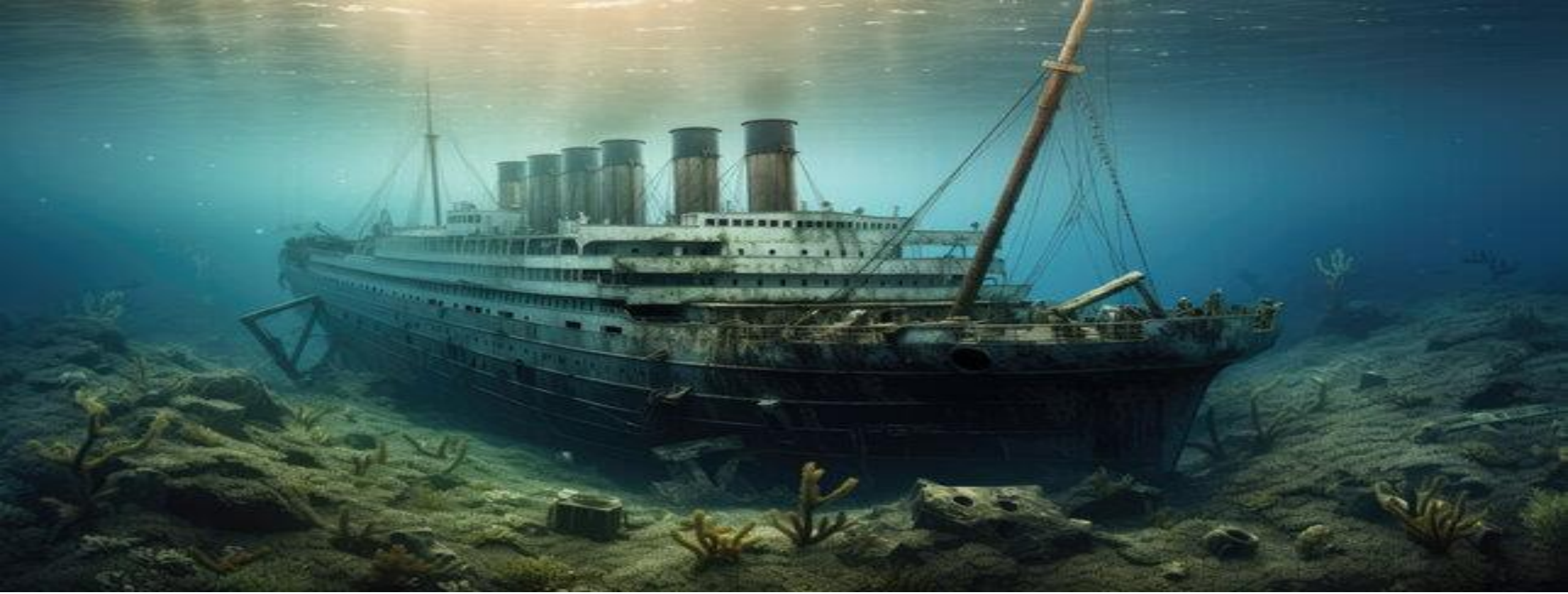

ANALYZING THE TITANIC DATASET

DAY 17 & 18 OF 30DAYSDUCKDBCHALLENGE

QUESTIONS & FINDINGS 👉

DuckDB

Question 1: Perform an analysis of survival rates based on fare in the Titanic dataset. Utilize the NTILE window function to evenly bucket passengers into 6 bins. Calculate statistics for each bin, including survival rates. Examine if there is a correlation between fare amounts and survival. Note any inconsistencies or noise in the fare column and present your findings.

My Notebook

Run my_db

```
1  WITH fare_bin AS (
```

Query executed in 676 ms. Row count: 6

Fare_Bin	Total_Passengers	Survived	Survival_Rate	Fare_Amount	Minimum_Fare	Maximum_Fare
1	219	53	0.24	1,487.64	0.00	7.78
2	218	45	0.21	1,735.05	7.78	8.66
3	218	69	0.32	2,529.28	8.66	14.45
4	218	94	0.43	4,431.32	14.45	26.00
5	218	94	0.43	7,547.05	26.00	53.10
6	218	139	0.64	25,820.14	53.10	512.33

FINDINGS ON Q1

1. As seen in the result, there is a general trend of increasing survival rates as fare amounts increase across the fare bins.

2. The minimum fare and maximum fare also vary across the fare bins. The minimum fare is 0.00, while the maximum fare varies from $7.775 to $512.33. This suggests that there was a wide range of fares available to passengers, regardless of their financial status.

3. There is a positive correlation between fare amounts and survival. As the fare bin increases, so does the survival rate. This suggests that passengers who paid higher fares were more likely to survive the sinking of the Titanic. This is likely due to the fact that passengers who paid higher fares had access to lifeboats.

4. There are a few inconsistencies and noise in the fare column. These include:

   - ✓ **Outliers:** The highest fare bin (Fare_Bin 6) shows a significant jump in both average fare amount (53.1 to 512.33) and survival rate (0.4312 to 0.64).
   - ✓ **Wide Range of Fares:** The range of fares is quite wide, from $0.00 to $512.33. This suggests that there was a significant difference in the financial status of the passengers on board the Titanic.
   - ✓ **Minimum Fare of $0.00:** The minimum fare in the data is $0.00, which seems unusual. It is possible that this is due to an error in the data collection process, or it could be that a very small number of passengers paid no fare at all.

Question 2: Conduct an analysis of survival rates based on sex in the Titanic dataset. Calculate the percentage of passengers who survived versus those who did not survive, focusing on the distinction between males and females. Express the survival rates and highlight any significant differences in survival ratios between genders.

My Notebook

Run my_db

```
1 SELECT
```

Query executed in 1.81 s. Row count: 2

Sex	Total_Passengers	SurvivalRate	%Survival	%Non_Survival
female	466	0.83	82.62	17.38
male	843	0.13	12.93	87.07

FINDINGS ON Q2

As seen from the result, there is a significant difference in survival rates between the genders. Females had a much higher survival rate of 82.62% compared to males with a survival rate of 12.93%. This is likely due to a number of factors, including:

- The women were prioritized during evacuation.
- Men may have been more likely to try to save others, putting themselves at risk.

DuckDB

Question 3: Explore the relationship between survival and age in the Titanic dataset. Calculate the survival rate for different age groups, providing insights into how age correlates with the likelihood of survival. Consider any notable patterns or trends in survival based on age.

My Notebook

Run | my_db

```
1  SELECT
```

Query executed in 5.03 s. Row count: 7

AgeGroup	TotalPassengers	TotalSurvived	SurvivalRate
0-10	82	47	57.32
11-20	162	66	40.74
21-30	360	129	35.83
31-40	208	89	42.79
41-50	129	48	37.21
51+	95	37	38.95
Unknown	10	1	10.00

FINDINGS ON Q3

1. As seen from the result, there is a correlation between age and survival rate. As age increases, the survival rate decreases
2. Several notable patterns and trends emerge from the analysis of survival rates by age group:

- Children (aged 0-10) have the highest survival rate (57.3%). This suggests that the prioritization of women and children in the boarding of lifeboats was particularly effective in saving young lives.

- The survival rate for younger adults (aged 11-20) remains relatively high (40.7%). This indicates that younger individuals were generally more physically fit and able to withstand the challenges of survival.

- The survival rate declines steadily with increasing age, particularly for older adults (51+). This suggests that older adults may have been more vulnerable to the cold and exposure.

- A notable exception to the age-survival correlation is the Survival Rate for the Age Group "Unknown" (Age 50+). With a survival rate of 10%, this group had a lower survival rate than the group immediately preceding it (Age 41-50). It's possible that this anomaly could be due to factors such as a smaller sample size or underlying health conditions within this group.