CSE 351 Final Project: Titanic
Who Will Survive?

Joshua Jacob

Project Overview

The Titanic disaster of April 15, 1912 claimed over 1,500 lives out of the 2,224 onboard.

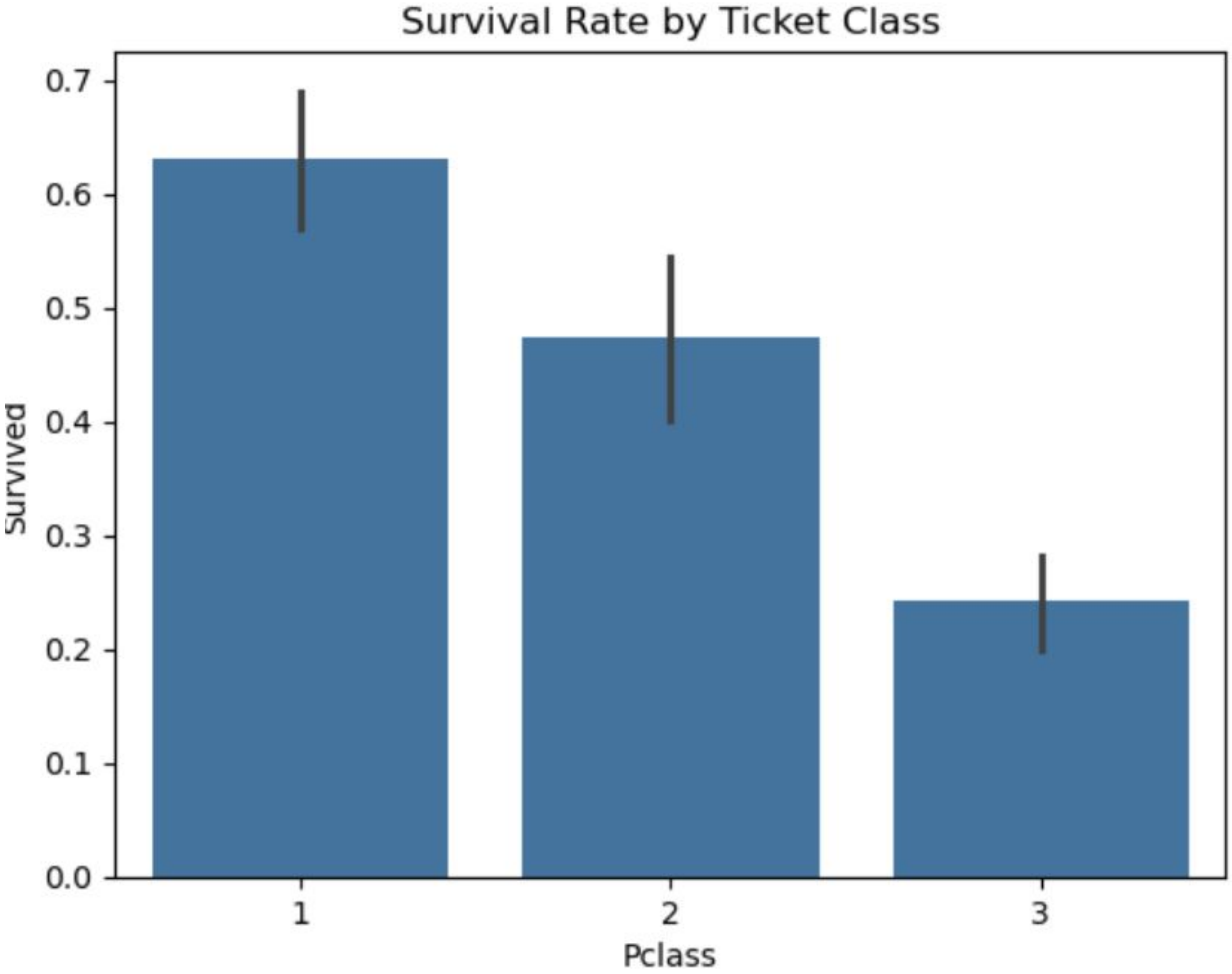
In this project, I will use passenger data to explore which factors most influenced survival whether if it was it luck, status, family size, or another variable.

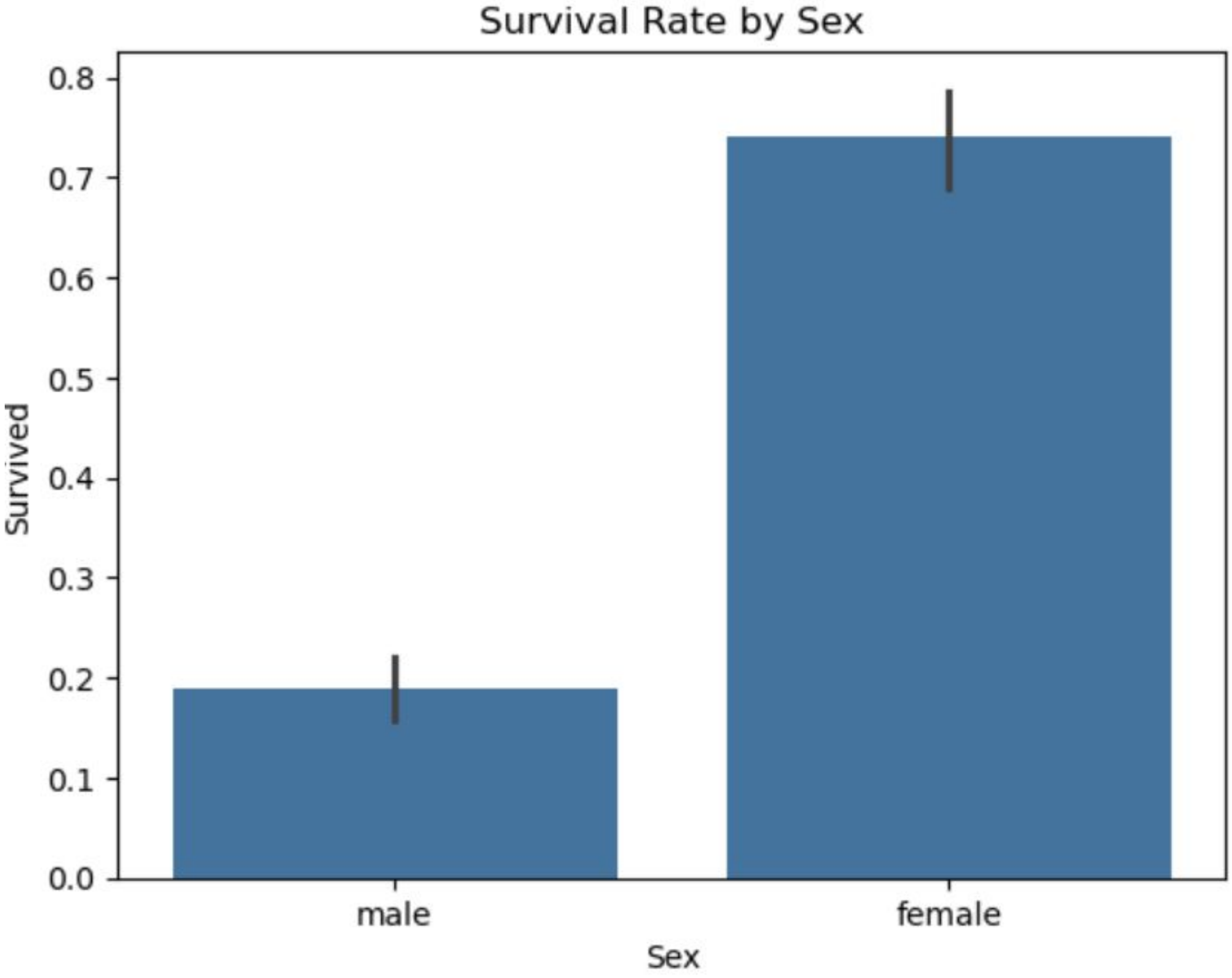
CLEANING THE DATA

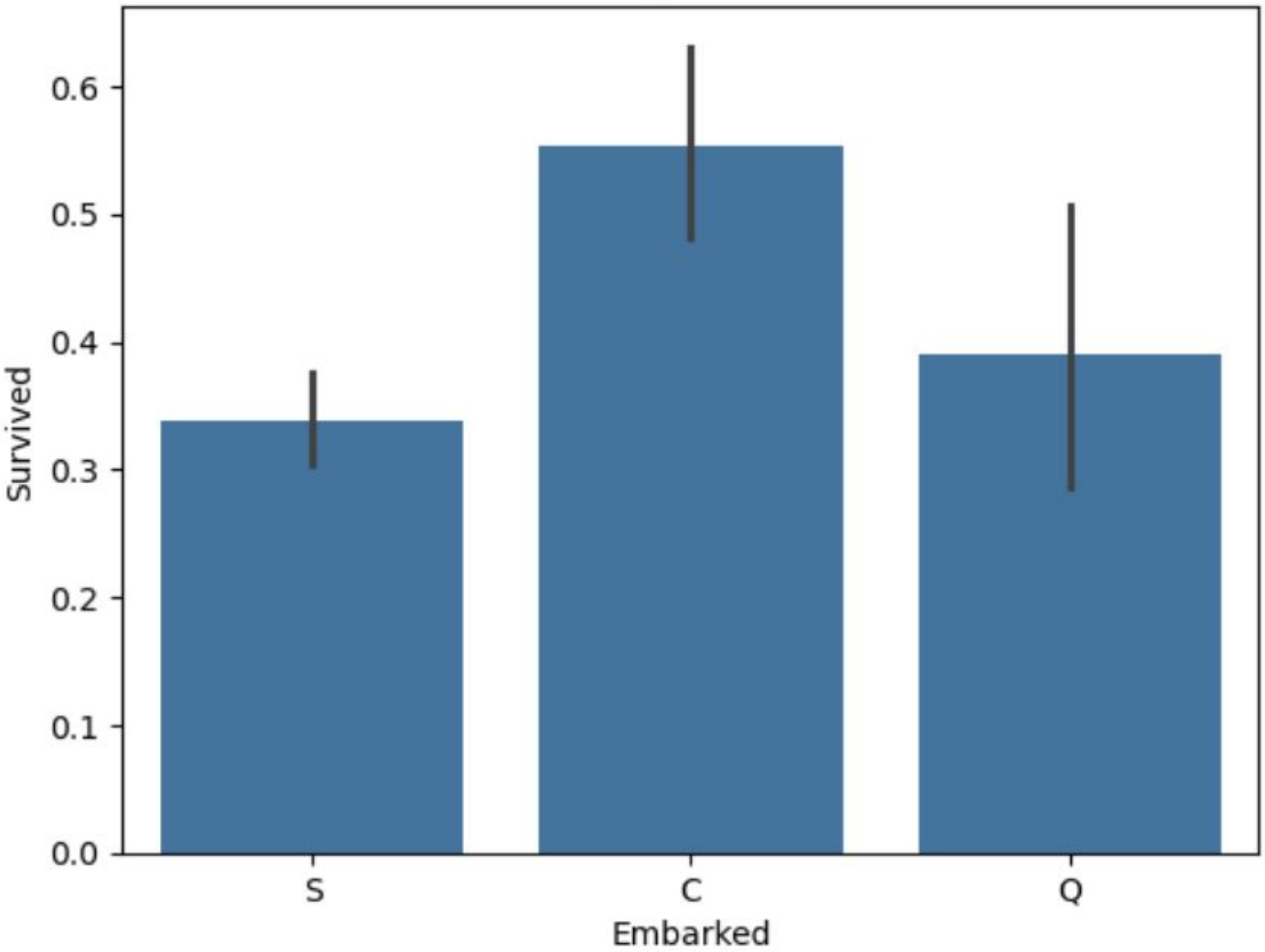
- I removed the Cabin column entirely because over three-quarters of its values were missing. For the Age field (about 20 % missing), I filled each empty entry with the overall mean age.
- The two missing Embarked entries were replaced with the most common embarkation port, “S.” In the test set, a single missing Fare was imputed by taking the average fare of passengers in the same class who boarded at the same port.

OUTLIER DETECTION

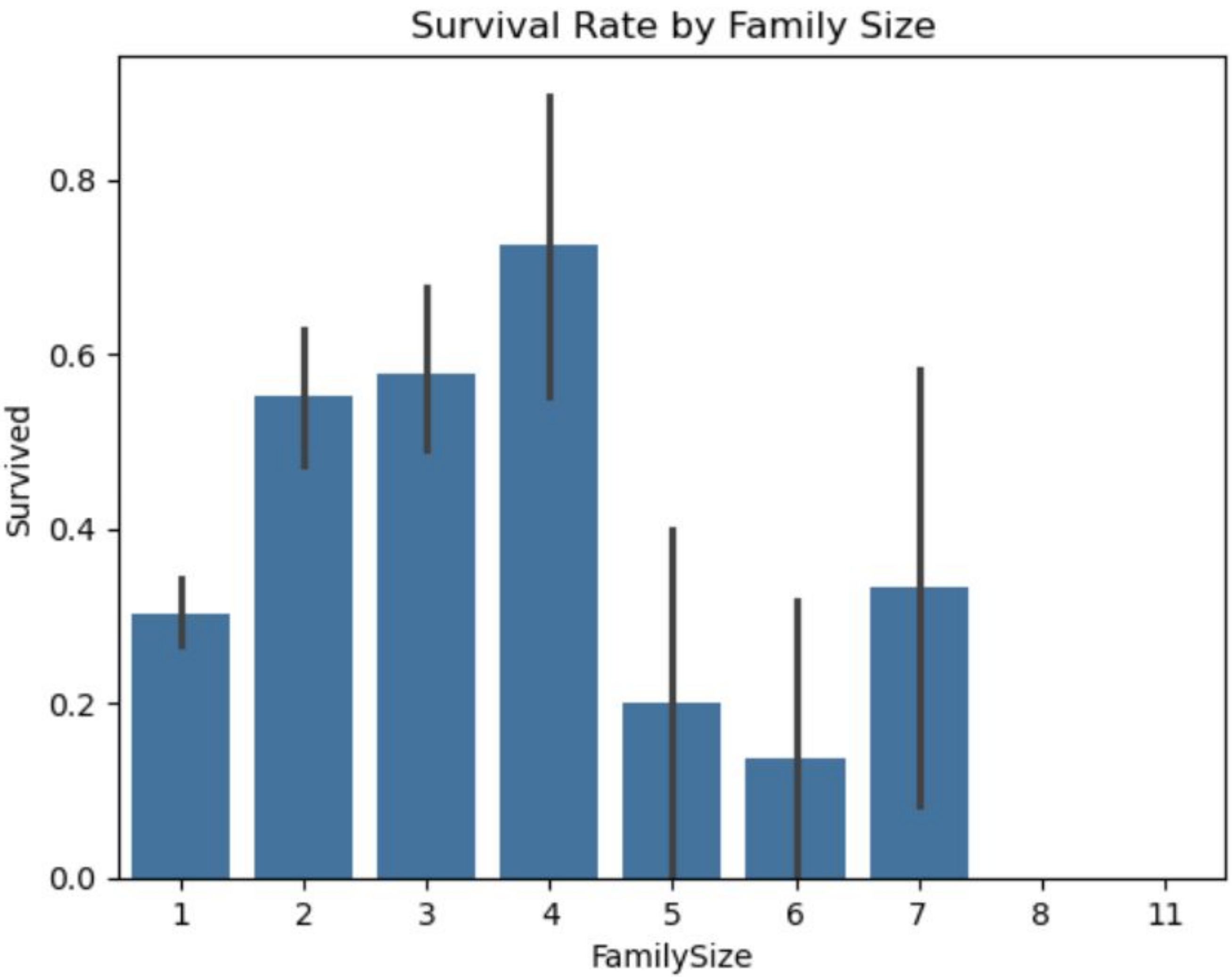
- Reviewed the ten highest and lowest values for Age, Fare, SibSp, and Parch
- Displayed key context (Pclass, Embarked, Survived) alongside each extreme
- Noted that high fares and zero fares aligned with expected passenger groups
- Observed that family counts stayed within reasonable limits

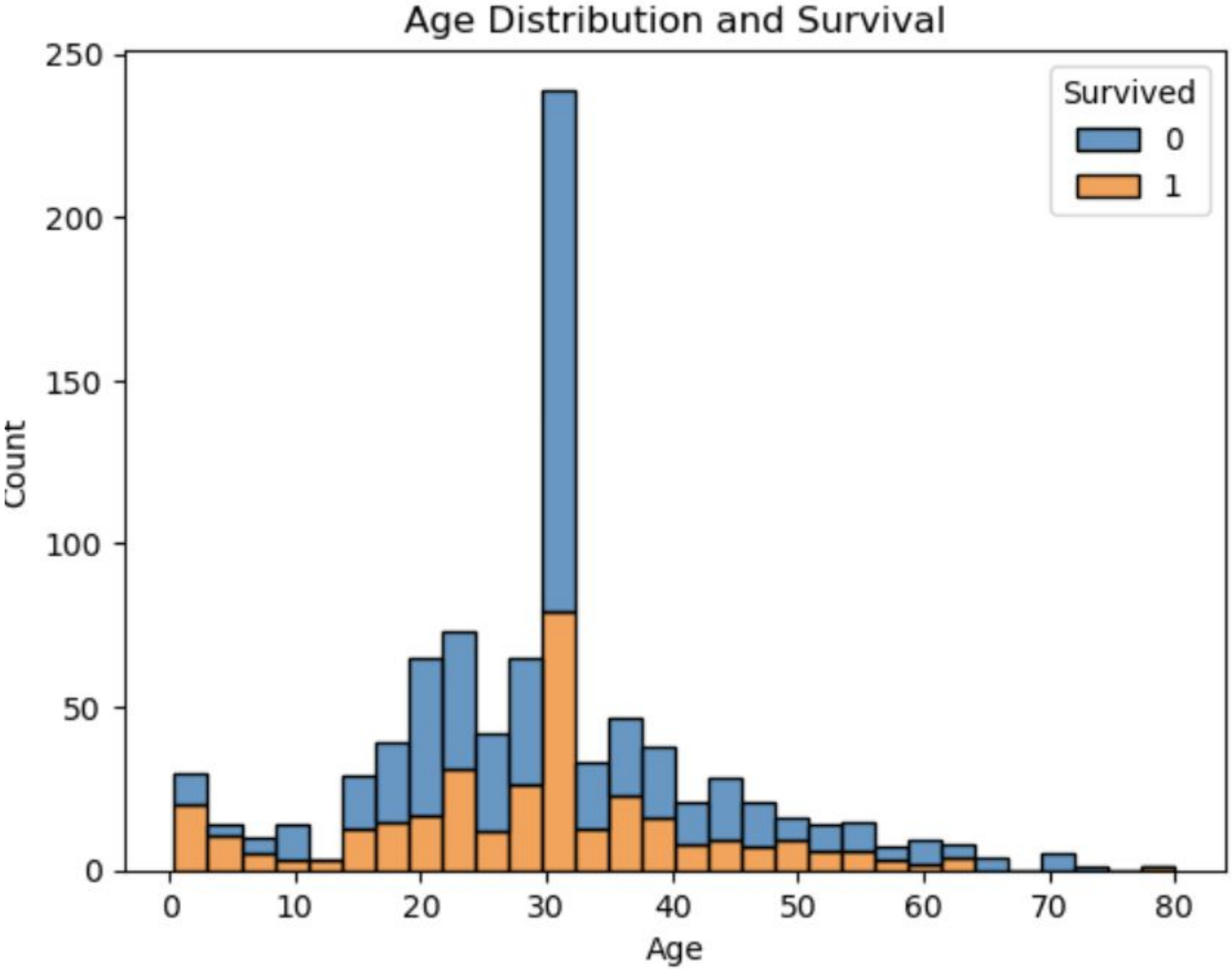

Survival Rate by Ticket Class
Survived
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1
2
3
Pclass

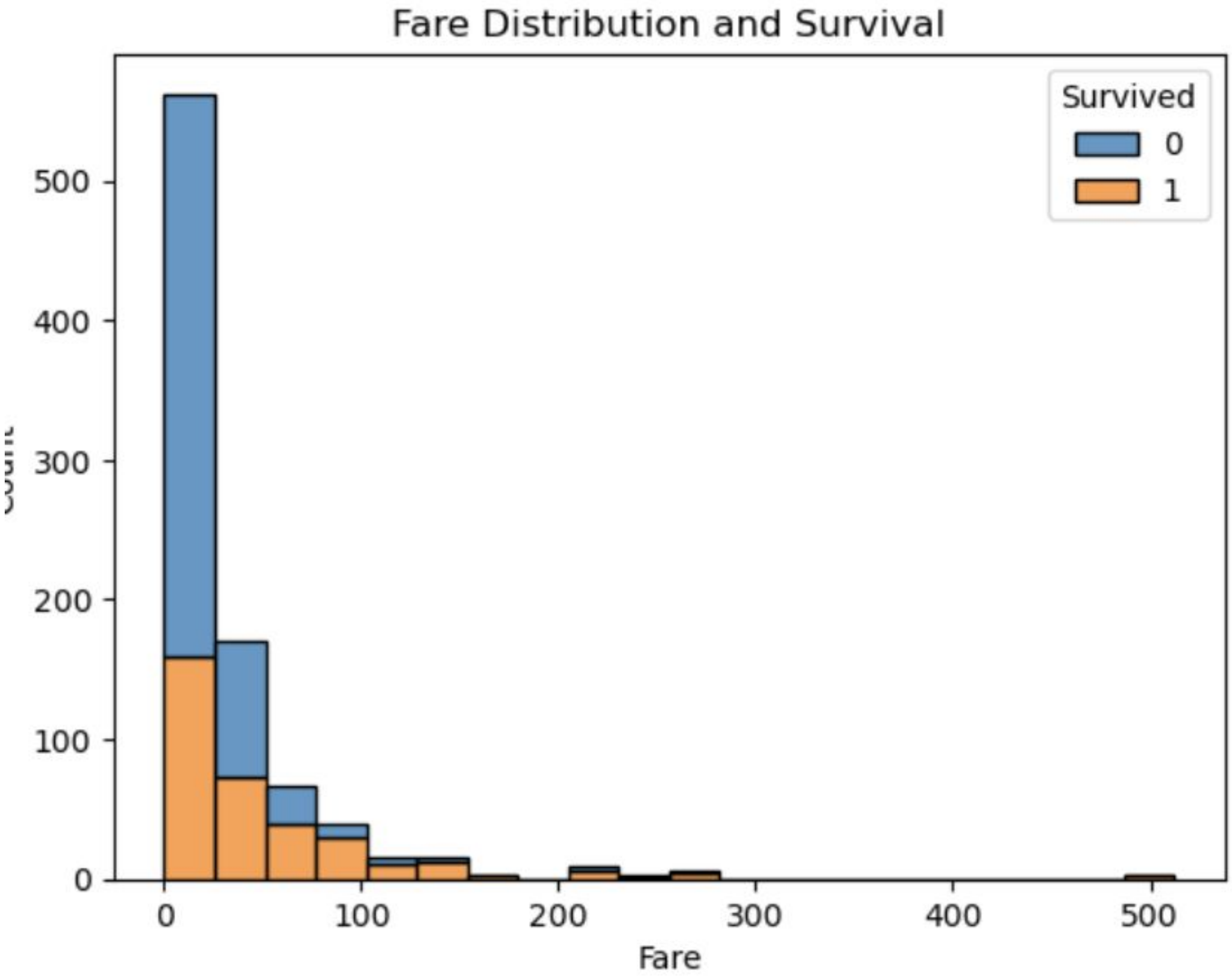

Survival Rate by Sex
Survived
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
male
female
Sex

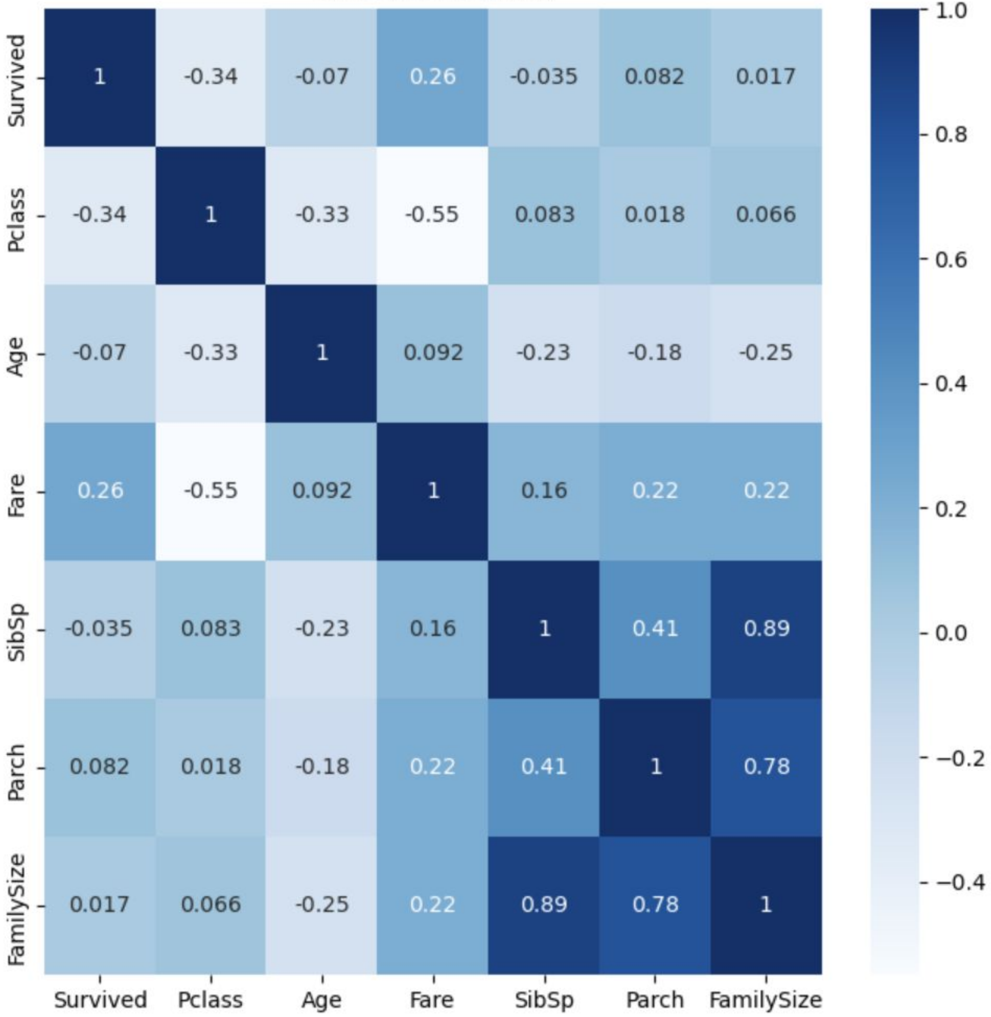


Survival Rate by Embarkation Port


Survival Rate by Family Size
Survived
0.0
0.2
0.4
0.6
0.8
FamilySize
1
2
3
4
5
6
7
8
11

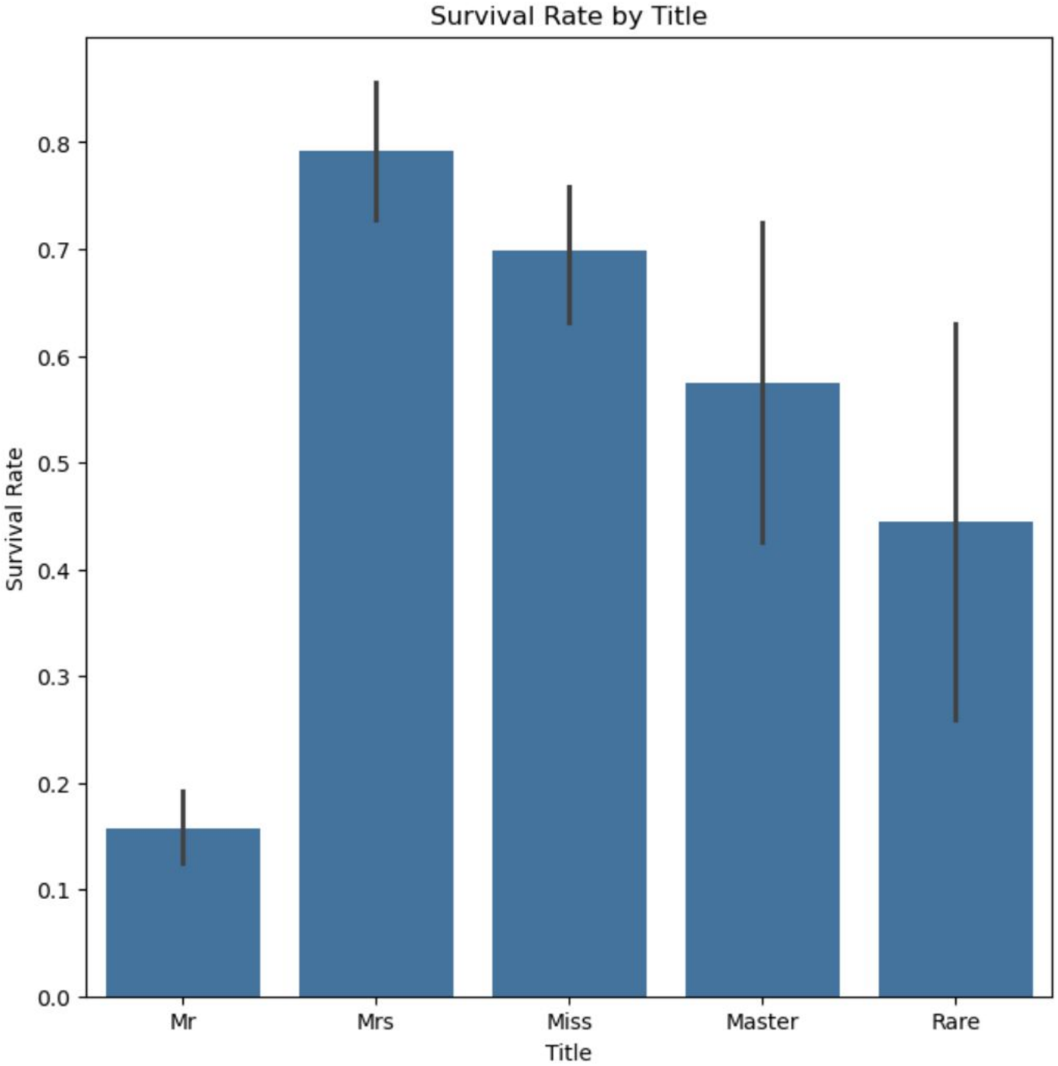

Age Distribution and Survival
Survived
🟦 0
🟧 1
Count
250
200
150
100
50
0
0
10
20
30
40
50
60
70
80
Age


Fare Distribution and Survival
Survived
0
1
500
400
300
200
100
0
0
100
200
300
400
500
Fare



Correlation Matrix

	Survived	Pclass	Age	Fare	SibSp	Parch	FamilySize
Survived	1	-0.34	-0.07	0.26	-0.035	0.082	0.017
Pclass	-0.34	1	-0.33	-0.55	0.083	0.018	0.066
Age	-0.07	-0.33	1	0.092	-0.23	-0.18	-0.25
Fare	0.26	-0.55	0.092	1	0.16	0.22	0.22
SibSp	-0.035	0.083	-0.23	0.16	1	0.41	0.89
Parch	0.082	0.018	-0.18	0.22	0.41	1	0.78
FamilySize	0.017	0.066	-0.25	0.22	0.89	0.78	1


Survival Rate by Title
Survival Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Mr
Mrs
Miss
Master
Rare
Title

I trained three classifiers on our cleaned Titanic data:

- Logistic Regression

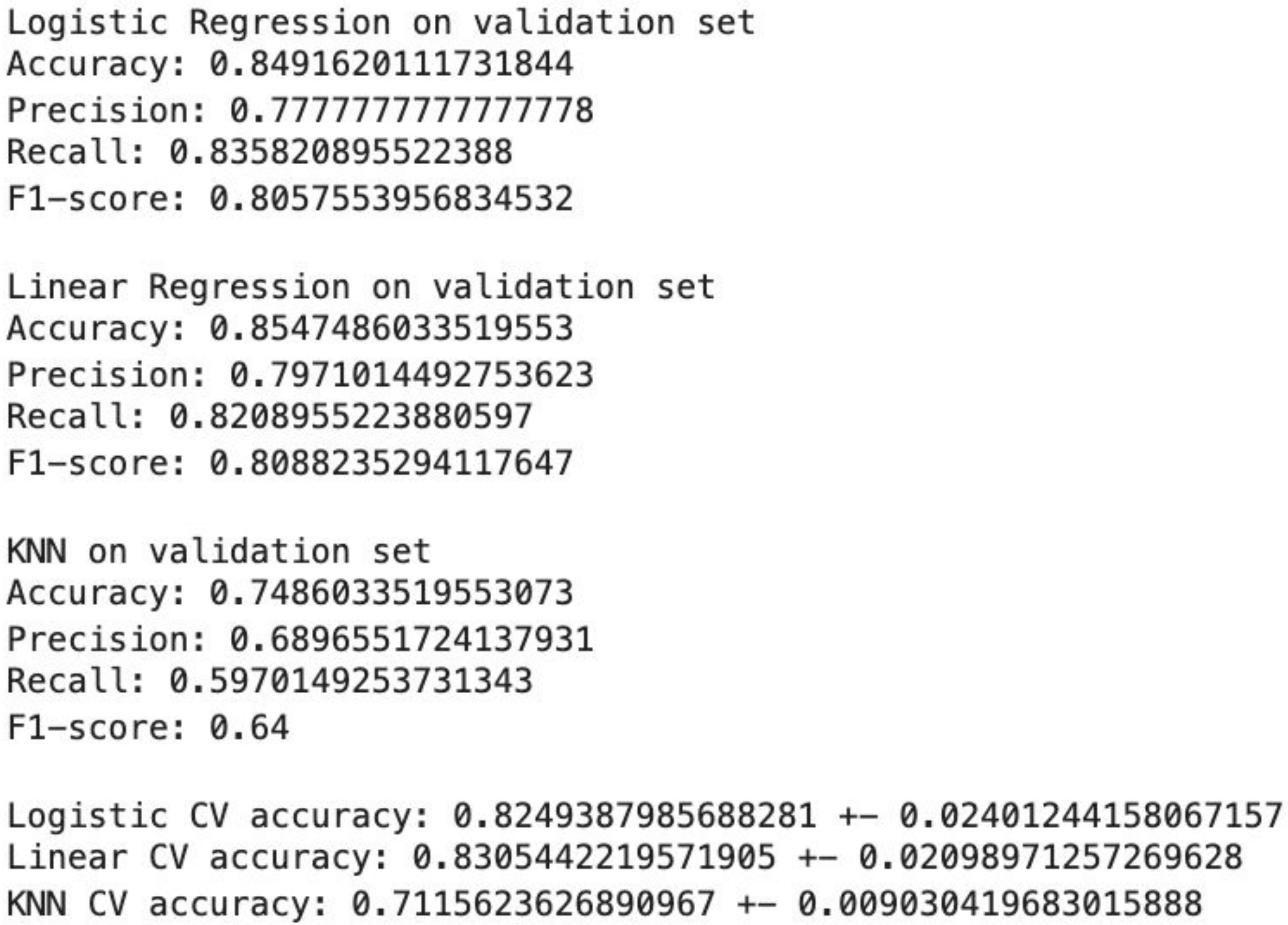
  – Estimates each passenger’s survival odds via a logistic (sigmoid) function on a weighted sum of features.

- Linear Regression (thresholded at 0.5)

  – Fits a continuous score to the 0/1 labels and then calls anything ≥ 0.5 a “survivor.”

- K-Nearest Neighbors (k=6)

  – Finds the six closest passengers in feature space and predicts by majority vote.



```
Logistic Regression on validation set
Accuracy: 0.8491620111731844
Precision: 0.7777777777777778
Recall: 0.835820895522388
F1-score: 0.8057553956834532

Linear Regression on validation set
Accuracy: 0.8547486033519553
Precision: 0.7971014492753623
Recall: 0.8208955223880597
F1-score: 0.8088235294117647

KNN on validation set
Accuracy: 0.7486033519553073
Precision: 0.6896551724137931
Recall: 0.5970149253731343
F1-score: 0.64

Logistic CV accuracy: 0.8249387985688281 +- 0.02401244158067157
Linear CV accuracy: 0.8305442219571905 +- 0.020989712572696 28
KNN CV accuracy: 0.7115623626890967 +- 0.009030419683015888
```

ERROR ANALYSIS

- Linear Regression

  – Linear Regression looked great on that single split but didn’t hold up when we tested more broadly, suggesting it was fitting to that specific subset.

- K-Nearest Neighbors

  – K-Nearest Neighbors jumped around because its “nearest neighbor” logic is very sensitive to how you measure distances—small changes in scaling or feature choice can swing its performance.

- Logistic Regression

  – Logistic Regression stayed the most consistent. Its drop from hold-out to cross-validation was alright, which means it’s capturing the main trends without over-tuning to one particular slice of data.

SUMMARY

Data Cleaning

– Dropped Cabin; imputed Age, Embarked, Fare

Key Insights from EDA

– Wealth (Pclass/Fare), gender, and age (children/elderly) drove survival rates

– No critical outliers or data errors found

Modeling & Performance

– Evaluated Logistic Regression, thresholded Linear Regression, K-NN (k=6)

– Logistic Regression most consistent; Linear best on a single split but less stable