QBS121_final

2025-03-01

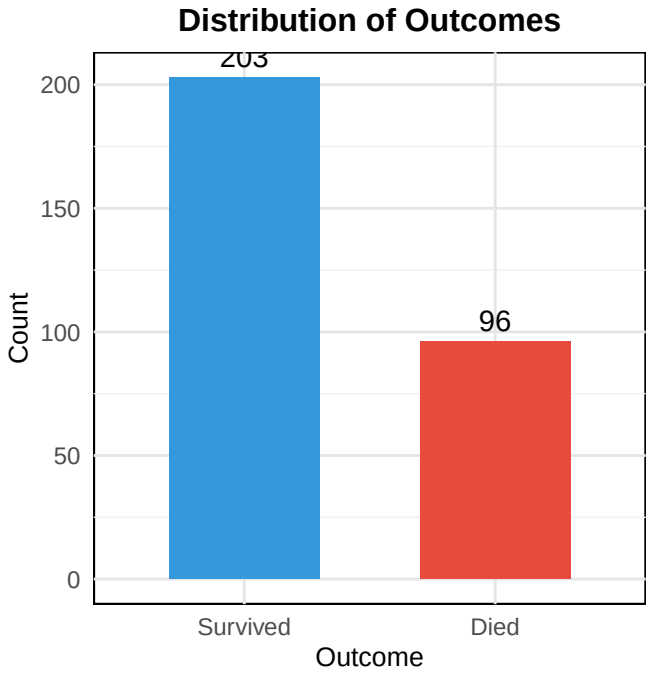
Data EDA

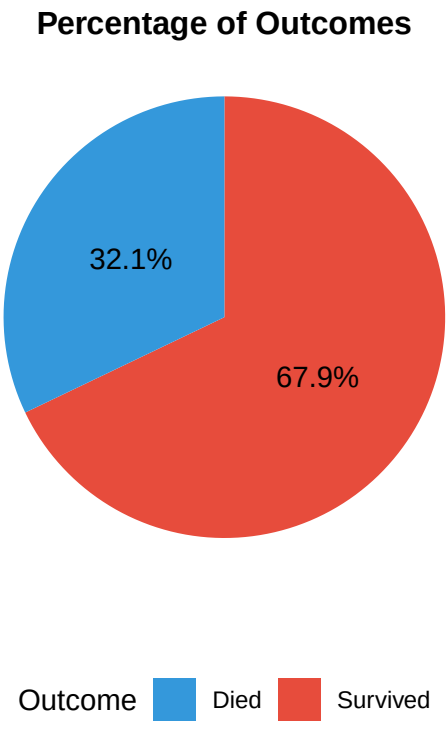
Load the data

Visualization of variables

```
## Loading required package: viridisLite
```

Visualization of the target variable

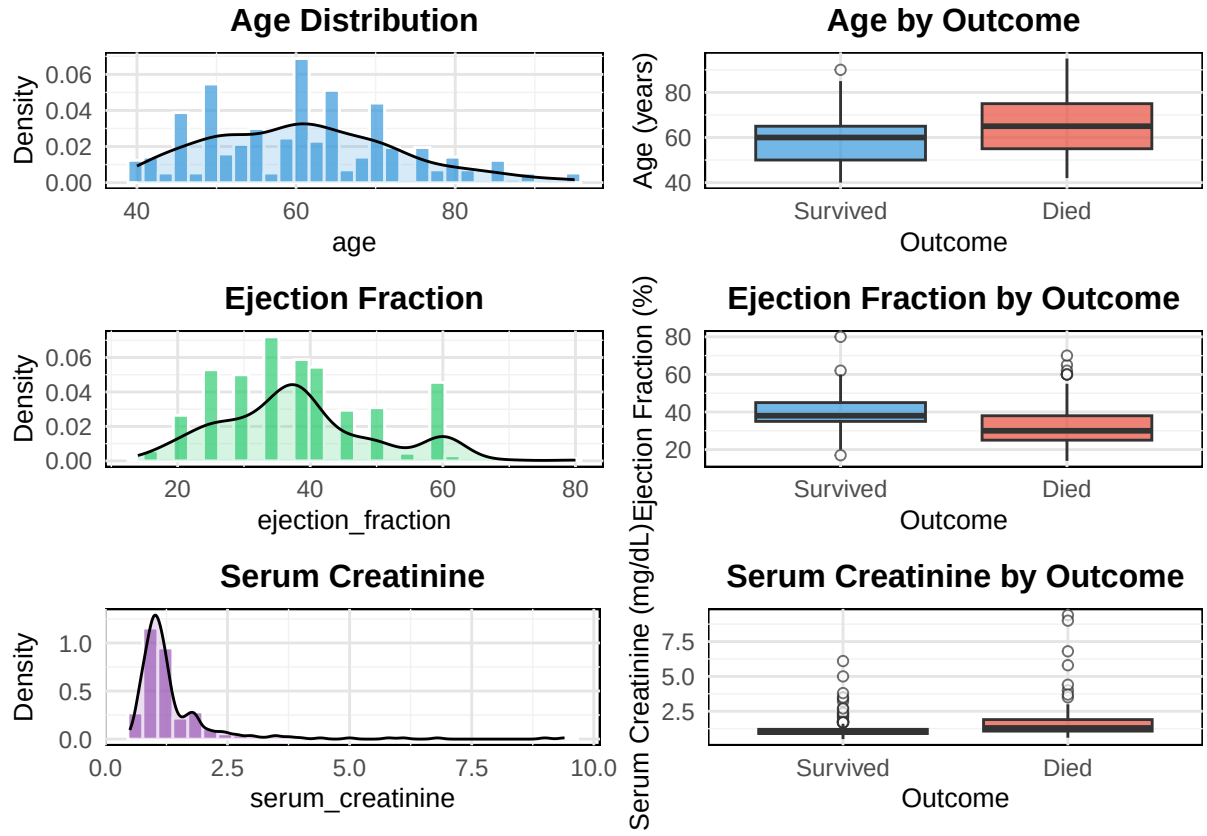




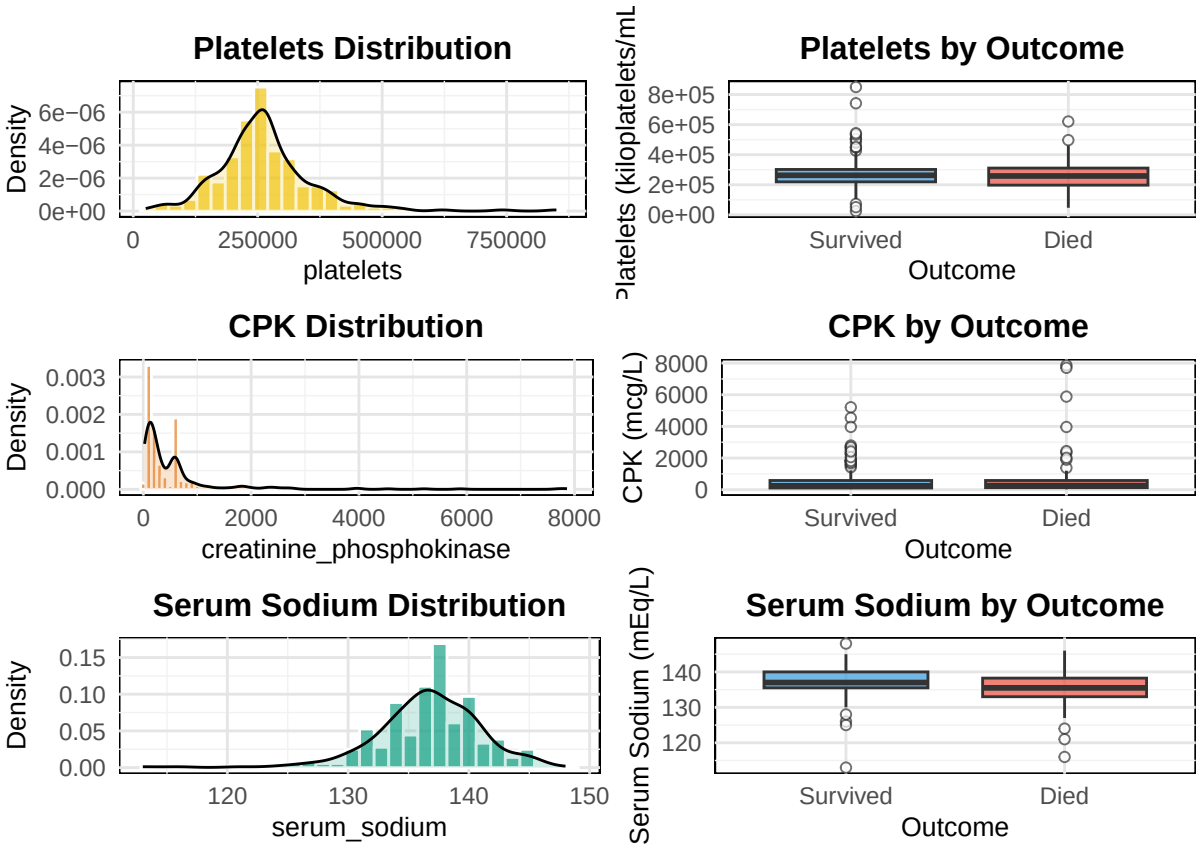
Explore numerical variables

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

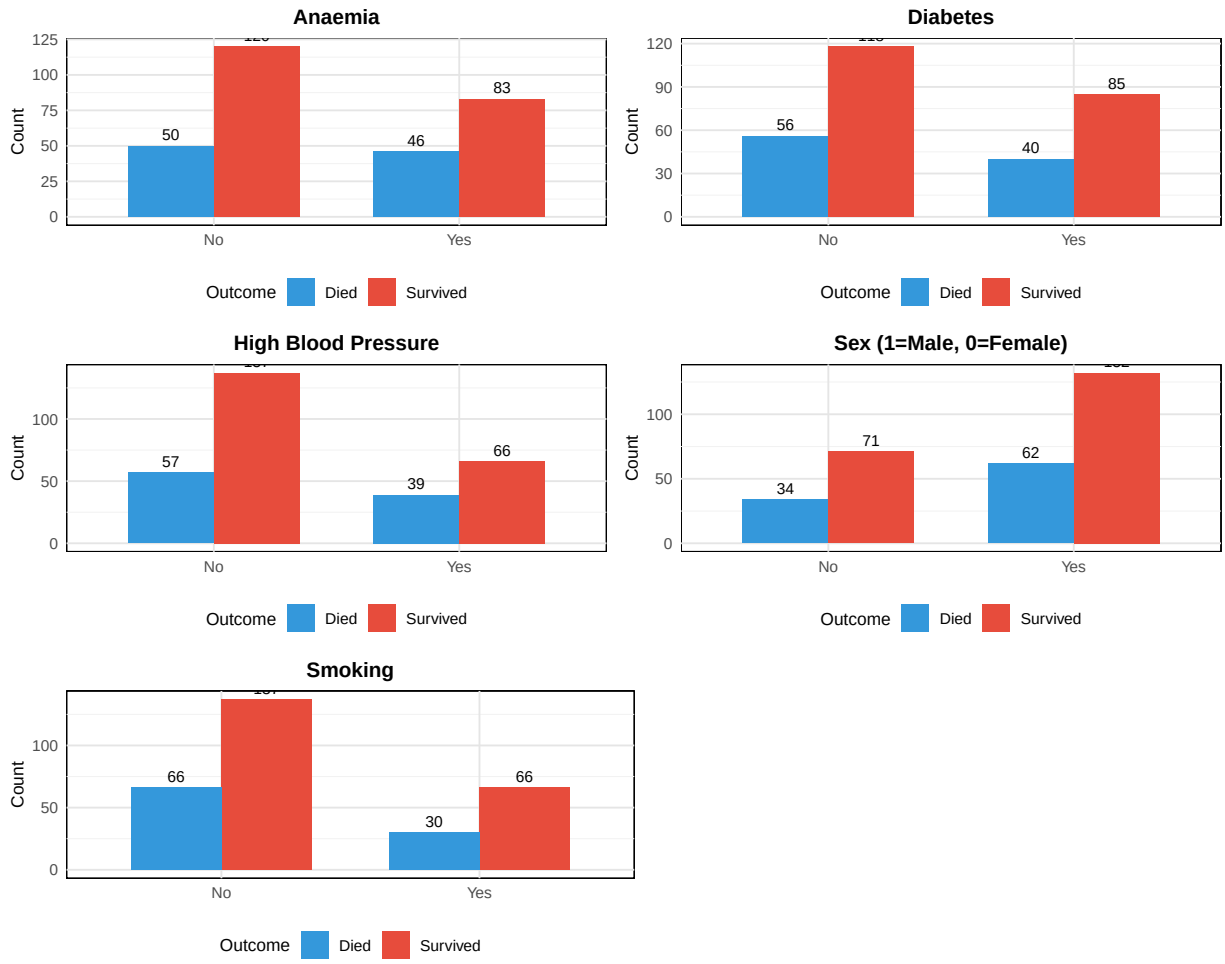
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



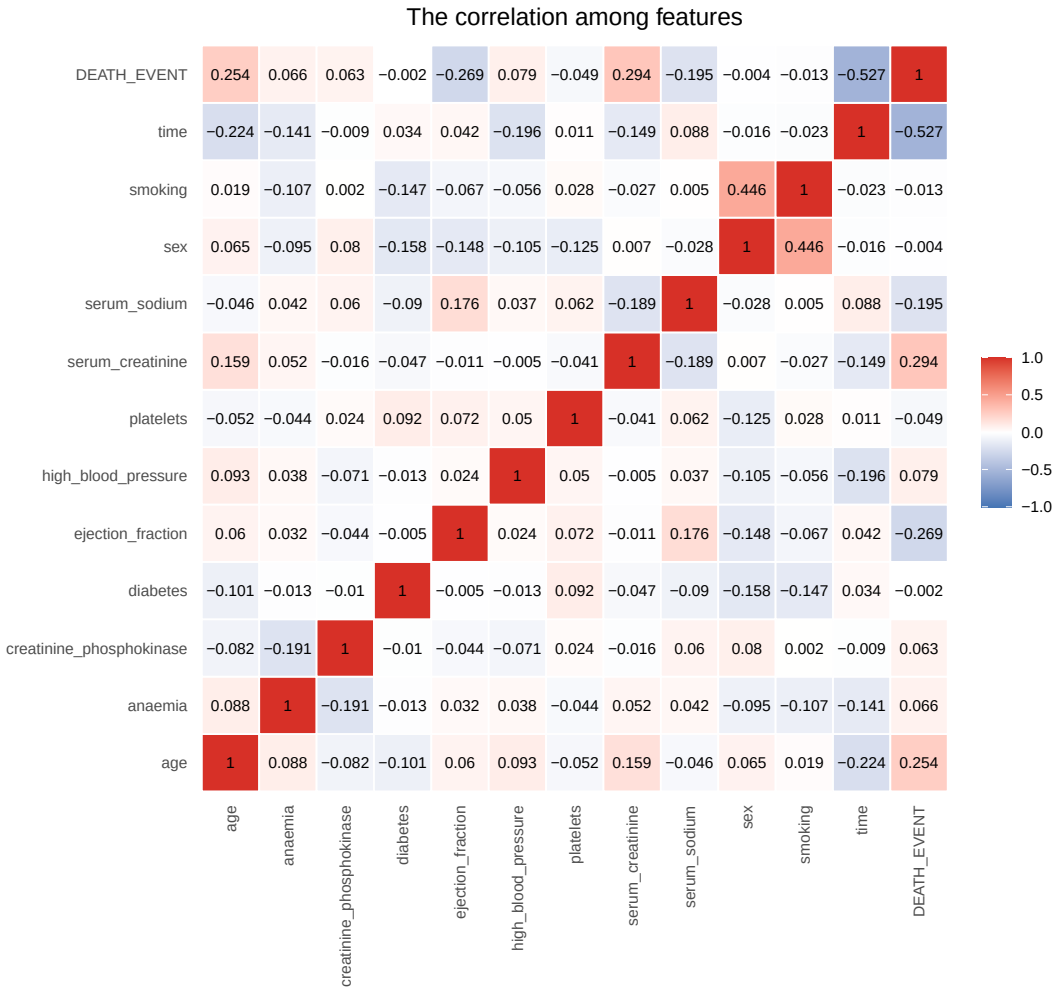

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Explore categorical variables



Correlation analysis



Exploratory Data Analysis Summary

This exploratory data analysis examines a dataset of 299 patients who had heart failure, focusing on patient characteristics and mortality outcomes.

Dataset Overview

The dataset consists of 13 clinical features collected during patient follow-up periods:

- Demographic: age (years), sex (binary: 1=male, 0=female)
- Clinical conditions: anaemia, diabetes, high blood pressure, smoking (all binary: 1=yes, 0=no)
- Heart measurements: ejection fraction (percentage), creatinine phosphokinase (CPK, mcg/L)
- Blood measurements: platelets (kiloplatelets/mL), serum creatinine (mg/dL), serum sodium (mEq/L)
- Outcome: time (follow-up period in days), death event (1=yes, 0=no)

Target Variable Distribution

The dataset shows an imbalanced distribution of the target variable (death event):

- 203 patients survived (67.9%)
- 96 patients died (32.1%)

Numerical Variables

Analysis of key numerical variables reveals several patterns:

Age: Patients' ages range from approximately 40 to 95 years, with a roughly normal distribution. Deceased patients tend to be slightly older than survivors, though there is substantial overlap.

Ejection Fraction: Shows a multimodal distribution with peaks around 20%, 38%, and 60%. Lower ejection fraction values appear associated with higher mortality, supporting its known clinical significance in heart failure.

Serum Creatinine: Exhibits a right-skewed distribution with most values below 2.5 mg/dL. Deceased patients show slightly higher creatinine levels, indicating possible renal dysfunction association with mortality.

Platelets: Normally distributed around 250,000 kiloplatelets/mL, with no clear difference between survivors and non-survivors.

CPK: Highly right-skewed with most values below 2000 mcg/L but some extreme outliers exceeding 7000. No substantial difference observed between outcome groups.

Serum Sodium: Normally distributed around 135-140 mEq/L, with deceased patients showing slightly lower values, suggesting hyponatremia might be associated with poorer outcomes.

Categorical Variables

The categorical variable analysis shows several patterns:

Anaemia: Present in 129 patients (43%), with higher mortality proportion among anaemic patients.

Diabetes: Present in 125 patients (42%), with slightly higher mortality proportion.

High Blood Pressure: Present in 105 patients (35%), with slightly higher mortality rate among those without hypertension.

Sex: The dataset includes 194 males (65%) and 105 females. Males show a slightly higher mortality rate.

Smoking: Present in 96 patients (32%), with smokers showing lower mortality compared to non-smokers.

Correlation Analysis

The correlation heatmap reveals several significant relationships:

Strong correlations:

- Death event and time (-0.53): Negative correlation indicating patients who died had shorter follow-up periods
- Sex and smoking (0.45): Males more likely to be smokers
- Death event and ejection fraction (-0.27): Lower ejection fraction associated with higher mortality

- Death event and serum creatinine (0.29): Higher serum creatinine associated with higher mortality
- Age and death event (0.25): Older age associated with higher mortality

Moderate correlations:

- High blood pressure and time (-0.20): Patients with hypertension had shorter follow-up periods
- Serum sodium and death event (-0.20): Lower sodium levels associated with mortality
- Anaemia and creatinine phosphokinase (-0.19): Inverse relationship

Key Findings

1. The three strongest predictors of mortality appear to be time (follow-up period), ejection fraction, and serum creatinine.
2. Age shows a positive correlation with mortality risk, confirming the expected clinical relationship.
3. Ejection fraction is inversely related to mortality, consistent with heart failure pathophysiology.
4. Serum abnormalities (creatinine, sodium) show associations with mortality, suggesting the importance of metabolic factors.
5. Categorical variables (anaemia, diabetes, hypertension, sex, smoking) show some patterns with mortality but with weaker associations than continuous measurements.

These findings align with clinical understanding of heart failure risk factors and highlight the complex interplay between demographic, cardiac, and metabolic parameters in predicting mortality in heart failure patients.