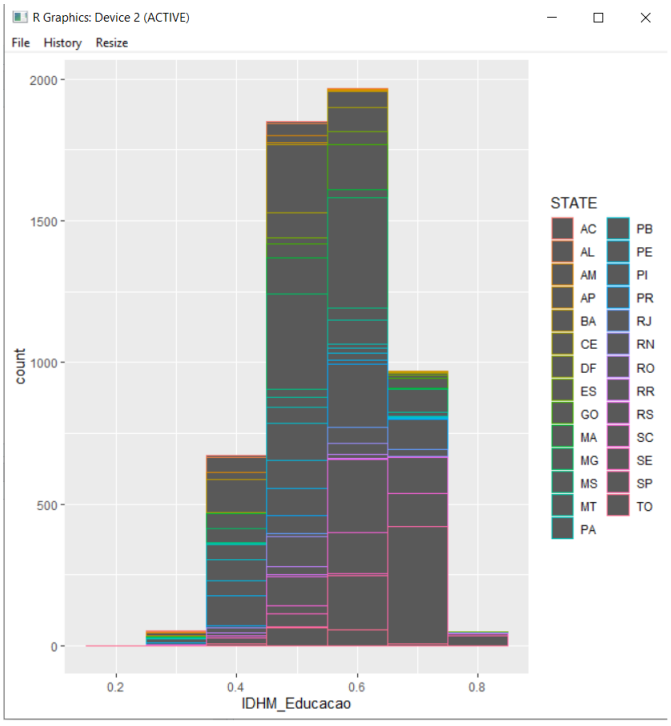
Fonte: O autor

Figura 3 – Distribuição do IDH de Longevidade nas cidades brasileiras



Fonte: O autor

Figura 4 – Distribuiçãa do IDH de Educação nas cidades brasileiras



Fonte: O autor

Dado os gráficos acima, foram realizados cálculos para conseguir realizar conclusões

mais assertivas. Foram realizados cálculos de média simples e desvio padrão (desigualdade entre os dados, sendo 0 mais próximo da igualdade e 1 mais próximo da desigualdade).

Calculando a média de cada critério do IDH, foi constatado que a renda teve a nota 0.65, a longevidade teve a nota 0.80 e a educação teve a nota 0.55.

Agora calculando o desvio padrão de cada critério do IDH, foi constatado que a renda teve a nota 0.08, a longevidade teve a nota 0.04 e a educação teve a nota 0.09.

4 Conclusão

Nesse projeto, foi-se iniciado por definições de EDA, seus métodos e definições básicas. Logo após, continuando com a importação de dados, definindo os principais atributos disponíveis na massa de dados disponibilizados para que, com isso, seja possível trabalhar com a metodologia de EDA por completa.

Com isso, geramos questionamentos iniciais e buscamos visualizar os dados a fim de respondê-las. Nessa seção, iremos concluir os raciocínios e analisar se foi possível responder aos questionamentos ou se serão formados novos questionamentos.

Analisando os gráficos fornecidos, percebe-se que a discrepância de IDH é devido, principalmente, à variação de educação, que possui a menor média e o maior desvio padrão entre os estados, resultando no maior grau de variação entre as cidades do Brasil (entre 0.3 e 0.8).

Também é possível responder que não há dados de IDH maiores que 1, visto que é uma padronização a nível global.