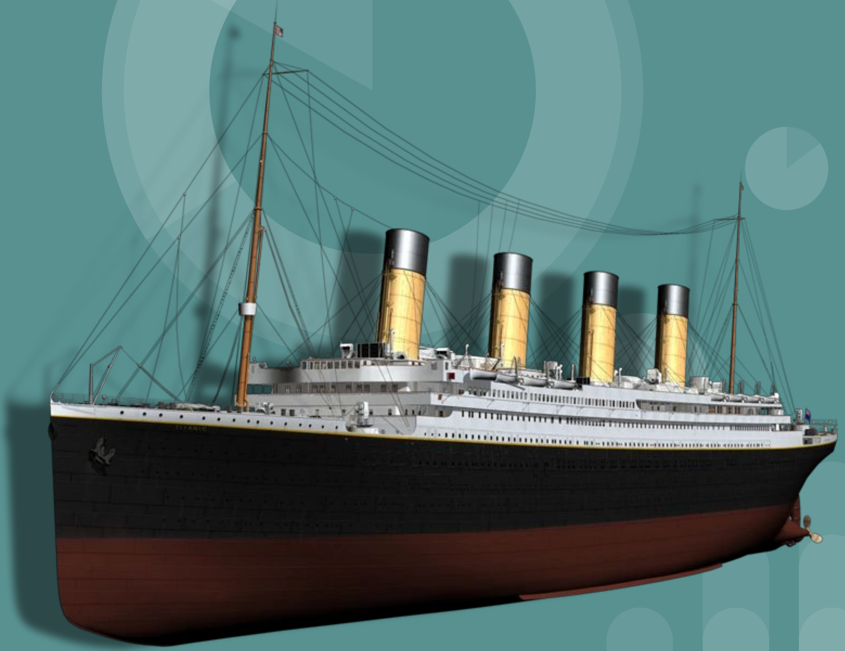
Oliver

# Titanic Survival Data Analysis



Joseph Milotta, Luis Carlos Lopez, Leonardo Rios, & Oliver von Mizener

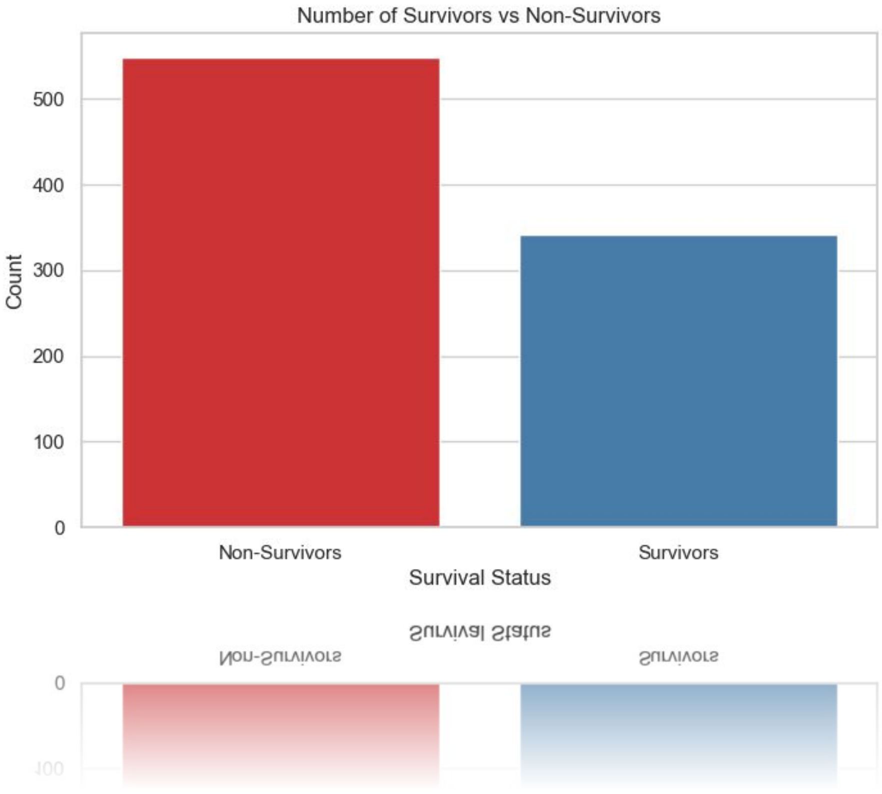
Project 3 Group 4

Oliver

# Introduction

The Titanic dataset offers a fascinating blend of historical data and analytical challenges, making it an ideal learning opportunity. Our interest in this dataset stems from its potential for exploring patterns, uncovering insights, and honing our technical skills.

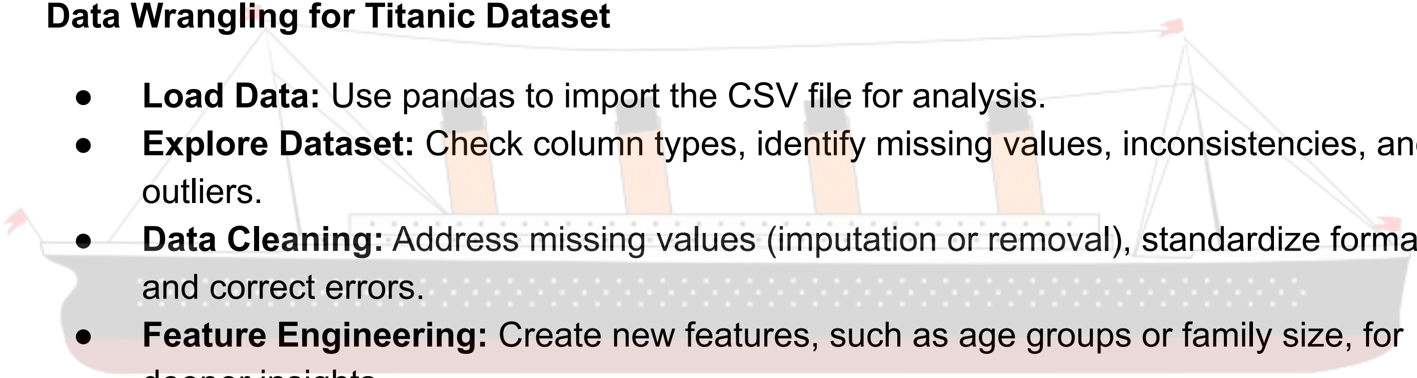
By working on this project, we aimed to apply our expertise in data engineering, including automated ETL pipelines and data cleaning, while creating innovative visualizations like heat maps, boxenplots and correlation matrices to tell a compelling data story. This project represents both a deep dive into the dataset and a journey of skill development.



Oliver

# Data Wrangling

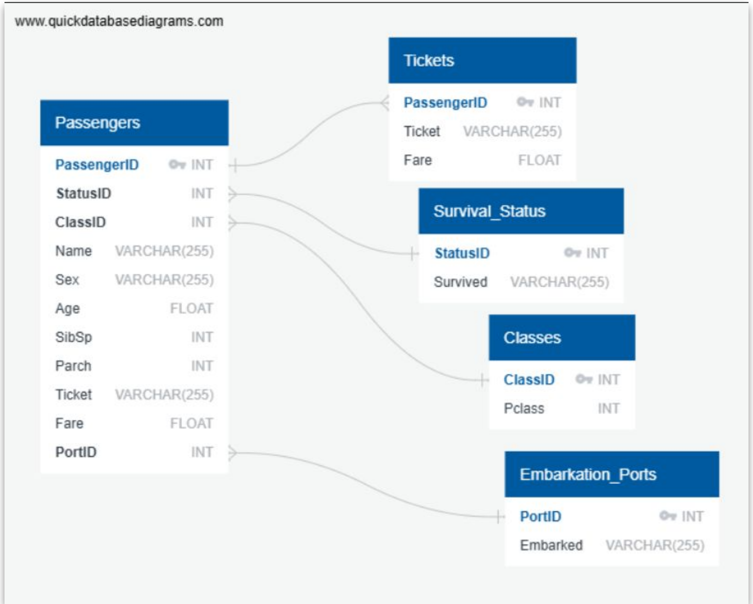
### Data Wrangling for Titanic Dataset



- **Load Data:** Use pandas to import the CSV file for analysis.
- **Explore Dataset:** Check column types, identify missing values, inconsistencies, and outliers.
- **Data Cleaning:** Address missing values (imputation or removal), standardize formats, and correct errors.
- **Feature Engineering:** Create new features, such as age groups or family size, for deeper insights.
- **Save Cleaned Data:** Export the cleaned dataset to a new CSV file for the ETL pipeline.
- **Visualization**: Render visualizations to help understand the scope of who survived and what circumstances would lead to a likely survival.

Oliver

# Database Design



| Passengers | |
|---|---|
| PassengerID | INT |
| StatusID | INT |
| ClassID | INT |
| Name | VARCHAR(255) |
| Sex | VARCHAR(255) |
| Age | FLOAT |
| SibSp | INT |
| Parch | INT |
| Ticket | VARCHAR(255) |
| Fare | FLOAT |
| PortID | INT |

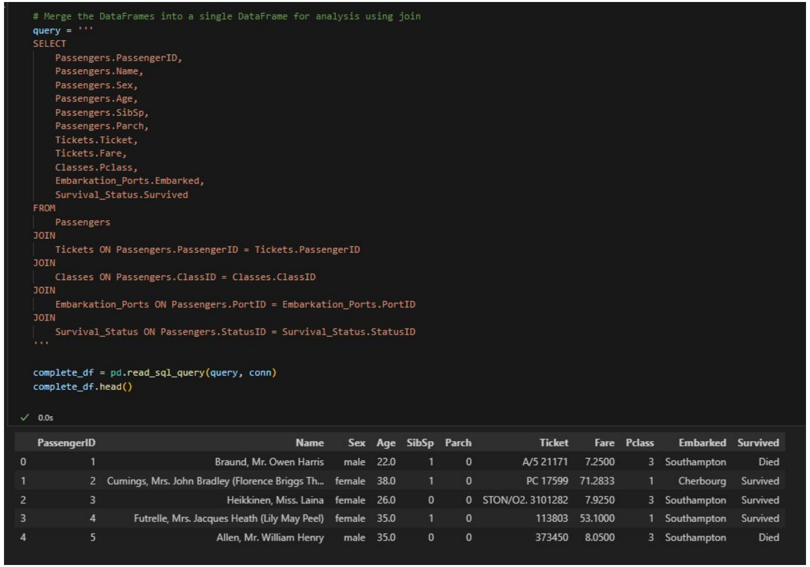
| Tickets | |
|---|---|
| PassengerID | INT |
| Ticket | VARCHAR(255) |
| Fare | FLOAT |

| Survival_Status | |
|---|---|
| StatusID | INT |
| Survived | VARCHAR(255) |

| Classes | |
|---|---|
| ClassID | INT |
| Pclass | INT |

| Embarkation_Ports | |
|---|---|
| PortID | INT |
| Embarked | VARCHAR(255) |

ERD Design using Quickdbd



```
# Merge the DataFrames into a single DataFrame for analysis using join
query = '''
SELECT
    Passengers.PassengerID,
    Passengers.Name,
    Passengers.Sex,
    Passengers.Age,
    Passengers.SibSp,
    Passengers.Parch,
    Tickets.Ticket,
    Tickets.Fare,
    Classes.Pclass,
    Embarkation_Ports.Embarked,
    Survival_Status.Survived
FROM
    Passengers
JOIN
    Tickets ON Passengers.PassengerID = Tickets.PassengerID
JOIN
    Classes ON Passengers.ClassID = Classes.ClassID
JOIN
    Embarkation_Ports ON Passengers.PortID = Embarkation_Ports.PortID
JOIN
    Survival_Status ON Passengers.StatusID = Survival_Status.StatusID
'''

complete_df = pd.read_sql_query(query, conn)
complete_df.head()
```

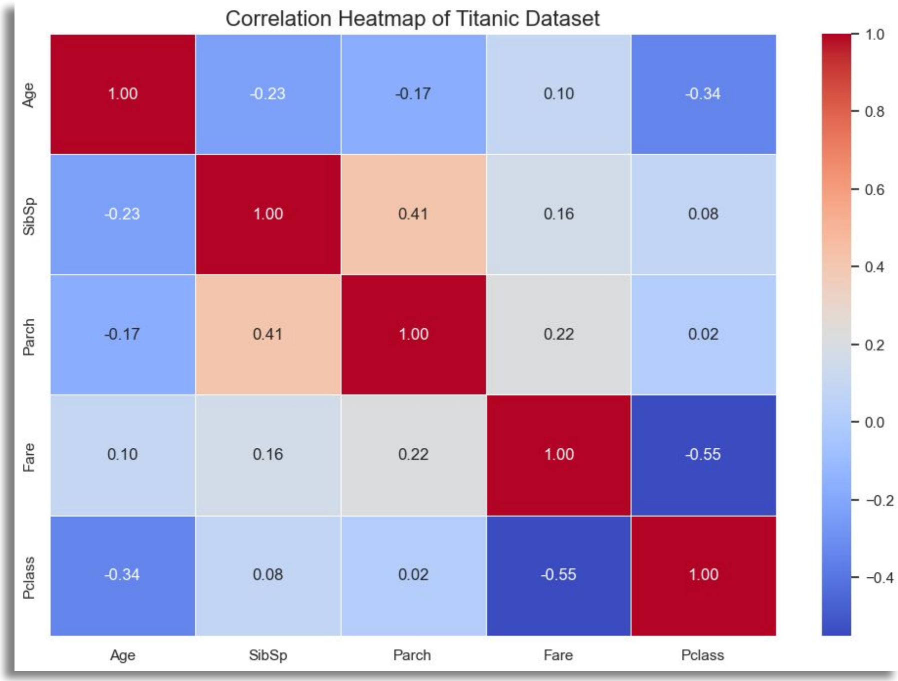
✓ 0.0s

| | PassengerID | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Pclass | Embarked | Survived |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | 3 | Southampton | Died |
| 1 | 2 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | 1 | Cherbourg | Survived |
| 2 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | 3 | Southampton | Survived |
| 3 | 4 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | 1 | Southampton | Survived |
| 4 | 5 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | 3 | Southampton | Died |

Pandas Dataframe created from tables in SQLite database

Luis

# Heatmap



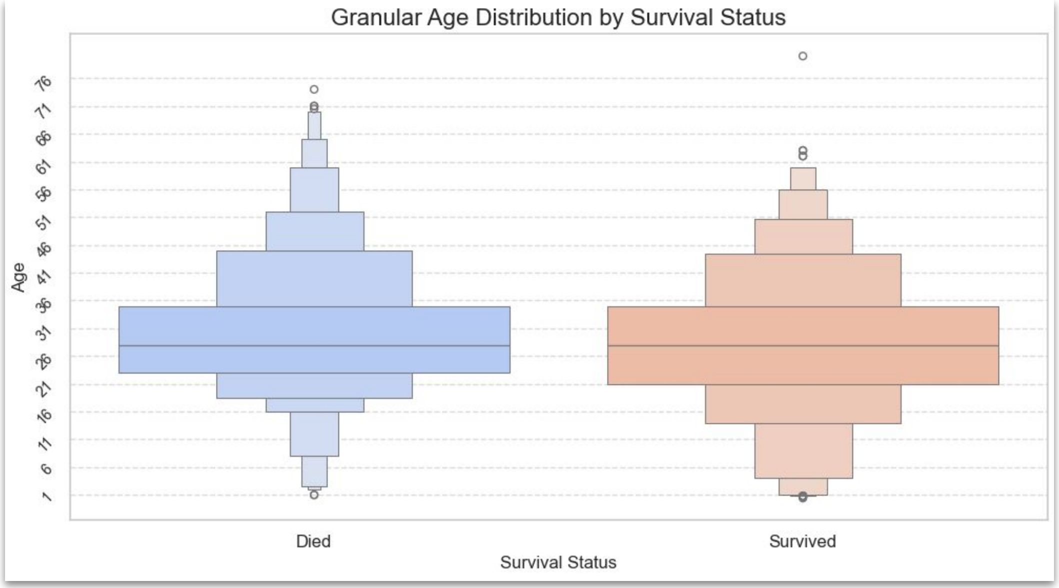
A heatmap of Titanic survival statistics highlights patterns and correlations between factors like age, gender, and passenger class.

It visually reveals trends, such as higher survival rates for females and children, especially in first class, while males and older passengers faced lower survival odds.

This makes it easier to identify key survival factors at a glance.

Oliver

# Boxenplot

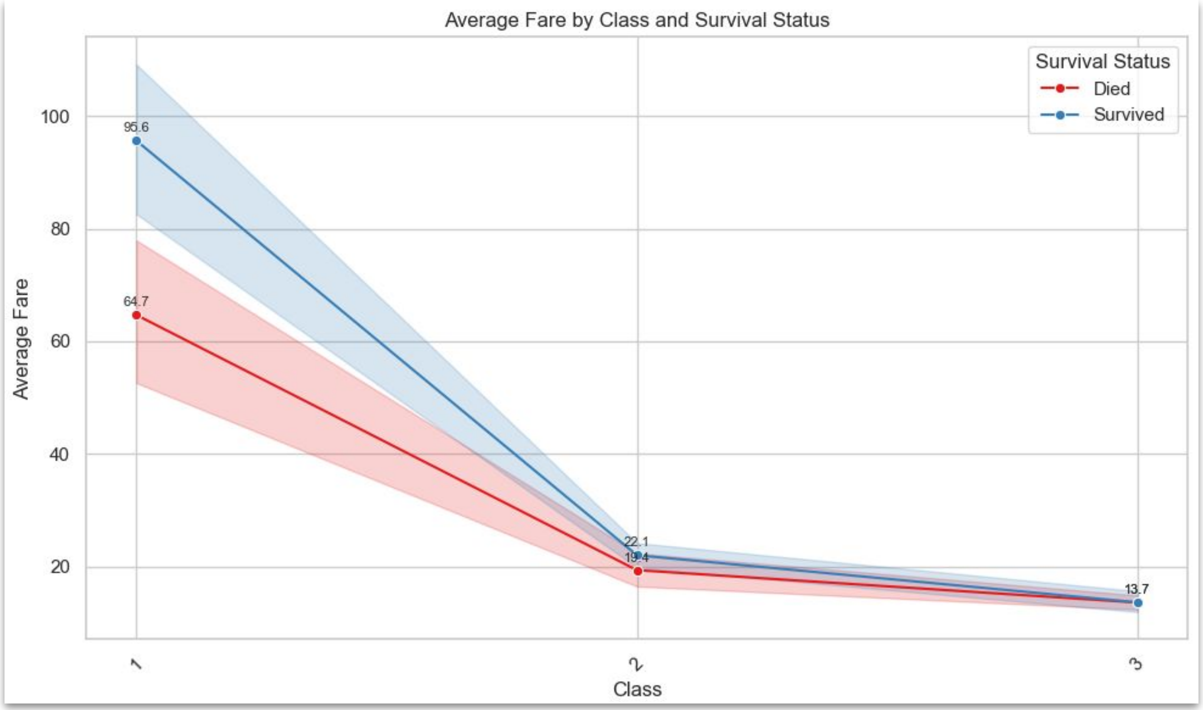


The Boxenplot revealed that younger passengers, especially children under 10, had higher survival rates, reflecting the "women and children first" evacuation rule. Survival rates declined for older age groups, highlighting age as a key factor in Titanic survival outcomes. The Boxenplot (or Letter Value Plot) does a better job of weighting the data to understand where a majority of survivors and fatalities were. The shape here indicates an exponential decline in survival by older passengers.

Luis

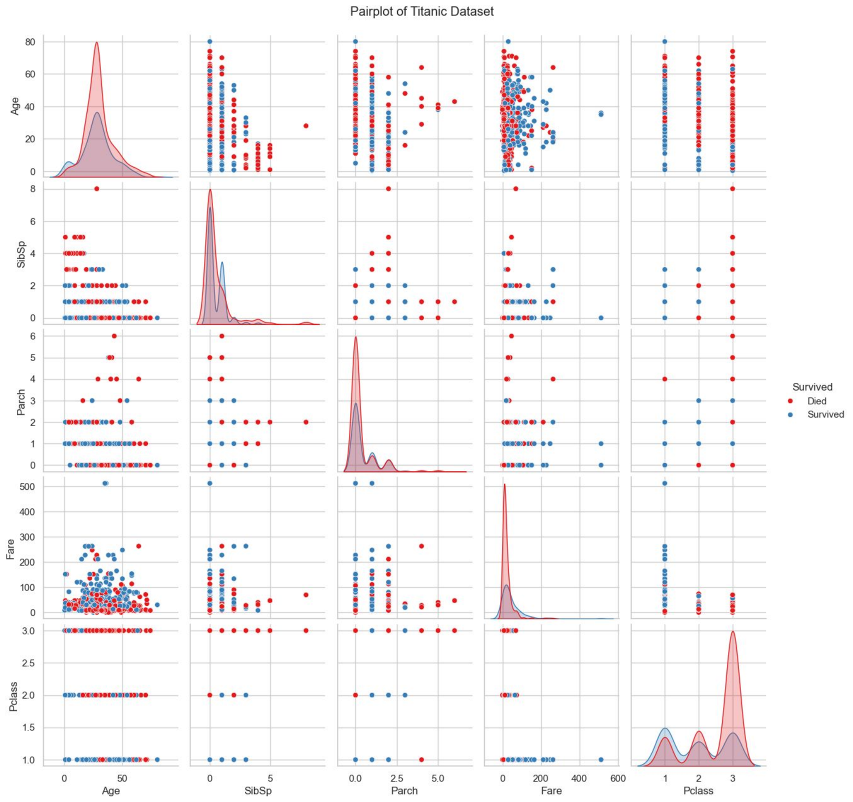
# Line Plot

**Average Fare by Class and Survival Status**
This graph shows the average fare paid by Titanic passengers across different classes, segmented by survival status. Passengers who survived generally paid higher fares, especially in 1st class where the average fare for survivors was around $95.6 compared to $64.7 for those who died. In 2nd and 3rd class, the fare difference between survivors and non-survivors is smaller. This suggests a correlation between higher fare (and class) and a greater chance of survival.



# Pairplot

Joe



- **Age Distribution:**
  - The KDE plot in the top left shows the distribution of age. It suggests a right-skewed distribution, meaning there are more younger passengers.
  - Comparing the KDEs for "Died" and "Survived" (red and blue lines), it looks like a higher proportion of younger passengers survived compared to older passengers.

- **SibSp Distribution:**
  - The KDE plot for SibSp shows that most passengers had 0 or 1 siblings/spouses.
  - The scatter plots of SibSp vs. other variables show that most data points are clustered at low SibSp values.

- **Parch Distribution:**
  - Similar to SibSp, the Parch distribution is skewed towards 0 and 1, indicating most passengers traveled with few parents/children.

- **Fare Distribution:**
  - The KDE plot for Fare shows a strong right skew, meaning most passengers paid low fares, with a few paying very high fares.
  - The scatter plots of Fare vs. other variables show that higher fares are associated with certain classes and survival patterns.

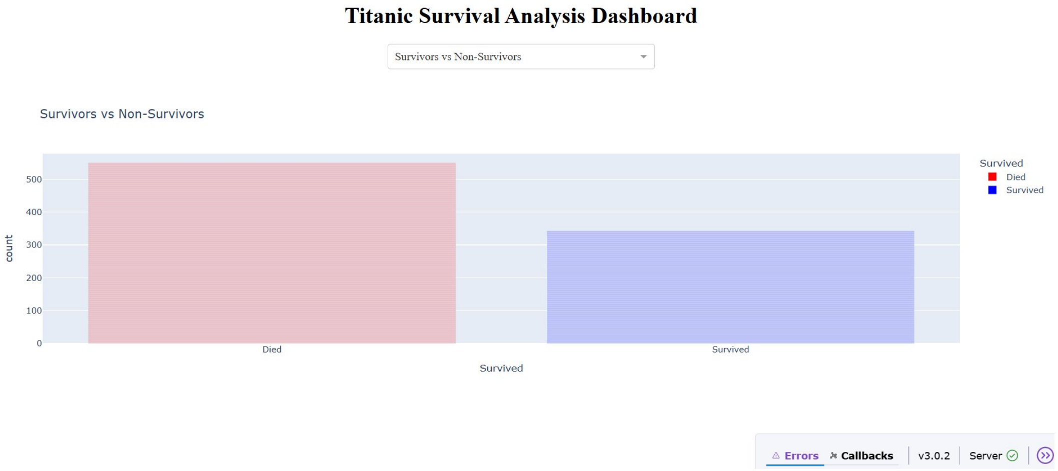
- **Pclass Distribution:**
  - The KDE plot for Pclass shows distinct peaks at 1, 2, and 3, indicating the discrete nature of passenger classes.
  - The scatter plots with Pclass show clear separation between classes, and survival rates vary significantly by class.

- **Relationships Between Variables:**
  - **Age vs. Survived:** As mentioned, younger passengers seem to have a higher survival rate.
  - **Pclass vs. Survived:** First-class passengers (Pclass=1) had a much higher survival rate than lower classes.
  - **Fare vs. Pclass:** Higher fares are strongly associated with first class.
  - **SibSp/Parch vs. Survived:** There might be some patterns, but it's less clear-cut than age or class.

**In essence, the pair plot provides a quick visual overview of the relationships between several key variables in the Titanic dataset, revealing patterns related to survival.**

Joe

# DASH - External Software Introduced



Titanic Survival Analysis Dashboard

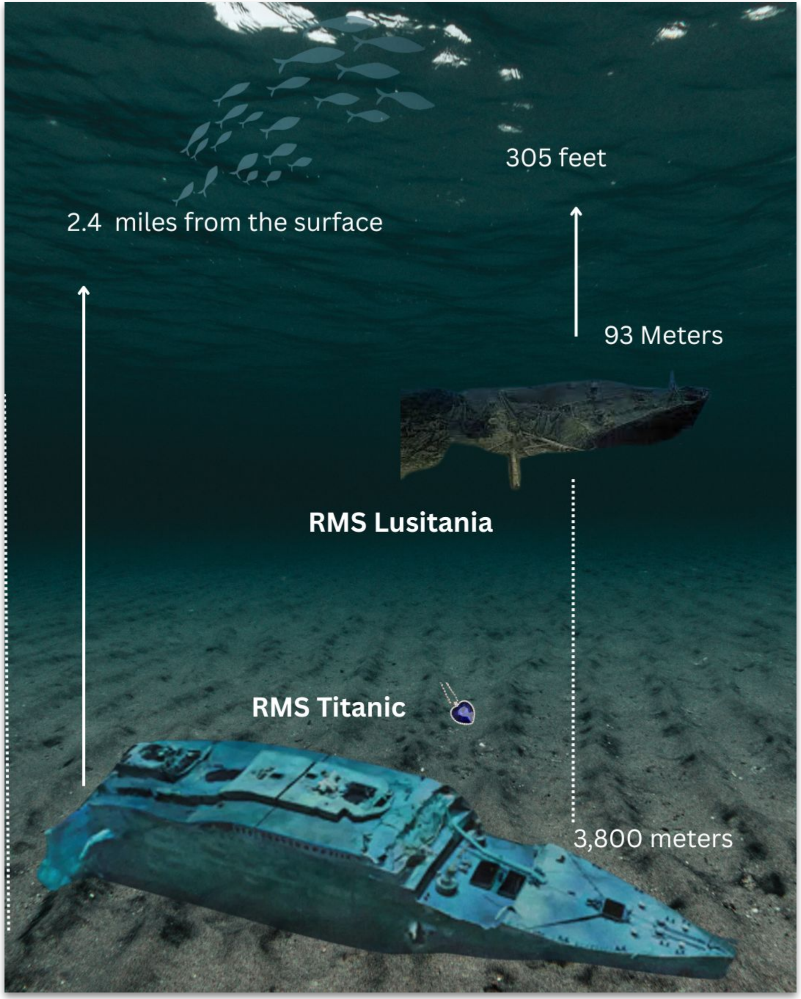
Survivors vs Non-Survivors

**Titanic Survival Analysis Dashboard**
An interactive data visualization tool built using Dash, designed to explore key patterns in the Titanic dataset. The dashboard allows users to select from a variety of graphs—such as survival rates by class, gender, age group, fare, and embarkation port—using a simple dropdown menu. It includes detailed visual insights like correlation heatmaps, box plots, and scatter plots to help uncover relationships between different variables and survival outcomes. Perfect for presentations, educational purposes, or quick exploratory data analysis.

http://127.0.0.1:8051/

# Potential Future Changes

For this project we focused only on one dataset, but we discussed that in future versions we would like to add more visualizations, or other datasets from different shipwrecks to compare data.



Leo

# Other User Interfaces Ideas

Involved using the the Titanic ship layout  to help visualize the data to enhanced the story telling aspect of data analytics. It would have provided  a unique and immersive perspective on the Titanic dataset by mapping the information directly onto a visual representation of the ship. This approach could have highlighted patterns and insights more effectively, such as passenger demographics, survival rates by deck or cabin class, or even the spatial distribution of lifeboat access.

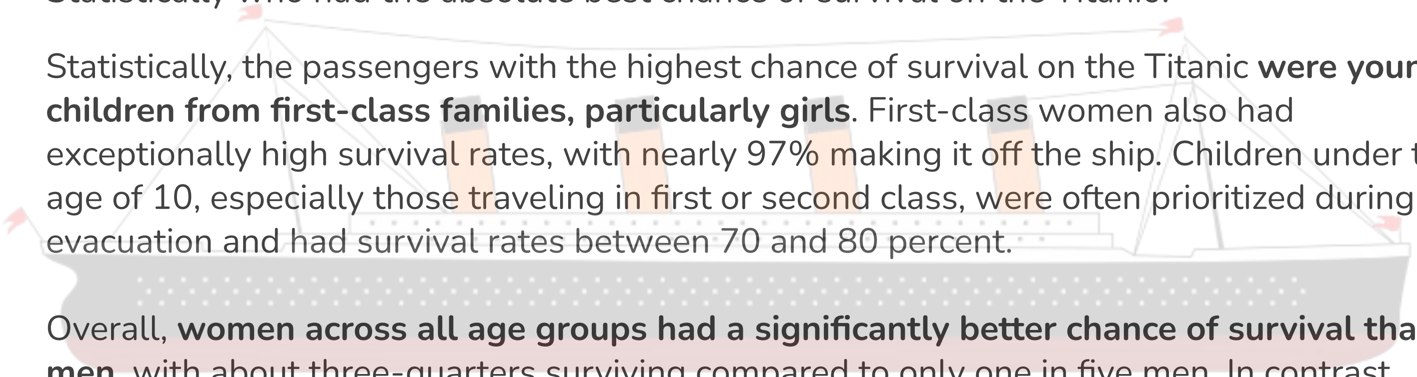
By overlaying data onto the ship's layout, the story-telling aspect of data analytics becomes more compelling and intuitive. For instance, instead of interpreting raw numbers or static charts, viewers could see the physical locations of passengers, visualize the impact of class divisions, and understand the dynamics of the tragedy in a way that resonates on a human level.

http://127.0.0.1:5500/Titanic.html

Leo

# Conclusions: *Women and Children First*

Statistically who had the absolute best chance of survival on the Titanic?



Statistically, the passengers with the highest chance of survival on the Titanic **were young children from first-class families, particularly girls**. First-class women also had exceptionally high survival rates, with nearly 97% making it off the ship. Children under the age of 10, especially those traveling in first or second class, were often prioritized during evacuation and had survival rates between 70 and 80 percent.

Overall, **women across all age groups had a significantly better chance of survival than men**, with about three-quarters surviving compared to only one in five men. In contrast, third-class men, older passengers, and lower-ranking crew members had the lowest survival odds due to limited access to lifeboats and delayed evacuation. Ultimately, being young, female, and in first class provided the greatest advantage for survival during the Titanic disaster.