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

20 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

20 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 e devidamente tratada (tendo-os consistentes), 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 relacionamentos anteriormente. Não há um processo exato há ser seguido, tendo em vista que cada caso é um caso. Contudo, há métodos que podem auxiliar nesse processo.

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.

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

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 Caso 1: Análise do IDH

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.

Com isso, surge-se uma questão: "Será que as cidades mais populosas são as que possuem o maior IDH?".

...

3.2 Caso 2: Análise de Estrangeiros

No mesmo contexto do caso 1, outra questão deve ser feita: "Um estrangeiro, de modo geral, tende a ir à cidades com maiores índices de IDH ou cidades mais populosas para morar?".

...

4 Conclusão