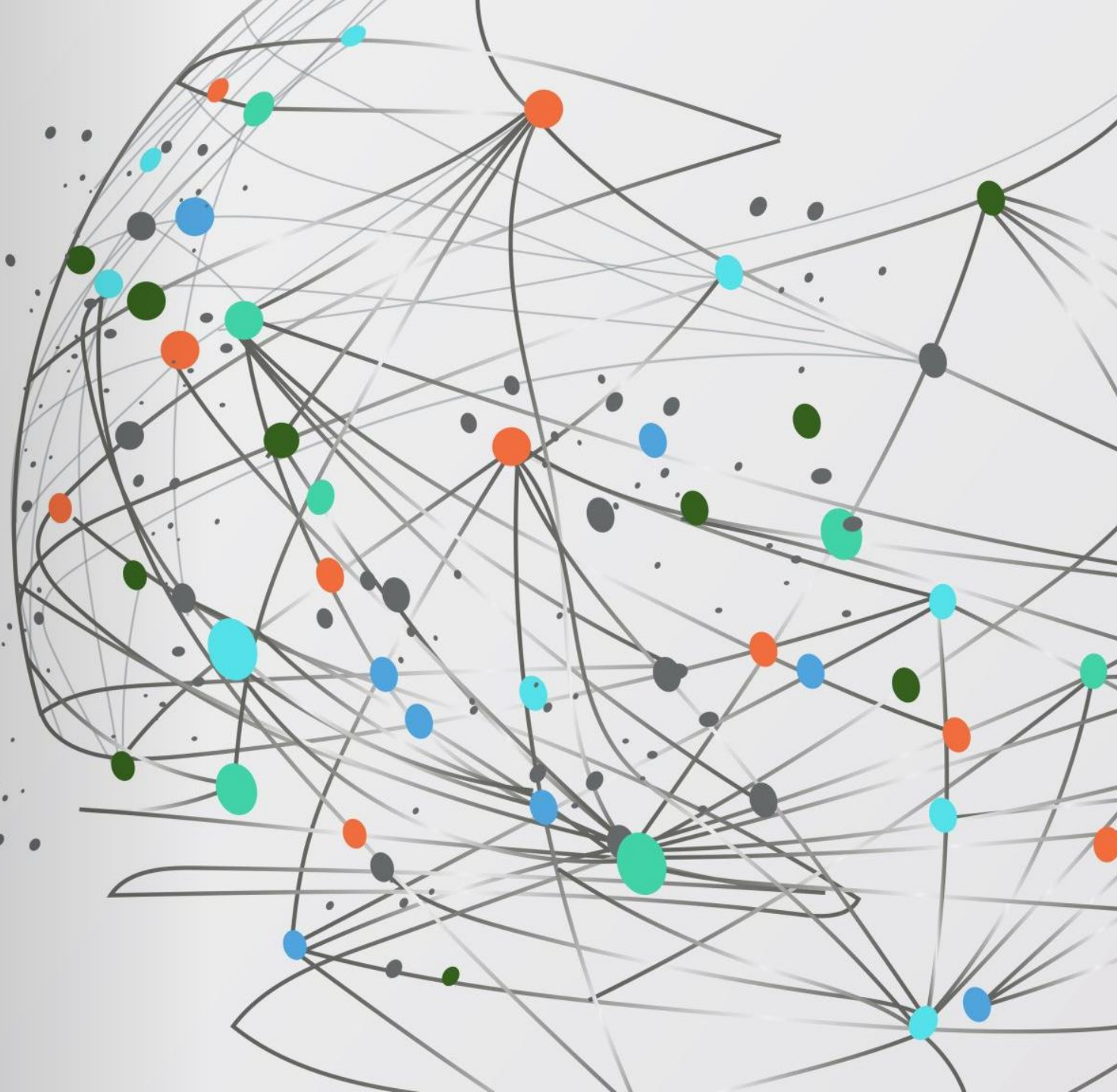


Haberman's Cancer Survival - Exploratory Data Analysis (EDA)

GO MY CODE PROJECT.

BY: ANGUISTA KUPEKA



1. Introduction

- Breast cancer remains one of the most prevalent cancers worldwide, making early detection and accurate prognosis critical for patient survival. This project focuses on analyzing the **Haberman's Cancer Survival** dataset, which contains historical data on breast cancer patients who underwent surgery at the University of Chicago's Billings Hospital between **1958 and 1970**. The goal is to uncover patterns that could help predict the likelihood of a patient surviving five years or more post-surgery, thereby aiding in clinical decision-making and improving patient outcomes.

2. Business Understanding


Understanding survival rates after breast cancer surgery is essential for healthcare providers to optimize treatment and improve patient outcomes. Accurate predictions can assist in:



Personalized Treatment Planning: Tailoring post-operative care based on individual survival risk.



Resource Allocation: Directing medical resources and attention to high-risk patients.



Improved Patient Outcomes: Identifying critical factors affecting recovery to enhance long-term survival rates.

Primary Objective:



The primary objective of this analysis is to identify the key factors influencing survival rates among breast cancer patients post-surgery. By understanding these factors, medical professionals can develop targeted treatment strategies, allocate resources more effectively, and improve overall patient care. Additionally, insights from this analysis can inform future research and support clinical decision-making.



By leveraging historical patient data, this project aims to explore patterns that could guide future clinical decisions and patient counseling.





3. Objective

- The specific goals of this exploratory data analysis are to:
- **Explore** the dataset to uncover trends and patterns related to patient survival.
- **Identify** the key features that influence the five-year survival rate post-surgery.
- **Predict** survival likelihood using variables such as age, year of operation, and the number of positive axillary lymph nodes.
- This analysis also aims to highlight potential data biases and uncover critical medical insights, particularly regarding the role of axillary lymph nodes in survival outcomes.
-

4. Data Understanding

- **Data Source:** Haberman's Cancer Survival Dataset (sourced from Kaggle).



Features:



Age: Patient's age at the time of surgery (numerical).



Operation Year: Year of the surgery (numerical, representing year - 1900).



Axillary Nodes: Number of positive axillary lymph nodes detected (numerical).



Survival Status: Target variable indicating survival status (1 = survived 5+ years, 0 = did not survive 5 years).

Dataset Overview:



Total records: 305 rows



Total features: 4 columns



No missing values detected.



The target variable was re-coded from {1, 2} to binary {1 = survived, 0 = did not survive}.



Initial Observations:

The majority of patients were between 30 and 83 years old.

Most surgeries took place between 1958 and 1970.

Axillary nodes counts varied, but most patients had fewer than 5 positive nodes.

5. Data Preparation



Data Cleaning:

- Renamed columns for clarity.
- Re-coded the Survival Status to binary values (1 = survived, 0 = did not survive).
- Verified data integrity and removed **19 duplicate records**.
- Identified outliers in features like Age and Axillary Nodes.

Data Transformation:

- Converted data types where necessary.
- No categorical variables were present; all features were numerical.
- Scaling was considered but deemed unnecessary for this exploratory analysis phase.



Feature Engineering:



No new features were engineered at this stage.

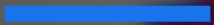


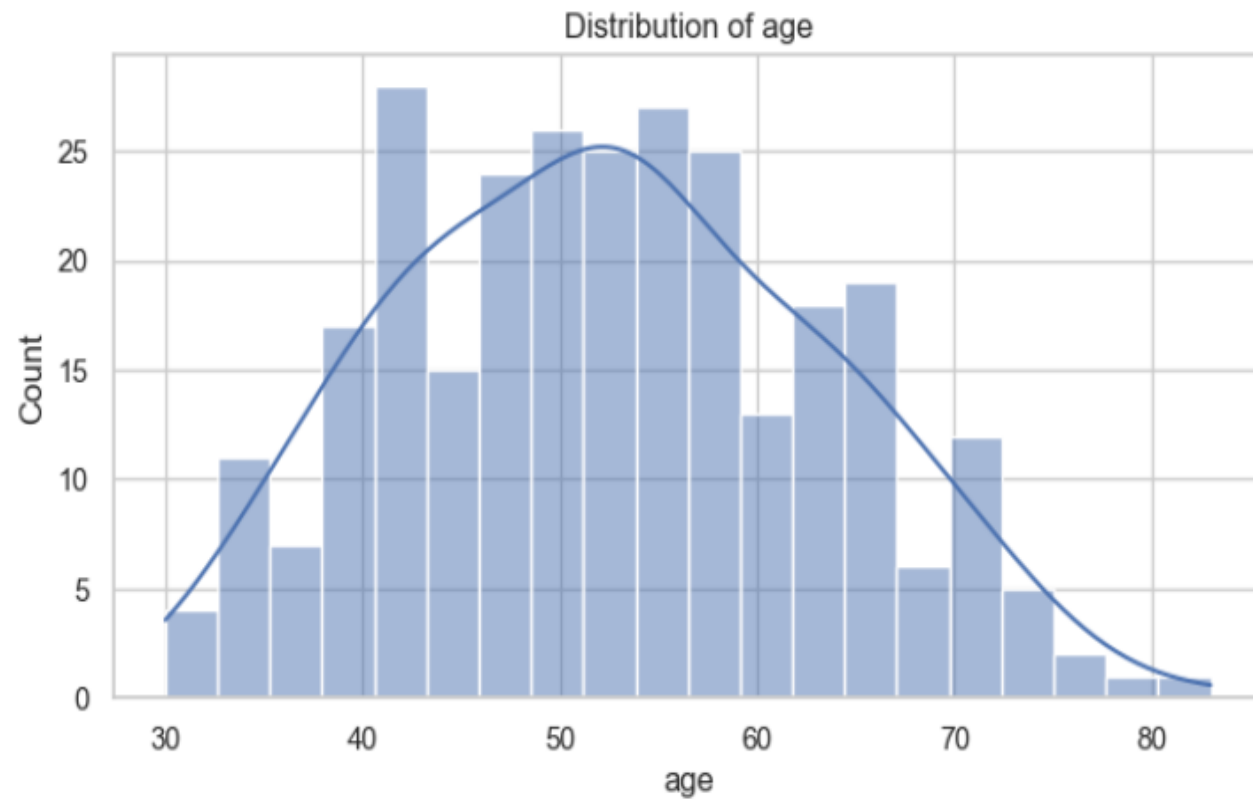
Future work may include deriving additional features (e.g., age brackets, node categories) for predictive modeling.

Data Analysis



Univariate Analysis:





1. Age:

o

Most patients were between 40 and 60 years old.

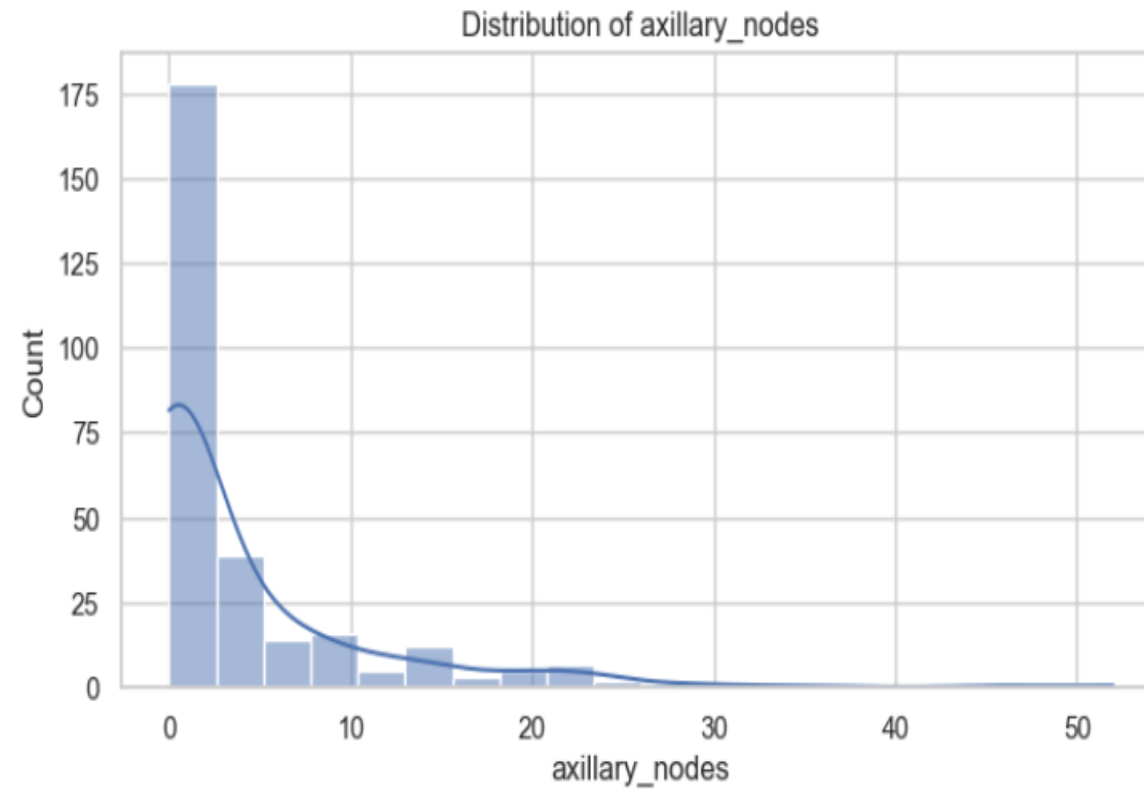
2. Operation Year:



o

Surgeries were evenly distributed between 1958 and 1970.

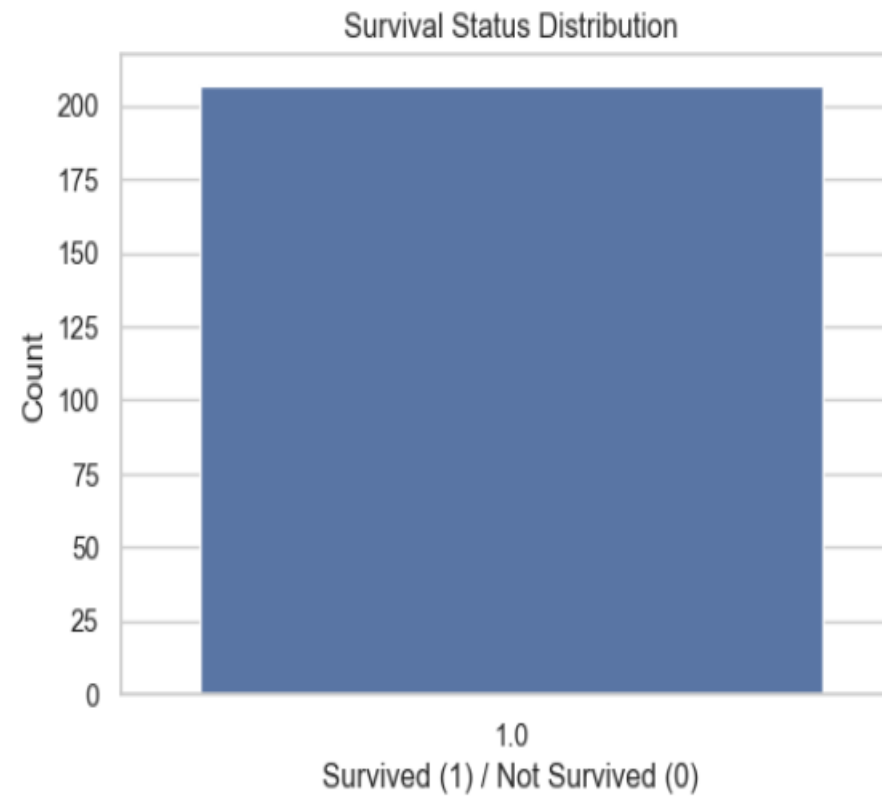
3. Axillary Nodes:



o

Majority of patients had fewer than 5 positive axillary nodes.

4. Survival Status:



o

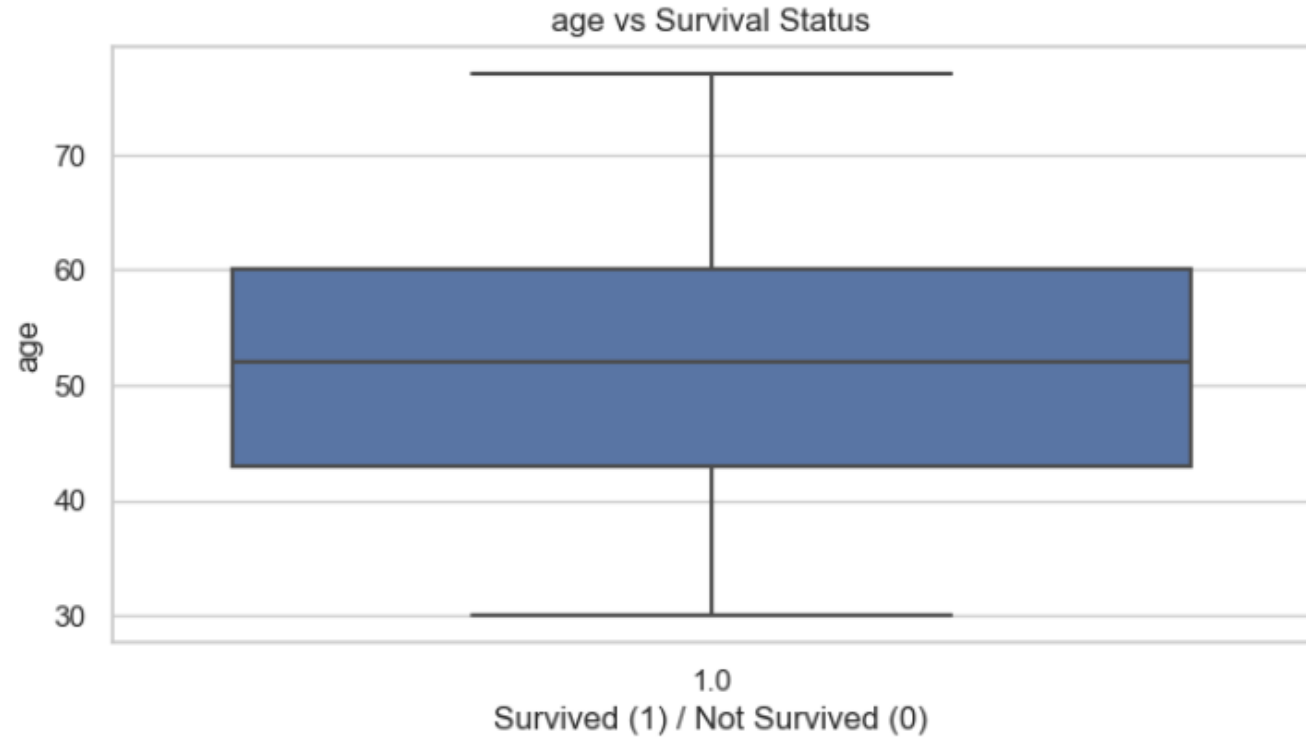
About 68% of patients survived 5 years or longer post-surgery.



Bivariate Analysis:



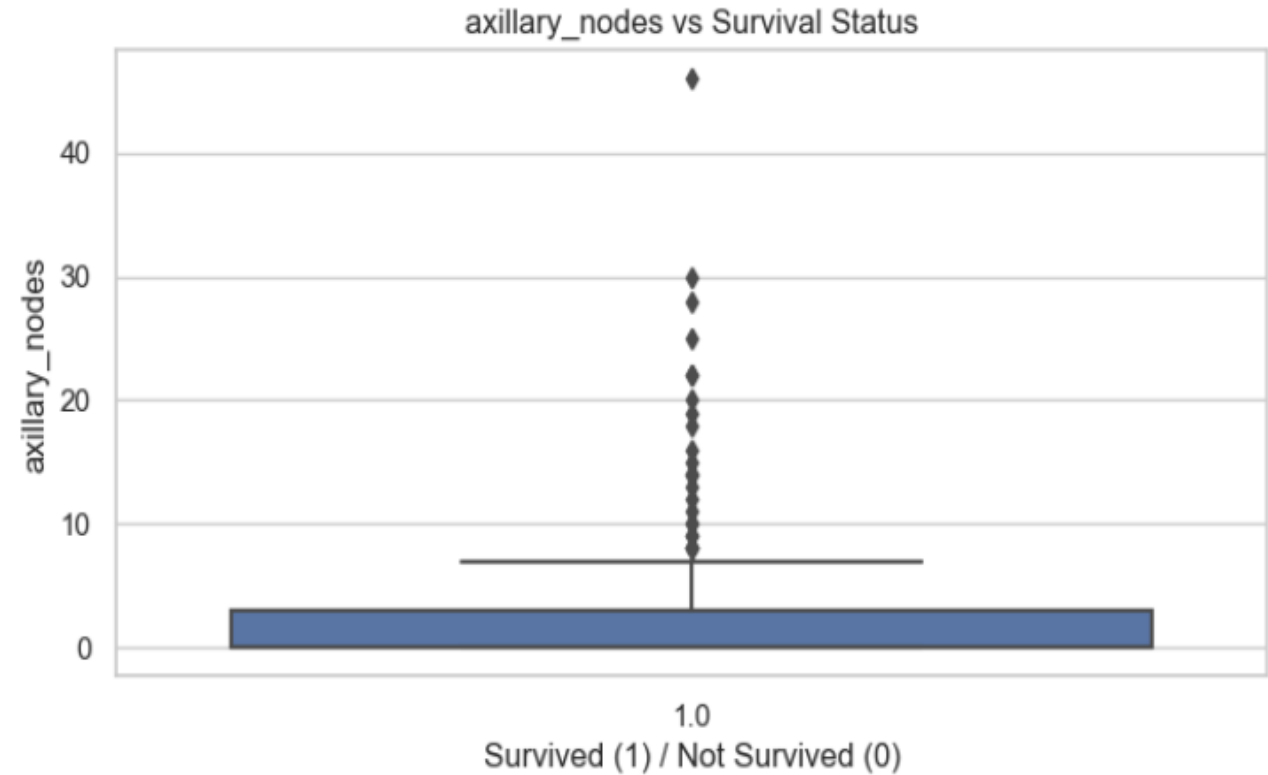
1. Age vs. Survival:



o

Younger patients had higher survival rates.

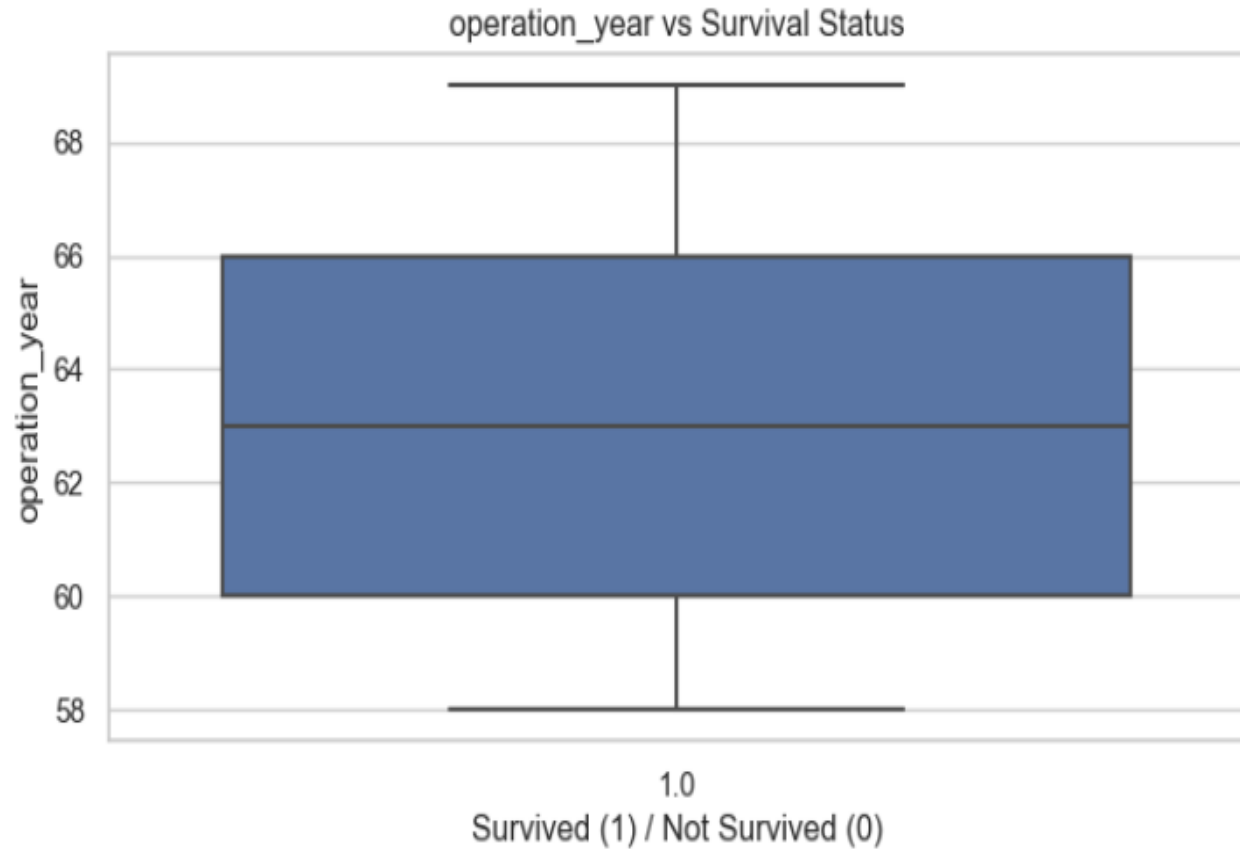
2. Axillary Nodes vs. Survival:



o

Patients with fewer positive axillary nodes had a significantly higher chance of survival.

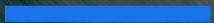
3. Operation Year vs. Survival:



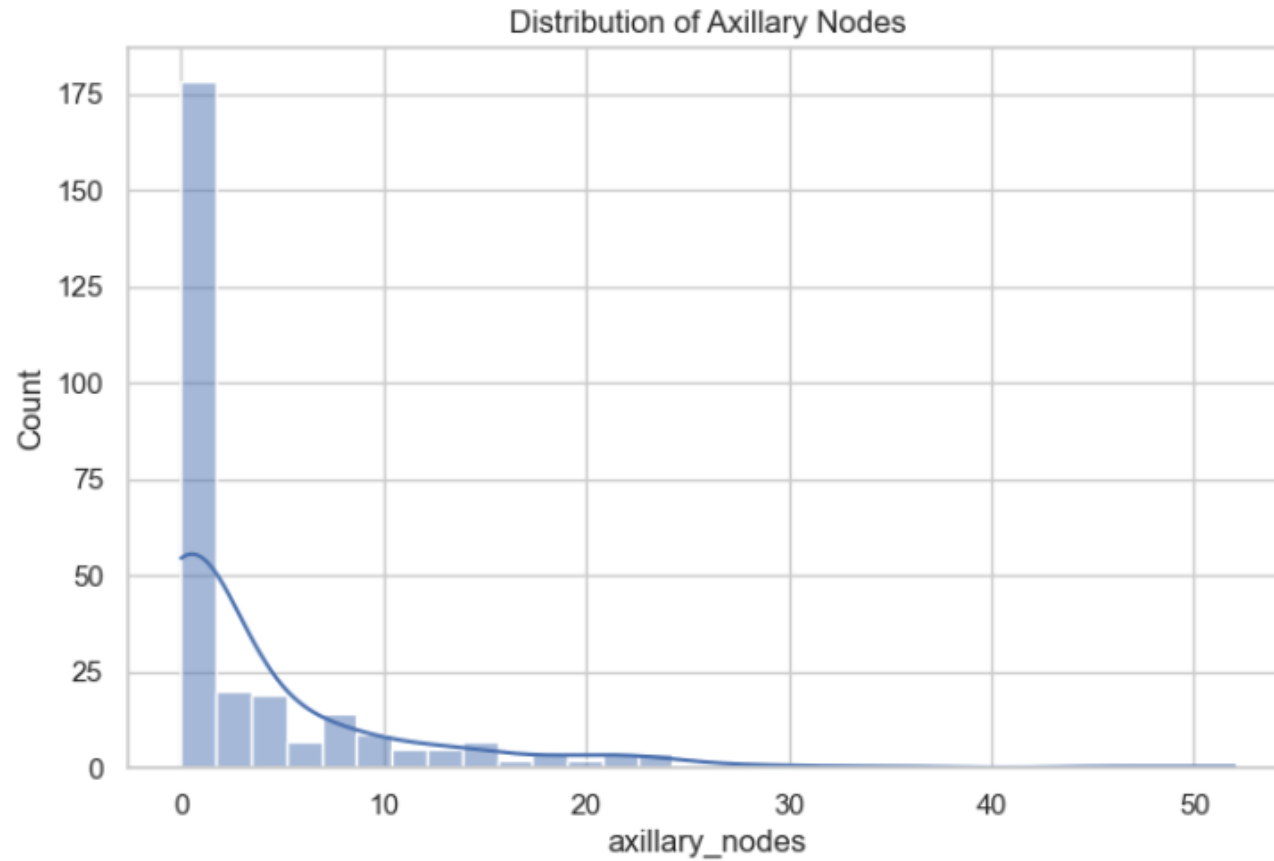
o

The year of surgery had minimal impact on survival rates.

Visualizations:



- 1. Distribution of Axillary Nodes:



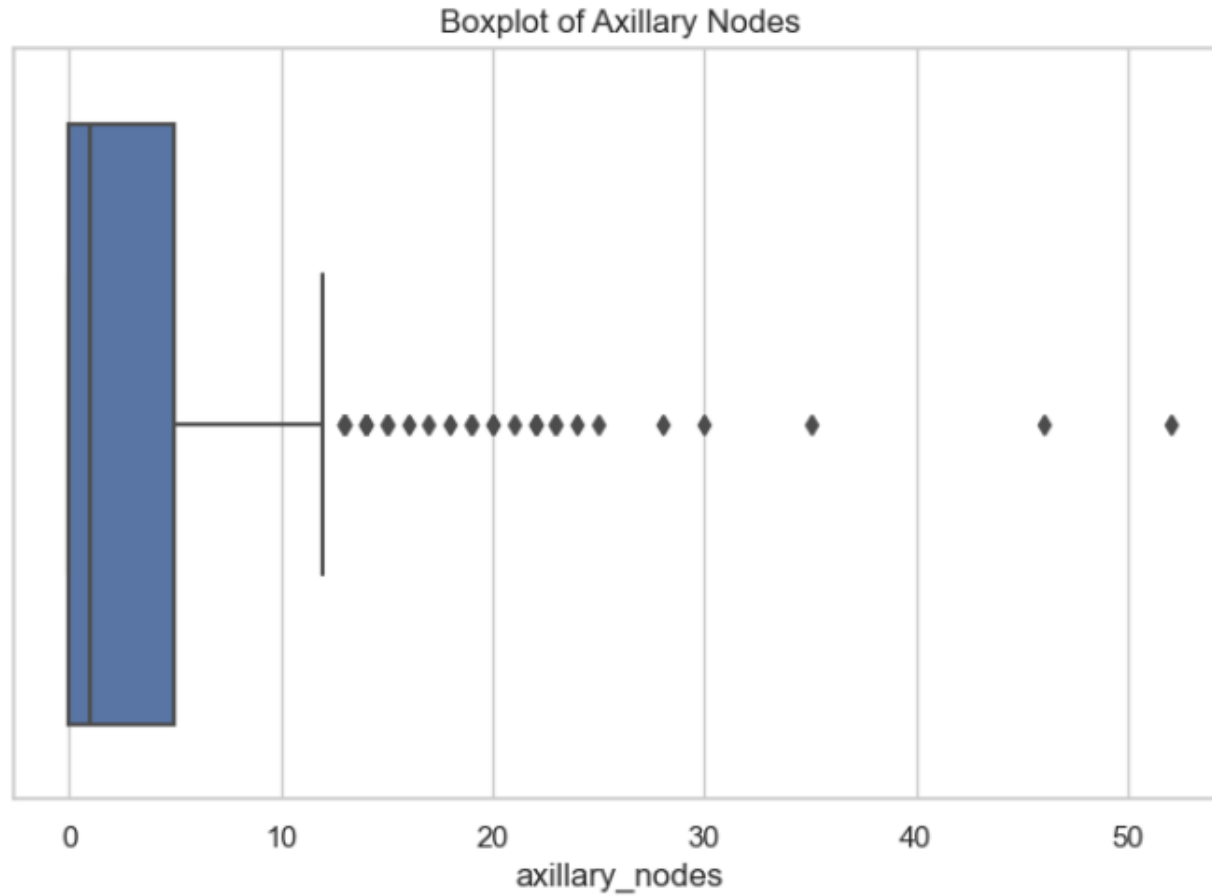
■

Showed skewed distribution of axillary nodes, with many patients having zero positive nodes.

Histograms:

—

- 2. Box Plots of Axillary Nodes:



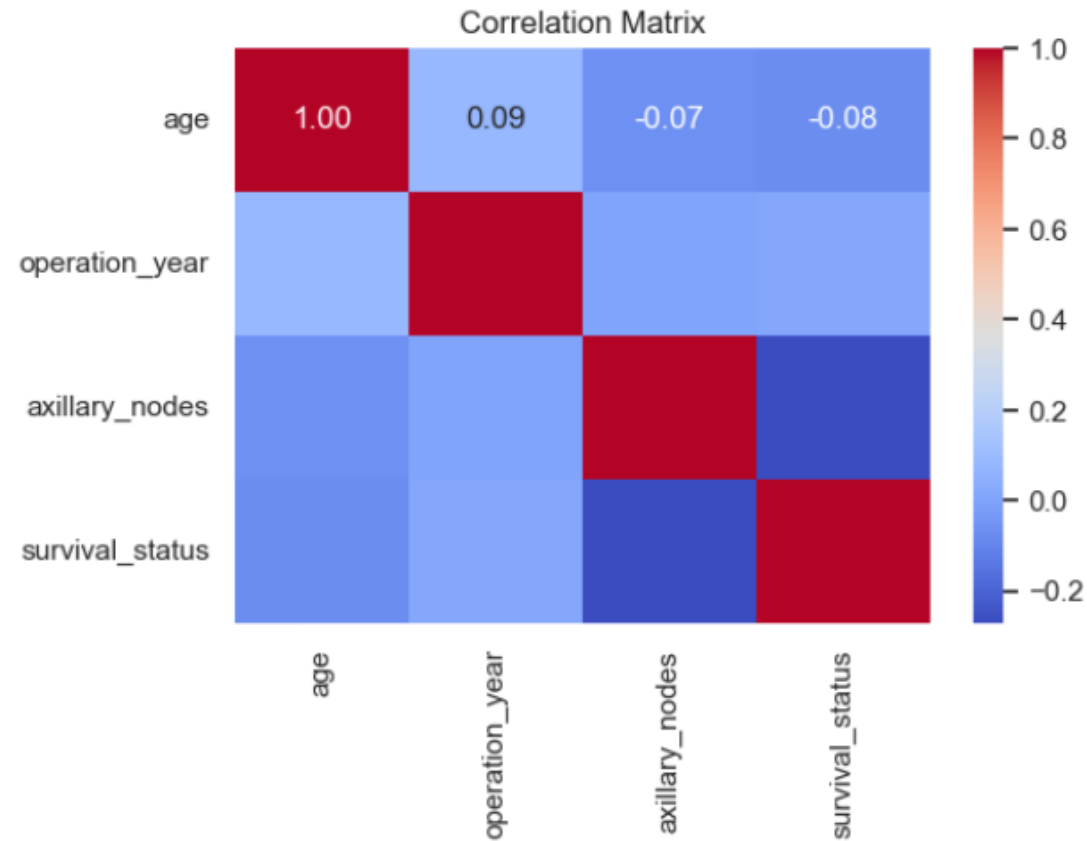
■

Highlighted outliers in axillary node counts among non-survivors.

Boxplots:

—

- 3. Correlation Matrix:



Heatmaps:

■

Revealed weak correlations between most features, except for axillary nodes, which had a notable negative correlation with survival.



Key Findings:

- **Age and axillary nodes** are significant predictors of survival.
- Patients with **fewer positive axillary nodes** had better survival rates.
- **Operation year** had no significant influence on survival outcomes.
- There is a **class imbalance** in the target variable, with more patients surviving than not.
-

7. Conclusion

- This analysis reveals that the **number of positive axillary nodes** is the most critical factor in predicting breast cancer survival post-surgery. Younger patients also show better survival rates, while the year of operation does not significantly affect outcomes.

Key takeaways:

01

Patients with **fewer positive axillary nodes** are more likely to survive beyond five years.

02

Age plays a role, with younger patients generally having better outcomes.

03

There is a notable **class imbalance** in the dataset that should be addressed in predictive modeling efforts.

04

These findings can help medical professionals focus on critical factors like lymph node involvement when planning post-operative care and follow-ups.



8. Recommendations



Business-Focused Recommendations:



Early Detection and Screening: Promote regular screenings, especially for women over **40 years old**, to detect cancer early and reduce the likelihood of positive axillary nodes.



Targeted Post-Surgery Care: Implement personalized post-operative care plans for patients with higher axillary node counts to improve survival chances.



Patient Education: Educate patients on the importance of **early detection** and the impact of positive axillary nodes on survival outcomes.



Resource Allocation: Allocate medical resources more efficiently by prioritizing high-risk patients (those with numerous positive axillary nodes) for intensive follow-up and monitoring.

Technical Recommendations:

Further Modeling:

Apply machine learning models (e.g., **Logistic Regression**, **Random Forest**, **Decision Trees**) to predict survival probabilities.
Address the class imbalance using techniques like **SMOTE** or **class weighting**.

Feature Scaling (if needed):

If advanced algorithms like **SVM** or **KNN** are used, apply scaling techniques such as **StandardScaler** or **MinMaxScaler**.

Data Enhancement:

Incorporate more recent data reflecting modern surgical techniques and treatments.
Explore adding features such as **tumor size**, **hormone receptor status**, and **genetic markers** to improve predictive power.

THANK YOU!

EMAIL: ANGUISTAK@GMAIL.COM

GIT HUB:
[HTTPS://GITHUB.COM/ANGUISTA/GO-MY-
CODE-PROJECT](https://github.com/anguista/go-my-code-project)