SQL QUERIES OF TITANIC SURVIVAL PREDICTION

DESCRIPTIVE ANALYSIS:-

1. How many passengers are there in the dataset?

Query:-SELECT COUNT(*) AS total_passengers FROM TitanicDataset;

Output:-892

2. What are the distinct values of socio-economic class (Pclass)?

Query:-SELECT DISTINCT c3 FROM TitanicDataset;

Output:-3

1

2

3.What is the average age of passengers? What is the minimum and maximum age?

```
Query:-SELECT AVG(c5) AS average_age,
      MIN(c5) AS minimum_age,
      MAX(c5) AS maximum_age
FROM TitanicDataset;
```

Output:-average_age=0

minimum_age=4

maximum_age=78

SURVIVAL ANALYSIS:-

1.What is the overall survival rate of passengers?

```
Query:- SELECT
    AVG(c1) AS overall_survival_rate
FROM
    TitanicDataset;
```

Output:- overall_survival_rate=445.5

2.How many passengers survived from each socio-economic class?

Query:- SELECT

```
    c1,
    SUM(c1) AS survived_count,
    COUNT(*) AS total_passengers,
    (SUM(c1) * 1.0 / COUNT(*)) AS survival_rate
FROM
    TitanicDataset
GROUP BY
    c1;
```

Output:-

c1	survived_count	total_passengers	survival_rate
1	1	1	1
10	10	1	10
100	100	1	100
101	101	1	101
102	102	1	102
103	103	1	103

3.What is the survival rate by gender?

Query:- SELECT

```
    c1,
    SUM(c1) AS survived_count,
    COUNT(*) AS total_passengers,
    (SUM(c1) * 1.0 / COUNT(*)) AS survival_rate
FROM
    TitanicDataset
GROUP BY
    c1;
```

Output:

c1	survived_count	total_passengers	survival_rate
1	1	1	1
10	10	1	10
100	100	1	100
101	101	1	101
102	102	1	102
103	103	1	103

-

Query:-SELECT

c2 AS Sex,

SUM(c1) AS survived_count,

COUNT(*) AS total_passengers,

(SUM(c1) * 1.0 / COUNT(*)) AS survival_rate

FROM

TitanicDataset

GROUP BY

c2;

Output:-

Sex	survived_count	total_passengers	survival_rate
0	245412	549	447.0163934426229
1	151974	342	444.36842105263156
Survived	0	1	0

AGE Distribution:-

1.How is the age distribution of passengers? Are there any age groups more likely to survive?

Query:- SELECT

CASE

WHEN c5 < 18 THEN '0-17'

WHEN c5 BETWEEN 18 AND 29 THEN '18-29'

WHEN c5 BETWEEN 30 AND 39 THEN '30-39'

WHEN c5 BETWEEN 40 AND 49 THEN '40-49'

```
        WHEN c5 BETWEEN 50 AND 59 THEN '50-59'
        WHEN c5 >= 60 THEN '60+'
        ELSE 'Unknown'
    END AS Age_Group,
    COUNT(*) AS Passengers_Count
FROM
    TitanicDataset
GROUP BY
    c5
ORDER BY
    c5;
```

Output:-

Age_Group	Passengers_Count
60+	1
60+	314
60+	577

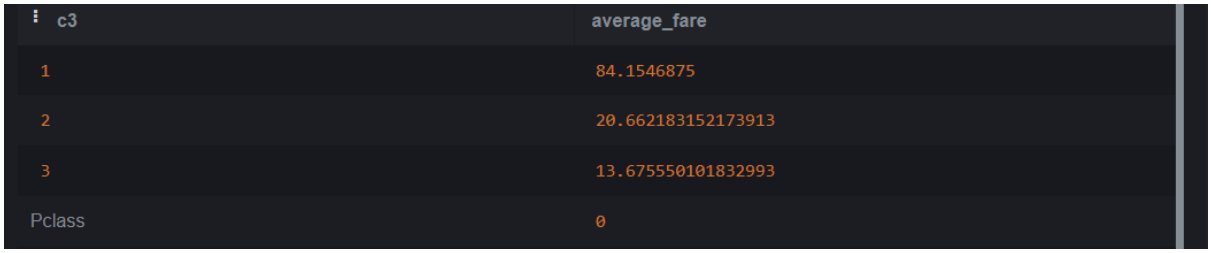
FARE ANALYSIS:-

1.What is the average fare paid by passengers in each socio-economic class?

Query:-

```
SELECT
    c3,
    AVG(c10) AS average_fare
FROM
    TitanicDataset
GROUP BY
    c3;
```

Output:



c3	average_fare
1	84.1546875
2	20.662183152173913
3	13.675550101832993
Pclass	0

-

2.How does the fare vary for passengers who survived vs. those who didn't?

Query:- SELECT

c2,

AVG(c10) AS average_fare

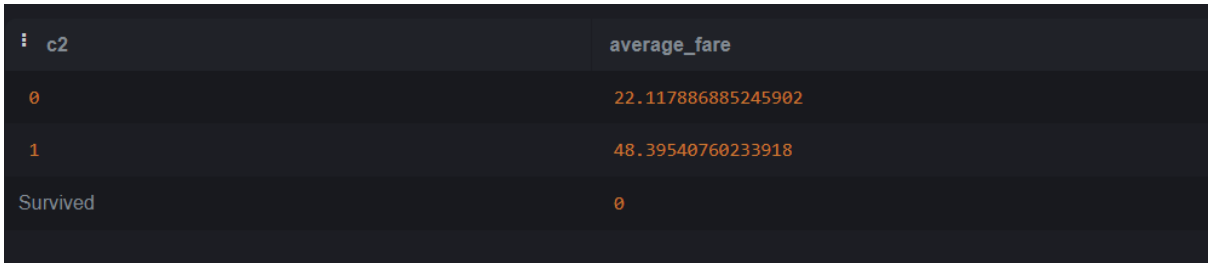
FROM

TitanicDataset

GROUP BY

c2;

Output:-



c2	average_fare
0	22.117886885245902
1	48.39540760233918
Survived	0

FAMILY ANALYSIS:-

1.How many passengers were traveling with family members (siblings, spouses, parents, children)?

Query:- SELECT

COUNT(*) AS passengers_with_family

FROM

TitanicDataset

WHERE

c6 > 0 OR c7 > 0;

Output:- passengers_with_family=755

2.Did passengers traveling with family members have a higher survival rate?

Query:- WITH FamilyPassengers AS (

```
    SELECT
        CASE
            WHEN c6 > 0 OR c7 > 0 THEN 'With Family'
            ELSE 'Alone'
        END AS Travel_Status,
        c2
    FROM
        TitanicDataset
)
SELECT
    Travel_Status,
    AVG(c2) AS Survival_Rate
FROM
    FamilyPassengers
GROUP BY
    Travel_Status;
```

Output:-

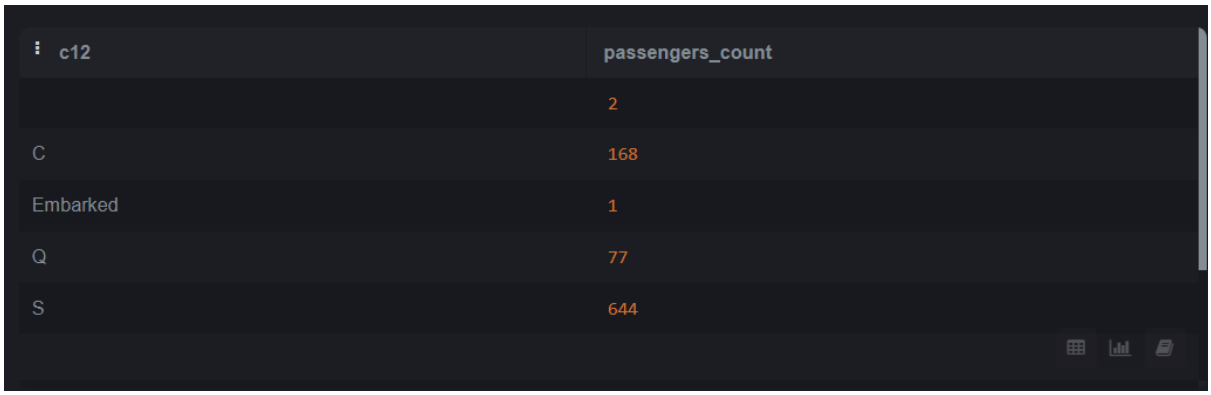
Travel_Status	Survival_Rate
Alone	0.25547445255474455
With Family	0.4066225165562914

PORT OF EMBARKATION:-

1.How many passengers embarked from each port (Cherbourg, Queenstown, Southampton)?

```
Query:- SELECT
    c12,
    COUNT(*) AS passengers_count
FROM
    TitanicDataset
GROUP BY
    c12;
```

Output:-



c12	passengers_count
	2
C	168
Embarked	1
Q	77
S	644

2.Is there a correlation between the port of embarkation and survival rate?

Query:- SELECT

```
c12,
AVG(c2) AS Survival_Rate
```

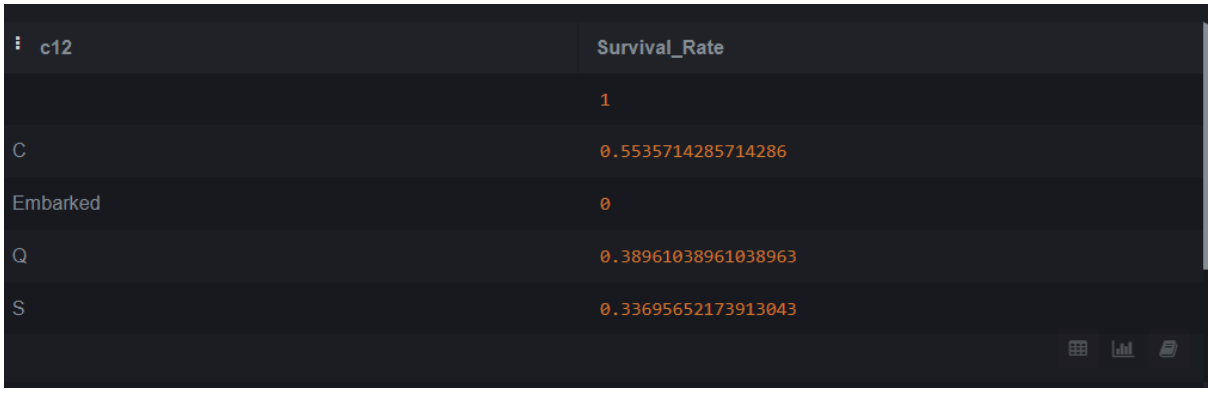
FROM

```
TitanicDataset
```

GROUP BY

```
c12;
```

Output:-



c12	Survival_Rate
	1
C	0.5535714285714286
Embarked	0
Q	0.38961038961038963
S	0.33695652173913043

MISSING DATA:-

1.Are there any missing values in the dataset? If so, in which columns?

Query:- SELECT

```
CASE
    WHEN COUNT(*) = COUNT(c6) THEN 'No missing values'
    ELSE 'Missing values'
```

```
    END AS missing_status,
    c6 AS column_name
FROM
    TitanicDataset;
```

Output:-

No Missing Values

CORRELATION ANALYSIS:-

1.Is there a correlation between socio-economic class and survival rate?

Query:-

```
SELECT
    c3,
    AVG(c2) AS Survival_Rate
FROM
    TitanicDataset
GROUP BY
    c3;
```

Output:-

c3	Survival_Rate
1	0.6296296296296297
2	0.47282608695652173
3	0.24236252545824846
Pclass	0

2.Is there a correlation between age and survival rate?

Query:-

```
SELECT
    CASE
        WHEN c5 < 18 THEN '0-17'
        WHEN c5 BETWEEN 18 AND 29 THEN '18-29'
        WHEN c5 BETWEEN 30 AND 39 THEN '30-39'
```

```
        WHEN c5 BETWEEN 40 AND 49 THEN '40-49'
        WHEN c5 BETWEEN 50 AND 59 THEN '50-59'
        WHEN c5 >= 60 THEN '60+'
        ELSE 'Unknown'
    END AS Age_Group,
    AVG(c2) AS Survival_Rate
FROM
    TitanicDataset
GROUP BY
    Age_Group
ORDER BY
    MIN(c5);
```

Output:-

Age_Group	Survival_Rate
60+	0.3834080717488789

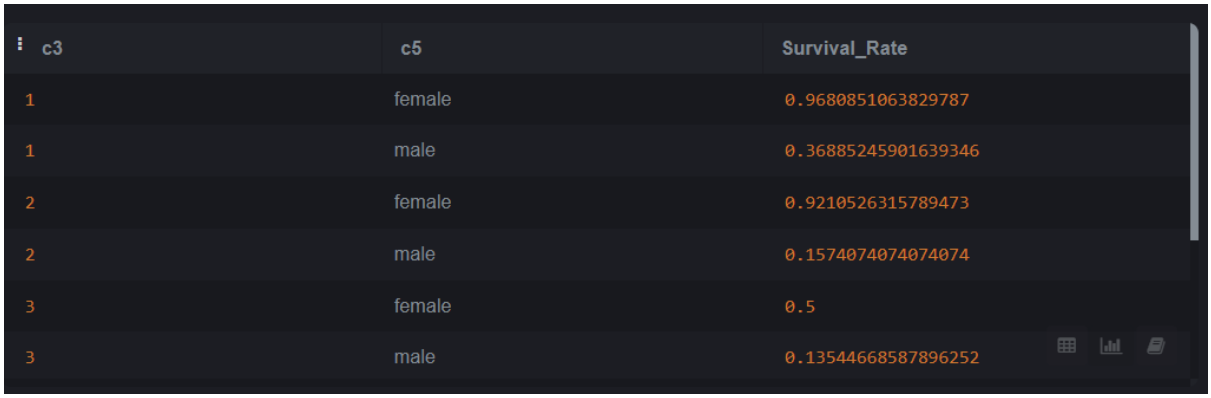
CLASS AND GENDER ANALYSIS:-

1.What is the survival rate for each combination of socio-economic class and gender?

Query:-

```
SELECT
    c3,
    c5,
    AVG(c2) AS Survival_Rate
FROM
    TitanicDataset
GROUP BY
    c3, c5;
```

Output:-



c3	c5	Survival_Rate
1	female	0.9680851063829787
1	male	0.36885245901639346
2	female	0.9210526315789473
2	male	0.1574074074074074
3	female	0.5
3	male	0.13544668587896252

COMPARATIVE ANALYSIS:-

1.How does the survival rate differ between adults and children?

Query:-

```
SELECT
  CASE
    WHEN c6 < 18 THEN 'Child'
    ELSE 'Adult'
  END AS Age_Group,
  AVG(c2) AS Survival_Rate
FROM
  TitanicDataset
GROUP BY
  Age_Group;
```

Output:-

Age_Group	Survival_Rate
Adult	0.3923076923076923
Child	0.359504132231405

2.How does the survival rate differ between passengers traveling alone and those traveling with family?

Query:-

```
SELECT
  CASE
    WHEN c6 + c7 > 0 THEN 'With Family'
```

```
        ELSE 'Alone'
    END AS Travel_Status,
    AVG(c2) AS Survival_Rate
FROM
    TitanicDataset
GROUP BY
    Travel_Status;
```

Output:-

Travel_Status	Survival_Rate
Alone	0.2536231884057971
With Family	0.40716180371352784