Análise Exploratória, Visualização de Dados e Enriquecimento de Dados

Desenvolvimento de Projetos Integrador

29/09/2021

1. Introdução

A finalidade da ***Análise Exploratória de Dados e Visualização (AED&V)*** é examinar os dados previamente à aplicação de qualquer técnica estatística. Desta forma o ***Cientista de Dados*** consegue um entendimento básico de seus dados e das relações existentes entre as variáveis analisadas. Para que qualquer aplicação de ***Machine Learning*** ou ***Deep Learning*** tenham sucesso é fundamental uma compreensão dos dados e correção de anomalias que possam existir.

As tabelas e gráficos a seguir são exemplos de exploração e visualização de dados.

Variable	Mean	Std Dev	Minimum	Maximum	N	N Miss	Lower Quartile	Median	Upper Quartile
NU_IDADE	22.0203777	7.1247618	7.0000000	83.0000000	1068424	724	18.0000000	19.0000000	23.0000000
NU_NOTA_CN	493.7970775	75.5839644	298.2000000	869.6000000	1065367	3781	434.9000000	486.4000000	545.9000000
NU_NOTA_CH	538.0041788	83.4692876	315.1000000	850.4000000	1067190	1958	474.5000000	542.6000000	602.5000000
NU_NOTA_LC	523.0988405	66.1977152	299.6000000	793.1000000	1058146	11002	480.2000000	526.9000000	569.9000000
NU_NOTA_MT	526.5660121	105.9486192	317.7000000	996.1000000	1053469	15679	443.0000000	507.7000000	594.5000000
NU_NOTA_REDACAO	559.4155431	158.6428051	120.0000000	1000.00	1065981	3167	460.0000000	560.0000000	640.0000000
NOTA_MEDIA	528.1742580	78.0013599	302.8800000	852.5800000	1065139	4009	470.1000000	518.2000000	576.3800000

Tabela 1: Estatísticas de resumo das variáveis de notas dos ENEMs 2017-2019

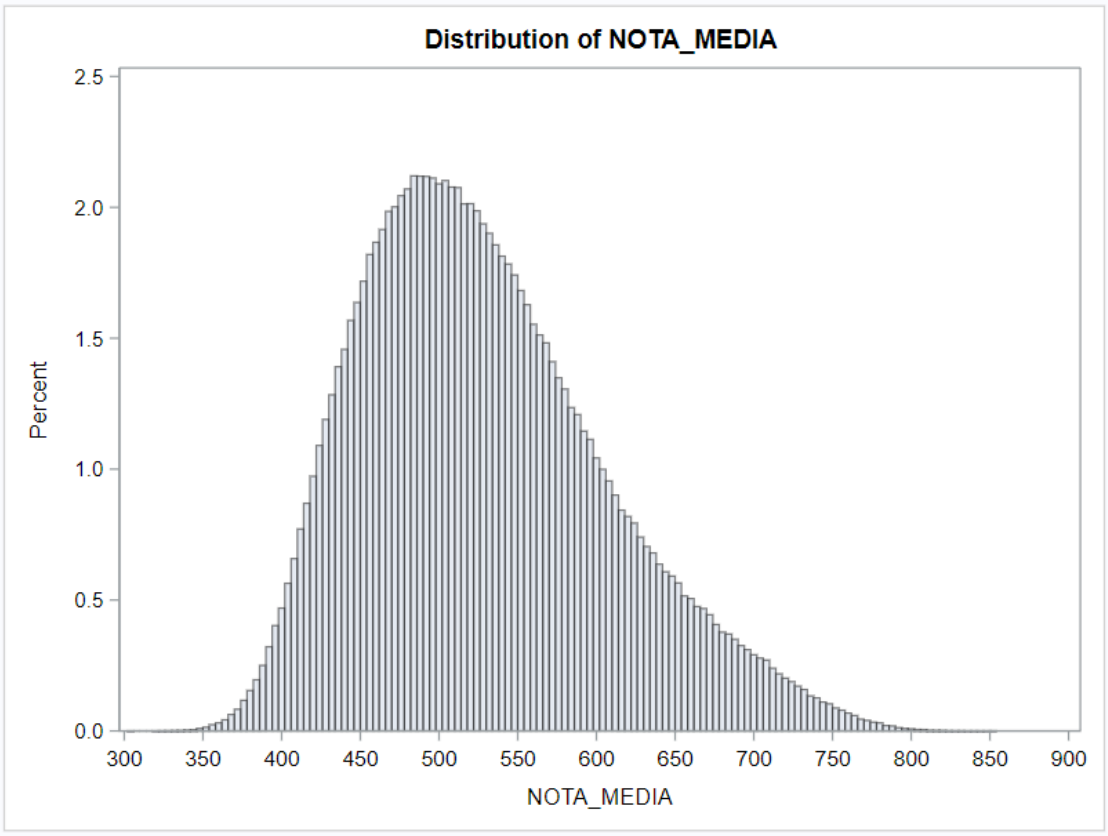

Gráfico 1: Histograma da variável Nota Média dos ENEMs 2017-2019

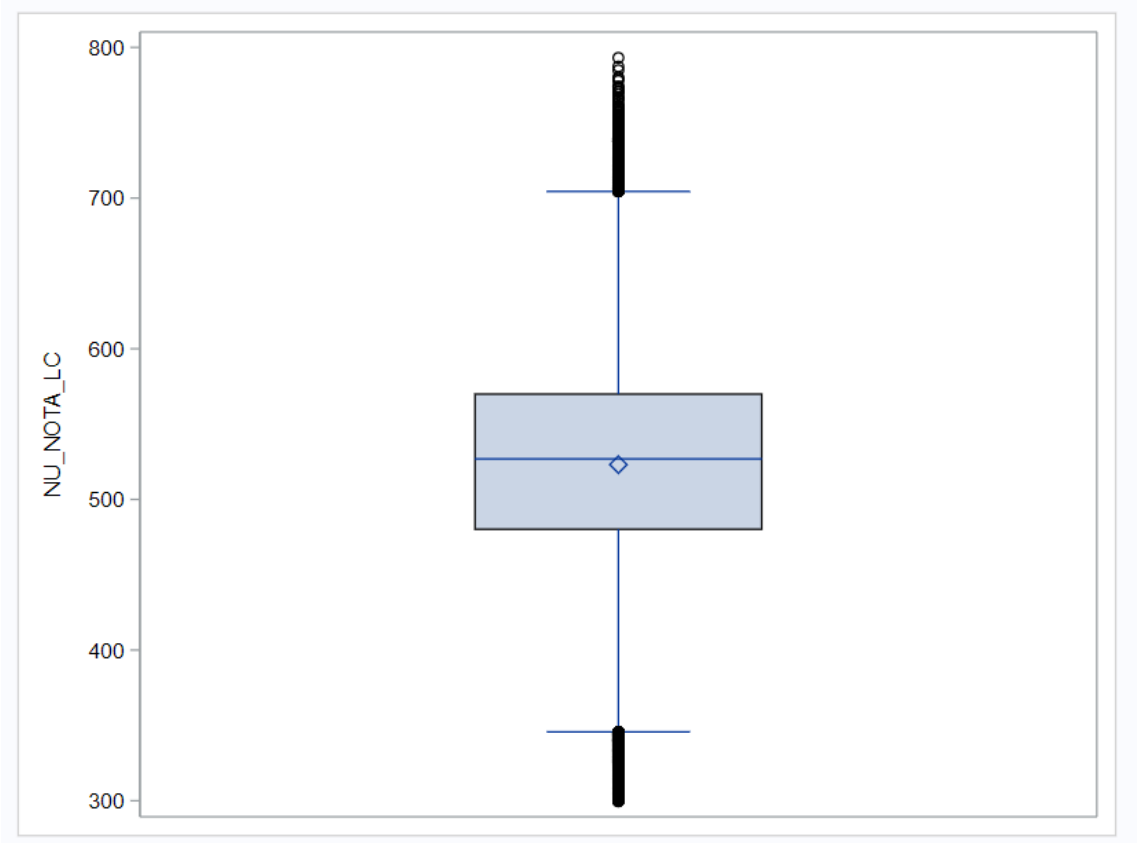

Gráfico 2: Boxplot da variável Linguagens e Códigos dos ENEMs 2017-2019

Observando a tabela e os gráficos anteriores constata-se anomalias nos dados que precisam ser analisados e corrigidos, se for o caso.

2. Desenvolvimento do trabalho – parte 1

Com base no conjunto de dados disponibilizado para análise e utilizando *qualquer solução de software* faça:

a) Utilize a planilha com uma amostra de dados disponibilizada para você. Esta planilha está identificada com seu nome na pasta "**X:\Dados\CIA036-2021.2\ENEM**" (**Drive CIA)** no Google Sala de Aula (**Tema 05**, item 01. Bases de Dados para atividades de Análise Exploratória e Enriquecimento de Dados (Pré-processamento);
b) Desenvolva uma ***Análise Exploratória de Dados e Visualização*** de **todas** as variáveis qualitativas e quantitativas, *que forem pertinentes*;
c) Faça um relatório apresentando suas análises e identificando ***todas as anomalias encontradas nos dados***. Você pode utilizar o ***template em Latex*** disponibilizado para os exercícios da disciplina ou desenvolver em outro editor de textos. Sempre siga os tópicos do template do Latex;

3. Desenvolvimento do trabalho – parte 2

Com base no ***relatório desenvolvido na parte 1 do trabalho*** e utilizando *qualquer solução de software* faça:

a) Desenvolva procedimentos para ***correção (imputação)*** de todas as anomalias que você identificou;
b) Refaça a sua ***Análise Exploratória de Dados e Visualização*** de **todas** as variáveis qualitativas e quantitativas, **agora com os dados corrigidos** (imputados);
c) Compare os resultados;
d) ***Complemente o seu relatório iniciado na parte 1*** deste trabalho, detalhando os seus procedimentos para correção (imputação) dos dados;
e) Faça a conclusão do seu trabalho;
f) Faça no ***máximo 15 slides*** sobre os pontos mais relevantes do seu trabalho para apresentação do desenvolvimento e conclusão do seu trabalho. Caso você não se sinta confortável para apresentação na sala de aula, grave um vídeo com a sua apresentação e faça o upload no Google Sala de Aula,