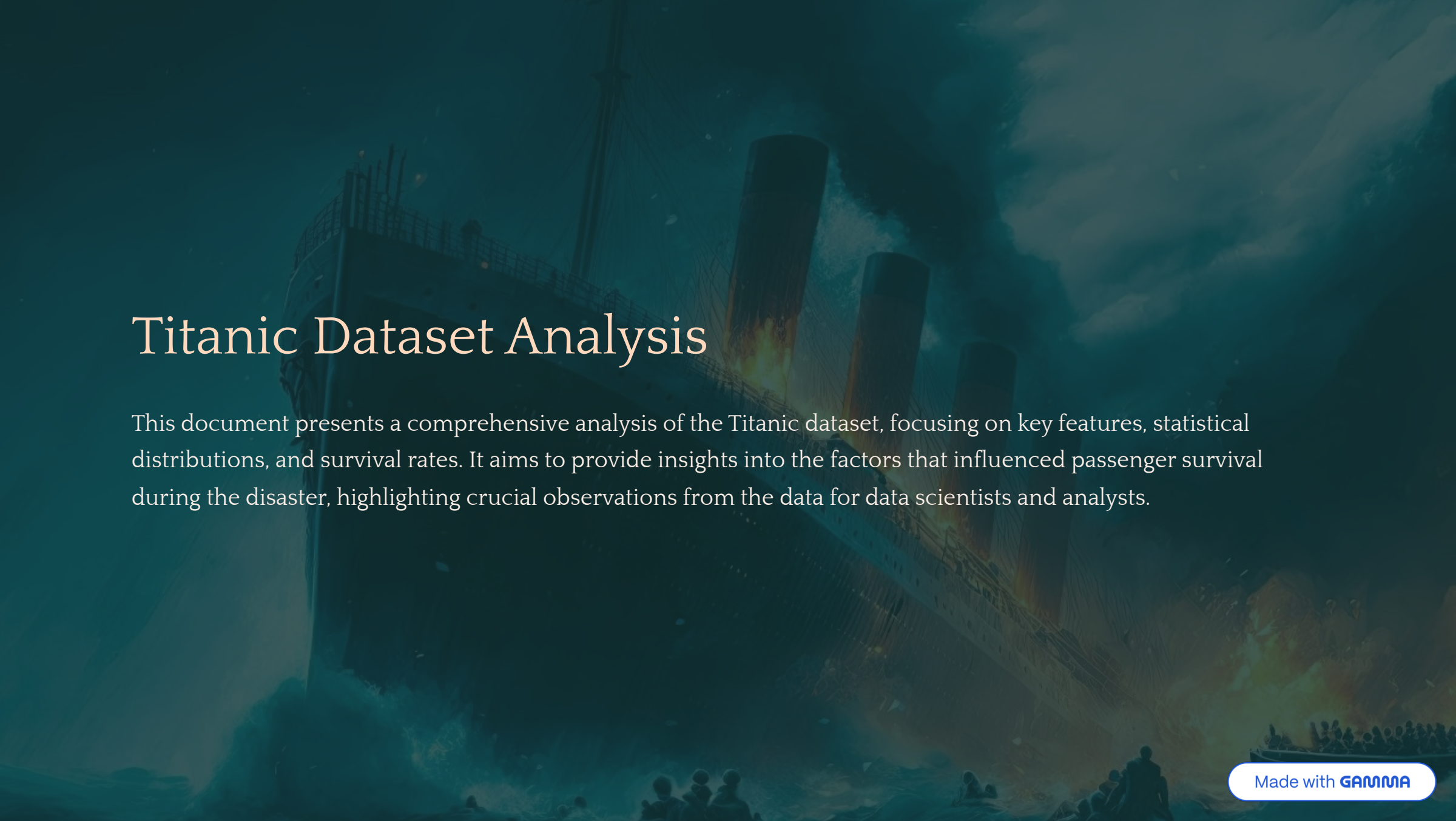

Titanic Dataset Analysis

This document presents a comprehensive analysis of the Titanic dataset, focusing on key features, statistical distributions, and survival rates. It aims to provide insights into the factors that influenced passenger survival during the disaster, highlighting crucial observations from the data for data scientists and analysts.

Titanic Dataset Analysis

This analysis is based on the Kaggle Titanic Competition dataset, which provides information on passengers aboard the RMS Titanic. The goal is to explore the dataset's characteristics and uncover patterns related to passenger survival. It serves as a foundational overview before deeper predictive modeling.

Dataset Overview

The Titanic dataset is composed of 891 rows, each representing a passenger, and 13 columns detailing various attributes. The target variable, 'Survived', indicates whether a passenger survived (1) or not (0). Notably, the dataset contains significant missing values, especially in the 'Cabin' and 'Age' columns, which requires careful handling during data preprocessing.

- **Rows:** 891
- **Columns:** 13
- **Target Variable:** Survived (0 = No, 1 = Yes)
- **Missing Values:**
  - Age: 177 missing
  - Cabin: 687 missing
  - Embarked: 2 missing

Key Columns

Understanding the role of each column is crucial for effective analysis. Features like 'Pclass', 'Sex', and 'Age' are particularly influential in predicting survival, while 'Fare' provides insight into socio-economic status. Family-related columns such as 'SibSp' and 'Parch' are essential for assessing group dynamics and their impact on survival outcomes.

- **PassengerId** – Unique ID
- **Pclass** – Ticket class (1 = Upper, 2 = Middle, 3 = Lower)
- **Name** – Passenger name
- **Sex** – Male/Female
- **Age** – Age in years
- **SibSp/Parch** – Number of siblings/spouses and parents/children aboard, respectively
- **Fare** – Ticket fare
- **Embarked** – Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Basic Statistics

A preliminary look at numerical features reveals distinct characteristics. The mean age of passengers was approximately 30 years, but the wide range from infants to seniors indicates a diverse age demographic. Similarly, ticket fares varied significantly, with most passengers paying relatively low amounts, but a few paid exceptionally high fares, highlighting economic disparities among passengers.

Age

- Mean: ~29.7 years
- Range: 0.42 – 80 years

Fare

- Mean: $32.20
- Max: $512.33
- 75% paid under $31

Categorical Distributions

The distribution of categorical variables offers critical insights into the passenger demographics. A significant majority of passengers were male and embarked from Southampton, reflecting common travel patterns of the era. The prevalence of third-class passengers underscores the Titanic's appeal across various economic strata, though this also suggests potential implications for survival rates.

Sex

- Male: 577
- Female: 314

Embarked

- Southampton (S): 644
- Cherbourg (C): 168
- Queenstown (Q): 77

Pclass

- Class 3: Most common (491 passengers)
- Class 1: Least common (216 passengers)

Observations

Initial observations suggest several factors that might influence survival. The predominance of male passengers in third class indicates a demographic imbalance that could affect overall survival statistics. The extensive missing data in 'Cabin' strongly correlates with third-class passengers, where cabin information might not have been meticulously recorded or was non-existent, posing a challenge for predictive models. The diverse age and fare ranges hint at complex interactions between socio-economic status and survival.

- Most passengers were in **3rd class** and **male**.
- A large number of passengers didn't have **Cabin** data (likely 3rd class).
- Age and fare show wide ranges, indicating varied socio-economic backgrounds.
- Missing values in **Age** and **Cabin** may impact modeling accuracy.

Survival Analysis: Introduction

Survival analysis delves into the critical outcome of whether a passenger survived the Titanic disaster. This section will present key findings by examining how different passenger attributes, such as gender and age, correlate with survival rates. These insights are crucial for understanding the human impact of the tragedy and for building predictive models.

Gender vs Survival

Gender played a significant role in survival rates, aligning with the "women and children first" protocol. Females had a considerably higher survival rate compared to males, indicating that gender was a primary factor determining who survived the disaster. This disparity is a key finding for predictive modeling and historical understanding.

Sex	Did Not Survive	Survived
Female	81	233
Male	468	109

> **Conclusion:** Females had a much higher survival rate than males.

Age vs Survival

Age also significantly influenced survival outcomes. Children, particularly those aged 0-12, showed a better chance of survival, reinforcing the "women and children first" policy. Conversely, the 21-40 age group experienced the highest number of fatalities, which could be attributed to this demographic's larger population size and potentially less priority during evacuation efforts.

Age Group	Did Not Survive	Survived
0–12	29	40
13–20	68	42
21–40	357	205
41–60	78	50
61–80	17	5

> **Conclusion:** Children (0–12) had a better survival chance. Most deaths occurred in the 21–40 age group.