**Report - Logistic Regression Model**

**Data description:**

* **This report analyses the Titanic dataset to build a model predicting passenger survival.**
  + **Response Variable (Y): Survived (a binary variable where 1 = Survived, 0 = Did Not Survive).**
  + **Predictor Variables (X): Pclass (Passenger Class: 1, 2, or 3), Sex (male or female), Age (age in years), and SibSp (number of siblings/spouses aboard).**

**Shapiro-Wilk Test**

* **Variables not normally distributed: Age, Fare (both p-value < 0.001)**

| **Variable** | **Median** | **IQR (Q1, Q3)** |
| --- | --- | --- |
| **Age** | **28** | **(20.1, 38.0)** |
| **Fare** | **14.45** | **(7.91, 31.00)** |

* **Descriptive Statistics (Categorical Variables)**

|  |  |
| --- | --- |
| **Variable** | **n (%)** |
| **Survived = 0** | **549 (61.6%)** |
| **Survived = 1** | **342 (38.4%)** |
| **Sex = Male** | **577 (64.8%)** |
| **Sex = Female** | **314 (35.2%)** |
| **Pclass = 1** | **216 (24.2%)** |
| **Pclass = 2** | **184 (20.7%)** |
| **Pclass = 3** | **491 (55.1%)** |

**Data Preprocessing and Imputation:**

**The initial dataset contained 891 observations. Significant preprocessing was required before modelling.**

* **Missing Data: The Cabin column was removed due to excessive missingness (687 missing values). The Age and Embarked columns had 177 and 2 missing values, respectively .**
* **Missing Data Analysis: A test for data Missing Completely at Random (MCAR) was rejected (p < 0.05), indicating that the missing data was not random. Further Chi-square tests revealed that missingness in the Age variable was significantly associated with other variables like Pclass, SibSp, and Embarked. This suggests the data is**

**Missing at Random (MAR), which justifies using imputation.**

* **Imputation Strategy: A simple imputation method was used instead of regression imputation since the variables were skewed. Missing Age values were filled with the median age of 28, and the two missing Embarked values were filled with the mode.**
* **Influential Points: After fitting an initial model, Cook's distance was used to identify highly influential data points that could disproportionately affect the model's coefficients. A total of 67 influential points were identified .**

**A graph of a number of objects

AI-generated content may be incorrect.**

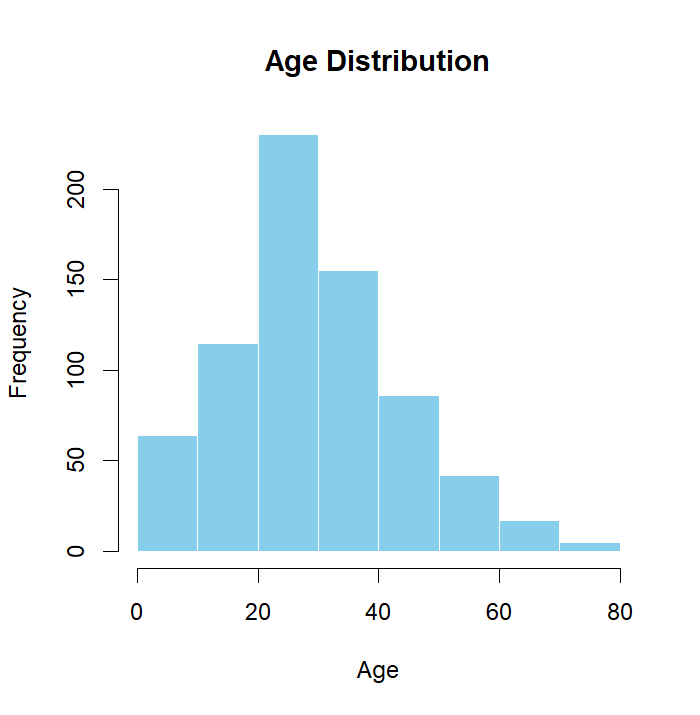
* **The data had no duplicated row. Additionally, the dataset was checked for any illogical values and none were found. Thus, the dataset was deemed suitable for immediate analysis.**

**Univariate Analysis with Graphs**

**Histogram:**

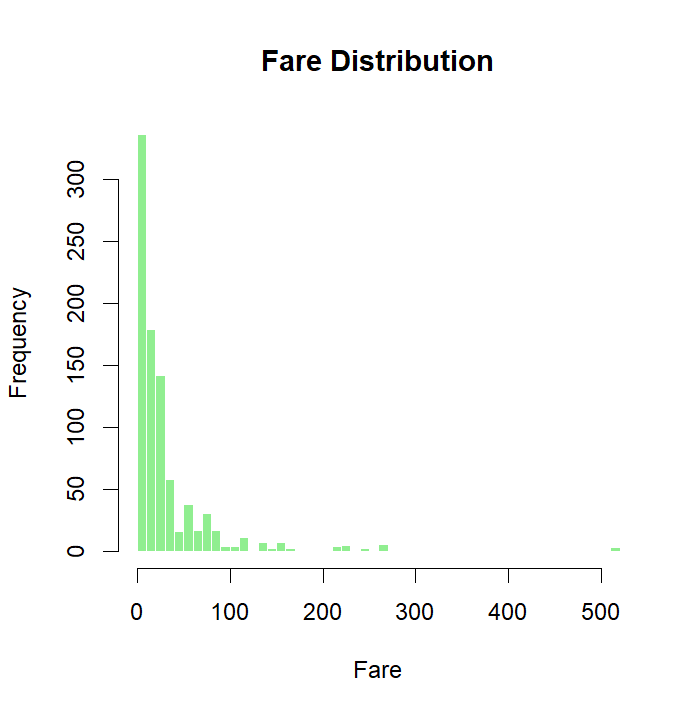
**Age**

**Displays a unimodal but right-skewed distribution, with most passengers between 20–40 years.**

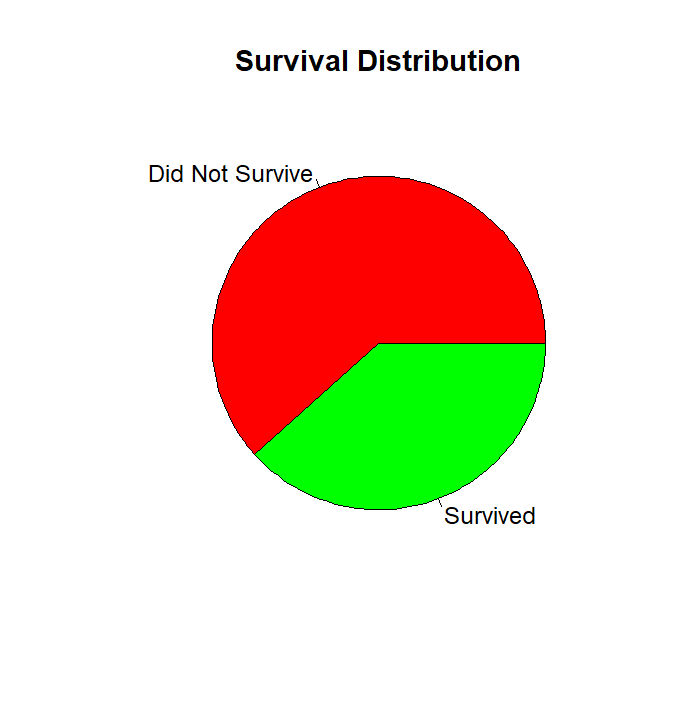
****

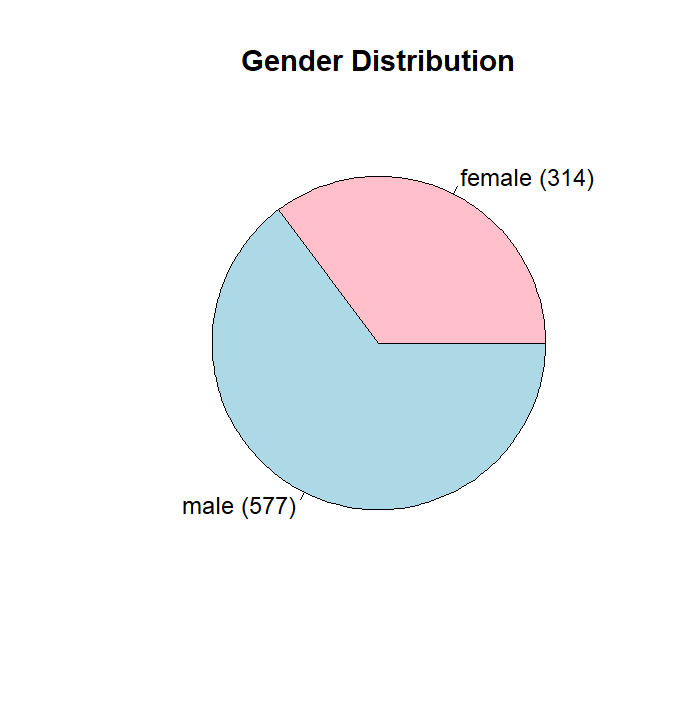
**Fare**

**Most fares are clustered below 50, with a few extreme outliers (up to 512).**

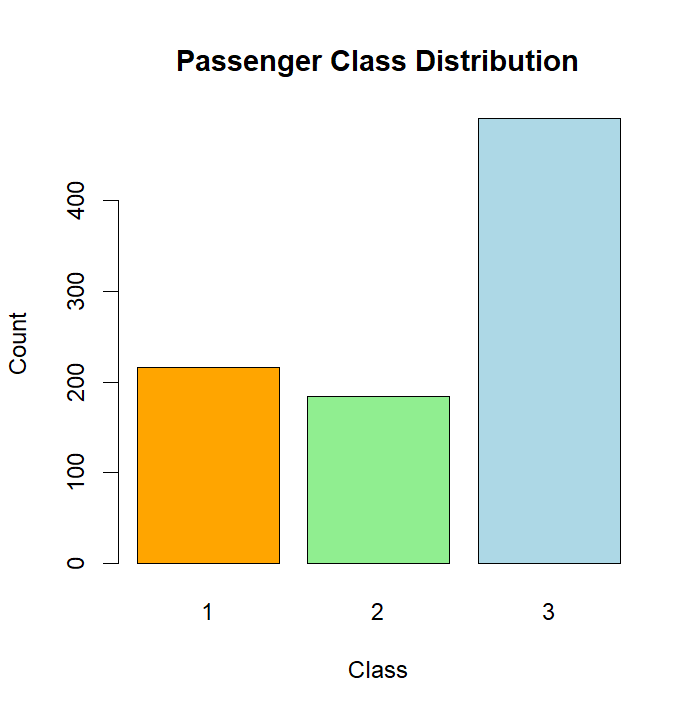
****

**Pie Charts:**

****

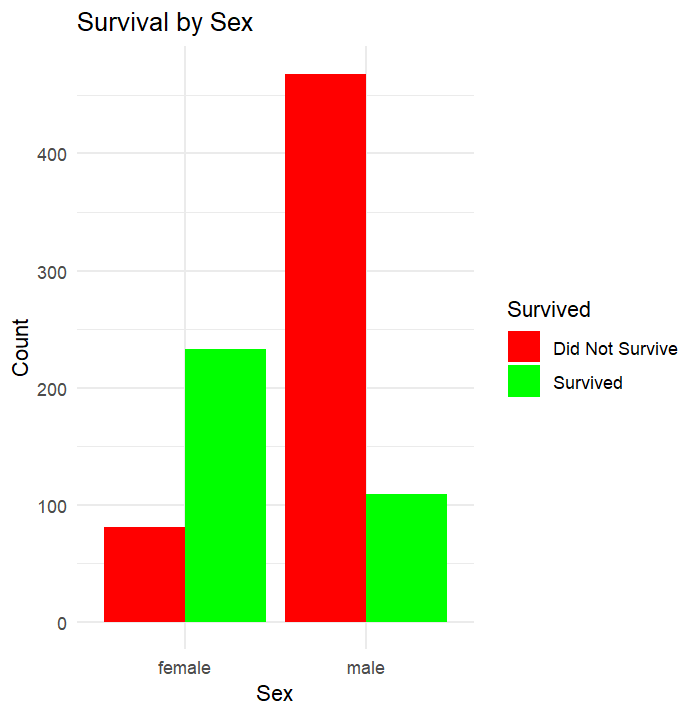
****

**Bar plots:**

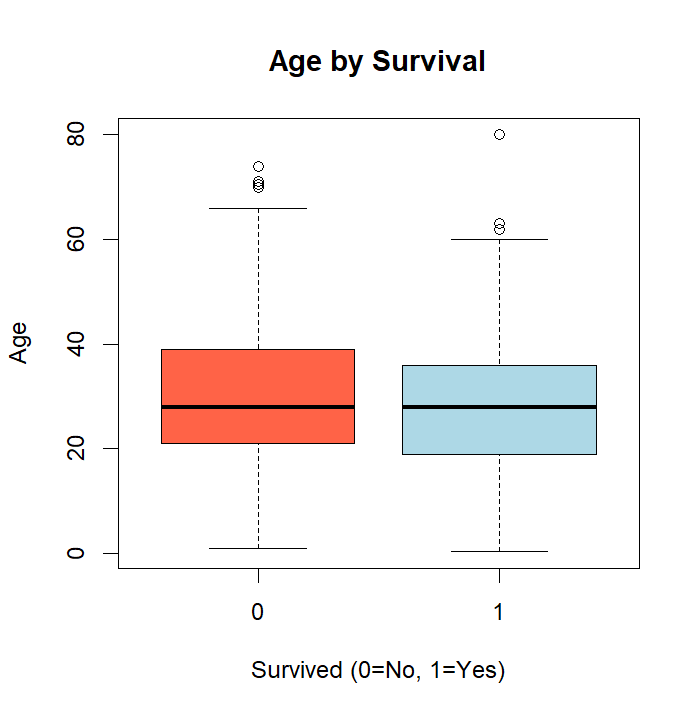
****

**A graph of a bar graph

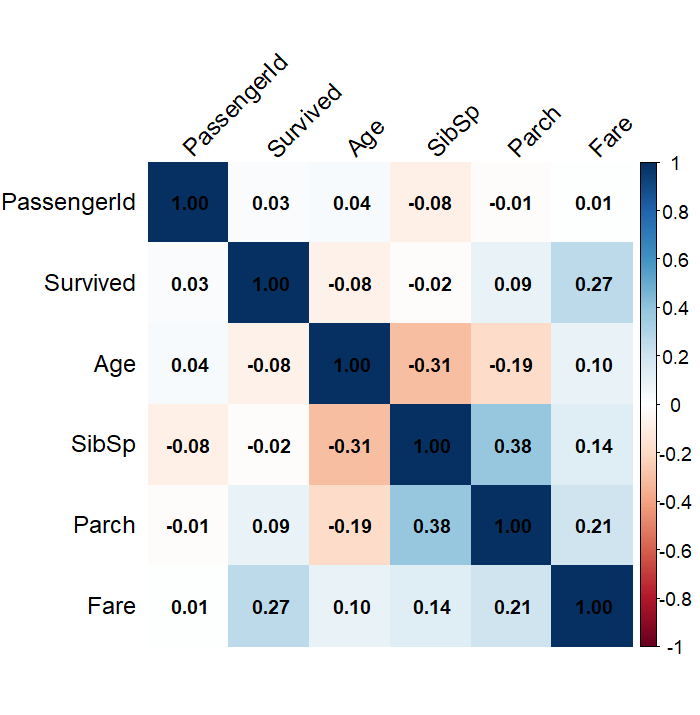
AI-generated content may be incorrect.**

****

**Box plot**

****

**Correlation plot**

****

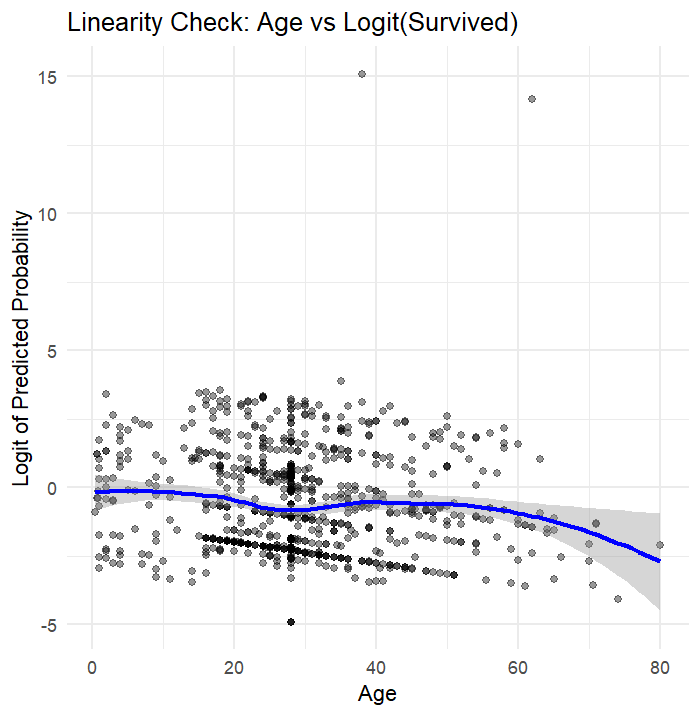
**Model form:**

**We wish to fit a logistic regression model of the form:**

**log(P/1-P)) = β0​+β1​x1​+β2​x2​+...+ϵ**

**where Y is the binary outcome Survived, and the x variables are the predictors. The model predicts the log-odds of survival.**

**Linearity:**

****

* **If the line is roughly straight (linear), this suggests that the assumption of linearity in the logit (required for logistic regression for continuous predictors like Age) is reasonably satisfied.**
* **If the line is curved (concave/convex), it indicates non-linearity, meaning Age might not have a strictly linear relationship with the logit of survival.**

**Deviance residual plot:**

**A graph with black dots and red line

AI-generated content may be incorrect.**

**Pearson residual plot:**

**A graph with black dots and numbers

AI-generated content may be incorrect.**

**Model With All Data (Before Removing Influential Points):**

**Model Results (Final Model After Backward Selection)**

| **Term** | **Estimate** | **Std. Error** | **z value** | **p-value** | **Odds Ratio (95% CI)** |
| --- | --- | --- | --- | --- | --- |
| **(Intercept)** | **4.036** | **0.463** | **8.716** | **< 0.001 \*\*\*** | **56.62 (23.51, 144.77)** |
| **Pclass2** | **-1.150** | **0.307** | **-3.747** | **< 0.001 \*\*\*** | **0.317 (0.172, 0.574)** |
| **Pclass3** | **-2.359** | **0.282** | **-8.357** | **< 0.001 \*\*\*** | **0.095 (0.054, 0.162)** |
| **Sexmale** | **-2.864** | **0.229** | **-12.511** | **< 0.001 \*\*\*** | **0.057 (0.036, 0.088)** |
| **Age** | **-0.0418** | **0.0093** | **-4.514** | **< 0.001 \*\*\*** | **0.959 (0.941, 0.976)** |
| **SibSp** | **-0.313** | **0.116** | **-2.693** | **0.007 \*\*** | **0.731 (0.573, 0.905)** |

**Interpretation of Coefficients**

***Model With Influential Points:*  
Being male reduces the odds of survival by approximately 94% compared to females, holding other factors constant. Passengers in 2nd class have about 68% lower odds of survival, and those in 3rd class have about 90% lower odds compared to 1st-class passengers. Each additional year of age decreases the odds of survival by roughly 4%, and each additional sibling or spouse aboard reduces the odds by about 27%.**

**Model Fit Statistics:**

* **Null deviance: 876.75 (657 df)**
* **Residual deviance: 567.26 (652 df)**
* **AIC: 579.26**
* **McFadden’s Pseudo R²: 0.353 (35.3% of variation explained)**

**Goodness-of-Fit & Diagnostics:**

* **ROC AUC: 0.8538 → Excellent discrimination**
* **Hosmer–Lemeshow Test: χ² = 26.56, df = 8, p = 0.00084 → Some lack of calibration**
* **VIF: All predictors < 2 → No multicollinearity issues**

**Classification Performance (Test Data)/Confusion matrix**

|  | **Actual = 0** | **Actual = 1** |
| --- | --- | --- |
| **Predicted = 0** | **89** | **24** |
| **Predicted = 1** | **13** | **40** |

* **Accuracy: 78.4%**
* **Sensitivity: 62.5%**
* **Specificity: 87.3%**

**Model After Removing Influential Points (Cook’s Distance > 4/n)**

**A total of 67 influential points were removed to improve model stability.**

**Model Results (Refit After Backward Selection)**

| **Term** | **Estimate** | **Std. Error** | **z value** | **p-value** | **Odds Ratio (95% CI)** |
| --- | --- | --- | --- | --- | --- |
| **(Intercept)** | **3.989** | **0.411** | **9.710** | **< 0.001 \*\*\*** | **54.13 (24.63, 121.09)** |
| **Pclass2** | **-1.199** | **0.272** | **-4.404** | **< 0.001 \*\*\*** | **0.301 (0.182, 0.498)** |
| **Pclass3** | **-2.324** | **0.252** | **-9.210** | **< 0.001 \*\*\*** | **0.098 (0.059, 0.160)** |
| **Sexmale** | **-2.764** | **0.201** | **-13.763** | **< 0.001 \*\*\*** | **0.063 (0.043, 0.091)** |
| **Age** | **-0.0398** | **0.0081** | **-4.891** | **< 0.001 \*\*\*** | **0.961 (0.945, 0.977)** |
| **SibSp** | **-0.321** | **0.105** | **-3.064** | **0.002 \*\*** | **0.726 (0.581, 0.891)** |
|  |  |  |  |  |  |

**Interpretation of Coefficients:**

**The interpretations remain very similar. Being male still reduces the odds of survival by around 94%. Passengers in 2nd class have roughly 70% lower odds, and 3rd-class passengers about 90% lower odds than those in 1st class. Each year increase in age decreases survival odds by ~4%, and each additional sibling or spouse aboard reduces the odds by ~27%. Removing influential points slightly improves model stability but does not change the direction or practical meaning of the effects.**

**Model Fit Statistics:**

* **Null deviance: 1098.10 (823 df)**
* **Residual deviance: 730.75 (818 df)**
* **AIC: 742.75**
* **McFadden’s Pseudo R²: 0.375 (Improved from 0.353)**

**Goodness-of-Fit & Diagnostics:**

* **ROC AUC: 0.8538 (no loss in discrimination)**
* **Hosmer–Lemeshow Test: p < 0.01 → slight lack of fit persists, but improved calibration visually**
* **VIF: All predictors < 2 → No multicollinearity**

**Classification Performance (Clean Test Data) /Confusion matrix:**

|  | **Actual = 0** | **Actual = 1** |
| --- | --- | --- |
| **Predicted = 0** | **95** | **16** |
| **Predicted = 1** | **15** | **53** |

* **Accuracy: 82.6% (↑ from 78.4%)**
* **Sensitivity: 76.8% (↑)**
* **Specificity: 86.4% (slightly ↓)**

**Conclusion:**

**Both models (before and after removing influential points) identified the same key predictors of survival: Pclass, Sex, Age, and SibSp. The direction and magnitude of effects remained consistent, with Sex (male) and Pclass3 showing the strongest negative impact on survival probability.**

**Although the Hosmer–Lemeshow test (p < 0.01) indicates slight lack of fit, the model shows excellent discrimination (ROC AUC = 0.854), stable coefficients, and no multicollinearity, making it robust and suitable for predicting survival.**

**After removing influential points:**

* **Model fit improved (lower AIC, higher pseudo R²).**
* **Predictive performance improved, with accuracy increasing from 78.4% to 82.6% and sensitivity (true positive rate) increasing from 62.5% to 76.8%.**
* **Odds ratios remained similar, confirming the robustness of the model.**

**Overall, the final model (after removing influential points) provides a more stable and better-fitting model for predicting Titanic survival while retaining the same key insights about risk factors.**