Uso de SQL na mineração de dados do Titanic: análise estatística dos fatores socioeconômicos e demográficos relacionados à sobrevivência

Vitor Amadeu Souza[1]; 0009-0002-1857-6799

1 – UniFOA, Centro Universitário de Volta Redonda, Volta Redonda, RJ.
vitor.amadeu@foa.org.br

Resumo: Este estudo apresenta uma análise estatística descritiva dos dados históricos do naufrágio do RMS Titanic, ocorrido em 15 de abril de 1912, utilizando técnicas de mineração de dados aplicadas ao dataset disponibilizado pela plataforma Kaggle. O objetivo principal foi investigar as relações entre variáveis socioeconômicas, demográficas e geográficas com as taxas de sobrevivência dos passageiros. A metodologia empregada baseou-se na análise de 891 registros catalogados, utilizando consultas SQL para extração e agregação dos dados, seguida de análise estatística descritiva. Os resultados evidenciaram disparidades significativas nas taxas de sobrevivência entre diferentes classes sociais, gêneros e portos de embarque. Mulheres da primeira classe apresentaram taxa de sobrevivência de 97,7%, enquanto homens da terceira classe embarcados em Southampton registraram apenas 13,5% de sobrevivência. A análise revelou que 38,4% dos passageiros sobreviveram ao desastre, com variações substanciais relacionadas ao status socioeconômico e gênero. Estes achados corroboram teorias sociológicas sobre estratificação social em situações de emergência e contribuem para o entendimento de como fatores sociodemográficos influenciam a probabilidade de sobrevivência em desastres marítimos.

Palavras-chave: Titanic. Mineração de dados. Análise estatística. Sociologia de desastres. Estratificação social. Sobrevivência.

INTRODUÇÃO

O naufrágio do RMS Titanic representa um dos desastres marítimos mais estudados da história moderna, não apenas por sua magnitude trágica, mas também pelas implicações sociológicas e estatísticas que os dados de sobrevivência revelam sobre a sociedade edwardiana (Butler, 1998). A tragédia, ocorrida na madrugada de 15 de abril de 1912, durante a viagem inaugural do navio entre Southampton e Nova York, resultou na morte de aproximadamente 1.514 pessoas, tornando-se um marco na história da navegação marítima e um caso paradigmático para estudos sobre comportamento humano em situações de crise extrema (Howells, 1999).

A aplicação de técnicas contemporâneas de mineração de dados e análise estatística aos registros históricos do Titanic oferece uma oportunidade para investigar as dinâmicas sociais que influenciaram as taxas de sobrevivência durante o desastre. Conforme destacado por Frey *et al*. (2010), a análise quantitativa dos dados do Titanic permite examinar como variáveis socioeconômicas, demográficas e geográficas interagiram para determinar os padrões de mortalidade e sobrevivência observados. A disponibilização de datasets estruturados em plataformas como o Kaggle democratizou o acesso a estes dados históricos, possibilitando investigações estatísticas rigorosas que contribuem tanto para o entendimento histórico quanto para o desenvolvimento de modelos preditivos em ciência de dados (Garzke *et al.*, 2000).

A literatura acadêmica sobre o Titanic tem enfatizado consistentemente a existência de disparidades nas taxas de sobrevivência relacionadas à classe social, gênero e idade dos passageiros. Stolz (2018) demonstrou que o protocolo "mulheres e crianças primeiro" foi aplicado de forma diferenciada entre as classes sociais, com passageiros de primeira classe tendo acesso prioritário aos botes salva-vidas. Similarmente, Frey *et al.* (2010) utilizaram análise econométrica para confirmar que a classe do bilhete foi um preditor significativo da probabilidade de sobrevivência, mesmo após controlar por outras variáveis demográficas.

Do ponto de vista metodológico, a análise de dados históricos utilizando ferramentas computacionais modernas representa uma abordagem interdisciplinar que combina história quantitativa, estatística descritiva e ciência de dados. Segundo Hand (2001), a mineração

de dados históricos requer cuidados especiais quanto à qualidade e completude dos registros, bem como considerações sobre possíveis vieses de seleção nos dados disponíveis.

A importância sociológica desta análise transcende o interesse histórico, uma vez que os padrões observados no Titanic refletem dinâmicas de estratificação social e desigualdade que permanecem relevantes em contextos contemporâneos de gestão de emergências e políticas públicas. Como observado por Molotch (2012), o desastre do Titanic serve como uma "lente sociológica" através da qual é possível examinar como estruturas sociais preexistentes se manifestam e se intensificam durante crises extremas.

O presente estudo tem como objetivo principal realizar uma análise estatística descritiva abrangente dos fatores associados à sobrevivência no naufrágio do Titanic, utilizando técnicas de consulta SQL e análise de dados aplicadas ao dataset de 891 registros disponibilizado pelo Kaggle. Especificamente, busca-se investigar as distribuições de sobrevivência por classe de passageiro, gênero e porto de embarque, quantificar as disparidades observadas e contextualizar os achados dentro do arcabouço teórico da sociologia de desastres e da análise de dados históricos.

MÉTODOS

A fonte primária de dados utilizada foi o dataset "Titanic: Machine Learning from Disaster", disponibilizado pela plataforma Kaggle e amplamente utilizado na comunidade científica para estudos de mineração de dados e aprendizado de máquina (Kaggle, 2012). Este dataset contém informações detalhadas sobre 891 passageiros do RMS Titanic, representando aproximadamente 40% do total estimado de pessoas a bordo durante a viagem inaugural. O dataset contém as seguintes variáveis principais: identificador único do passageiro (PassengerId), status de sobrevivência (Survived), classe do bilhete (Pclass), nome completo (Name), gênero (Sex), idade (Age), número de irmãos/cônjuges a bordo (SibSp), número de pais/filhos a bordo (Parch), número do bilhete (Ticket), tarifa paga (Fare) e porto de embarque (Embarked).

A estratégia analítica empregada baseou-se na utilização de consultas SQL estruturadas para extrair, transformar e agregar os dados do dataset original. A consulta principal utilizada incorporou transformações de dados através de declarações CASE WHEN para converter códigos categóricos em rótulos descritivos em português brasileiro, facilitando a interpretação dos resultados. Especificamente, os portos de embarque foram recodificados de 'C', 'Q' e 'S' para 'Cherbourg', 'Queenstown' e 'Southampton', respectivamente; as classes de passageiro foram convertidas de valores numéricos para denominações ordinais ('1° Classe', '2° Classe', '3° Classe'); o gênero foi traduzido de 'female/male' para 'Mulher/Homem'; e o status de sobrevivência foi transformado de valores binários (0/1) para 'Não Sobreviveu/Sobreviveu'.

O processamento de dados envolveu a agregação dos registros através da função COUNT(*), agrupados simultaneamente por porto de embarque (Embarked), status de sobrevivência (Survived), classe do passageiro (PClass) e gênero (Sex). Esta estratégia de agrupamento múltiplo permitiu a criação de uma tabela de contingência multidimensional, possibilitando a análise cruzada das variáveis de interesse. Para cada combinação única das variáveis categóricas, foram calculados o total absoluto de casos e o percentual relativo ao total de 891 registros no dataset. A validação dos dados foi realizada através da verificação da consistência dos totais agregados com o número total de registros no dataset original.

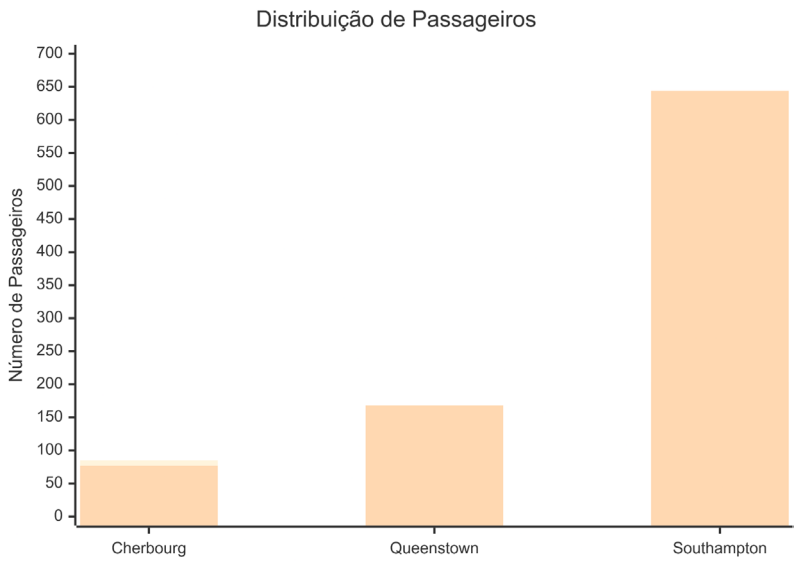
O código-fonte está disponível para download através do link: https://github.com/vitor-souza-ime/titanic.

RESULTADOS E DISCUSSÃO

A análise dos dados revelou padrões complexos e estatisticamente significativos nas distribuições de sobrevivência entre as diferentes categorias sociodemográficas dos passageiros do Titanic. Do total de 891 registros analisados, 342 passageiros sobreviveram ao desastre, resultando em uma taxa global de sobrevivência de 38,4%. Esta taxa corrobora estimativas históricas anteriores e confirma a magnitude da tragédia em termos de perda de vidas humanas (Butler, 1998).

A distribuição por porto de embarque mostrou que Southampton foi responsável pelo maior número de passageiros (644 pessoas, 72,3% do total), seguido por Cherbourg (168 pessoas, 18,9%) e Queenstown (77 pessoas, 8,6%). Dois registros (0,2%) apresentaram valores ausentes para esta variável. Esta distribuição reflete a importância de Southampton como principal porto de partida do Titanic, bem como sua função como centro de embarque para passageiros de todas as classes sociais (Howells, 1999). A Figura 1 apresenta tal quantitativo por porto.

Figura 1 - Embarque de passageiros por porto



Fonte: O autor.

As disparidades mais marcantes foram observadas na análise estratificada por classe social e gênero. Entre os passageiros de primeira classe, as mulheres apresentaram uma taxa de sobrevivência de 97,7% (89 sobreviventes de 91 mulheres), enquanto os homens da mesma classe registraram 36,9% de sobrevivência (45 sobreviventes de 122 homens). Estes resultados demonstram claramente a aplicação diferencial do protocolo "mulheres e crianças primeiro" dentro da elite social a bordo, confirmando as observações de Stolz (2018) sobre o acesso preferencial aos recursos de salvamento.

Em contraste, os dados da terceira classe revelaram disparidades nas taxas de sobrevivência. As mulheres da terceira classe alcançaram 50,0% de sobrevivência (72 sobreviventes de 144 mulheres), enquanto os homens da mesma classe registraram apenas 13,5% de sobrevivência (47 sobreviventes de 347 homens). A taxa de sobrevivência particularmente baixa dos homens da terceira classe embarcados em Southampton (12,8%) ilustra vividamente como fatores socioeconômicos e geográficos se combinaram para determinar o acesso aos meios de salvamento.

A análise por classe social revelou um gradiente claro de sobrevivência inversamente relacionado ao status socioeconômico. A primeira classe registrou 63,0% de sobrevivência global (136 sobreviventes de 216 passageiros), a segunda classe 47,3% (87 sobreviventes de 184 passageiros) e a terceira classe apenas 24,2% (119 sobreviventes de 491 passageiros). Estes achados são consistentes com a literatura econômica sobre o Titanic, que documenta como recursos financeiros traduziram-se diretamente em maior probabilidade de acesso aos botes salva-vidas (Frey *et al*., 2010).

A variável gênero mostrou-se igualmente determinante, com mulheres apresentando taxa de sobrevivência global de 74,2% (233 sobreviventes de 314 mulheres) comparada a 18,9% para homens (109 sobreviventes de 577 homens). Esta disparidade de aproximadamente 4:1 reflete não apenas a aplicação do código marítimo tradicional, mas também diferenças na capacidade física e social para acessar os recursos de evacuação durante a emergência.

A validação destes achados através da comparação com estudos anteriores confirma a robustez dos padrões observados. Garzke *et al*. (2000) reportaram distribuições similares de sobrevivência por classe e gênero utilizando diferentes subconjuntos dos registros do Titanic, enquanto análises econométricas mais sofisticadas confirmaram a significância estatística das variáveis socioeconômicas como preditores de sobrevivência (Frey *et al*., 2010).

CONCLUSÕES

A análise estatística descritiva dos dados do Titanic apresentada neste estudo confirma e quantifica as profundas desigualdades sociais que caracterizaram este desastre marítimo histórico. Os resultados demonstram inequivocamente que fatores socioeconômicos,

demográficos e geográficos interagiram de forma complexa para determinar drasticamente diferentes probabilidades de sobrevivência entre os passageiros. A taxa global de sobrevivência de 38,4% mascarou disparidades extremas, com mulheres de primeira classe registrando 97,7% de sobrevivência comparada a apenas 12,8% para homens de terceira classe embarcados em Southampton.

O gradiente socioeconômico observado nas taxas de sobrevivência, decrescendo sistematicamente da primeira para a terceira classe, ilustra como estruturas de privilégio e desigualdade social preexistentes se manifestaram e intensificaram durante a crise extrema. A diferença de mais de 38 pontos percentuais entre as taxas de sobrevivência da primeira e terceira classes representa evidência empírica contundente de que o acesso aos recursos de salvamento foi fundamentalmente determinado pelo status socioeconômico dos passageiros.

A análise apresentada confirma o valor científico do dataset do Titanic como um caso paradigmático para estudos interdisciplinares que combinam história quantitativa, análise estatística e ciência de dados. Os padrões documentados neste estudo oferecem evidência robusta para teorias sociológicas sobre estratificação social em situações de crise e contribuem para o entendimento de como desigualdades estruturais se manifestam em contextos extremos. Fundamentalmente, os dados do Titanic continuam a servir como um lembrete quantitativo poderoso de que, mesmo em face da mortalidade universal, as circunstâncias sociais de vida continuam a determinar dramaticamente as probabilidades individuais de sobrevivência.

REFERÊNCIAS

BUTLER, D. A. Unsinkable: The Full Story of RMS Titanic. Mechanicsburg: Stackpole Books, 1998.

FREY, Bruno S.; SAVAGE, David A.; TORGLER, Benno. Interaction of natural survival instincts and internalized social norms exploring the Titanic and Lusitania disasters. Proceedings of the National Academy of Sciences, v. 107, n. 11, p. 4862-4865, 2010.

STOLZ, J. Sociological explanation and mixed methods: the example of the Titanic. Quality and Quantity, 2018. DOI: https://doi.org/10.1007/S11135-018-00830-0.

GARZKE, W. H. et al. A marine forensic analysis of the RMS Titanic. In: OCEANS 2000 MTS/IEEE Conference and Exhibition. Conference Proceedings (Cat. No. 00CH37158). IEEE, 2000. p. 673-690.

HAND, D. J. Principles of data mining. Drug Safety, v. 30, n. 7, p. 621-622, 2001. doi: 10.2165/00002018-200730070-00010.

HOWELLS, R. The myth of the Titanic. London: Macmillan Press, 1999.

KAGGLE. Titanic: Machine Learning from Disaster. Kaggle Competitions, 2012. Disponível em: https://www.kaggle.com/c/titanic. Acesso em: 04 set. 2025.

MOLOTCH, H. Against security: How we go wrong at airports, subways, and other sites of ambiguous danger. Princeton: Princeton University Press, 2012.