**ANALYZING SURVIVAL FACTORS IN THE TITANIC DISASTER: A DATA-DRIVEN APPROACH**

**Table of Contents**

[1. Introduction](#_6bff6a9c5ozj) 2

[2. Data Cleaning and Feature Engineering](#_8phx33kquccr) 2

[3. Exploratory Data Analysis and Visualization](#_783v4duzxx85) 3

[3.1 Age Distribution of Passengers](#_p7060i2wan9g) 3

[3.2 Survival Rate by Passenger Class](#_buyc31xpp9ab) 3

[3.3 Age by Survival Status (Box Plot)](#_nm3pook0wobm) 4

[3.4 Passenger Class Distribution (Pie Chart)](#_qg6hxpvqi4y) 4

[3.5 Age vs. Fare by Survival Status (Scatter Plot)](#_sveu0w3eis60) 5

[3.6 Correlation Heatmap](#_2u1vumtluc70) 5

[4. Predictive Modeling](#_c4wsahtwj60z) 6

[5. Conclusion](#_ml7ie0q6x9s3) 6

[References](#_vyjz556n1l82) 7

# 

# 1. Introduction

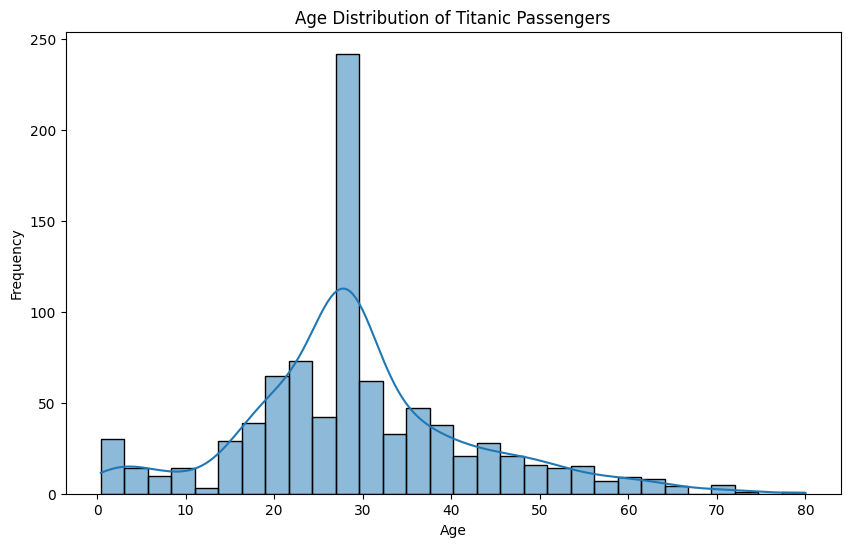
This study explores data on factors determining passengers' likelihood of survival during the Titanic disaster. It analyzes the general features and lifeless factors of the passengers and Titanic through statistical representation through measurements of age gender, fare, and division of classes. The results outline specific factors that may influence the choice of characteristics and the consequent quality of passenger experience.

# 2. Data Cleaning and Feature Engineering

Data cleaning has been applied to the dataset to handle missing values that affect the analysis. In the binary feature “Age”, “Embarked,” and numerical feature “Fare”, all the missing values are imputed using the medium and mode values respectively. Other useful columns including “Cabin” and “Ticket” in this case have been excluded (May *et al*., 2022). This led to the engineering of a new feature called “FamilySize” to also improve predictions.

# 3. Exploratory Data Analysis and Visualization

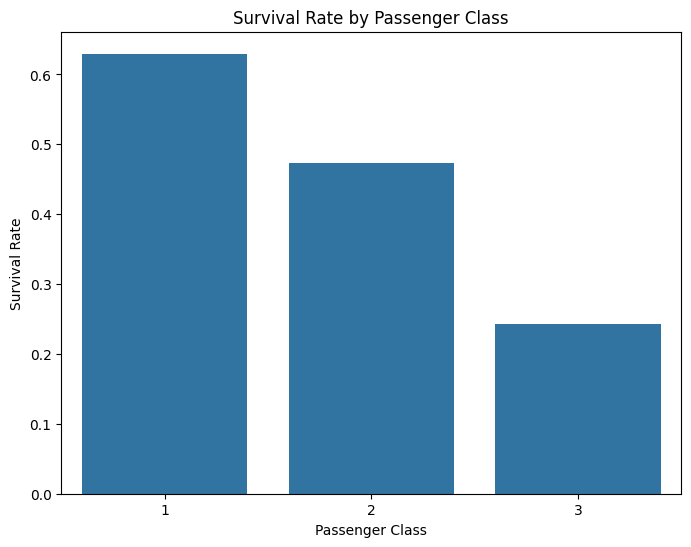
## 3.1 Age Distribution of Passengers



**Figure 1: Age Distribution of Titanic Passengers**

The figure presents the age pyramids of the Titanic’s passengers. It is positively skewed with the highest frequency in passengers in their early 30s and very few elderly passengers. This pattern indicates the dominance of young people among the passengers.

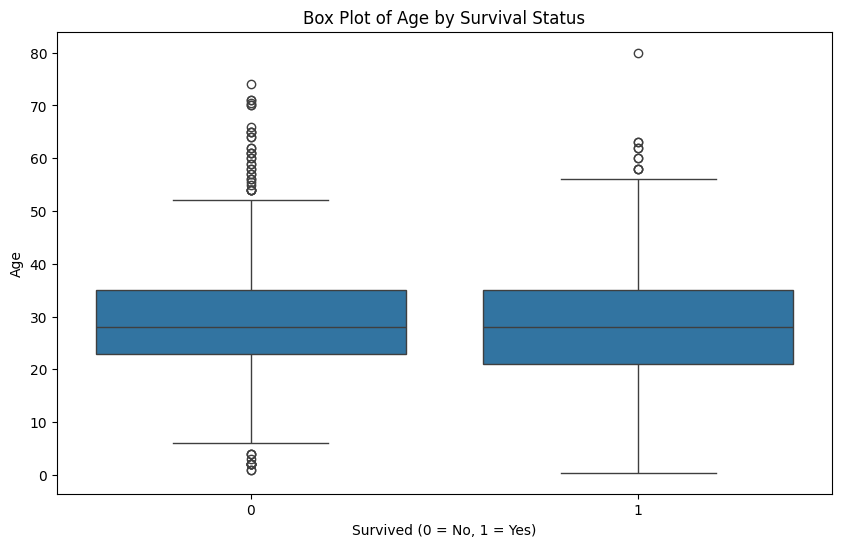
## 3.2 Survival Rate by Passenger Class



**Figure 2: Survival Rate by Passenger Class**

On the same plot, the survival rate is displayed depending on the passenger class. The first class has over 60% chance of passing to the next phase, the second class around 45% while the third class has about 25%.

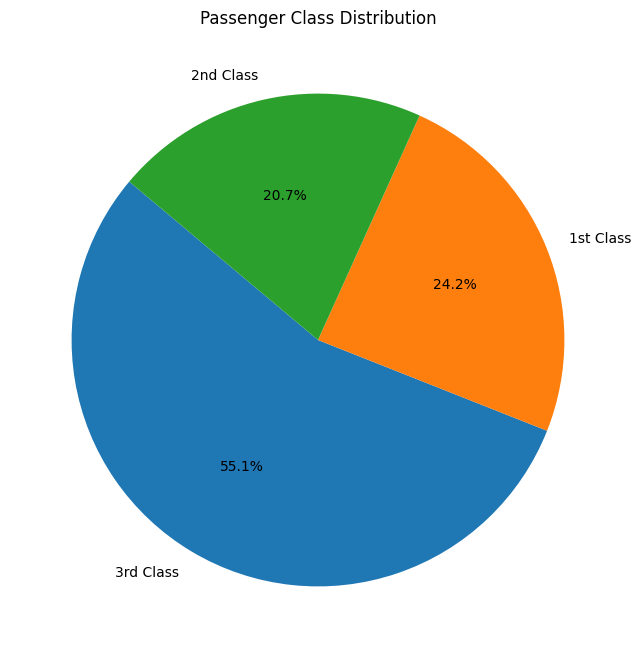
## 3.3 Age by Survival Status (Box Plot)



**Figure 3: Box Plot of Age by Survival Status**

The box plot reveals the non-survivors have been older and a higher number of patients had values above 60. The median ages of Survivors & Opponents are 28-30 years & Survivors had less no of elder persons up to 80 age only is an outlier.

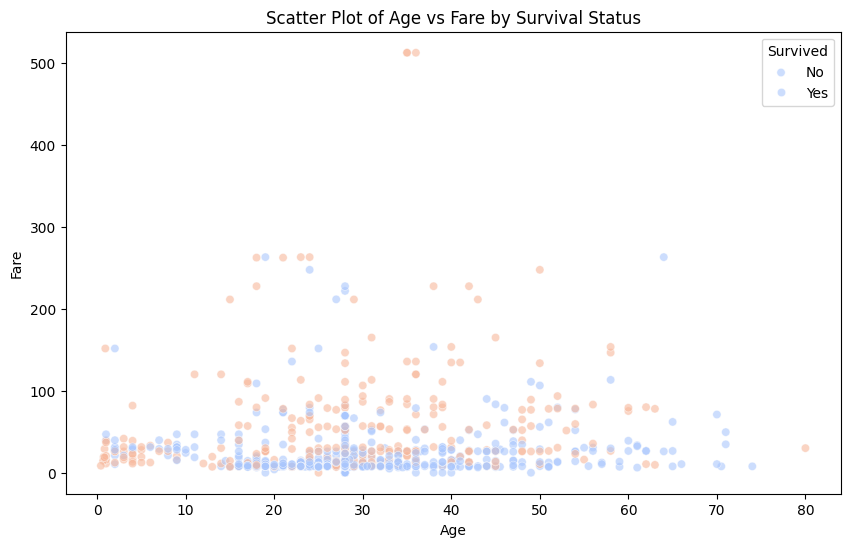
## 3.4 Passenger Class Distribution (Pie Chart)



**Figure 4: Passenger Class Distribution**

The given plot illustrates passengers’ distribution by class. The third class stands at 55.1% while the first class is at 24.2% and the second class is at 40.7%.

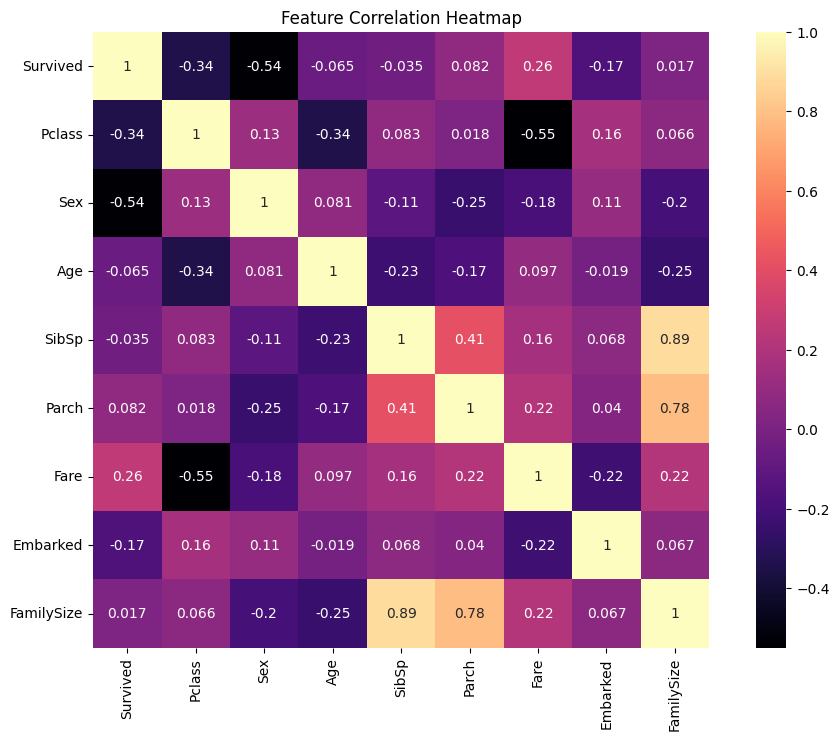
## 3.5 Age vs. Fare by Survival Status (Scatter Plot)



**Figure 5: Scatter Plot of Age vs Fare by Survival Status**

The dispersion of the survivors and non-survivors can be practically equal in fares less than 100 indicating that no special fare-to-age relationship can be determined by the survival rate.

## 3.6 Correlation Heatmap

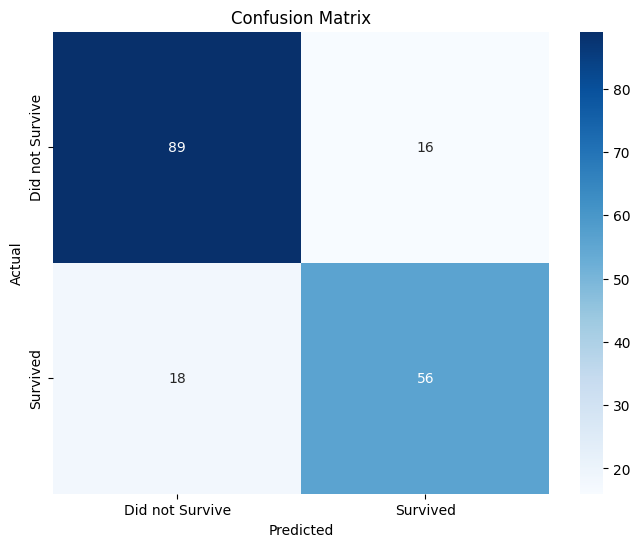


**Figure 6: Feature Correlation Heatmap**

This plot is a feature correlation heatmap. The highest positive correlation coefficient is between SibSp and FamilySize equal to 0.89. The lowest positive correlation is between Pclass and Fare with a coefficient of -0.55.

# 4. Predictive Modeling

To predict Titanic passenger survival, analysis has been done using a Random Forest Classifier and it has been trained in an 80:20 manner (McKinney, 2022). In the Random Forest Classifier model, an accuracy of 81% has been obtained which is relatively good for predicting the performance of the model on unseen data. The confusion matrix showed equal responsiveness and precise outputs relevant to both groups of patients as well as non-survivors, highlighting the model’s applicability.



**Figure 7: Confusion Matrix**

The plot described here is a confusion matrix plot, a visual way to depict how well a classification model is working. It has true negatives of 89 false positives of 16 false negatives of 18 true positives of 56 when it comes to model accuracy.

# 5. Conclusion

The study gave information on the survival indicators, of which passenger class, gender, and fare posed highly significant effects on survival. The Random Forest model generated high accuracy, thus confirming these patterns (José, 2021). This study demonstrates that insight into the important variables that inform decision-making and shape the behavior of systems can be gained through secondary analysis of data and could be applied in understanding real-life, complicated, and, often, tragic events, such as the Titanic sinking.

# References

José, U., 2021. Python programming for data analysis. Springer Nature.

May, R.M., Goebbert, K.H., Thielen, J.E., Leeman, J.R., Camron, M.D., Bruick, Z., Bruning, E.C., Manser, R.P., Arms, S.C. and Marsh, P.T., 2022. MetPy: A meteorological Python library for data analysis and visualization. Bulletin of the American Meteorological Society, 103(10), pp.E2273-E2284.

McKinney, W., 2022. Python for data analysis. " O'Reilly Media, Inc.".