Teoria do Aprendizado Estatístico

Tarefa 1: Base de Dados

Banco de dados sobre a economia global (World Economic Dataset)

Iris Durante Alvim do Nascimento

Waleska Mayara Silva Reis

1. Introdução

A aplicação da Teoria do Aprendizado Estatístico é importante para compreender como os modelos podem extrair padrões e realizar previsões a partir de conjuntos de dados. Os conceitos envolvidos englobam probabilidade, variabilidade, funções estatísticas e a relação entre variáveis, sendo fundamentais para garantir análises consistentes e confiáveis.

2. Desenvolvimento teórico – Banco de Dados

O Banco de Dados escolhido foi o Word Economic Dataset, importado do Kaggle. Este banco fornece dados sobre a econômica global e, apesar, de não ter data definida o dataset foi escolhido pelo fato de que essa base é completa e possui dados qualitativos e quantitativos (discretos e contínuos). Além disso, a partir deste banco é possível realizar diferentes tipos de análises, como comparar quantos países diferentes tem o mesmo idioma pátrio.

O Ahmad Firman, engenheiro elétrico e autor do dataset, coletou os dados por meio de web scraping e, depois, organizou os dados pelo Github, deixando-os prontos para análise. Não há afirmações diretas que este banco de dados tenha sido usado para trabalhos acadêmicos.

Como afirma o autor do dataset:

> Os dados de ambas as fontes foram limpos e transformados usando Python (pandas, solicitações) em um pipeline ETL. Campos JSON aninhados (por exemplo, moedas, idiomas, latlng) foram compactados em strings simples ou valores numéricos. Listas vazias ou valores ausentes foram padronizados como Nenhum. Finalmente, os dois conjuntos de dados foram mesclados em um usando o nome do país como chave primária.

Em resumo, Ahmad Firman destaca os processos necessários para a realização das análises de dados do dataset em questão. No Kaggle, o autor disponibiliza os dados já preparados para análise, a fim de facilitar a consulta em diferentes áreas. Contudo, neste trabalho optou-se por utilizar a tabela em formato CSV contendo os dados brutos, com o objetivo de realizar uma

análise própria, desenvolvida desde o início, em vez de recorrer diretamente à versão já tratada disponibilizada na plataforma.

3. Metodologia

Neste trabalho, a resolução utilizou dados para instruir o banco de dados com variáveis definidas, que foram manipuladas em Excel e R. Embora os dados sejam destinados a regressão linear, aplica-se métodos de estatística para normalizá-los e, assim, poderem ser utilizados para média e mediana, além de fornecer valores em caractere.

Entretanto, na avaliação da dupla, mesmo que o dataset possua licença autorizada pelo Kaggle e a plataforma seja reconhecida internacionalmente, inclusive fora da área de Ciência de Dados, considera-se que o material não é totalmente confiável, uma vez que foram identificados espaços vazios que precisaram ser anulados durante o processamento no código.

Para este trabalho foram definidos alguns passos para realizar as análises estatísticas:

A. Leitura dos dados a partir de read.csv() ou read.table();
B. Resumo dos dados utilizando a função summary (média, mediana, mínimo, máximo e
C. valores ausentes (NA));
D. Medidas de tendência central e dispersão: Foram calculadas a média, mediana, moda, variância e desvio padrão, de modo a compreender tanto o comportamento central dos dados quanto sua variabilidade.
E. Tabelas de frequência:

    - Para as variáveis qualitativas, elaborou-se uma tabela representando o número de países por continente;
    - Para as variáveis quantitativas, construiu-se uma tabela indicando a quantidade de países de acordo com faixas populacionais.

F. histograma do PIB: Foi elaborado um histograma para examinar a distribuição do Produto Interno Bruto (PIB) entre os países da base de dados.
G. Distribuição de variáveis adicionais: A variável taxa de desemprego também foi analisada por meio de histogramas, permitindo visualizar a forma de sua distribuição, utilizando a distribuição GAMMA.
H. Relação entre variáveis qualitativas: Foi desenvolvida uma tabela de contingência (tabela cruzada) para investigar a relação entre idiomas e regiões (continentes).
I. Associação entre variáveis quantitativas: A relação entre o PIB e a população foi estudada, evidenciando a dependência entre os dois indicadores.
J. Associação entre variáveis qualitativas: A ligação entre região e idioma foi analisada a partir de técnicas de cruzamento de variáveis categóricas.

K. Associação entre variáveis qualitativas e quantitativas: Por fim, foi construído um Boxplot do PIB por região, permitindo comparar a distribuição do PIB entre diferentes continentes e identificar possíveis discrepâncias.

3.1 Descrição das variáveis

O conjunto de dados foi construído a partir da combinação de duas fontes principais:

1. TradingEconomics.com – responsável pelos indicadores econômicos:

   - GDP (PIB)
   - GDP.Growth (Crescimento do PIB)
   - Interest.Rate (Taxa de Juros)
   - Inflation.Rate (Inflação)
   - Jobless.Rate (Taxa de Desemprego)
   - Gov..Budget (Orçamento do Governo)
   - Debt.GDP (Dívida/PIB)
   - Current.Account (Conta Corrente)
   - Population (População)

2. Informações em nível de país, que incluem:

   - name (Nome do país)
   - currency (Moeda)
   - capital (Capital)
   - languages (Idiomas)
   - region (Região)

No total, o dataset possui 173 linhas (países) e 14 variáveis, sendo 9 delas de indicadores econômicos e 5 de características gerais de cada país.

Além disso as variáveis do tipo foram codificadas categoricamente (labl in coding) com a função “factor”:

```
name <- factor(dados$name)
currency <- factor(dados$currency)
capital <- factor(dados$capital)
languages <- factor(dados$capital)
region <- factor(dados$region)
```

3.2 Medidas de tendência central e dispersão

A tabela apresenta as medidas estatísticas descritivas das variáveis econômicas do conjunto de dados. Observa-se, por exemplo, que o PIB (GDP) possui média muito elevada (637,5) e grande dispersão, com desvio-padrão de 2703,9, o que indica forte variação entre os países. Já o crescimento do PIB (GDP.Growth) tem média de 1,1, mediana próxima (0,7) e desvio-padrão relativamente baixo (3,8), sugerindo maior homogeneidade. A taxa de juros (Interest.Rate) apresenta média de 7,8 e desvio-padrão de 8,1, mostrando variação moderada. A inflação (Inflation.Rate) também é variável, com média de 8,3 e desvio-padrão mais alto (21,3). A taxa de desemprego (Jobless.Rate) é mais estável, com média de 7,3 e menor dispersão (6,0). O orçamento do governo (Gov..Budget) mostra tendência negativa, com média de -2,8, indicando déficits em boa parte dos países. A dívida em relação ao PIB (Debt.GDP) tem média de 60,9, mediana de 54,9 e desvio-padrão de 38,9, refletindo diferenças significativas entre economias. A conta corrente (Current.Account) tem média próxima de zero (-0,54), mas alta variabilidade (desvio-padrão 8,9). Por fim, a população (Population) apresenta média de 45,5 milhões, mediana de 10,5 milhões e desvio-padrão de 155,9, mostrando que há países muito populosos que elevam a média em relação à mediana:

A data.frame: 9 × 6

Variavel	Media	Mediana	Moda	Variancia	Desvio_Padrao
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
GDP	637.5086705	72.000	2.00	7.311242e+06	2703.930867
GDP.Growth	1.1156436	0.700	0.80	1.459824e+01	3.820764
Interest.Rate	7.8022222	5.250	2.15	6.673770e+01	8.169314
Inflation.Rate	8.3486628	3.350	4.30	4.565489e+02	21.367006
Jobless.Rate	7.3630994	5.200	5.20	3.679634e+01	6.065999
Gov..Budget	-2.8228313	-3.175	-3.20	1.819649e+01	4.265735
Debt.GDP	60.9458788	54.900	69.00	1.520910e+03	38.998840
Current.Account	-0.5452695	-1.000	1.30	7.869843e+01	8.871213
Population	45.5700578	10.580	0.78	2.430383e+04	155.896870

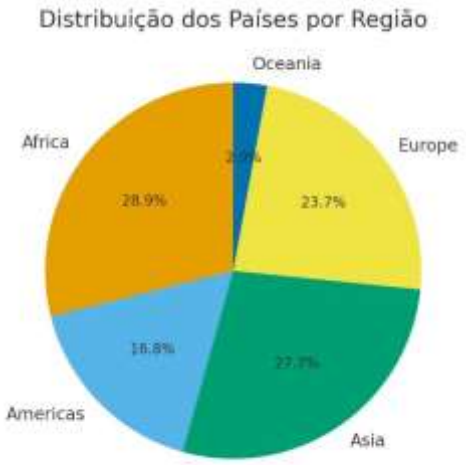
3.3 Tabela de Frequência

Tabela de frequência de números para cada continentes:

	Região	Frequência	Relativa...
1	Africa	50	28.90
2	Americas	29	16.76
3	Asia	48	27.75
4	Europe	41	23.70
5	Oceania	5	2.89

A partir do gráfico foi feita a seguinte análise:

1. África possui a maior frequência absoluta com 50 ocorrências, correspondendo a 28,90% do total. Isso indica que quase um terço dos dados analisados estão concentrados neste continente.
2. Ásia vem em seguida, com 48 ocorrências, equivalente a 27,75% do total. Apesar de ter uma frequência ligeiramente menor que a da África, sua representatividade é praticamente equivalente, sugerindo que esses dois continentes juntos concentram mais da metade dos dados.
3. Europa apresenta 41 ocorrências, correspondendo a 23,70%, mostrando uma presença significativa, mas inferior à da África e Ásia.
4. Américas têm 29 ocorrências, representando 16,76%, o que indica uma participação menor em comparação com os continentes mencionados anteriormente.
5. Oceania é o continente com menor frequência, com apenas 5 ocorrências, equivalente a 2,89% do total, mostrando que os números correspondentes a esta região são escassos no conjunto de dados analisado.



- A distribuição não é uniforme, sendo mais concentrada na África e Ásia, seguidas da Europa, Américas e, por último, Oceania.
- A soma das frequências relativas é aproximadamente 100%, confirmando que os dados apresentados representam a totalidade das ocorrências.
- Este tipo de tabela é útil para identificar padrões regionais e pode auxiliar em análises estatísticas descritivas ou em decisões baseadas na representatividade de cada continente.

Tabela de frequência da população (quantas países tem determinada população):

	Classe	Frequencia	Frequencia.Relativa....
1	[0,10]	83	47.98
2	(10,50]	60	34.68
3	(50,100]	15	8.67
4	(100,500]	13	7.51
5	(500,1.5e+03]	2	1.16

Antes de concluir a análise é preciso detalhar a estrutura desta tabela: cada linha da tabela representa uma classe de população, a frequência absoluta de países nessa faixa populacional e a frequência relativa associada:

1. Classe (Faixa populacional): a tabela divide a população dos países em intervalos. Esses intervalos são expressos em formato de intervalo [0,10] significa de 0 até 10 milhões de habitantes.
2. Frequência: a frequência indica quantos países estão incluídos em cada faixa populacional:

    - Para a faixa [0,10] milhões de habitantes, existem 83 países.
    - Para a faixa [10,50] milhões de habitantes, 60 países se enquadram.
    - Para a faixa [50,100] milhões, são 15 países.
    - Para a faixa [100,500] milhões, existem 13 países.
    - E para a faixa [500, 1.5e03] milhões (ou seja, de 500 milhões a 1.5 bilhões de habitantes), há apenas 2 países.

3. Frequência Relativa: a frequência relativa é expressa em percentual e indica a proporção de países dentro de cada faixa em relação ao total de países. A soma das frequências relativas deve ser 100%, representando o total da população distribuída nas classes. Para cada classe, temos a seguinte distribuição:

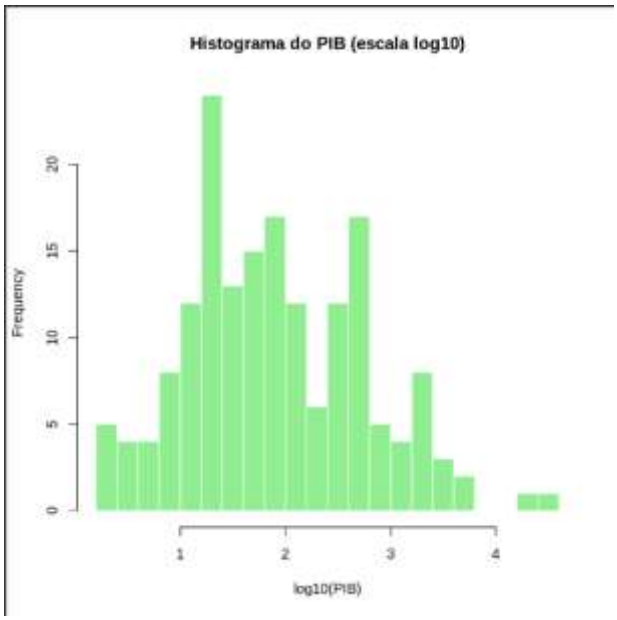
    - [0,10] milhões: 47,98% dos países.
    - [10,50] milhões: 34,68% dos países.
    - [50,100] milhões: 8,67%.
    - [100,500] milhões: 7,51%.
    - [500,1.5e03] milhões: 1,16%.

Contudo, a análise dos dados:

A maioria dos países tem uma população relativamente baixa, com quase 48% deles possuindo até 10 milhões de habitantes. Em seguida, a faixa de 10 a 50 milhões engloba cerca de 35% dos países. As classes com populações maiores, como 50 a 100 milhões e 100 a 500 milhões, têm uma porcentagem significativamente menor de países. As faixas de população

extremamente alta, como a de 500 milhões a 1.5 bilhões, incluem apenas uma pequena quantidade de países, evidenciando que são poucas as nações com populações muito grandes.

3.4 Histograma



É apresentado um histograma que representa a distribuição do Produto Interno Bruto (PIB) de uma série de países, mas de uma maneira transformada, utilizando uma escala logarítmica do PIB, ou seja, os valores do PIB foram convertidos usando o logaritmo na base 10. Isso é frequentemente feito para lidar com a grande variação nos valores do PIB entre os países.

Estrutura do Gráfico:

- Eixo X (log10(PIB)): Este eixo mostra os valores do PIB em escala logarítmica. O eixo vai de 1 a 4, o que indica uma variação no PIB entre 10^1 (10 milhões) e 10^4 (10.000 milhões, ou 10 bilhões);
- Eixo Y (Frequency): O eixo vertical representa a frequência de países cujos valores de PIB se enquadram em cada intervalo no eixo X;
- Barras: Cada barra do histograma representa um intervalo de valores do logaritmo do PIB e a altura da barra reflete quantos países possuem um PIB dentro dessa faixa logarítmica.

Análise:

a) Distribuição do PIB: O gráfico mostra que a maioria dos países têm um PIB que está em torno de valores menores, representados pela área entre log10(PIB) = 1 e 2. Isso indica que muitos países têm PIBs em torno de 10 milhões a 100 milhões

de unidades monetárias. Essa concentração de países com PIBs menores é comum, pois a maioria das economias globais são menores em termos de PIB.

b) Desvio para PIBs mais altos: A partir de log10(PIB) = 2 a 3, ou seja, PIBs entre 100 milhões e 1.000 milhões (1 bilhão), a distribuição tende a ser mais dispersa, com algumas barras mais altas, indicando que um número significativo de países possui PIBs nesta faixa. Esses países representam economias maiores, mas ainda assim não são tão numerosos quanto os países de PIB mais baixo.
c) Outliers e concentração nas faixas baixas e médias: O gráfico mostra que poucos países estão nas faixas de PIB muito alto, acima de log10(PIB) = 3 (acima de 1 bilhão de PIB). Essa distribuição assimétrica (com uma concentração de países na parte inferior do gráfico) sugere que há uma grande disparidade entre as economias mais fracas e as mais fortes.
d) Piada das barras extremas: A barra mais alta no gráfico está concentrada em torno de log10(PIB) = 1, indicando que a maioria dos países tem PIBs mais baixos (entre 10 milhões a 100 milhões). Além disso, as barras à direita do gráfico, na região do PIB mais alto, são muito baixas, indicando que há uma quantidade muito pequena de países com PIB extremamente alto.

3.5 Associação de duas variáveis quantitativas

Correlação entre PIB e População

```
    Pearson's product-moment correlation

data:  dados$GDP and dados$Population
t = 8.9621, df = 171, p-value = 5.361e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4544535 0.6589432
sample estimates:
      cor
0.5653223
```

A análise da correlação entre o PIB e a população dos países, utilizando o coeficiente de correlação, revelou um valor de aproximadamente 0,5653. Esse resultado indica uma correlação moderada e positiva entre as duas variáveis, sugerindo que, de maneira geral, países com maior população tendem a apresentar PIBs mais elevados.

O teste de significância estatística apresentou um p-valor extremamente baixo ($5{,}361 \times 10^{-16}$), o que indica que essa relação observada tenha ocorrido por acaso. Além disso, o intervalo de confiança de 95% para o coeficiente de correlação está entre 0,4545 e 0,6589, reforçando a consistência do resultado.

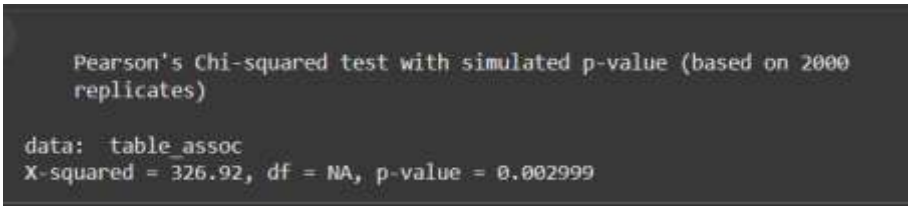
Associação de PIB por continentes

Essa imagem apresenta uma análise do PIB de diferentes continentes, com várias estatísticas, incluindo valores mínimos, máximos, média e quartis para cada um. Os continentes listados são África, Américas, Ásia, Europa e Oceania:

Os dados do PIB dos continentes mostram que a África apresenta o menor PIB médio (54,98 bilhões de dólares) e variação de 2 a 400 bilhões. As Américas têm o maior PIB médio (1.324 bilhões) e extrema disparidade. A Ásia possui o PIB máximo mais alto (18.740 bilhões) e uma média de 836,6 bilhões. A Europa apresenta PIB médio de 654,1 bilhões. Já a Oceania, com média de 412,2 bilhões.

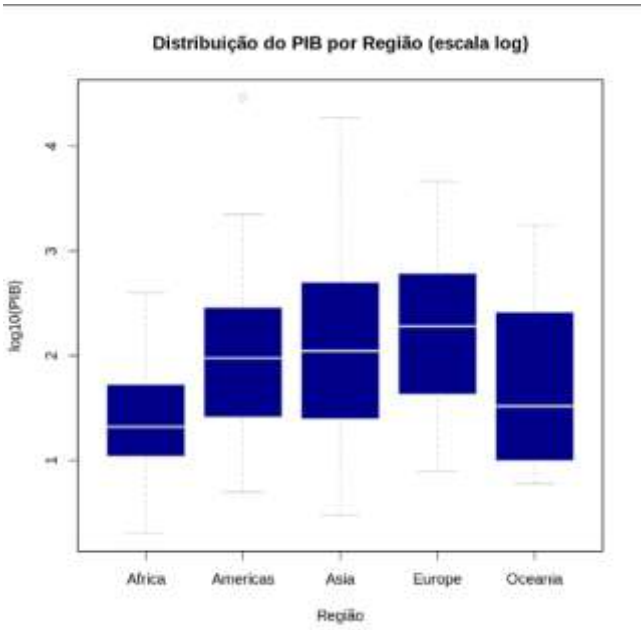
3.6 Associação de variáveis qualitativas

associação entre as variáveis continente e idioma



A partir de um teste qui-quadrado, conduzido para verificar a associação entre as variáveis "continente" e "idioma", seu resultado foi de 326,92, acompanhado de um p-valor de 0,002999. Este p-valor, abaixo do nível de significância usual de 0,05, indica que a hipótese nula de independência entre as variáveis pode ser rejeitada. Em outras palavras, a distribuição dos idiomas entre os continentes analisados não é aleatória, sugerindo que o continente exerce influência sobre os idiomas falados em cada região.

3.7 BoxPlot



BoxPlot é um gráfico de caixa que ilustra, nesse caso, a distribuição do PIB por região, utilizando uma escala logarítmica (log10) para a variável PIB. A análise da distribuição revela que, em termos gerais, as regiões com os maiores valores médios de PIB são a Ásia e a Europa, seguidas pelas Américas. A Oceania apresenta uma distribuição de PIB mais concentrada, com valores tendendo para o intervalo inferior, enquanto a África possui uma dispersão considerável, refletindo a grande variação econômica entre os países do continente.

3.8 Tabela de contingência

Idiomas x continentes

	Africa	Americas	Asia	Europe	Oceania	total coluna
Afrikaans	2	0	0	0	0	2
Albanian	0	0	0	2	0	2
Arabic	12	0	13	0	0	25
Aymara	0	2	0	0	0	2
Chinese	0	0	2	0	0	2
Dutch	0	1	0	1	0	2
English	16	8	6	3	4	37
French	14	1	0	3	1	19
German	0	0	0	5	0	5
Greek	0	0	0	2	0	2
Guaraní	0	2	0	0	0	2
Portuguese	3	1	1	1	0	6
Romanian	0	0	0	2	0	2
Russian	0	0	3	1	0	4
Spanish	0	14	0	1	0	15
total linha	47	29	25	21	5	127

A tabela apresentada corresponde a uma análise de contingência entre idiomas e continentes, destacando a distribuição de falantes de diferentes línguas ao redor do mundo. Os idiomas estão organizados em linhas, enquanto os continentes são representados por colunas.

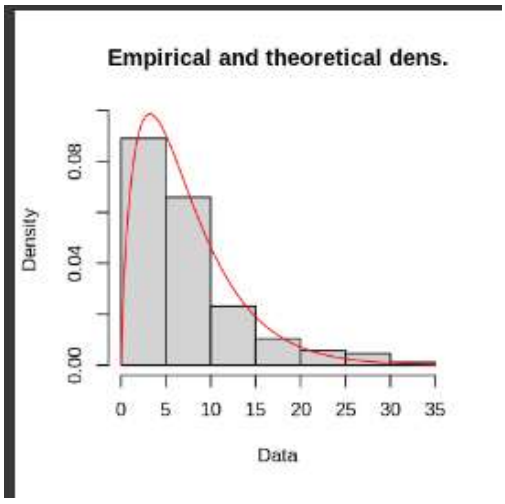
O idioma árabe destaca-se pela sua presença em três continentes: África, Ásia e Europa, com 12 falantes na África, 13 na Ásia e nenhum na Europa, totalizando 25 falantes. O inglês possui a maior dispersão, com 16 falantes na Europa, 1 na Ásia, 4 nas Américas e 1 na Oceania, somando 37 falantes no total. O francês também possui uma distribuição considerável, com 14 falantes nas Américas e 3 na Europa, totalizando 19 falantes.

Outros idiomas, como o alemão e o chinês, têm presença restrita a um ou dois continentes, com o alemão sendo falado principalmente na Europa (5 falantes) e o chinês com 2 falantes na Ásia e 2 na Europa. O português aparece principalmente na África, com 3 falantes, e nas Américas, com 1 falante, refletindo sua distribuição geográfica restrita a esses dois continentes.

Idiomas como guaraní e romeno apresentam falantes limitados a apenas um continente: o guaraní com 1 falante nas Américas e o romeno com 2 falantes na Europa. O espanhol tem uma forte presença nas Américas, com 14 falantes, mas com praticamente nenhum falante na Europa e em outros continentes.

A análise se limita aos idiomas com pelo menos 2 falantes em cada continente, como o árabe, o inglês e o francês, o que permite identificar as línguas de maior relevância geográfica.

3.9 Distribuição da taxa de desemprego



A análise da distribuição da taxa de desemprego mostra que a maior parte dos dados se concentra em taxas baixas, principalmente entre 0% e 5%, indicando que, em períodos normais, o desemprego tende a ser reduzido. Observa-se também uma cauda longa à direita, refletindo casos ocasionais de desemprego elevado, típicos de crises ou recessões. Esse

padrão caracteriza uma distribuição assimétrica positiva, comum em fenômenos econômicos. A linha de densidade teórica acompanha bem a distribuição observada, e, para modelar esse comportamento, foi utilizada a distribuição Gama, por ser a mais adequada para esse tipo de distribuição, capturando tanto a concentração em valores baixos quanto a dispersão em valores mais altos.

4. Conclusão

Este trabalho aplicou a Teoria do Aprendizado Estatístico para analisar o *World Economic Dataset*, um conjunto de dados sobre a economia global. A metodologia incluiu a leitura e resumo dos dados, cálculo de medidas de tendência central e dispersão, criação de tabelas de frequência e análise de associação entre variáveis.

O Produto Interno Bruto (PIB) apresentou grande variação, com alguns países muito ricos elevando a média. A população também variou bastante: a maioria dos países possui menos habitantes, o que torna a média (45,5 milhões) muito maior que a mediana (10,5 milhões). Já a taxa de desemprego e o crescimento do PIB mostraram menor dispersão.

A distribuição de países por continente foi desigual, com África e Ásia concentrando mais da metade da amostra. A análise da população por faixas revelou que a maioria dos países tem até 10 milhões de habitantes, enquanto poucos possuem populações muito grandes, de 500 milhões a 1,5 bilhões.

A correlação entre PIB e população foi moderada e positiva, indicando que países mais populosos tendem a ter PIB mais alto. O teste qui-quadrado mostrou uma relação significativa entre continente e idioma, rejeitando a hipótese de distribuição aleatória. Já o Boxplot do PIB por região indicou que Ásia e Europa apresentam maiores medianas, enquanto a África apresenta grande disparidade econômica.

Em resumo, o estudo permitiu identificar padrões e compreender melhor as relações econômicas e demográficas globais. Embora o conjunto de dados tenha algumas lacunas, os resultados destacam como a análise estatística é útil para revelar a complexidade da economia mundial e das populações dos países.