Visualização de dados multidimensionais do Censo Nacional de 2010 e seleção de Modelos Representativos

Palavras-Chave: Clusterização, Visualização Dados, Dados Multidimensionais

Autores:

Marcos Vinícius Vieira Takahashi da Silva – FT, UNICAMP

Prof. Dr. Luis Augusto Angelotti Meira– FT, UNICAMP

INTRODUÇÃO:

A importância do uso de dados na tomada de decisão é amplamente discutida na literatura acadêmica e empresarial. De acordo com Davenport e Harris (2007), no livro *Competing on Analytics: The New Science of Winning*, a capacidade de coletar e analisar dados de forma eficaz permite que as organizações obtenham uma vantagem competitiva significativa ao tomar decisões mais informadas e precisas.

Temos no Brasil acesso a diversos dados demográficos, socioeconômicos e censitários provenientes de pesquisas realizadas pelo Instituto Brasileiro de Geografia e Estatística (IBGE), porém o atual formato de consulta a esses dados têm um empecilho que é a dificuldade de entendimento dos dados da maneira que são disponibilizados nas plataformas oficiais: via API pública ou via consulta às bases brutas em tabelas *.csv* ou *excel*. Este trabalho tem o objetivo de analisar os dados disponíveis no site do IBGE, tratá-los, criar formas de visualização das variáveis escolhidas e aplicar técnicas de seleção de modelos representativos.

METODOLOGIA:

Para a primeira etapa desta pesquisa precisamos definir e extrair os dados do site do IBGE. Podemos encontrar resultados dos censos de 2000, 2010 e 2022. Devido à pandemia do Coronavírus em 2020 o censo foi adiado e ainda não estava completo na data de início deste trabalho. Optamos, portanto, por utilizar os dados de 2010.

O próximo passo foi entender os requisitos da última etapa do processo que seria a redução de cenários, para que a base inicial estivesse adequada para a aplicação da metodologia desejada. O trabalho "*Selection of representative models for decision analysis under uncertainty*" partiu de um conjunto original de 214 cenários, também chamados no contexto de petróleo, de modelos e reduziu para 9 cenários. O conjunto de dados de cada cenário possuía dimensão igual a 4. Já no trabalho "*Improving representativeness in a scenario reduction process to aid decision making in petroleum fields*", o número de cenários originais se manteve em torno de 214, mas os dados passaram a ter dimensão até 29. O número de cenários reduzidos variou de 2 até 25.

Por esta razão, decidimos buscar no site do IBGE dados que tivessem dimensões compatíveis, já que a mesma técnica de redução de cenários será aplicada neste trabalho. Segundo o site do IBGE, "O Setor Censitário é a menor porção de área utilizada pelo IBGE para planejar, coletar e disseminar os resultados dos Censos e Pesquisas Estatísticas". Cada setor está em um distrito, que por sua vez está em um município que por fim está em uma Unidade Federativa (UF), popularmente chamada de estado.

Em razão da proximidade geográfica das UFs e municípios em relação à sede do campus da instituição de ensino dos responsáveis por esta pesquisa, analisamos as opções e escolhemos o município de São Paulo que contém 94 distritos como objeto de estudo.

Para Kotler e Keller (2012), os fatores pessoais determinantes para definir o perfil de um consumidor são idade, estágio no ciclo de vida, gênero, ocupação, circunstâncias econômicas, personalidade e autoimagem. Dessa forma, buscamos no IBGE uma base que contém alguns desses atributos relacionados em uma única tabela, unindo gênero, cor ou raça e idade. Tal tabela contém o número total de indivíduos em cada um dos grupos cruzando essas três variáveis.

Avançada esta etapa, o próximo desafio foi organizar e higienizar os dados. Apesar da categorização de distritos dentro do município, as bases trazem outro formato, que é a divisão de município em setores censitários que o compõe. Assim, utilizando uma técnica de cruzamento de tabelas, através de uma base que continha a informação sobre a qual distrito cada setor censitário pertencia, pudemos somar as populações dos setores para chegar nas estatísticas de cada distrito.

Dessa forma, obtivemos uma base com 94 linhas em um eixo (os 94 distritos) e 255 colunas (255 grupos compostos por um intervalo de idade, um gênero e uma cor ou raça). Para melhor visualizarmos os dados, criamos colunas auxiliares que somam as variáveis por grupo. Por exemplo, soma da população masculina, soma da população negra, soma da população entre 30 e 34 anos. Isso foi feito para que além da visão multidimensional dos dados, pudéssemos também visualizar cada variável separadamente. Todas as técnicas de manipulação de dados foram feitas na linguagem de programação *Python* com auxílio da biblioteca *Pandas*.

Assim, ainda com a linguagem *Python* e agora com auxílio da biblioteca *Seaborn* geramos as visualizações separadas por cada esfera das variáveis socioeconômicas, veja na figura 1.

Figura 1: Exemplo das visualizações de idade, cor ou raça e gênero utilizando o distrito São Mateus.

Com a visualização definida, pudemos gerar arquivos de entrada para a metodologia de redução de cenários a partir das novas variáveis criando grupos e testando gradativa e iterativamente o software *RMFinder* para o objeto deste estudo. Em sequência tivemos que parametrizar as funções para aceitarem a saída do software a fim de visualizar as características de cada modelo.

RESULTADOS E DISCUSSÃO:

Fizemos quatro baterias de testes de execução do software *RMFinder*, que permite a definição de diversas variáveis de entrada a fim de otimizar o seu resultado. O primeiro teste foi utilizando o conjunto de dados mais simples, com a divisão da população apenas em masculina e feminina, e adicionando uma variável categórica comum para todos os elementos que é obrigatória para execução do programa. O segundo teste foi seguindo o método do primeiro, com a divisão da população entre os cinco grupos de cor ou raça com adição do variável categórica obrigatória.

O próximo conjunto de dados utilizado para execução do programa foi com a divisão dos grupos em intervalos de idade de 5 anos dos 0 aos 59 anos, um intervalo de 10 anos dos 60 aos 69 e todos os grupos restantes a partir da idade de 70 anos. Veja na figura 2.

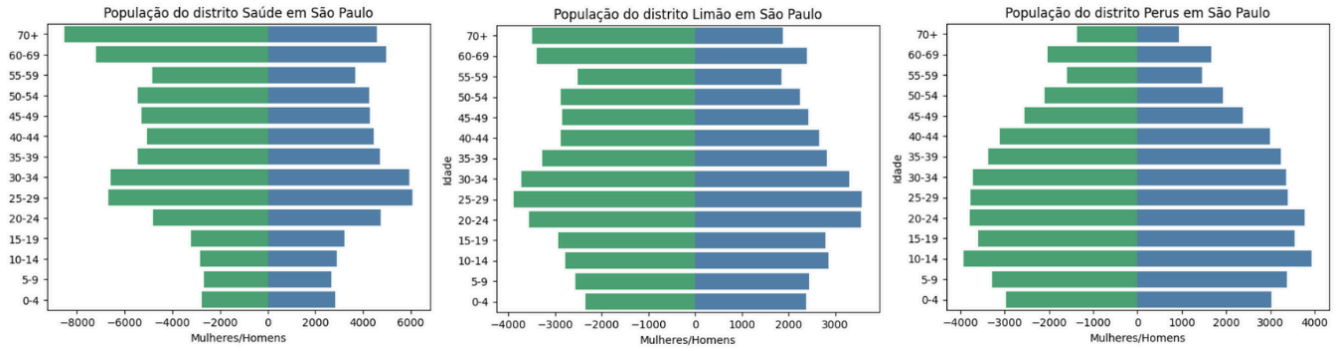

Figura 2: Três pirâmides etárias de três distritos selecionados pelo software de redução de cenários *RMFinder*.

O objetivo é encontrar um conjunto reduzido de distritos que represente bem o todo. Nas três imagens da Figura 2, temos um distrito com pirâmide etária invertida (topo mais largo que a base), a segunda possui uma pirâmide com base e topo parecidos. Já a terceira imagem possui uma pirâmide etária mais tradicional. Observe que este estudo encontrou três distritos com características bem distintas em relação à pirâmide etária.

Repetimos o mesmo processo para os dados com a divisão total de variáveis, onde cada grupo tinha um intervalo de idade, um gênero e uma cor ou raça definidos. Essa última execução trouxe uma situação nova pois o software nunca tinha sido executado com dados multidimensionais com 250 dimensões. O software *RMFinder* não foi capaz de gerar o relatório de correlação entre as variáveis. Entretanto, o arquivo de saída com os identificadores dos modelos selecionados foi gerado e pudemos utilizar as funções criadas aqui para visualizar as distribuições da população e mais uma vez a ferramenta de visualização se mostrou assertiva para compreendermos as peculiaridades de cada distrito.

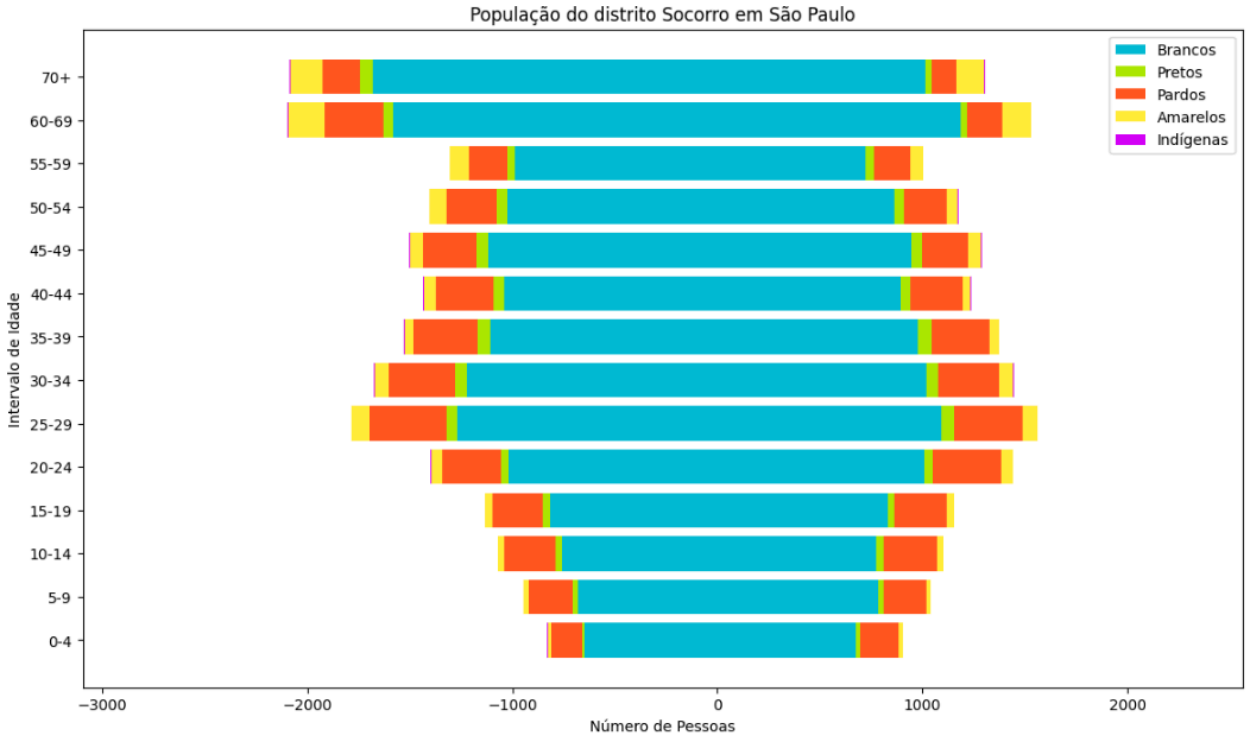

Figura 3: Visualização multidimensional do distrito Socorro.

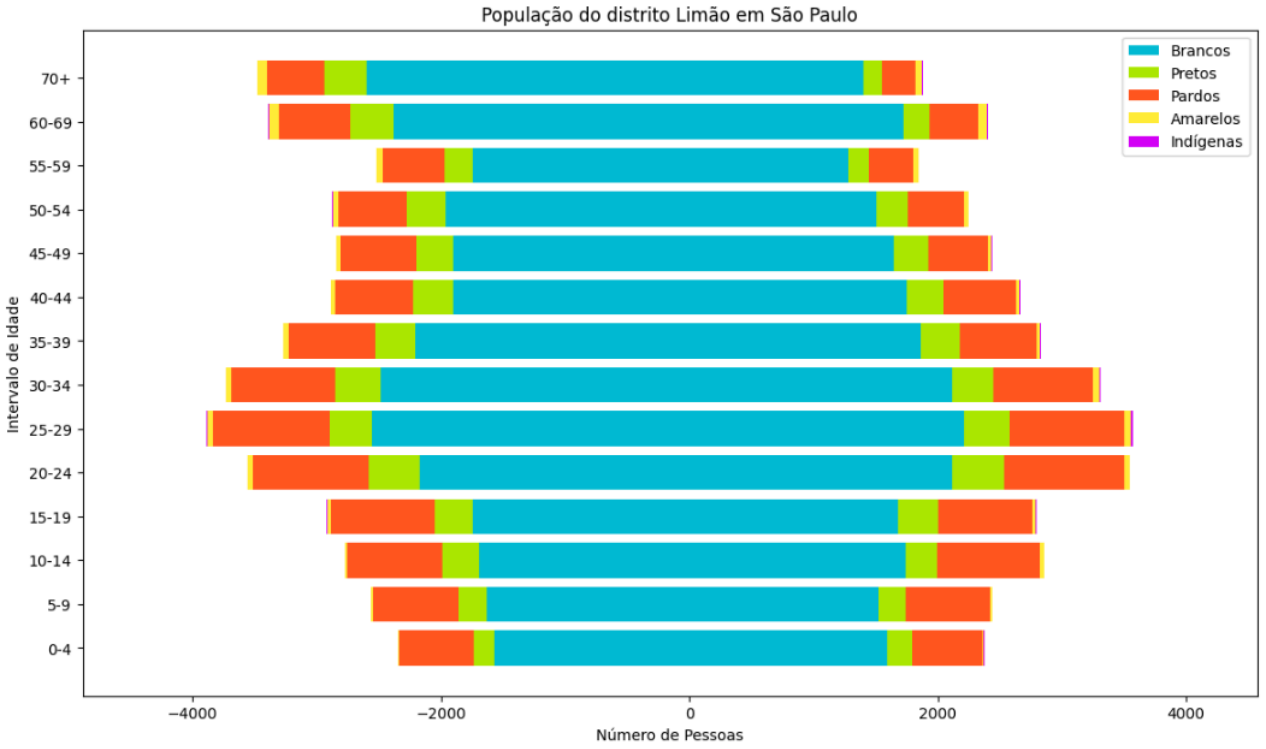

Figura 4: Visualização multidimensional do distrito Limão.

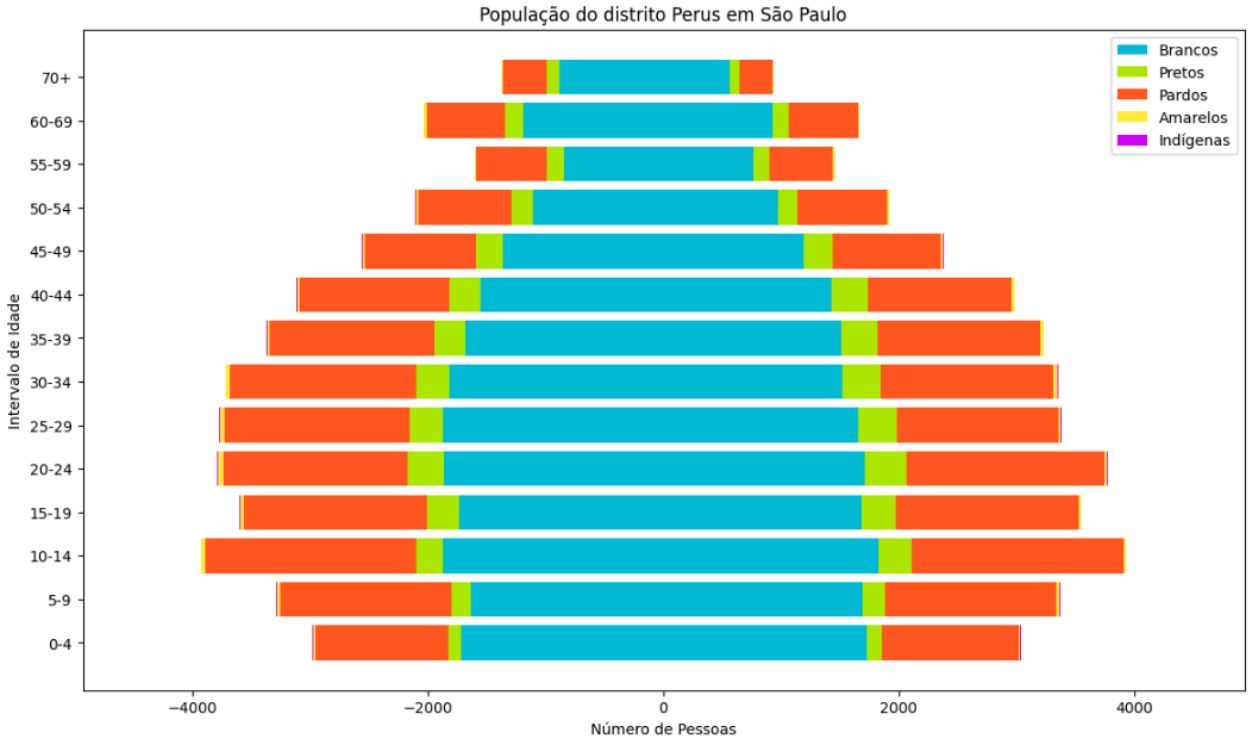

Figura 5: Visualização multidimensional do distrito Perus

CONCLUSÕES:

Podemos analisar nas figuras 3, 4 e 5 retratos bem diferentes entre si, como por exemplo uma grande parcela de população de raça amarela no distrito Socorro nas faixas de idade mais longevas, se comparadas a outros distritos. Ao comparar a solução de modelos representativos entre o conjunto de dados apenas com a separação por intervalo de idade e a divisão entre todos os grupos, podemos perceber que dois dos três distritos são os mesmos, Limão e Perus, o que mostra que o peso de gênero e cor ou raça é pouco representativo na escolha dos distritos. Podemos confirmar essa percepção ao analisar o gráfico de dispersão da população feminina e masculina gerado no relatório da metodologia de redução de cenários que demonstra uma distribuição aproximada entre 50% e 55% para a população feminina e 45% e 50% para a masculina.

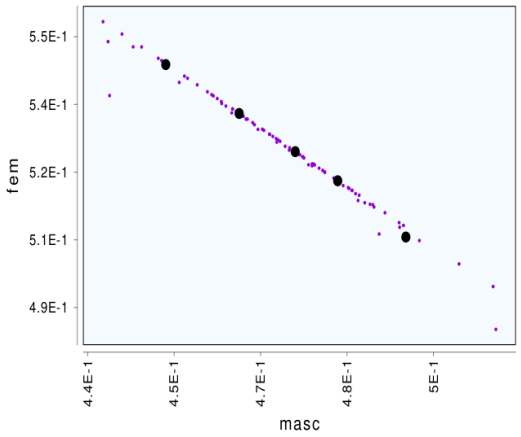

Figura 6: Gráfico de dispersão para 5 modelos reduzidos comparando a população feminina e masculina

Com base no estudo, notamos que a metodologia de redução de cenários se mostrou eficaz para os conjuntos de dados experimentados, assim como as visualizações criadas.

BIBLIOGRAFIA

Instituto Brasileiro de Geografia e Estatística. Ibge - instituto brasileiro de geografia e estatística, 2024. Acesso em 08/03/2024.

Instituto Brasileiro de Geografia e Estatística. Censo brasileiro de 2010. Rio de Janeiro: IBGE, 2012.

Philip Kotler and Kevin Lane Keller. Administração de Marketing. Pearson, São Paulo, 14ª ed., 2012.

Davenport, T. H., Harris, J. G. *Competing on Analytics: The New Science of Winning*. Harvard Business Review Press, 2007.

Luis A.A. Meira, Guilherme P. Coelho, Antonio Alberto S. Santos, and Denis J. Schiozer. Selection of representative models for decision analysis under uncertainty. Computers Geosciences, 88:67–82, 2016.

Luis A.A. Meira, Guilherme P. Coelho, Celmar G. da Silva, João L.A. Abreu, Antonio A.S. Santos, and Denis J. Schiozer. Improving representativeness in a scenario reduction process to aid decision making in petroleum fields. Journal of Petroleum Science and Engineering, 184:106398, 2020.