Cirrhosis Dataset

Exploratory Data Analysis Report

Generated: February 09, 2026 at 03:31 PM

1. EXECUTIVE SUMMARY

This report presents a comprehensive Exploratory Data Analysis (EDA) of the Cirrhosis dataset. The dataset contains **418** patient records with **20** features including demographic information, clinical symptoms, and laboratory measurements. The analysis covers data quality assessment, statistical summaries, distribution analysis, correlation studies, and feature relationships with the target variable (Status).

2. DATASET OVERVIEW

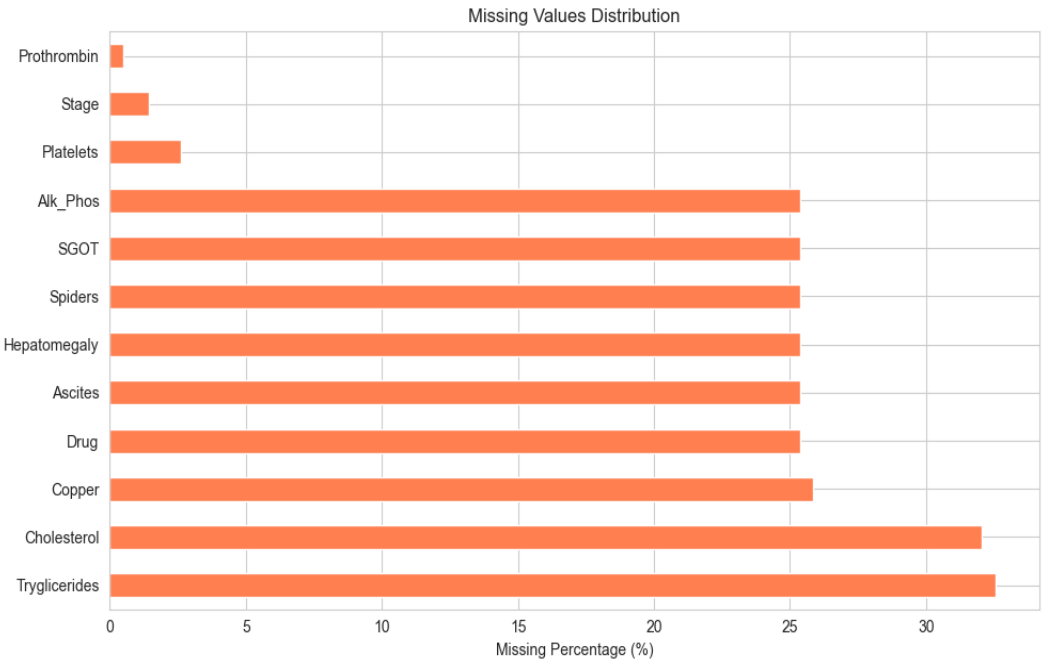
Metric	Value
Total Records	418
Total Features	20
Numeric Features	13
Categorical Features	7
Duplicate Rows	0
Memory Usage (MB)	0.20

3. MISSING VALUES ANALYSIS

Column	Missing Count	Percentage (%)
Tryglicerides	136	32.54
Cholesterol	134	32.06
Copper	108	25.84

Drug	106	25.36
Ascites	106	25.36
Hepatomegaly	106	25.36
Spiders	106	25.36
SGOT	106	25.36
Alk_Phos	106	25.36
Platelets	11	2.63
Stage	6	1.44
Prothrombin	2	0.48

Figure 1: Missing Values Distribution



4. NUMERIC FEATURES STATISTICS

Feature	Mean	Std	Min	Max	Skewness
ID	209.50	120.81	1.00	418.00	0.00
N_Days	1917.78	1104.67	41.00	4795.00	0.47
Age	18533.35	3815.85	9598.00	28650.00	0.09
Bilirubin	3.22	4.41	0.30	28.00	2.72
Cholesterol	369.51	231.94	120.00	1775.00	3.41
Albumin	3.50	0.42	1.96	4.64	-0.47
Copper	97.65	85.61	4.00	588.00	2.30

Alk_Phos	1982.66	2140.39	289.00	13862.40	2.99
SGOT	122.56	56.70	26.35	457.25	1.45
Tryglicerides	124.70	65.15	33.00	598.00	2.52
Platelets	257.02	98.33	62.00	721.00	0.63
Prothrombin	10.73	1.02	9.00	18.00	2.22
Stage	3.02	0.88	1.00	4.00	-0.50

4.1 Numeric Distributions Visualization

Status:

Value	Count	Percentage (%)

C	232	55.50
D	161	38.52
CL	25	5.98

Drug:

Value	Count	Percentage (%)
D-penicillamine	158	37.80
Placebo	154	36.84

Sex:

Value	Count	Percentage (%)
F	374	89.47
M	44	10.53

Ascites:

Value	Count	Percentage (%)
N	288	68.90
Y	24	5.74

Hepatomegaly:

Value	Count	Percentage (%)
Y	160	38.28
N	152	36.36

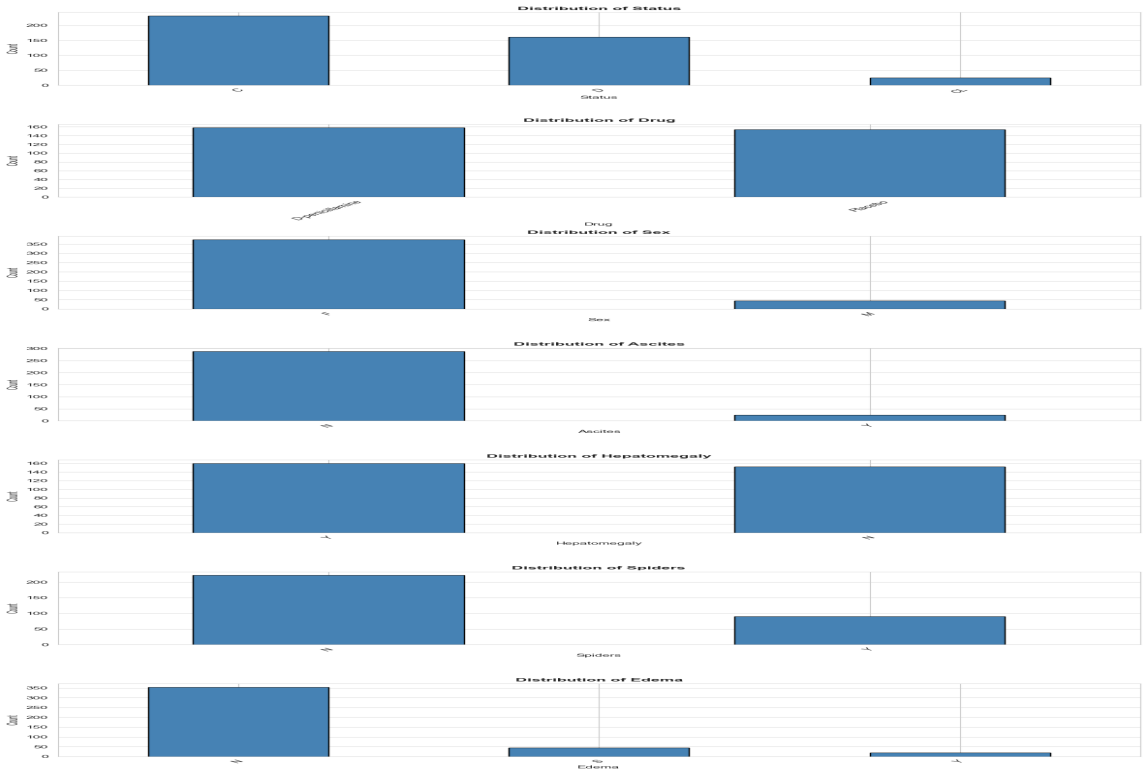
Spiders:

Value	Count	Percentage (%)
N	222	53.11
Y	90	21.53

Edema:

Value	Count	Percentage (%)
N	354	84.69
S	44	10.53
Y	20	4.78

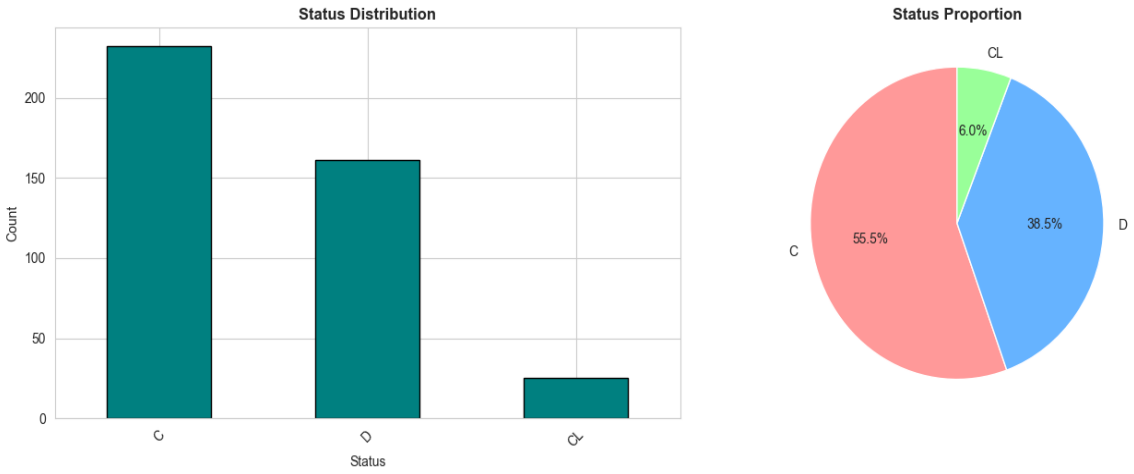
5.1 Categorical Distributions Visualization



6. TARGET VARIABLE ANALYSIS (Status)

Status	Count	Percentage (%)
C	232	55.50
D	161	38.52
CL	25	5.98

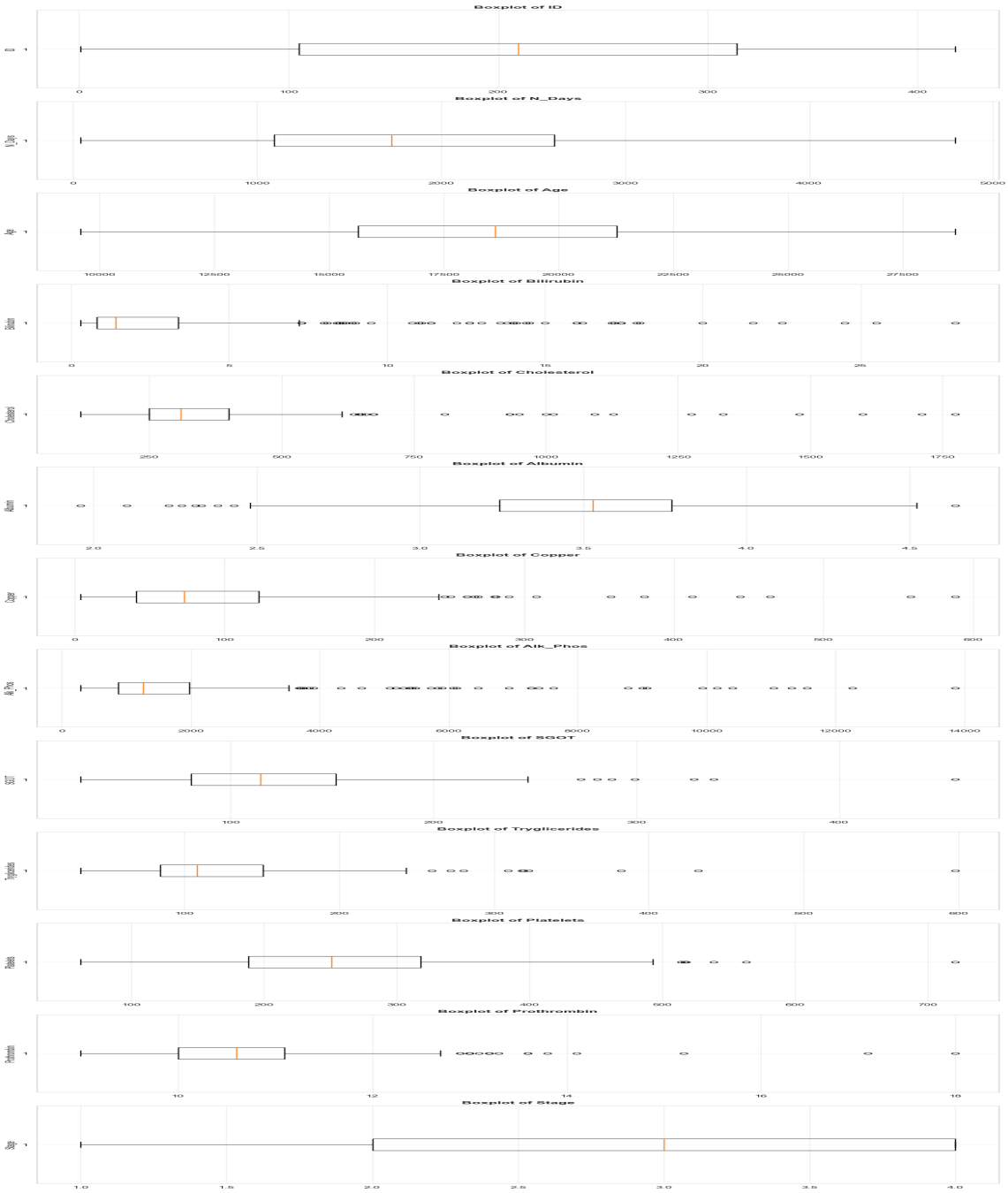
Figure 2: Target Distribution



7. OUTLIER DETECTION (IQR Method)

Feature	Outlier Count	Percentage (%)
Bilirubin	46	11.00
Cholesterol	20	4.78
Albumin	9	2.15
Copper	17	4.07
Alk_Phos	35	8.37
SGOT	7	1.67
Tryglicerides	10	2.39
Platelets	6	1.44
Prothrombin	18	4.31

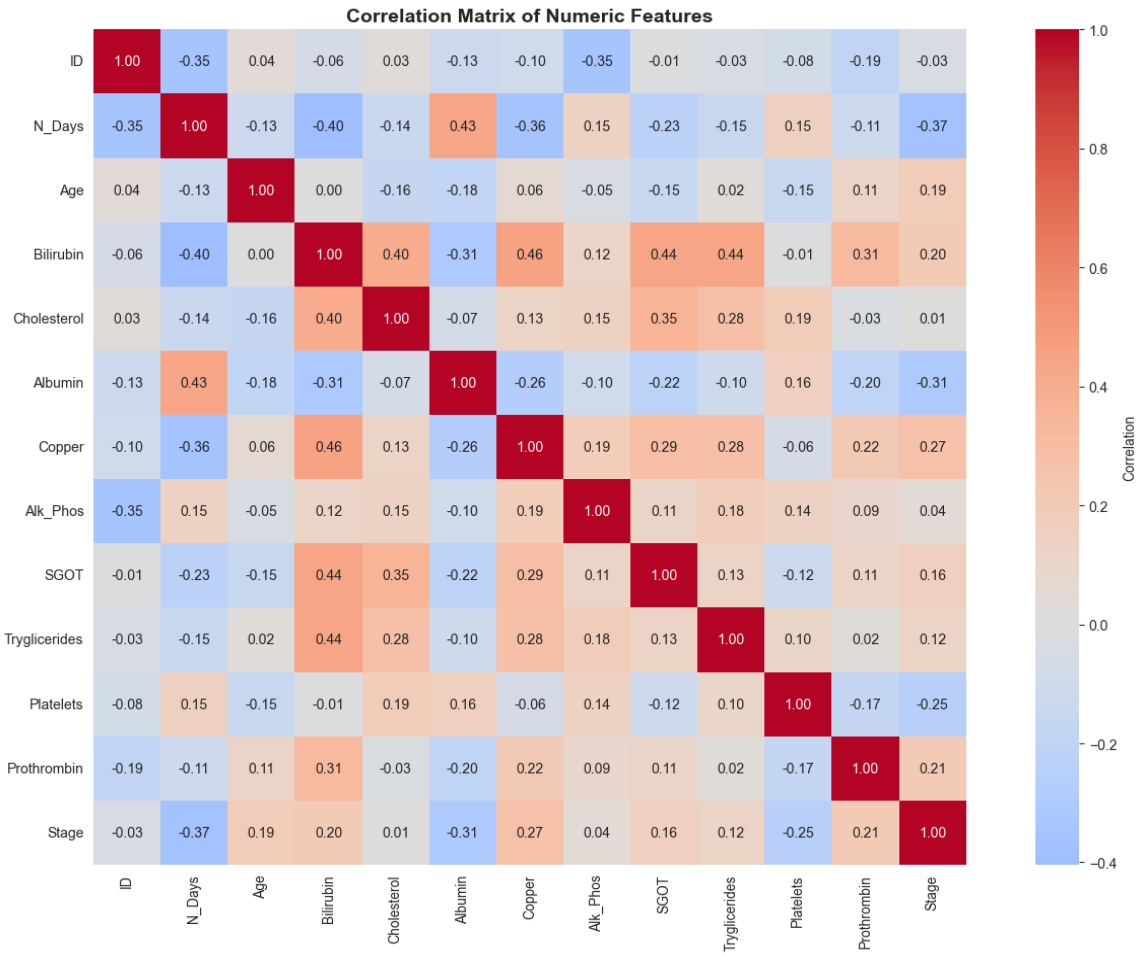
Figure 3: Boxplots of Numeric Features


Boxplot of ID
ID
0
100
200
300
400
Boxplot of N_Days
N_Days
0
1000
2000
3000
4000
5000
Boxplot of Age
Age
10000
12500
15000
17500
20000
22500
25000
27500
Boxplot of Bilirubin
Bilirubin
0
5
10
15
20
25
Boxplot of Cholesterol
Cholesterol
250
500
750
1000
1250
1500
1750
Boxplot of Albumin
Albumin
2.0
2.5
3.0
3.5
4.0
4.5
Boxplot of Copper
Copper
0
100
200
300
400
500
600
Boxplot of Alk_Phos
Alk_Phos
0
2000
4000
6000
8000
10000
12000
14000
Boxplot of SGOT
SGOT
100
200
300
400
Boxplot of Tryglicerides
Tryglicerides
100
200
300
400
500
600
Boxplot of Platelets
Platelets
100
200
300
400
500
600
700
Boxplot of Prothrombin
Prothrombin
10
12
14
16
18
Boxplot of Stage
Stage
1.0
1.5
2.0
2.5
3.0
3.5
4.0

8. CORRELATION ANALYSIS

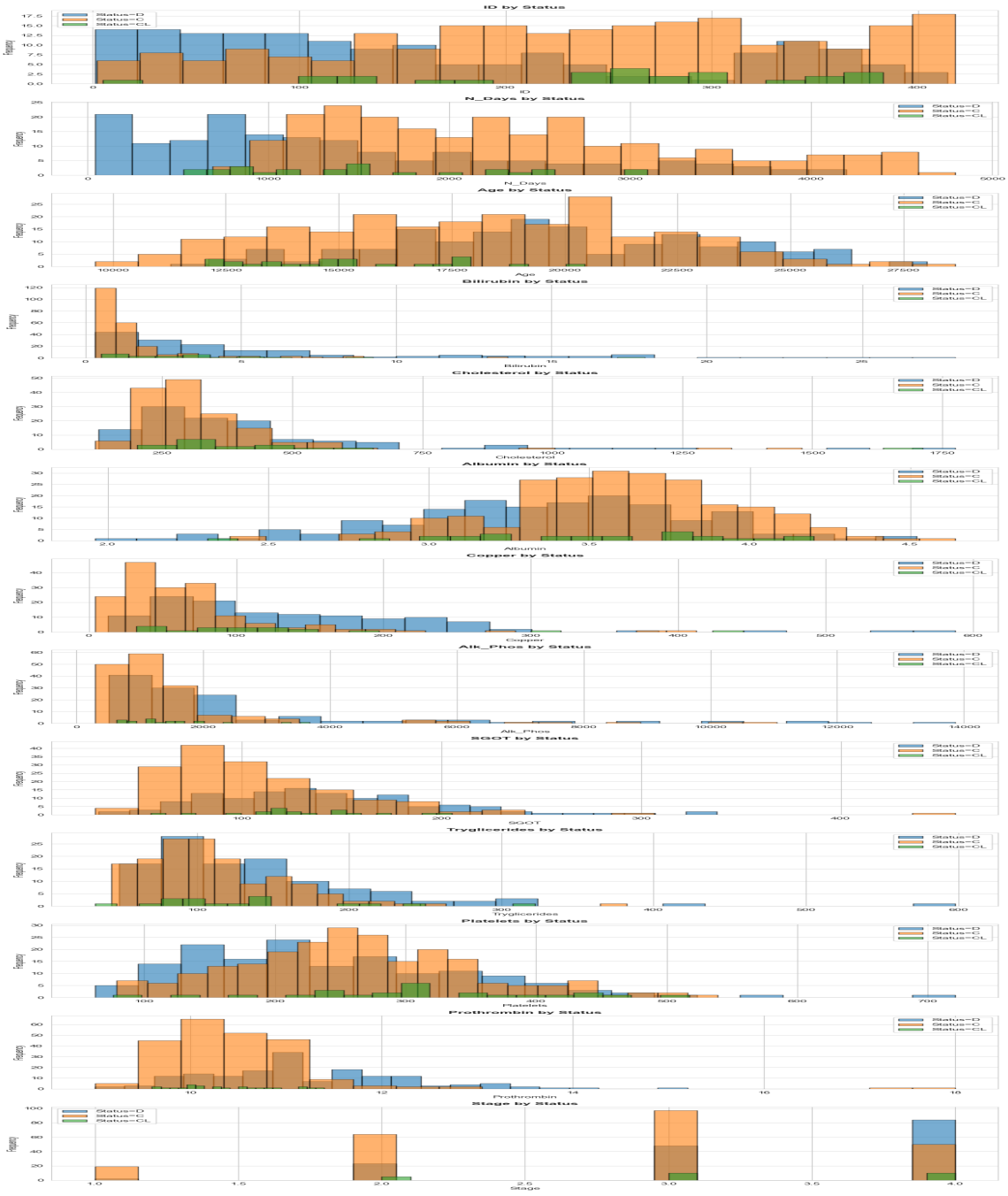
No strong correlations (|r| > 0.7) detected.

Figure 4: Correlation Matrix Heatmap

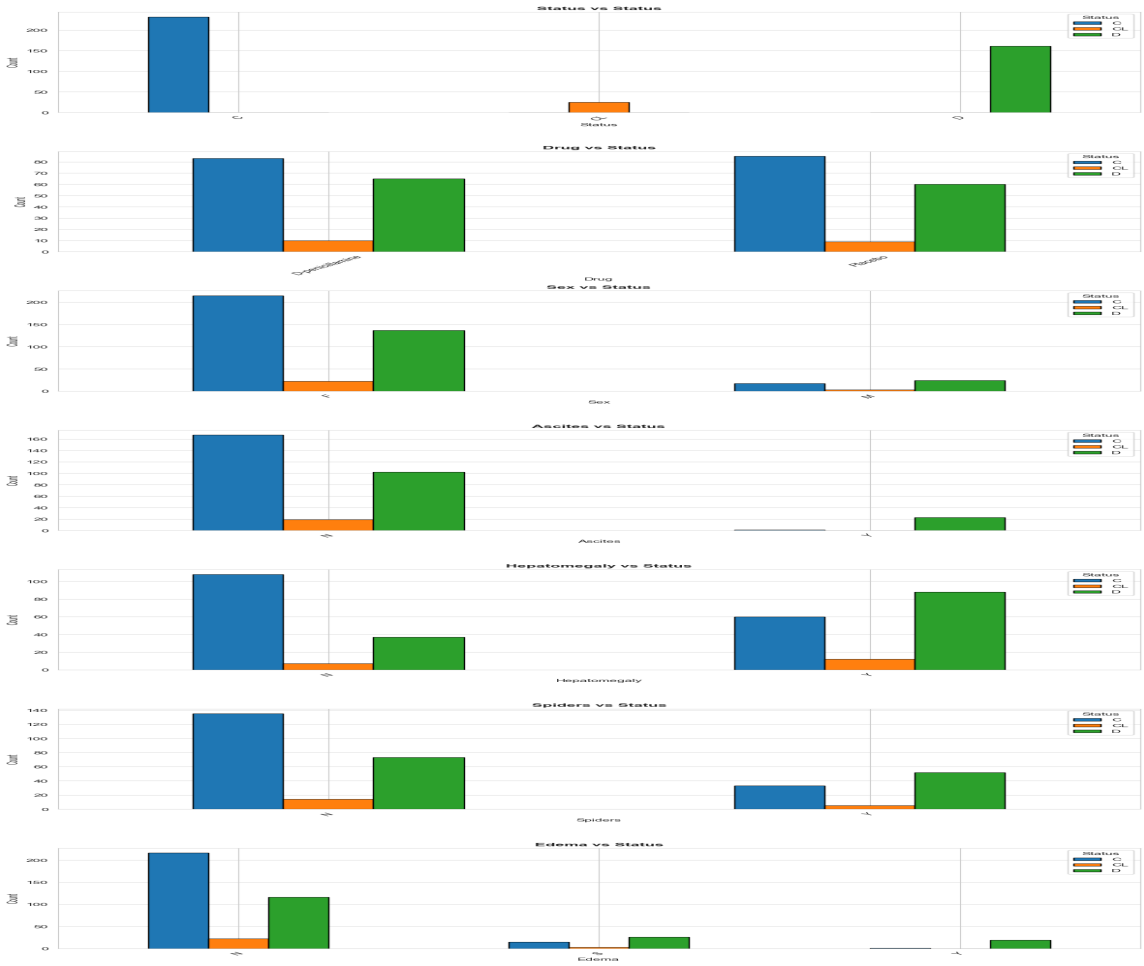


9. FEATURE RELATIONSHIPS WITH TARGET

9.1 Numeric Features vs Target



9.2 Categorical Features vs Target



10. DATA QUALITY SUMMARY

Metric	Value
Total Records	418
Total Features	20
Complete Rows (No Missing)	276
Rows with Missing Values	142
Duplicate Rows	0
Overall Completeness (%)	87.64

11. KEY FINDINGS AND RECOMMENDATIONS

Key Findings:
• The dataset contains **418** patient records with comprehensive clinical information
• **1033** missing values detected in 12 features
• Target variable (Status) has **3** classes with distribution: C: 232, D: 161, CL: 25
• Outliers detected in numeric features using IQR method

Recommendations:
• Handle missing values through imputation or removal based on analysis
• Consider feature scaling for machine learning models
• Investigate and handle outliers appropriately
• Use highly correlated features carefully to avoid multicollinearity
• Consider feature engineering for improved model performance

Report generated on February 09, 2026 at 03:31 PM
This report is based on exploratory data analysis of the Cirrhosis dataset.