WEEK 4 ASSIGNMENT

Exploratory Data Analysis on Titanic Dataset

Objective

The primary objective of this analysis is to conduct an in-depth Exploratory Data Analysis (EDA) on the Titanic dataset. The focus is on:

- Understanding data distributions
- Identifying and handling missing values
- Detecting outliers
- Exploring relationships between variables
- Visualizing patterns through plots such as histograms, box plots, and heatmaps

---

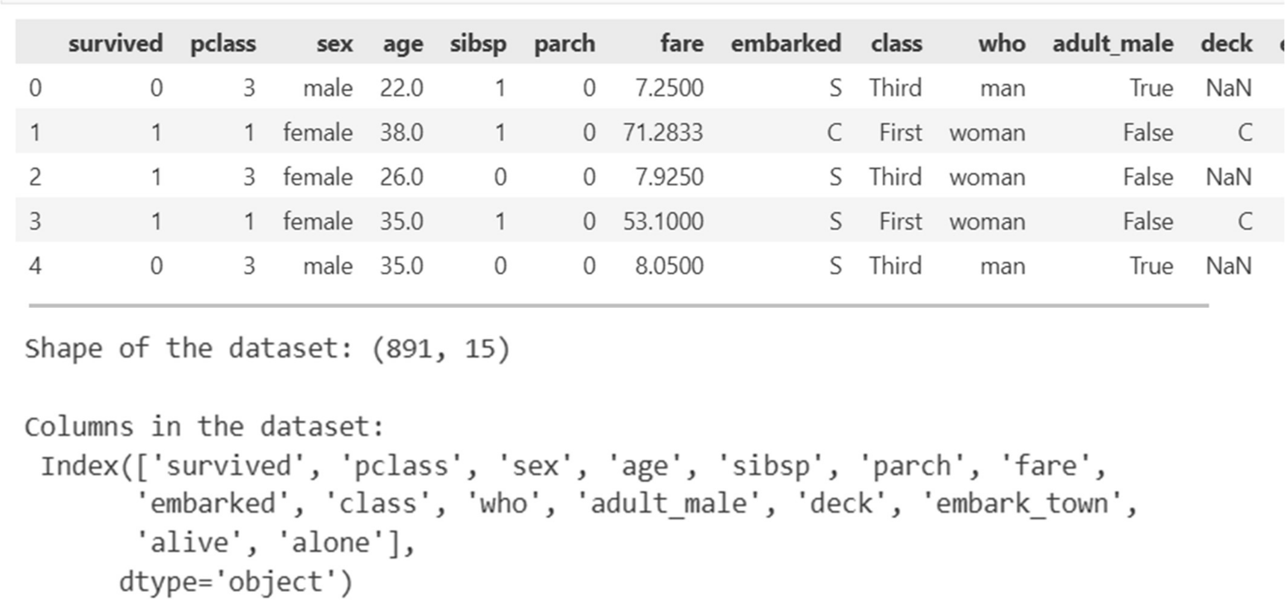
1. Importing Required Libraries

Essential libraries for data manipulation, visualization, and imputation are imported in this section.

2. Dataset Overview

Basic information about the Titanic dataset including:

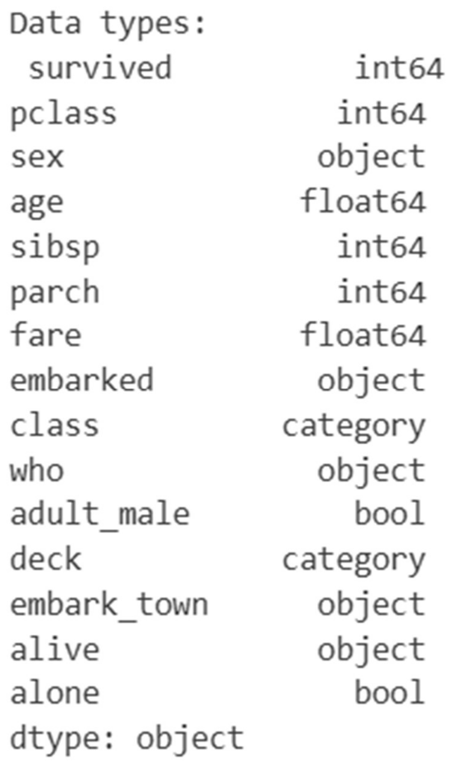
- First few rows
- Shape
- Column names
- Data types
- Descriptive statistics



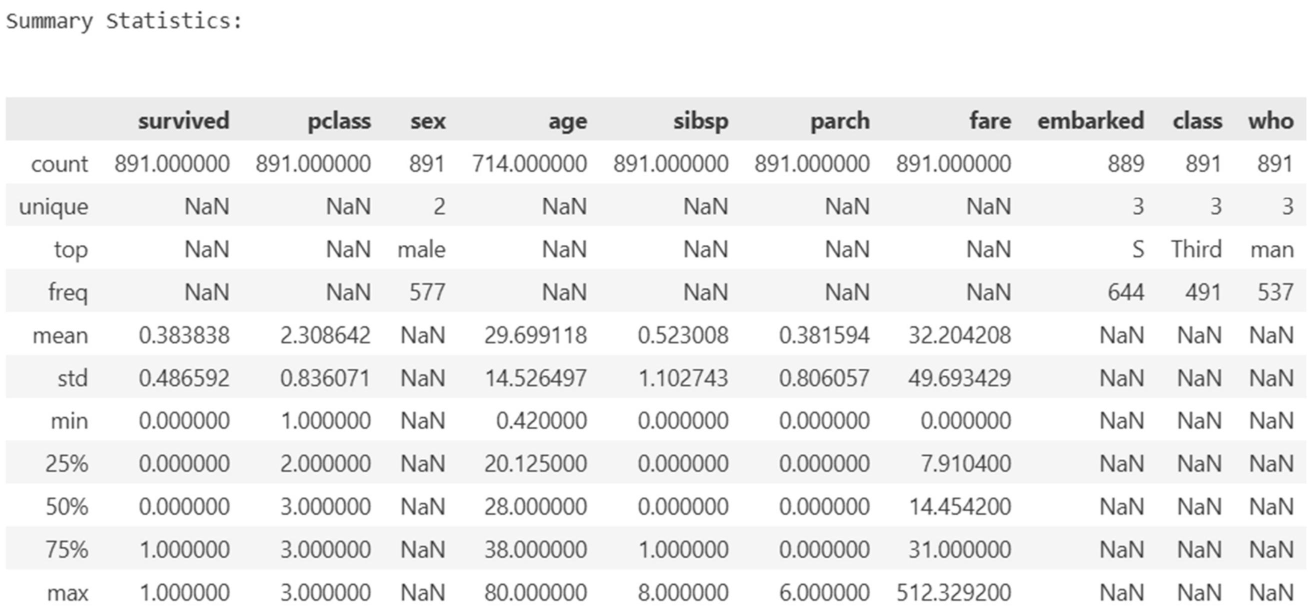
	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN

```
Shape of the dataset: (891, 15)

Columns in the dataset:
 Index(['survived', 'pclass', 'sex', 'age', 'sibsp', 'parch', 'fare',
       'embarked', 'class', 'who', 'adult_male', 'deck', 'embark_town',
       'alive', 'alone'],
      dtype='object')
```



```
Data types:
 survived          int64
pclass            int64
sex              object
age             float64
sibsp             int64
parch             int64
fare            float64
embarked         object
class          category
who              object
adult_male         bool
deck           category
embark_town      object
alive            object
alone              bool
dtype: object
```



Summary Statistics:

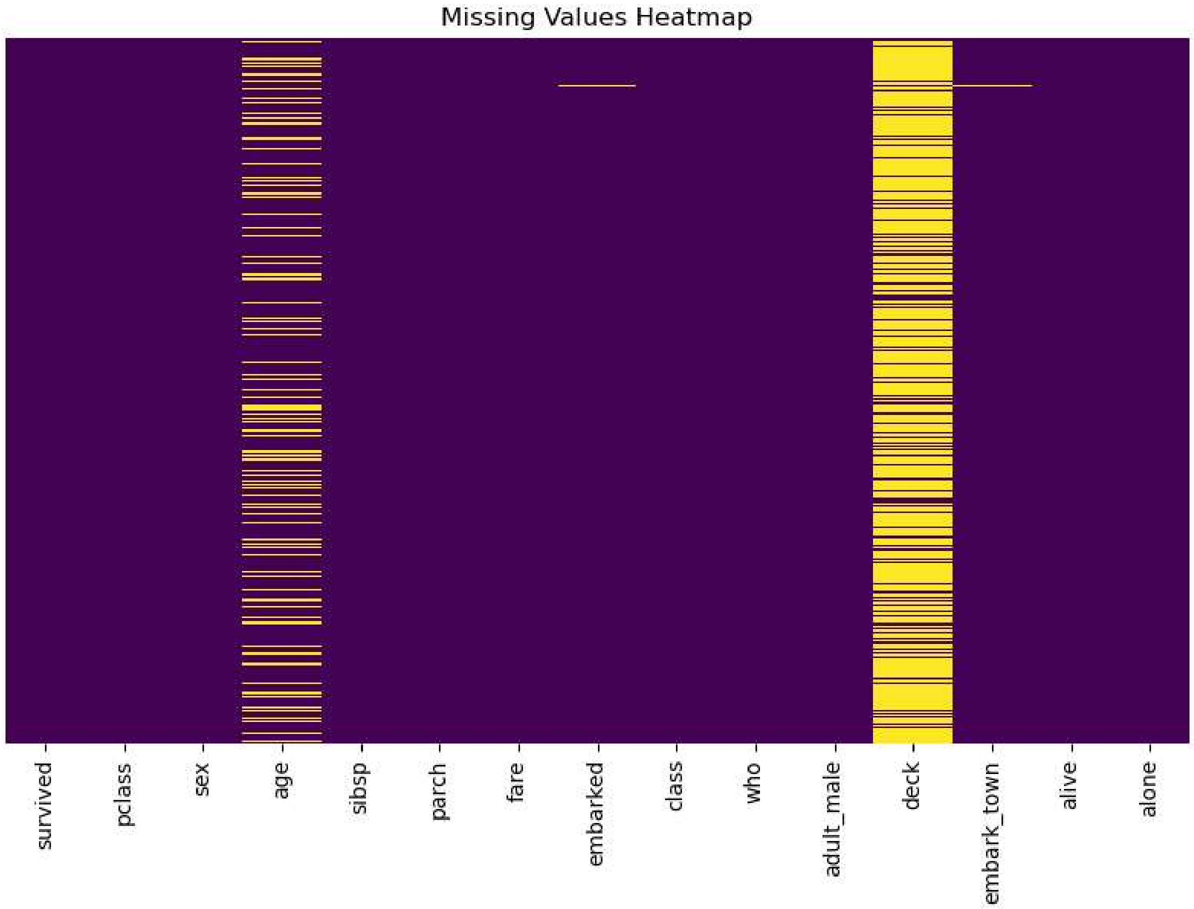
	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who
count	891.000000	891.000000	891	714.000000	891.000000	891.000000	891.000000	889	891	891
unique	NaN	NaN	2	NaN	NaN	NaN	NaN	3	3	3
top	NaN	NaN	male	NaN	NaN	NaN	NaN	S	Third	man
freq	NaN	NaN	577	NaN	NaN	NaN	NaN	644	491	537
mean	0.383838	2.308642	NaN	29.699118	0.523008	0.381594	32.204208	NaN	NaN	NaN
std	0.486592	0.836071	NaN	14.526497	1.102743	0.806057	49.693429	NaN	NaN	NaN
min	0.000000	1.000000	NaN	0.420000	0.000000	0.000000	0.000000	NaN	NaN	NaN
25%	0.000000	2.000000	NaN	20.125000	0.000000	0.000000	7.910400	NaN	NaN	NaN
50%	0.000000	3.000000	NaN	28.000000	0.000000	0.000000	14.454200	NaN	NaN	NaN
75%	1.000000	3.000000	NaN	38.000000	1.000000	0.000000	31.000000	NaN	NaN	NaN
max	1.000000	3.000000	NaN	80.000000	8.000000	6.000000	512.329200	NaN	NaN	NaN

3. Missing Value Analysis

Evaluate the amount and proportion of missing data in each column. A heatmap is used to visualize the missing patterns.

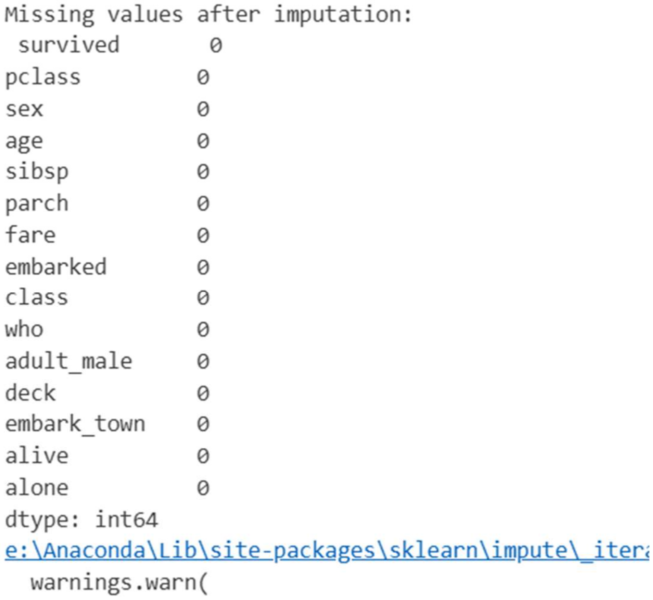
```
Missing values in each column:
 survived         0
pclass          0
sex             0
age           177
sibsp           0
parch           0
fare            0
embarked        2
class           0
who             0
adult_male      0
deck          688
embark_town     2
alive           0
alone           0
dtype: int64
```

```
Percentage of missing values in each column:
 survived        0.000000
pclass          0.000000
sex             0.000000
age            19.865320
sibsp           0.000000
parch           0.000000
fare            0.000000
embarked        0.224467
class           0.000000
who             0.000000
adult_male      0.000000
deck           77.216611
embark_town     0.224467
alive           0.000000
alone           0.000000
dtype: float64
```



4. Missing Value Imputation

- Numerical values are imputed using IterativeImputer with a RandomForestRegressor.
- Categorical columns are imputed using the mode.

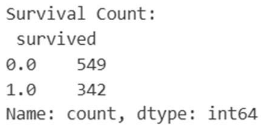


```
Missing values after imputation:
 survived      0
pclass        0
sex           0
age           0
sibsp         0
parch         0
fare          0
embarked      0
class         0
who           0
adult_male    0
deck          0
embark_town   0
alive         0
alone         0
dtype: int64
e:\Anaconda\Lib\site-packages\sklearn\impute\_iter
  warnings.warn(
```

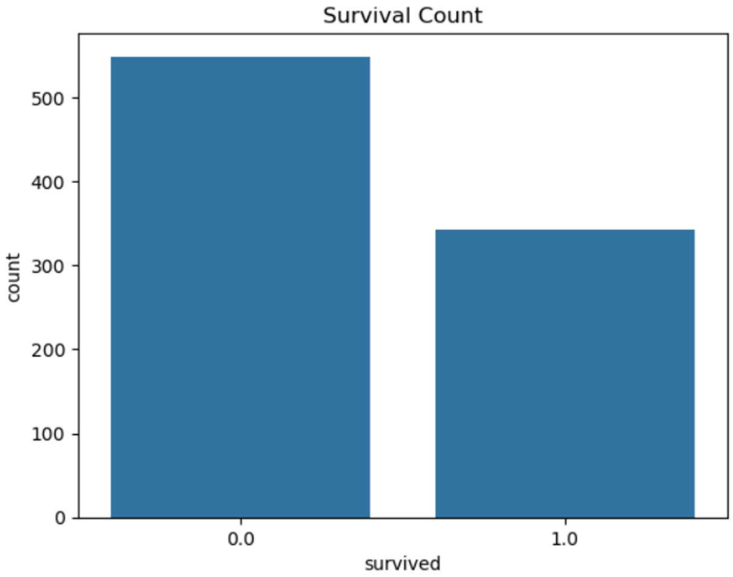
5. Univariate Analysis

5.1. Survival Count

Shows the distribution of survivors and non-survivors.

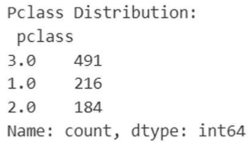


```
Survival Count:
 survived
0.0    549
1.0    342
Name: count, dtype: int64
```

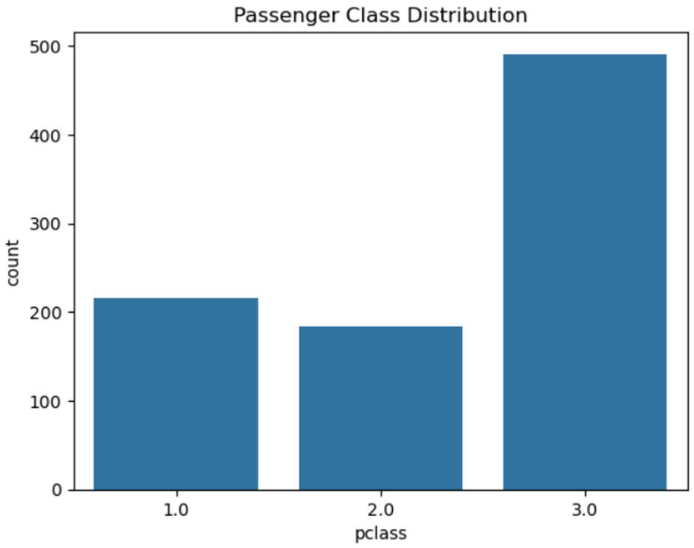


5.2. Passenger Class Distribution

Analyzes the distribution of passenger classes.

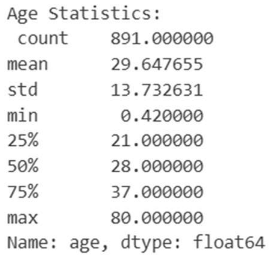


```
Pclass Distribution:
 pclass
3.0    491
1.0    216
2.0    184
Name: count, dtype: int64
```

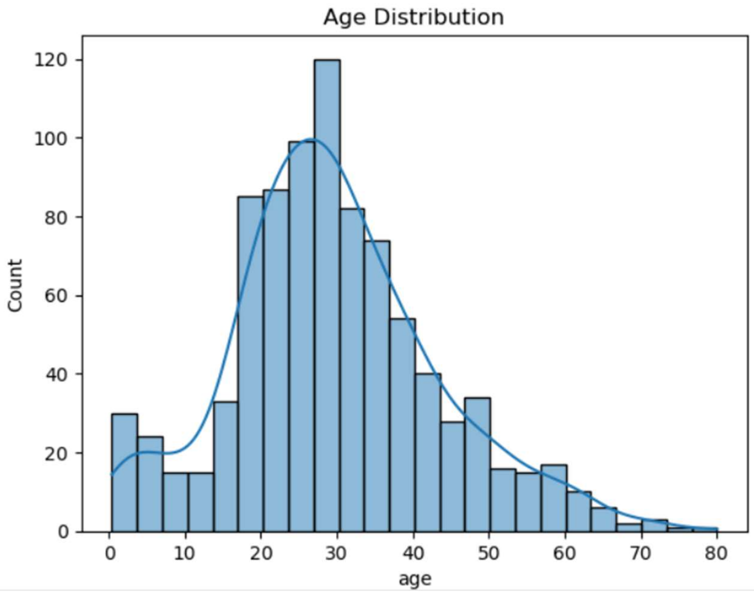


5.3. Age Distribution

Histogram with KDE plot to show the age distribution.

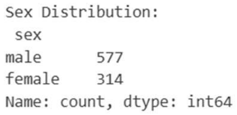


```
Age Statistics:
 count    891.000000
mean      29.647655
std       13.732631
min        0.420000
25%       21.000000
50%       28.000000
75%       37.000000
max       80.000000
Name: age, dtype: float64
```

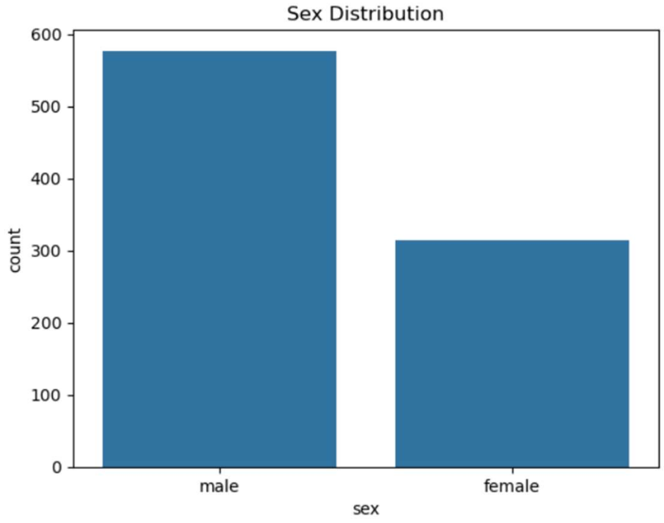


5.4. Gender Distribution

Analyzes the sex distribution of passengers.



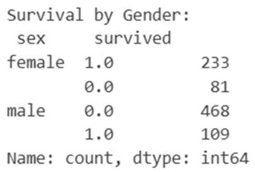
```
Sex Distribution:
 sex
male      577
female    314
Name: count, dtype: int64
```



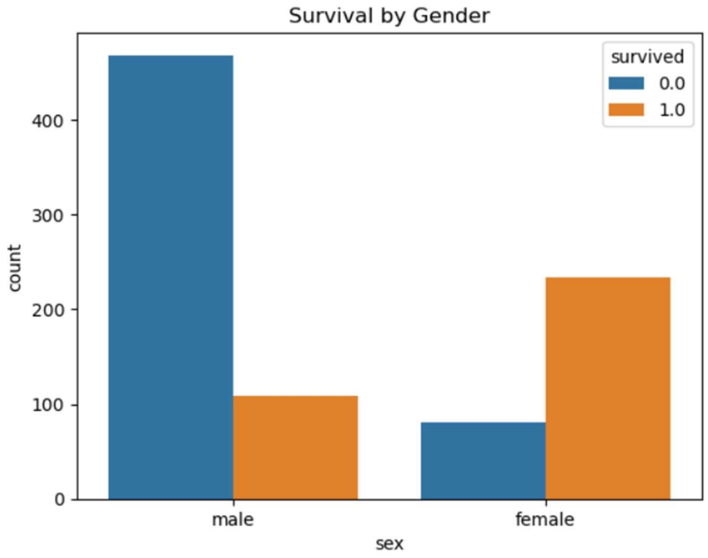
6. Bivariate Analysis

6.1. Survival by Gender

Cross-tabulation and bar chart showing survival rates by gender.

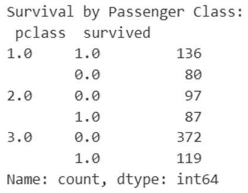


```
Survival by Gender:
 sex     survived
female  1.0         233
        0.0          81
male    0.0         468
        1.0         109
Name: count, dtype: int64
```

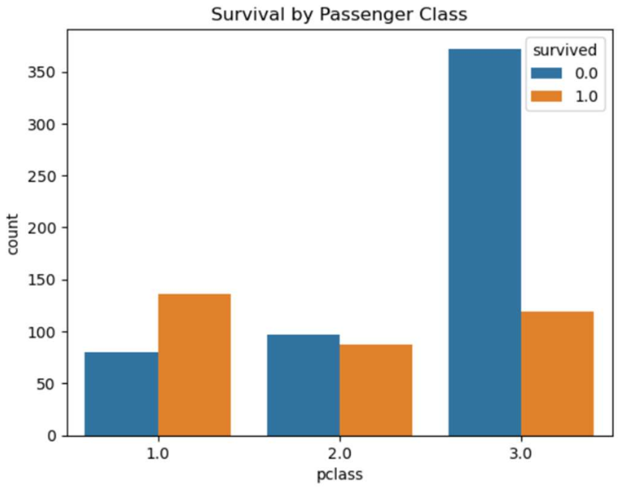


6.2. Survival by Passenger Class

Visual analysis of survival rates based on class.

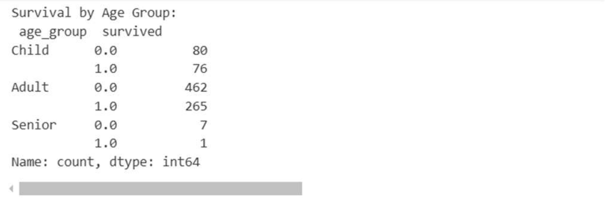


```
Survival by Passenger Class:
 pclass  survived
1.0     1.0         136
        0.0          80
2.0     0.0          97
        1.0          87
3.0     0.0         372
        1.0         119
Name: count, dtype: int64
```

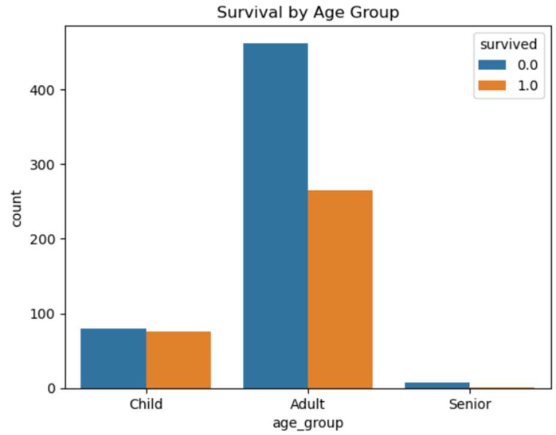


6.3. Survival by Age Group

Categorizes passengers into age groups and compares survival rates.



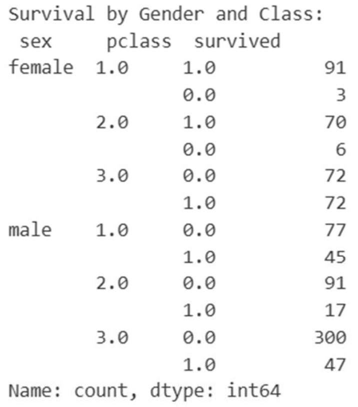
```
Survival by Age Group:
 age_group  survived
Child      0.0          80
           1.0          76
Adult      0.0         462
           1.0         265
Senior     0.0           7
           1.0           1
Name: count, dtype: int64
```



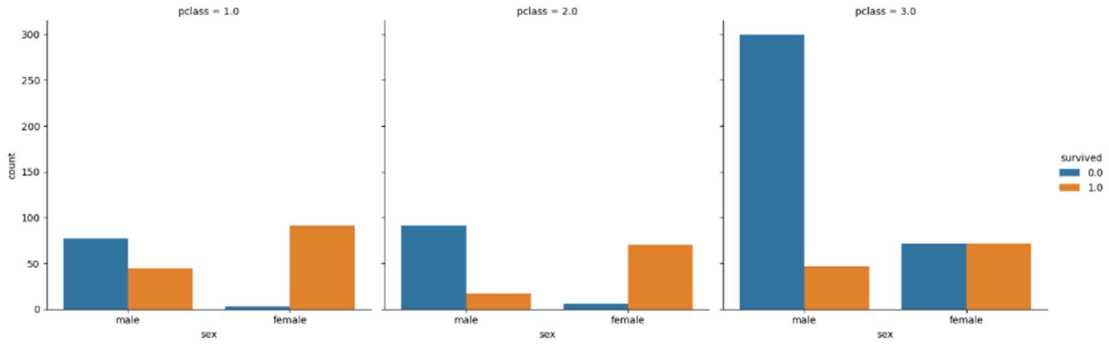
7. Multivariate Analysis

7.1. Survival by Gender and Class

Shows a more granular breakdown using combinations of gender and class.



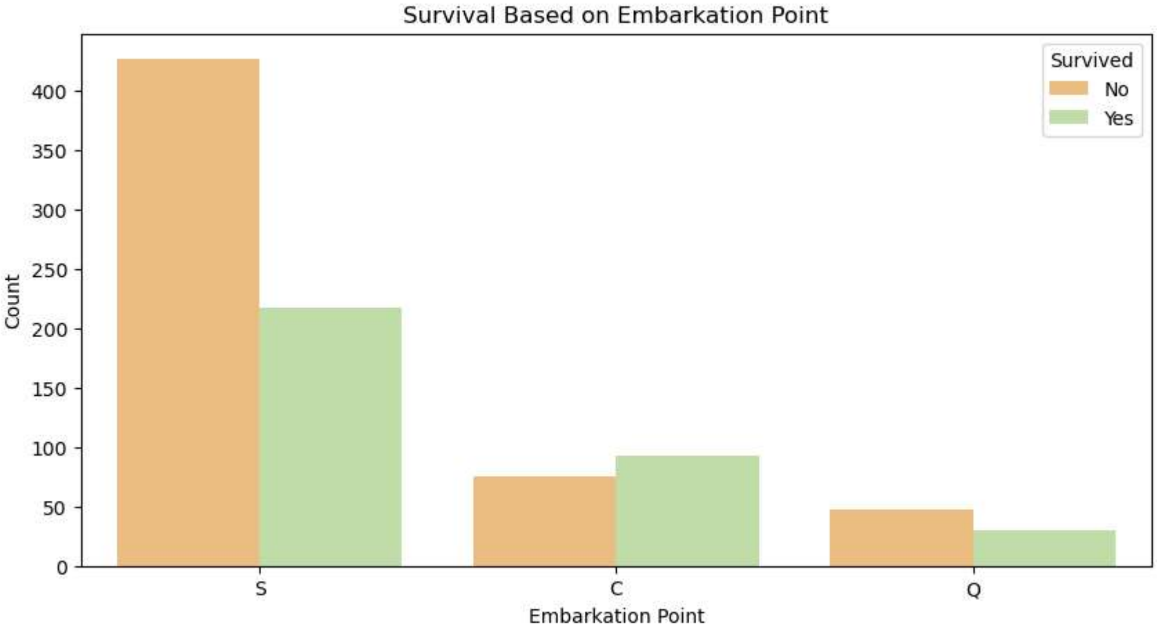
```
Survival by Gender and Class:
 sex     pclass  survived
female  1.0     1.0          91
                0.0           3
        2.0     1.0          70
                0.0           6
        3.0     0.0          72
                1.0          72
male    1.0     0.0          77
                1.0          45
        2.0     0.0          91
                1.0          17
        3.0     0.0         300
                1.0          47
Name: count, dtype: int64
```



8. Additional Insights

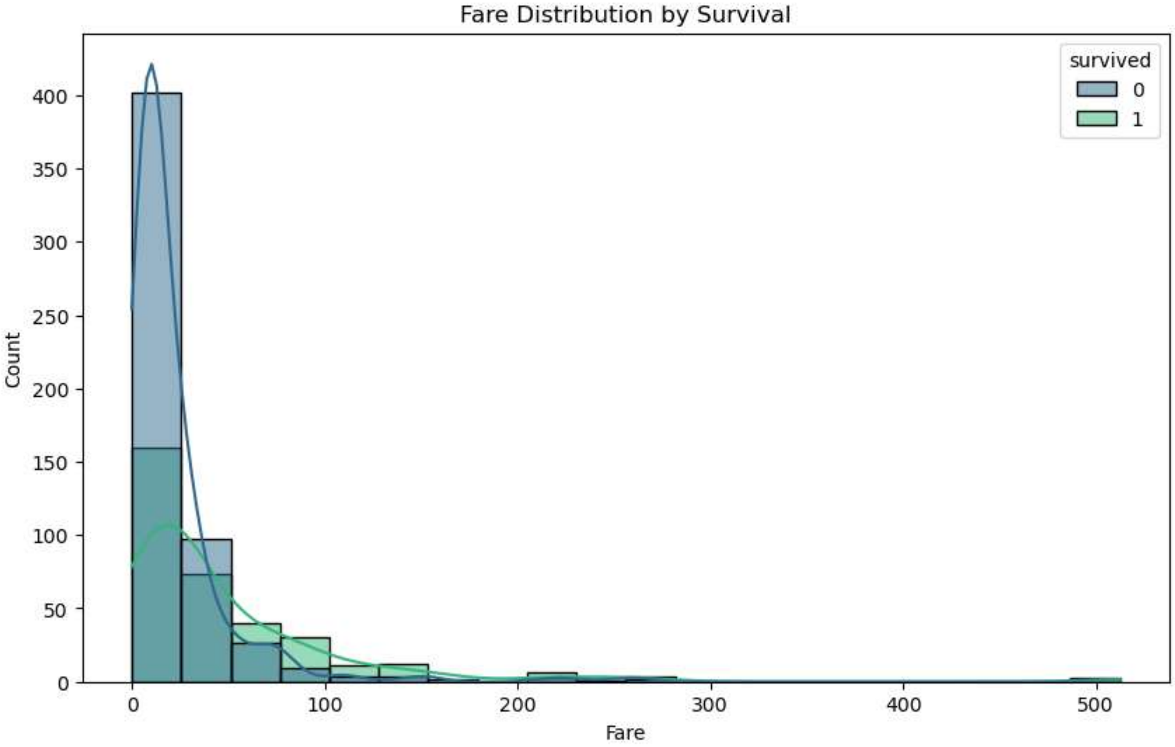
8.1. Survival Based on Embarkation Point

Examines how embarkation location correlates with survival.



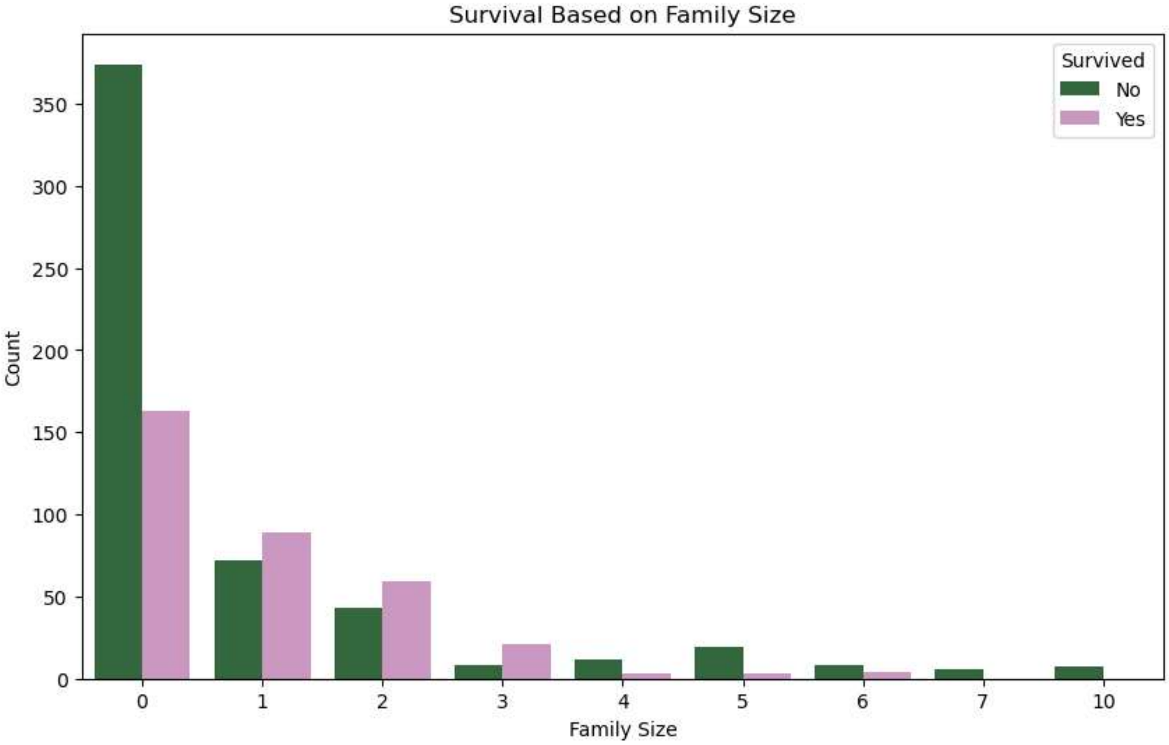
8.2. Fare Distribution by Survival

Shows how fare amounts affect survival rates.



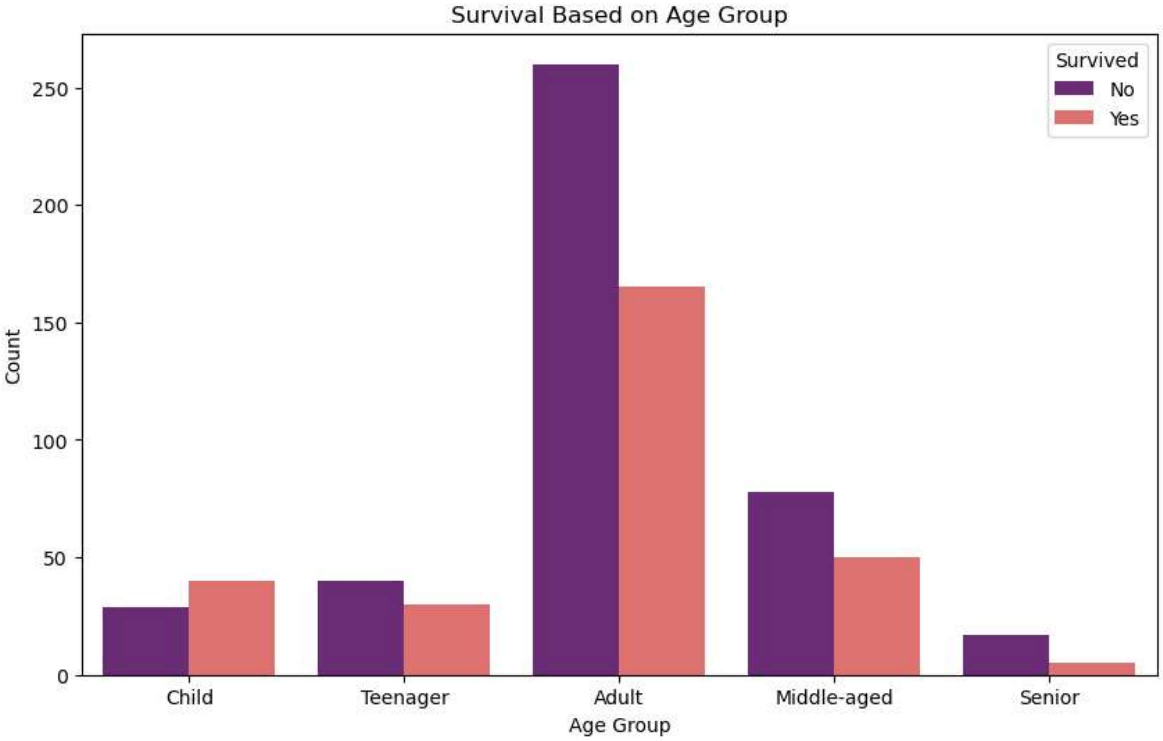
8.3. Family Size and Survival

Combines sibling/spouse and parent/child data to analyze family size effect.



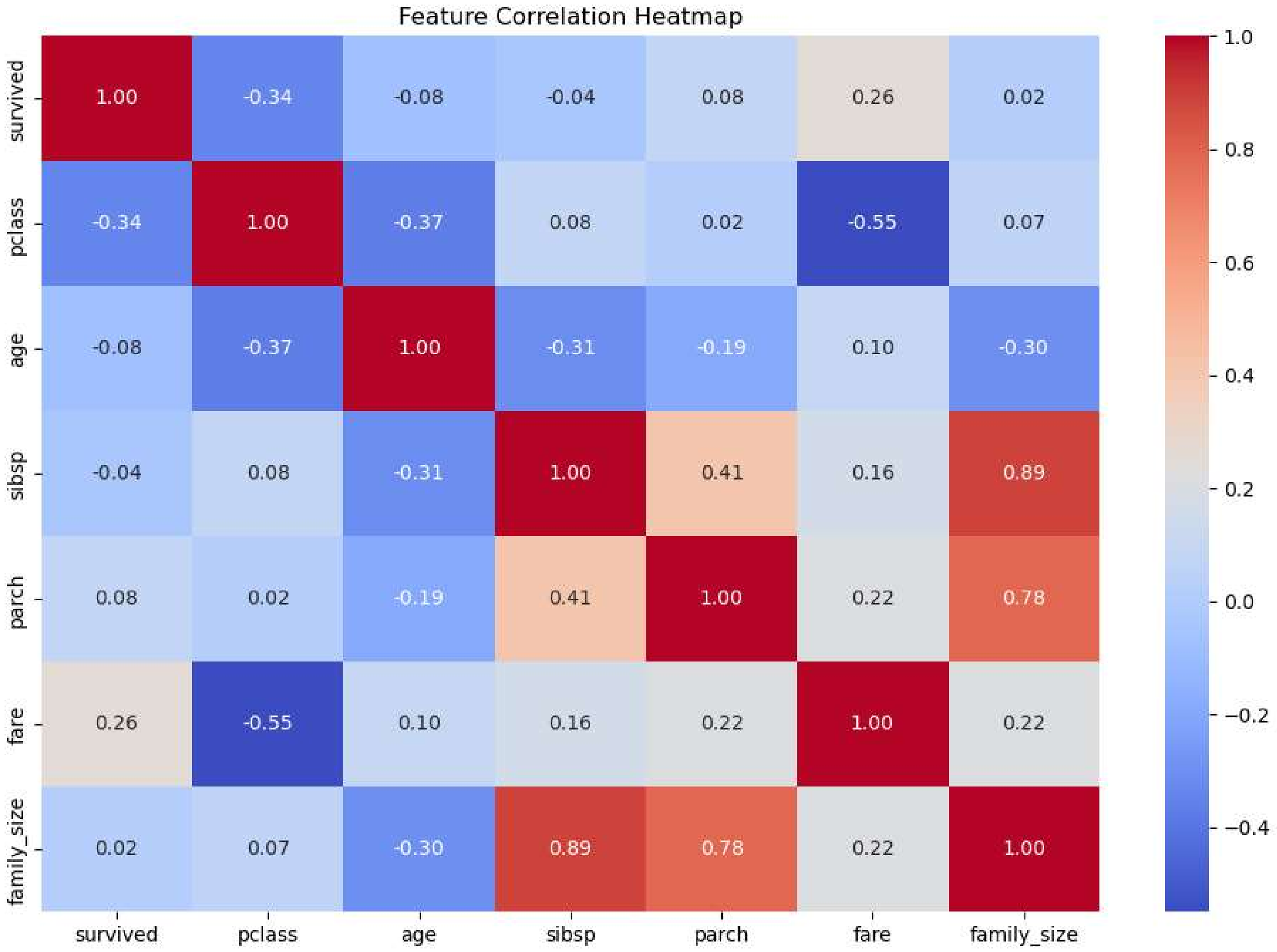
8.4. Survival by Detailed Age Groups

Uses more specific age bins for better granularity.



9. Correlation Heatmap

Visualizes correlation between numerical variables using a heatmap.



Feature Correlation Heatmap

	survived	pclass	age	sibsp	parch	fare	family_size
survived	1.00	-0.34	-0.08	-0.04	0.08	0.26	0.02
pclass	-0.34	1.00	-0.37	0.08	0.02	-0.55	0.07
age	-0.08	-0.37	1.00	-0.31	-0.19	0.10	-0.30
sibsp	-0.04	0.08	-0.31	1.00	0.41	0.16	0.89
parch	0.08	0.02	-0.19	0.41	1.00	0.22	0.78
fare	0.26	-0.55	0.10	0.16	0.22	1.00	0.22
family_size	0.02	0.07	-0.30	0.89	0.78	0.22	1.00

10. Key Findings

- **Gender:** Females had a higher survival rate than males.
- **Class:** First-class passengers were more likely to survive.
- **Age:** Children had better chances of survival.
- **Fare & Family Size:** Higher fare and smaller family size increased survival probability.
- **Embarkation Point:** Passengers from certain locations had varied survival outcomes.