Professor Nelio Machado

Projeto alternativo: Dados do Titanic

Objetivo: Carregar os dados do Titanic no Databricks e analisar os dados usando Python.
Fonte dos dados: Titanic, que pode ser encontrado no Kaggle ou várias outras fontes de dados.

Atividades:

A. Redefinir a coluna Survived, que possui os valores {0, 1} para {'Não_Sobreviveu", "Sobreviveu"}, sendo 0 = 'Não_Sobreviveu" e 1 = 'Sobreviveu".

- A redefinição da coluna Survived para os valores {0: "Não_Sobreviveu", 1: "Sobreviveu"} foi implementada e pode ser conferida no notebook anexado ao e-mail.

B. Delete as colunas 'PassengerID' e 'Ticket'. Estamos deletando estas colunas por entendermos que são colunas que não são importantes para a área de negócio;

- A exclusão das colunas *PassengerID* e *Ticket* foi realizada, conforme indicado, por não serem relevantes para a análise. A implementação pode ser conferida no notebook anexado ao e-mail.

C. Verifique a quantidade de missing values para cada coluna do dataframe. Qual a coluna com maior número de missing values?

- A coluna com o maior número de valores ausentes foi a **Cabin**, com 687 entradas faltando. Isso pode ser justificado pela natureza dos dados, uma vez que nem todos os passageiros do Titanic tinham cabine registrada.

D. De forma sistemática, sugira como podemos tratar os missing values das colunas do dataframe? Na sequência, faça o tratamento dos missing values do dataframe;

- Age: Valores ausentes preenchidos com a média, embora essa abordagem seja simples e pode não capturar a verdadeira variabilidade da idade, ela é útil para preservar a estrutura dos dados.
- Cabin: Valores ausentes substituídos por "Unknown" devido à alta quantidade de dados faltantes, preservando os registros.

- Embarked: Valores ausentes preenchidos com a moda, representando o comportamento mais frequente dos passageiros.

E. **Calcule estatísticas descritivas como sum, skew, furt, count, mean, std, min, 25%, 50%, 75% e max para as colunas do dataframe. Qual a interpretação?**

- A análise dos dados revela uma grande diversidade de passageiros, com predominância de jovens e tarifas baixas, embora uma pequena parcela tenha pago por serviços de luxo na primeira classe. A maioria dos passageiros estava viajando sozinho ou com poucos familiares. A desigualdade social a bordo é evidente, com uma forte concentração de passageiros nas classes mais baixas (Pclass 2 e 3) e uma distribuição de tarifas que favorece as camadas sociais mais baixas.

F. Construa a matriz de correlação para as colunas do dataframe. Quais colunas são altamente correlacionadas? Quais colunas possuem fraca correlação com a coluna 'Survived'?

- Fare (valor pago pela passagem) tem uma correlação moderada positiva com Survived (0.26), indicando que passageiros que pagaram mais por suas passagens tinham uma maior chance de sobreviver.
- Pclass (classe do passageiro) tem uma correlação negativa moderada com Survived (-0.34), sugerindo que os passageiros nas classes mais baixas tinham menos chances de sobrevivência.

G. Qual a porcentagem de homens que sobreviveram? Qual a porcentagem de mulheres que sobreviveram?

- Mulheres: 74.2% das mulheres sobreviveram.
- Homens: Apenas 18.9% dos homens sobreviveram.

- Esses resultados indicam uma clara disparidade entre os sexos, possivelmente devido a protocolos de evacuação que priorizavam mulheres e crianças.

H. Qual a média do valor pago (coluna 'Fare') por sexo? Na média, homens e mulheres pagaram o mesmo valor?

- Mulheres: 44.48
- Homens: 25.52

➢ Embora a diferença não seja exorbitante, ela sugere que as mulheres, em geral, pagaram um valor mais alto pelas passagens, o que pode refletir um padrão de classe social mais alta ou preferências por acomodações mais luxuosas.

I. Qual a média do valor pago (coluna 'Fare') por 'Pclass'? Qual sua interpretação dos resultados?

- Pclass 1: 84.15
- Pclass 2: 20.66
- Pclass 3: 13.68

➢ Pclass 1: Passageiros dessa classe pagaram significativamente mais, refletindo o custo elevado associado a serviços de maior luxo e exclusividade.
➢ Pclass 2 e 3: Apresentam valores médios menores, evidenciando a diferença de custo e serviços oferecidos entre as classes econômicas.

J. Qual a média do valor pago (coluna 'Fare') por 'Embarked'? Qual sua interpretação dos resultados?

- C: 59.95
- S: 27.24
- Q: 13.28

➢ Cherbourg (C) tem a tarifa média mais alta, sugerindo que os passageiros desse porto eram de classes sociais mais altas.
➢ Southampton (S) apresenta uma média intermediária, com uma maior diversidade no perfil de passageiros.
➢ Queenstown (Q) tem a menor tarifa média, sugerindo passagens mais acessíveis, possivelmente para passageiros de classe mais baixa.

K. Quantas pessoas embarcaram (count') por cada tipo de 'Embarked'?

- Q: 77 passageiros
- C: 168 passageiros
- S: 646 passageiros

L. Qual sua conclusão quando analisamos (count e média) a coluna 'Fare' por 'Embarked', 'Sex' e 'Pclass'?

- Cherbourg (C): Passageiros deste porto pagaram, em média, mais pelas passagens, com destaque para as mulheres na Pclass 1, que têm o maior valor médio de 'Fare'. Isso sugere que Cherbourg está associado a passageiros de maior classe econômica, especialmente em relação às mulheres, que podem ter viajado mais em cabines de alto custo ou tido maior acesso a serviços premium.
- Southampton (S): Aqui, observamos uma divisão de preços maior, com mulheres na Pclass 1 pagando significativamente mais que os homens, refletindo talvez um perfil mais diverso de classe social. Os valores para as classes 2 e 3 são mais acessíveis, com as maiores diferenças entre os sexos nas classes mais altas.
- Queenstown (Q): Este porto tem os valores mais baixos de 'Fare', especialmente na Pclass 3, indicando que os passageiros de Queenstown tinham, em sua maioria, passagens mais econômicas.

M. Você consegue identificar outliers na coluna 'Fare'?

- Sim, outliers identificados: 116

N. Construa a coluna adicional 'Fare_Range' para categorizar os valores da coluna 'Fare' em {'Alto', 'Medio', 'Baixo'}.

- A criação da coluna adicional *Fare_Range*, categorizando os valores da coluna *Fare* como {'Alto', 'Medio', 'Baixo'}, foi realizada. O código correspondente encontra-se no notebook anexado ao e-mail.