UFCD 5417: Projeto 1
Análise de dados do Titanic

v1.0

Nelson Santos
nelson.santos.0001376@edu.atec.pt

ATEC — 17 de dezembro de 2024

Índice

1 Introdução

Este projeto tem como objetivo explorar e analisar o conjunto de dados "Titanic", que contém informações sobre os passageiros do famoso navio e os sobreviventes do desastre. A análise será realizada utilizando as bibliotecas `Pandas` e `Matplotlib` do `Python 3`, permitindo a leitura, manipulação e visualização dos dados para uma análise descritiva e exploratória.

1.1 Titanic

O *Royal Mail Ship Titanic* (RMS) Titanic era um barco transatlântico Britânico que se afundou na noite de 14 para 15 de abril de 1912, após colidir com um iceberg no Oceano Atlântico Norte, durante sua viagem inaugural entre Southampton (Inglaterra) e Nova Iorque (EUA). Na época, o Titanic era considerado o maior e mais luxuoso navio de passageiros do mundo, e foi projetado para ser inafundável. No entanto, a colisão com o iceberg causou danos significativos ao casco do navio, levando-o a afundar em menos de três horas.

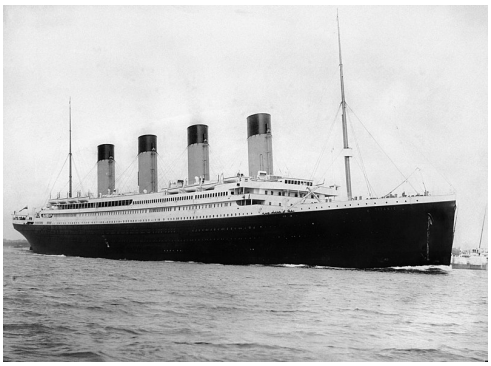

Figura 1: RMS Titanic. Fonte: Wikipédia

O navio transportava cerca de 2.200 pessoas, incluindo passageiros e membros da tripulação. Desses, aproximadamente 1.500 perderam a vida, tornando-se um dos desastres marítimos mais fatais da história moderna. Os passageiros estavam a bordo em várias classes: primeira, segunda e terceira classe, e também havia uma grande diversidade de nacionalidades, com alguns passageiros embarcando em diferentes portos durante a viagem.

1.2 Passageiros e dados relacionados

O conjunto de dados "Titanic" que será utilizado neste projeto contém informações sobre um subconjunto dos passageiros do navio, incluindo dados como: idade, sexo, classe de embarque, etc. Estes dados fornecem uma visão detalhada dos passageiros e suas características, permitindo uma análise mais profunda sobre os fatores que influenciaram a sobrevivência no desastre.

O conjunto de dados contém uma amostra de cerca de 400 passageiros, apesar de o Titanic ter aproximadamente 2.200 pessoas a bordo. Esta discrepância ocorre por várias razões. Em primeiro lugar, o conjunto de dados disponível foi extraído como uma amostra representativa dos passageiros para fins educativos e de análise, sem a intenção de incluir todos os passageiros do navio. Em segundo lugar, muitos dados sobre os passageiros não estavam completos ou não foram registados na época, como informações sobre idade ou classe, o que contribuiu para a redução do número de entradas.

Esses dados permitirão explorar a relação entre fatores como classe de embarque, idade, sexo e a sobrevivência dos passageiros. Além disso, as informações sobre os portos de embarque serão importantes para analisar possíveis padrões de sobrevivência com base na origem dos passageiros.

O conjunto de dados pode ser obtido na plataforma Kaggle, onde estão disponíveis para download em formato CSV. Também será disponibilizado um ficheiro CSV com os dados.

2 Estrutura dos Dados

O conjunto de dados contém informações sobre os passageiros, incluindo se sobreviveram ou não, entre outras variáveis. De seguida encontra-se uma descrição das colunas do dataset. A Tabela 1 apresenta a mesma informação de forma mais compacta.

Coluna	Tipo de Dado	Unidade	Exemplo	Descrição
PassengerId	`inteiro`	–	1	Identificador único do passageiro
Pclass	`inteiro`	–	3	Classe do passageiro (1, 2, ou 3)
Name	`string`	–	Braund, Mr. Owen Harris	Nome do passageiro
Sex	`string`	–	male	Sexo do passageiro
Age	`float`	anos	22.0	Idade do passageiro
SibSp	`inteiro`	–	1	Número de irmãos/cônjuges a bordo (*Siblings/Spouse*)
Parch	`inteiro`	–	0	Número de pais/filhos a bordo (*Parents/Children*)
Ticket	`string`	–	A/5 21171	Número do bilhete do passageiro
Fare	`float`	libras	7.25	Tarifa paga pelo passageiro
Cabin	`string`	–	C85	Número da cabine, se disponível
Embarked	`string`	–	C	Porto de embarque (C = Cherbourg; Q = Queenstown; S = Southampton)
Survived	`inteiro`	–	0	Se o passageiro sobreviveu (0 = Não, 1 = Sim)

Tabela 1: Descrição das colunas do conjunto de dados Titanic

3 Atividades

Segue-se o conjunto de instruções para desenvolver a análise:

3.1 Leitura e exploração dos dados

- Carregar o ficheiro `CSV` utilizando o `pandas`
- Visualizar os primeiros e últimos registos para explorar a estrutura e o conteúdo dos dados
- Realizar uma análise descritiva inicial com as funções `.describe()` e `.info()`

3.2 Limpeza e pré-processamento de dados

- Verificar a existência de valores nulos e decidir o tratamento adequado (ex.: preenchimento ou eliminação)
- Converter colunas para os tipos de dados apropriados, se necessário
- Criar uma nova coluna chamada `Idade_Milissegundos`, onde cada valor será a idade do passageiro convertida para o número de milissegundos desde o *Epoch* (1 de Janeiro de 1970). Este processo permite uma formatação consistente das idades

3.3 Análise e manipulação de dados

Para esta secção, usar funções como `groupby`, `mean`, `sum`, entre outras, de `pandas`, para realizar análises agregadas.

- **Análise de Sobrevivência:**
  - Calcular a taxa de sobrevivência por classe (`Pclass`) e sexo (`Sex`)
  - Analisar a relação entre idade (`Age`) e sobrevivência
- **Análise de Tarifa e Classe:**
  - Calcular a tarifa média por classe e sexo
  - Identificar correlações entre a tarifa (`Fare`) e a sobrevivência

3.4 Visualização de dados

- **Tendências Temporais:** Criar gráficos de barra ou linha para mostrar a distribuição de sobreviventes por classe e sexo
- **Correlação entre Variáveis:** Utilizar gráficos de dispersão (scatter plots) para analisar a correlação entre `Age`, `Fare` e `Survived`
- **Distribuição das Variáveis:** Criar histogramas para visualizar a distribuição de `Age`, `Fare`, e `Survived`

3.5 Exportação dos resultados

- Guardar o `DataFrame` atualizado num novo ficheiro `Excel`, incluindo todas as colunas novas criadas durante a análise
- Exportar também gráficos relevantes para um relatório final

3.6 Armazenamento numa Base de Dados

Guardar os dados numa base de dados, utilizando um sistema de bases de dados (como `SQLite`, `PostgreSQL` ou `MySQL`) para guardar e consultar o conjunto de dados.

- Criar uma tabela onde cada coluna corresponde a uma das variáveis no conjunto de dados
- Inserir o conjunto de dados na tabela criada, assegurando que a nova coluna `Idade_Milissegundos` esteja formatada corretamente

3.7 Análise Adicional

Poderá adicionar as colunas que achar necessárias para a análise, bem como realizar outras análises que considere relevantes. A criatividade e originalidade serão valorizadas na avaliação do projeto.

4 Entregáveis e data de entrega

- **Código em Python:** O código deve estar comentado para explicar cada operação realizada
- **Gráficos:** Guardar os gráficos gerados e incluí-los no relatório
- **Relatório Final:** Elaborar um breve resumo das principais descobertas (em formato `Markdown` ou `Word`) com os gráficos criados, incluindo conclusões
- **Base de Dados:** Apresentar a base de dados com os dados inseridos
- **Data de entrega:** A data da entrega será ainda difundida após coordenação com os alunos

5 Grupos de Trabalho

Os grupos de trabalho para este projeto ainda serão a designar.

6 Matriz de avaliação

Tabela 2: Critérios de Avaliação do Projeto Titanic

Critério	**Descrição Detalhada**	**Pontos**
Qualidade do Código	• Estrutura e organização do código (2 pts) • Uso de ambientes virtuais (0,5 pts) • Utilização de Git com commits claros e bem documentados (1,5 pts) • Código comentado e boas práticas de programação (2 pts)	**6 pts**
Manipulação e Análise de Dados	• Criação e utilização correta da coluna `Idade_Milissegundos` (1 pt) • Realização das análises descritas (2 pts) • Uso adequado de funções `pandas` para agregações (1 pt)	**4 pts**
Visualização dos Dados	• Gráficos que mostrem as tendências descritas (1 pt) • Criatividade e clareza dos gráficos (0,5 pts) • Uso correto de bibliotecas como `Matplotlib` ou `Seaborn` (0,5 pts)	**2 pts**
Exportação e Armazenamento	• Exportação correta dos dados para `Excel` (0,5 pts) • Exportação de gráficos para o relatório final (0,5 pts) • Inserção do conjunto de dados numa base de dados com a nova coluna formatada (0,5 pts)	**1,5 pts**
Criatividade	• Elementos inovadores no projeto que se destaquem pela originalidade • Estes pontos são atribuídos pelo docente consoante a criatividade e originalidade do projeto	**2 pts**
Relatório Final	• Clareza e qualidade da documentação (1,5 pts) • Inclusão de gráficos e conclusões detalhadas (1 pt)	**2,5 pts**
Total		**20 pts**