EXPLORATORY DATA ANALYSIS USING R

HEART FAILURE CLINICAL RECORDS

Nada Haboudal

Contents

DATASET:
HEART FAILURE CLINICAL RECORDS. (2020). UCI MACHINE LEARNING REPOSITORY. HTTPS://DOI.ORG/10.24432/C5Z89R.

1.SUMMARY OF DATASET

Description: df [13 × 3]

Column <chr>	Type <chr>	Range <chr>
age	numeric	40 – 95
anaemia	Categorical	No, Yes
creatinine_phosphokinase	numeric	23 – 7861
diabetes	Categorical	No, Yes
ejection_fraction	numeric	14 – 80
high_blood_pressure	Categorical	No, Yes
platelets	numeric	25100 – 850000
serum_creatinine	numeric	0.5 – 9.4
serum_sodium	numeric	113 – 148
sex	Categorical	No, Yes
smoking	Categorical	No, Yes
time	numeric	4 – 285
DEATH_EVENT	Categorical	No, Yes

13 rows

2.1
EXPLORE FACTORS EFFECTING MORTALITY
CHECK FOR NORMALITY

Column <chr>	Shapiro_Wilk_p_value <dbl>	Normality <chr>
age	5.349669e-05	Not Normal
creatinine_phosphokinase	7.050459e-28	Not Normal
ejection_fraction	7.215954e-09	Not Normal
platelets	2.883451e-12	Not Normal
serum_creatinine	5.392797e-27	Not Normal
serum_sodium	9.214858e-10	Not Normal
time	6.284953e-09	Not Normal

7 rows

2.1
EXPLORE FACTORS EFFECTING MORTALITY
MANN-WHITNEY U (CONTINUES NON NORMAL DATA)
CHI-SQUARED (CATAGORICAL DATA)

Column	Test	P_Value	Statistical_Significance
time	Mann-Whitney U	6.852197e-21	TRUE
serum_creatinine	Mann-Whitney U	1.580998e-10	TRUE
ejection_fraction	Mann-Whitney U	7.368249e-07	TRUE
age	Mann-Whitney U	1.667518e-04	TRUE
serum_sodium	Mann-Whitney U	2.927557e-04	TRUE
high_blood_pressure	Chi-Square	2.141034e-01	FALSE
anaemia	Chi-Square	3.073161e-01	FALSE
platelets	Mann-Whitney U	4.255585e-01	FALSE
creatinine_phosphokinase	Mann-Whitney U	6.840400e-01	FALSE
smoking	Chi-Square	9.317653e-01	FALSE
diabetes	Chi-Square	1.000000e+00	FALSE
sex	Chi-Square	1.000000e+00	FALSE

2.2
EXPLORE FACTORS EFFECTING MORTALITY
LOGISTIC REGRESSION

```
Call:
glm(formula = DEATH_EVENT ~ age + ejection_fraction + serum_creatinine +
    serum_sodium + diabetes + smoking, family = binomial(link = "logit"),
    data = dataset_new)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.85113    4.48411   0.859    0.390
age                0.05273    0.01256   4.200 2.67e-05 ***
ejection_fraction -0.06739    0.01439  -4.684 2.81e-06 ***
serum_creatinine   0.62950    0.15787   3.988 6.68e-05 ***
serum_sodium      -0.04647    0.03240  -1.434    0.151
diabetesYes        0.16605    0.29092   0.571    0.568
smokingYes        -0.12361    0.30713  -0.402    0.687
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 375.35  on 298  degrees of freedom
Residual deviance: 302.55  on 292  degrees of freedom
AIC: 316.55

Number of Fisher Scoring iterations: 5

      (Intercept)              age ejection_fraction  serum_creatinine       serum_sodium       diabetesYes        smokingYes
       47.0461597        1.0541498        0.9348347        1.8766671        0.9545899        1.1806271        0.8837284
```

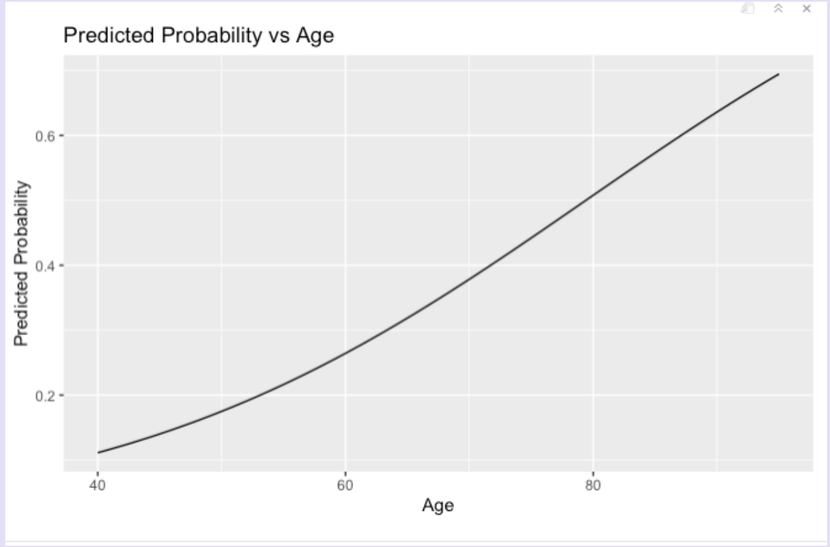
2.2 QUESTION: HOW DO MULTIPLE CLINICAL FACTORS SIMULTANEOUSLY IMPACT THE RISK OF DEATH IN HEART FAILURE PATIENTS?

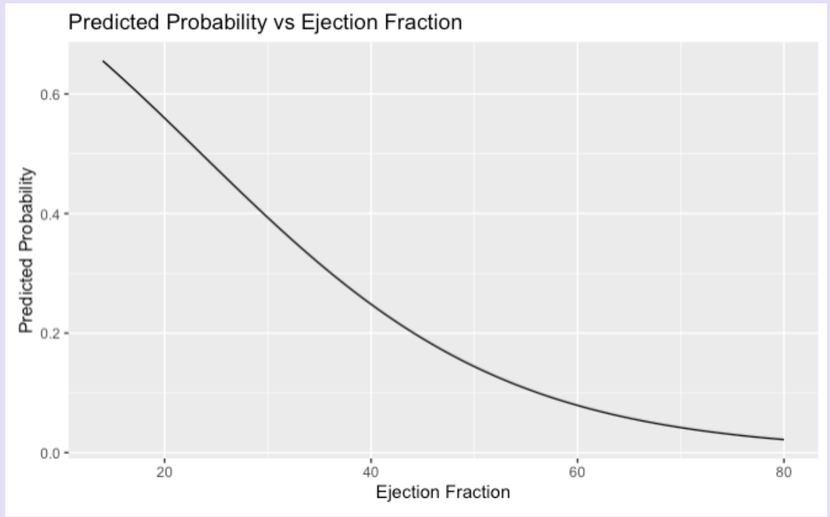
Based on the output of the logistic regression model:

- **Age:** The positive coefficient suggests that as age increases, so does the risk of death. It is statistically significant ($p < 0.001$), indicating a reliable predictor.
- **Ejection Fraction:** The negative coefficient indicates that higher the ejection fraction values are associated with a lower risk of death, and this predictor is also statistically significant ($p < 0.001$).
- **Serum Creatinine:** This has a positive coefficient and is statistically significant ($p < 0.001$), suggesting that higher serum creatinine levels increase the risk of death.
- **Serum Sodium:** The coefficient is negative, but it is not statistically significant ($p > 0.05$), implying that serum sodium levels may not be a reliable predictor of death risk in this model.
- **Diabetes:** The positive coefficient is not statistically significant ($p > 0.05$), suggesting that diabetes, as coded, may not be associated with the risk of death in this model.
- **Smoking:** The negative coefficient is not statistically significant ($p > 0.05$), indicating that smoking status may not be a significant predictor of death risk.

The significant predictors of mortality risk in this cohort of heart failure patients are age, ejection fraction, and serum creatinine levels.

Each increase in these parameters (age and serum creatinine) is associated with an increased risk of death, while higher ejection fraction is associated with decreased risk.





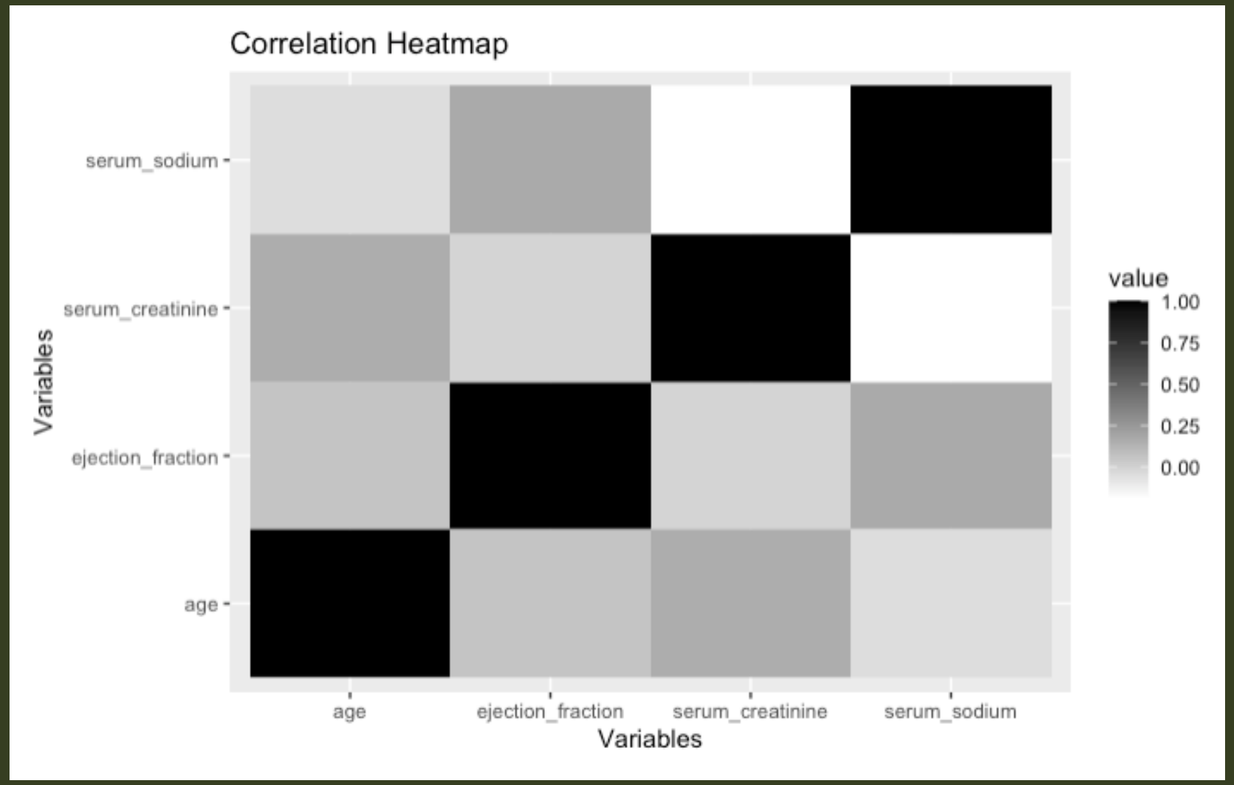
3.SUMMARY STATISTICS FOR DEPENDENT FEATURES

- Age: Patients who experienced death (DEATH_EVENT=1) tend to be older on average.
- Ejection Fraction: Patients who survived (DEATH_EVENT=0) have a higher average ejection fraction, indicating better heart function
- Serum Creatinine: Patients who did not survive have higher average serum creatinine levels, which may indicate kidney issues.

Summary Table		
DEATH_EVENT	0.0000000	1.000000
Mean_Age	58.7619064	65.215281
SD_Age	10.6378902	13.214556
Median_Age	60.0000000	65.000000
Mean_Ejection_Fraction	40.2660099	33.468750
SD_Ejection_Fraction	10.8599627	12.525303
Median_Ejection_Fraction	38.0000000	30.000000
Mean_Serum_Creatinine	1.1848768	1.835833
SD_Serum_Creatinine	0.6540827	1.468562
Median_Serum_Creatinine	1.0000000	1.300000

4.VISUALIZATIONS

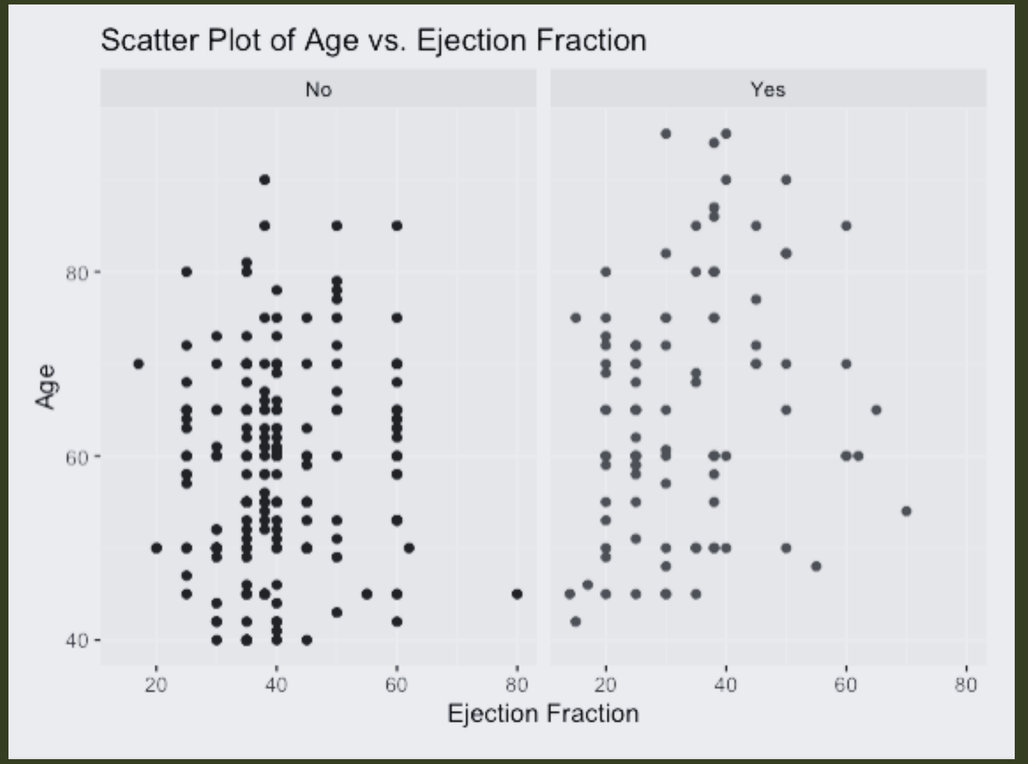
The heatmap indicates a stronger relationship between serum creatinine and ejection fraction than with age or serum sodium.



4.VISUALIZATIONS

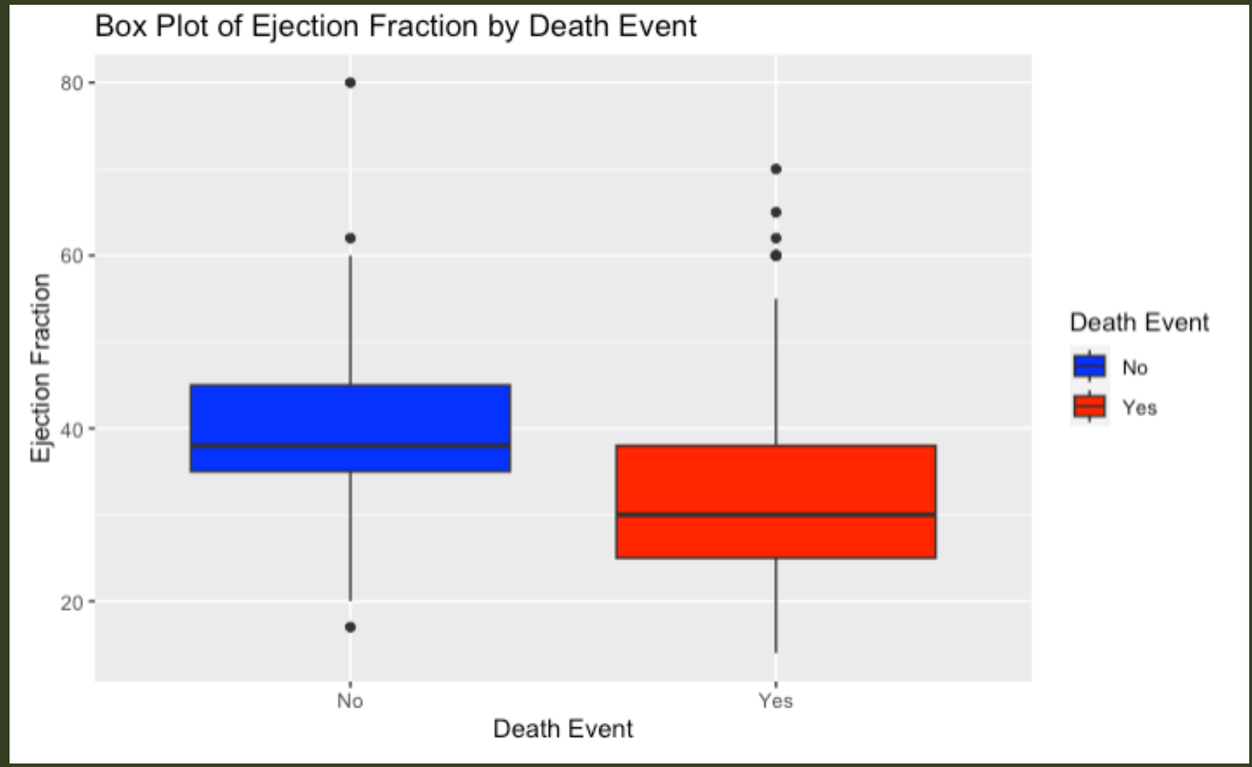
This scatter plot suggests no strong correlation between age and ejection fraction for those who did not experience a death event, as the points are spread throughout the plot.

For those with a death event, while also spread, there's a slight indication of higher age and lower ejection fraction occurrences together.



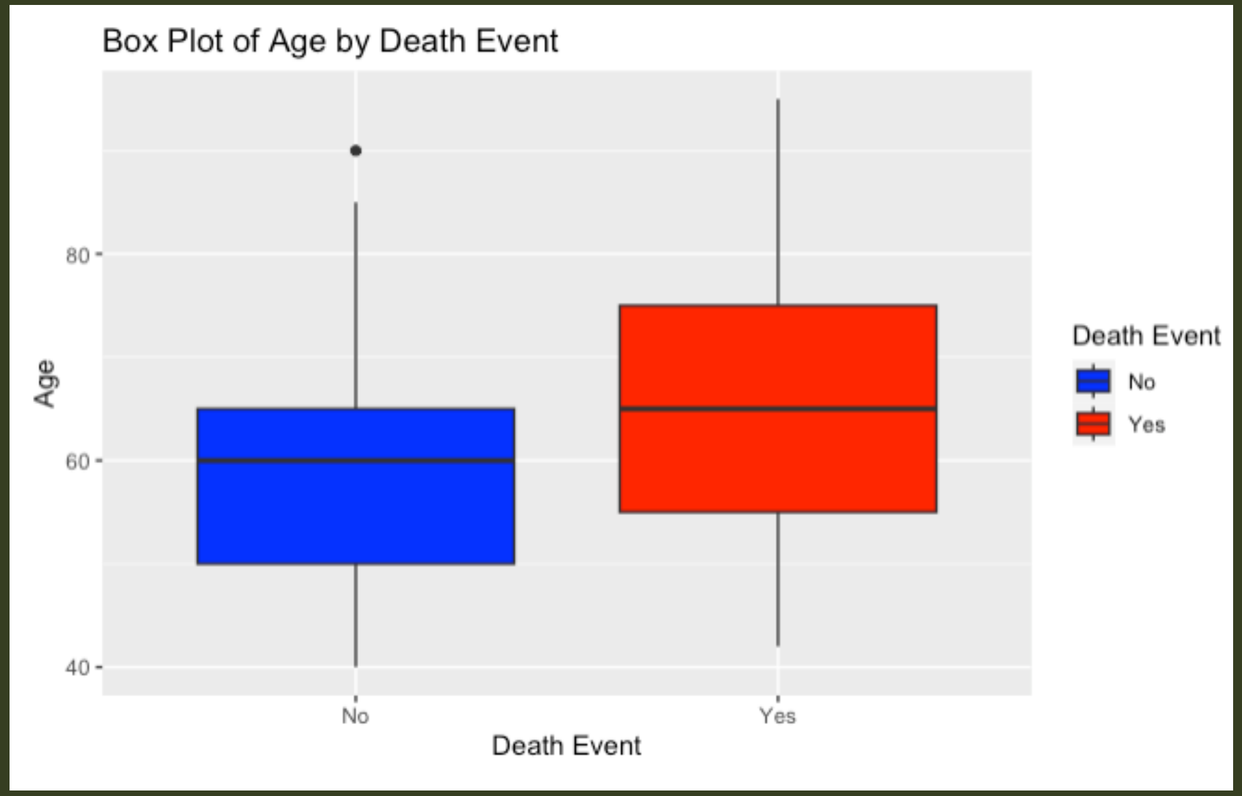
4.VISUALIZATIONS

Lower ejection fractions are more commonly associated with death events.



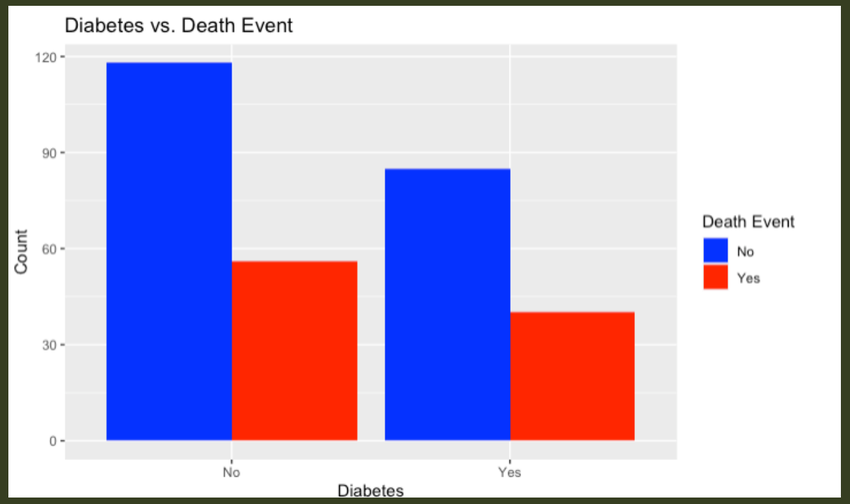
4.VISUALIZATIONS

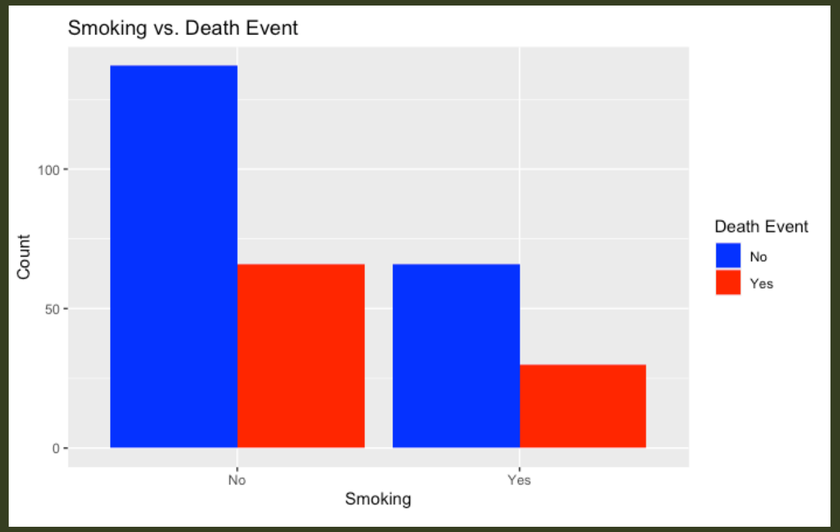
Those who experienced a death event are generally older.

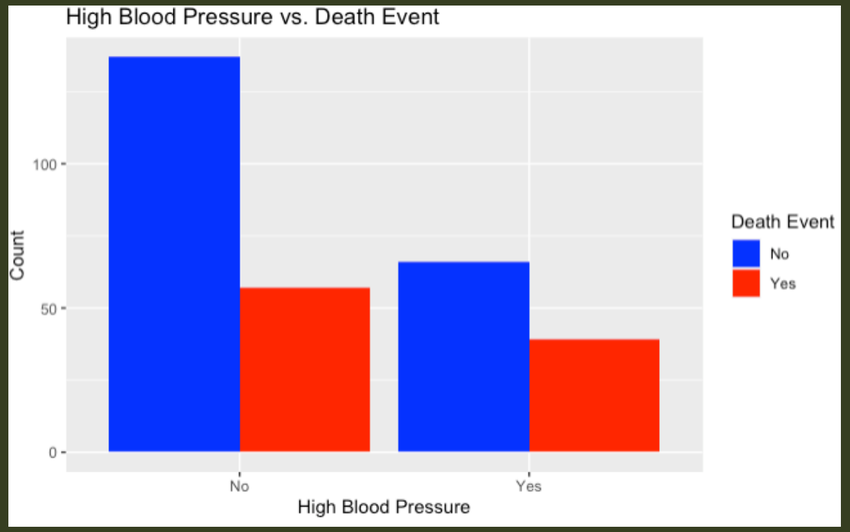


4.VISUALIZATIONS

High blood pressure, diabetes, and smoking
do not show a clear association with death events.







4.VISUALIZATIONS SUMMARY

These visualizations suggest that higher serum creatinine levels and lower ejection fractions are associated with a higher occurrence of death events.

High blood pressure, diabetes, and smoking status do not show a clear association with death events.

Age shows a trend where older individuals are more likely to have experienced a death event.

5.FUTURE CONSIDERATIONS

Transforming categorical variables like smoking, diabetes, and blood pressure into more granular numerical values could reveal more nuanced correlations with death events.

Smoking: Instead of a binary "Yes/No" for smoking, we could quantify this as "Years of Smoking," which would represent the total number of years an individual has been smoking. This provides a continuous variable that reflects both the duration and potential intensity of smoking habits.

Diabetes: For diabetes, a possible numerical transformation could be "Diabetes Control Score," which could be based on a composite measure of glycemic control (like HbA1c levels).
By quantifying these variables, we might be able to better capture their effects on the risk of death events and potentially uncover patterns that are not visible with categorical data.

It's essential to ensure that these numerical values are clinically meaningful and based on actual health records for accurate analysis.