Coleta e Análise Exploratória de Dados Secundários

Instruções:

a. **Objetivo**: O objetivo desta atividade é levantar, explorar, analisar, descrever e interpretar um conjunto de dados secundários. Esse tipo de dados são essenciais para o início de uma investigação, ou seja, para mapear o cenário em questão e auxiliar nas tomadas de decisão do gestor. Como objetivo colateral está o uso de pacote computacional em aplicações de estatística descritiva. Será importante também a apresentação adequada de relatório contendo a análise e suas conclusões. As normas serão aquelas aplicáveis a relatórios técnicos

b. **Problema proposto**: Considere que você é um pesquisador do Governo Federal, tendo sido encomendado à turma 2017 um estudo para identificar possíveis fatores influentes ou determinantes da mortalidade infantil entre todas as Microrregiões do Brasil (conforme classificação do IBGE) e de alguns municípios de MG. Há dois tios de mortalidade infantil a serem consideradas: a mortalidade neonatal, que ocorre entre 0 e 27 dias (incluindo os menores de um mês, com idade ignorada) e a pós-natal, que ocorre entre 1 e 11 meses (incluindo menores de 1 ano, ignorada a idade). A intenção é passar a uma abordagem preventiva, resolvendo as causas dos problemas a fim de diminuir custos no longo prazo.

c. **Bancos de dados**: O Ministério da Saúde indicou algumas variáveis que podem ser importantes para o estudo e os bancos de dados de onde elas podem ser retiradas. Recomendou-se que sejam utilizados os dados de 2010 a fim de facilitar sua pesquisa, já que o último Censo foi realizado nesse ano.

   i. **DataSUS**:

   - Quantidade de médicos.
   - Quantidade de famílias atendidas pelo Programa de Assistência à Saúde (PACS).
   - Quantidade de agentes de atenção básica à saúde do Programa de Assistência à Saúde (PACS).
   - Quantidade de leitos hospitalares.
   - Quantidade de equipamentos.
   - Quantidade de nascimentos.
   - Quantidade de óbitos.
   - Quantidade de produção hospitalar.
   - Quantidade de produção ambulatorial.
   - Cobertura vacinal, que é a razão de doses aplicadas pela população alvo.
   - Casos de AIDS de pessoas com 12 anos ou mais (total e por sexo do respondente)

   ii. **Censo de 2010 (IBGE)* ou IPEAData:**

   - População
   - População acima de 60 anos, inclusive (idade ≥ 60)
   - População entre 15 e 59 anos ($15 \leq$ idade ≤ 59)
   - População abaixo de 15 anos, exclusive (idade < 15)
   - Taxa de analfabetismo - 18 anos ou mais
   - % de 15 a 17 anos com fundamental completo
   - % de 18 a 24 anos com fundamental completo (2010)
   - % de 18 anos ou mais com fundamental completo (2010)
   - % de 25 anos ou mais com fundamental completo (2010)

- % de 18 anos ou mais com médio completo (2010)
- % de 25 anos ou mais com médio completo (2010)
- % de 25 anos ou mais com superior completo (2010)
- Expectativa de anos de estudo (2010)
- % de 6 a 14 anos no fundamental com 2 anos ou mais de atraso (2010)
- % de 18 a 24 anos no fundamental regular seriado
- Decil 20% da renda da população
- Decil 40% da renda da população
- Decil 60% da renda da população
- Decil 80% da renda da população
- Média de anos de escolaridade
- Média de escolaridade de pessoas com 15 ou mais anos (total e por sexo do respondente).
- Porcentagem de pessoas com acesso a água e esgoto
- Porcentagem de pessoas com acesso a energia elétrica
- PIB per capita
- Renda média de domicílios
- PIB total
- PIB do setor de agropecuária
- PIB do setor de serviços
- PIB do setor industrial
- IDH
- Índice Gini

(*) Acessar o Atlas do Desenvolvimento Humano do Brasil. Notar que as variáveis vêm por municípios, devendo ser somadas para agregá-las por microrregiões.

d. **Construção de variáveis**: O Ministro da Saúde solicitou especial atenção às variáveis listadas abaixo que devem ser calculadas a partir de variáveis dos bancos de dados citados no item ():
   - Taxa de Urbanização = população urbana dividida pela população total.
   - Densidade demográfica = população total dividida pela superfície do território(**)
   - Razão de dependência = (população abaixo de 15 anos, exclusive mais população acima de 60 anos, inclusive) divididas por população entre 15 e 59 anos.
   - Porcentagem da participação do PIB da agropecuárias.
   - Porcentagem da participação do PIB de serviços.
   - Porcentagem da participação do PIB da indústria.

(**) Superfície do território pode ser encontrada no site do IBGE (aba cidades).

e. **Divisão de tarefas:** Dada a dimensão do trabalho, cada aluno será responsável para obter e analisar os dados de parte das microrregiões do Brasil ou de municípios de MG. Encontre na página da disciplina sua quota de participação no trabalho.

f. **Coleta e análise dos dados:** Siga as seguintes recomendações para montagens do conjunto de dados e para análise dos dados:
   - Monte o conjunto de dados com todas as variáveis obtidas do banco de dados e com as variáveis calculadas.
   - Lembre-se de sempre que achar pertinente criar variáveis *per capita*. Ou seja, há variáveis que possuem clara influência do número da população, tal como *Quantidade de leitos* (considere a criação da variável

Quantidade de leitos *per capita* = Quantidade de leitos dividida por população)

- Sempre que possível compare seus resultados ou análise com os valores da variável de interesse agregados para Brasil e Minas Gerais. Use-os como *baseline*.
- Um dos objetivos do trabalho é identificar possíveis relações entre as variáveis explicativas e entre as variáveis explicativas e as respostas. Considere a discretização de algumas variáveis para auxiliar sua busca.
- Espera-se que sua análise utilize intensamente de visualização gráfica das variáveis (individual ou bivariada ou multivariada) e de medidas descritivas que possam conduzir futuras inferências sobre parâmetros populacionais importantes (mesmo aqueles que você não sabe como construir). Espera-se que você crie os gráficos apropriados aos dados e comente sobre qualquer coisa de interesse que veja nos gráficos, em particular se observar algum comportamento inesperado que possa fazê-lo sentir-se pouco à vontade para aplicar testes formais de inferência estatística.
- Você é encorajado a olhar sempre os dados e estabelecer conjecturas a serem posteriormente verificadas formalmente como consequência dessa análise exploratória. Assim, um dos resultados esperados é você apresentar essas suposições sobre o problema as quais você tenha percebido empiricamente.
- Redução de dimensionalidade é um estágio importante em quase todas as análises de dados. Há vários procedimentos estruturados com essa finalidade. É importante que, baseando-se em sua análise exploratória você identifique aquelas variáveis que poderiam ser mais importantes para explicar as respostas (taxas de mortalidade neonatal e pós-natal).
- Incentiva-se que sejam apresentadas variáveis adicionais que você julgue importantes para a solução do problema proposto. Tenha também em mente a pergunta sobre quais variáveis você introduziria no estudo para compreender melhor a população em questão.

g. **Conjunto de dados:** O conjunto de todos os dados devem ser entregues em arquivo de formato aberto (extensão .txt ou .csv). Devem ser apresentadas todas as variáveis retiradas dos bancos de dados e as variáveis construídas por meio de operações com as variáveis retiradas. Devem ser acrescentadas as variáveis: município (quando for o caso), microrregião, mesorregião, estado.

h. **Relatório**: A análise deverá ser apresentada na forma de relatório técnico, compreendendo o problema proposto, sua modelagem e resolução, bem como os resultados e sua análise. Isto é, o trabalho deverá ganhar um título. Um pequeno resumo vem em seguida, para que o eventual leitor tenha uma ideia geral do conteúdo do trabalho. O corpo do trabalho é o próximo, dividido em três partes clássicas: introdução, desenvolvimento e conclusão. Por fim, deverão ser apresentadas as referências bibliográficas (livros, revistas, relatórios, etc.) que foram consultadas. Indique também o software utilizado.

Além disso, o relatório deverá conter:

i. Identificação dos tipos de variáveis encontradas no Banco de Dados.

ii. Construção de tabelas e gráficos das variáveis de interesse. Organize as saídas de seu pacote e não apresente resultados que não utilizará ou comentará.
iii. Cálculo, onde aplicável, das medidas descritivas para cada variável (medidas de tendência central, de posição, de dispersão, etc.) individual ou por grupo.
iv. Breve comentário sobre os resultados obtidos.
v. Análise das relações e associações mais relevantes entre as variáveis apresentadas.
vi. Um resumo das principais conclusões a respeito dos dados apresentados, a partir da interpretação dos resultados obtidos.
vii. Indicação do tipo de levantamento adicional que poderia ser efetuado no sentido de melhorar as condições de interpretação dos dados.
viii. Apresentação das observações ou sugestões a respeito do presente trabalho.

i. **Avaliação:** O trabalho será avaliado com base nos seguintes quesitos:

Quesito		*Percentual*
Dados	Montagem do conjunto	10%
Resolução	Uso de estatística	40%
	Análise dos resultados	30%
Apresentação	Apresentação/Relatório	20%

j. **Recomendações adicionais:** O foco deverá sempre ser a análise das características dos dados, assim como o tratamento adequado dos valores relevantes do conjunto de dados selecionado, de maneira a extrair informações e a alicerçar conclusões.