**Exploratory Data Analysis: Titanic Dataset**

**What questions does this data analysis attempt to answer?**

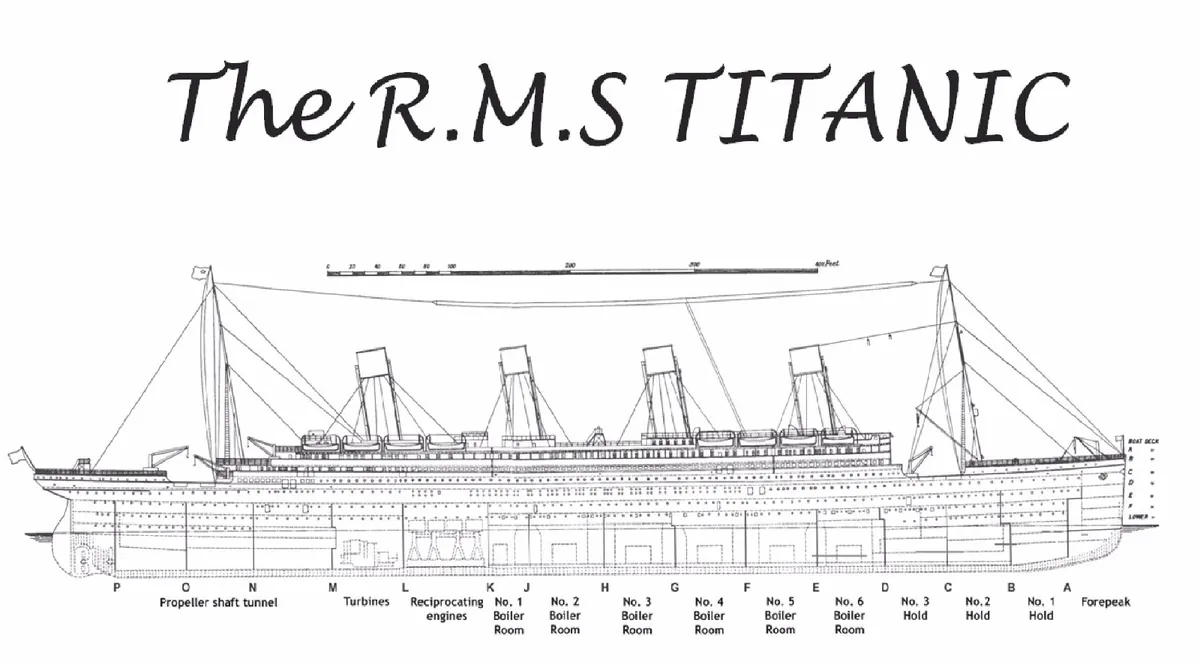
* Filling in the missing values in the ‘age’ column
  + Correlations
    - A graph of age and age

      Description automatically generated
    - Correlation coefficients:
    - Age and Fare: 0.179
    - Age and Parch: -0.151
    - Age and SibSp: -0.244
    - Age and Pclass: -0.41
    - **Age and Pclass seem to have the strongest correlation**
      * Fill in the age column based on the passenger’s class
  + Median age for each Pclass:
    - 1: 39.0
    - 2: 29.0
    - 3: 24.0

**Filling in the Fare column 🡪 the ‘Fare’ column contains a lot of zero values**

* Fare and Pclass: -0.5586 (correlation coefficient)
* Get the median and fill in the missing data

|  |  |
| --- | --- |
| 1 | 71.14165 |
| 2 | 13.00000 |
| 3 | 9.25625 |



A few things about the ‘cabin’ column:

* We do NOT need the specific cabin numbers 🡪 just the letter of each cabin is enough
* Created a new column called “Deck”

A graph of a distribution of deck

Description automatically generated

Deck column

|  |  |
| --- | --- |
| NaN | 1014 |
| C | 94 |
| B | 65 |
| D | 46 |
| E | 41 |
| A | 22 |
| F | 21 |
| G | 5 |
| T | 1 |

Deck – (median Fare of each deck)

|  |  |
| --- | --- |
| A | 35.0771 |
| B | 82.2667 |
| C | 86.2896 |
| D | 52.5542 |
| E | 53.1000 |
| F | 13.0000 |
| G | 16.7000 |
| T | 35.5000 |

Deck Pclass Count

|  |  |  |
| --- | --- | --- |
| A | 1 | 22 |
| B | 1 | 65 |
| C | 1 | 94 |
| D | 1 2 | 40 6 |
| E | 1 2 3 | 34 4  3 |
| F | 2 3 | 13 8 |
| G | 3 | 5 |
| T | 1 | 1 |

A diagram of different colors

Description automatically generated

* **Filled the Deck column based on Pclass**

**‘Survived’ column**

* Correlation coefficients – ‘sex’ and ‘survived’
  + sex\_male -0.543351
  + sex\_female 0.543351
  + More females survived than males
* **Sex is strongly correlated with survival rates**

Deck Survived = 1, Died = 0

|  |  |  |
| --- | --- | --- |
| A | 0  1 | 8  7 |
| B | 0  1 | 12  35 |
| C | 0  1 | 45  54 |
| D | 0  1 | 8  25 |
| E | 0  1 | 8  24 |
| F | 0  1 | 465  195 |
| G | 0  1 | 2  2 |
| T | 0  1 | 1  0 |

* **Deck T passengers all died**
* **Deck B and E passengers mostly survived**
* **Deck F passengers mostly died**

A graph of a number of blue and orange bars

Description automatically generated

**Filling in the ‘Survived’ column:**

1. **Deck F**
2. **Female**
3. **Pclass = 1**

**🡪 Most likely survived!**

**Three exceptions:**

1. Sandstrom, Miss. Beatrice Irene

|  |  |
| --- | --- |
| PassengerId | 1009 |
| Survived | NaN |
| Pclass | 3 |
| Name | Sandstrom, Miss. Beatrice Irene |
| Sex | Female |
| Age | 1.0 |
| SibSp | 1 |
| Parch | 1 |
| Ticket | PP 9549 |
| Fare | 16.7 |
| Cabin | G6 |
| Embarked | S |
| Deck | G |
| title | Miss. |

1. **Malachard, Mr. Noel**

PassengerId 1193

Survived NaN

Pclass 2

Name Malachard, Mr. Noel

Sex male

Age 29.0

SibSp 0

Parch 0

Ticket 237735

Fare 15.0458

Cabin D

Embarked C

Deck D

title Mr.

1. **Nourney, Mr. Alfred (Baron von Drachstedt)**

PassengerId 1297

Survived NaN

Pclass 2

Name Nourney, Mr. Alfred (Baron von Drachstedt")"

Sex male

Age 20.0

SibSp 0

Parch 0

Ticket SC/PARIS 2166

Fare 13.8625

Cabin D38

Embarked C

Deck D

title Mr.