Report on mimic-iii Data Set

Team SIX Member:

- Ananta Arora (SID: 100421624)
- Jinghao Chen (SID: 100406201)
- Roxanne Alvarez (SID: 100405742)
- Teshani Jayasinghe (SID: 100422405)

Contents

Summary

Our team conducted an exploratory data analysis on the Medical Information Mart for Intensive Care (MIMIC-III) dataset. MIMIC-III is a comprehensive health-related dataset that primarily focuses on patients admitted to the Intensive Care Unit at the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts, USA.

In this content, we will explore the dataset on mechanically ventilated ICU patients, examine each variable, and apply appropriate statistical tests to identify associations and differences between patient survival outcomes.

Our dataset consists of 18,883 observations and 70 variables. We dropped the variables 'RRT' and 'ventilation duration' because these variables are not obtained on the first day of ICU admission. Based on our observations, missing values predominantly occur in the vital signs and laboratory results groups. The team has decided to drop observations with at least one missing value row-wise, resulting in 12,799 remaining rows. Additionally, we removed 310 rows containing at least one value that falls outside the valid range. Consequently, the resulting cleaned dataset comprises 12,489 rows and 68 variables.

```
Vital signs
   missing_percentage  count
4               100.0     41
3                60.0      1
2                40.0     18
1                20.0    647
0                 0.0  18176
None
----------------------------

Laboratory results
   missing_percentage  count
7              100.00     50
6               85.71      9
5               71.43      6
4               57.14     15
3               42.86     28
2               28.57    968
1               14.29   4345
0                0.00  13462
None
```

We conducted a Chi-Square test for the categorical variables, and Mann-Whitney U test and Spearman correlation for the ordinal variables.

Regarding the continuous variables, we first conducted a Shapiro-Wilk test to check for normality, and the result was "significantly different from normal" for all variables. To confirm the result of the Shapiro-Wilk test, we created histograms and QQ plots for

visual inspection. As part of the analysis, we performed Levene's Test to assess the equality of variances between the two groups, survivors and non-survivors, and discovered that all continuous variables violated the assumption of equal variances.

Although some distributions appeared to be roughly normal, the presence of outliers caused the distribution shapes to be skewed. Therefore, we conducted a log transformation on the continuous variables and found that some of the variables became normally distributed after the transformation. For these variables, we performed a Welch T-Test using the normalized data. For the remaining continuous variables where the distribution remained non-normal, we conducted a non-parametric test called the Mann-Whitney U test using the values before transformation.

Additionally, for the variables that we were able to normalize the distribution through log transformation, we conducted a comparative analysis between Welch T-test and Mann-Whitney U test results.

Spearman correlation was also conducted for the continuous variables using the dataset before transformation. We present the pairs of variables exhibiting a moderate to strong correlation, based on the correlation coefficient.

For visualizations, we used bar graphs for categorical variables, density plots and scatter plot for ordinal variables, and box plots, scatter plot and density plots for continuous variables.

It is important to note that we attempted to normalize the continuous variables using log transformation, L2 normalization, BoxCox method, and min-max method. Although log transformation performed best, the distributions did not become normal for all the variables.

Finally, we created a table to illustrate the summary of all variables, data types, tests conducted, and the resulting p-value.

Data Exploration

General characteristics of the dataset

How many records do we have? How many variables?

```
# number of records and variables
df.shape
```

There are 18,883 records and 70 variables in the dataset.

```
# variable names
df.columns
```

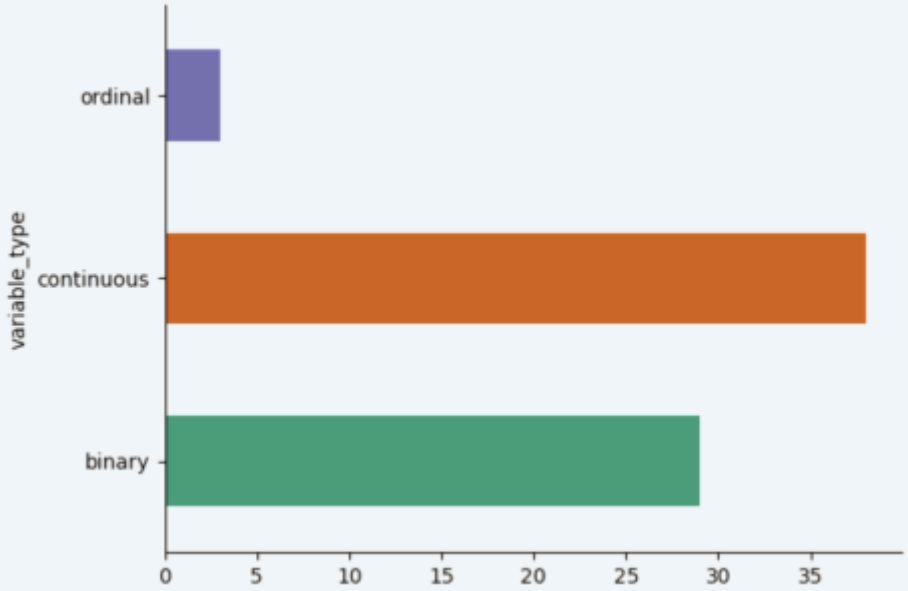
What are the variable names? Are they meaningful?

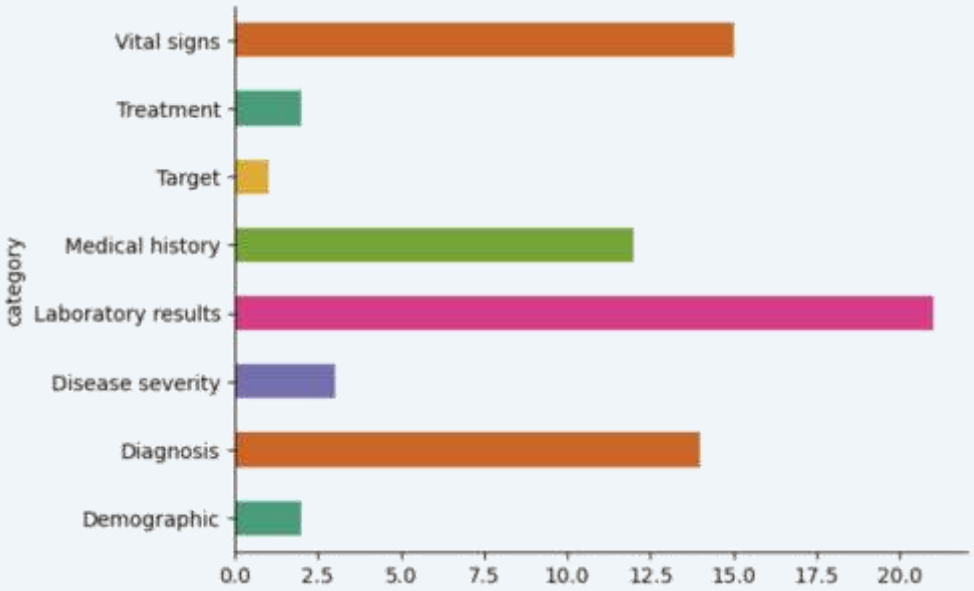
Below is the list of variable names in our dataset. We will learn later that by conducting the statistical tests, we are able to determine which variables give us meaningful insight into predicting the target variable.

```
Variables names:
Index(['Hospital Mortality', 'Age', 'Gender', 'Uncomplicated Hypertension',
       'Complicated Hypertension', 'Uncomplicated Diabetes',
       'Complicated Diabetes', 'Malignancy', 'Hematologic Disease',
       'Metastasis', 'Peripheral Vascular Disease', 'Hypothyroidism',
       'Chronic Heart Failure', 'Stroke', 'Liver Disease', 'SAPS II', 'SOFA',
       'OASIS', 'Sepsis', 'Any Organ Failure', 'Severe Respiratory Failure',
       'Severe Coagulation Failure', 'Severe Liver Failure',
       'Severe Cardiovascular Failure',
       'Severe Central Nervous System Failure', 'Severe Renal Failure',
       'Respiratory Dysfunction', 'Cardiovascular Dysfunction',
       'Renal Dysfunction', 'Hematologic Dysfunction', 'Metabolic Dysfunction',
       'Neurologic Dysfunction', 'Max Heart Rate', 'Min Heart Rate',
       'Mean Heart Rate', 'Max MAP', 'Min MAP', 'Mean MAP',
       'Max Systolic Pressure', 'Min Systolic Pressure',
       'Mean Systolic Pressure', 'Max Diastolic Pressure',
       'Min Diastolic Pressure', 'Mean Diastolic Pressure', 'Max Temperature',
       'Min Temperature', 'Mean Temperature', 'Max Lactate', 'Min Lactate',
       'Mean Lactate', 'Max pH', 'Min pH', 'Mean pH', 'Max Glucose',
       'Min Glucose', 'Mean Glucose', 'Max WBC', 'Min WBC', 'Mean WBC',
       'Max BUN', 'Min BUN', 'Mean BUN', 'Max Creatinine', 'Min Creatinine',
       'Mean Creatinine', 'Max Hemoglobin', 'Min Hemoglobin',
       'Mean Hemoglobin', 'Ventilation Duration (h)', 'RRT'],
      dtype='object')
```

What type is each variable?

Category	Count	Data Type
Demographic	2	1 continuous 1 binary
Diagnosis	14	14 binary
Disease severity	3	3 ordinal
Laboratory results	21	21 continuous
Medical history	12	12 binary
Vital signs	15	15 continuous
Hospital Mortality	1	1 binary
Total	1 target variable 67 predictors	27 binary 37 continuous 3 ordinal





How many unique values does each variable have?

```
# number of unique values per variable
df.nunique()
```

Hospital Mortality	2
Age	72
Gender	2
Uncomplicated Hypertension	2
Complicated Hypertension	2
Uncomplicated Diabetes	2
Complicated Diabetes	2
Malignancy	2
Hematologic Disease	2
Metastasis	2
Peripheral Vascular Disease	2
Hypothyroidism	2
Chronic Heart Failure	2
Stroke	2
Liver Disease	2
SAPS II	107
SOFA	23
OASIS	64
Sepsis	2
Any Organ Failure	2
Severe Respiratory Failure	2
Severe Coagulation Failure	2
Severe Liver Failure	2
Severe Cardiovascular Failure	2
Severe Central Nervous System Failure	2
Severe Renal Failure	2
Respiratory Dysfunction	2
Cardiovascular Dysfunction	2
Renal Dysfunction	2
Hematologic Dysfunction	2
Metabolic Dysfunction	2
Neurologic Dysfunction	2
Max Heart Rate	162
Min Heart Rate	132
Mean Heart Rate	12730
Max MAP	426
Min MAP	258
Mean MAP	14561
Max Systolic Pressure	205
Min Systolic Pressure	167
Mean Systolic Pressure	13300

Max Diastolic Pressure	179
Min Diastolic Pressure	92
Mean Diastolic Pressure	11778
Max Temperature	334
Min Temperature	369
Mean Temperature	11507
Max Lactate	239
Min Lactate	182
Mean Lactate	946
Max pH	89
Min pH	95
Mean pH	78
Max Glucose	573
Min Glucose	332
Mean Glucose	4984
Max WBC	553
Min WBC	445
Mean WBC	1974
Max BUN	167
Min BUN	151
Mean BUN	1077
Max Creatinine	144
Min Creatinine	123
Mean Creatinine	606
Max Hemoglobin	143
Min Hemoglobin	159
Mean Hemoglobin	934
Ventilation Duration (h)	5895
RRT	2

What value occurs most frequently, and how often does it occur?

```
# most frequently occuring value and the count
most_frequent_values = {}
for column in df.columns:
    most_common = df[column].value_counts().idxmax()
    count = df[column].value_counts().max()
    most_frequent_values[column] = {'value': most_common,
'count': count}

# DataFrame from the dictionary
result_df = pd.DataFrame(most_frequent_values).T

print(result_df)
```

	value	count
Hospital Mortality	0	15866
Age	77	504
Gender	M	11457
Uncomplicated Hypertension	0	10123
Complicated Hypertension	0	17418
Uncomplicated Diabetes	0	14980
Complicated Diabetes	0	17924
Malignancy	0	16871
Hematologic Disease	0	16070
Metastasis	0	18055
Peripheral Vascular Disease	0	17244
Hypothyroidism	0	17307
Chronic Heart Failure	0	14348
Stroke	0	17825
Liver Disease	0	17085
SAPS II	34	692
SOFA	4	2872
OASIS	35	972
Sepsis	0	16063
Any Organ Failure	1	9574
Severe Respiratory Failure	0	17663
Severe Coagulation Failure	0	18783
Severe Liver Failure	0	18663
Severe Cardiovascular Failure	0	16581
Severe Central Nervous System Failure	0	17784
Severe Renal Failure	0	17965
Respiratory Dysfunction	0	14003
Cardiovascular Dysfunction	0	16329
Renal Dysfunction	0	14244

Hematologic Dysfunction	0	16858
Metabolic Dysfunction	0	17010
Neurologic Dysfunction	0	17190
Max Heart Rate	88	534
Min Heart Rate	70	647
Mean Heart Rate	87	26
Max MAP	93	484
Min MAP	58	773
Mean MAP	74	26
Max Systolic Pressure	150	396
Min Systolic Pressure	85	601
Mean Systolic Pressure	108	26
Max Diastolic Pressure	80	572
Min Diastolic Pressure	45	863
Mean Diastolic Pressure	60	37
Max Temperature	37.5	645
Min Temperature	36.11111	607
Mean Temperature	36.94444	51
Max Lactate	2	461
Min Lactate	1	976
Mean Lactate	1.4	366
Max pH	7.44	1224
Min pH	7.32	963
Mean pH	7.38	1421
Max Glucose	165	177
Min Glucose	99	343
Mean Glucose	130	76
Max WBC	12.3	162
Min WBC	10.2	193
Mean WBC	9.7	94
Max BUN	15	1041
Min BUN	13	1128
Mean BUN	14	544
Max Creatinine	0.8	2237
Min Creatinine	0.7	2678
Mean Creatinine	0.8	1211
Max Hemoglobin	12.7	403
Min Hemoglobin	9.4	369
Mean Hemoglobin	9.7	138
Ventilation Duration (h)	4	354
RRT	0	18328

Are there missing observations (vertically and horizontally)? If so,

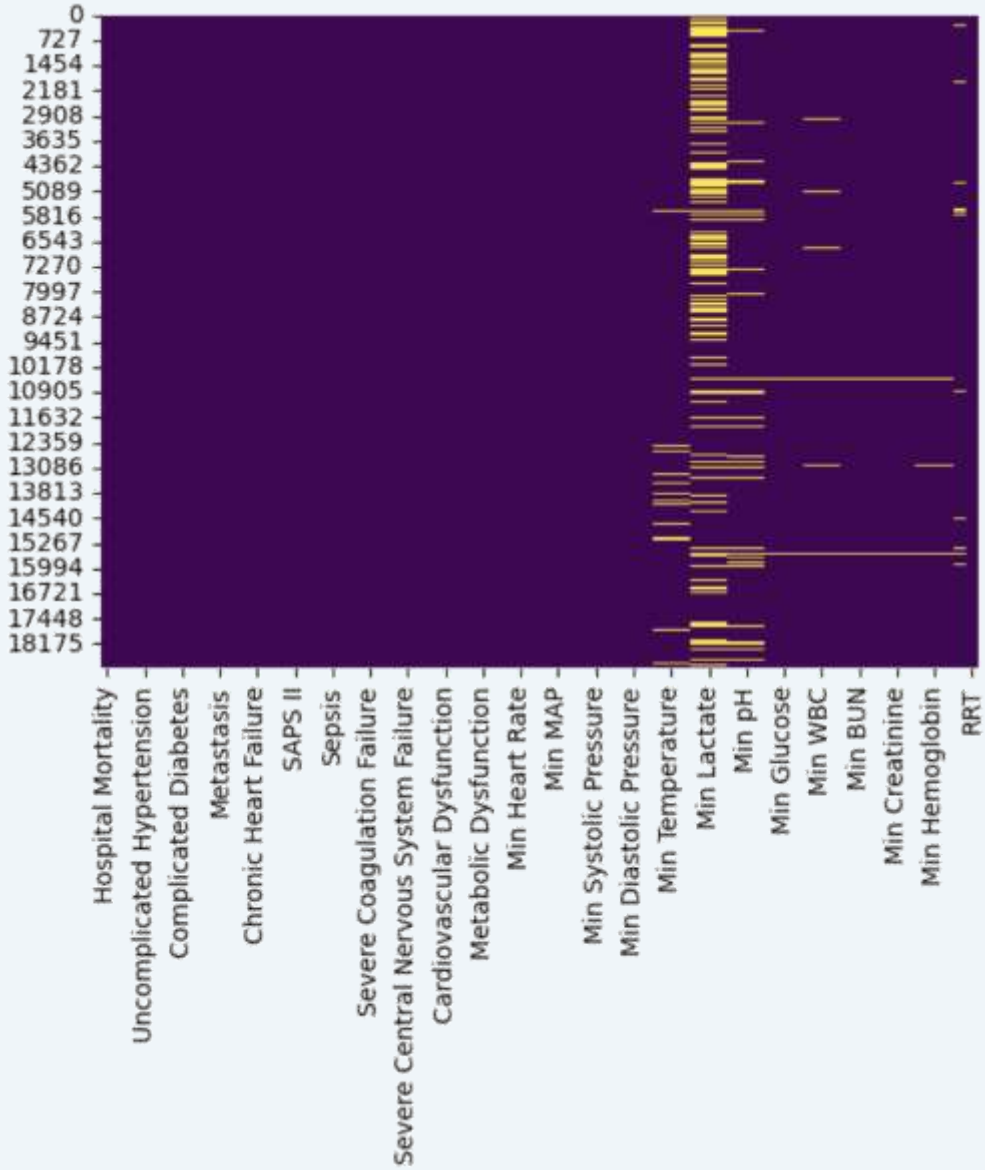
```
# missing values column-wise
na_count = df.isnull().sum() # total count
na_pct = (na_count/len(df))*100 # percentage

na_df = pd.DataFrame({'Count': na_count.values,
                      'Percentage (%)':
na_pct}).reset_index().rename(columns = {'index': 'Feature'})

na_df
```

how frequently does this occur?

The result below shows which variables have missing values and their corresponding count. We have also produced a heatmap to determine the pattern of missing values for all variables.



Feature	Count	Percentage (%) Missing
Max Heart Rate	41	0.22
Min Heart Rate	41	0.22
Mean Heart Rate	41	0.22
Max MAP	42	0.22
Min MAP	42	0.22
Mean MAP	42	0.22
Max Systolic Pressure	60	0.32
Min Systolic Pressure	60	0.32
Mean Systolic Pressure	60	0.32
Max Diastolic Pressure	61	0.32
Min Diastolic Pressure	61	0.32
Mean Diastolic Pressure	61	0.32
Max Temperature	687	3.64
Min Temperature	687	3.64
Mean Temperature	687	3.64
Max Lactate	5101	27.01
Min Lactate	5101	27.01
Mean Lactate	5101	27.01
Max pH	1173	6.21
Min pH	1173	6.21
Mean pH	1173	6.21
Max Glucose	75	0.40
Min Glucose	75	0.40
Mean Glucose	75	0.40
Max WBC	229	1.21
Min WBC	229	1.21
Mean WBC	229	1.21
Max BUN	95	0.50
Min BUN	95	0.50
Mean BUN	95	0.50
Max Creatinine	95	0.50
Min Creatinine	95	0.50
Mean Creatinine	95	0.50
Max Hemoglobin	91	0.48

Min Hemoglobin	91	0.48
Mean Hemoglobin	91	0.48
Ventilation Duration (h)	497	2.63

Vertical
The above table shows that Min Lactate, Max Lactate and Mean Lactate columns are missing 27% of their data.

Horizontal
The following table reveals that missing values predominantly occur in the vital signs and laboratory results groups. Notably, we found that 41 patients have completely missing vital signs data, and 50 patients have entirely missing laboratory results. We decided to remove 6,084 rows which have at least one missing value across all features.

Additionally, we found a literature[1] from the National Library of Medicine which contains the information about valid ranges for vital signs and laboratory results. Based on this, we have removed additional 310 observations where at least one of its values falls outside of the valid range.

[1] M A Papadakis et al., "Prognosis of Mechanically Ventilated Patients," *The Western Journal of Medicine* 159, no. 6 (December 1993): 659–64, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1022451/.

```
Demographic
   missing_percentage  count
0                 0.0  18883
None
----------------------------

Medical history
   missing_percentage  count
0                 0.0  18883
None
----------------------------

Disease severity
   missing_percentage  count
0                 0.0  18883
None
----------------------------

Diagnosis
   missing_percentage  count
0                 0.0  18883
None
----------------------------

Vital signs
   missing_percentage  count
4               100.0     41
3                60.0      1
2                40.0     18
1                20.0    647
0                 0.0  18176
None
----------------------------

Laboratory results
   missing_percentage  count
7              100.00     50
6               85.71      9
5               71.43      6
4               57.14     15
3               42.86     28
2               28.57    968
1               14.29   4345
0                0.00  13462
None
```

The distribution of the values of each variable employed in our models. All variables are present in every dataset.

Clinical Variables	Valid Range		MIMIC-III		MIMIC-IV		eICU	
	Lower	Upper	Survival 1	Death 2	Survival 1	Death 2	Survival 1	Death 2
Vital sign variables (7 variables)								
Heart rate (beats/min)	0	350	84.3 (16.1)	90.4 (19.8)	84.4 (16.7)	91.3 (19.8)	84.2 (17.1)	93.1 (20.8)
Diastolic blood pressure (mmHg)	0	375	61.4 (14.2)	56.8 (14.4)	63.3 (14.6)	57.5 (14.3)	67.9 (14.4)	60.9 (15.1)
Systolic blood pressure (mmHg)	0	375	124.5 (21.4)	116.4 (25.6)	122.6 (21.0)	113.5 (24.5)	126.1 (22.1)	115.4 (24.7)
Mean arterial pressure (mmHg)	14	330	80.5 (15.3)	75.4 (17.1)	79.3 (15.3)	73.8 (16.7)	83.9 (16.1)	76.0 (16.9)
Temperature (°C)	26	45	36.9 (0.6)	36.9 (1.0)	36.9 (0.5)	36.9 (0.8)	36.8 (0.5)	36.8 (1.1)
Peripheral oxygen saturation (%)	0	100	96.8 (2.6)	95.6 (6.4)	96.3 (2.6)	95.3 (6.1)	96.3 (3.2)	95.2 (7.2)
Respiratory rate (breaths/min)	0	300	19.9 (5.4)	21.4 (6.7)	19.7 (5.4)	21.4 (6.4)	19.5 (5.2)	21.9 (7.2)
Laboratory variables (16 variables)								
Albumin (g/dL)	0.6	6	2.6 (0.5)	2.5 (0.6)	2.8 (0.5)	2.7 (0.6)	2.6 (0.6)	2.4 (0.6)
Blood urea nitrogen (mg/dL)	0	250	36.9 (26.5)	50.4 (31.8)	38.0 (26.8)	44.6 (32.4)	26.3 (21.2)	39.8 (28.2)
Bilirubin (mg/dL)	0.1	60	3.6 (5.3)	10.5 (13.1)	3.9 (6.4)	7.9 (10.1)	1.4 (3.2)	3.2 (6.1)
Lactate (mmol/L)	0.4	30	1.9 (1.5)	3.8 (4.2)	1.9 (1.4)	3.5 (3.4)	2.1 (2.0)	5.5 (5.0)
Bicarbonate (mEq/L)	0	60	25.8 (4.7)	23.9 (5.7)	25.8 (5.3)	23.0 (5.7)	25.6 (4.9)	22.3 (6.0)
Band neutrophil (%)	0	100	5.2 (6.3)	6.4 (7.2)	4.6 (5.7)	5.2 (5.6)	8.6 (12.6)	12.2 (13.1)
Chloride (mEq/L)	50	175	105.3 (6.1)	103.9 (7.3)	103.6 (7.2)	103.3 (7.9)	104.3 (6.7)	106.8 (9.2)
Creatinine (mg/dL)	0.1	60	1.4 (1.4)	1.8 (1.3)	1.6 (1.5)	1.8 (1.3)	1.3 (1.4)	2.0 (1.6)
Glucose (mg/dL)	33	2000	131.6 (52.2)	136.9 (64.3)	140.6 (62.3)	147.8 (67.4)	144.2 (57.2)	149.9 (62.9)
Hemoglobin (g/dL)	0	25	9.6 (1.4)	9.6 (1.4)	9.1 (1.6)	8.9 (1.4)	10.2 (2.0)	9.8 (2.1)
Hematocrit (g/dL)	0	75	28.6 (3.9)	28.6 (3.9)	27.8 (4.6)	27.2 (4.3)	31.1 (6.1)	29.8 (6.5)
Platelet count (1000/mm ^{3})	0	2000	278.9 (192.2)	162.5 (138.9)	237.0 (174.2)	154.3 (134.8)	208.3 (115.7)	153.4 (105.5)
Potassium (mEq/L)	0	12	4.0 (0.5)	4.1 (0.6)	4.0 (0.5)	4.1 (0.6)	3.9 (0.5)	4.2 (0.8)
Partial thromboplastin time (s)	18.8	150	44.7 (23.7)	52.9 (28.0)	47.7 (24.4)	52.9 (26.7)	50.7 (27.7)	52.2 (27.9)
Sodium (mEq/L)	50	225	140.3 (5.2)	139.1 (6.0)	140.4 (5.9)	139.5 (6.7)	138.9 (5.7)	141.7 (8.1)
White blood cells (1000/mm ^{3})	0	1000	12.9 (8.5)	14.7 (9.8)	12.9 (8.9)	15.7 (13.7)	11.2 (6.0)	15.8 (9.7)

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8465577/

```
df = df.loc[
    ((df['Max Heart Rate'] >= 0) & (df['Max Heart Rate'] <= 350))
    & ((df['Min Heart Rate'] >= 0) & (df['Max Heart Rate'] <= 350))
    & ((df['Mean Heart Rate'] >= 0) & (df['Mean Heart Rate'] <= 350))
    & ((df['Max MAP']  >= 14) & (df['Max MAP'] <= 330))
    & ((df['Min MAP']  >= 14) & (df['Min MAP'] <= 330))
    & ((df['Mean MAP']  >= 14) & (df['Mean MAP'] <= 330))
    & ((df['Min Systolic Pressure'] >= 0) & (df['Min Systolic Pressure'] <= 375))
    & ((df['Max Systolic Pressure'] >= 0) & (df['Max Systolic Pressure'] <= 375))
    & ((df['Mean Systolic Pressure'] >= 0) & (df['Mean Systolic Pressure'] <= 375))
    & ((df['Min Diastolic Pressure'] >= 0) & (df['Min Diastolic Pressure'] <= 375))
    & ((df['Max Diastolic Pressure'] >= 0) & (df['Max Diastolic Pressure'] <= 375))
    & ((df['Mean Diastolic Pressure'] >= 0) & (df['Mean Diastolic Pressure'] <= 375))
    & ((df['Min Temperature'] >= 26)& (df['Min Temperature'] <= 45))
    & ((df['Max Temperature'] >= 26)& (df['Max Temperature'] <= 45))
    & ((df['Mean Temperature'] >= 26)& (df['Mean Temperature'] <= 45))
    & ((df['Min pH']  >= 0)& (df['Min pH'] <= 14))
    & ((df['Max pH']  >= 0)& (df['Max pH'] <= 14))
    & ((df['Mean pH']  >= 0)& (df['Mean pH'] <= 14))
    & ((df['Min Lactate']  >= 0.4)& (df['Min Lactate'] <= 30))
    & ((df['Max Lactate']  >= 0.4)& (df['Max Lactate'] <= 30))
    & ((df['Mean Lactate']  >= 0.4)& (df['Mean Lactate'] <= 30))
    & ((df['Min Glucose'] >= 33)& (df['Min Glucose'] <= 2000))
    & ((df['Max Glucose'] >= 33)& (df['Max Glucose'] <= 2000))
    & ((df['Mean Glucose'] >= 33)& (df['Mean Glucose'] <= 2000))
    & ((df['Min WBC'] >= 0)& (df['Min WBC'] <= 1000))
    & ((df['Max WBC'] >= 0)& (df['Max WBC'] <= 1000))
    & ((df['Mean WBC'] >= 0)& (df['Mean WBC'] <= 1000))
    & ((df['Min BUN'] >= 0)& (df['Min BUN'] <= 250))
    & ((df['Max BUN'] >= 0)& (df['Max BUN'] <= 250))
    & ((df['Mean BUN'] >= 0)& (df['Mean BUN'] <= 250))
    & ((df['Min Creatinine'] >= 0.1)& (df['Min Creatinine'] <= 60))
    & ((df['Max Creatinine'] >= 0.1)& (df['Max Creatinine'] <= 60))
    & ((df['Mean Creatinine'] >= 0.1)& (df['Mean Creatinine'] <= 60))
    & ((df['Min Hemoglobin'] >= 0) & (df['Min Hemoglobin'] <= 25))
    & ((df['Max Hemoglobin'] >= 0) & (df['Max Hemoglobin'] <= 25))
    & ((df['Mean Hemoglobin'] >= 0) & (df['Mean Hemoglobin'] <= 25))
]
```

Examine descriptive statistics for each variable

Categorical Variables

How many distinct values or "levels" does the variable exhibit

There are ~30 categorical values with two levels/ distinct values: 0 or 1

Variable	0
Hospital Mortality	2
Gender	2
Uncomplicated Hypertension	2
Complicated Hypertension	2
Uncomplicated Diabetes	2
Complicated Diabetes	2
Malignancy	2
Hematologic Disease	2
Metastasis	2
Peripheral Vascular Disease	2
Hypothyroidism	2
Chronic Heart Failure	2
Stroke	2
Liver Disease	2
Sepsis	2
Any Organ Failure	2
Severe Respiratory Failure	2
Severe Coagulation Failure	2
Severe Liver Failure	2
Severe Cardiovascular Failure	2
Severe Central Nervous System Failure	2
Severe Renal Failure	2
Respiratory Dysfunction	2
Cardiovascular Dysfunction	2
Renal Dysfunction	2
Hematologic Dysfunction	2
Metabolic Dysfunction	2
Neurologic Dysfunction	2
RRT	2

How often does each of these levels occur in the dataset?

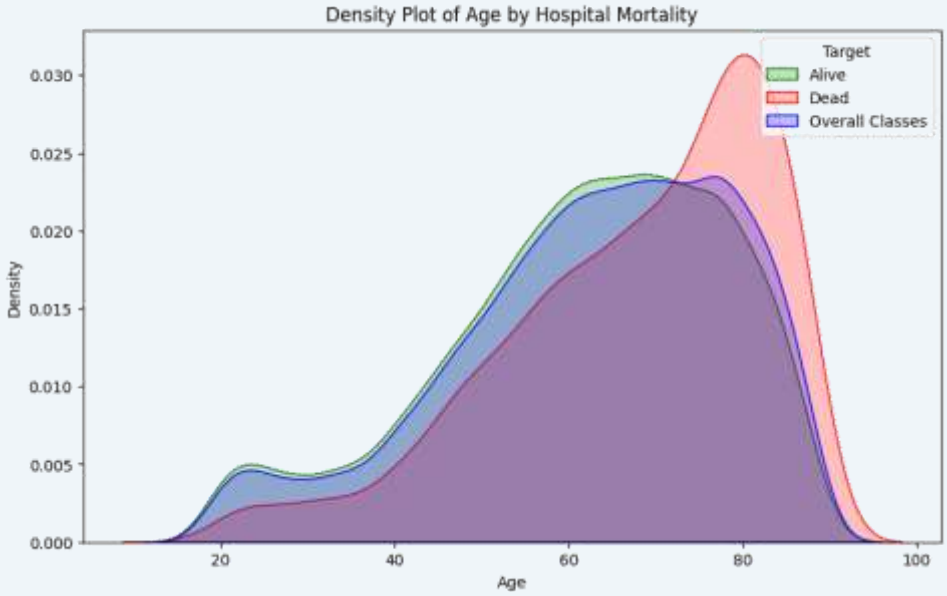
- Most of the categorical features have a higher percentage of class 0 than class 1
- Features like Uncomplicated_Hypertension and Any organ failure have a balance between the classes with ~50% distribution for each class.

	0%	1%
Hospital Mortality	84.02267	15.97733
Gender	0	0
Uncomplicated Hypertension	53.60907	46.39093
Complicated Hypertension	92.2417	7.758301
Uncomplicated Diabetes	79.33061	20.66939
Complicated Diabetes	94.92136	5.078642
Malignancy	89.34491	10.65509
Hematologic Disease	85.103	14.897
Metastasis	95.6151	4.384896
Peripheral Vascular Disease	91.32024	8.679765
Hypothyroidism	91.65387	8.346131
Chronic Heart Failure	75.98369	24.01631
Stroke	94.39708	5.602923
Liver Disease	90.47821	9.521792
Sepsis	85.06593	14.93407
Any Organ Failure	49.29831	50.70169
Severe Respiratory Failure	93.53916	6.460838
Severe Coagulation Failure	99.47042	0.529577
Severe Liver Failure	98.83493	1.165069
Severe Cardiovascular Failure	87.80914	12.19086
Severe Central Nervous System Failure	94.17995	5.82005
Severe Renal Failure	95.13848	4.861516
Respiratory Dysfunction	74.15665	25.84335
Cardiovascular Dysfunction	86.47461	13.52539
Renal Dysfunction	75.43293	24.56707
Hematologic Dysfunction	89.27607	10.72393
Metabolic Dysfunction	90.08103	9.918975
Neurologic Dysfunction	91.03426	8.965736
RRT	97.06085	2.939152

How does the behaviour of another variable, X, vary over *the levels of C?*

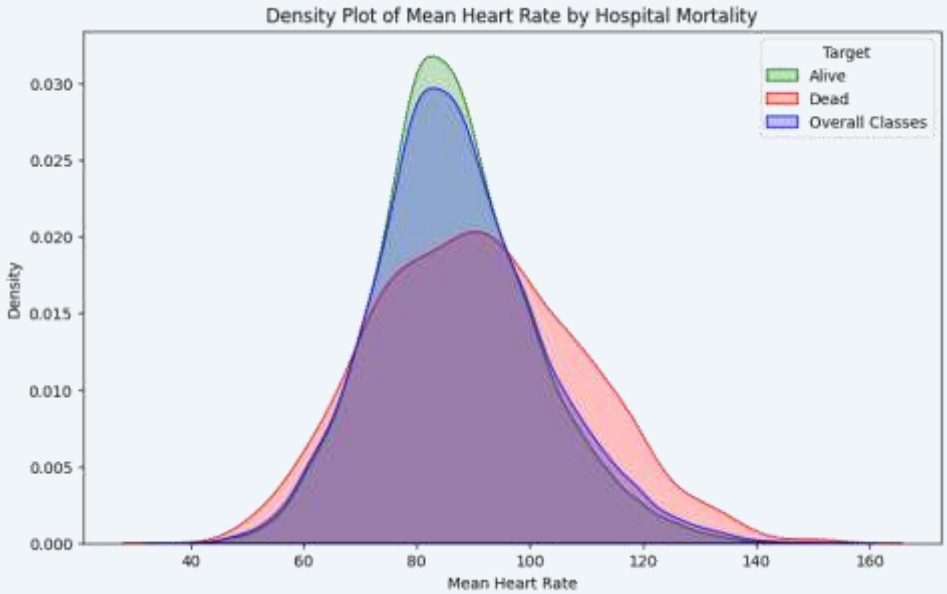
These are the three variables that have significant differences.

Age



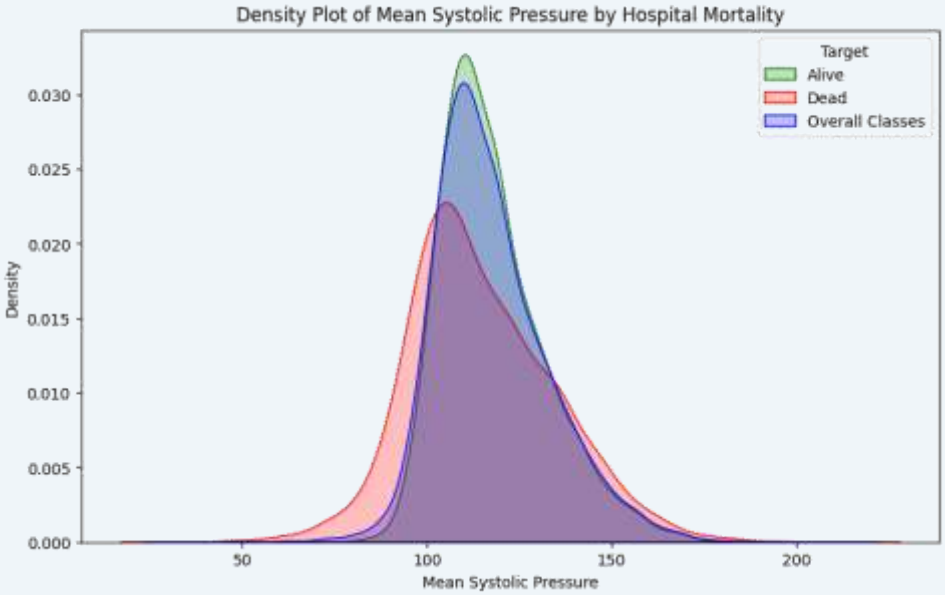
The density plot reveals that the probability of hospital mortality increases with age, particularly after 60. The density drops significantly after the peak. Additionally, we observed a higher density of non-surviving patients after the age of 70. On the other hand, before the age of 60, individuals who are admitted to the hospital and survived have a higher density rate. This could be attributed to the fact that patients before the age of 60 are still physically strong, while those over 60 are often retired and dealing with chronic illnesses, which are more challenging to cure.

Heart Rate



The density plot reveals that most patients who survived had lower heart rates than those without. The heart rate of the surviving patients peaked around 80, while that of those who passed away peaked around 90.

Glucose



According to the density plot of mean systolic pressure, there is a clear association between systolic pressure and mortality. Patients with higher systolic pressure are more likely to survive, and the density between the alive and dead is significant.

Numerical Variable

What is the mean, median, standard deviation?

- The mean and median are similar for most of the variables, which means most of them follow a symmetric distribution.
- The variables like ventilation duration, WBC, Glucose, Systolic Pressure and Diastolic pressure do not seem to have a bell curve, which means they are affected by other factors.

	Mean	Median	Standard Deviation
Max Heart Rate	106.3032	104	20.20465
Min Heart Rate	71.28757	70	15.10189
Mean Heart Rate	87.3072	86	14.95483
Max MAP	108.467	102	28.96581
Min MAP	56.98575	58	13.49352
Mean MAP	78.04275	76.5067	10.28279
Max Systolic Pressure	152.2893	149	24.06168
Min Systolic Pressure	87.81793	87	17.48621
Mean Systolic Pressure	117.3882	114.8611	15.61045
Max Diastolic Pressure	82.91058	80	17.99794
Min Diastolic Pressure	43.40686	44	10.78942
Mean Diastolic Pressure	60.14084	59.13841	9.790386
Max Temperature	37.71966	37.7	0.826505
Min Temperature	36.02669	36.11111	0.971659
Mean Temperature	36.94948	36.95024	0.701227
Max Lactate	3.414678	2.5	2.865615
Min Lactate	1.834647	1.4	1.462907
Mean Lactate	2.569431	2	1.954506
Max pH	7.435479	7.44	0.071786
Min pH	7.297606	7.31	0.104431
Mean pH	7.370085	7.38	0.068992
Max Glucose	189.3344	173	86.83328
Min Glucose	109.1762	102	37.4836
Mean Glucose	145.9804	135.9	46.27097
Max WBC	15.0611	13.7	10.262
Min WBC	10.94903	10	7.456914
Mean WBC	12.94895	11.9	8.551809
Max BUN	24.74585	18	19.45151
Min BUN	20.76075	16	16.87047
Mean BUN	22.71265	17	17.98949
Max Creatinine	1.360863	1	1.334245
Min Creatinine	1.122919	0.8	1.093559
Mean Creatinine	1.23829	0.9	1.200874
Max Hemoglobin	12.39754	12.3	1.927305
Min Hemoglobin	9.758317	9.6	2.217876
Mean Hemoglobin	10.95254	10.7	1.847817

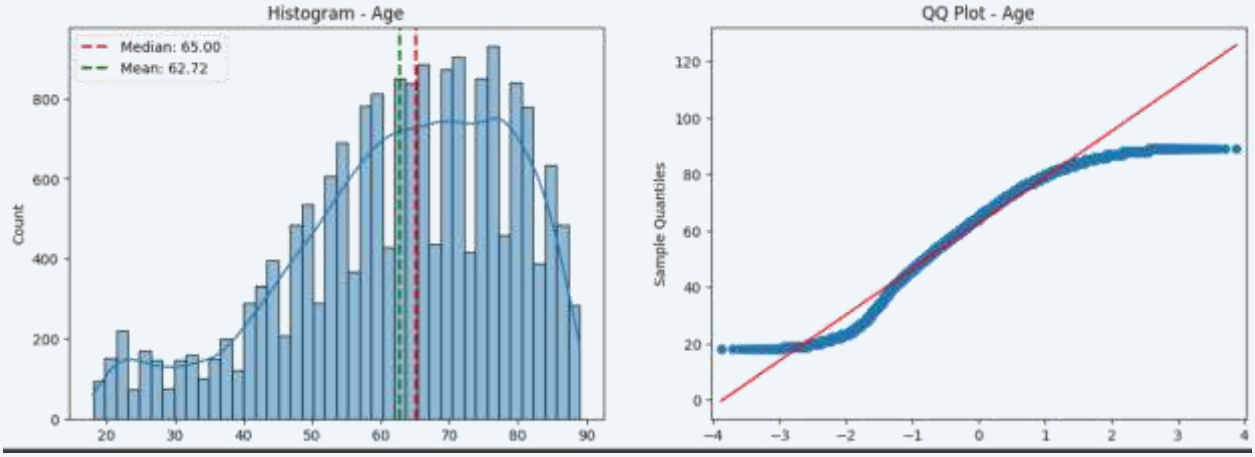
Ventilation Duration (h)	81.15984	19	154.8556
SAPS II	37.99121	36	14.81331
SOFA	5.04226	4	3.286905
OASIS	35.41169	35	8.258749

Normalization

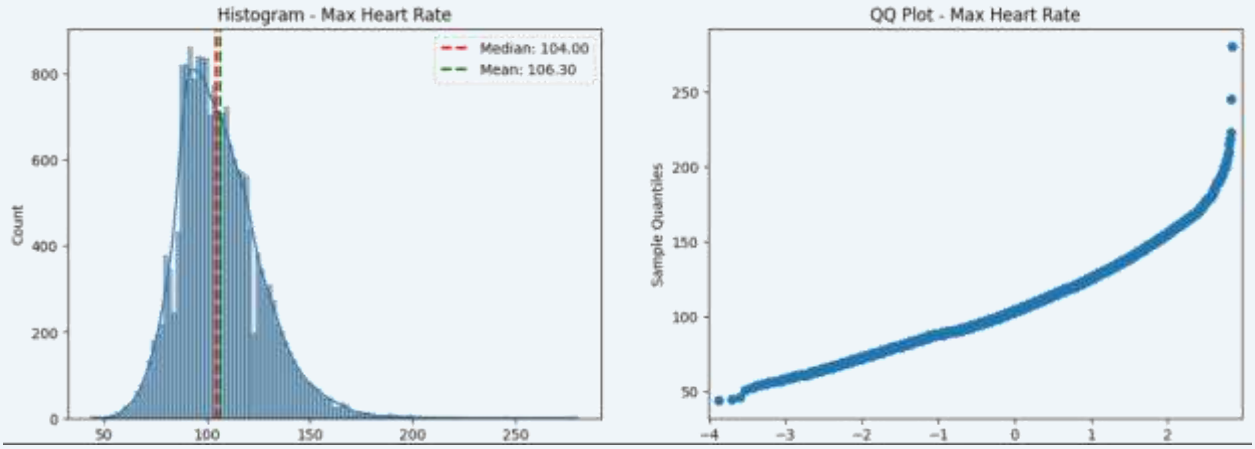
Normality check

Histograms and QQ Plots

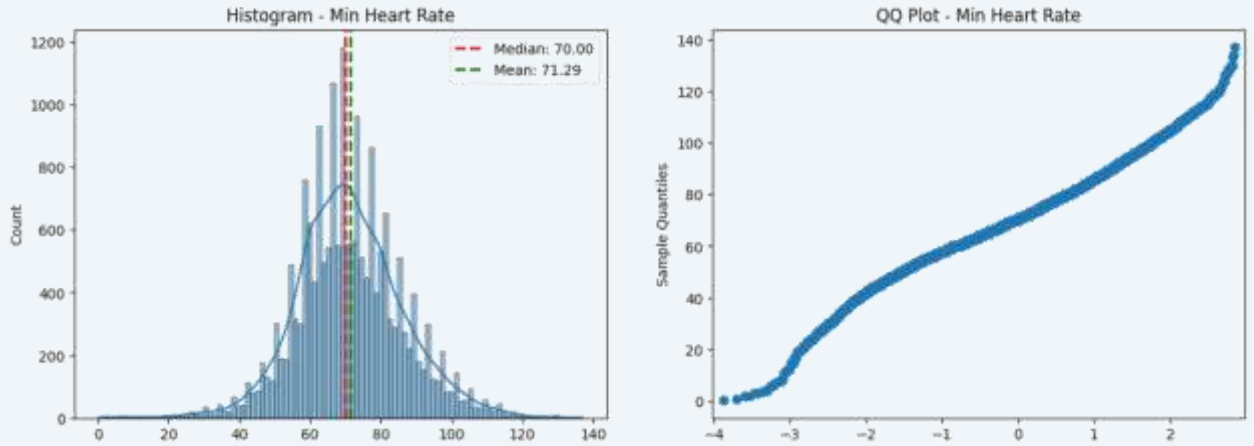
When assessing the normality of continuous variables, we created both histograms and quantile-quantile (Q-Q) plots.



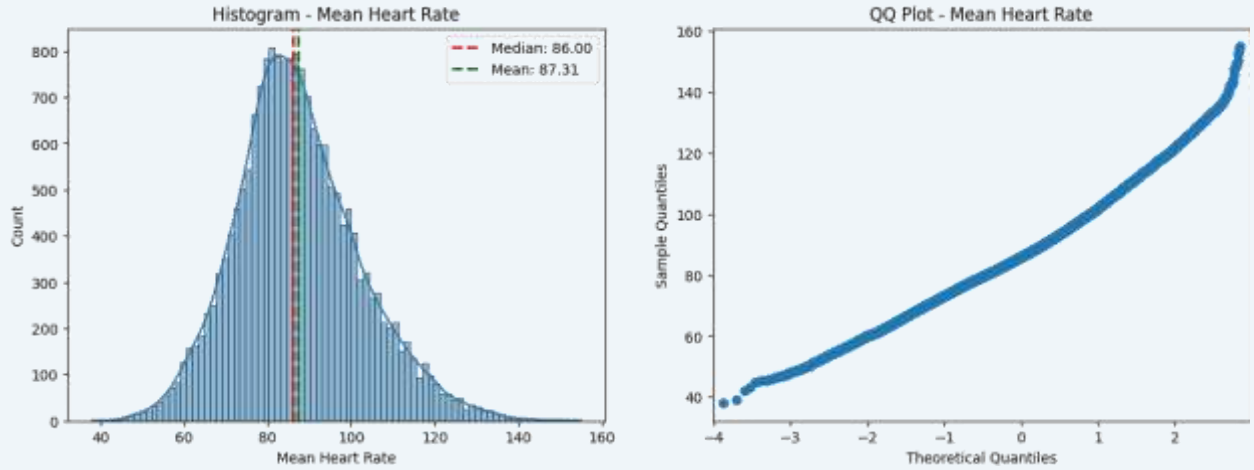
The histogram representing the age variables exhibited a bimodal distribution, indicating that there were two distinct peaks in the data. The Q-Q plot indicated that the age variable deviates from a normal distribution. This can be seen in the graph where points are departing from red line.



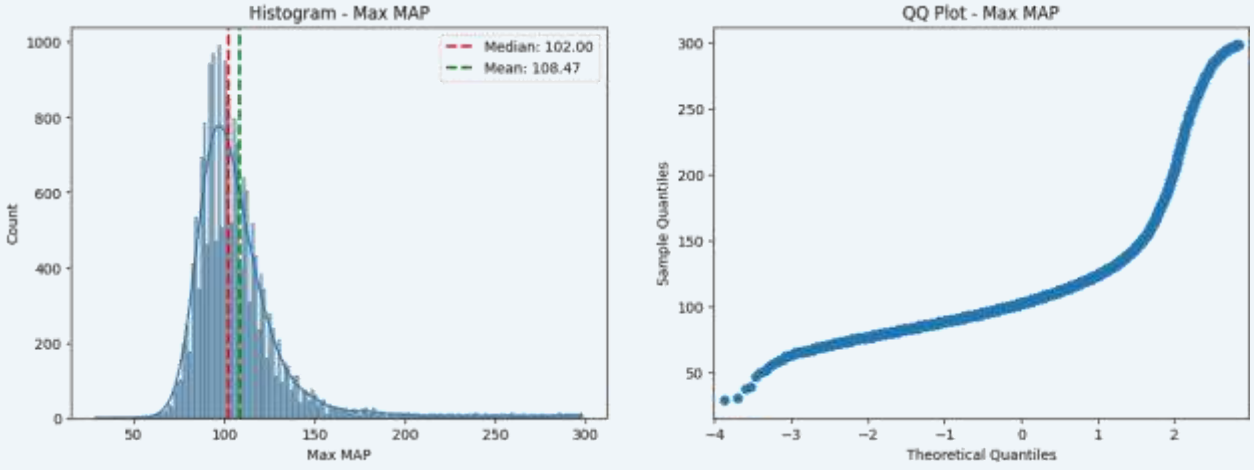
The histogram for the Max Heart Rate is a right skewed curve, which can be confirmed by the QQ plot where the data departs from the straight line at the right side.



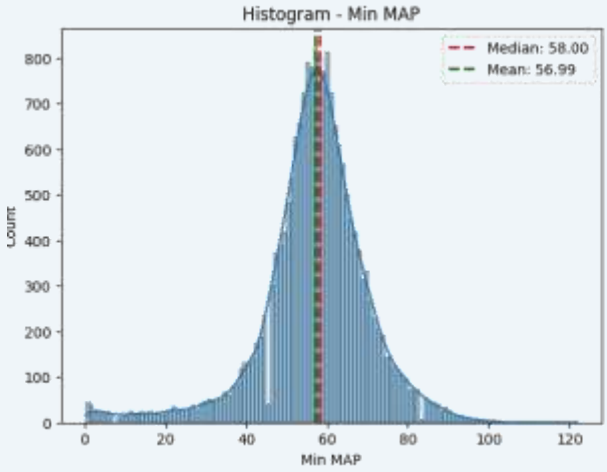
The histogram of the Min Heart Rate follows a normal distribution approximately. This can also be seen in the QQ Plot where the points are falling to a straight line.

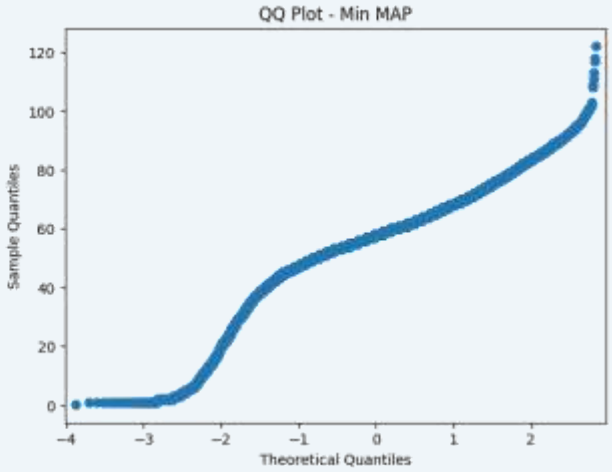


The histogram of the Mean Heart Rate follows a normal distribution approximately. This can also be seen in the QQ Plot where the points are falling to a straight line.

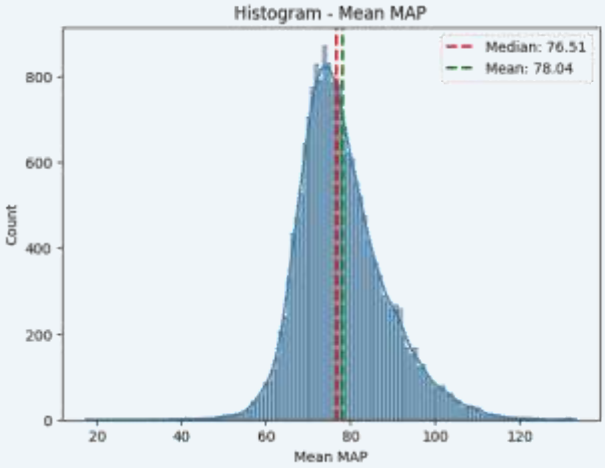


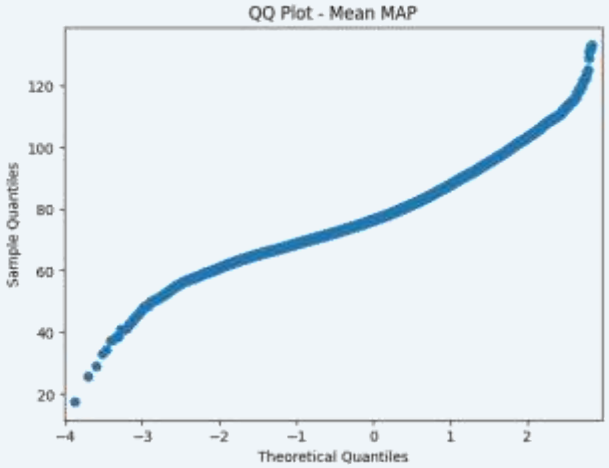
The histogram for the Max MAP is a right skewed curve, which can be confirmed by the QQ plot where the points form a dip on the right side of the plot.



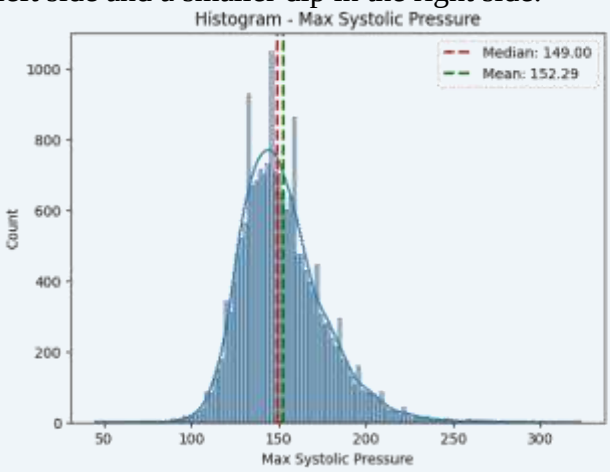


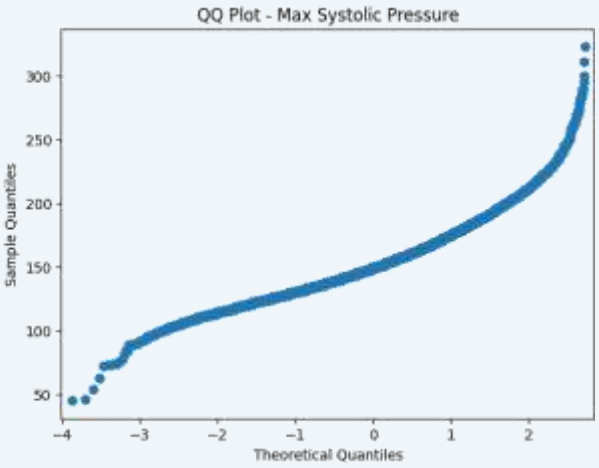
The histogram for the Min MAP is a left skewed curve, which can be confirmed by the QQ plot where the points form a dip on the left side of the plot.



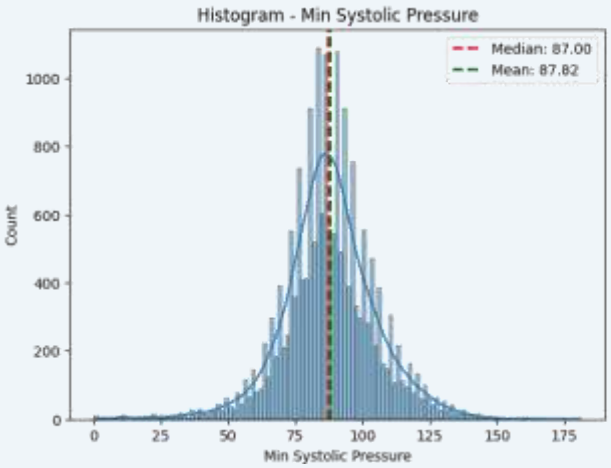


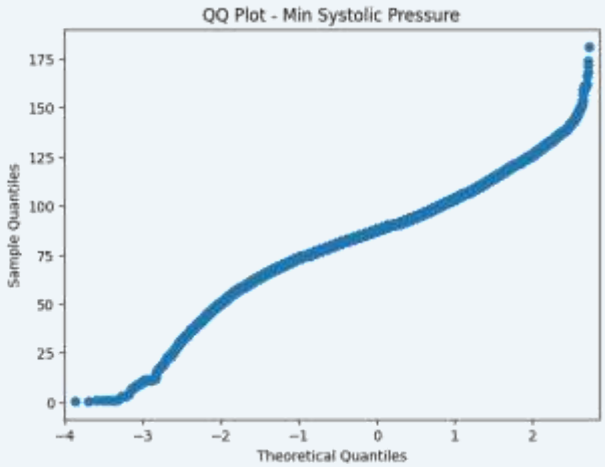
The histogram of the Mean MAP has a longer tail on the left side and shorter tail on the right side, which is clear in the QQ plot where the points from larger dip on the left side and a smaller dip in the right side.



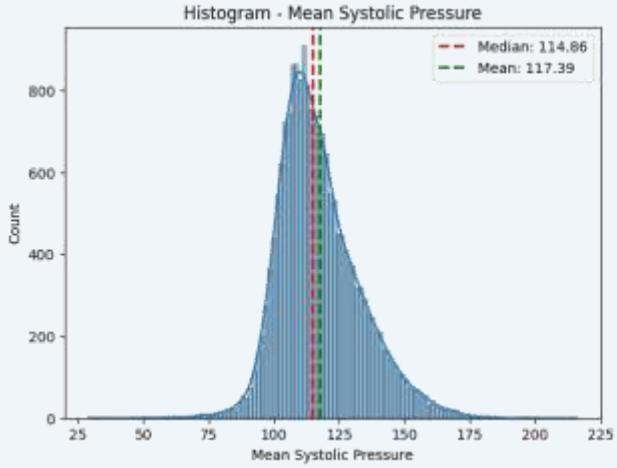


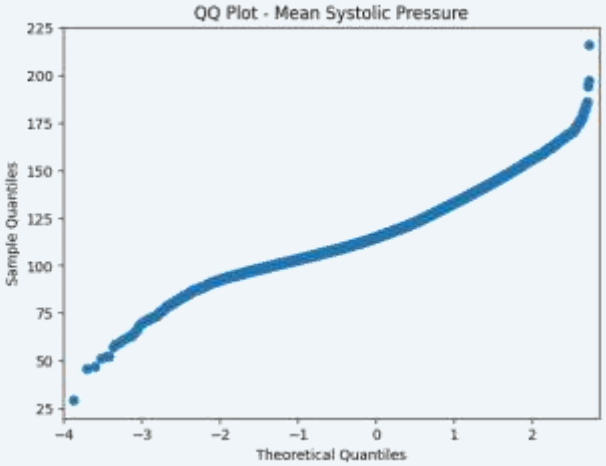
The Max Systolic Pressure has a longer tail on the right side which is confirmed in the QQ plot where there is a larger dip formed in the right side.



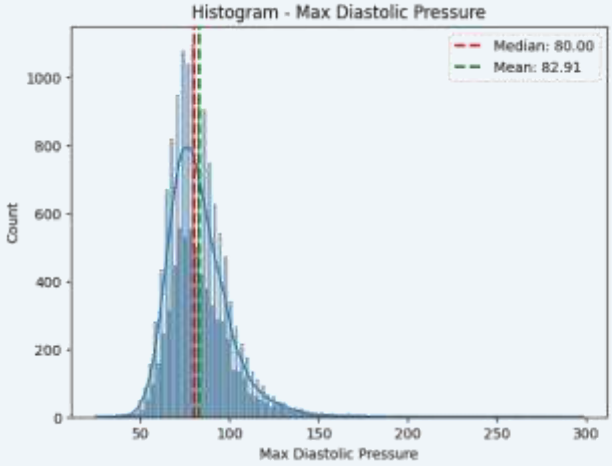


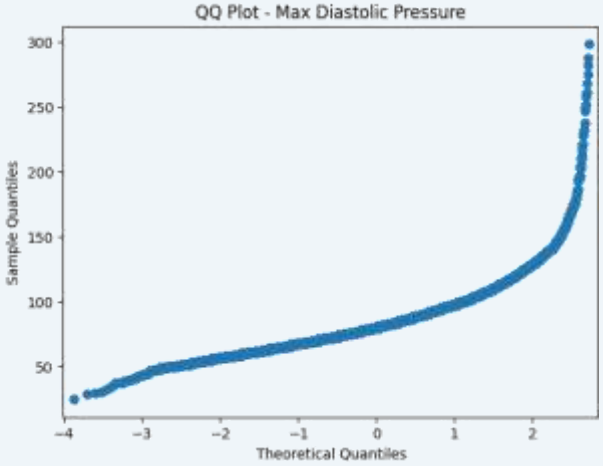
The Min Systolic pressure forms a normal distribution approximately.



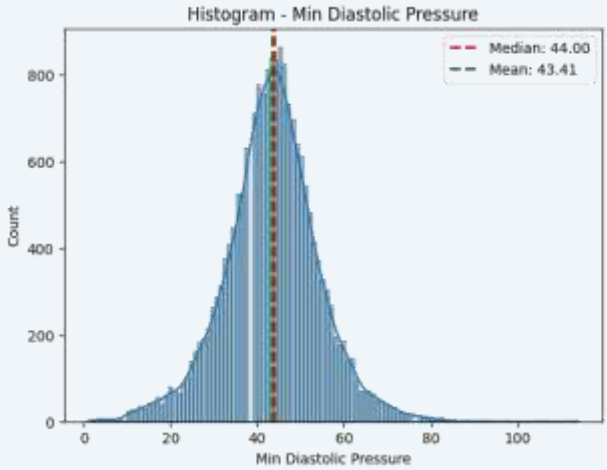


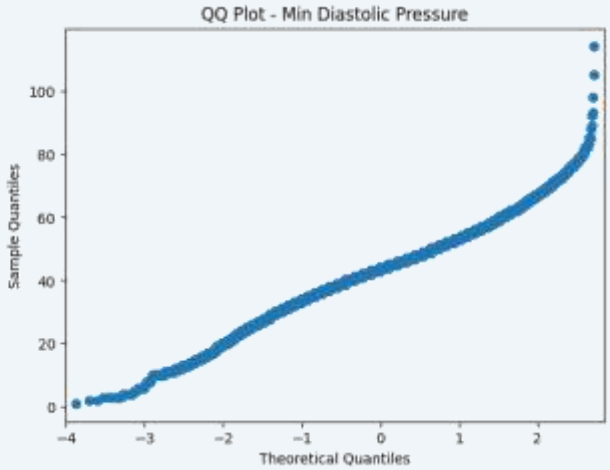
The Mean Systolic pressure forms a normal distribution approximately.



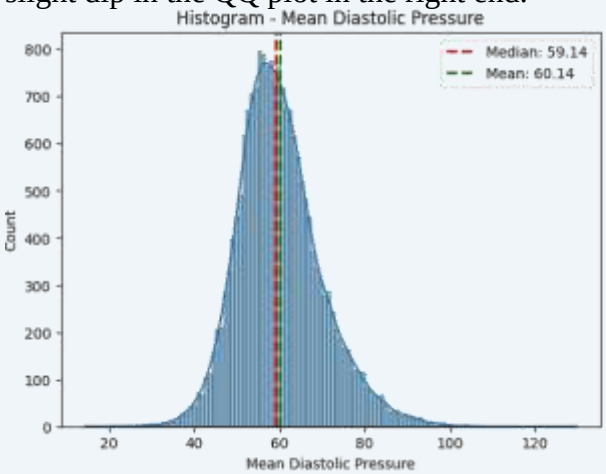


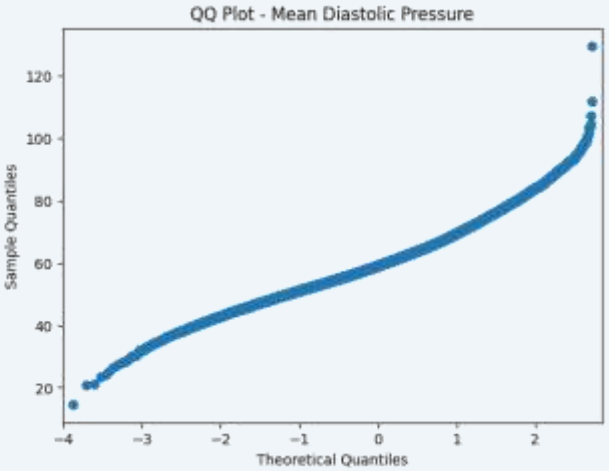
The max Diastolic pressure forms right skewed curve, which can also be seen in the QQ plot where the points form a dip in the right end.



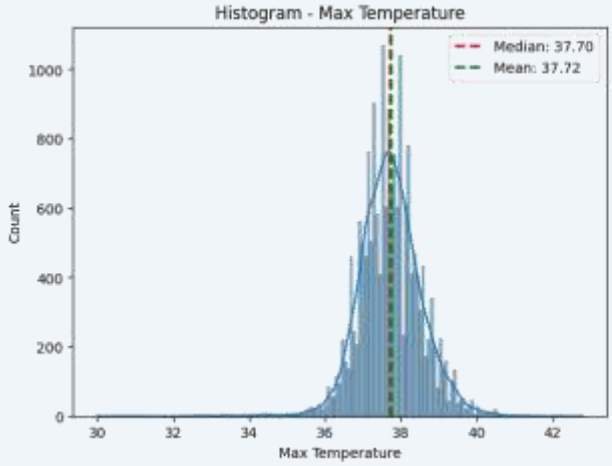


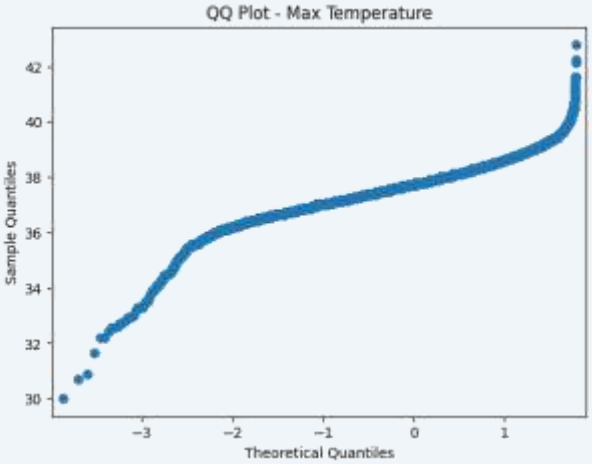
The Min Diastolic pressure forms a normal curve approximately, although there is slight dip in the QQ plot in the right end.



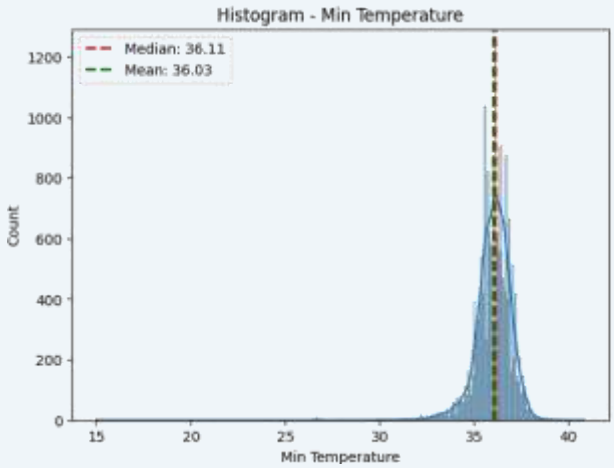


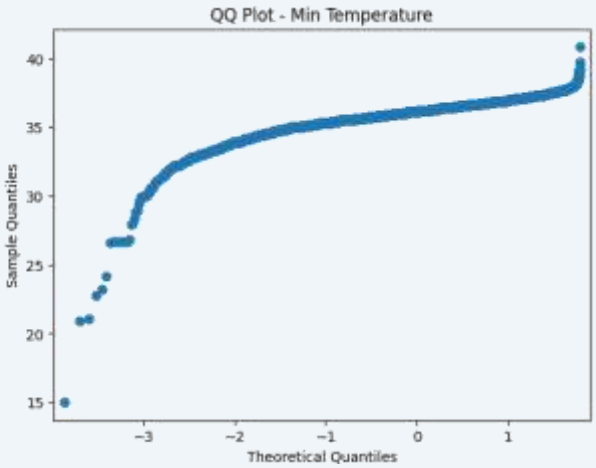
The Mean Diastolic pressure forms a normal curve approximately, although there is slight dip in the QQ plot in the right end.



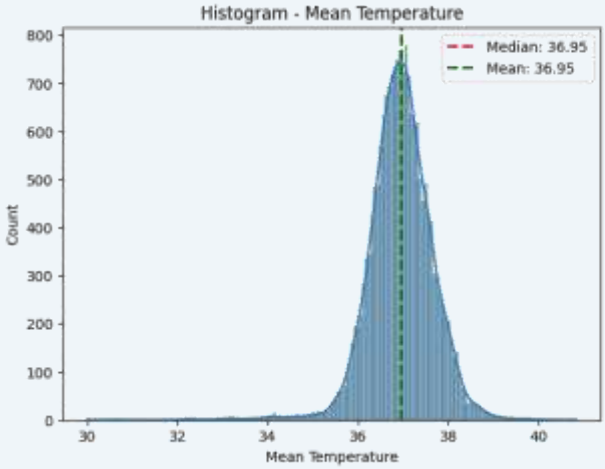


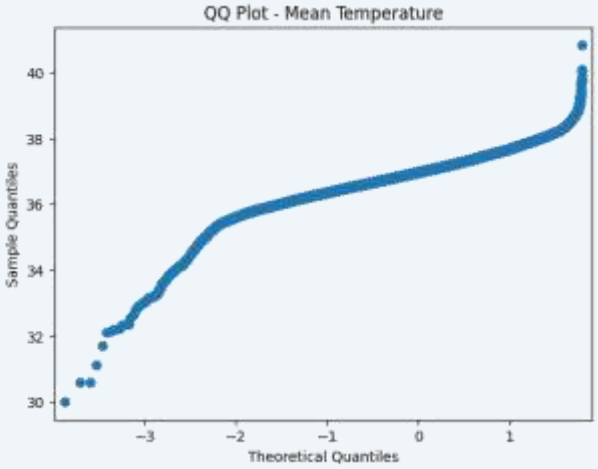
The Max Temperature follows a normal distribution approximately.



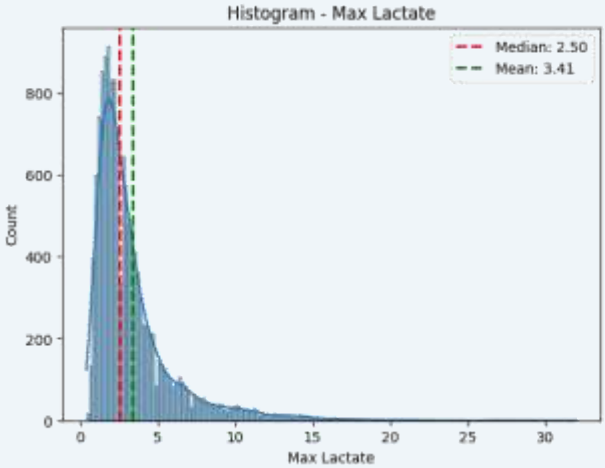


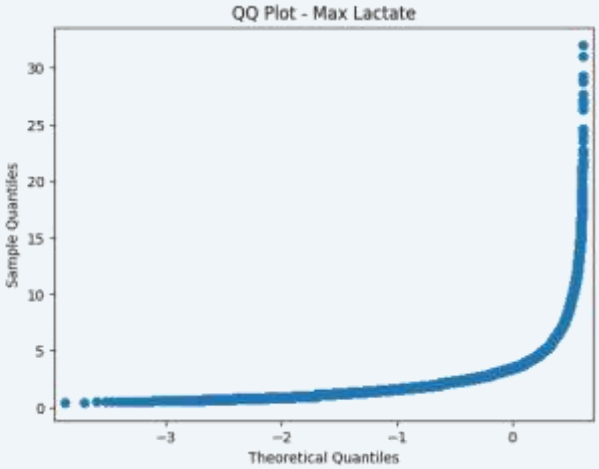
The Min Temperature forms a normal distribution approximately.



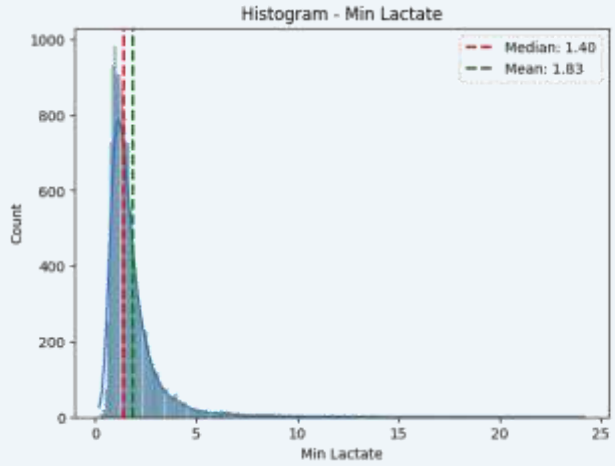


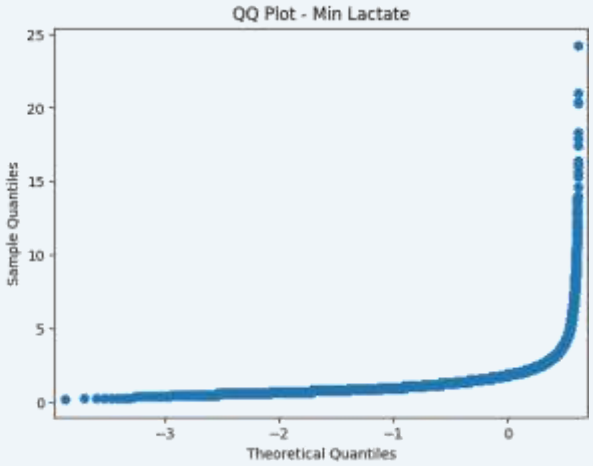
The Mean Temperature follows a normal distribution approximately.



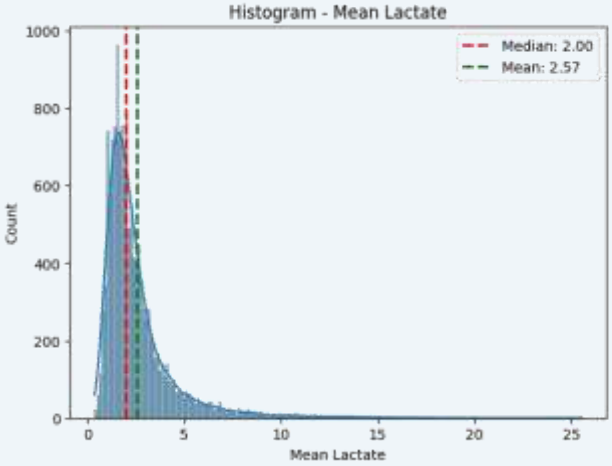


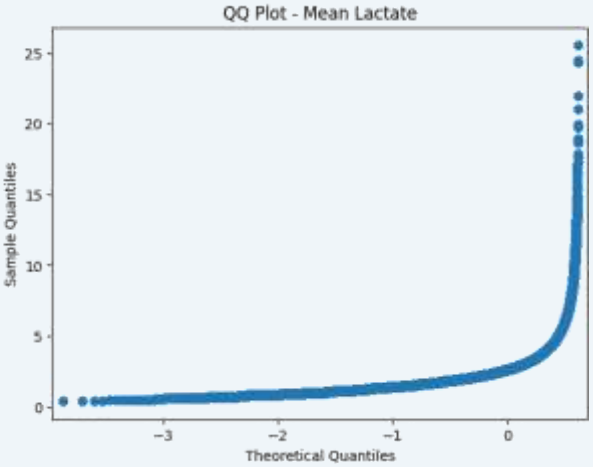
The Max Lactate is extremely right skewed, which is the reason the points form a larger dip in the right end of the QQ plot.



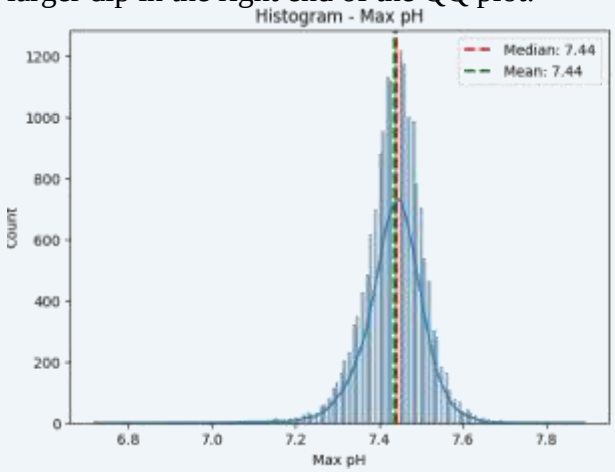


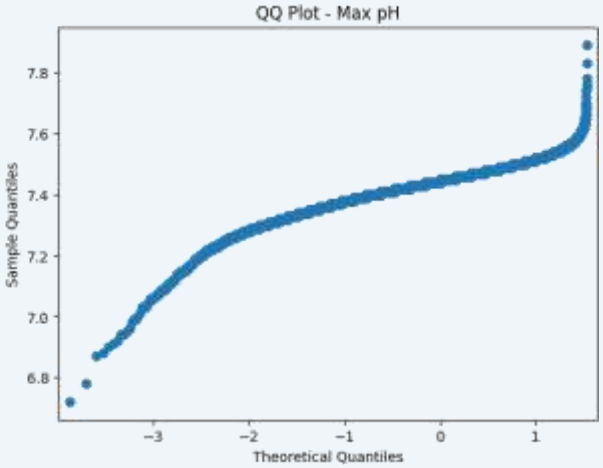
The Min Lactate is extremely right skewed, which is the reason the points form a larger dip in the right end of the QQ plot.



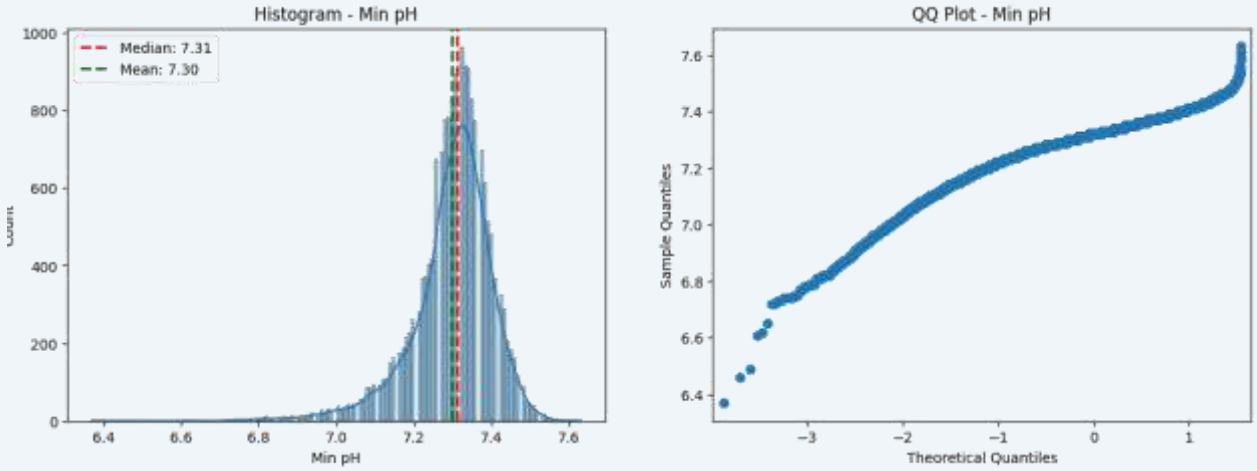


The Mean Lactate is extremely right skewed, which is the reason the points form a larger dip in the right end of the QQ plot.

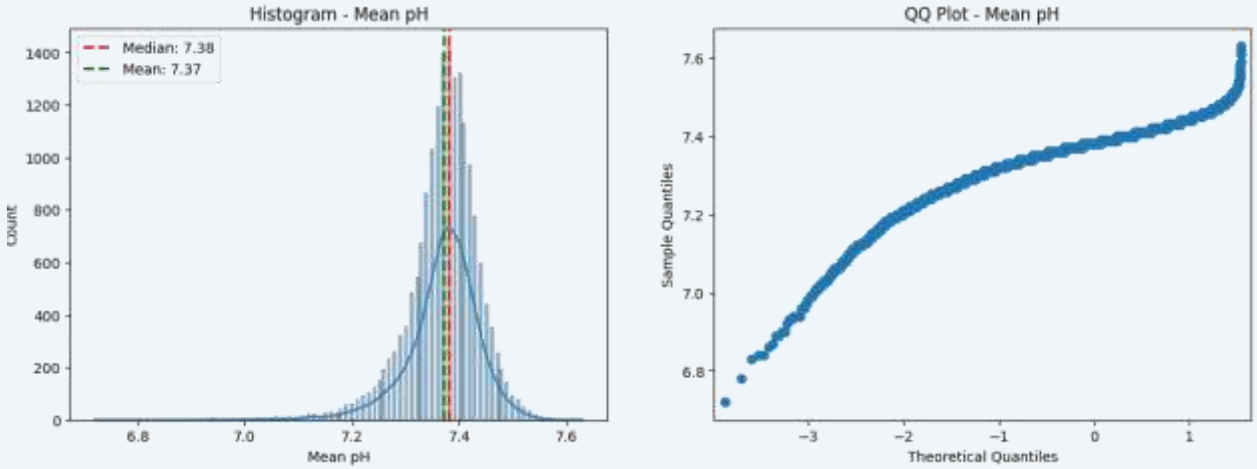




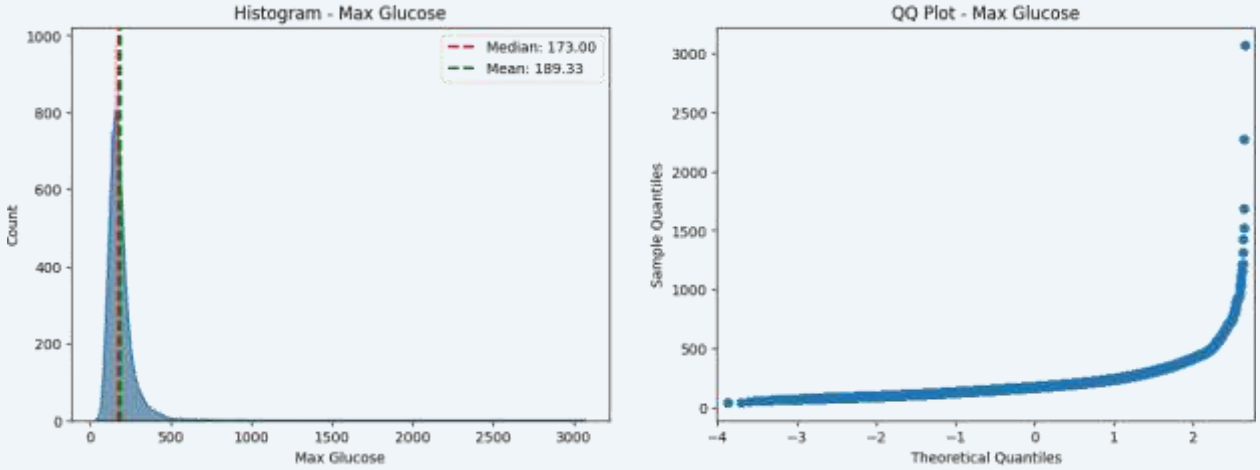
The Max pH follows a normal distribution approximately.



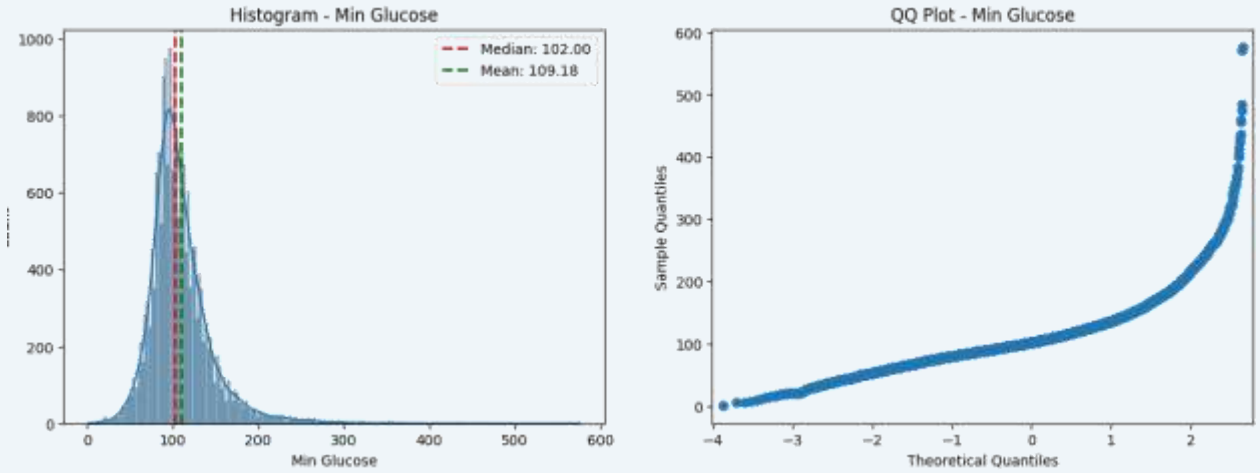
The Min pH follows a normal distribution approximately.



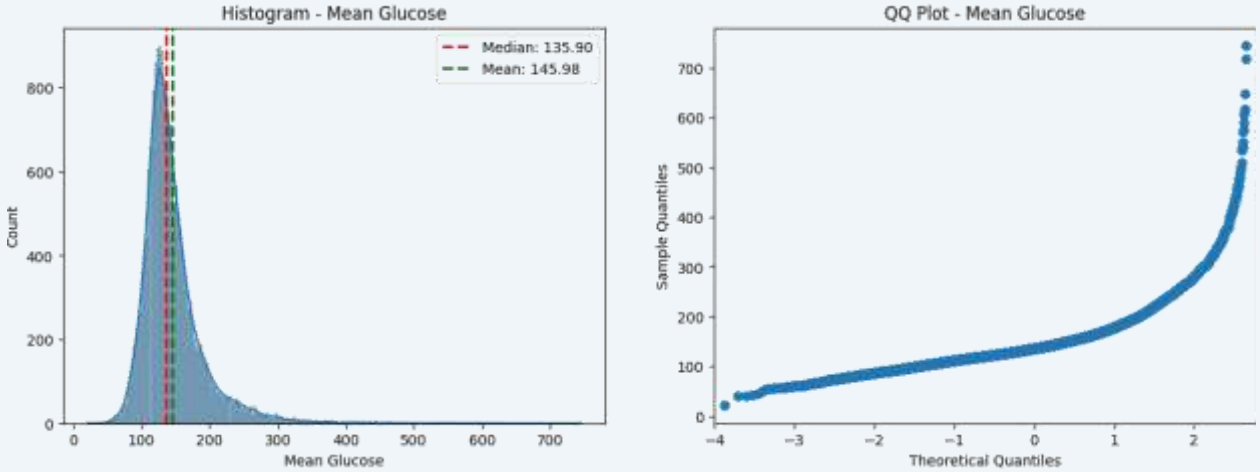
The Mean pH follows a normal distribution approximately.



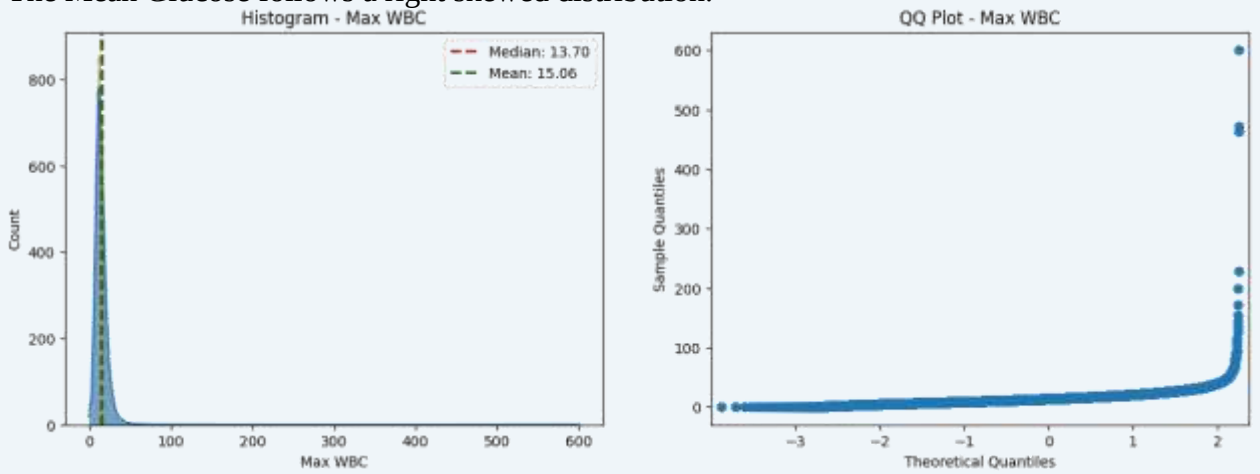
The Max Glucose follows a right skewed distribution.



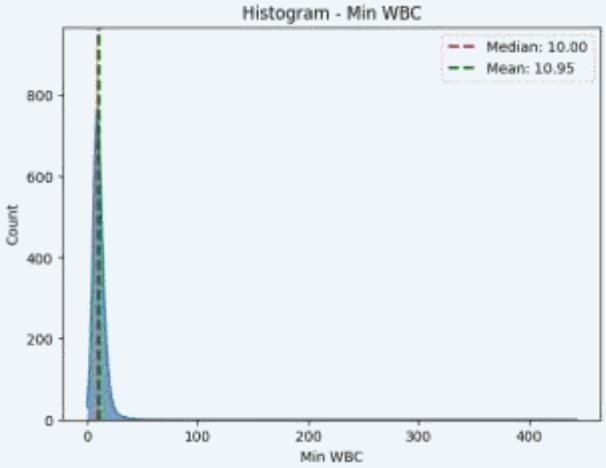
The Min Glucose follows a right skewed distribution.

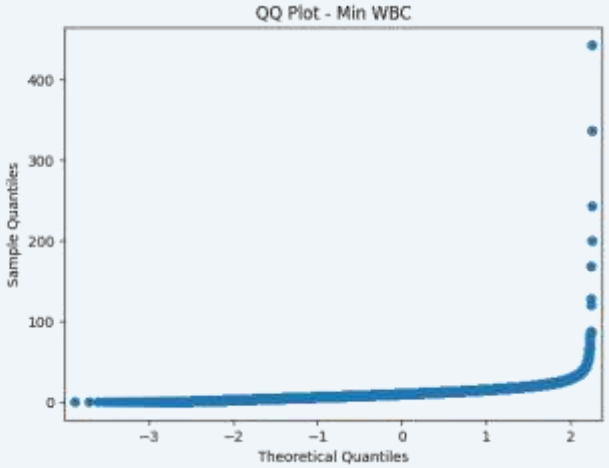


The Mean Glucose follows a right skewed distribution.

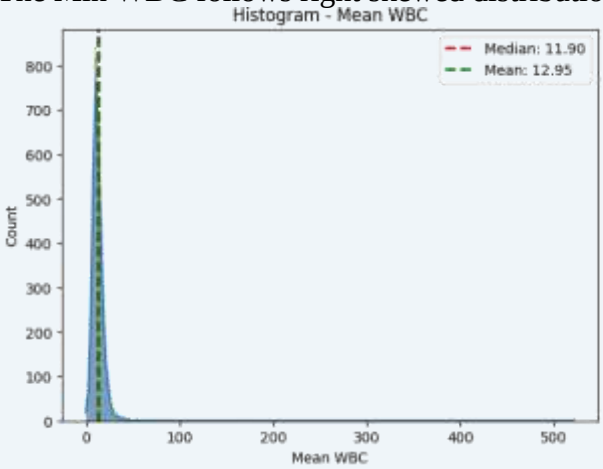


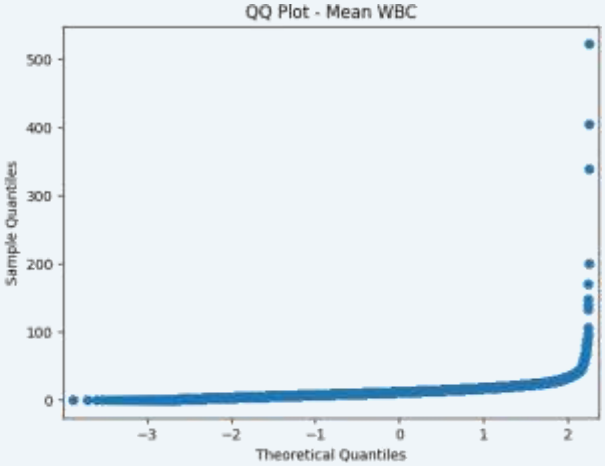
The Max WBC follows right skewed distribution.



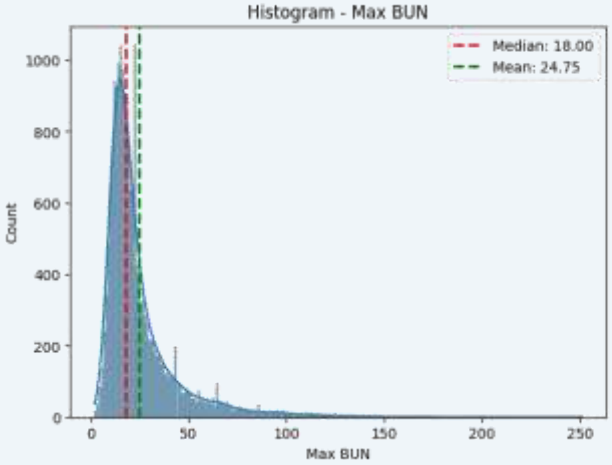


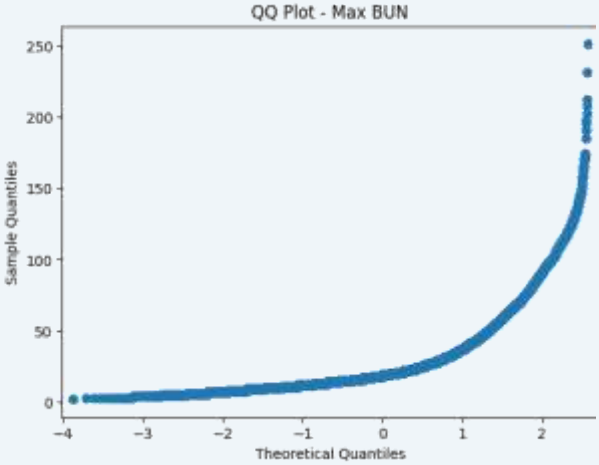
The Min WBC follows right skewed distribution.



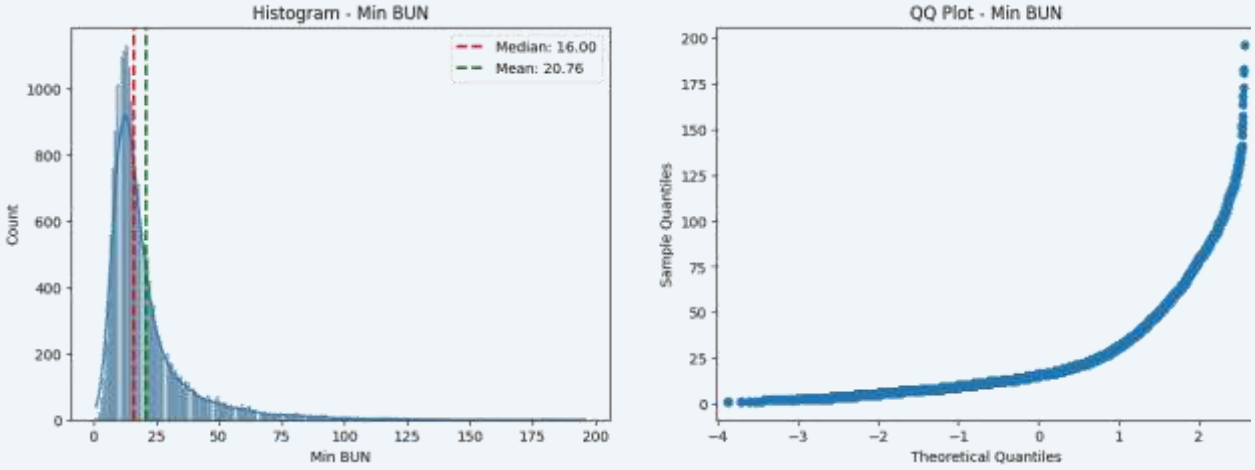


The Mean WBC follows right skewed distribution.

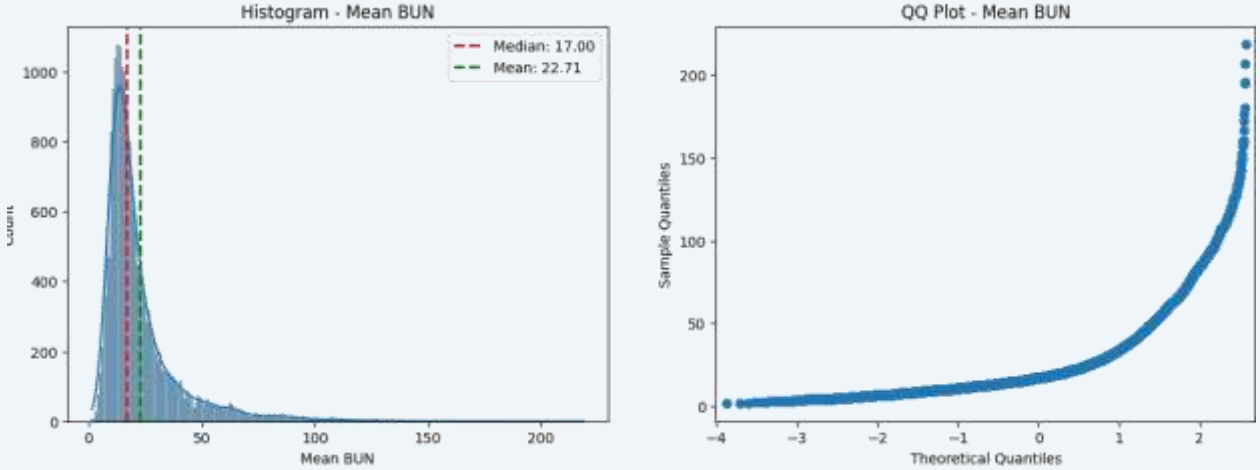




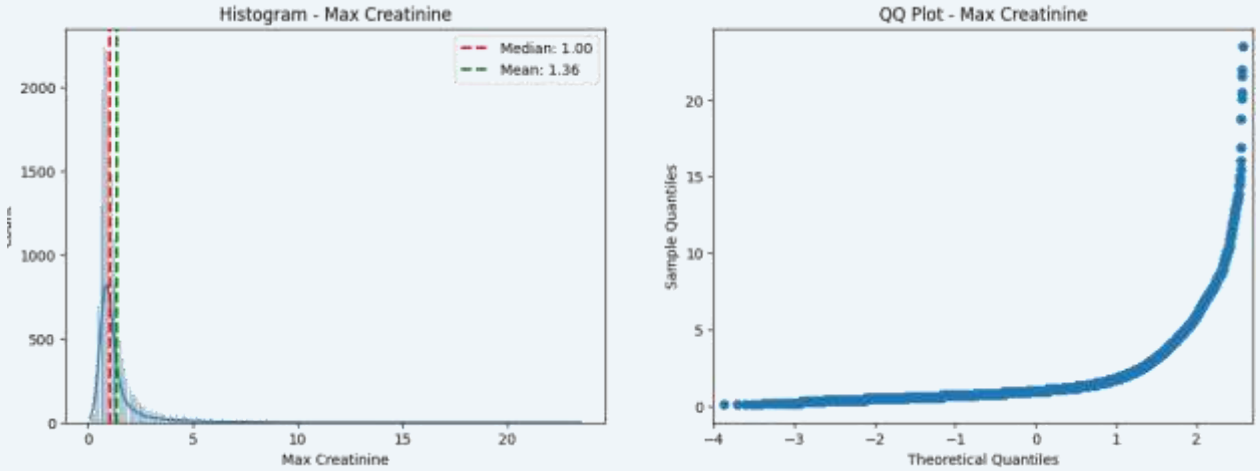
The Max BUN follows a right-skewed distribution.



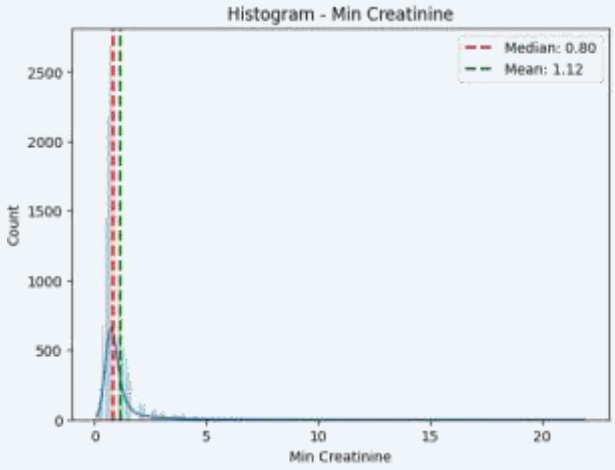
The Min BUN follows a right skewed distribution.

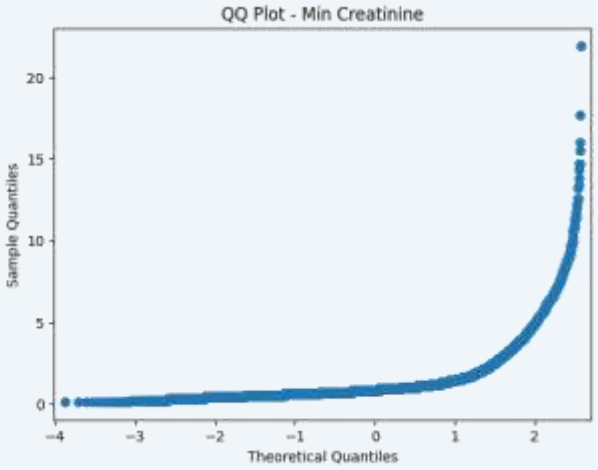


The Mean BUN follows a right skewed distribution.

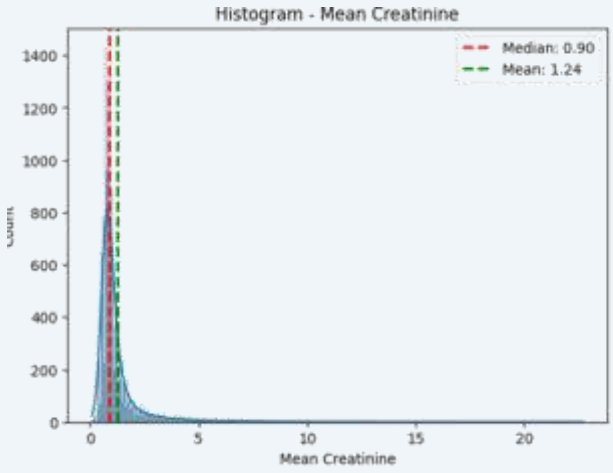


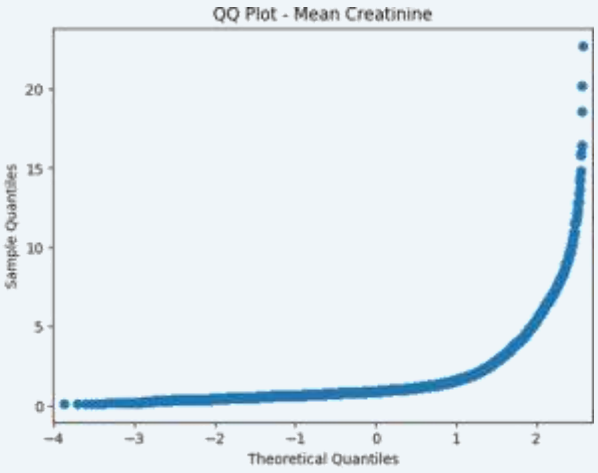
The Max Creatinine follows a right skewed curve.



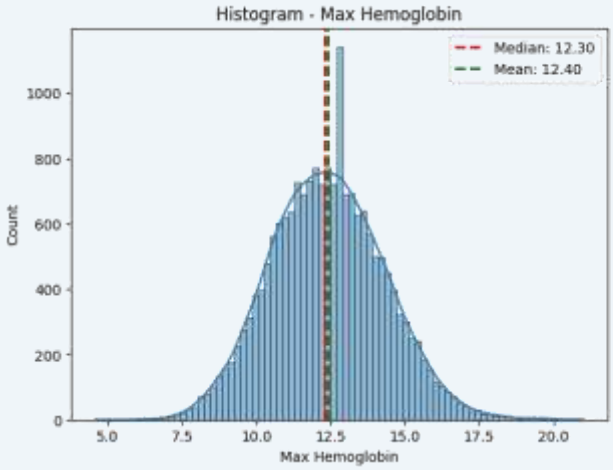


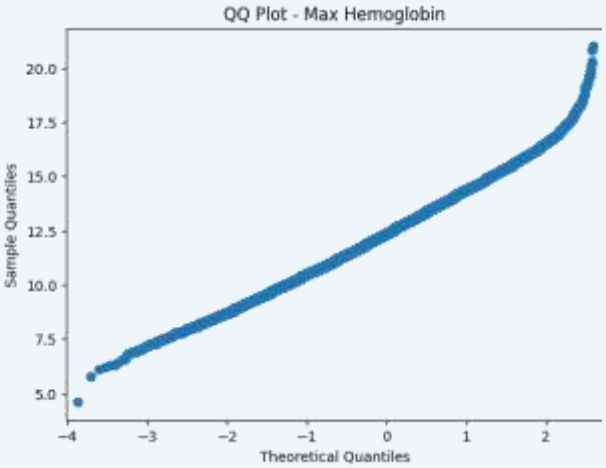
The Min Creatinine follows a right skewed curve.



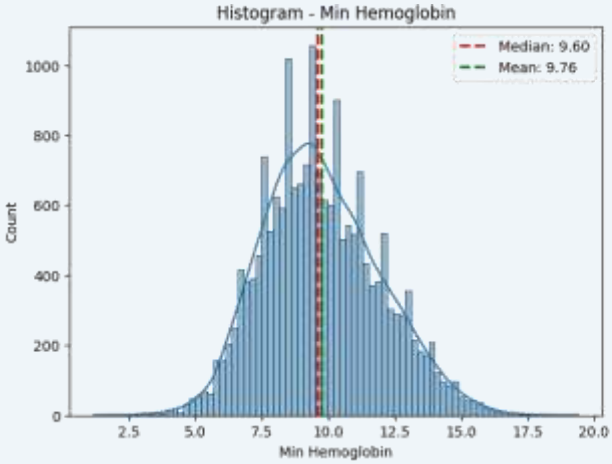


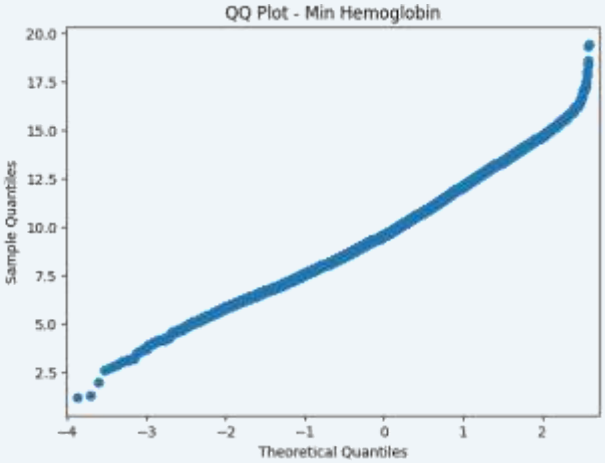
The Mean Creatinine follows a right skewed curve.



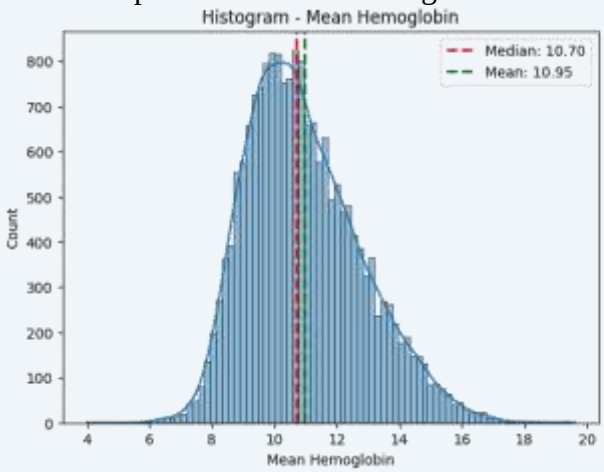


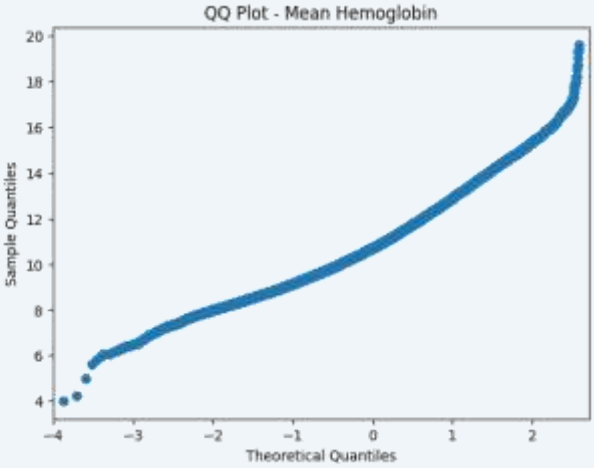
The Max Hemoglobin follows a normal distribution. This is apparent in the QQ plot where the points falls on to a straight line.





The Min Hemoglobin follows a normal distribution. This is apparent in the QQ plot where the points fall on to a straight line.





The Mean Hemoglobin follows a normal distribution approximately.

Shapiro Wilk test

We performed the Shapiro-Wilk test to confirm the normality of the above variables.

Shapiro-Wilk test for Age:
Statistic: 0.9588286876678467
p-value: 0.0

The following columns are all presented p-value with 0.

Shapiro-Wilk test for Max Hemoglobin:
Statistic: 0.9976174831390381
p-value: 1.2826694035654755e-15

The distribution of Max Hemoglobin is significantly different from normal.

Shapiro-Wilk test for Min Hemoglobin:
Statistic: 0.9931529760360718
p-value: 2.2844843665866908e-27

The distribution of Min Hemoglobin is significantly different from normal.

Shapiro-Wilk test for Mean Hemoglobin:
Statistic: 0.9779326319694519
p-value: 1.5414283107572988e-44

The distribution of Mean Hemoglobin is significantly different from normal.

/usr/local/lib/python3.10/dist-packages/scipy/stats/_morestats.py:1882: UserWarning: p-value may not be accurate for N > 5000.

warnings.warn("p-value may not be accurate for N > 5000.")

Upon reviewing the test results, we observed that none of the variables conform to a normal distribution. This discrepancy may be due to a failure to meet one of the test's conditions. Specifically, the sample size exceeds 5000, which violates a condition for this test."

Normalization

Since the continuous variables do not follow a normal distribution, we used four methods to normalize them as our purpose was to perform parametric tests on these variables.

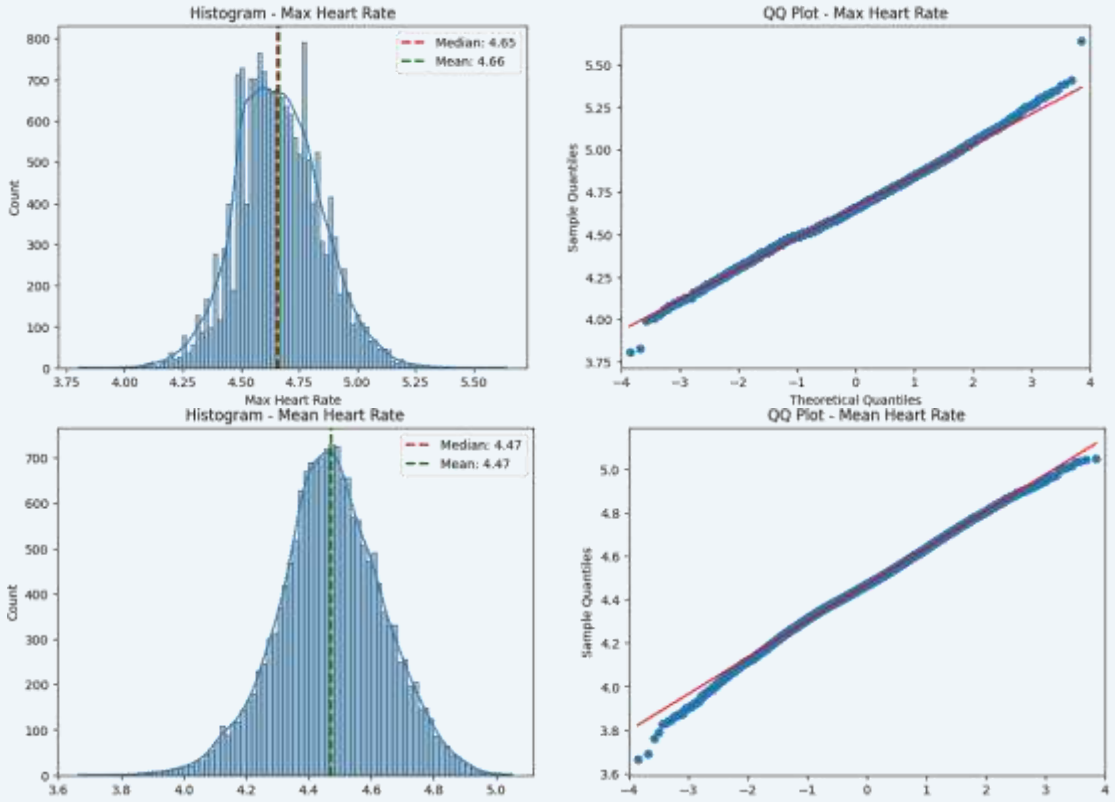
Method 1 – Log Transformation Method

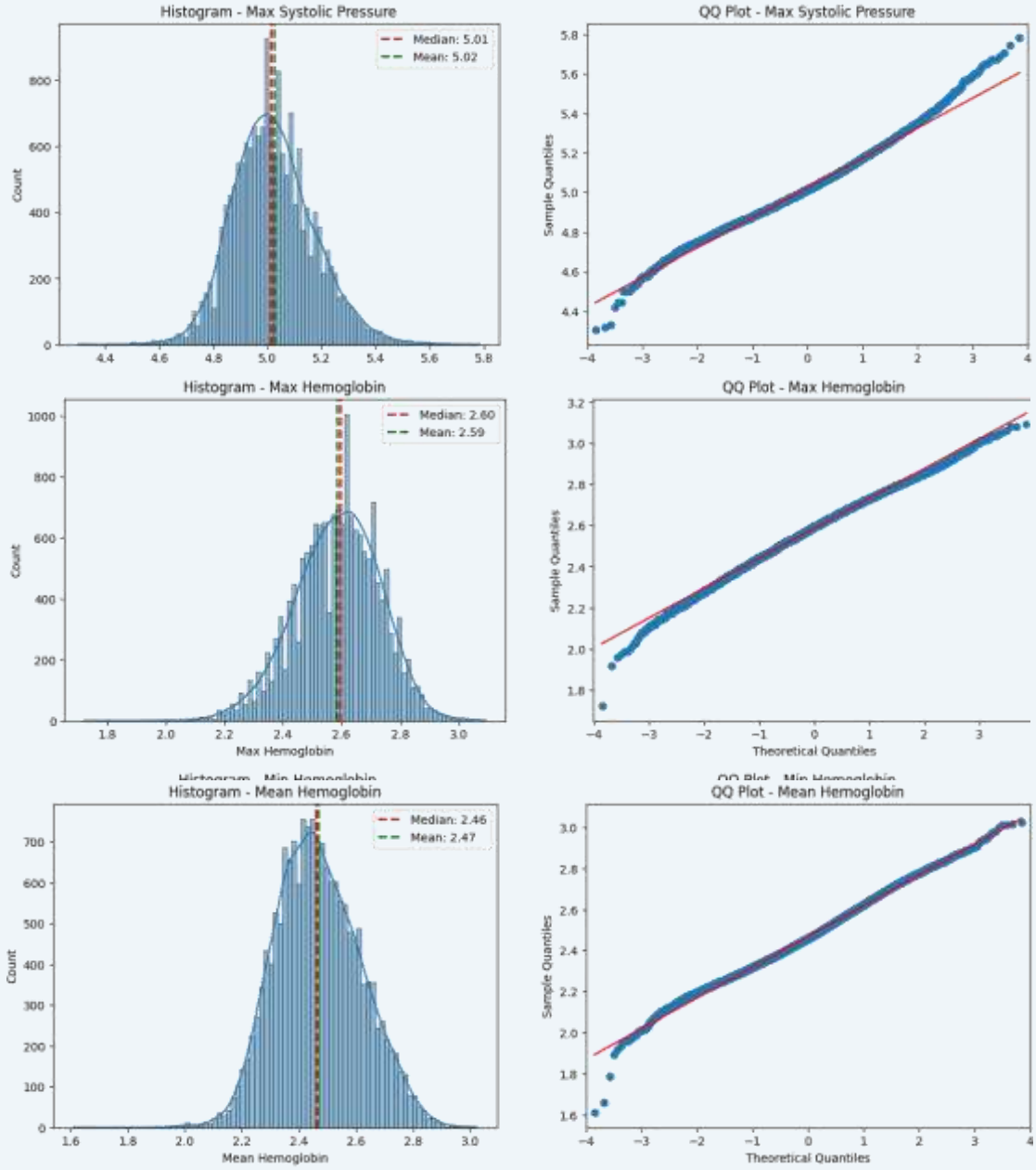
We applied log transformation method to achieve normal distributions in the variables.

Result

In the case of Max Heart Rate, Mean Heart Rate, Max Systolic Pressure, Max Hemoglobin and Mean Hemoglobin, we were able to achieve distributions that closely resemble a normal distribution.

For the remaining variables, there was still significant deviation from the normal distribution.



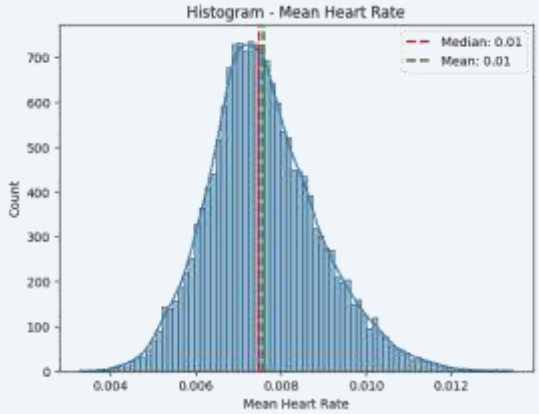


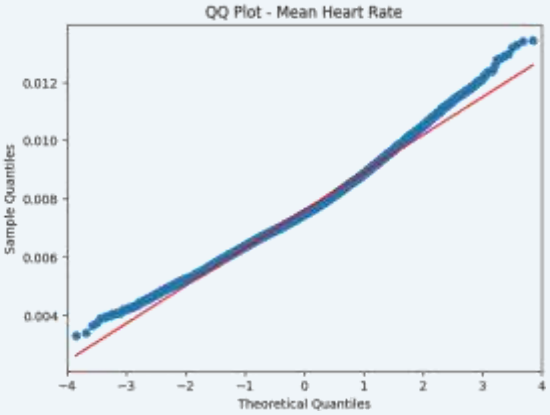
However, the Shapiro-Wilk test conducted after the Log Transformation indicated that all the variables exhibit distributions significantly different from the normal distribution."

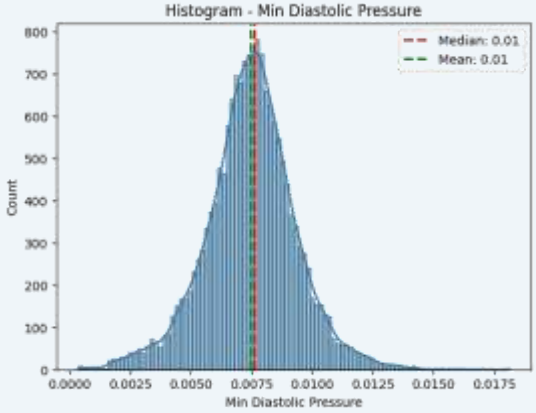
Method 2 – L2 Normalization

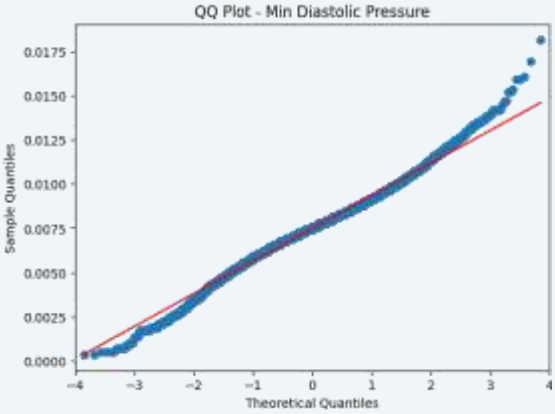
We used the Euclidean normalization method to achieve a normal distribution for the data. The L2 norm is a measure of the magnitude (length) of a vector and is defined as the square root of the sum of the squared elements of the vector.

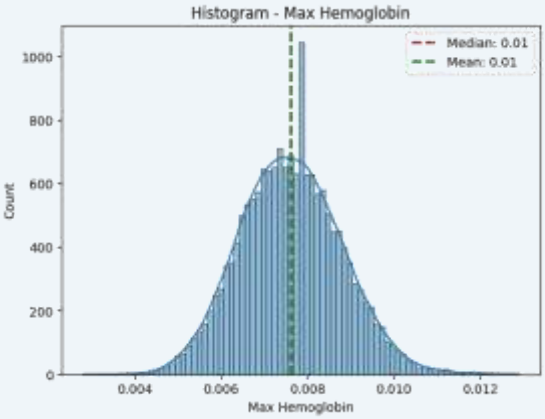
Among the variables we examined, Mean Heart Rate, Min Diastolic Pressure, Max Hemoglobin, and Min Hemoglobin exhibited distributions that closely resembled a normal distribution. However, for the remaining variables, there was significant deviation from the normal distribution.

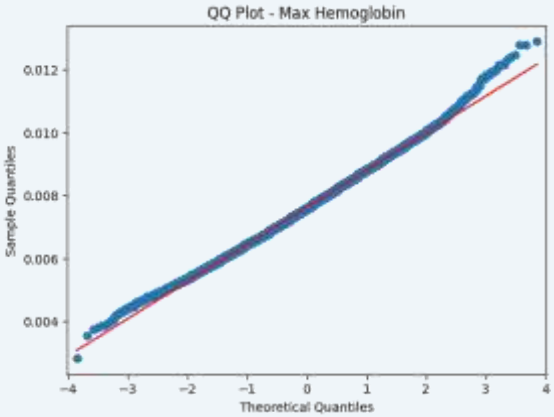

Histogram - Mean Heart Rate
Median: 0.01
Mean: 0.01
Count
Mean Heart Rate

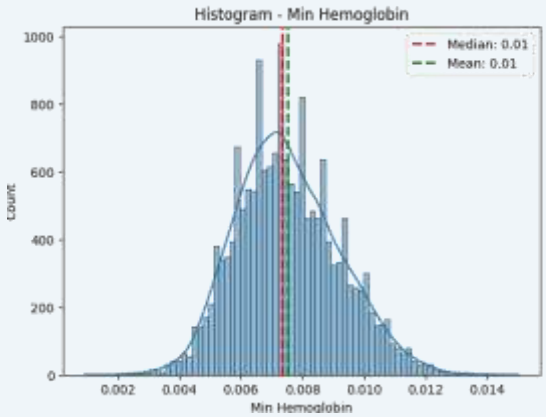

QQ Plot - Mean Heart Rate
Sample Quantiles
Theoretical Quantiles

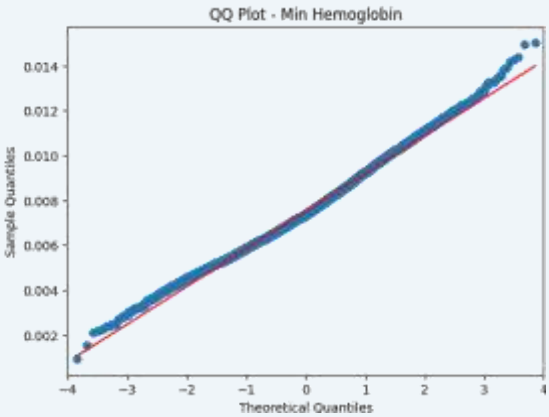

Histogram - Min Diastolic Pressure
Median: 0.01
Mean: 0.01
Count
Min Diastolic Pressure


QQ Plot - Min Diastolic Pressure
Sample Quantiles
Theoretical Quantiles


Histogram - Max Hemoglobin
Median: 0.01
Mean: 0.01
Count
Max Hemoglobin


QQ Plot - Max Hemoglobin
Sample Quantiles
Theoretical Quantiles

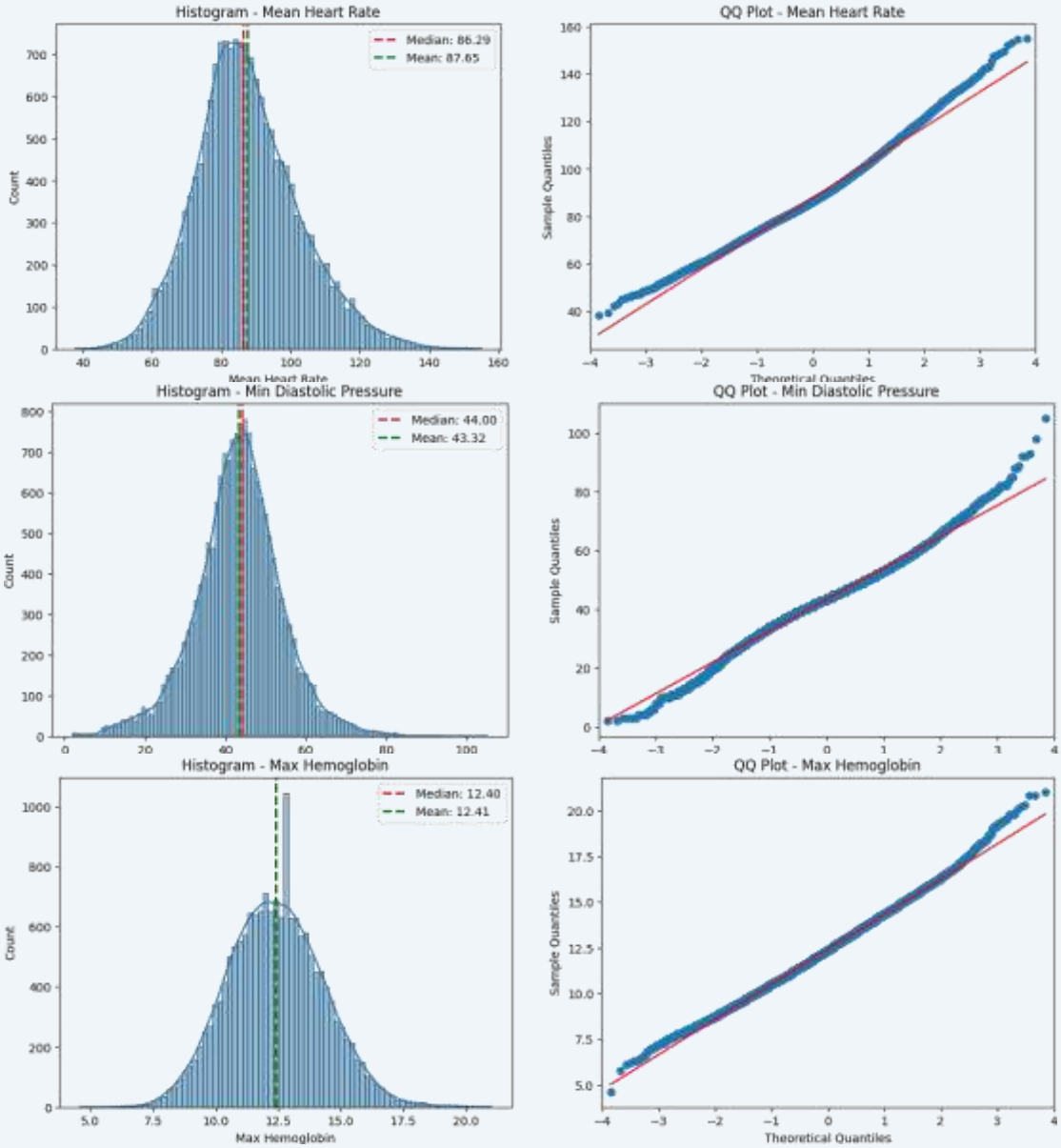

Histogram - Min Hemoglobin
Median: 0.01
Mean: 0.01
Count
Min Hemoglobin

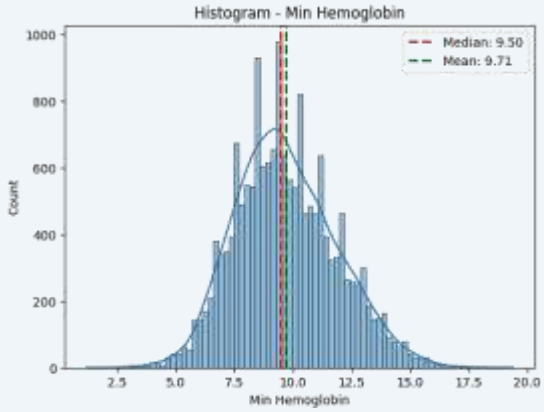

QQ Plot - Min Hemoglobin
Sample Quantiles
Theoretical Quantiles

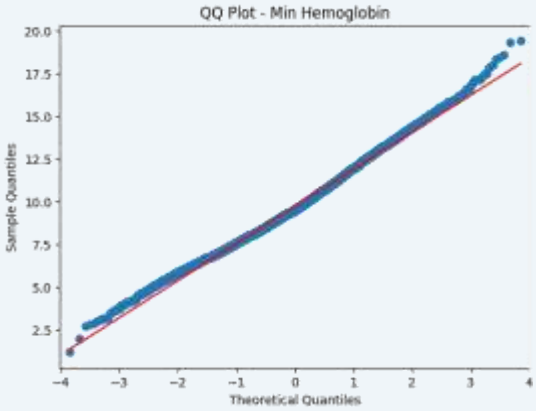
Method 3 – BoxCox Method

To achieve normal distributions for the variables, we utilized the BoxCox transformation method. To align the data with a more Gaussian-like distribution, improving the suitability for statistical analyses

In our analysis, we observed that Mean Heart Rate, Min Diastolic Pressure, Max Hemoglobin, and Min Hemoglobin displayed distributions that closely resembled a normal distribution. However, for the remaining variables, there was significant deviation from the normal distribution.

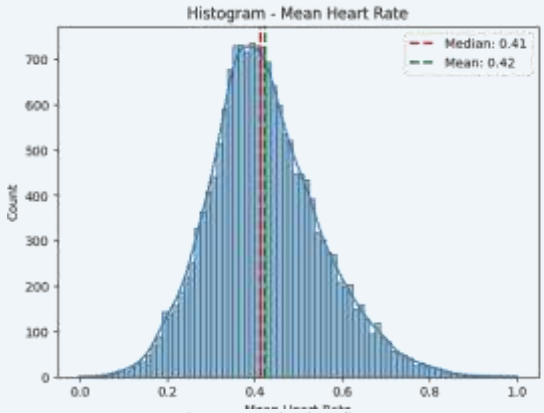


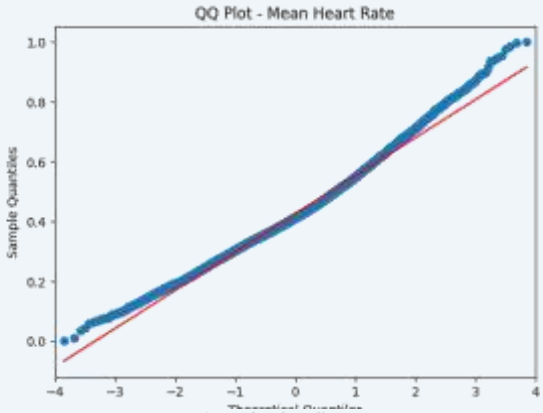


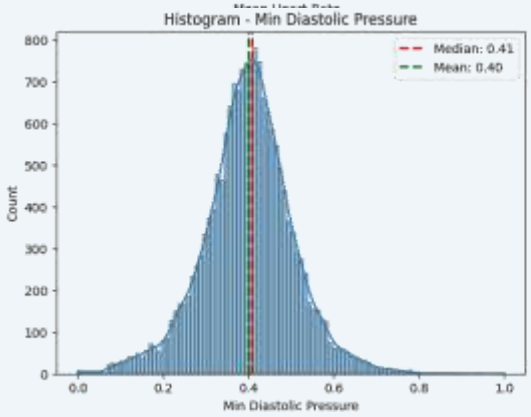


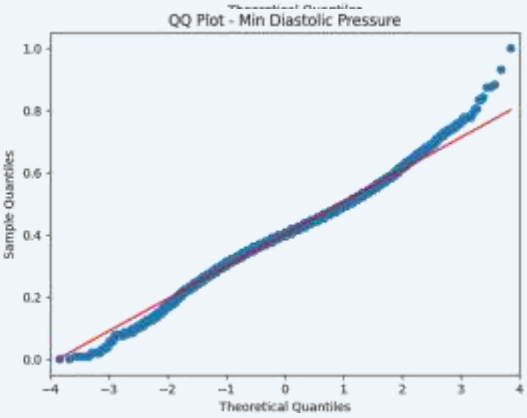
Method 4 – Min-Max Method

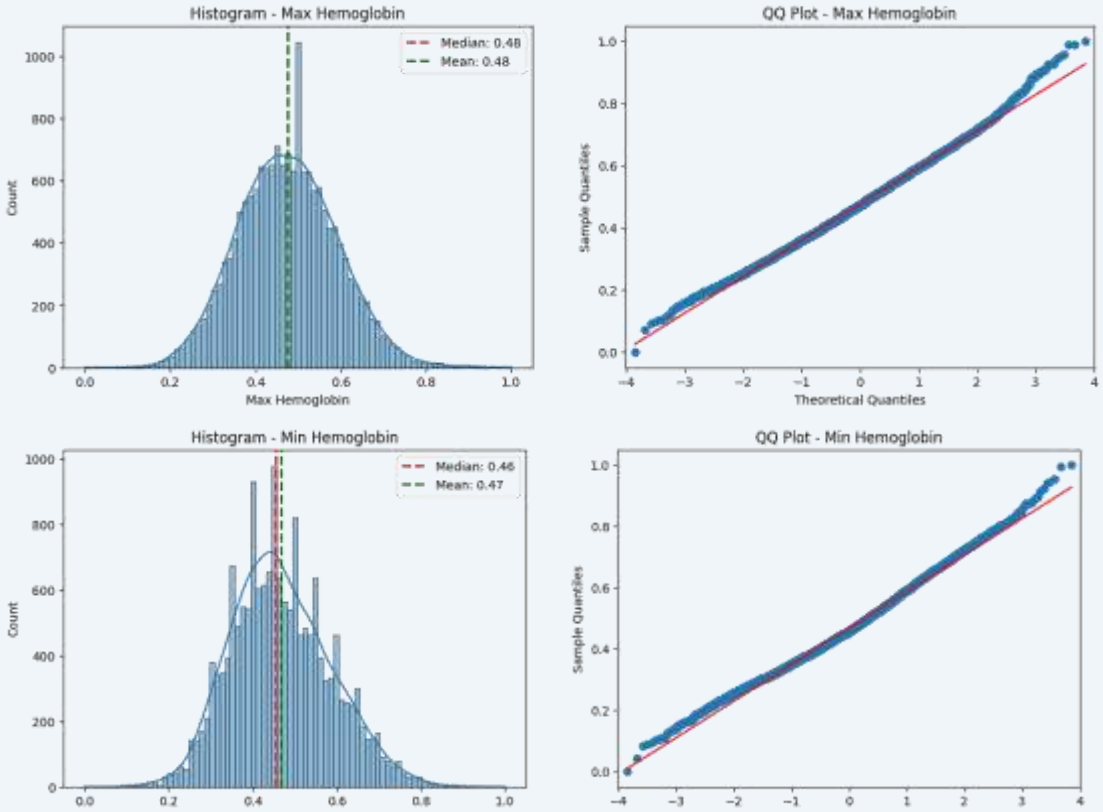
We applied the Min-Max method to achieve normal distributions in the variables. Among the variables we examined, Mean Heart Rate, Min Diastolic Pressure, Max Hemoglobin, and Min Hemoglobin exhibited distributions that closely resembled a normal distribution. However, for the remaining variables, there was significant deviation from the normal distribution.











Statistical Tests

Ordinal Variables

Mann-Whitney U test

We used the Mann-Whitney U and Spearman Correlation tests for the ordinal variables. Note that all assumptions listed below are met.

Mann-Whitney U Assumptions:

1. Observations in each group are independent of each other.
2. Random sampling is assumed.
3. The variables are ordinal or interval.
4. Shapes of the distributions in the two groups are similar.

H0 = The distributions of the living group are equal to those of the deceased group.
Ha = There is a difference between the distributions of the two groups.

Since the p-value is <.05 for all 3 variables, we have enough evidence to reject the null hypothesis. Therefore, can conclude that there is a difference between the distributions of the two groups, dead and alive.

Variable	Category	Data_Type	Type_of_Test	P-value
OASIS	Severity	Ordinal	Mann-Whitney U	6.7278E-308
SOFA	Severity	Ordinal	Mann-Whitney U	3.5648E-178
SAPS II	Severity	Ordinal	Mann-Whitney U	0

Spearman Correlation

We divided the dataset into two groups, survivors and non-survivors, and calculated the Spearman correlation coefficient for each pair of variables within each group.

Spearman Correlation Assumption:

1. The variables are either ordinal or continuous data that follow a monotonic relationship.

H0: There is no monotonic association between the two variables.
Ha: There is a monotonic association between the two variables.

The result of the correlation test reveals that the correlation coefficients in the non-survivor groups are higher compared to the survivor groups which suggests that the relationship between variables differ between the two groups.

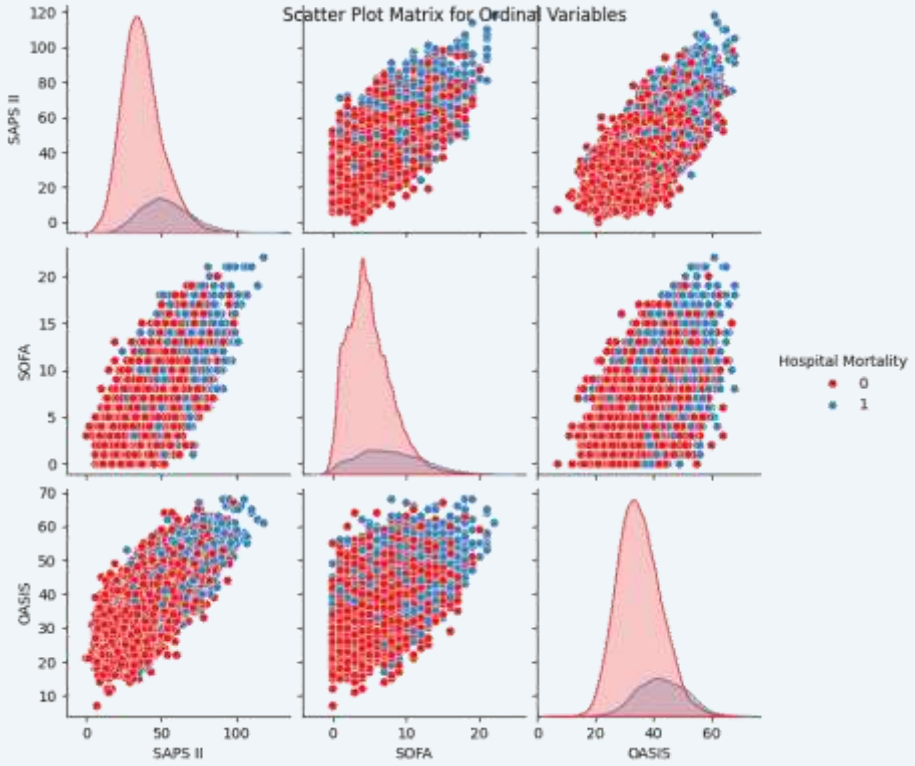
The resulting correlation coefficients suggests that there is moderate positive monotonic relationship between SAPS II and SOFA, and SAPS II and OASIS and a weak positive monotonic relationship between SOFA and OASIS. With the significance level at 0.05, the p-value tells us that the results are statistically significant.

Variable_1	Variable_2	Test	Correlation	P_value	Group
SAPS II	SOFA	Spearman	0.596778	0.000000e+00	Survivors
SAPS II	SOFA	Spearman	0.675700	7.755076e-288	Non-Survivors
SAPS II	OASIS	Spearman	0.598171	0.000000e+00	Survivors
SAPS II	OASIS	Spearman	0.674522	1.814579e-286	Non-Survivors
SOFA	OASIS	Spearman	0.360364	0.000000e+00	Survivors
SOFA	OASIS	Spearman	0.453597	5.383949e-110	Non-Survivors

Visualization of Ordinal Variables

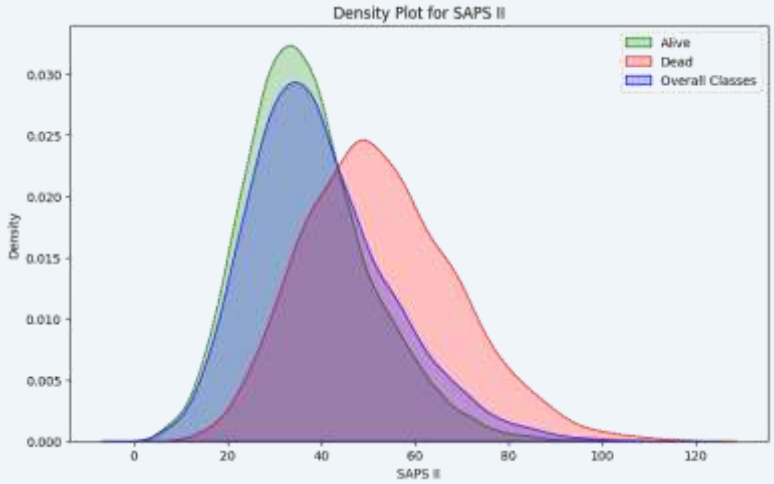
Scatter Plot

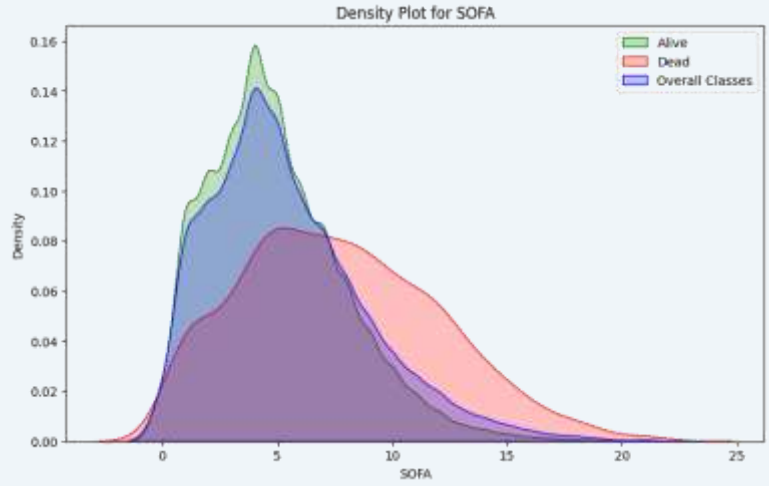
The scatter plot illustrates that there is a positive relationship between each pair of the ordinal variables.

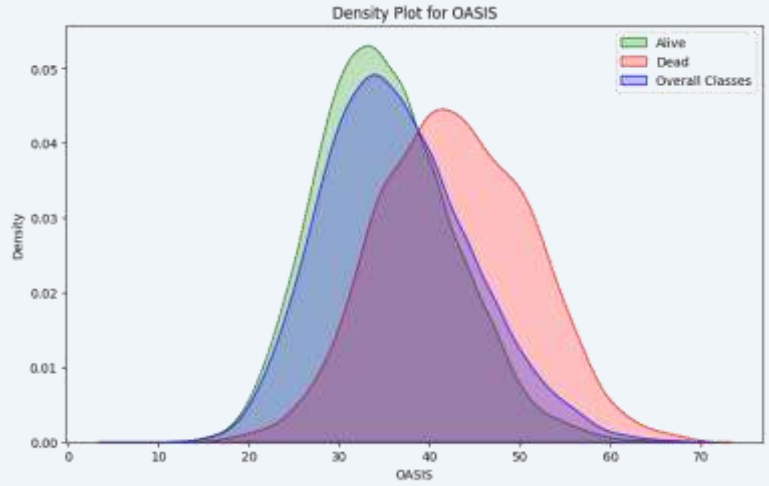


Density Plots

The density plots for all three severity-of-illness score variables reveal that the variability of scores for non-survivors is much more spread out compared to survivors.







Categorical Variables

Chi-Square test

With respect to the categorical variables, we used the Chi-Square Test of Independence. Note that all assumptions listed below are met.

Assumptions:

1. The observations in the independence table must be independent.
2. The variables are categorical.
3. Random sampling is assumed.
4. Each cell in the contingency is >5.

H0 = The presence of the medical condition and the survival outcome of the patients are independent
Ha = There is a significant association between the presence of the medical condition and the survival outcome of the patients.

For all categorical variables, except uncomplicated diabetes, complicated diabetes, peripheral vascular disease and hypothyroidism, the resulting p-value is < 0.05. This implies that, for the mentioned variables, we failed to reject the null hypothesis and that there is no significant association between the presence of these medical conditions and the survival outcome of the patients. For the rest of the variables with p-value < 0.05, we can assume that there is significant association.

Variable	Category	Data_Type	Type_of_Test	P-value
Gender	Demographic	Categorical	Chi-Square	1.9244E-07
Neurologic Dysfunction	Diagnosis	Categorical	Chi-Square	2.5955E-12
Metabolic Dysfunction	Diagnosis	Categorical	Chi-Square	7.9043E-57
Sepsis	Diagnosis	Categorical	Chi-Square	1.074E-140
Severe Respiratory Failure	Diagnosis	Categorical	Chi-Square	9.0631E-63
Severe Coagulation Failure	Diagnosis	Categorical	Chi-Square	5.1543E-23
Severe Liver Failure	Diagnosis	Categorical	Chi-Square	1.3535E-40
Any Organ Failure	Diagnosis	Categorical	Chi-Square	4.884E-136
Severe Central Nervous System Failure	Diagnosis	Categorical	Chi-Square	2.9968E-16
Severe Renal Failure	Diagnosis	Categorical	Chi-Square	7.6697E-96
Respiratory Dysfunction	Diagnosis	Categorical	Chi-Square	5.314E-115
Cardiovascular Dysfunction	Diagnosis	Categorical	Chi-Square	2.104E-169
Renal Dysfunction	Diagnosis	Categorical	Chi-Square	9.653E-122
Severe Cardiovascular Failure	Diagnosis	Categorical	Chi-Square	2.131E-204
Hematologic Dysfunction	Diagnosis	Categorical	Chi-Square	8.0735E-44
Liver Disease	Medical history	Categorical	Chi-Square	1.2689E-55
Stroke	Medical history	Categorical	Chi-Square	5.5547E-08
Chronic Heart Failure	Medical history	Categorical	Chi-Square	1.1952E-08
Hypothyroidism	Medical history	Categorical	Chi-Square	0.35780794
Peripheral Vascular Disease	Medical history	Categorical	Chi-Square	0.70785101
Metastasis	Medical history	Categorical	Chi-Square	3.6667E-26
Malignancy	Medical history	Categorical	Chi-Square	1.6983E-27
Hematologic Disease	Medical history	Categorical	Chi-Square	2.1735E-34
Uncomplicated Diabetes	Medical history	Categorical	Chi-Square	0.57953094
Complicated Hypertension	Medical history	Categorical	Chi-Square	0.00511494
Uncomplicated Hypertension	Medical history	Categorical	Chi-Square	1.0947E-13
Complicated Diabetes	Medical history	Categorical	Chi-Square	0.46633284

Visualization of Categorical Variables

Bar Graphs

The following visualization of ordinal variables provides valuable insights into the differences in proportions between individuals with medical conditions and those without relative to the patient survival outcome.

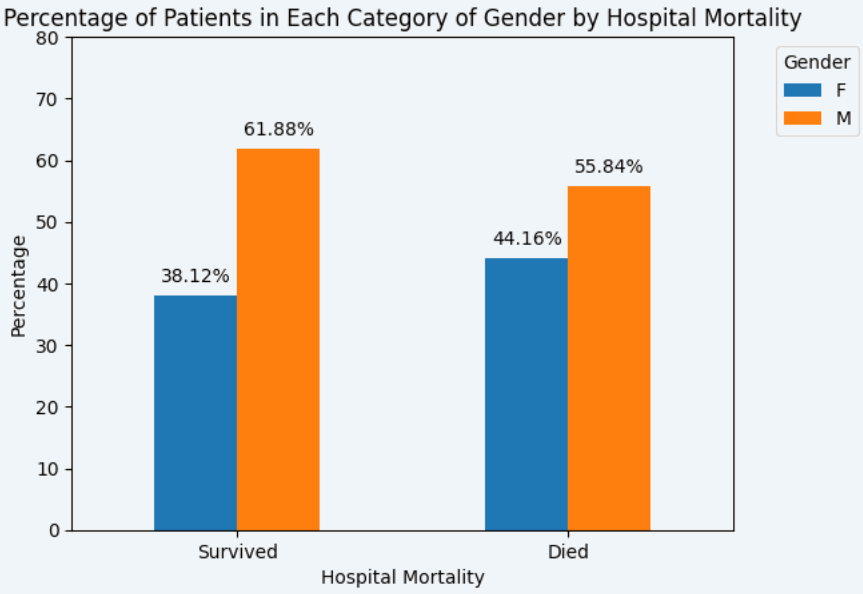
In summary, these are the critical findings we can infer from the graphs:

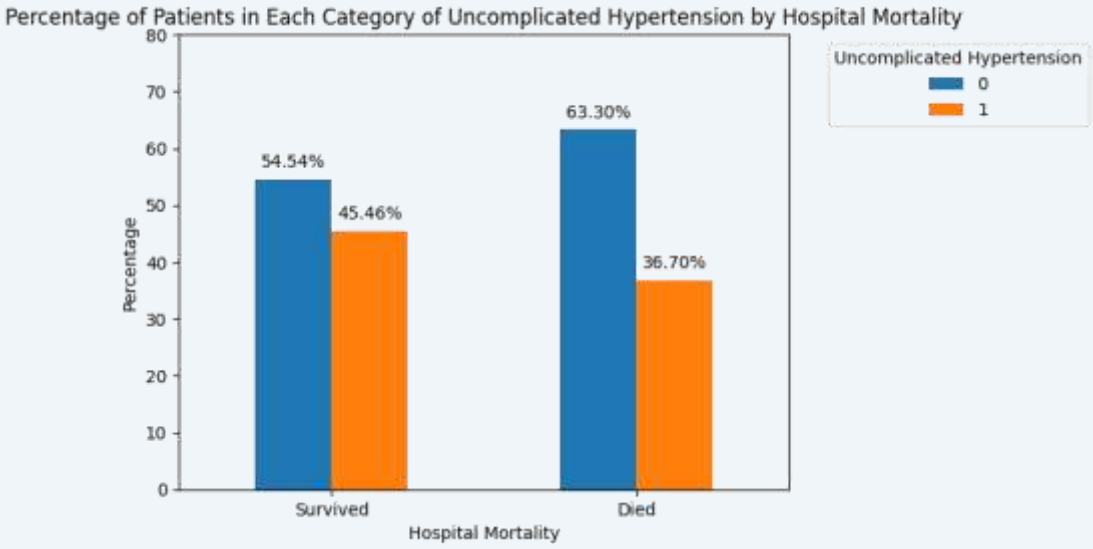
- There are more males than females in both the non-survivor and survivor groups.

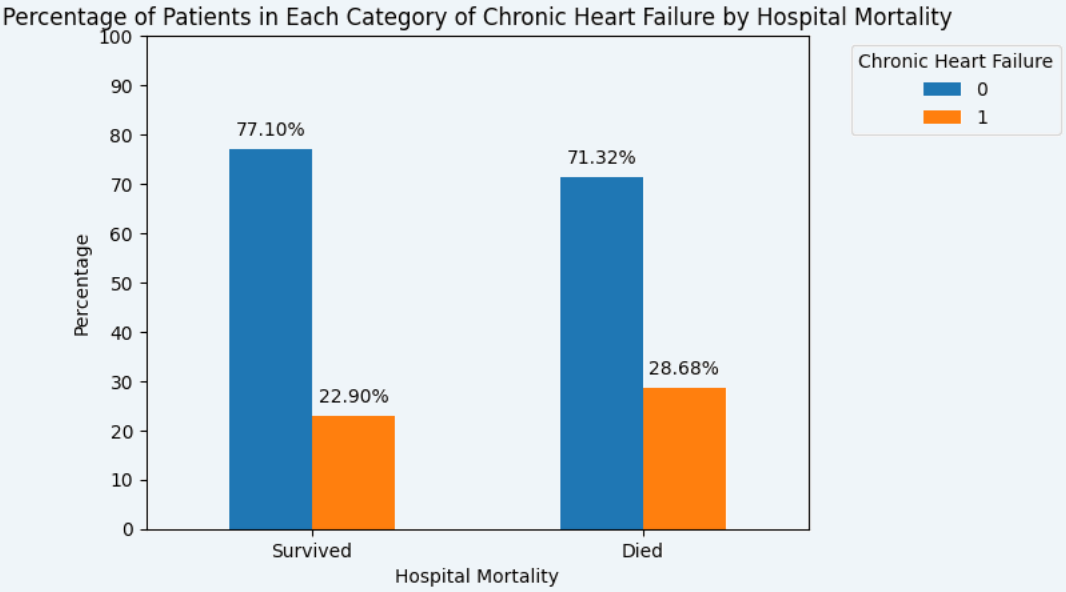
The following statistics relate to the non-survivor group:

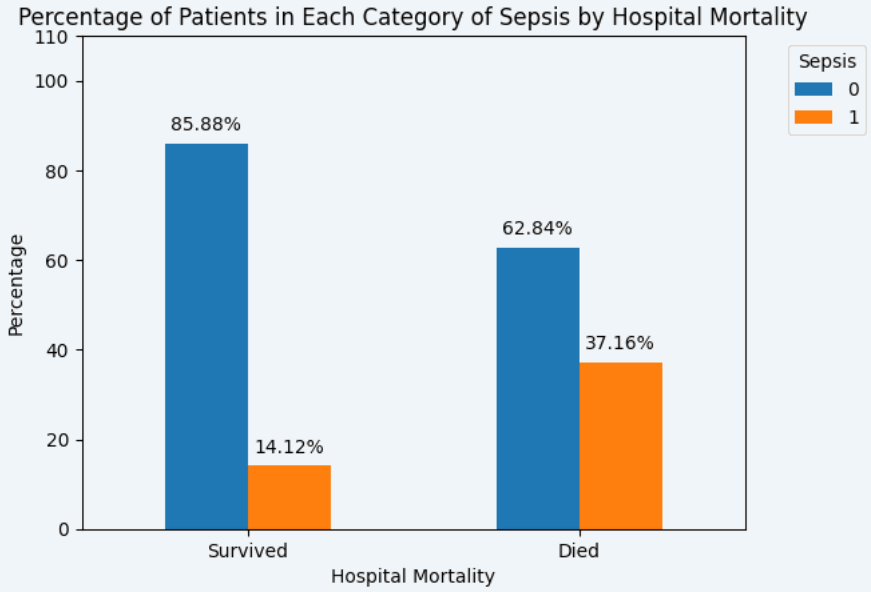
- **80%** have any organ failure
- **50%** have respiratory dysfunction
- **49%** have renal dysfunction.
- **37%** have uncomplicated hypertension
- **37%** have sepsis
- **36%** have severe cardiovascular failure
- **36%** have cardiovascular dysfunction
- **29%** have chronic heart failure

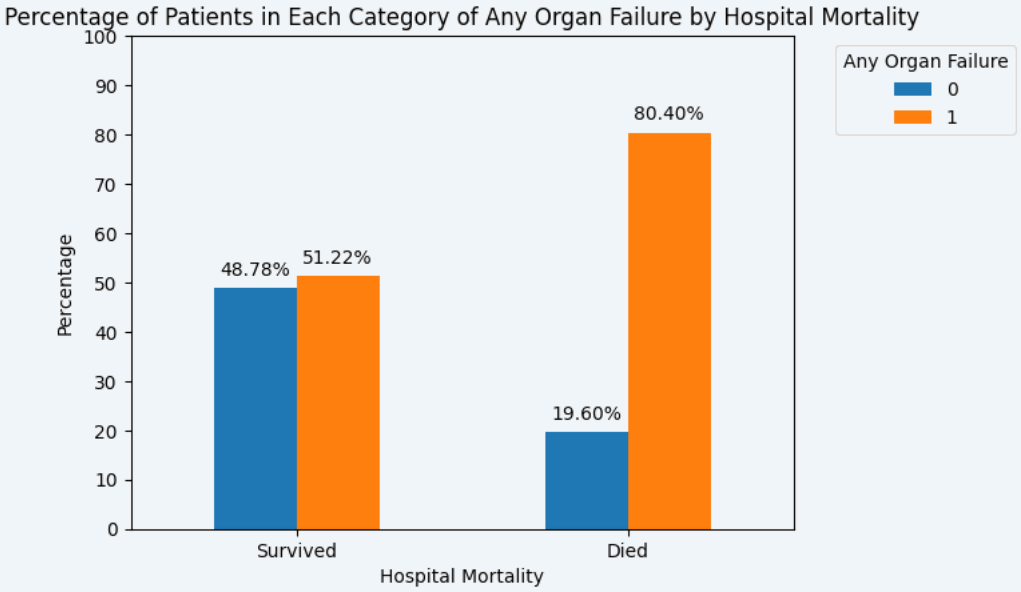
- **25%** have hematologic diseases
- **23%** have metabolic dysfunction
- **21%** have hematologic dysfunction
- **20%** have liver disease
- **19%** have uncomplicated diabetes
- **19%** have malignancy
- **17%** have severe respiratory failure

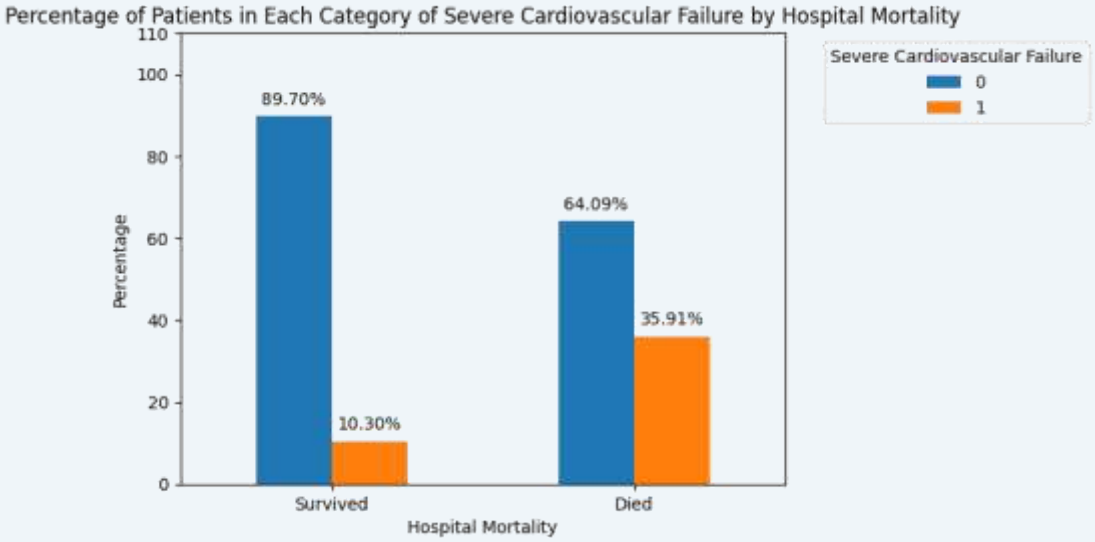




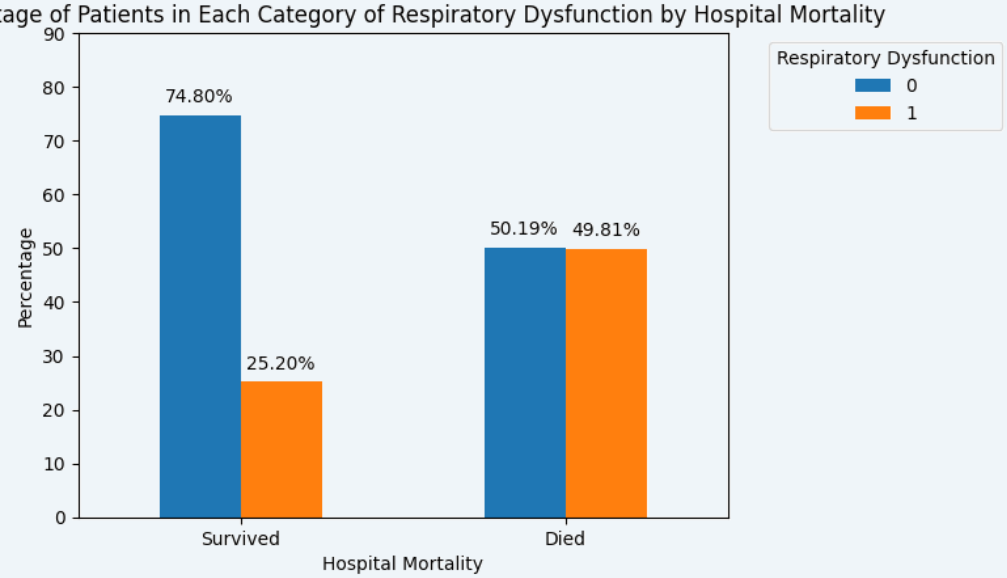

Percentage of Patients in Each Category of Chronic Heart Failure by Hospital Mortality
Chronic Heart Failure
0
1
Percentage
Hospital Mortality
Survived
Died
77.10%
22.90%
71.32%
28.68%


Percentage of Patients in Each Category of Sepsis by Hospital Mortality
Sepsis
0
1
Percentage
Hospital Mortality
Survived
Died
85.88%
14.12%
62.84%
37.16%

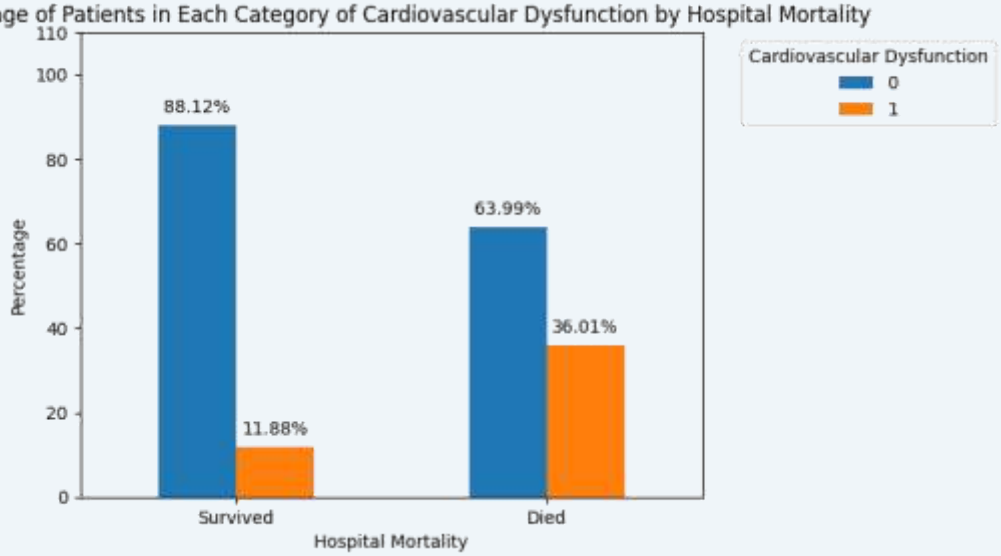

Percentage of Patients in Each Category of Any Organ Failure by Hospital Mortality
Any Organ Failure
0
1
Percentage
Hospital Mortality
Survived
Died
48.78%
51.22%
19.60%
80.40%


Percentage of Patients in Each Category of Severe Cardiovascular Failure by Hospital Mortality
Severe Cardiovascular Failure
0
1
Percentage
Hospital Mortality
Survived
Died
89.70%
10.30%
64.09%
35.91%

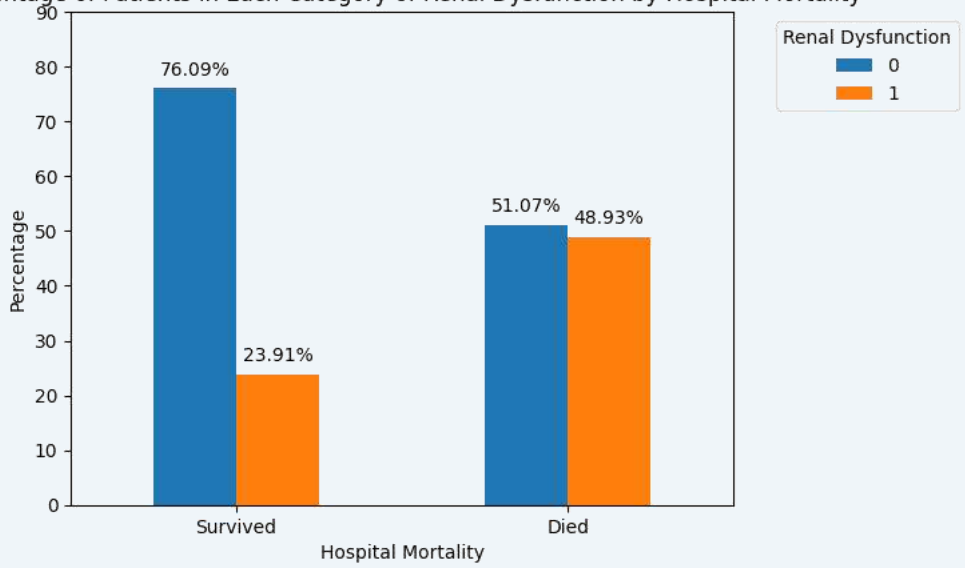
Percentage of Patients in Each Category of Respiratory Dysfunction by Hospital Mortality

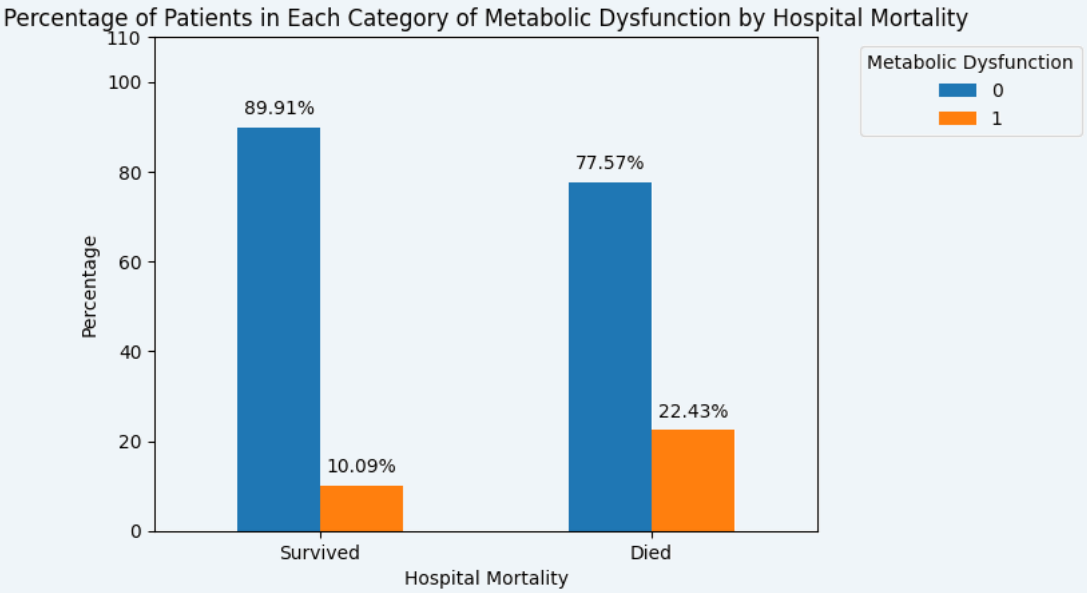


Percentage of Patients in Each Category of Cardiovascular Dysfunction by Hospital Mortality



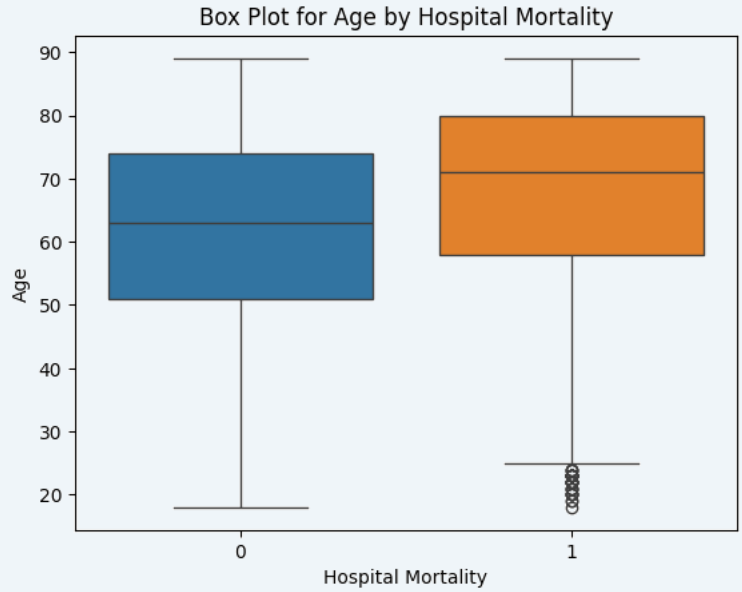
Percentage of Patients in Each Category of Renal Dysfunction by Hospital Mortality



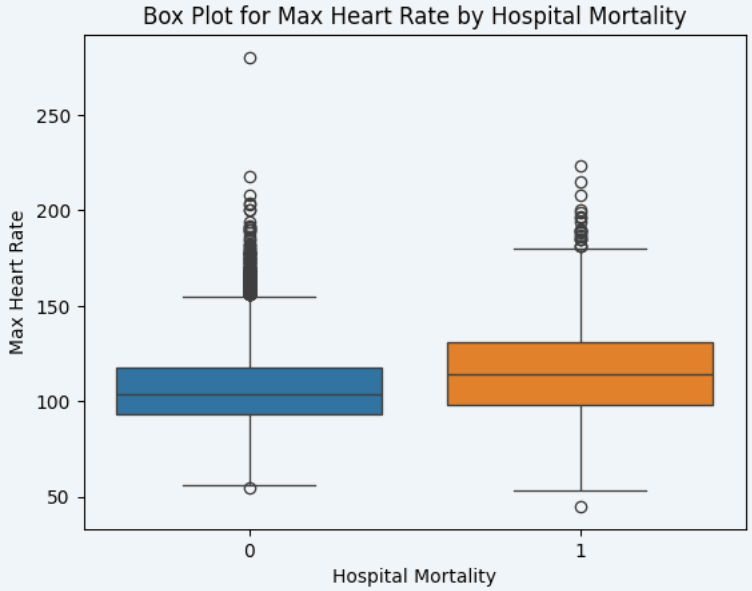

Percentage of Patients in Each Category of Metabolic Dysfunction by Hospital Mortality
Metabolic Dysfunction
0
1
Percentage
110
100
80
60
40
20
0
89.91%
10.09%
77.57%
22.43%
Survived
Died
Hospital Mortality

Box Plots

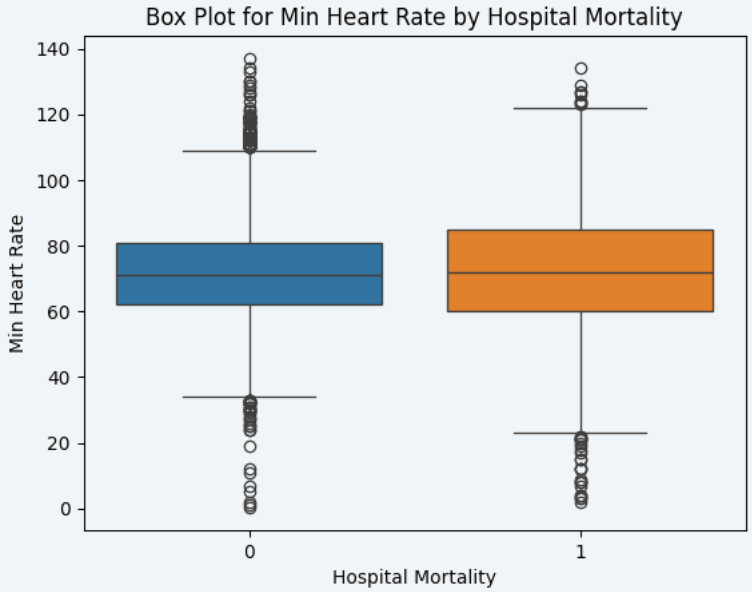
The median age is higher among non-survivors compared to survivors, with similar variability observed in both groups.



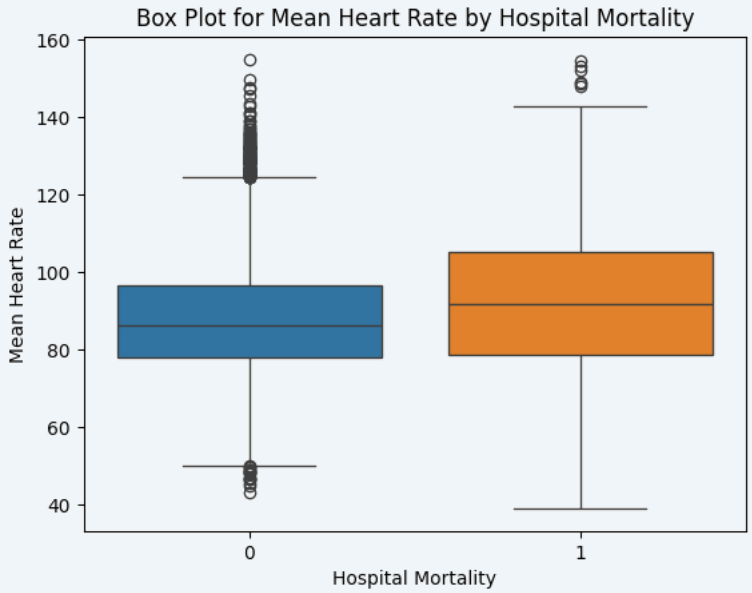
Non-survivors exhibit a slightly higher maximum heart rate compared to survivors. In the survivor group, data points are more concentrated around the center, while outliers contribute to the right skewed distribution observed in both groups.



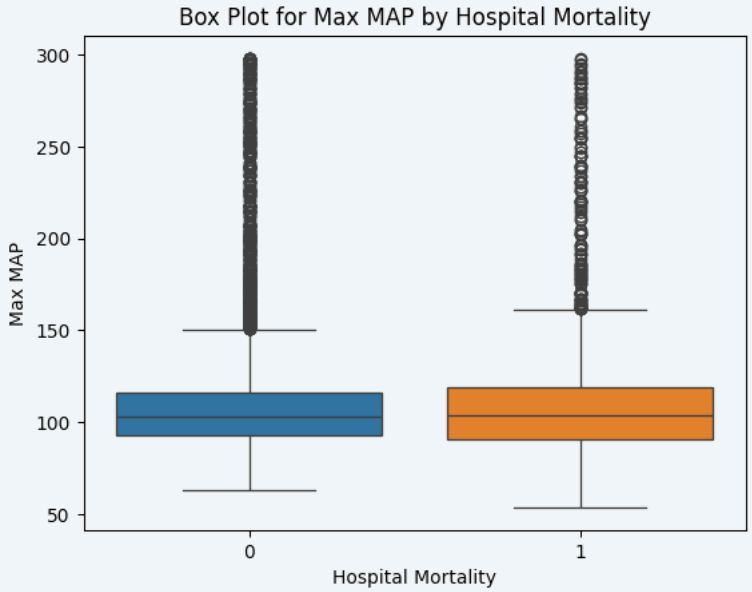
The median of the minimum heart rate is nearly identical for the two groups. The box plot illustrates a tighter concentration of data points around the median for the survivor group, resulting in a much narrower shape.



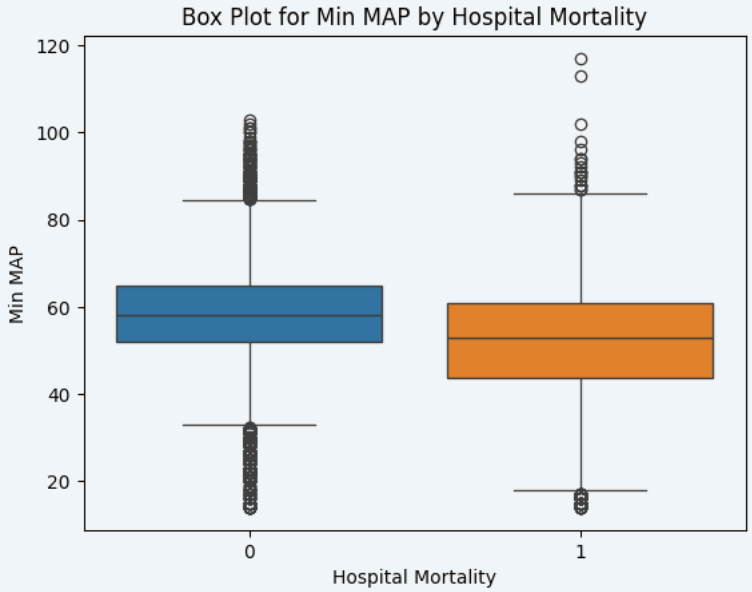
The box plot illustrates a slightly higher mean heart rate in the non-survivor group. Outliers are contributing to a right-skewed shape in the survivor group, while the non-survivor group exhibits a more normal distribution.



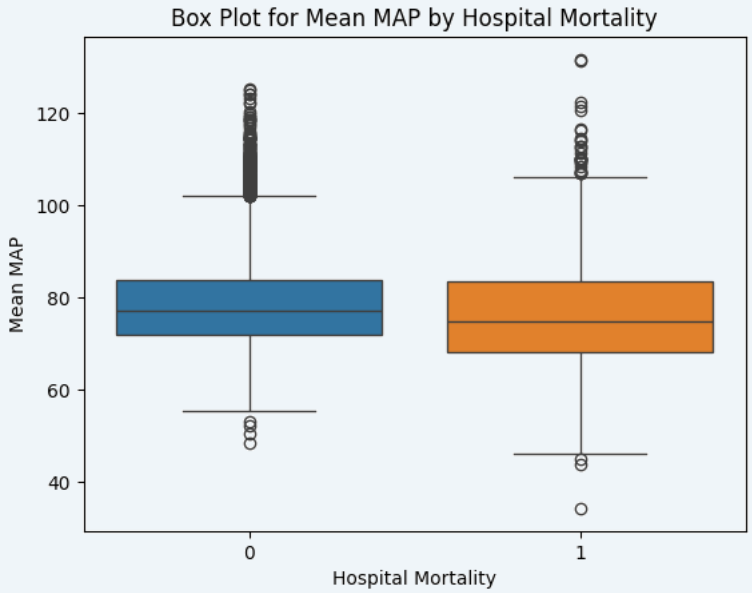
The medians of the maximum Mean Arterial Pressure (MAP) in the two groups are identical, but the shape of the distribution in both groups is highly influenced by outliers, causing it to become skewed.



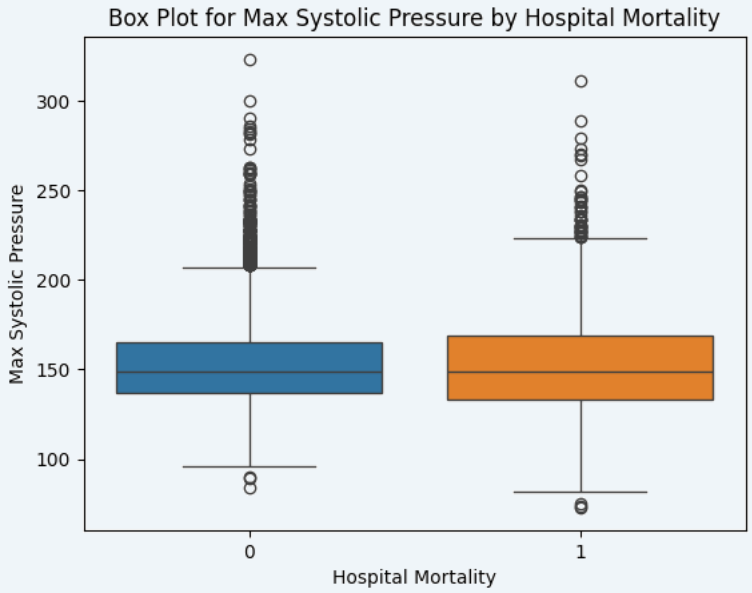
The median of min MAP is higher, and the data points are more concentrated around the median for the survivor group.



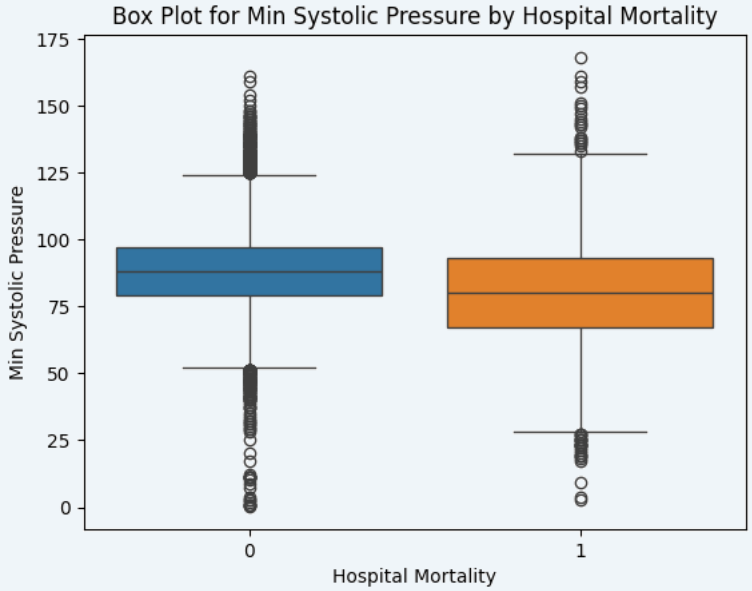
The variability is higher in the non-survivor group, but the median in both groups is almost the same.



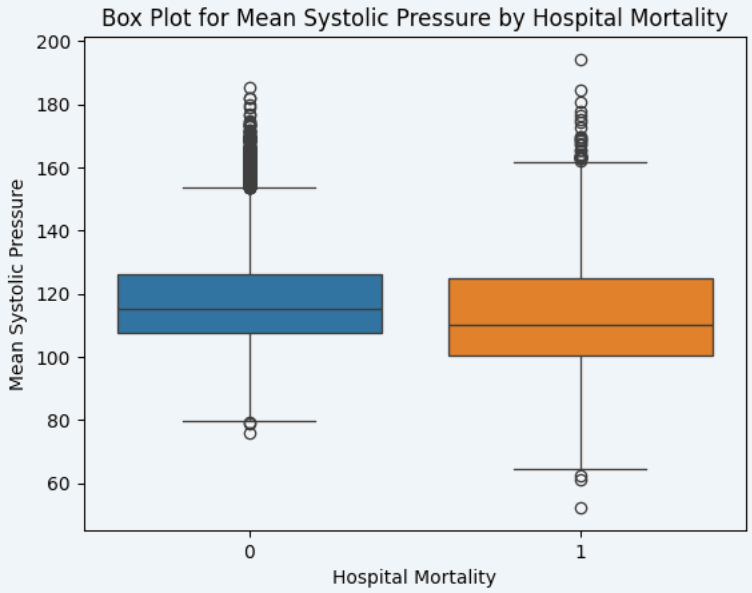
The variability is lower in the survivor group for max systolic pressure and the medial on both groups are almost the same.



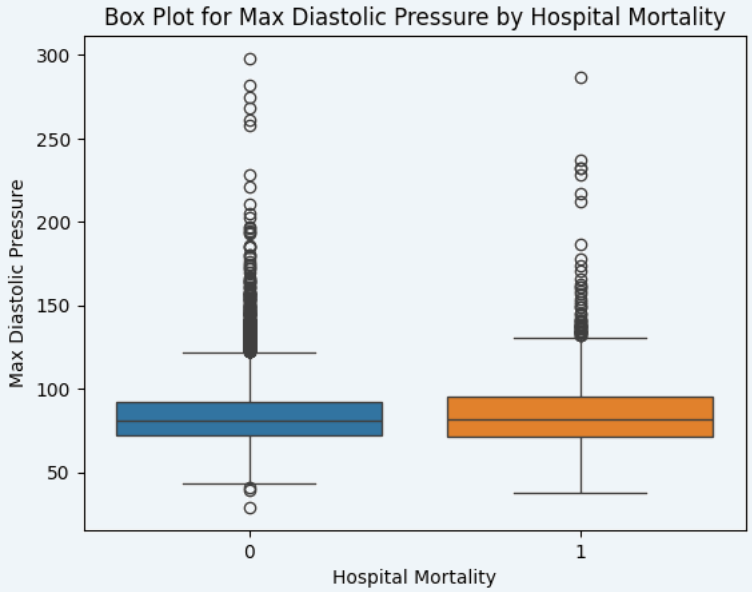
The min systolic pressure is higher in the survivor group and the datapoints are more concentrated around the median.



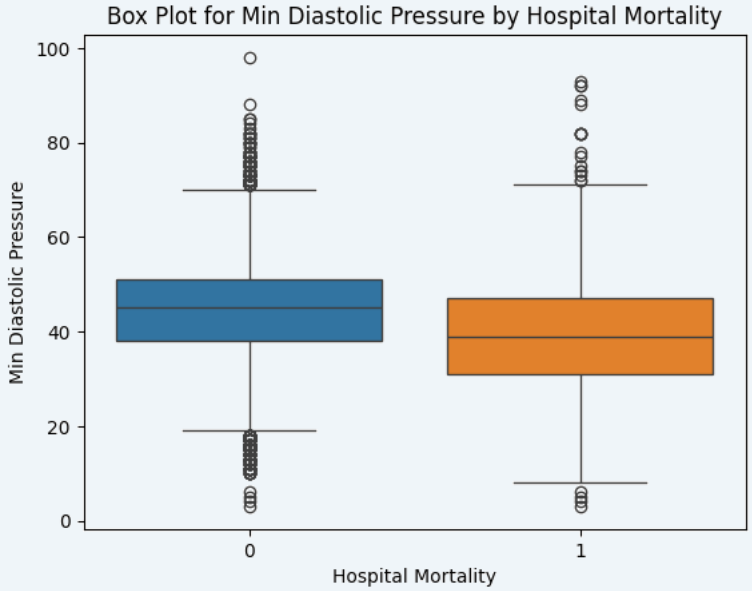
The variability is more in the non-survivor group and the median is less compared to the survivor group.



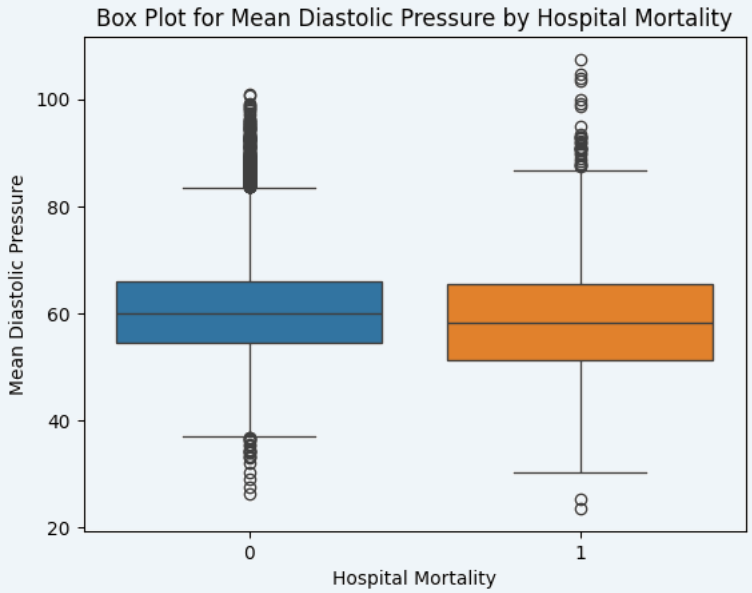
The median of max diastolic pressure in both groups are almost the same but the data points are more concentrated in the center in the survivor group. The outliers cause the shape to be right skewed in both groups.



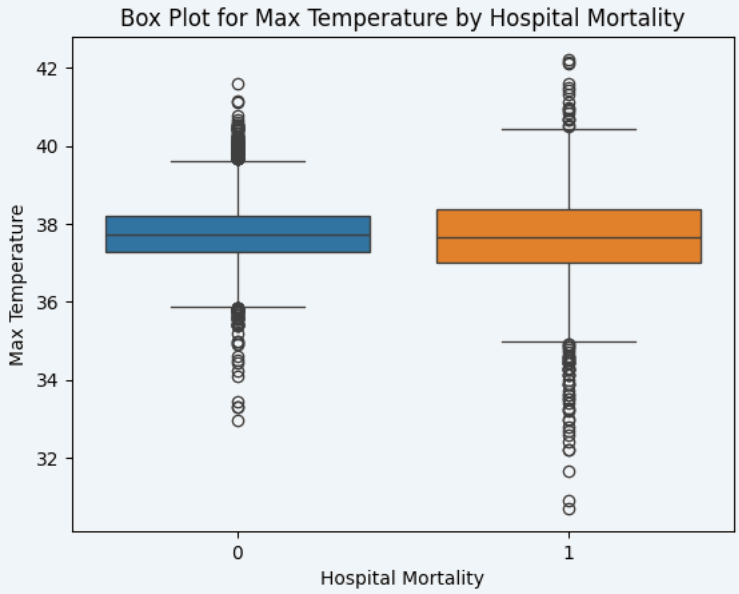
The min diastolic pressure is lower in the non-survivor group. The data points are more concentrated around the center in the survivor group. Outliers are causing the distribution of the non-survivor group to be right skewed. The survivor group data appears normally distributed.



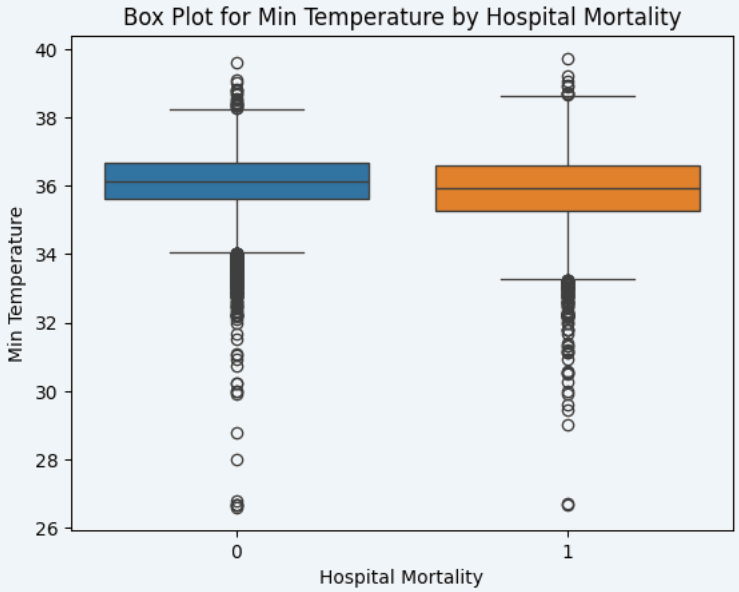
The variability in the survivor is less compared to the non-survivor group while the median in both groups is nearly identical.



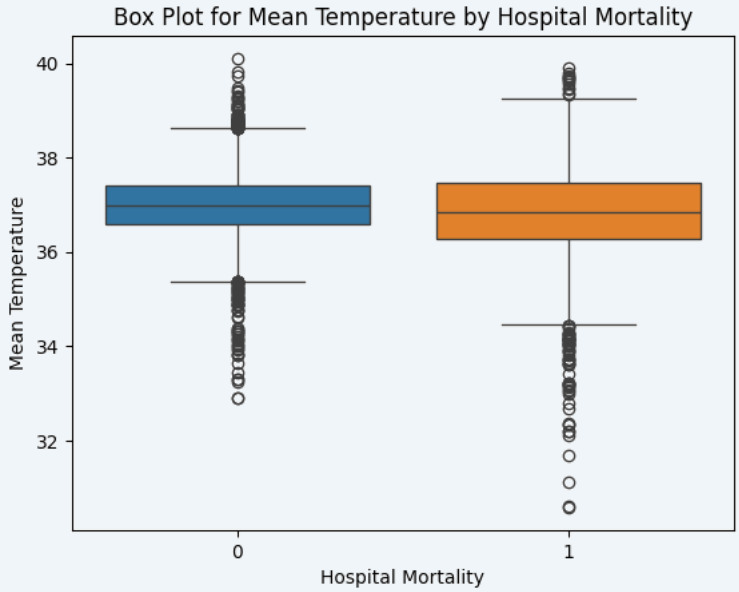
The variability in the non-survivor group is greater, while the median in both groups is identical. Outliers are influencing the distribution shape of the non-survivor data to become left skewed.



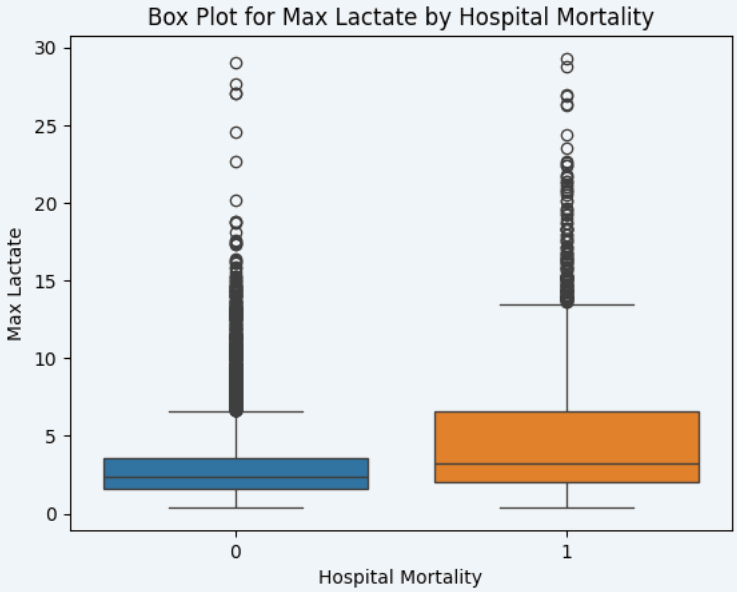
The outliers are causing the shape of the distribution in both groups to be left skewed.



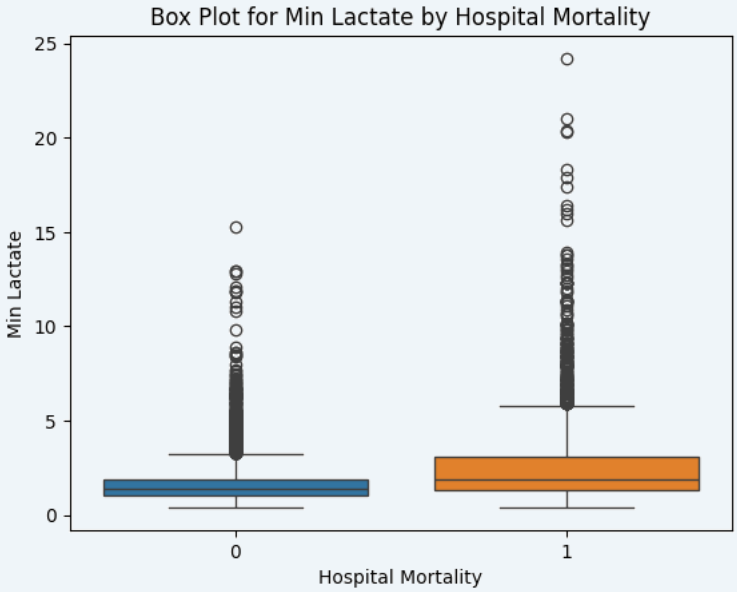
The outliers in the mean temperature are causing the distribution shape of non-survivors to become left skewed.



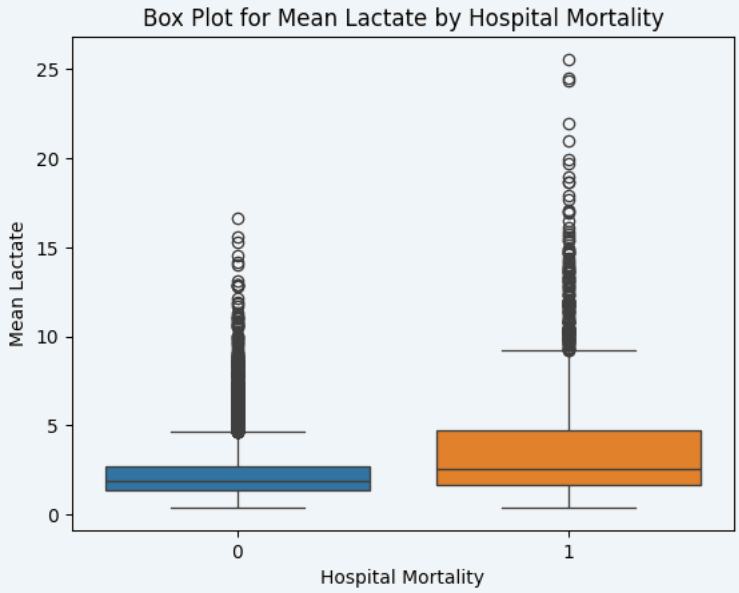
The datapoints in the survivor group is concentrated tightly around the center. The outliers are causing the distribution to be right skewed in both groups.



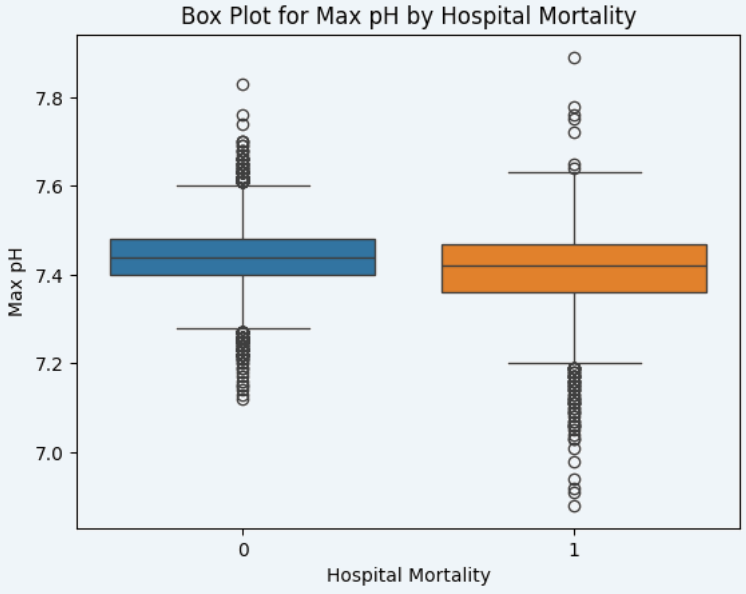
The data points in the survivor group are much more concentrated around the median in the survivor group. The variability in the non-survivor group is higher.



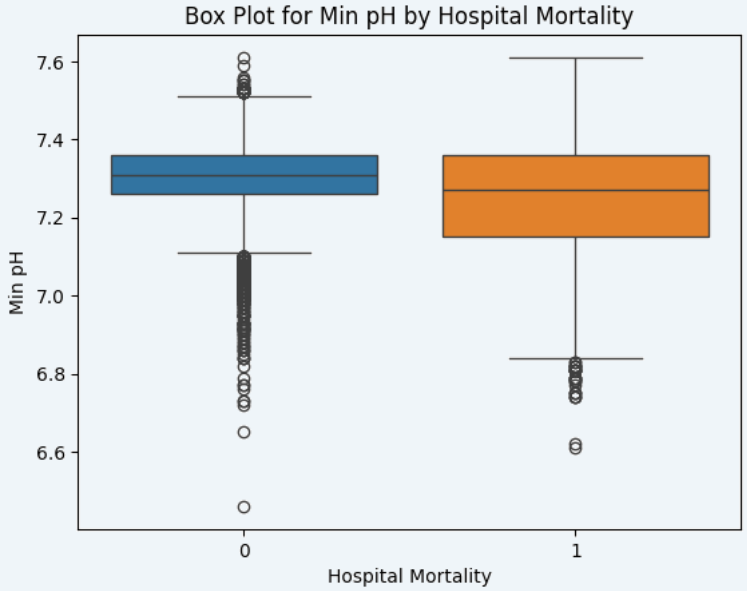
We can say the same observations as the min lactate.



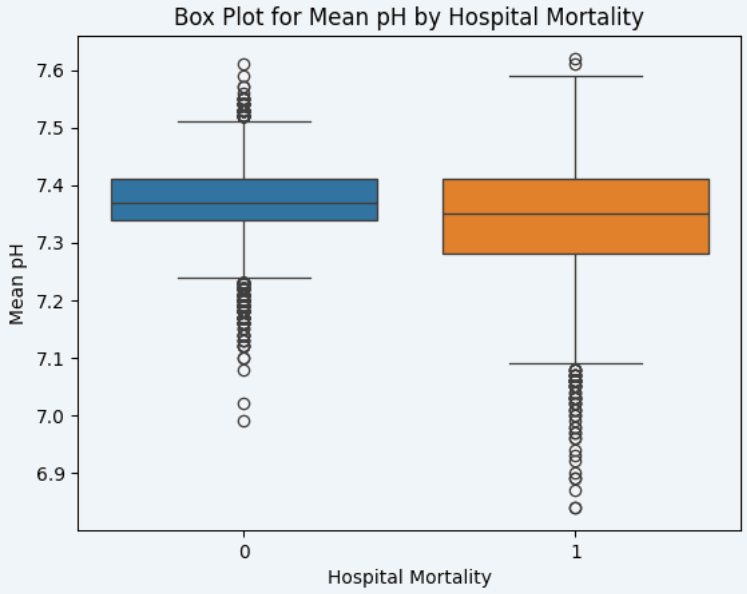
The outliers are causing the distribution to be skewed in both groups.



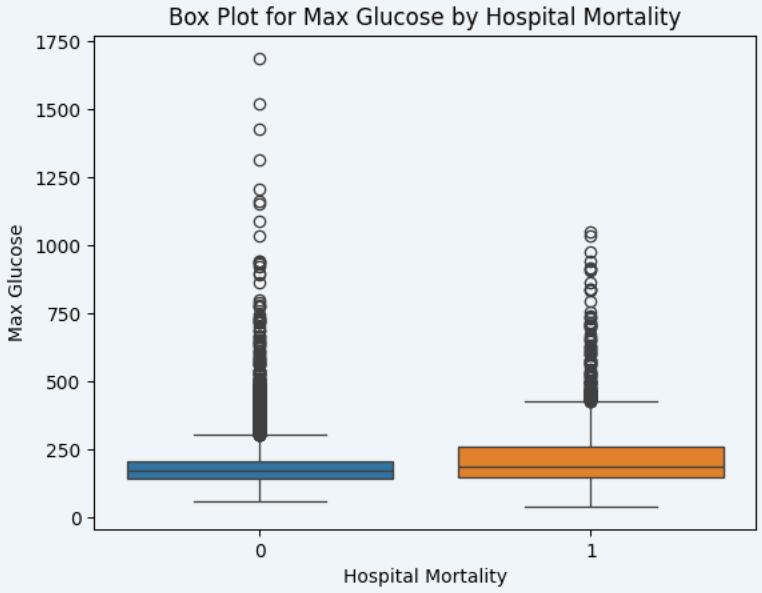
The outliers are causing the distribution shape to become left skewed in both groups.



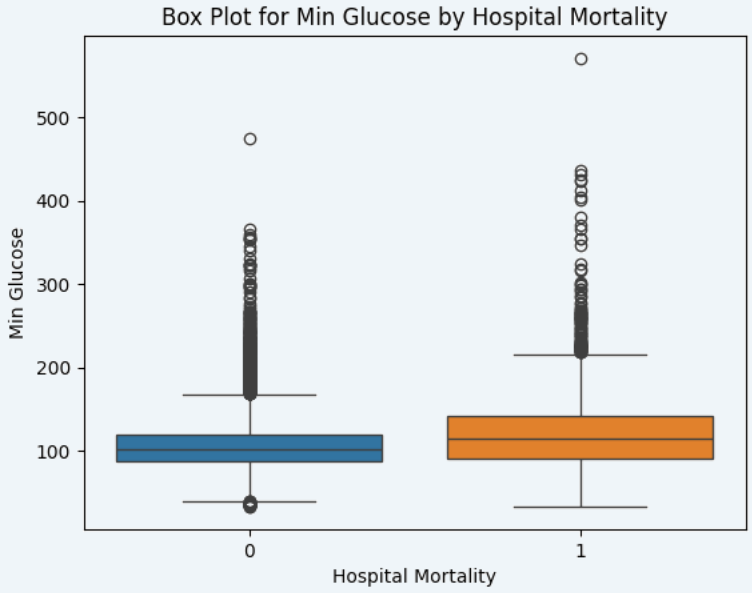
Both groups have a left skewed distribution because of the presence of outliers.



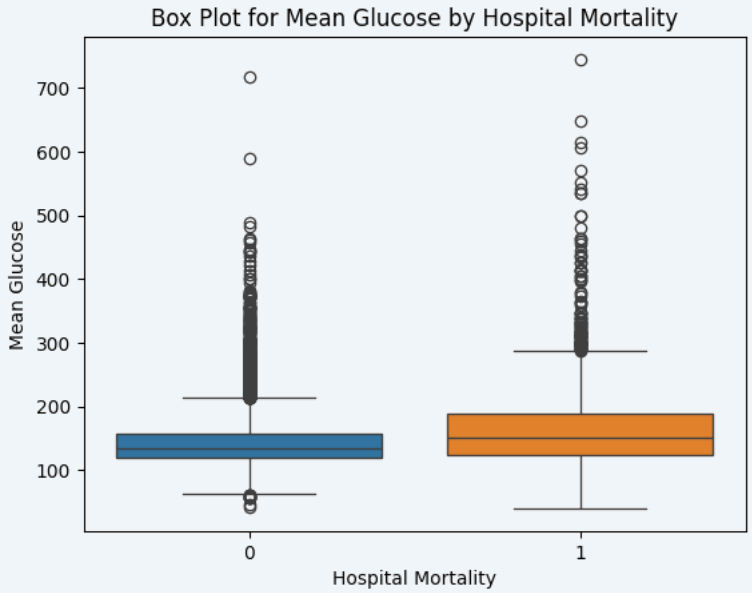
The variability is higher in the survivor group.



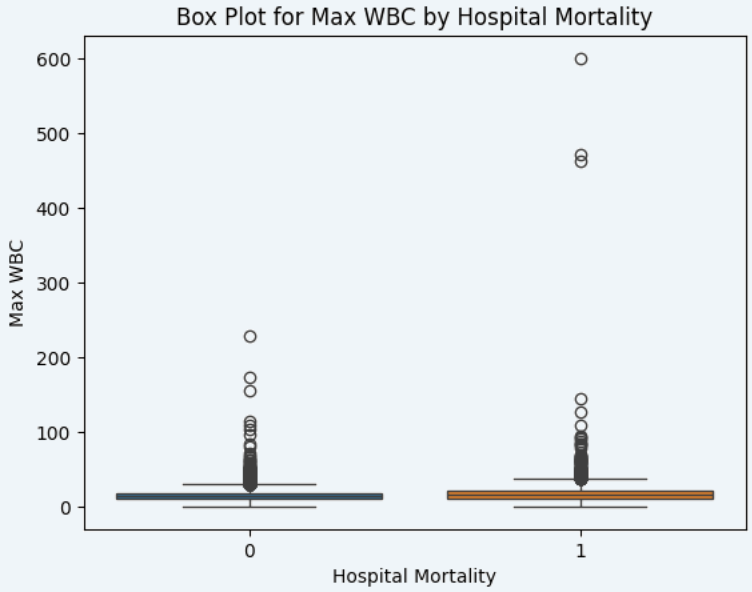
The distribution shape of the two groups is influenced by outliers.

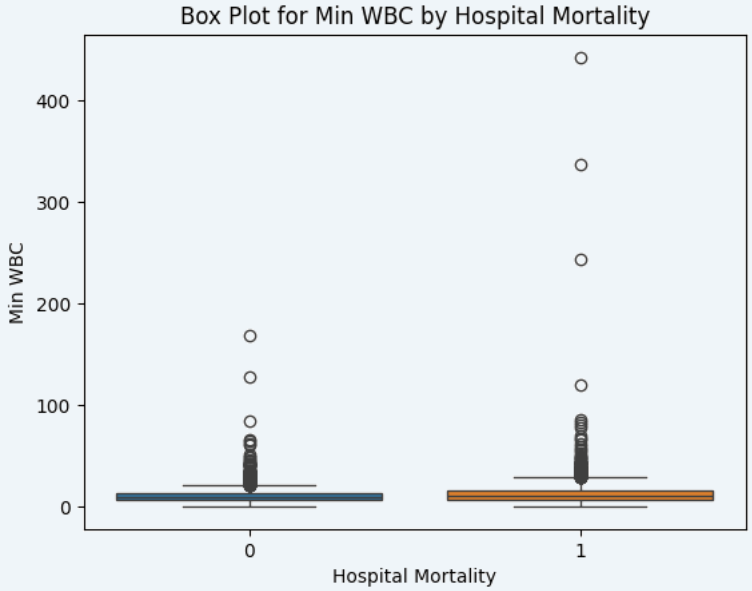


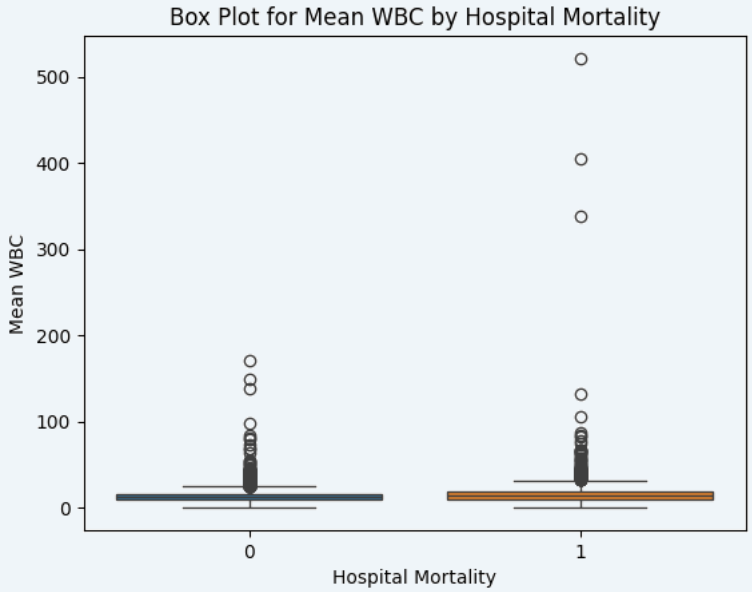
The distribution shape of the two groups is influenced by outliers.



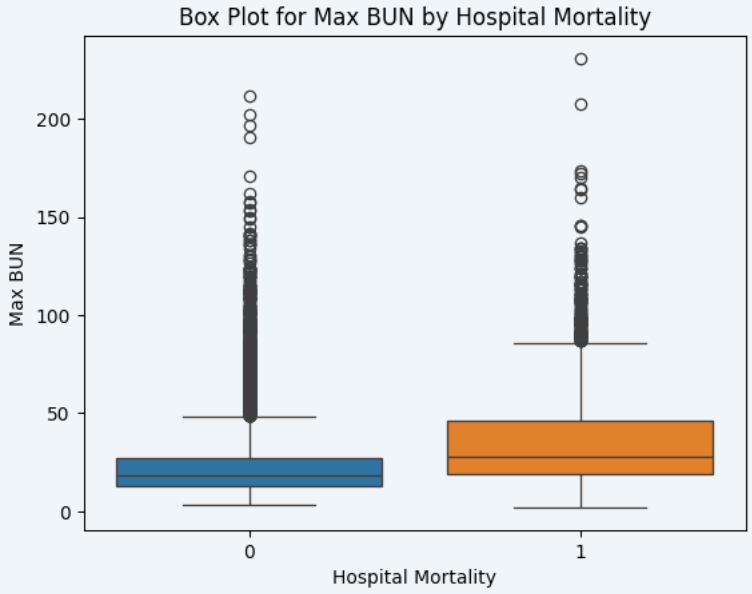
With respect to WBC, the distribution shape is greatly influenced by the outliers in the data of the two groups.

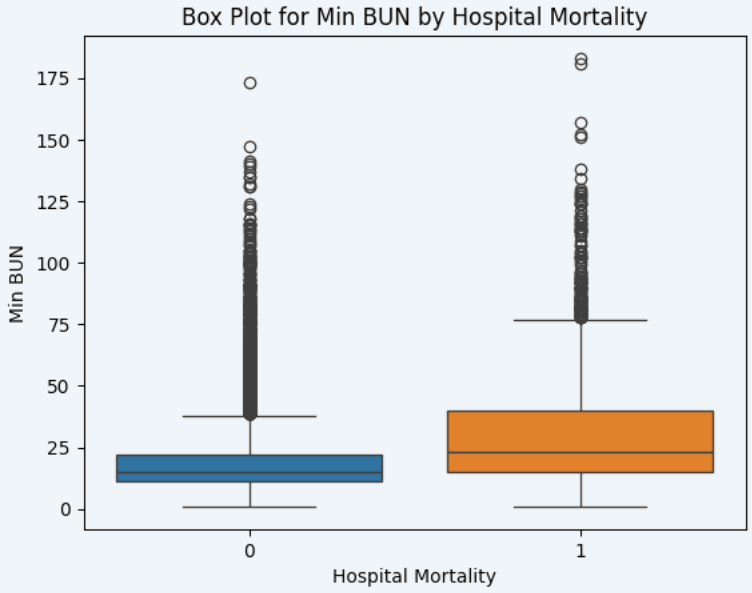

Box Plot for Max WBC by Hospital Mortality
Max WBC
Hospital Mortality
0
1
0
100
200
300
400
500
600


Box Plot for Min WBC by Hospital Mortality
Min WBC
Hospital Mortality
0
1
0
100
200
300
400

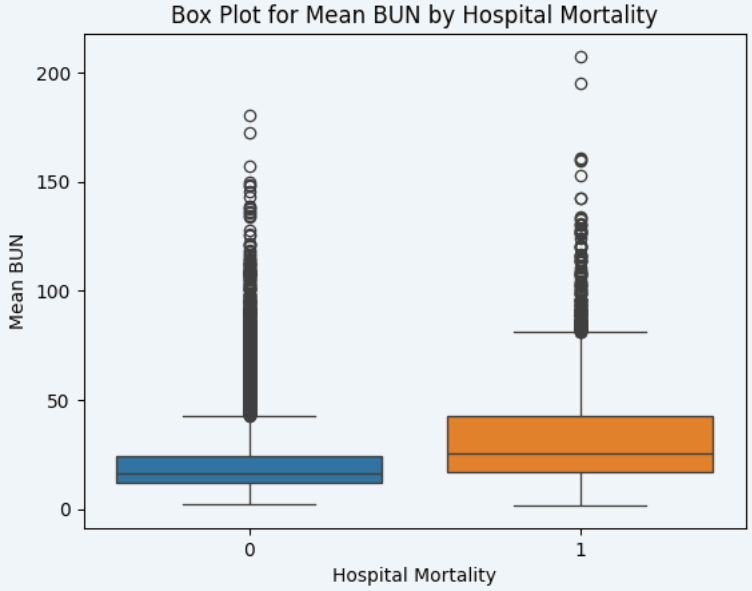

Box Plot for Mean WBC by Hospital Mortality
Mean WBC
Hospital Mortality
0
1
0
100
200
300
400
500

The data points of the max BUN in the survivor group are tightly concentrated around the center. In both groups, the outliers are causing the distribution shape to become right skewed.



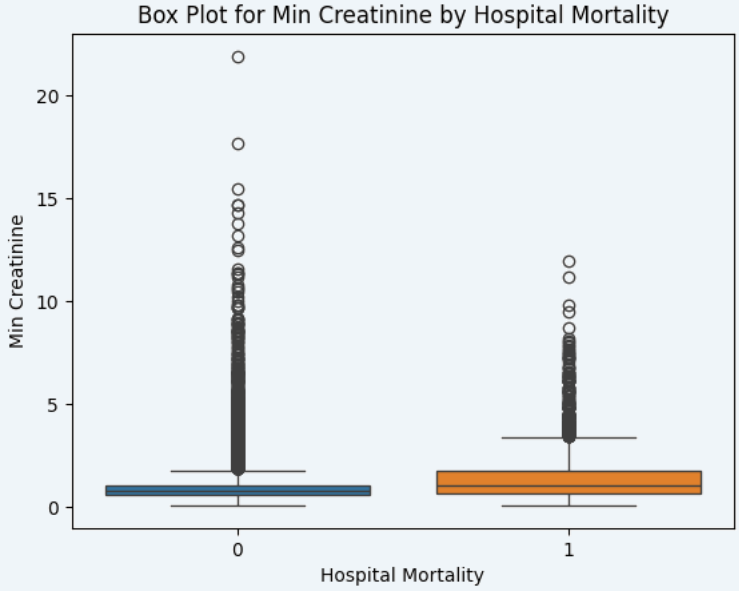


The data points of the mean BUN in the survivor group are tightly concentrated around the center. In both groups, the outliers are causing the distribution shape to become right skewed.



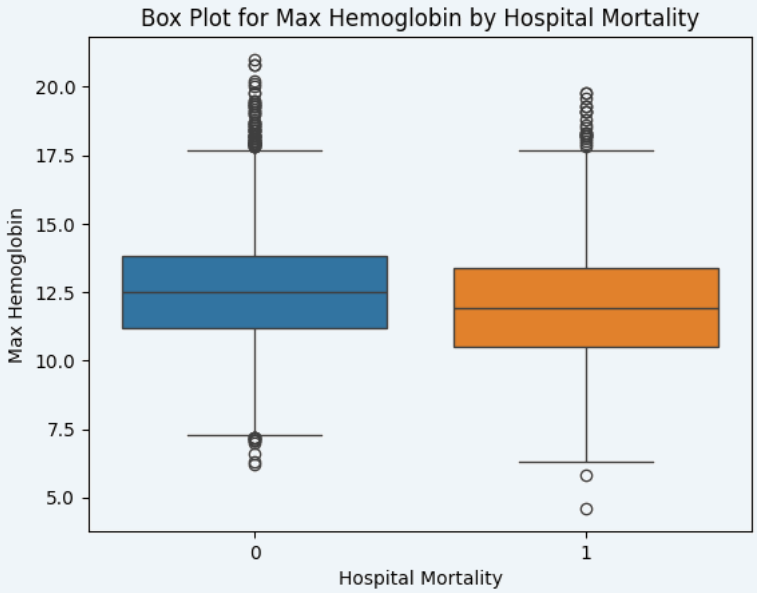
The variability of max creatinine is high in the non-survivor group, but it is even higher in the survivor group. The shapes of both groups are highly positively skewed.

The variability of min creatinine is high in the non-survivor group, but it is even higher in the survivor group. The shapes of both groups are highly positively skewed.

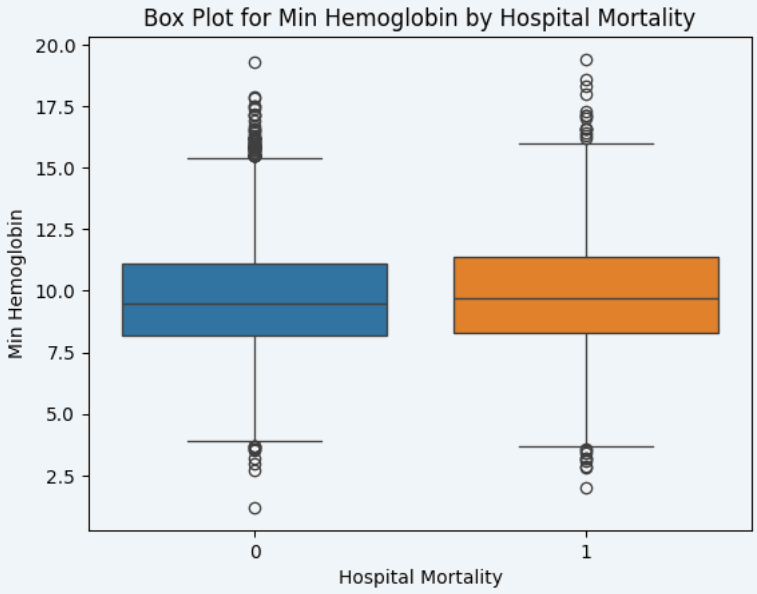


The variability of mean creatinine is high in the non-survivor group, but it is even higher in the survivor group. The shapes of both groups are highly positively skewed.

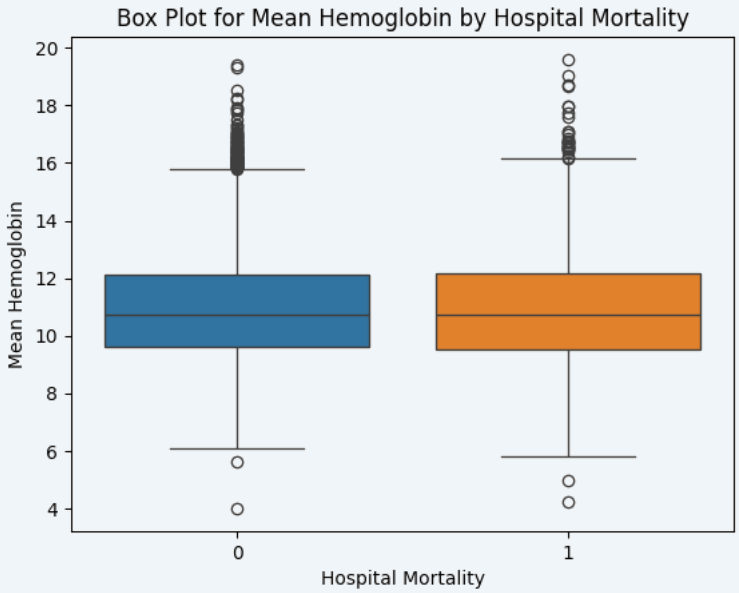
The median of max hemoglobin for non-survivor group is slightly lower compared to the survivor group.



The median of min hemoglobin for non-survivor group is slightly higher compared to the survivor group.



The shape of the distribution for mean hemoglobin is similar is both groups.



Continuous Variables

After the Log Transformation the variables Max Heart Rate, Mean Heart Rate, Mean MAP, Mean Systolic Pressure, Mean Diastolic Pressure, Mean BUN, Max Haemoglobin, Mean Haemoglobin were identified to have a normal distribution.
We performed the Levene Test for homogeneity to identify the suitable parametric test that should be performed on these variables.

Levene Test for homogeneity

We performed Levene test to assess the equality of variances between the two groups of Hospital Mortality in all the Continuous variables. It is a statistical test used to determine whether the variances of two or more groups are statistically significantly different from each other.
This was tested to determine whether it is appropriate to proceed with parametric statistical tests that assume equal variances across groups, or whether alternative methods that do not assume equal variances should be used.
Note that all assumptions listed below are met.

- The observations within each group or category are assumed to be independent.
- The variables being analysed are quantitative.

H0 = The population variances are equal among the two groups of mortality
Ha = The population variances are not equal among the two groups of mortality
$\alpha = 0.05$
Based on the summary table below all the variables have a p – values less that set $\alpha = 0.05$. Hence, we rejected the null hypothesis, concluding that in none of the variables the population variances are not equal among the two groups of mortality.

Variable	Levene Statistic	P - value	Homogeneity
Age	9.766476	$1.78e^{-03}$	No
Max Heart Rate	228.955	$2.87e^{-51}$	No
Min Heart Rate	381.173	$1.21e^{-83}$	No
Mean Heart Rate	242.539	$3.55e^{-54}$	No
Max MAP	51.128	$9.14e^{-13}$	No
Min MAP	280.885	$2.31e^{-62}$	No
Mean MAP	128.331	$1.32e^{-29}$	No
Max Systolic Pressure	151.111	$1.57e^{-34}$	No
Min Systolic Pressure	309.132	$2.24e^{-68}$	No
Mean Systolic Pressure	225.881	$1.31e^{-50}$	No
Max Diastolic Pressure	67.658	$2.14e^{-16}$	No
Min Diastolic Pressure	167.035	$5.76e^{-38}$	No
Mean Diastolic Pressure	77.282	$1.67e^{-18}$	No
Max Temperature	586.364	$1.26e^{-126}$	No

Min Temperature	329.538	$1.04e^{-72}$	No
Mean Temperature	644.301	$1.23e^{-138}$	No
Max Lactate	850.402	$6.38e^{-181}$	No
Min Lactate	820.759	$7.00e^{-175}$	No
Mean Lactate	1041.35	$1.66e^{-219}$	No
Max pH	364.313	$4.45e^{-80}$	No
Min pH	940.224	$3.84e^{-199}$	No
Mean pH	1061.326	$1.64e^{-223}$	No
Max Glucose	241.999	$4.63e^{-54}$	No
Min Glucose	361.134	$2.10e^{-79}$	No
Mean Glucose	393.565	$2.93e^{-86}$	No
Max WBC	146.193	$1.81e^{-33}$	No
Min WBC	224.203	$2.99e^{-50}$	No
Mean WBC	190.961	$4.07e^{-43}$	No
Max BUN	373.597	$4.84e^{-82}$	No
Min BUN	453.03	$8.89e^{-09}$	No
Mean BUN	417.703	$2.39e^{-91}$	No
Max Creatinine	149.299	$3.86e^{-34}$	No
Min Creatinine	167.901	$3.74e^{-38}$	No
Mean Creatinine	158.743	$3.53e^{-36}$	No
Max Hemoglobin	59.074	$1.63e^{-14}$	No
Min Hemoglobin	15.581	$7.95e^{-05}$	No
Mean Hemoglobin	39.784	$2.93e^{-10}$	No

Parametric Tests

Since we identified from the Levene Test that none of variables have equal variances among the two groups of mortality, we decided to perform the Welch's t-test as it is used to compare the means of two independent groups when the assumption of equal population variances is violated.

Welch's t-test

We decided to use the Welch's t-test on the variables Max Heart Rate, Mean Heart Rate, Mean MAP, Mean Systolic Pressure, Mean Diastolic Pressure, Mean BUN, Max Haemoglobin, Mean Haemoglobin as they followed a normal distribution closely after the Log Transformation.
Note the data after Log Transformation were used in performing Welch's t Test.
The below assumptions were met by the variables.

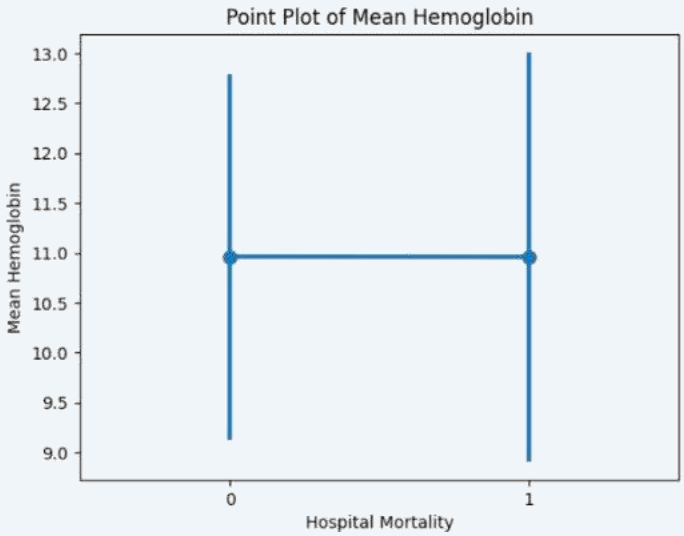
- The observations within each group are assumed to be independent.
- The data for each group should be approximately normally distributed.

H0 = The means for the two groups of mortality are equal.
Ha = The means for the two groups of mortality are not equal.
$\alpha = 0.05$

Based on the summary table of results below we can see the p values of Max Heart Rate, Mean Heart Rate, Mean MAP, Mean Systolic Pressure, Mean Diastolic Pressure, Mean BUN and Mean Hemoglobin are less than 0.05 level, concluding the difference between survivors and non-survivors for each variable is statistically significant.

However, in the variable Mean Haemoglobin the p value is greater than 0.05 level, where we accept the null hypothesis concluding the means of survivors and non-survivors are equal.

Variable	T-statistic	P-value	Conclusion
Max Heart Rate	-14.42535	$1.5137e^{-45}$	The difference is statistically significant.
Mean Heart Rate	-9.820534	$2.11329e^{-22}$	The difference is statistically significant.
Mean MAP	9.96346	$5.40936e^{-23}$	The difference is statistically significant.
Mean Systolic Pressure	11.31721	$4.935026e^{-29}$	The difference is statistically significant.
Mean Diastolic Pressure	8.72391	$4.528457e^{-18}$	The difference is statistically significant.
Mean BUN	-27.185729	$4.778132e^{-145}$	The difference is statistically significant.
Max Haemoglobin	9.4272588	$8.44463e^{-21}$	The difference is statistically significant.
Mean Haemoglobin	0.8239296	0.410047	There is no significant difference



The above test results can also be verified by the point plot which shows the means of two groups of mortality are same.

Non-Parametric Test

Mann-Whitney U Test

Since the shape of the distribution for the remaining continuous variables are not normal, we decided to use a non-parametric test called the Mann-Whitney on the dataset before log transformation. Note that all assumptions listed below are met.

Assumptions:

1. Observations in each group are independent of each other.
2. Random sampling is assumed.
3. The variables are ordinal or interval.
4. Shapes of the distributions in the two groups are similar.

H0 = The distributions of the living group are equal to the distributions of the deceased group.
Ha = There is a difference between the distributions of the two groups.

Based on the summary table only Min Heart Rate, Max MAP, Max Systolic Pressure and Max Diastolic Pressure have a p-value of > 0.05. For the said variables, we can assume that the distributions of the living group are equal to the distributions of the deceased group. We can also refer to the density plots in the Appendix to confirm this assumption. For the rest of the variables, results show that there is a difference between the distributions of the two groups.

Variable	Category	Data_Type	Type_of_Test	P-value
Age	Demographic	Continuous	Mann-Whitney U	2.15253E-66
Max WBC	Laboratory results	Continuous	Mann-Whitney U	1.11179E-13
Min WBC	Laboratory results	Continuous	Mann-Whitney U	2.28103E-15
Mean WBC	Laboratory results	Continuous	Mann-Whitney U	3.78164E-15
Max BUN	Laboratory results	Continuous	Mann-Whitney U	2.0237E-170
Min Creatinine	Laboratory results	Continuous	Mann-Whitney U	3.9911E-94
Mean Creatinine	Laboratory results	Continuous	Mann-Whitney U	4.6758E-110
Min Hemoglobin	Laboratory results	Continuous	Mann-Whitney U	0.000440385
Mean Glucose	Laboratory results	Continuous	Mann-Whitney U	2.59408E-57
Min BUN	Laboratory results	Continuous	Mann-Whitney U	1.7652E-161
Min Glucose	Laboratory results	Continuous	Mann-Whitney U	2.28413E-38
Max Creatinine	Laboratory results	Continuous	Mann-Whitney U	6.4357E-118
Mean pH	Laboratory results	Continuous	Mann-Whitney U	2.20523E-47
Max Glucose	Laboratory results	Continuous	Mann-Whitney U	3.59421E-31
Max Lactate	Laboratory results	Continuous	Mann-Whitney U	6.17776E-84
Mean Lactate	Laboratory results	Continuous	Mann-Whitney U	1.6949E-114
Max pH	Laboratory results	Continuous	Mann-Whitney U	4.31198E-29
Min pH	Laboratory results	Continuous	Mann-Whitney U	2.1447E-47
Min Lactate	Laboratory results	Continuous	Mann-Whitney U	1.4045E-153
OASIS	Severity	Ordinal	Mann-Whitney U	6.7278E-308
SOFA	Severity	Ordinal	Mann-Whitney U	3.5648E-178
SAPS II	Severity	Ordinal	Mann-Whitney U	0
Mean Temperature	Vital signs	Continuous	Mann-Whitney U	1.31255E-15
Min Temperature	Vital signs	Continuous	Mann-Whitney U	4.34424E-17
Min MAP	Vital signs	Continuous	Mann-Whitney U	5.90274E-94
Min Heart Rate	Vital signs	Continuous	Mann-Whitney U	0.640581755
Max MAP	Vital signs	Continuous	Mann-Whitney U	0.74015476
Max Systolic Pressure	Vital signs	Continuous	Mann-Whitney U	0.099642427
Min Systolic Pressure	Vital signs	Continuous	Mann-Whitney U	1.35005E-91
Min Diastolic Pressure	Vital signs	Continuous	Mann-Whitney U	3.89039E-87
Max Diastolic Pressure	Vital signs	Continuous	Mann-Whitney U	0.073008961
Max Temperature	Vital signs	Continuous	Mann-Whitney U	1.13908E-05

Comparison of Mann Whitney & Welch T test

We performed the Mann Whitney U test for the same variables (Max Heart Rate, Mean Heart Rate, Mean MAP, Mean Systolic Pressure, Mean Diastolic Pressure, Mean BUN, Max Haemoglobin, Mean Haemoglobin) that Welch T test was performed, to compare the results from two different tests. The null and alternative hypothesis were same as above with the $\alpha = 0.05$.

Considering the below results obtained from the Mann-Whitney test, we can observe that it leads to the similar conclusion derived from the Welch t-test discussed earlier. For the Mean Haemoglobin the means of survivors and non -survivors are equal, whereas for the other variables the means are statistically significant between survivors and non-survivors.

Variable	U-statistic	P-value	Conclusion
Max Heart Rate	8640638.5	$7.52e^{-61}$	The difference is statistically significant.
Mean Heart Rate	9409815	$3.95e^{-30}$	The difference is statistically significant.
Mean MAP	12733640.5	$2.12e^{-25}$	The difference is statistically significant.
Mean Systolic Pressure	13073048.5	$1.22e^{-36}$	The difference is statistically significant.
Mean Diastolic Pressure	12469046	$4.03e^{-18}$	The difference is statistically significant.
Mean BUN	6923118	$2.92e^{-169}$	The difference is statistically significant.
Max Haemoglobin	12656776	$3.71e^{-23}$	The difference is statistically significant.
Mean Haemoglobin	11249336	0.502333204	There is no significant difference

Spearman Correlation

The below correlation matrix illustrates the pair of variables per group that exhibit a moderate to strong positive or negative correlation according to Spearman correlation coefficient.

Spearman Correlation Assumption:

1. The variables are either ordinal or continuous data that follow a monotonic relationship.

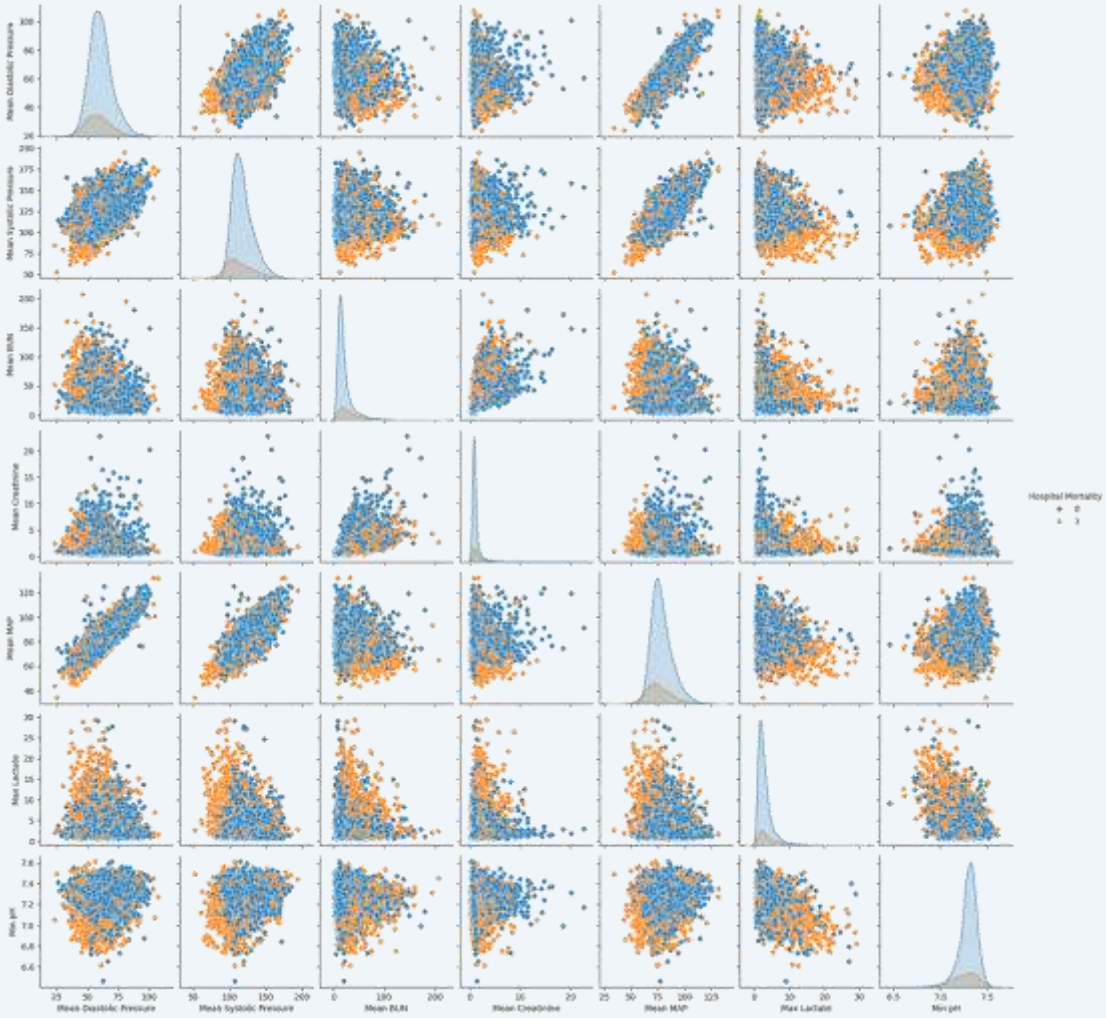
H0: There is no monotonic association between the two variables.
Ha: There is a monotonic association between the two variables.

With the significance level at 0.05, the results are statistically significant.

	Variable_1	Variable_2	Test	Correlation	P_Value	Group
1	Mean MAP	Mean Diastolic Pressure	Spearman	0.8582	0.000000e+00	Survivors
3	Mean MAP	Mean Diastolic Pressure	Spearman	0.8569	0.000000e+00	Non-Survivors
5	Min Diastolic Pressure	Min MAP	Spearman	0.8189	0.000000e+00	Non-Survivors
7	Min Diastolic Pressure	Min MAP	Spearman	0.8148	0.000000e+00	Survivors
9	Min Systolic Pressure	Min MAP	Spearman	0.7987	0.000000e+00	Non-Survivors
11	Mean MAP	Mean Systolic Pressure	Spearman	0.7556	0.000000e+00	Non-Survivors
13	Max MAP	Max Diastolic Pressure	Spearman	0.7452	0.000000e+00	Non-Survivors
15	Max Diastolic Pressure	Max MAP	Spearman	0.7362	0.000000e+00	Survivors
17	Mean Creatinine	Max BUN	Spearman	0.7351	0.000000e+00	Non-Survivors
19	Mean Creatinine	Mean BUN	Spearman	0.7328	0.000000e+00	Non-Survivors
21	Min Creatinine	Min BUN	Spearman	0.7320	0.000000e+00	Non-Survivors
23	Mean BUN	Min Creatinine	Spearman	0.7314	0.000000e+00	Non-Survivors
25	Max Creatinine	Max BUN	Spearman	0.7298	0.000000e+00	Non-Survivors
27	Min Creatinine	Max BUN	Spearman	0.7186	0.000000e+00	Non-Survivors
29	Mean BUN	Mean Creatinine	Spearman	0.7161	0.000000e+00	Survivors
31	Max Creatinine	Mean BUN	Spearman	0.7145	0.000000e+00	Non-Survivors
33	Min MAP	Min Systolic Pressure	Spearman	0.7143	0.000000e+00	Survivors
35	Mean Creatinine	Min BUN	Spearman	0.7138	0.000000e+00	Non-Survivors
37	Max BUN	Max Creatinine	Spearman	0.7094	0.000000e+00	Survivors
39	Min Creatinine	Min BUN	Spearman	0.7048	0.000000e+00	Survivors
41	Max BUN	Mean Creatinine	Spearman	0.7046	0.000000e+00	Survivors
43	Mean MAP	Mean Systolic Pressure	Spearman	0.7038	0.000000e+00	Survivors
45	Min BUN	Mean Creatinine	Spearman	0.7038	0.000000e+00	Survivors
47	Max Creatinine	Mean BUN	Spearman	0.7032	0.000000e+00	Survivors
49	Mean BUN	Min Creatinine	Spearman	0.6957	0.000000e+00	Survivors
51	Min BUN	Max Creatinine	Spearman	0.6833	8.287704e-297	Non-Survivors
53	Max Systolic Pressure	Max MAP	Spearman	0.6739	9.774152e-286	Non-Survivors
55	Max Creatinine	Min BUN	Spearman	0.6734	0.000000e+00	Survivors
57	Max BUN	Min Creatinine	Spearman	0.6680	0.000000e+00	Survivors
59	Max MAP	Max Systolic Pressure	Spearman	0.6467	0.000000e+00	Survivors
61	Min Diastolic Pressure	Min Systolic Pressure	Spearman	0.6334	2.256757e-242	Non-Survivors
63	Min Diastolic Pressure	Mean MAP	Spearman	0.6235	0.000000e+00	Survivors
65	Max Systolic Pressure	Mean MAP	Spearman	0.6105	1.279377e-220	Non-Survivors
67	Mean Diastolic Pressure	Min MAP	Spearman	0.6095	0.000000e+00	Survivors
69	Max Diastolic Pressure	Mean MAP	Spearman	0.6088	0.000000e+00	Survivors
71	Min Diastolic Pressure	Mean MAP	Spearman	0.6041	7.549115e-215	Non-Survivors
73	Max Diastolic Pressure	Mean MAP	Spearman	0.5847	4.401237e-198	Non-Survivors
75	Max MAP	Mean Diastolic Pressure	Spearman	0.5596	4.030514e-178	Non-Survivors
77	Max Diastolic Pressure	Max Systolic Pressure	Spearman	0.5512	8.346475e-172	Non-Survivors
79	Mean Systolic Pressure	Min MAP	Spearman	0.5493	2.152481e-170	Non-Survivors
81	Min Diastolic Pressure	Min Systolic Pressure	Spearman	0.5186	0.000000e+00	Survivors
83	Mean MAP	Min Systolic Pressure	Spearman	0.5106	0.000000e+00	Survivors
85	Mean Diastolic Pressure	Min MAP	Spearman	0.5080	6.528197e-142	Non-Survivors
87	Min Systolic Pressure	Mean MAP	Spearman	0.5078	8.496490e-142	Non-Survivors
89	Max Lactate	Min pH	Spearman	-0.5302	9.810846e-157	Non-Survivors

Visualization of Continuous Variables

Scatter Plot



Density Plots

Please refer to the Appendix.

Results Summary Table

Summary table for statistical tests conducted.

Variable	Category	Data_Type	Type_of_Test	P-value
Gender	Demographic	Categorical	Chi-Square	1.92E-07
Age	Demographic	Continuous	Mann-Whitney U	2.15E-66
Neurologic Dysfunction	Diagnosis	Categorical	Chi-Square	2.6E-12
Metabolic Dysfunction	Diagnosis	Categorical	Chi-Square	7.9E-57
Sepsis	Diagnosis	Categorical	Chi-Square	1.1E-140
Severe Respiratory Failure	Diagnosis	Categorical	Chi-Square	9.06E-63
Severe Coagulation Failure	Diagnosis	Categorical	Chi-Square	5.15E-23
Severe Liver Failure	Diagnosis	Categorical	Chi-Square	1.35E-40
Any Organ Failure	Diagnosis	Categorical	Chi-Square	4.9E-136
Severe Central Nervous System Failure	Diagnosis	Categorical	Chi-Square	3E-16
Severe Renal Failure	Diagnosis	Categorical	Chi-Square	7.67E-96
Respiratory Dysfunction	Diagnosis	Categorical	Chi-Square	5.3E-115
Cardiovascular Dysfunction	Diagnosis	Categorical	Chi-Square	2.1E-169
Renal Dysfunction	Diagnosis	Categorical	Chi-Square	9.7E-122
Severe Cardiovascular Failure	Diagnosis	Categorical	Chi-Square	2.1E-204
Hematologic Dysfunction	Diagnosis	Categorical	Chi-Square	8.07E-44
Max WBC	Laboratory results	Continuous	Mann-Whitney U	1.11E-13
Min WBC	Laboratory results	Continuous	Mann-Whitney U	2.28E-15
Mean WBC	Laboratory results	Continuous	Mann-Whitney U	3.78E-15
Max BUN	Laboratory results	Continuous	Mann-Whitney U	2E-170
Min Creatinine	Laboratory results	Continuous	Mann-Whitney U	3.99E-94
Mean Creatinine	Laboratory results	Continuous	Mann-Whitney U	4.7E-110
Min Hemoglobin	Laboratory results	Continuous	Mann-Whitney U	0.00044
Mean Glucose	Laboratory results	Continuous	Mann-Whitney U	2.59E-57
Min BUN	Laboratory results	Continuous	Mann-Whitney U	1.8E-161
Min Glucose	Laboratory results	Continuous	Mann-Whitney U	2.28E-38
Max Creatinine	Laboratory results	Continuous	Mann-Whitney U	6.4E-118
Mean pH	Laboratory results	Continuous	Mann-Whitney U	2.21E-47
Max Glucose	Laboratory results	Continuous	Mann-Whitney U	3.59E-31
Max Hemoglobin	Laboratory results	Continuous	Welch's T-test	8.44E-21
Mean BUN	Laboratory results	Continuous	Welch's T-test	4.8E-145
Max Lactate	Laboratory results	Continuous	Mann-Whitney U	6.18E-84
Mean Hemoglobin	Laboratory results	Continuous	Welch's T-test	0.410047
Mean Lactate	Laboratory results	Continuous	Mann-Whitney U	1.7E-114
Max pH	Laboratory results	Continuous	Mann-Whitney U	4.31E-29
Min pH	Laboratory results	Continuous	Mann-Whitney U	2.14E-47
Min Lactate	Laboratory results	Continuous	Mann-Whitney U	1.4E-153
Liver Disease	Medical history	Categorical	Chi-Square	1.27E-55
Stroke	Medical history	Categorical	Chi-Square	5.55E-08
Chronic Heart Failure	Medical history	Categorical	Chi-Square	1.2E-08

Hypothyroidism	Medical history	Categorical	Chi-Square	0.357808
Peripheral Vascular Disease	Medical history	Categorical	Chi-Square	0.707851
Metastasis	Medical history	Categorical	Chi-Square	3.67E-26
Malignancy	Medical history	Categorical	Chi-Square	1.7E-27
Hematologic Disease	Medical history	Categorical	Chi-Square	2.17E-34
Uncomplicated Diabetes	Medical history	Categorical	Chi-Square	0.579531
Complicated Hypertension	Medical history	Categorical	Chi-Square	0.005115
Uncomplicated Hypertension	Medical history	Categorical	Chi-Square	1.09E-13
Complicated Diabetes	Medical history	Categorical	Chi-Square	0.466333
OASIS	Severity	Ordinal	Mann-Whitney U	6.7E-308
SOFA	Severity	Ordinal	Mann-Whitney U	3.6E-178
SAPS II	Severity	Ordinal	Mann-Whitney U	0
Mean Heart Rate	Vital signs	Continuous	Welch's T-test	2.11E-22
Mean MAP	Vital signs	Continuous	Welch's T-test	5.41E-23
Mean Systolic Pressure	Vital signs	Continuous	Welch's T-test	4.94E-29
Mean Diastolic Pressure	Vital signs	Continuous	Welch's T-test	4.53E-18
Mean Temperature	Vital signs	Continuous	Mann-Whitney U	1.31E-15
Min Temperature	Vital signs	Continuous	Mann-Whitney U	4.34E-17
Min MAP	Vital signs	Continuous	Mann-Whitney U	5.9E-94
Min Heart Rate	Vital signs	Continuous	Mann-Whitney U	0.640582
Max MAP	Vital signs	Continuous	Mann-Whitney U	0.740155
Max Systolic Pressure	Vital signs	Continuous	Mann-Whitney U	0.099642
Min Systolic Pressure	Vital signs	Continuous	Mann-Whitney U	1.35E-91
Min Diastolic Pressure	Vital signs	Continuous	Mann-Whitney U	3.89E-87
Max Diastolic Pressure	Vital signs	Continuous	Mann-Whitney U	0.073009
Max Temperature	Vital signs	Continuous	Mann-Whitney U	1.14E-05
Max Heart Rate	Vital signs	Continuous	Welch's T-test	1.51E-45

Literature Review

Feature Explanations in Recurrent Neural Networks for Predicting Risk of Mortality in Intensive Care Patients

Summary:
A recent study has proposed a framework that uses a Recurrent Neural Network (RNN) architecture to predict the risk of mortality in patients receiving intensive care. This is achieved by processing a time series of vital signs and laboratory results, which helps to produce a predicted mortality risk score. Additionally, the model identifies the contribution of each input feature to the prediction. The framework has been evaluated using three critical care databases - MIMIC-III, MIMIC-IV, and eICU - which yielded consistent results, establishing the robustness of the approach. The model also explains its predictions, which can aid clinical staff in understanding mortality-related factors.
How does this article help the report?
The article discusses using datasets, particularly Mimic iii, to predict mortality risk. It also highlights the importance of the vital signs and laboratory variables. In Table 2, the authors indicate the valid range for each variable in both categories. This information can be used to identify values that fall outside the valid range in our Mimic iii datasets. With this information, we can remove values outside the valid range. Those values may be caused by various reasons, such as machine malfunction, and they are not considered outliers.

Table 2. The distribution of the values of each variable employed in our models. All variables are present in every dataset.

Clinical Variables	Valid Range		MIMIC-III		MIMIC-IV		eICU	
	Lower	Upper	Survival [1]	Death [2]	Survival [1]	Death [2]	Survival [1]	Death [2]
Vital sign variables (7 variables)								
Heart rate (beats/min)	0	350	84.3 (16.1)	90.4 (19.8)	84.4 (16.7)	91.3 (19.8)	84.2 (17.1)	93.1 (20.8)
Diastolic blood peessure (mmHg)	0	375	61.4 (14.2)	56.8 (14.4)	63.3 (14.6)	57.5 (14.3)	67.9 (14.4)	60.9 (15.1)
Systolic blood pressure (mmHg)	0	375	124.5 (21.4)	116.4 (25.6)	122.6 (21.0)	113.5 (24.5)	126.1 (22.1)	115.4 (24.7)
Mean arterial pressure (mmHg)	14	330	80.5 (15.3)	75.4 (17.1)	79.3 (15.3)	73.8 (16.7)	83.9 (16.1)	76.0 (16.9)
Temperature (°C)	26	45	36.9 (0.6)	36.9 (1.0)	36.9 (0.5)	36.9 (0.8)	36.8 (0.5)	36.8 (1.1)
Peripheral oxygen saturation (%)	0	100	96.8 (2.6)	95.6 (6.4)	96.3 (2.6)	95.3 (6.1)	96.3 (3.2)	95.2 (7.2)
Respiratory rate (breaths/min)	0	300	19.9 (5.4)	21.4 (6.7)	19.7 (5.4)	21.4 (6.4)	19.5 (5.2)	21.9 (7.2)
Laboratory variables (16 variables)								
Albumin (g/dL)	0.6	6	2.6 (0.5)	2.5 (0.6)	2.8 (0.5)	2.7 (0.6)	2.6 (0.6)	2.4 (0.6)
Blood urea nitrogen (mg/dL)	0	250	36.9 (26.5)	50.4 (31.8)	38.0 (26.8)	44.6 (32.4)	26.3 (21.2)	39.8 (28.2)
Bilirubin (mg/dL)	0.1	60	3.6 (5.3)	10.5 (13.1)	3.9 (6.4)	7.9 (10.1)	1.4 (3.2)	3.2 (6.1)
Lactate (mmol/L)	0.4	30	1.9 (1.5)	3.8 (4.2)	1.9 (1.4)	3.5 (3.4)	2.1 (2.0)	5.5 (5.0)
Bicarbonate (mEq/L)	0	60	25.8 (4.7)	23.9 (5.7)	25.8 (5.3)	23.0 (5.7)	25.6 (4.9)	22.3 (6.0)
Band neutrophil (%)	0	100	5.2 (6.3)	6.4 (7.2)	4.6 (5.7)	5.2 (5.6)	8.6 (12.6)	12.2 (13.1)
Chloride (mEq/L)	50	175	105.3 (6.1)	103.9 (7.3)	103.6 (7.2)	103.3 (7.9)	104.3 (6.7)	106.8 (9.2)
Creatinine (mg/dL)	0.1	60	1.4 (1.4)	1.8 (1.3)	1.6 (1.5)	1.8 (1.3)	1.3 (1.4)	2.0 (1.6)
Glucose (mg/dL)	33	2000	131.6 (52.2)	136.9 (64.3)	140.6 (62.3)	147.8 (67.4)	144.2 (57.2)	149.9 (62.9)
Hemoglobin (g/dL)	0	25	9.6 (1.4)	9.6 (1.4)	9.1 (1.6)	8.9 (1.4)	10.2 (2.0)	9.8 (2.1)
Hematocrit (g/dL)	0	75	28.6 (3.9)	28.6 (3.9)	27.8 (4.6)	27.2 (4.3)	31.1 (6.1)	29.8 (6.5)
Platelet count (1000/mm^3)	0	2000	278.9 (192.2)	162.5 (138.9)	237.0 (174.2)	154.3 (134.8)	208.3 (115.7)	153.4 (105.5)
Potassium (mEq/L)	0	12	4.0 (0.5)	4.1 (0.6)	4.0 (0.5)	4.1 (0.6)	3.9 (0.5)	4.2 (0.8)
Partial thromboplastin time (s)	18.8	150	44.7 (23.7)	52.9 (28.0)	47.7 (24.4)	52.9 (26.7)	50.7 (27.7)	52.2 (27.9)
Sodium (mEq/L)	50	225	140.3 (5.2)	139.1 (6.0)	140.4 (5.9)	139.5 (6.7)	138.9 (5.7)	141.7 (8.1)
White blood cells (1000/mm^3)	0	1000	12.9 (8.5)	14.7 (9.8)	12.9 (8.9)	15.7 (13.7)	11.2 (6.0)	15.8 (9.7)

Prognosis of Mechanically Ventilated Patients

Summary:
The prognosis of patients who need mechanical ventilation is affected by various factors, including age, serum albumin levels, APACHE II scores, and the timing of mechanical ventilation after CPR. The study shows that the one-year mortality rate is high, especially among patients aged over 70. Additionally, the APACHE II scoring system is an effective tool to predict the outcomes of mechanically ventilated patients,

both on an individual hospital level and as a whole group among veterans receiving intensive care.

How does this article help the report?

In our report, we decided to remove the "ventilation duration" column as not all patients require mechanical ventilation upon admission to the ICU. Recent research has suggested that ventilation has little to no impact on patient survival, making it safe to eliminate this column.

Additionally, the article confirms our previous findings regarding the correlation between hospital mortality and patients' ages. Our density plot of age by hospital mortality revealed a significant drop in density after the age of 70, indicating that younger patients have a higher likelihood of survival.

Machine Learning Prediction Models for Mechanically Ventilated Patients: Analyses of the MIMIC-III Database

The research utilized the MIMIC-III database to predict hospital mortality in ICU patients undergoing mechanical ventilation. Seven machine learning models were employed, with the XGBoost model proving the most effective. The five leading predictors of hospital mortality were age, respiratory dysfunction, SAPS II score, maximum hemoglobin, and minimum lactate. The study reveals that these factors are likely closely correlated with hospital mortality in mechanically ventilated ICU patients. However, external validation is essential to confirm these findings.

How does this article help the report?

This text contains important information for the second part of the assignment. According to the article, the XGBoost machine learning model achieved the highest AUCs (areas under the curve of the receiver operating characteristic). This is the model that we are going to use in our second part of assignment. Additionally, the article highlights that age is the most significant factor, followed by respiratory dysfunction, SAPS II score, maximum hemoglobin, and minimum lactate. We will use our dataset to confirm if we obtain similar results.

References

Papadakis, M A, et al. “Prognosis of Mechanically Ventilated Patients.” *PubMed*, vol. 159, no. 6, 1 Dec. 1993, pp. 659–64. Accessed 17 Mar. 2024.

Tan, Yvette, et al. “Predicting Respiratory Decompensation in Mechanically Ventilated Adult ICU Patients.” *Frontiers in Physiology*, vol. 14, 14 Apr. 2023, https://doi.org/10.3389/fphys.2023.1125991. Accessed 30 Nov. 2023.

Thanakron Na Pattalung, et al. “Feature Explanations in Recurrent Neural Networks for Predicting Risk of Mortality in Intensive Care Patients.” *Journal of Personalized Medicine*, vol. 11, no. 9, 19 Sept. 2021, pp. 934–934, https://doi.org/10.3390/jpm11090934. Accessed 17 Mar. 2024.

Thanos Gentimis, et al. *Predicting Hospital Length of Stay Using Neural Networks on MIMIC III Data*. 1 Nov. 2017, https://doi.org/10.1109/dasc-picom-datacom-cyberscitec.2017.191. Accessed 21 May 2023.

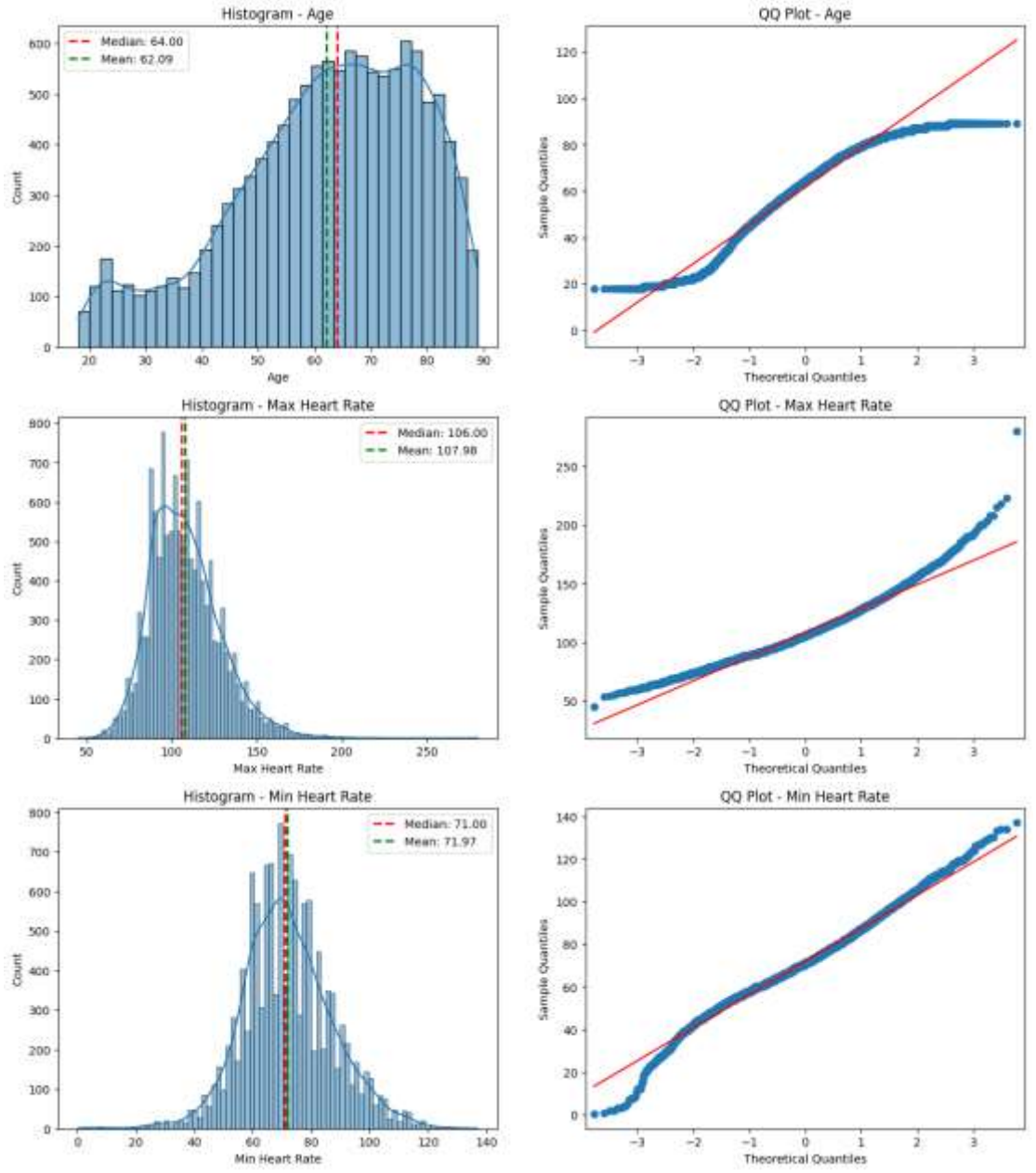
Zhu, Yibing, et al. “Machine Learning Prediction Models for Mechanically Ventilated Patients: Analyses of the MIMIC-III Database.” *Frontiers in Medicine*, vol. 8, 1 July 2021, https://doi.org/10.3389/fmed.2021.662340.

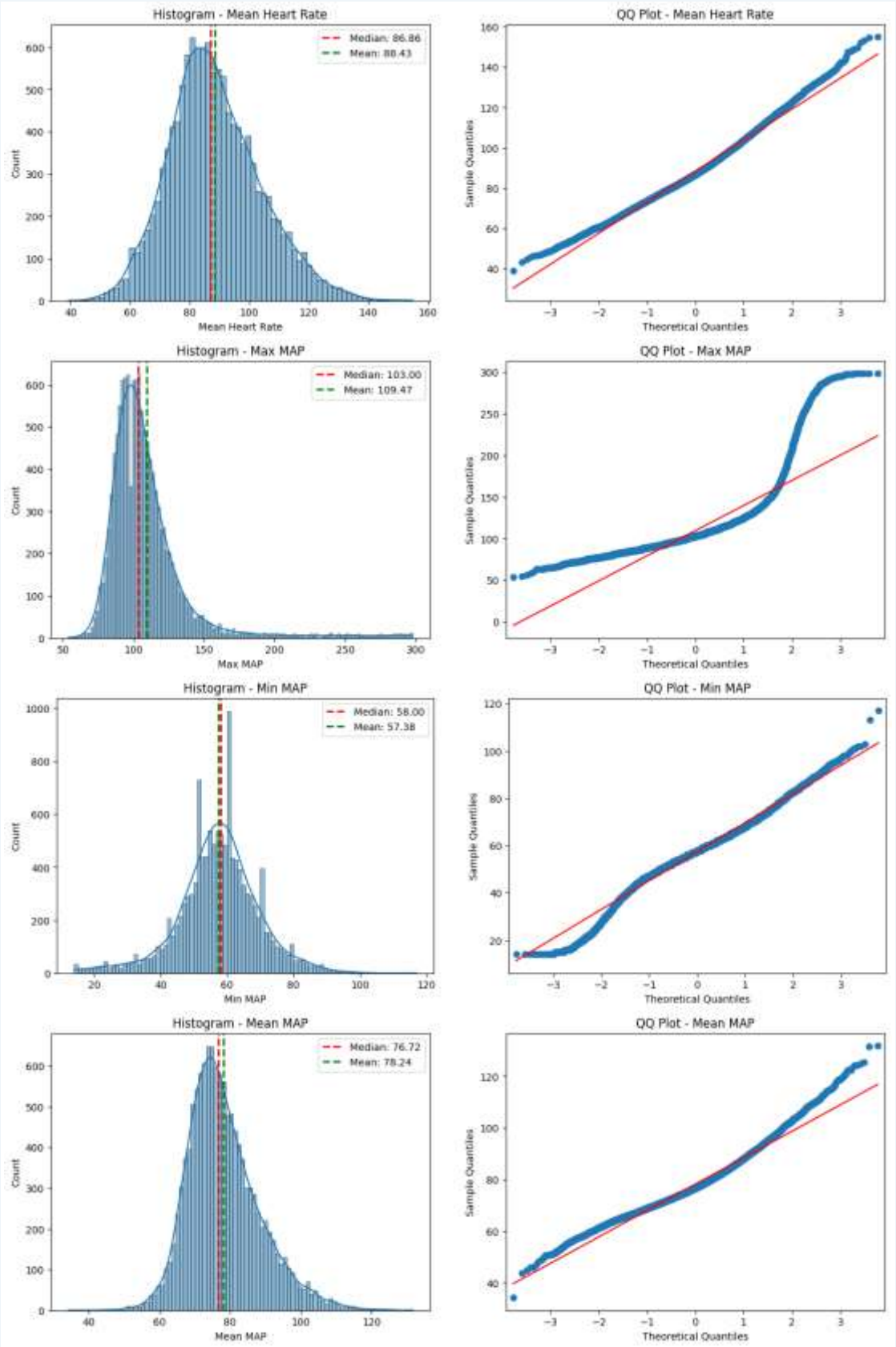
Appendix

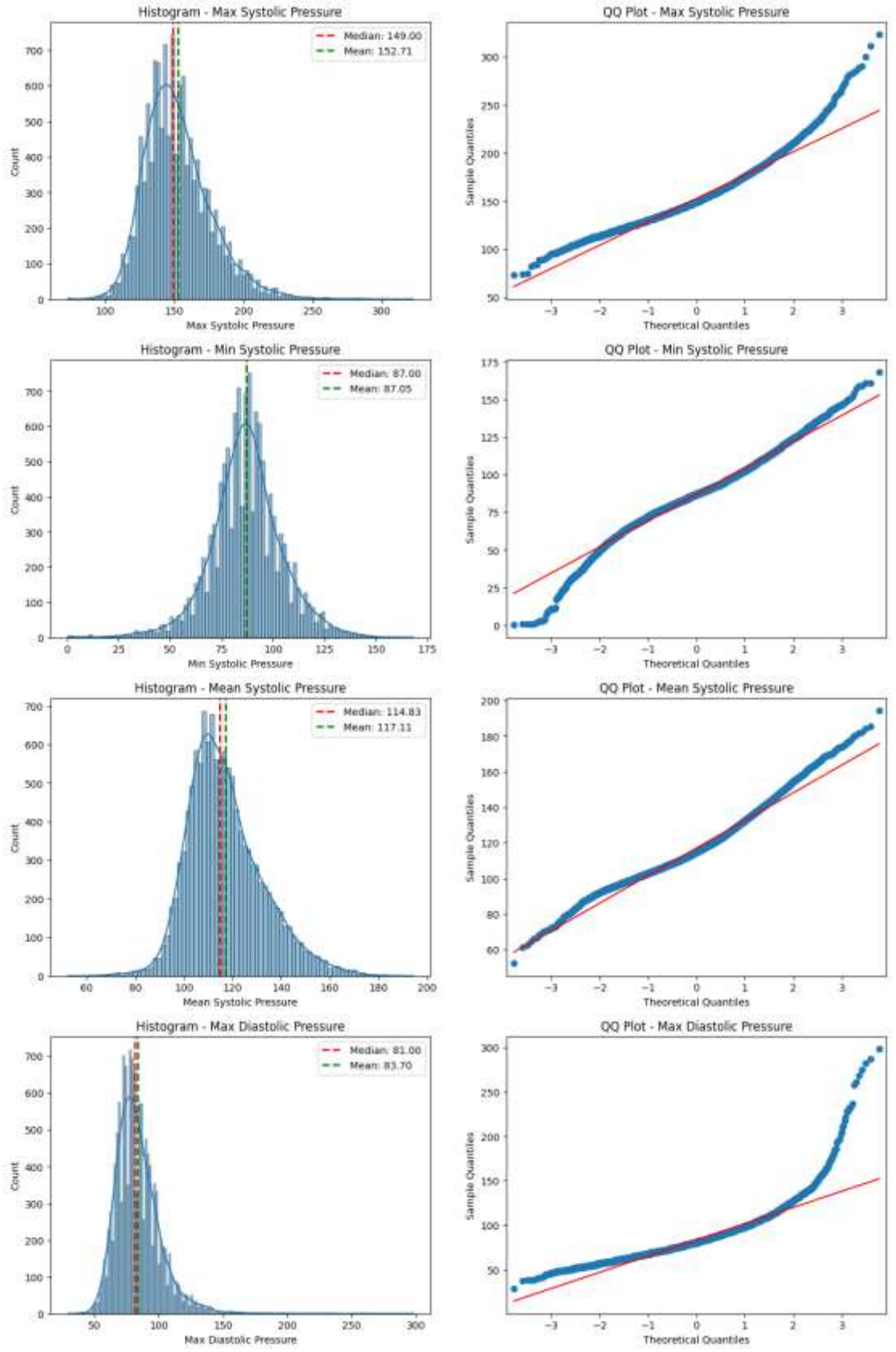
Normalization

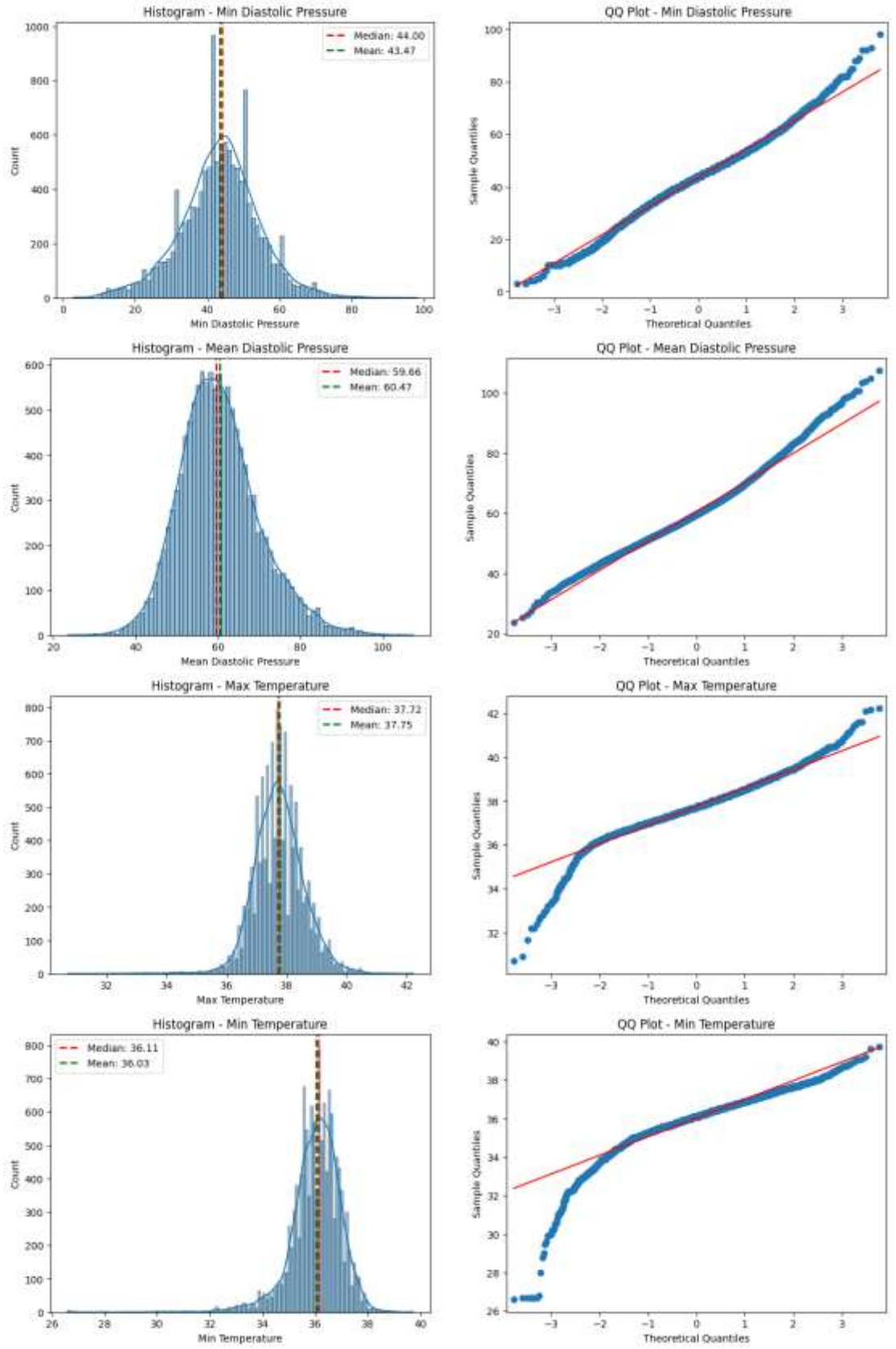
Before normalization

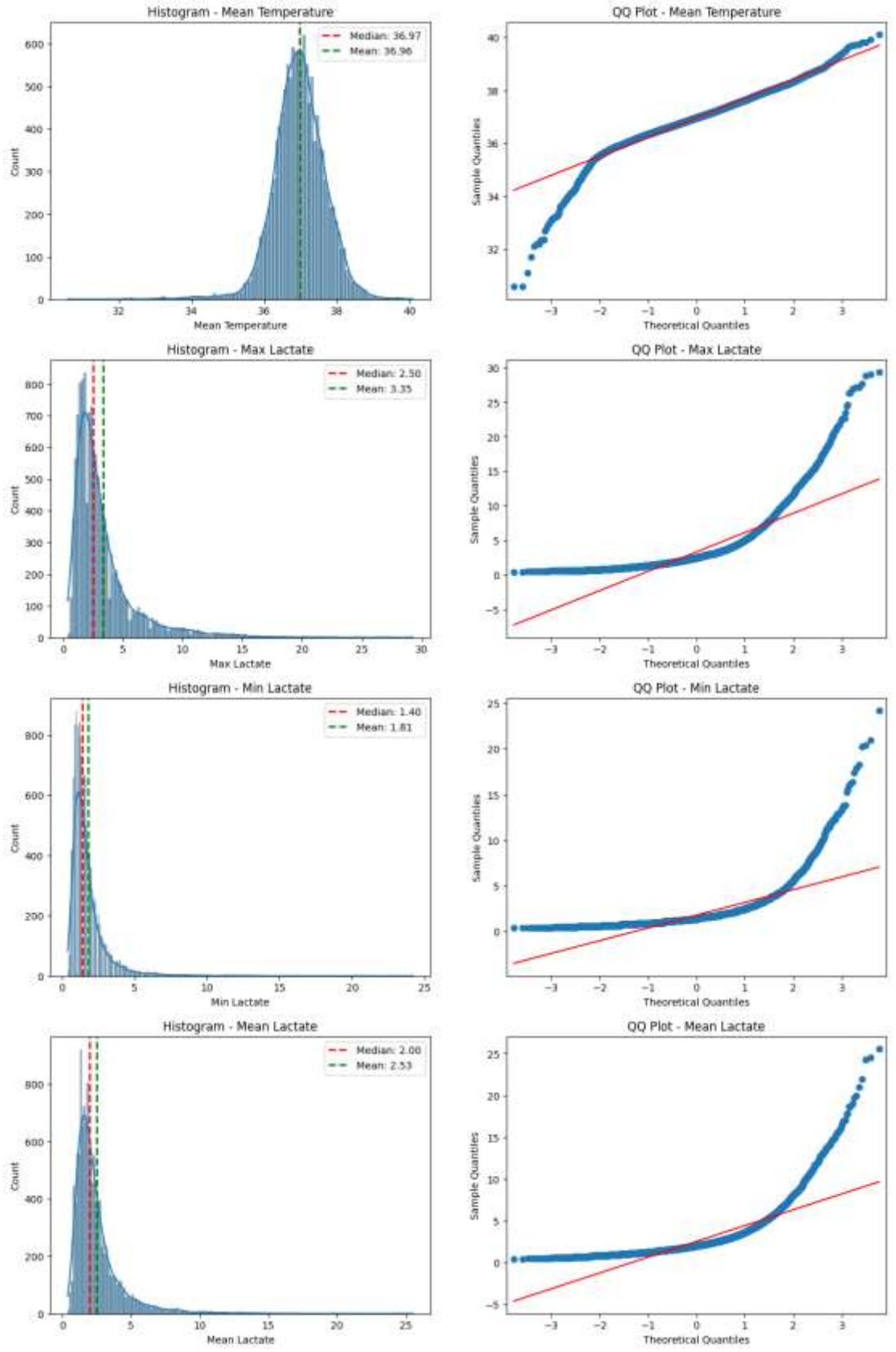
Histograms and QQ Plots

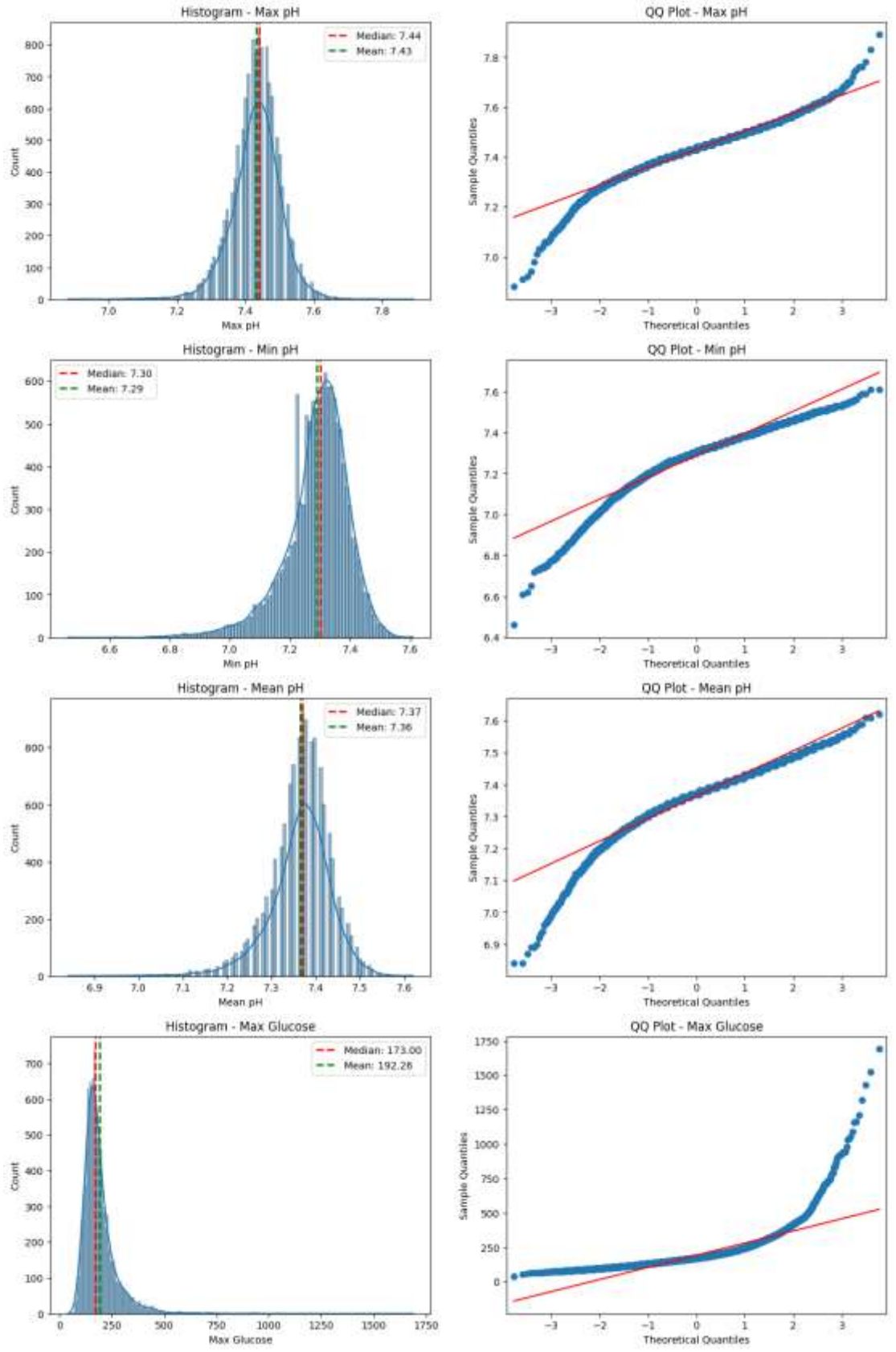


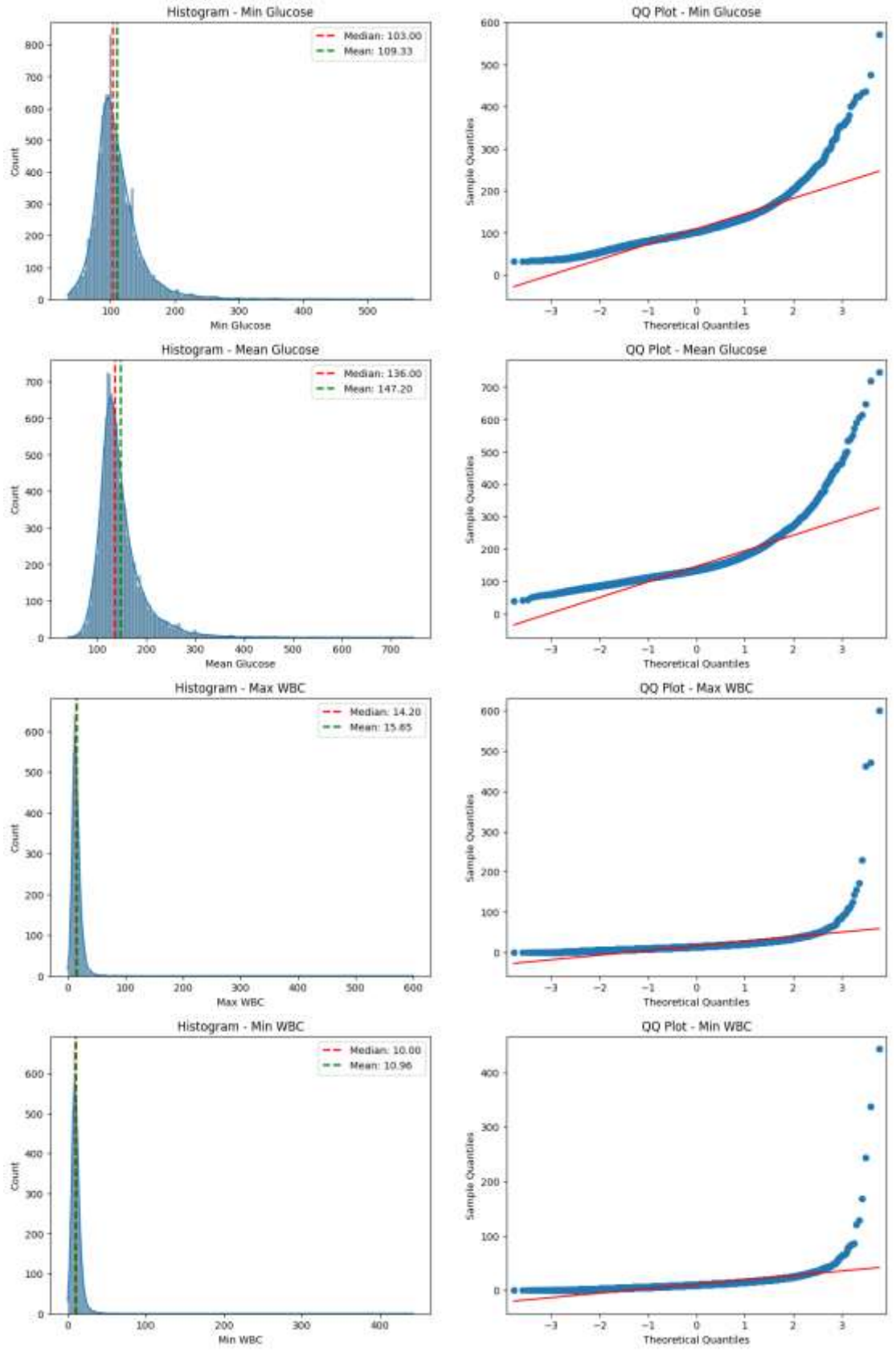

Histogram - Mean Heart Rate
Median: 86.86
Mean: 88.43
Count
Mean Heart Rate
QQ Plot - Mean Heart Rate
Sample Quantiles
Theoretical Quantiles
Histogram - Max MAP
Median: 103.00
Mean: 109.47
Count
Max MAP
QQ Plot - Max MAP
Sample Quantiles
Theoretical Quantiles
Histogram - Min MAP
Median: 58.00
Mean: 57.38
Count
Min MAP
QQ Plot - Min MAP
Sample Quantiles
Theoretical Quantiles
Histogram - Mean MAP
Median: 76.72
Mean: 78.24
Count
Mean MAP
QQ Plot - Mean MAP
Sample Quantiles
Theoretical Quantiles

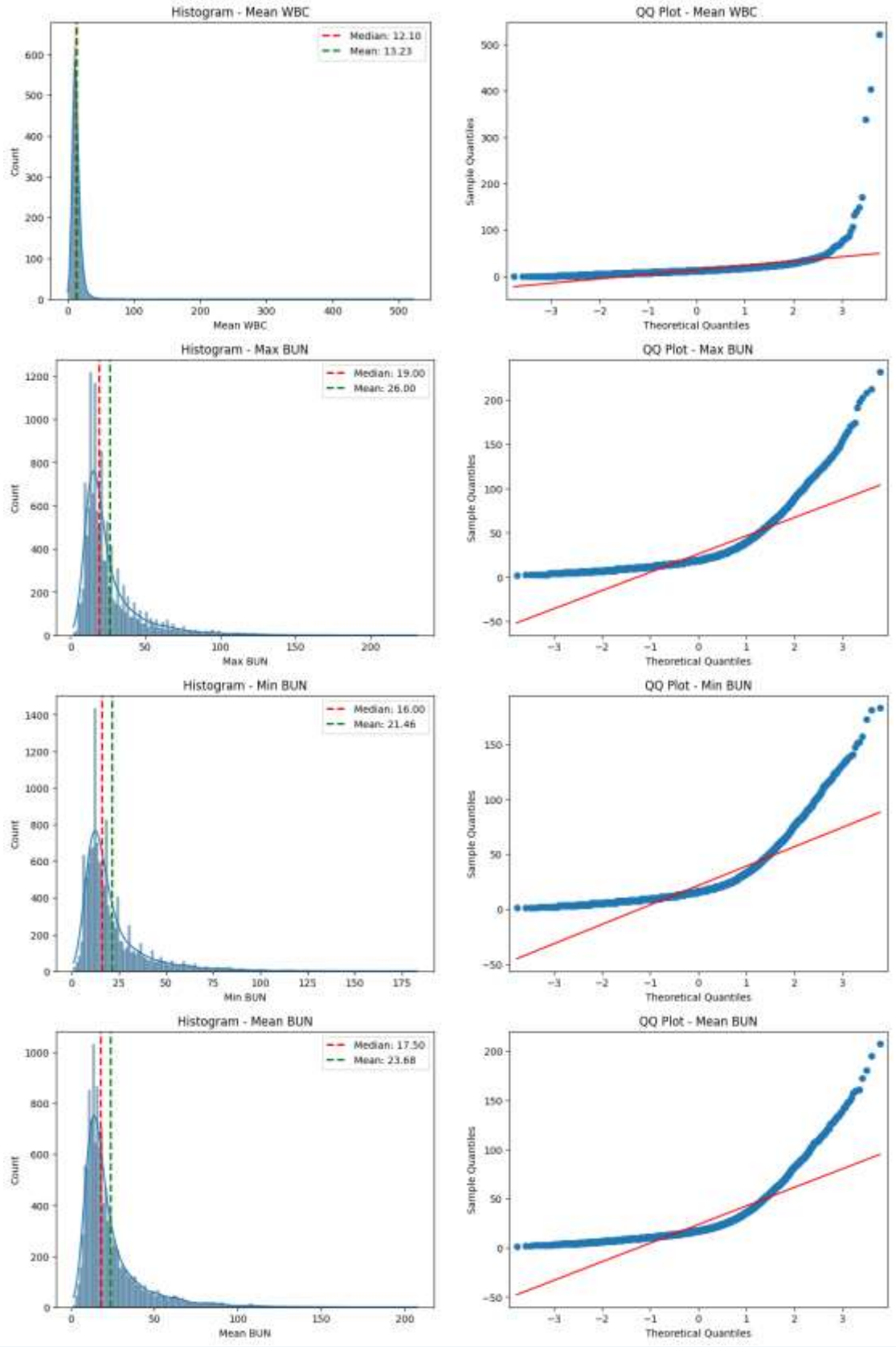

Histogram - Max Systolic Pressure
Median: 149.00
Mean: 152.71
Count
Max Systolic Pressure
QQ Plot - Max Systolic Pressure
Sample Quantiles
Theoretical Quantiles
Histogram - Min Systolic Pressure
Median: 87.00
Mean: 87.05
Count
Min Systolic Pressure
QQ Plot - Min Systolic Pressure
Sample Quantiles
Theoretical Quantiles
Histogram - Mean Systolic Pressure
Median: 114.83
Mean: 117.11
Count
Mean Systolic Pressure
QQ Plot - Mean Systolic Pressure
Sample Quantiles
Theoretical Quantiles
Histogram - Max Diastolic Pressure
Median: 81.00
Mean: 83.70
Count
Max Diastolic Pressure
QQ Plot - Max Diastolic Pressure
Sample Quantiles
Theoretical Quantiles

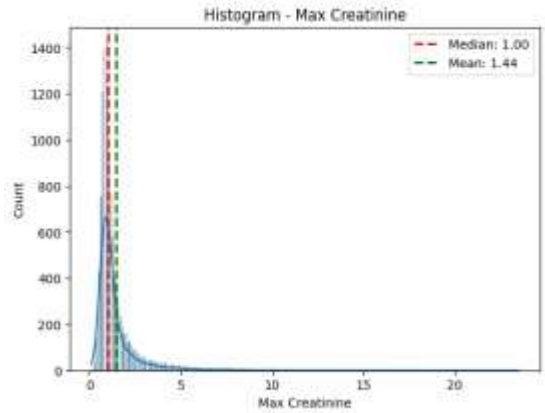

Histogram - Min Diastolic Pressure
Median: 44.00
Mean: 43.47
Count
Min Diastolic Pressure
QQ Plot - Min Diastolic Pressure
Sample Quantiles
Theoretical Quantiles
Histogram - Mean Diastolic Pressure
Median: 59.66
Mean: 60.47
Count
Mean Diastolic Pressure
QQ Plot - Mean Diastolic Pressure
Sample Quantiles
Theoretical Quantiles
Histogram - Max Temperature
Median: 37.72
Mean: 37.75
Count
Max Temperature
QQ Plot - Max Temperature
Sample Quantiles
Theoretical Quantiles
Histogram - Min Temperature
Median: 36.11
Mean: 36.03
Count
Min Temperature
QQ Plot - Min Temperature
Sample Quantiles
Theoretical Quantiles

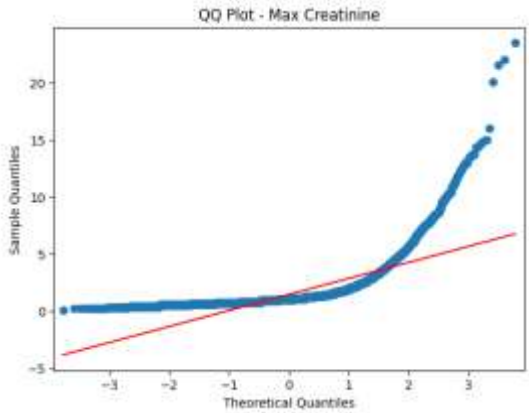

Histogram - Mean Temperature
Median: 36.97
Mean: 36.96
Count
Mean Temperature
QQ Plot - Mean Temperature
Sample Quantiles
Theoretical Quantiles
Histogram - Max Lactate
Median: 2.50
Mean: 3.35
Count
Max Lactate
QQ Plot - Max Lactate
Sample Quantiles
Theoretical Quantiles
Histogram - Min Lactate
Median: 1.40
Mean: 1.81
Count
Min Lactate
QQ Plot - Min Lactate
Sample Quantiles
Theoretical Quantiles
Histogram - Mean Lactate
Median: 2.00
Mean: 2.53
Count
Mean Lactate
QQ Plot - Mean Lactate
Sample Quantiles
Theoretical Quantiles

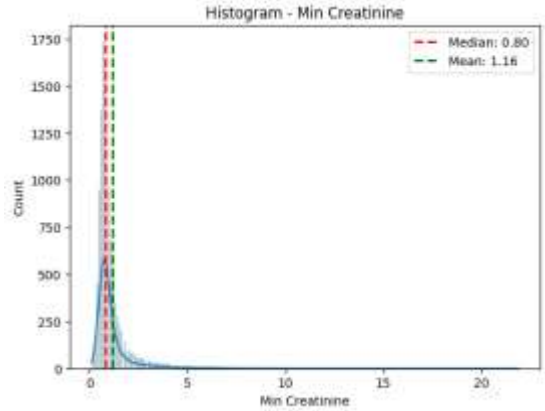

Histogram - Max pH
Median: 7.44
Mean: 7.43
Count
Max pH
QQ Plot - Max pH
Sample Quantiles
Theoretical Quantiles
Histogram - Min pH
Median: 7.30
Mean: 7.29
Count
Min pH
QQ Plot - Min pH
Sample Quantiles
Theoretical Quantiles
Histogram - Mean pH
Median: 7.37
Mean: 7.36
Count
Mean pH
QQ Plot - Mean pH
Sample Quantiles
Theoretical Quantiles
Histogram - Max Glucose
Median: 173.00
Mean: 192.26
Count
Max Glucose
QQ Plot - Max Glucose
Sample Quantiles
Theoretical Quantiles

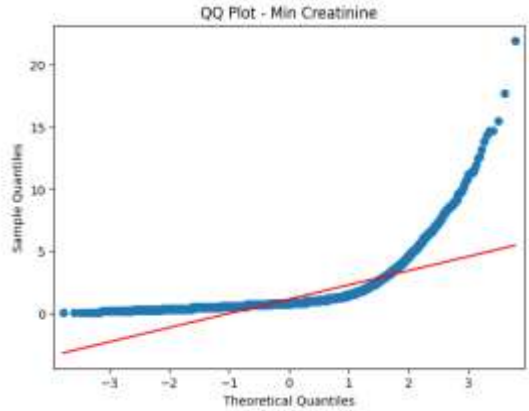

Histogram - Min Glucose
Median: 103.00
Mean: 109.33
Count
Min Glucose
QQ Plot - Min Glucose
Sample Quantiles
Theoretical Quantiles
Histogram - Mean Glucose
Median: 136.00
Mean: 147.20
Count
Mean Glucose
QQ Plot - Mean Glucose
Sample Quantiles
Theoretical Quantiles
Histogram - Max WBC
Median: 14.20
Mean: 15.65
Count
Max WBC
QQ Plot - Max WBC
Sample Quantiles
Theoretical Quantiles
Histogram - Min WBC
Median: 10.00
Mean: 10.96
Count
Min WBC
QQ Plot - Min WBC
Sample Quantiles
Theoretical Quantiles

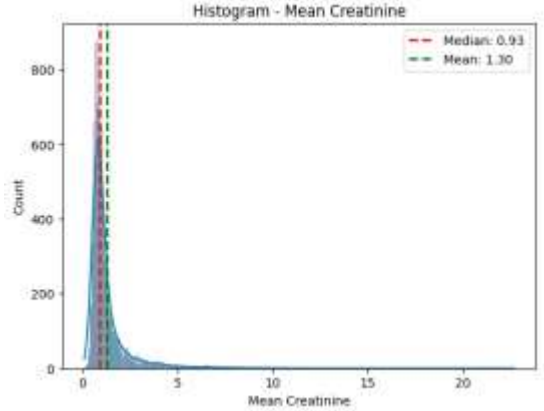

Histogram - Mean WBC
Median: 12.10
Mean: 13.23
Count
Mean WBC
QQ Plot - Mean WBC
Sample Quantiles
Theoretical Quantiles
Histogram - Max BUN
Median: 19.00
Mean: 26.00
Count
Max BUN
QQ Plot - Max BUN
Sample Quantiles
Theoretical Quantiles
Histogram - Min BUN
Median: 16.00
Mean: 21.46
Count
Min BUN
QQ Plot - Min BUN
Sample Quantiles
Theoretical Quantiles
Histogram - Mean BUN
Median: 17.50
Mean: 23.68
Count
Mean BUN
QQ Plot - Mean BUN
Sample Quantiles
Theoretical Quantiles

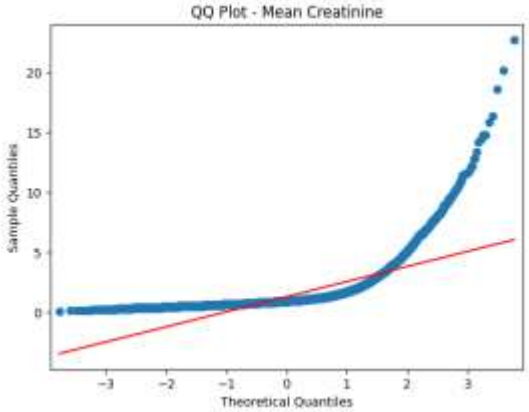

Histogram - Max Creatinine
Median: 1.00
Mean: 1.44
Count
Max Creatinine

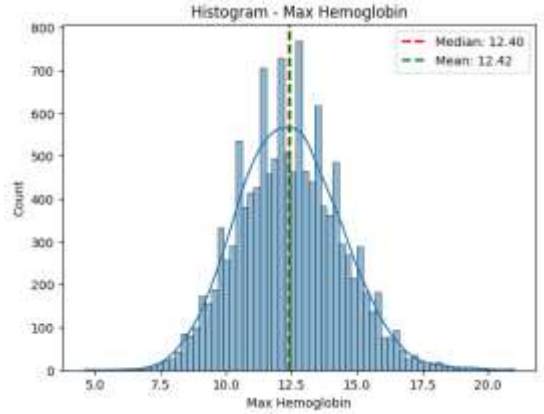

QQ Plot - Max Creatinine
Sample Quantiles
Theoretical Quantiles

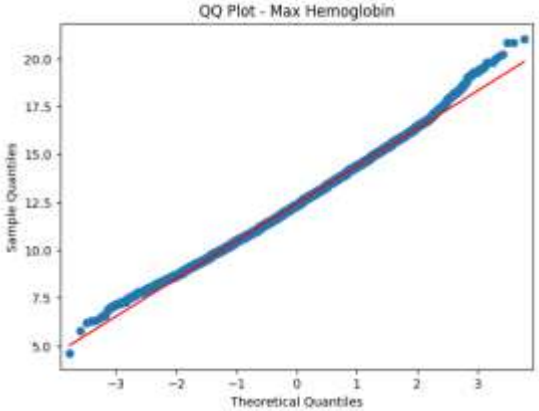

Histogram - Min Creatinine
Median: 0.80
Mean: 1.16
Count
Min Creatinine

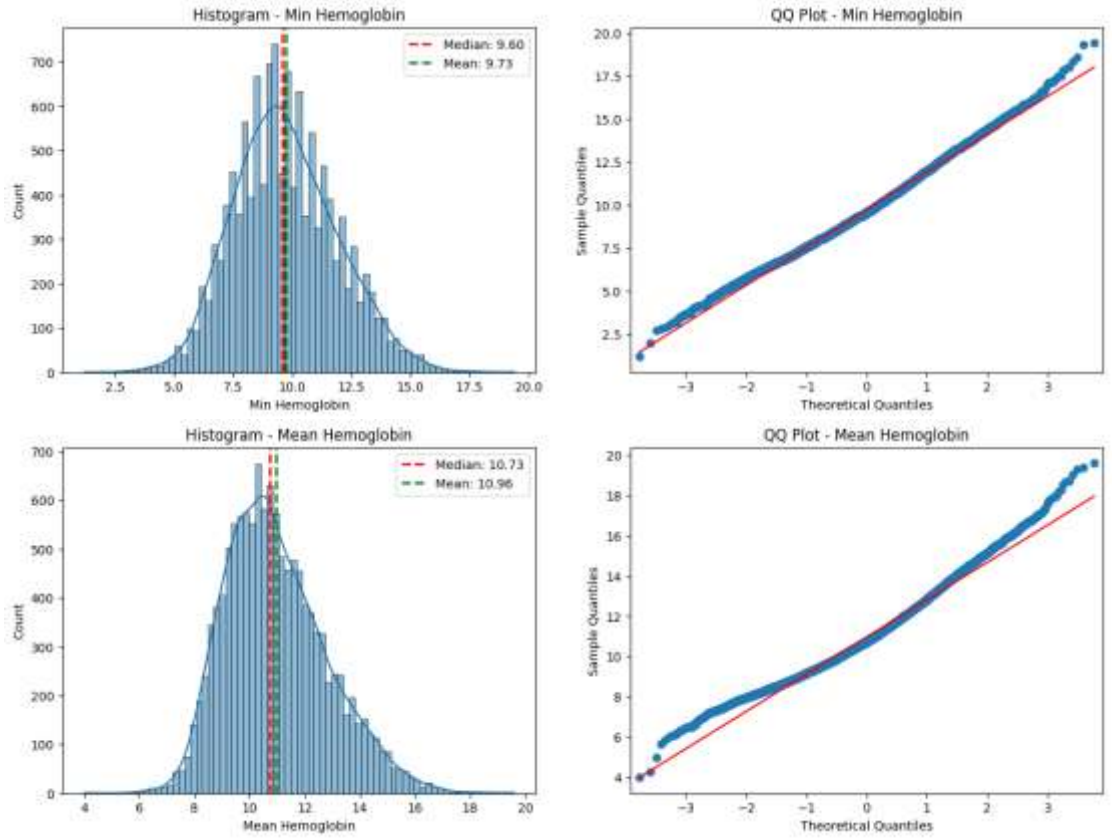

QQ Plot - Min Creatinine
Sample Quantiles
Theoretical Quantiles


Histogram - Mean Creatinine
Median: 0.93
Mean: 1.30
Count
Mean Creatinine


QQ Plot - Mean Creatinine
Sample Quantiles
Theoretical Quantiles

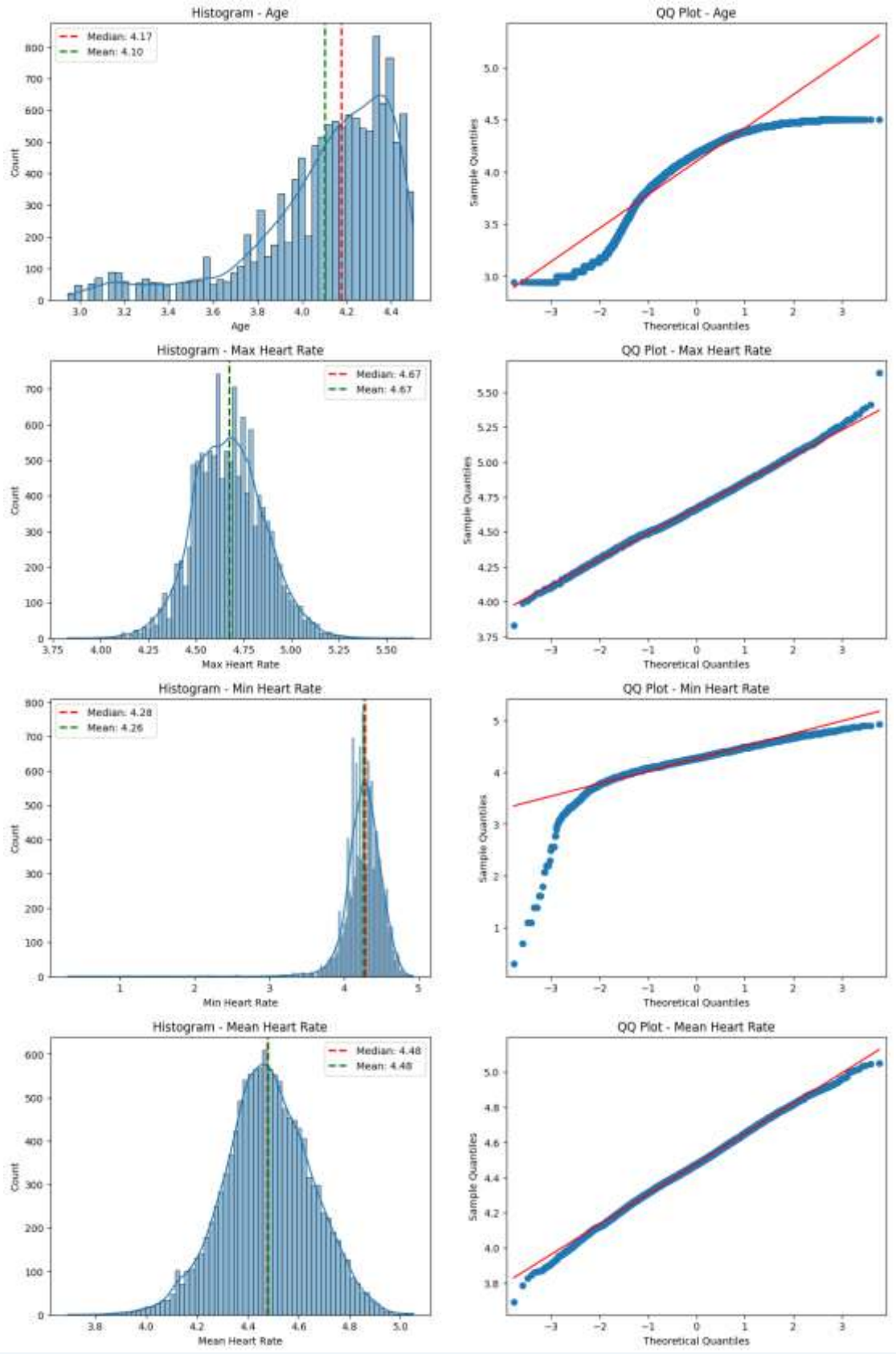

Histogram - Max Hemoglobin
Median: 12.40
Mean: 12.42
Count
Max Hemoglobin

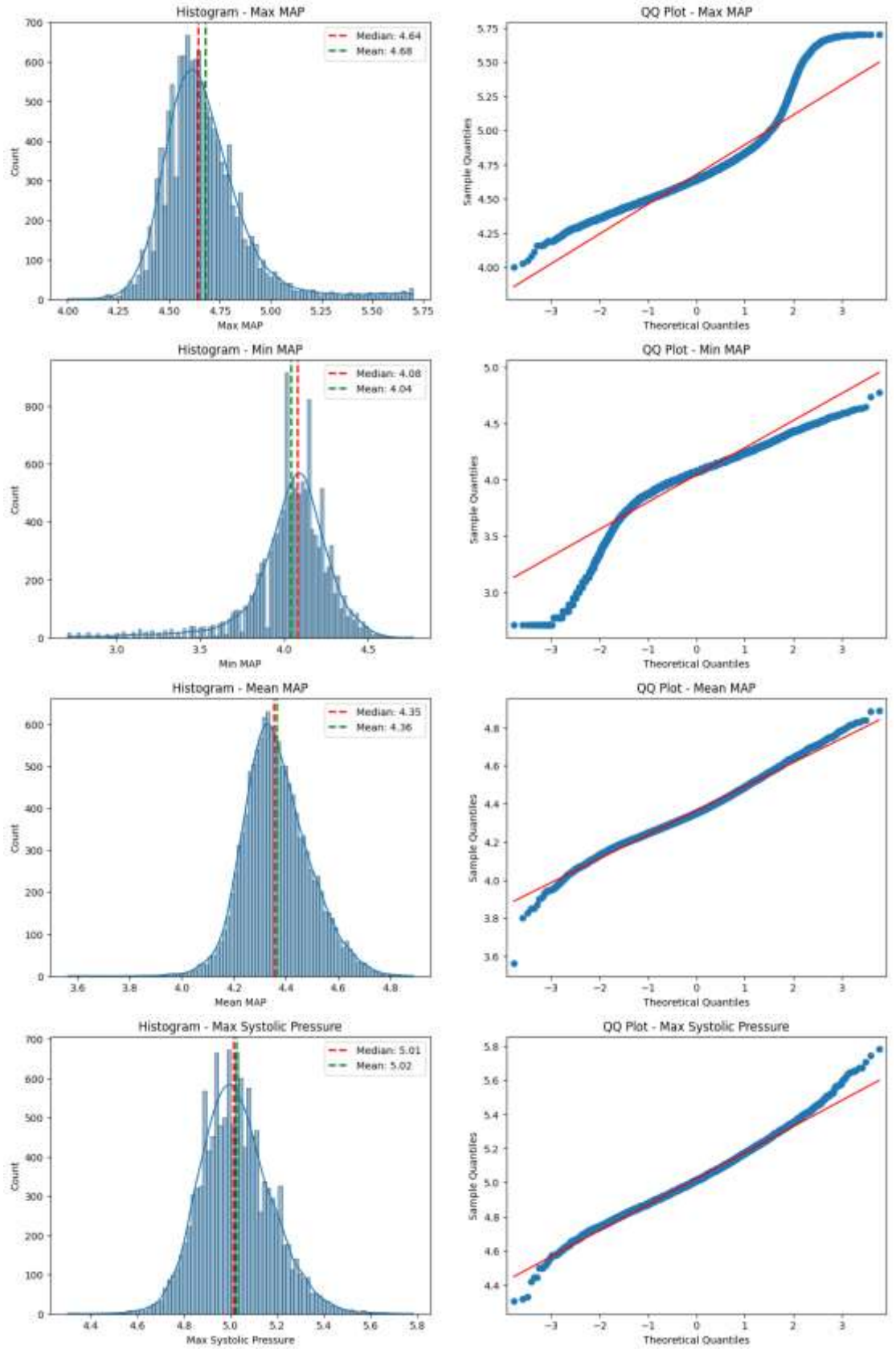

QQ Plot - Max Hemoglobin
Sample Quantiles
Theoretical Quantiles

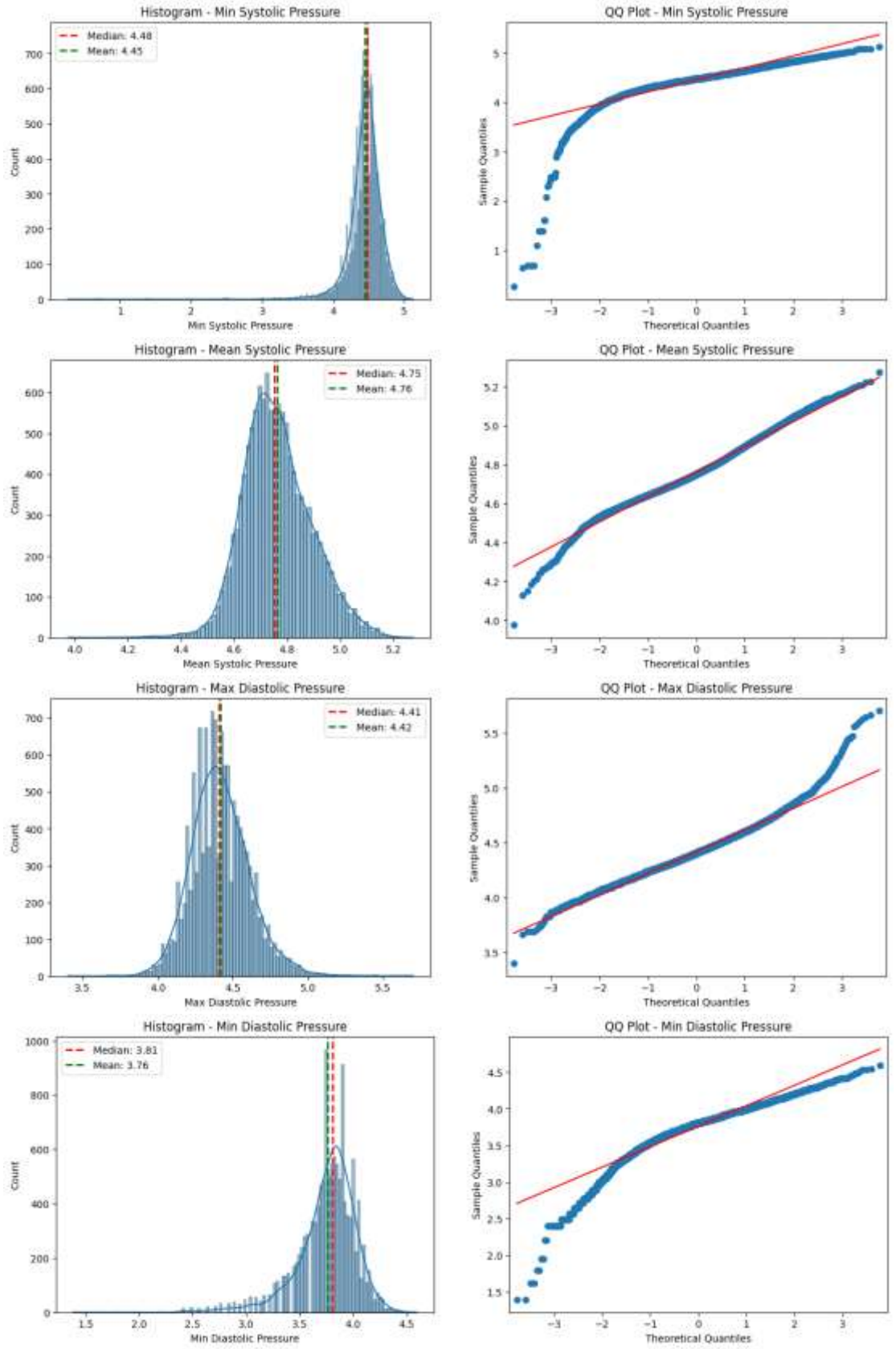

Histogram - Min Hemoglobin
Median: 9.60
Mean: 9.73
Count
Min Hemoglobin
QQ Plot - Min Hemoglobin
Sample Quantiles
Theoretical Quantiles
Histogram - Mean Hemoglobin
Median: 10.73
Mean: 10.96
Count
Mean Hemoglobin
QQ Plot - Mean Hemoglobin
Sample Quantiles
Theoretical Quantiles

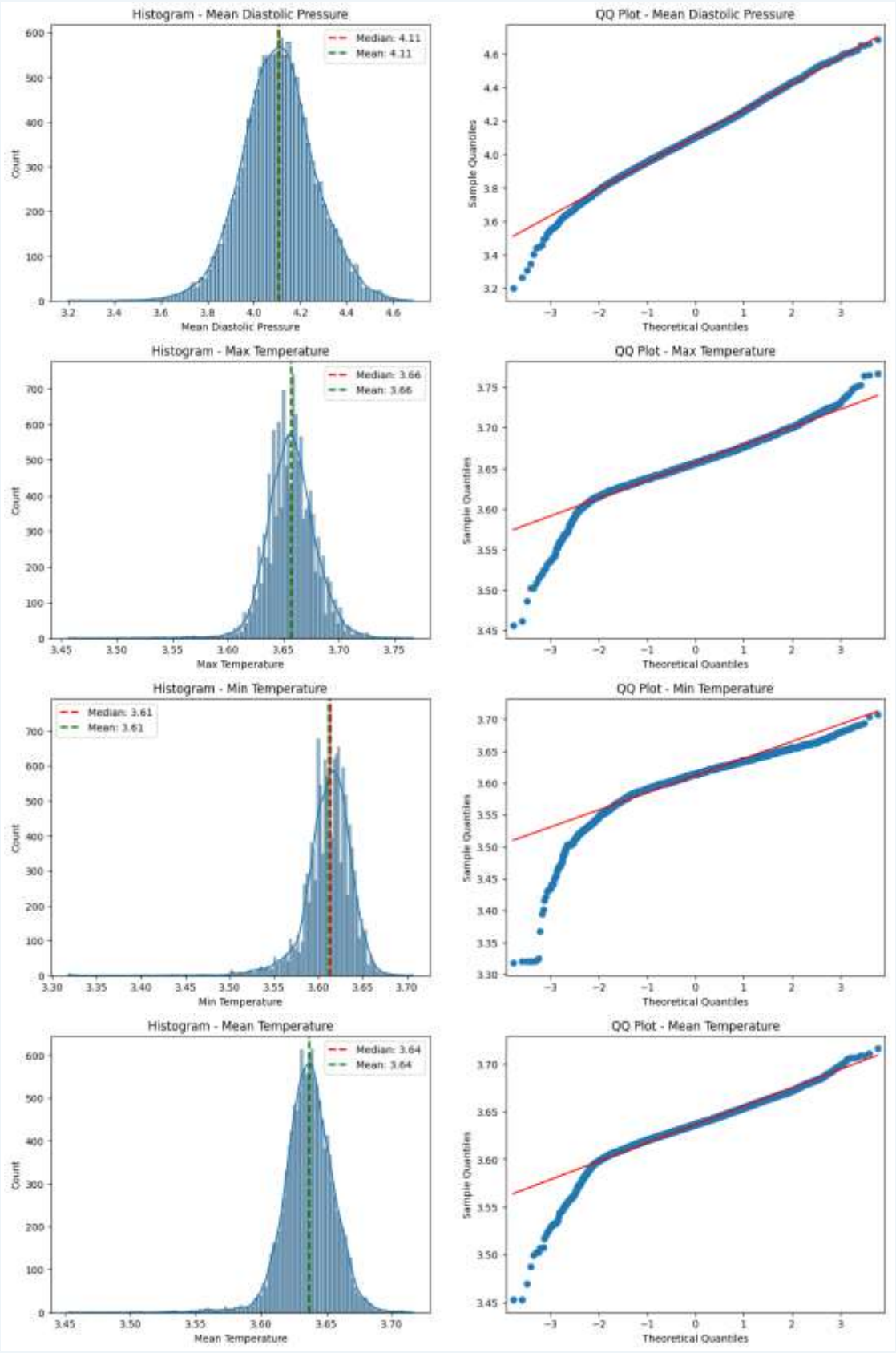
After normalization

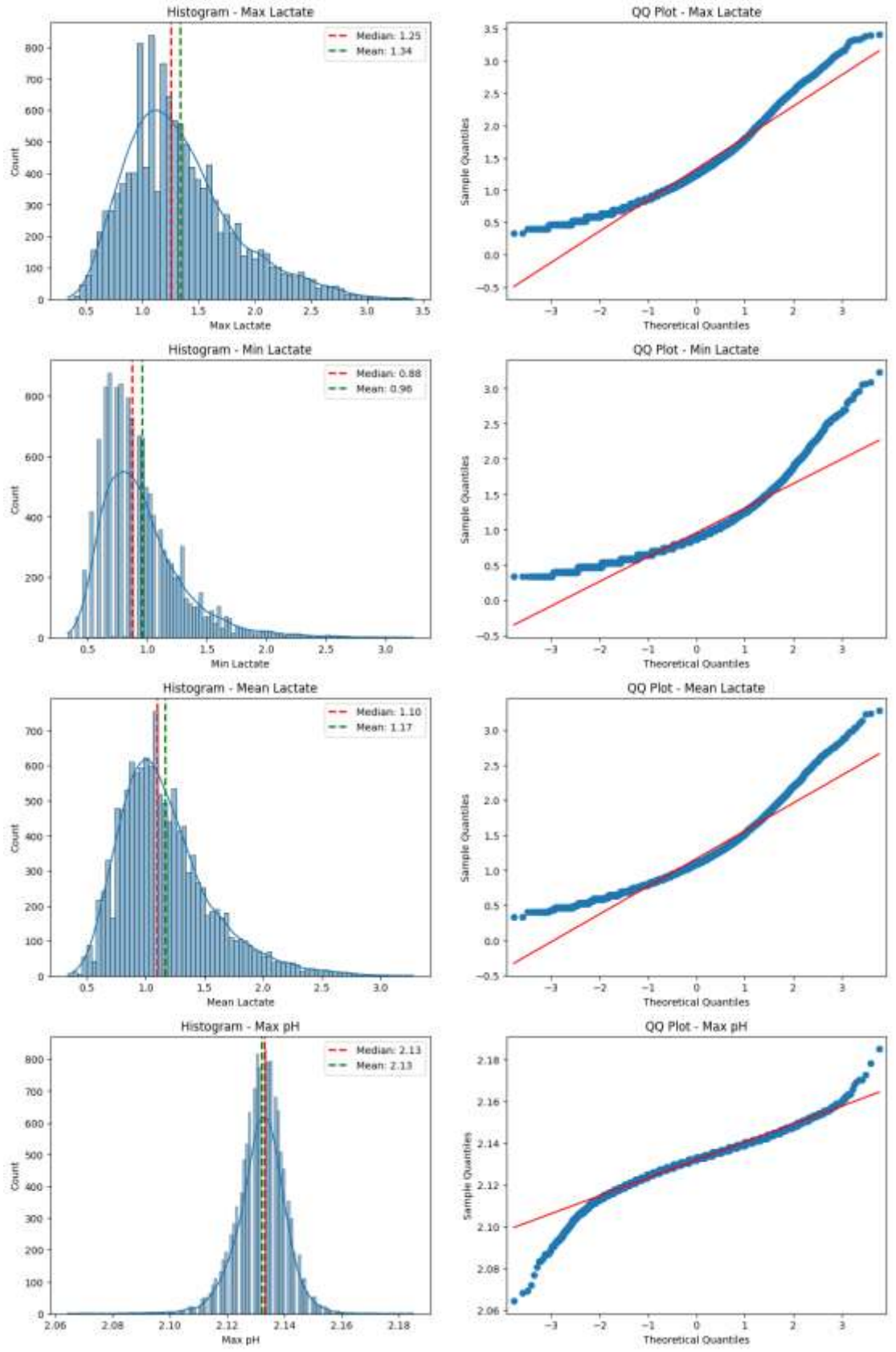
Method 1 – Log Transformation Method

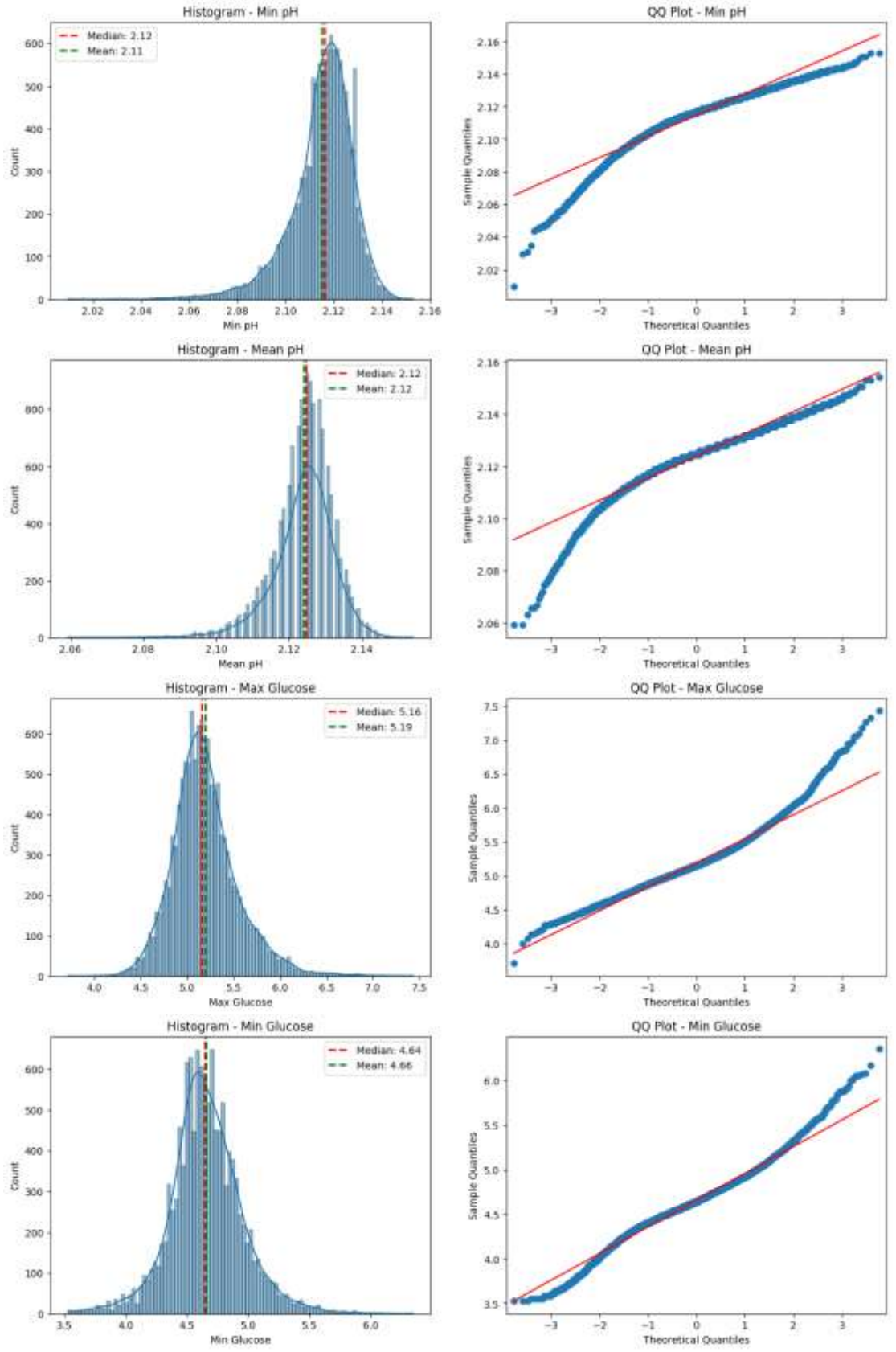


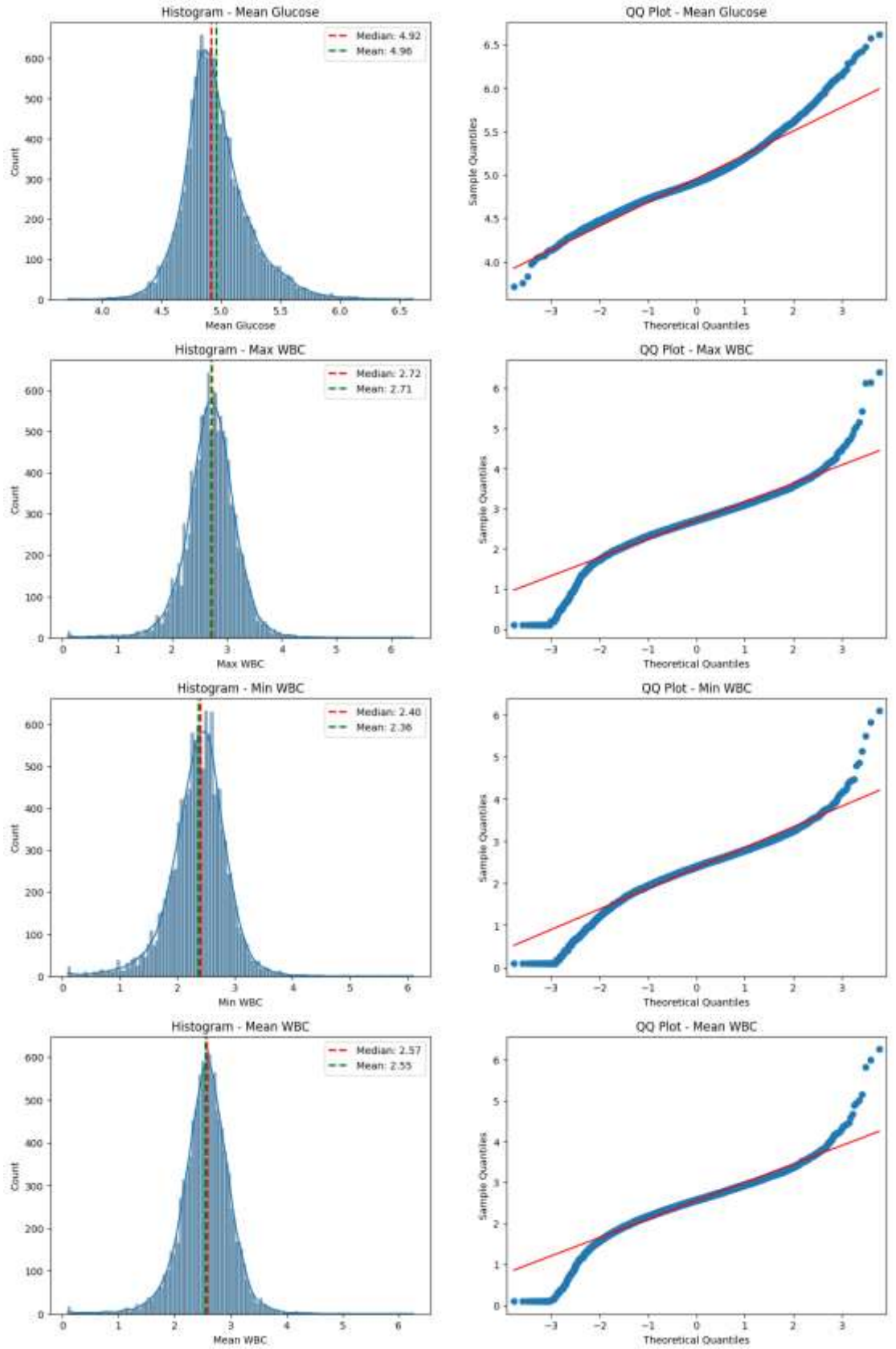

Histogram - Max MAP
Median: 4.64
Mean: 4.68
Count
Max MAP
QQ Plot - Max MAP
Sample Quantiles
Theoretical Quantiles
Histogram - Min MAP
Median: 4.08
Mean: 4.04
Count
Min MAP
QQ Plot - Min MAP
Sample Quantiles
Theoretical Quantiles
Histogram - Mean MAP
Median: 4.35
Mean: 4.36
Count
Mean MAP
QQ Plot - Mean MAP
Sample Quantiles
Theoretical Quantiles
Histogram - Max Systolic Pressure
Median: 5.01
Mean: 5.02
Count
Max Systolic Pressure
QQ Plot - Max Systolic Pressure
Sample Quantiles
Theoretical Quantiles

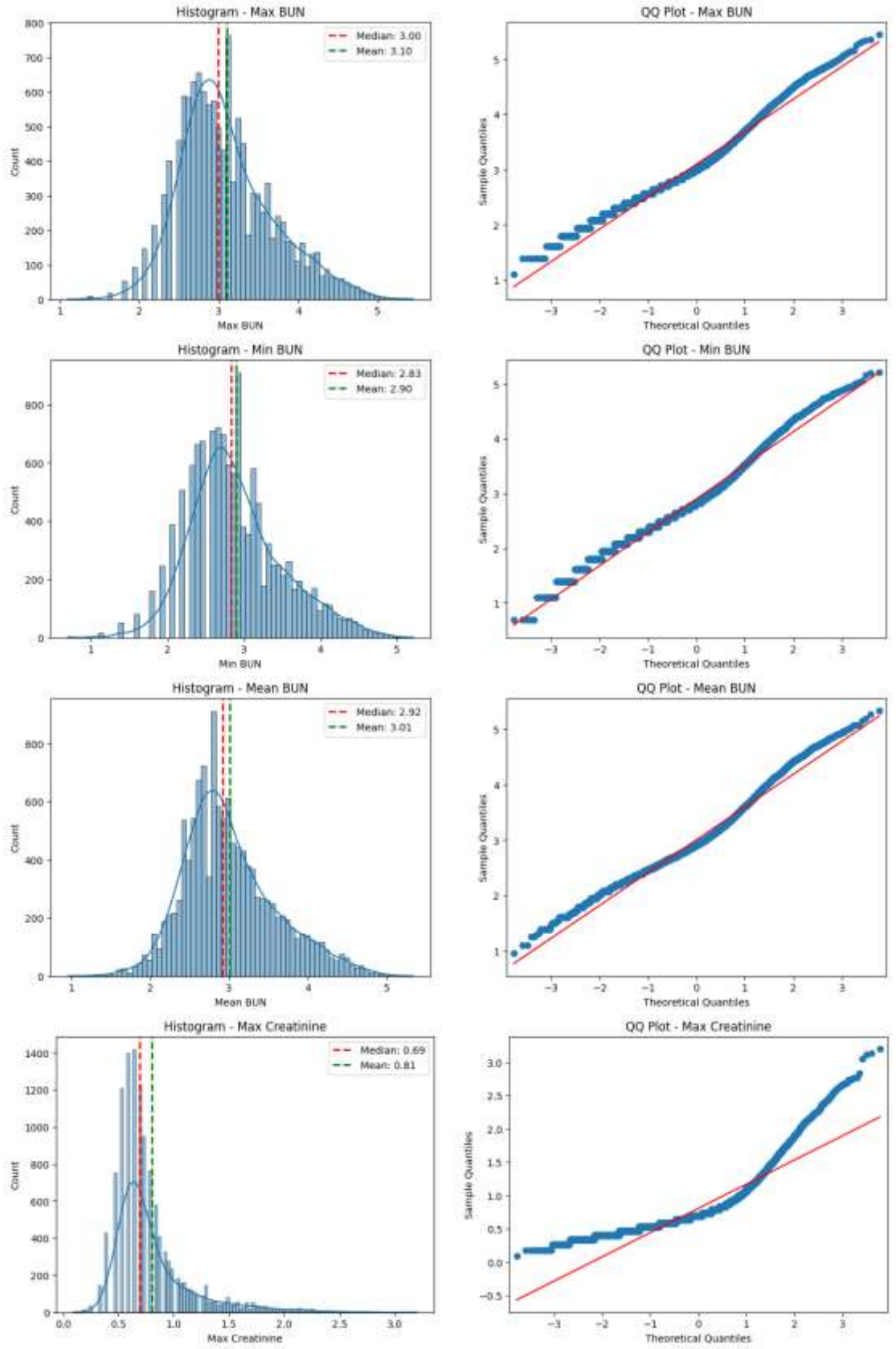

Histogram - Min Systolic Pressure
Median: 4.48
Mean: 4.45
Count
Min Systolic Pressure
QQ Plot - Min Systolic Pressure
Sample Quantiles
Theoretical Quantiles
Histogram - Mean Systolic Pressure
Median: 4.75
Mean: 4.76
Count
Mean Systolic Pressure
QQ Plot - Mean Systolic Pressure
Sample Quantiles
Theoretical Quantiles
Histogram - Max Diastolic Pressure
Median: 4.41
Mean: 4.42
Count
Max Diastolic Pressure
QQ Plot - Max Diastolic Pressure
Sample Quantiles
Theoretical Quantiles
Histogram - Min Diastolic Pressure
Median: 3.81
Mean: 3.76
Count
Min Diastolic Pressure
QQ Plot - Min Diastolic Pressure
Sample Quantiles
Theoretical Quantiles

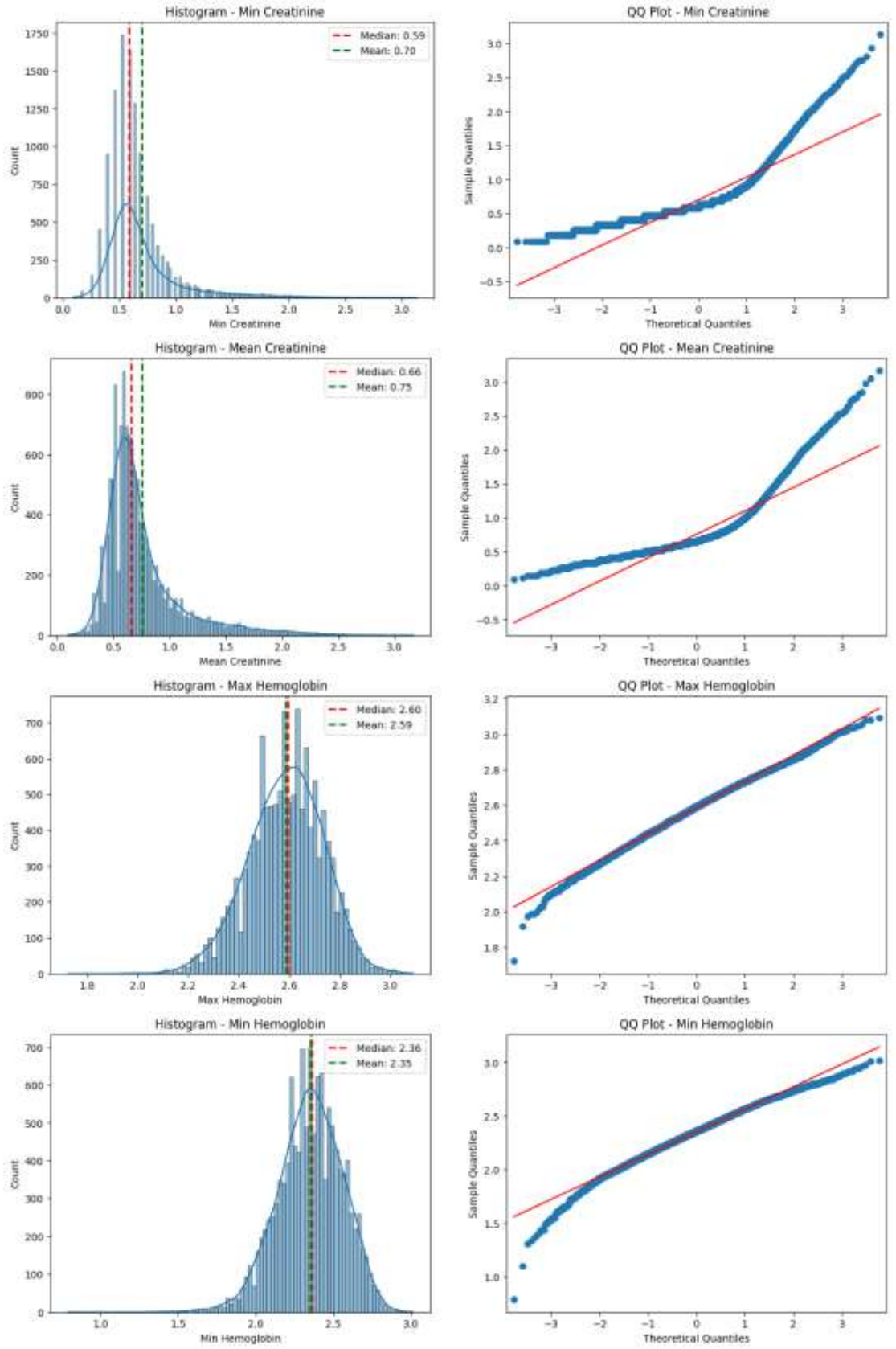

Histogram - Mean Diastolic Pressure
Median: 4.11
Mean: 4.11
Count
Mean Diastolic Pressure
QQ Plot - Mean Diastolic Pressure
Sample Quantiles
Theoretical Quantiles
Histogram - Max Temperature
Median: 3.66
Mean: 3.66
Count
Max Temperature
QQ Plot - Max Temperature
Sample Quantiles
Theoretical Quantiles
Histogram - Min Temperature
Median: 3.61
Mean: 3.61
Count
Min Temperature
QQ Plot - Min Temperature
Sample Quantiles
Theoretical Quantiles
Histogram - Mean Temperature
Median: 3.64
Mean: 3.64
Count
Mean Temperature
QQ Plot - Mean Temperature
Sample Quantiles
Theoretical Quantiles

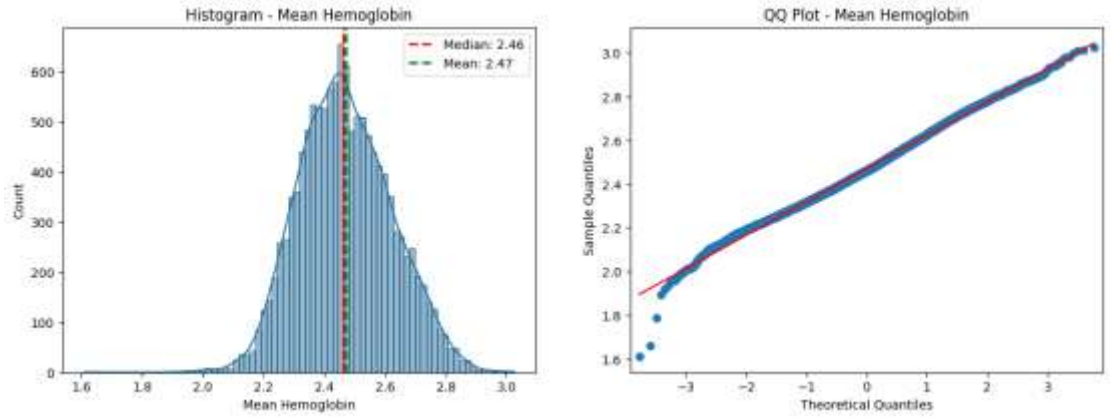

Histogram - Max Lactate
Median: 1.25
Mean: 1.34
Count
Max Lactate
QQ Plot - Max Lactate
Sample Quantiles
Theoretical Quantiles
Histogram - Min Lactate
Median: 0.88
Mean: 0.96
Count
Min Lactate
QQ Plot - Min Lactate
Sample Quantiles
Theoretical Quantiles
Histogram - Mean Lactate
Median: 1.10
Mean: 1.17
Count
Mean Lactate
QQ Plot - Mean Lactate
Sample Quantiles
Theoretical Quantiles
Histogram - Max pH
Median: 2.13
Mean: 2.13
Count
Max pH
QQ Plot - Max pH
Sample Quantiles
Theoretical Quantiles


Histogram - Min pH
Median: 2.12
Mean: 2.11
Count
Min pH
QQ Plot - Min pH
Sample Quantiles
Theoretical Quantiles
Histogram - Mean pH
Median: 2.12
Mean: 2.12
Count
Mean pH
QQ Plot - Mean pH
Sample Quantiles
Theoretical Quantiles
Histogram - Max Glucose
Median: 5.16
Mean: 5.19
Count
Max Glucose
QQ Plot - Max Glucose
Sample Quantiles
Theoretical Quantiles
Histogram - Min Glucose
Median: 4.64
Mean: 4.66
Count
Min Glucose
QQ Plot - Min Glucose
Sample Quantiles
Theoretical Quantiles


Histogram - Mean Glucose
Median: 4.92
Mean: 4.96
Count
Mean Glucose
QQ Plot - Mean Glucose
Sample Quantiles
Theoretical Quantiles
Histogram - Max WBC
Median: 2.72
Mean: 2.71
Count
Max WBC
QQ Plot - Max WBC
Sample Quantiles
Theoretical Quantiles
Histogram - Min WBC
Median: 2.40
Mean: 2.36
Count
Min WBC
QQ Plot - Min WBC
Sample Quantiles
Theoretical Quantiles
Histogram - Mean WBC
Median: 2.57
Mean: 2.55
Count
Mean WBC
QQ Plot - Mean WBC
Sample Quantiles
Theoretical Quantiles

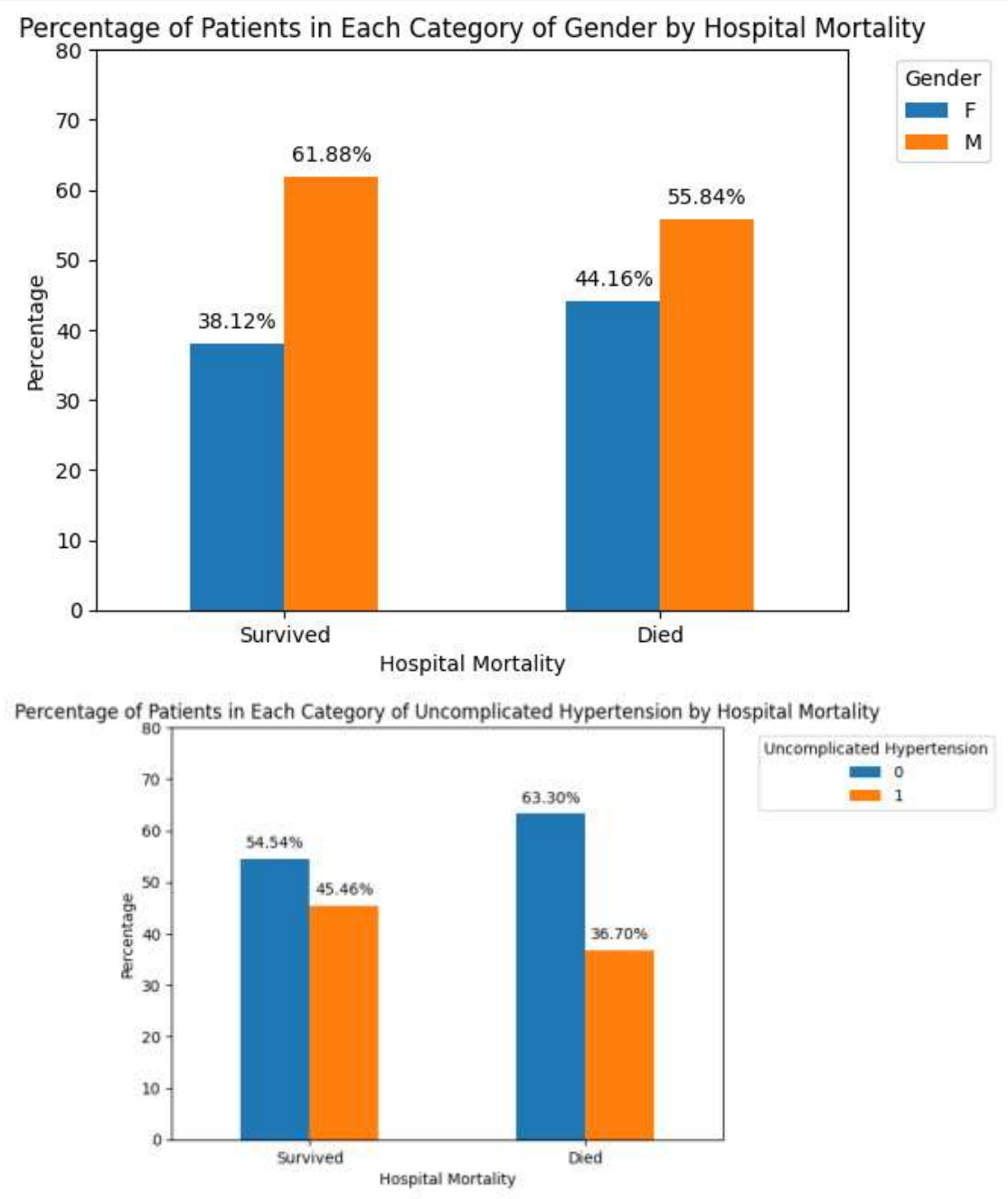

Histogram - Max BUN
Median: 3.00
Mean: 3.10
Count
Max BUN
QQ Plot - Max BUN
Sample Quantiles
Theoretical Quantiles
Histogram - Min BUN
Median: 2.83
Mean: 2.90
Count
Min BUN
QQ Plot - Min BUN
Sample Quantiles
Theoretical Quantiles
Histogram - Mean BUN
Median: 2.92
Mean: 3.01
Count
Mean BUN
QQ Plot - Mean BUN
Sample Quantiles
Theoretical Quantiles
Histogram - Max Creatinine
Median: 0.69
Mean: 0.81
Count
Max Creatinine
QQ Plot - Max Creatinine
Sample Quantiles
Theoretical Quantiles

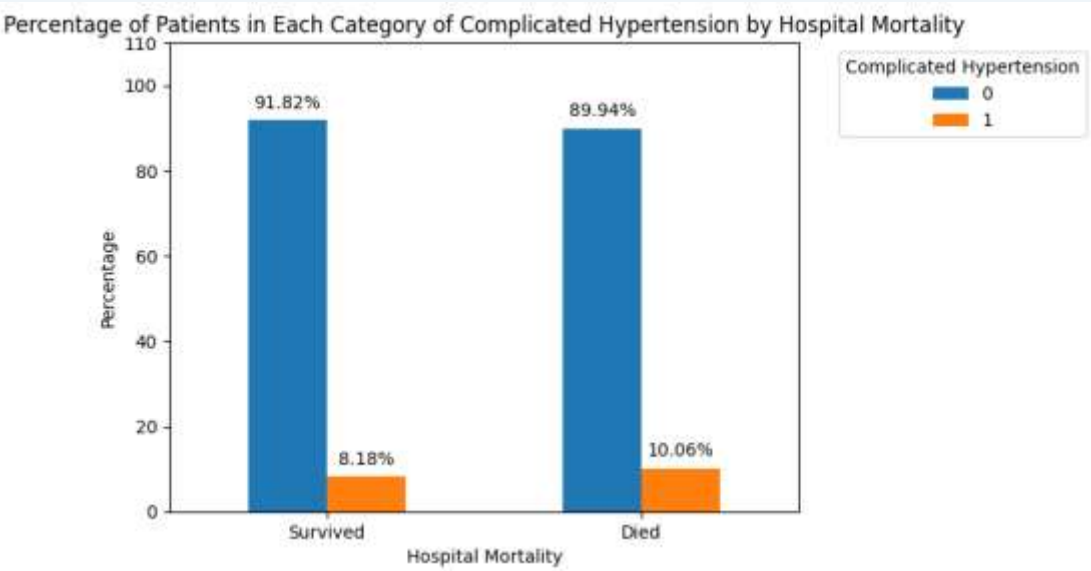

Histogram - Min Creatinine
Median: 0.59
Mean: 0.70
Count
Min Creatinine
QQ Plot - Min Creatinine
Sample Quantiles
Theoretical Quantiles
Histogram - Mean Creatinine
Median: 0.66
Mean: 0.75
Count
Mean Creatinine
QQ Plot - Mean Creatinine
Sample Quantiles
Theoretical Quantiles
Histogram - Max Hemoglobin
Median: 2.60
Mean: 2.59
Count
Max Hemoglobin
QQ Plot - Max Hemoglobin
Sample Quantiles
Theoretical Quantiles
Histogram - Min Hemoglobin
Median: 2.36
Mean: 2.35
Count
Min Hemoglobin
QQ Plot - Min Hemoglobin
Sample Quantiles
Theoretical Quantiles

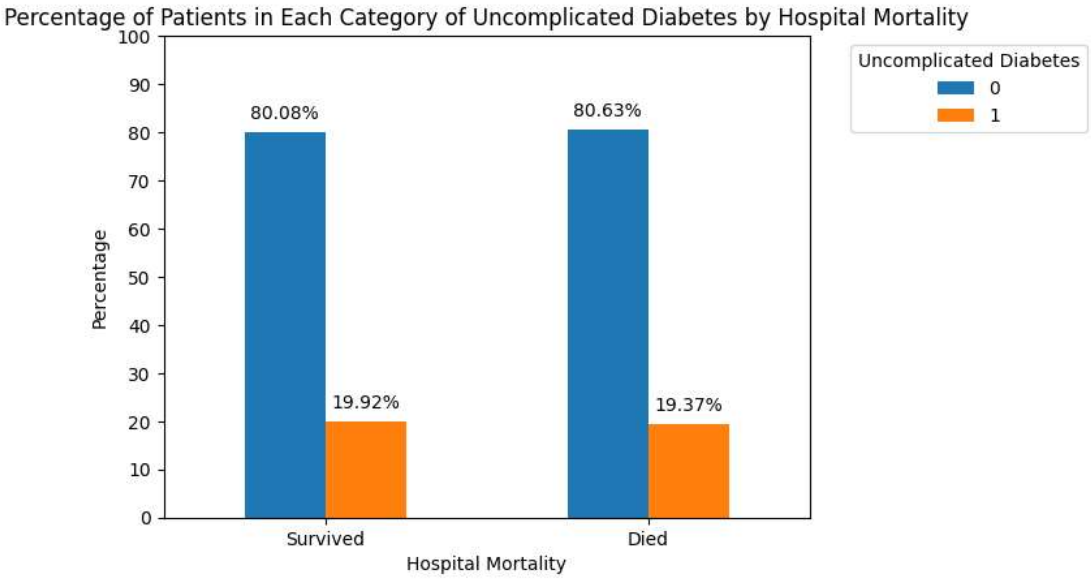

Histogram - Mean Hemoglobin
Median: 2.46
Mean: 2.47
Count
Mean Hemoglobin
QQ Plot - Mean Hemoglobin
Sample Quantiles
Theoretical Quantiles

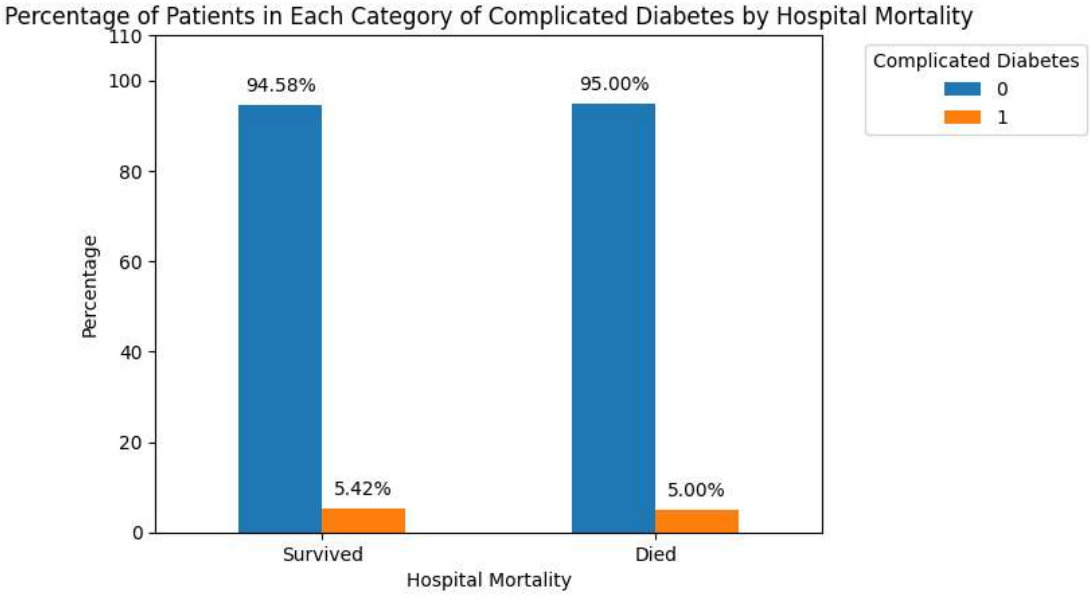
Visualization of Categorical Variables

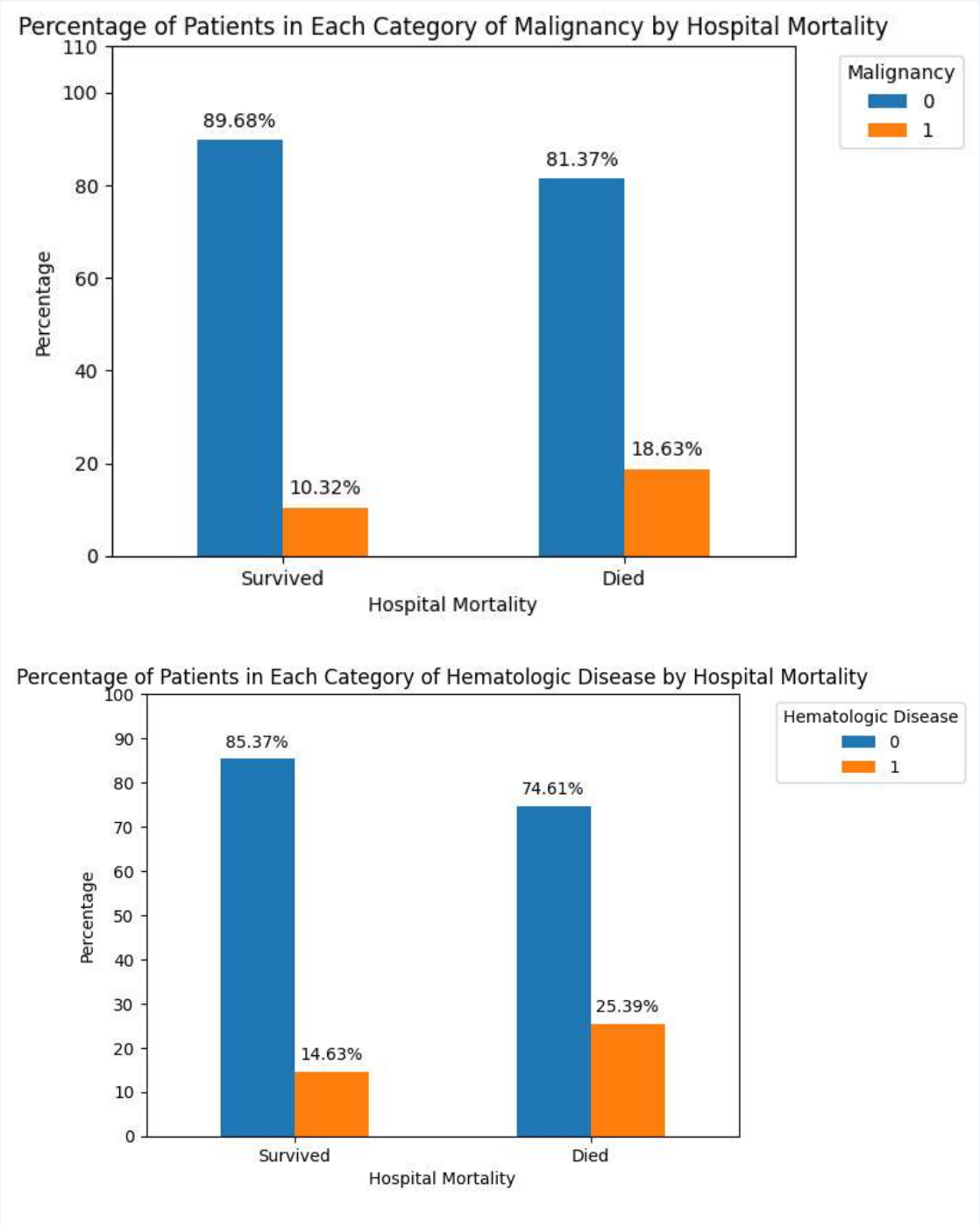
Bar Graphs

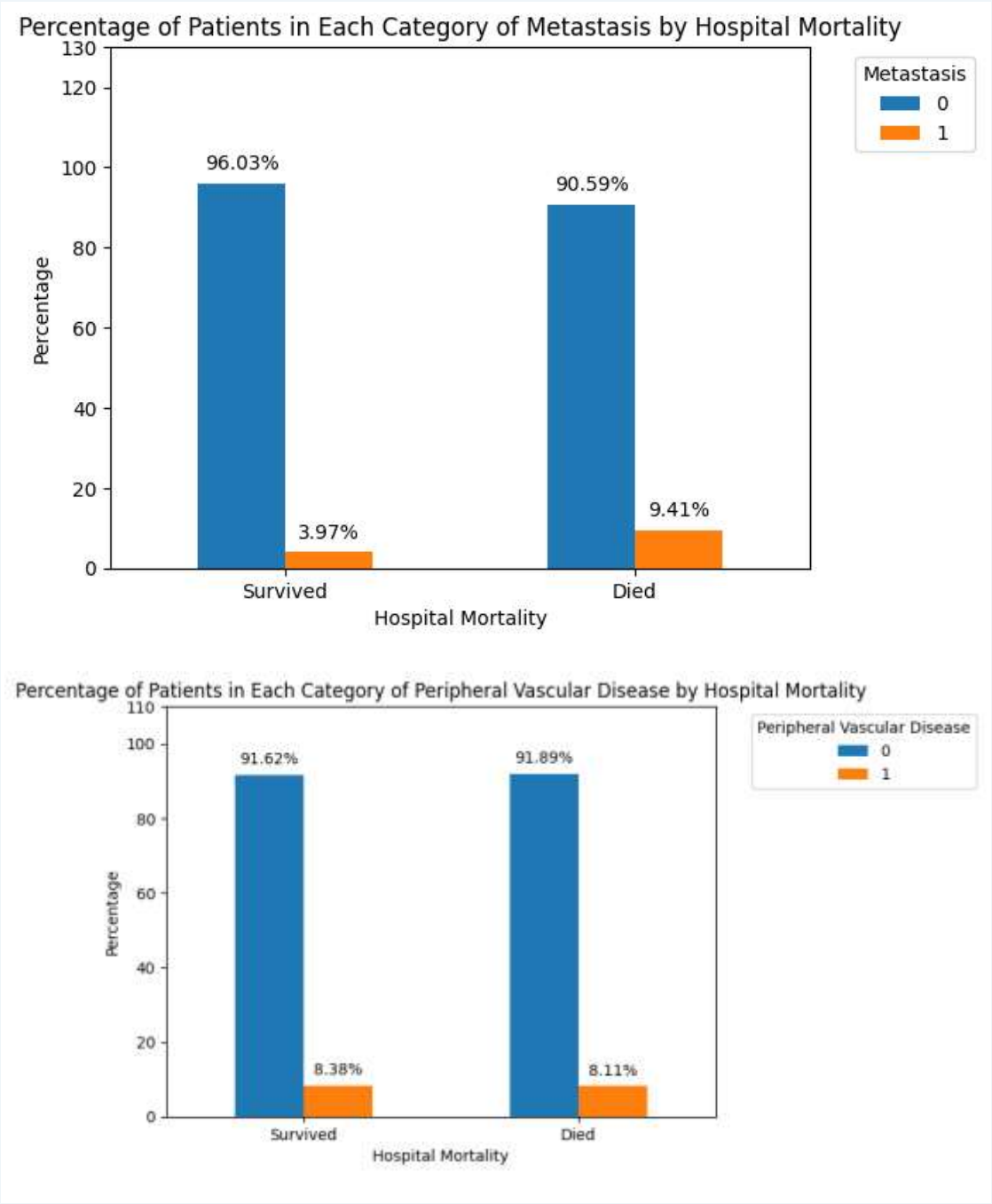


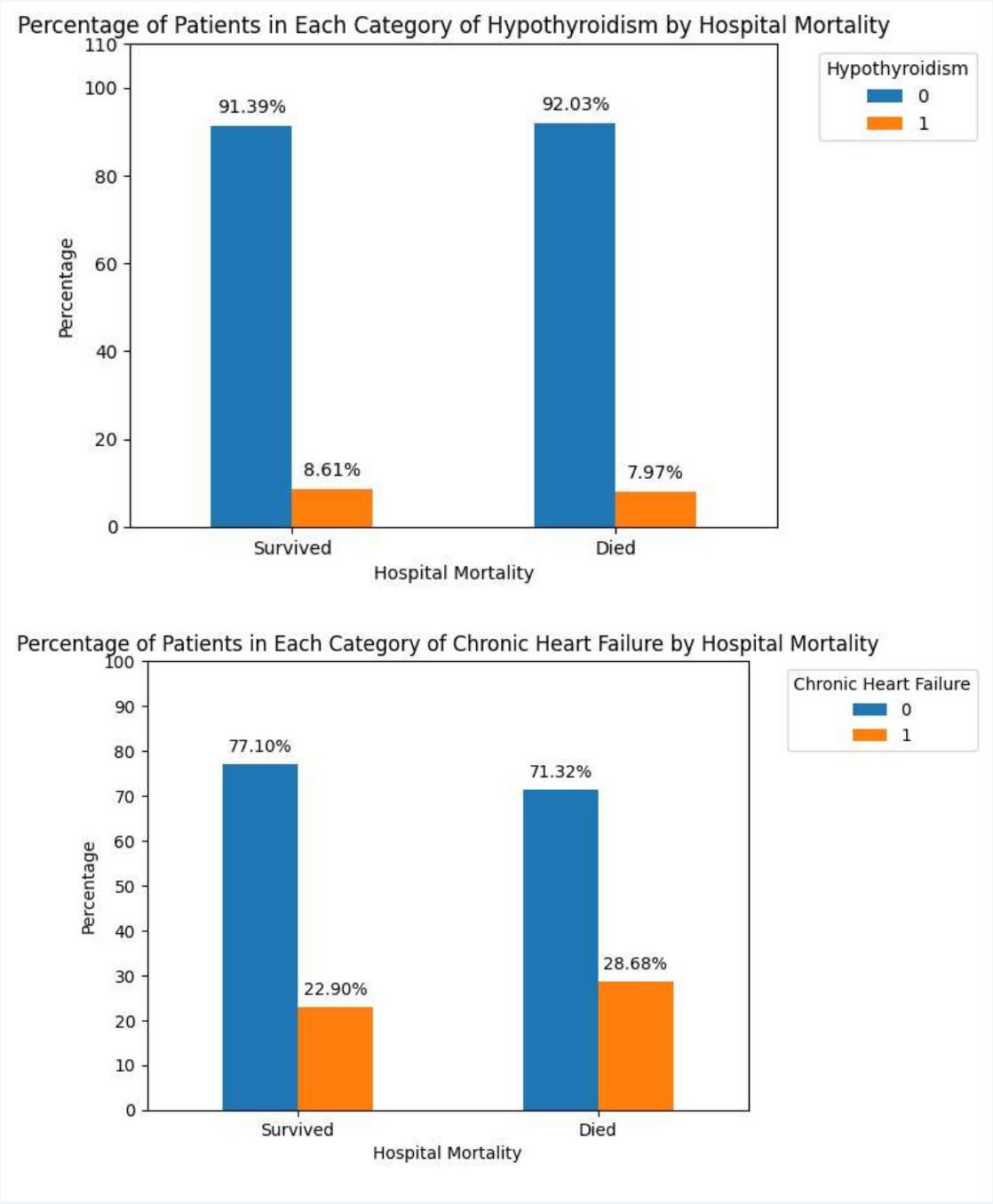

Percentage of Patients in Each Category of Complicated Hypertension by Hospital Mortality
Complicated Hypertension
0
1
Percentage
Hospital Mortality
Survived
Died
91.82%
8.18%
89.94%
10.06%

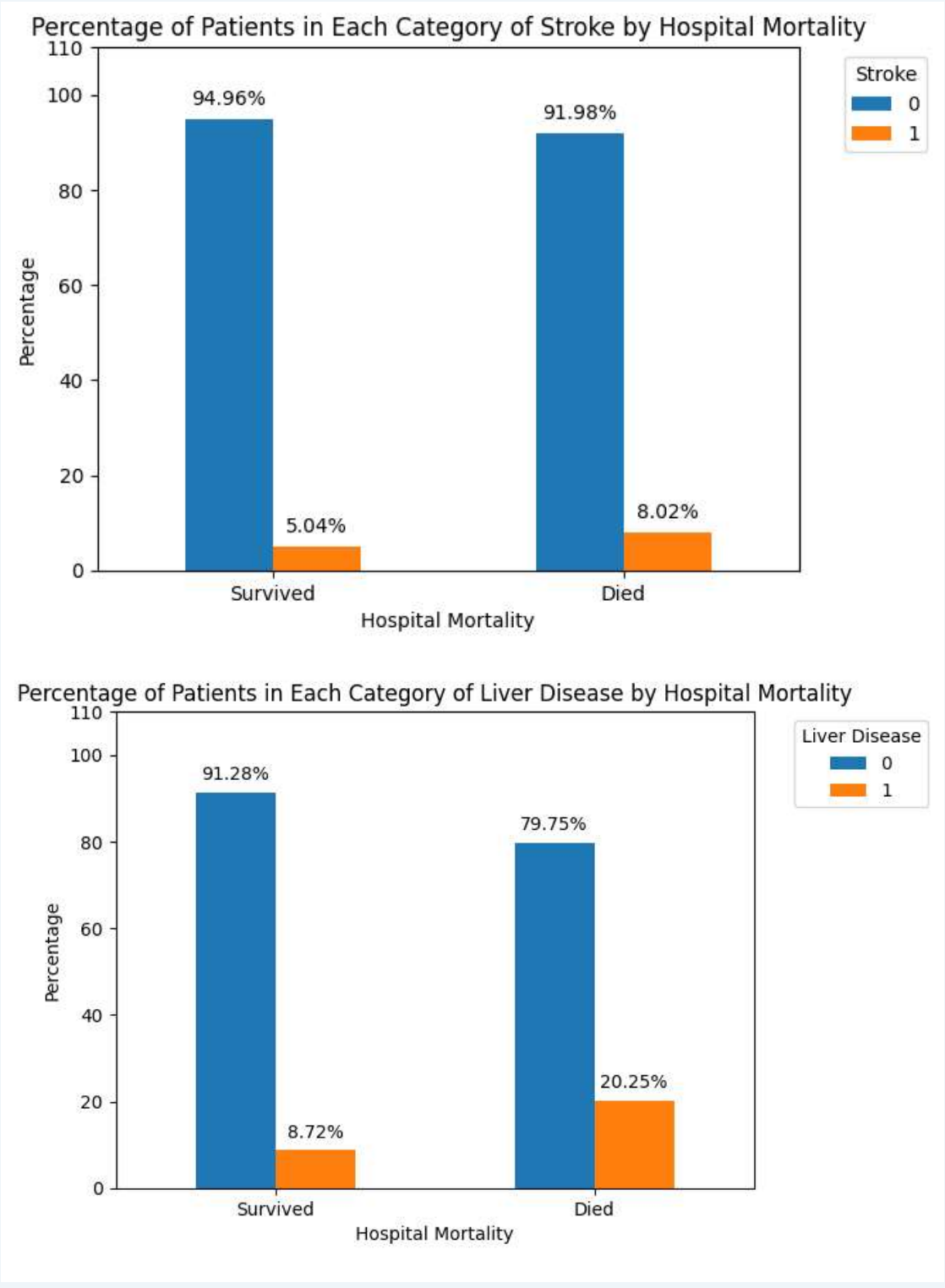

Percentage of Patients in Each Category of Uncomplicated Diabetes by Hospital Mortality
Uncomplicated Diabetes
0
1
Percentage
Hospital Mortality
Survived
Died
80.08%
19.92%
80.63%
19.37%

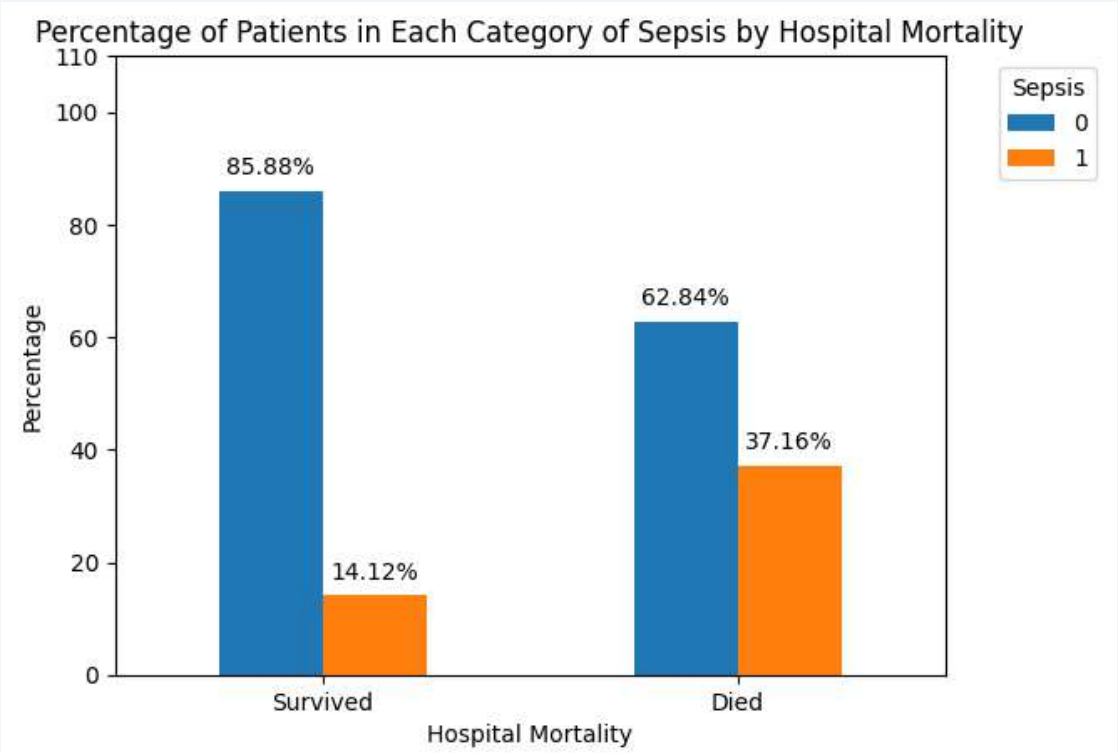

Percentage of Patients in Each Category of Complicated Diabetes by Hospital Mortality
Complicated Diabetes
0
1
Percentage
Hospital Mortality
Survived
Died
94.58%
5.42%
95.00%
5.00%

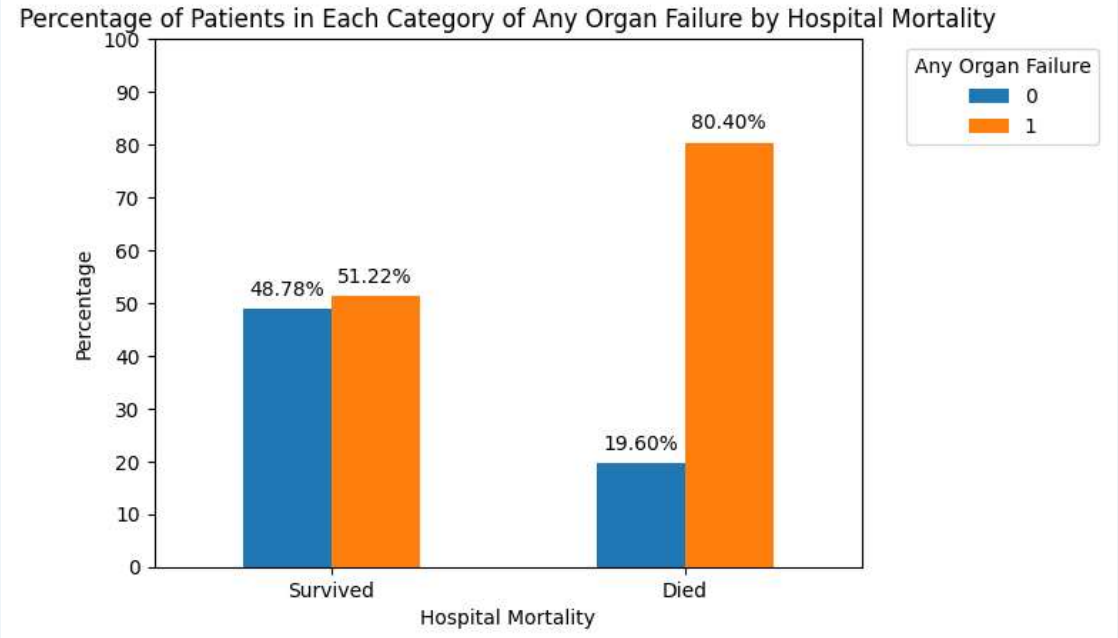

Percentage of Patients in Each Category of Malignancy by Hospital Mortality
Malignancy
0
1
Percentage
110
100
80
60
40
20
0
89.68%
10.32%
81.37%
18.63%
Survived
Died
Hospital Mortality
Percentage of Patients in Each Category of Hematologic Disease by Hospital Mortality
Hematologic Disease
0
1
Percentage
100
90
80
70
60
50
40
30
20
10
0
85.37%
14.63%
74.61%
25.39%
Survived
Died
Hospital Mortality

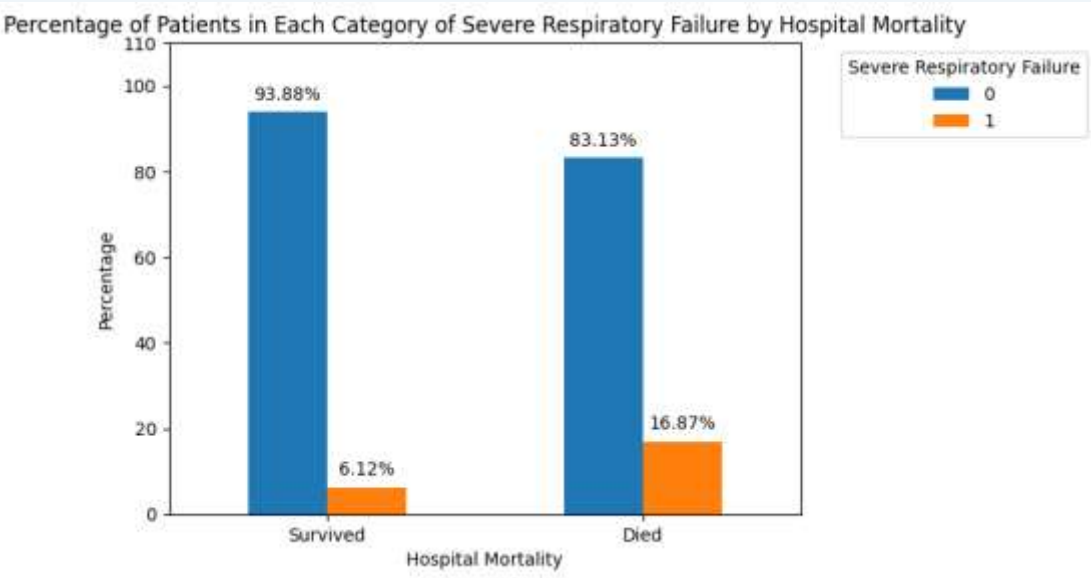

Percentage of Patients in Each Category of Metastasis by Hospital Mortality
Metastasis
0
1
96.03%
3.97%
90.59%
9.41%
Percentage
Survived
Died
Hospital Mortality
Percentage of Patients in Each Category of Peripheral Vascular Disease by Hospital Mortality
Peripheral Vascular Disease
0
1
91.62%
8.38%
91.89%
8.11%
Percentage
Survived
Died
Hospital Mortality

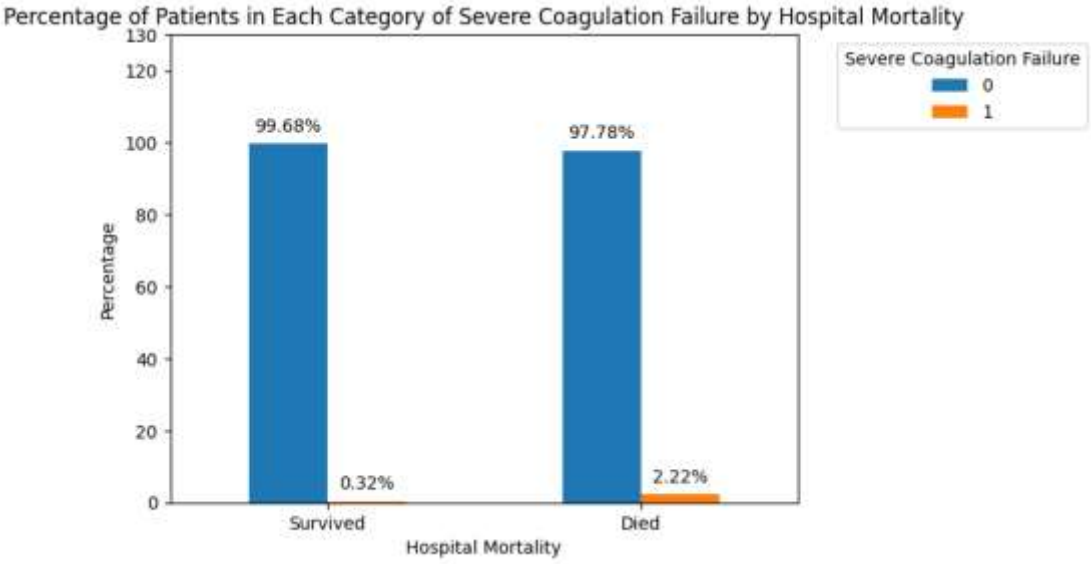

Percentage of Patients in Each Category of Hypothyroidism by Hospital Mortality
Hypothyroidism
0
1
Percentage
110
100
80
60
40
20
0
91.39%
8.61%
92.03%
7.97%
Survived
Died
Hospital Mortality
Percentage of Patients in Each Category of Chronic Heart Failure by Hospital Mortality
Chronic Heart Failure
0
1
Percentage
100
90
80
70
60
50
40
30
20
10
0
77.10%
22.90%
71.32%
28.68%
Survived
Died
Hospital Mortality

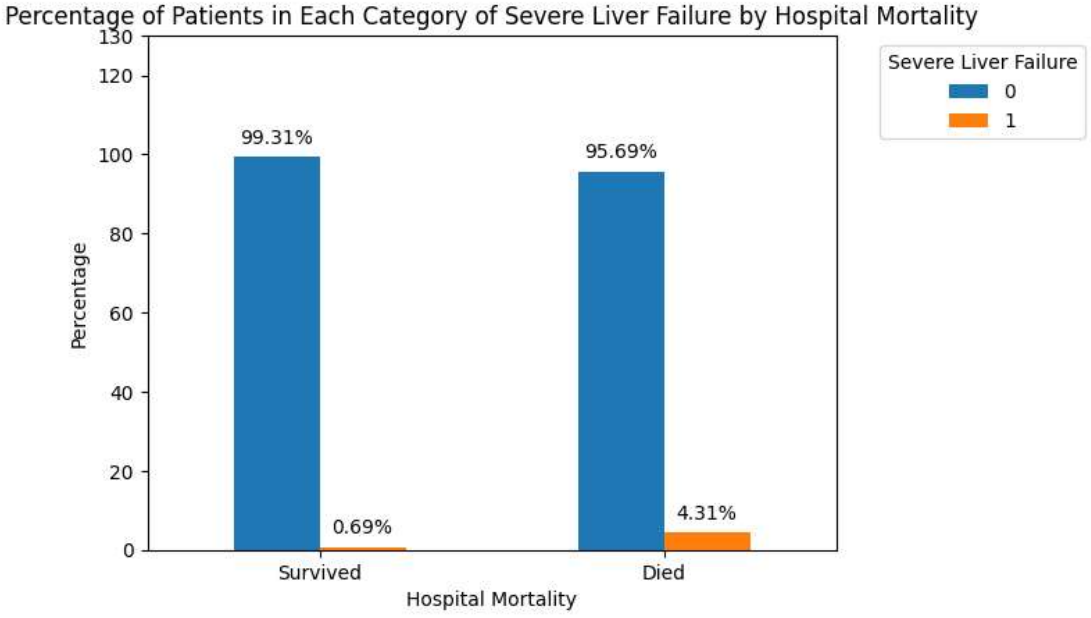

Percentage of Patients in Each Category of Stroke by Hospital Mortality
Stroke
0
1
Percentage
110
100
80
60
40
20
0
94.96%
5.04%
91.98%
8.02%
Survived
Died
Hospital Mortality
Percentage of Patients in Each Category of Liver Disease by Hospital Mortality
Liver Disease
0
1
Percentage
110
100
80
60
40
20
0
91.28%
8.72%
79.75%
20.25%
Survived
Died
Hospital Mortality

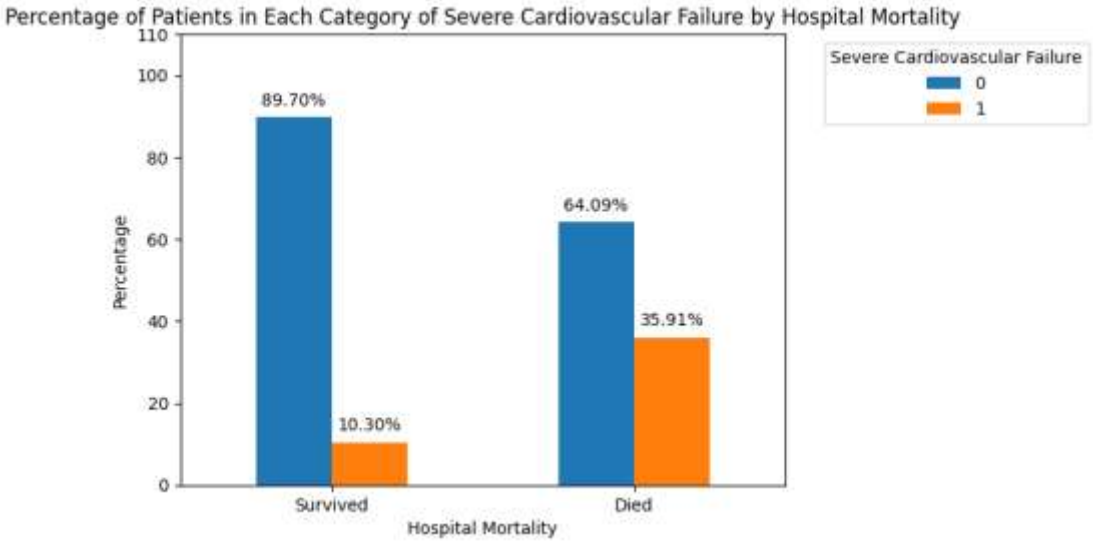

Percentage of Patients in Each Category of Sepsis by Hospital Mortality
Sepsis
0
1
85.88%
14.12%
62.84%
37.16%
Percentage
Survived
Died
Hospital Mortality

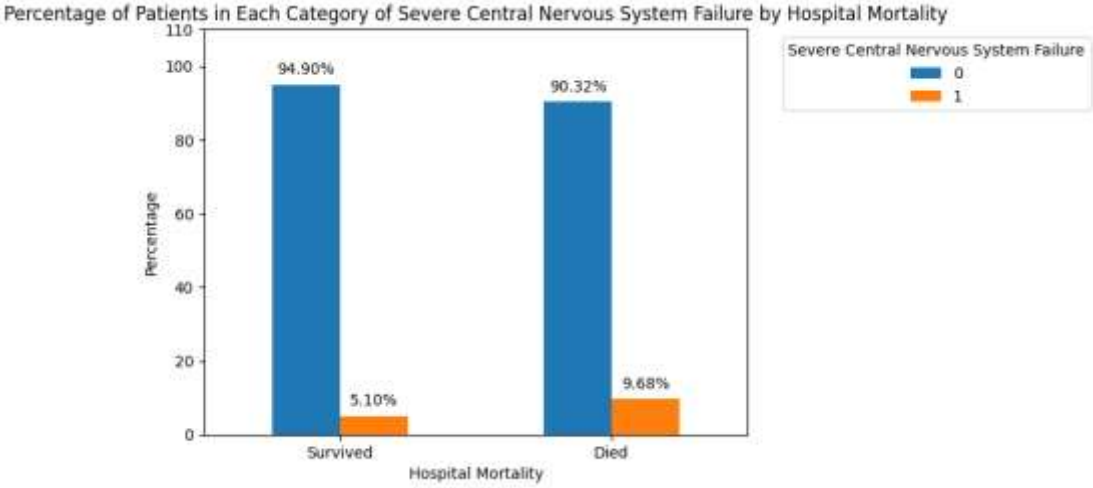

Percentage of Patients in Each Category of Any Organ Failure by Hospital Mortality
Any Organ Failure
0
1
48.78%
51.22%
19.60%
80.40%
Percentage
Survived
Died
Hospital Mortality

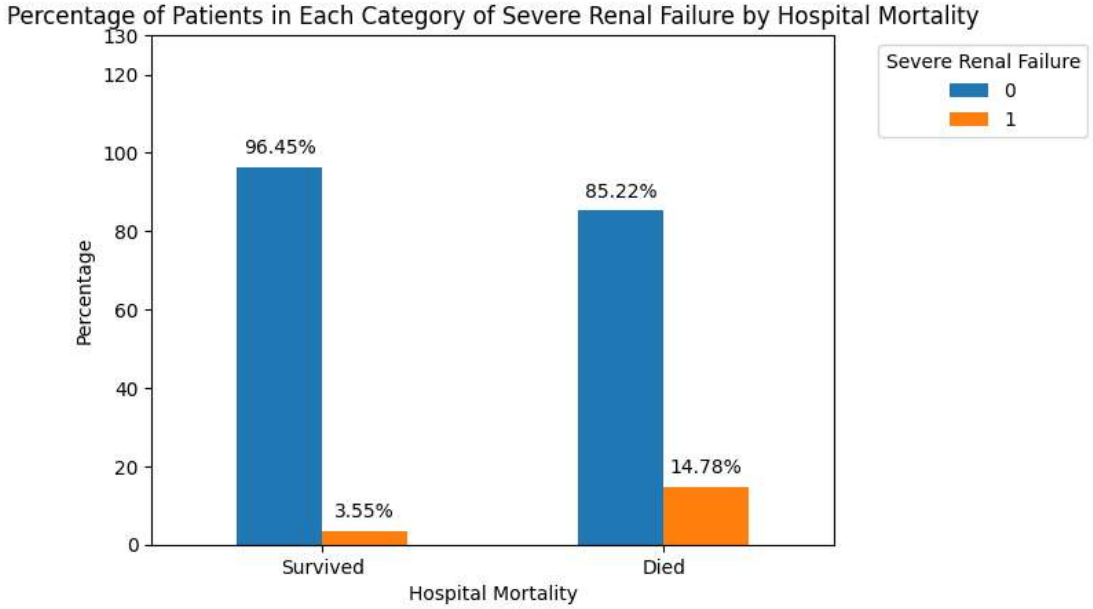

Percentage of Patients in Each Category of Severe Respiratory Failure by Hospital Mortality
Severe Respiratory Failure
0
1
93.88%
6.12%
83.13%
16.87%
Percentage
Survived
Died
Hospital Mortality


Percentage of Patients in Each Category of Severe Coagulation Failure by Hospital Mortality
Severe Coagulation Failure
0
1
99.68%
0.32%
97.78%
2.22%
Percentage
Survived
Died
Hospital Mortality

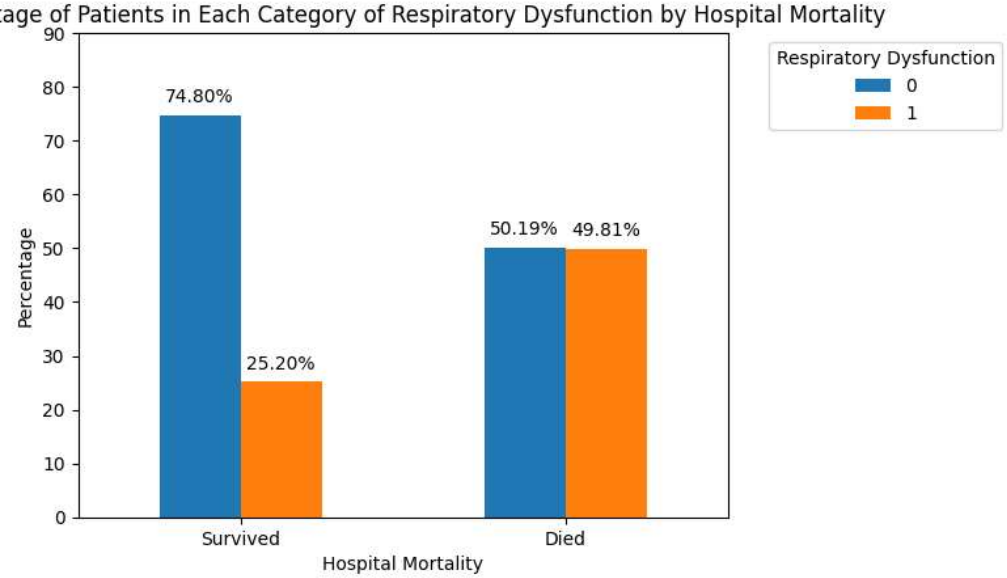

Percentage of Patients in Each Category of Severe Liver Failure by Hospital Mortality
Severe Liver Failure
0
1
99.31%
0.69%
95.69%
4.31%
Percentage
Survived
Died
Hospital Mortality


Percentage of Patients in Each Category of Severe Cardiovascular Failure by Hospital Mortality
Severe Cardiovascular Failure
0
1
Percentage
89.70%
10.30%
64.09%
35.91%
Survived
Died
Hospital Mortality

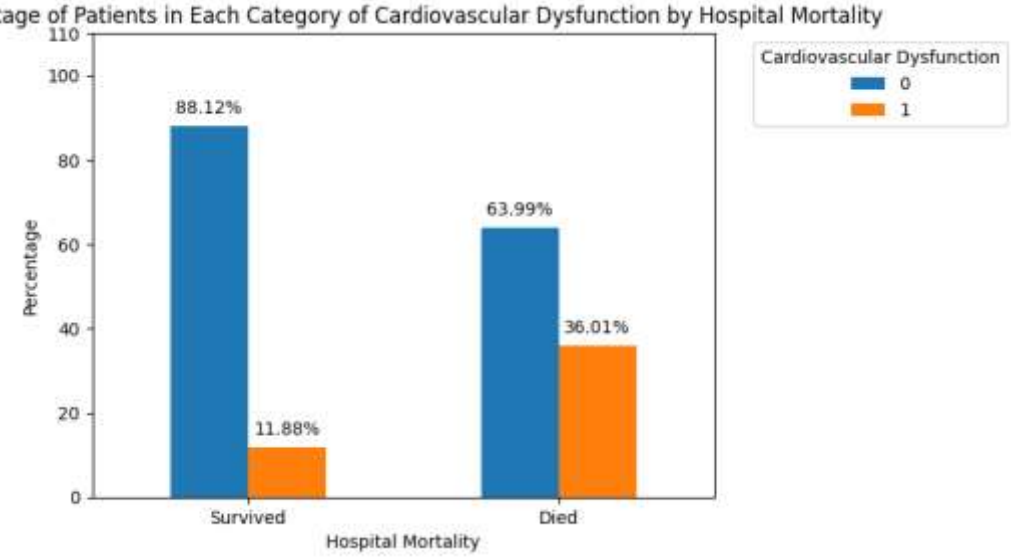

Percentage of Patients in Each Category of Severe Central Nervous System Failure by Hospital Mortality
Severe Central Nervous System Failure
0
1
Percentage
94.90%
5.10%
90.32%
9.68%
Survived
Died
Hospital Mortality


Percentage of Patients in Each Category of Severe Renal Failure by Hospital Mortality
Severe Renal Failure
0
1
Percentage
96.45%
3.55%
85.22%
14.78%
Survived
Died
Hospital Mortality

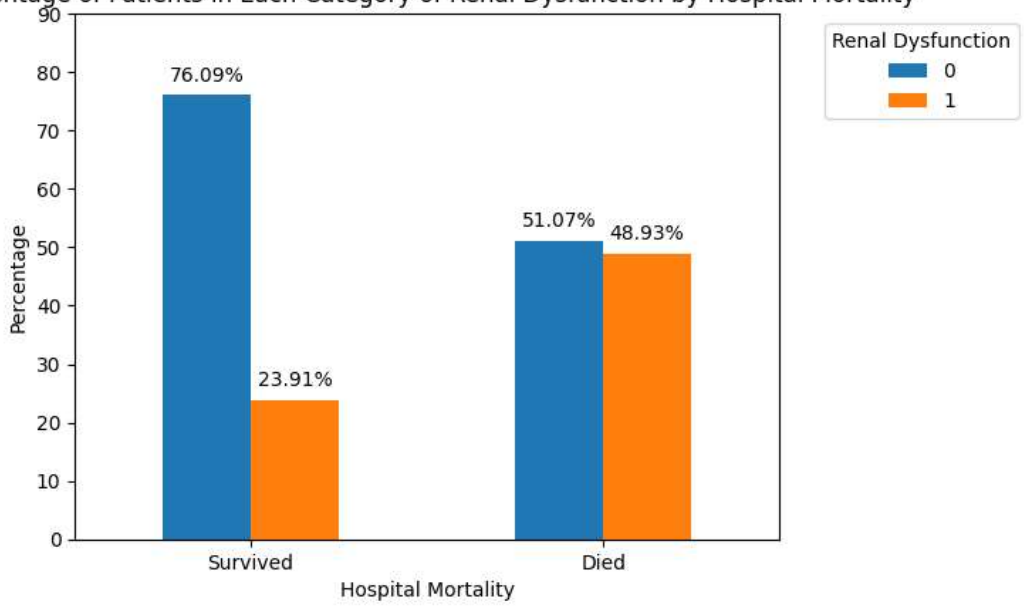
Percentage of Patients in Each Category of Respiratory Dysfunction by Hospital Mortality



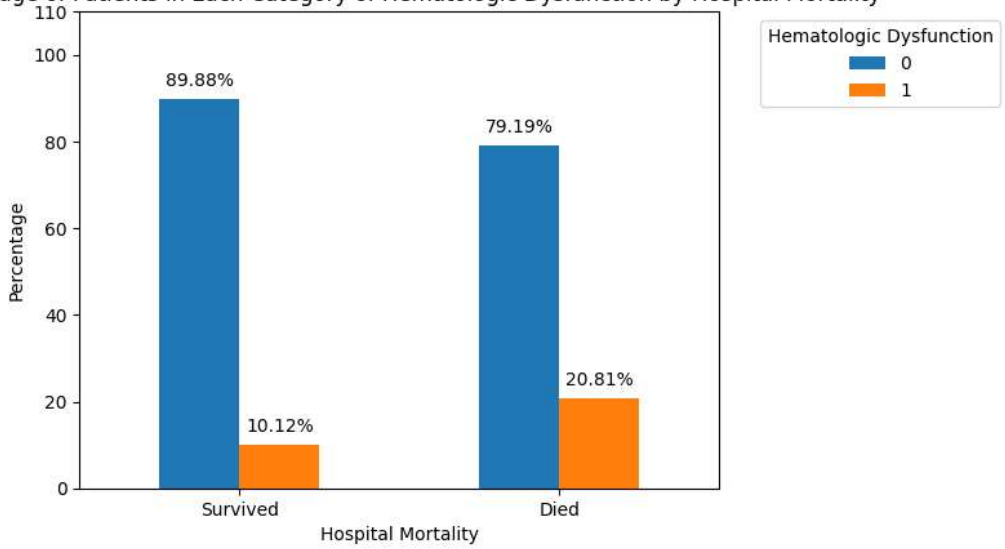
Percentage of Patients in Each Category of Cardiovascular Dysfunction by Hospital Mortality



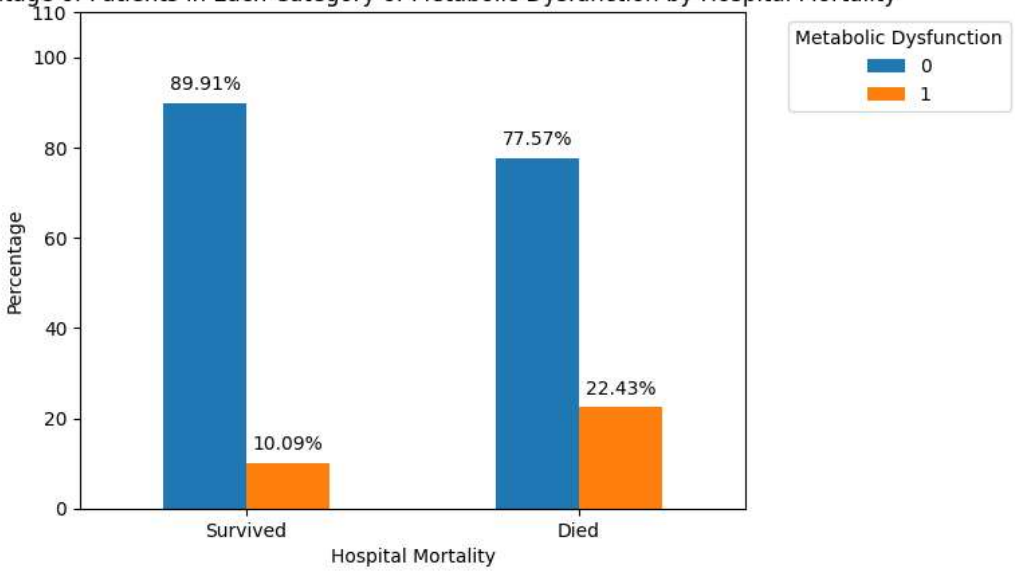
Percentage of Patients in Each Category of Renal Dysfunction by Hospital Mortality



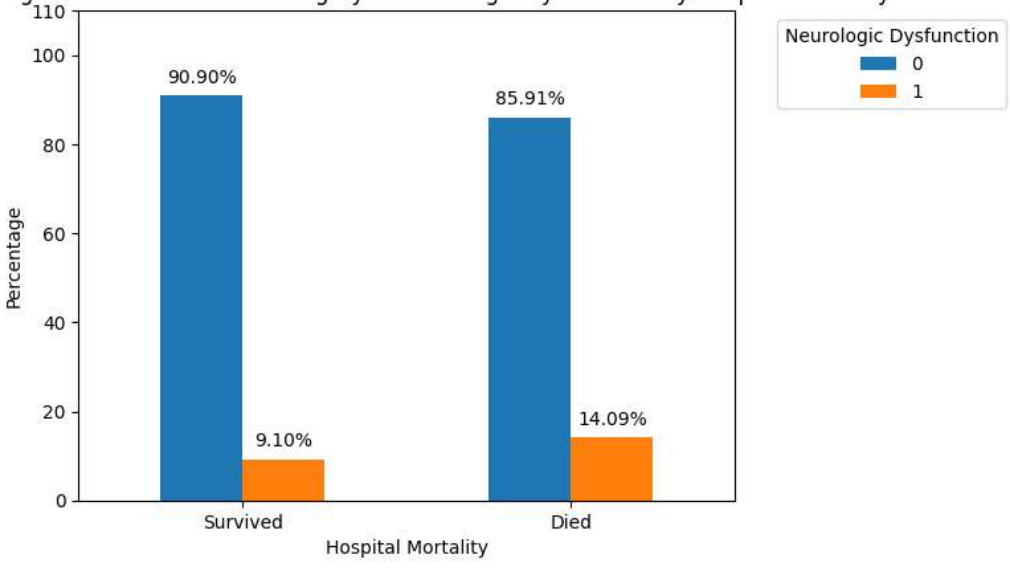
Percentage of Patients in Each Category of Hematologic Dysfunction by Hospital Mortality



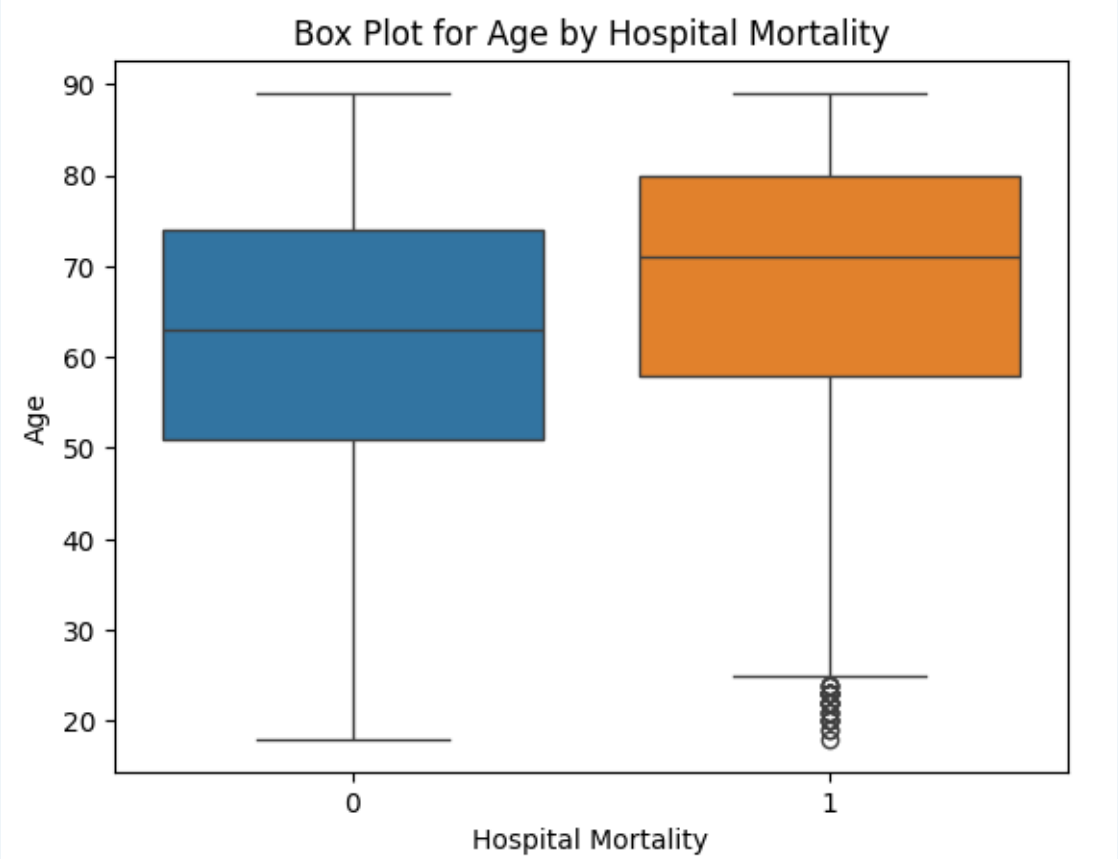
Percentage of Patients in Each Category of Metabolic Dysfunction by Hospital Mortality

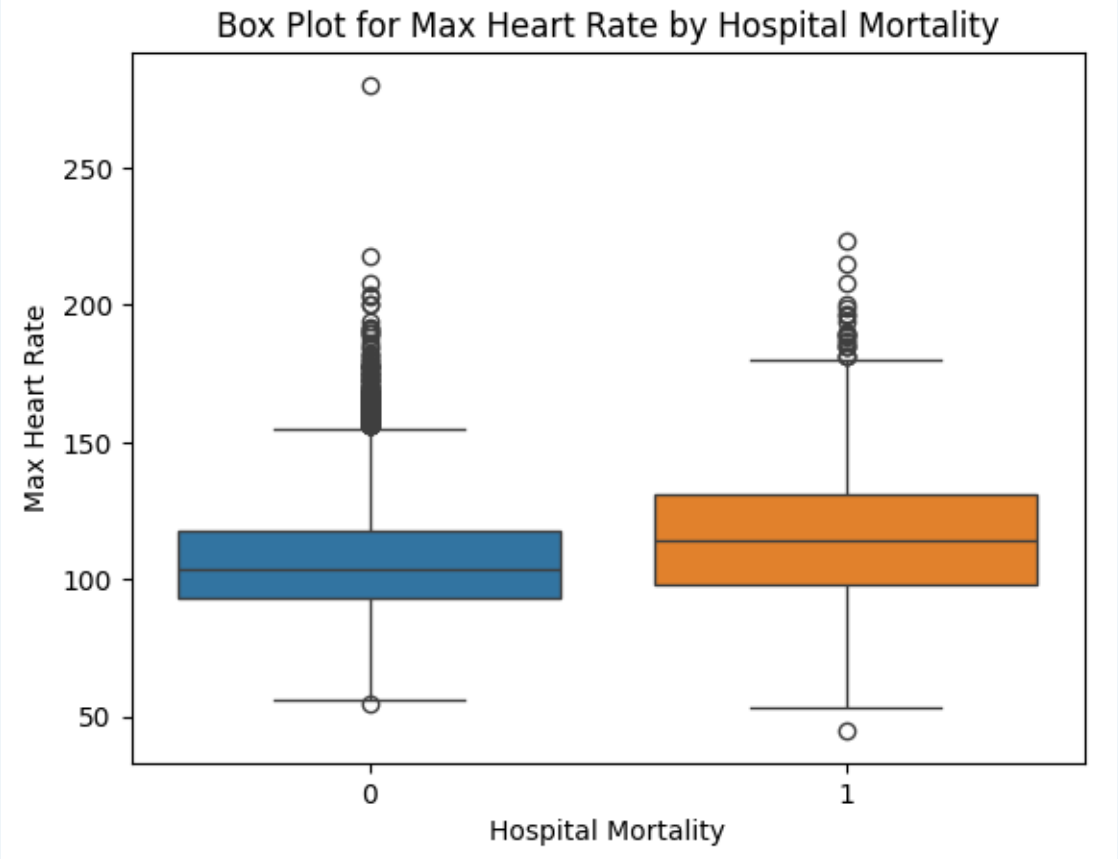


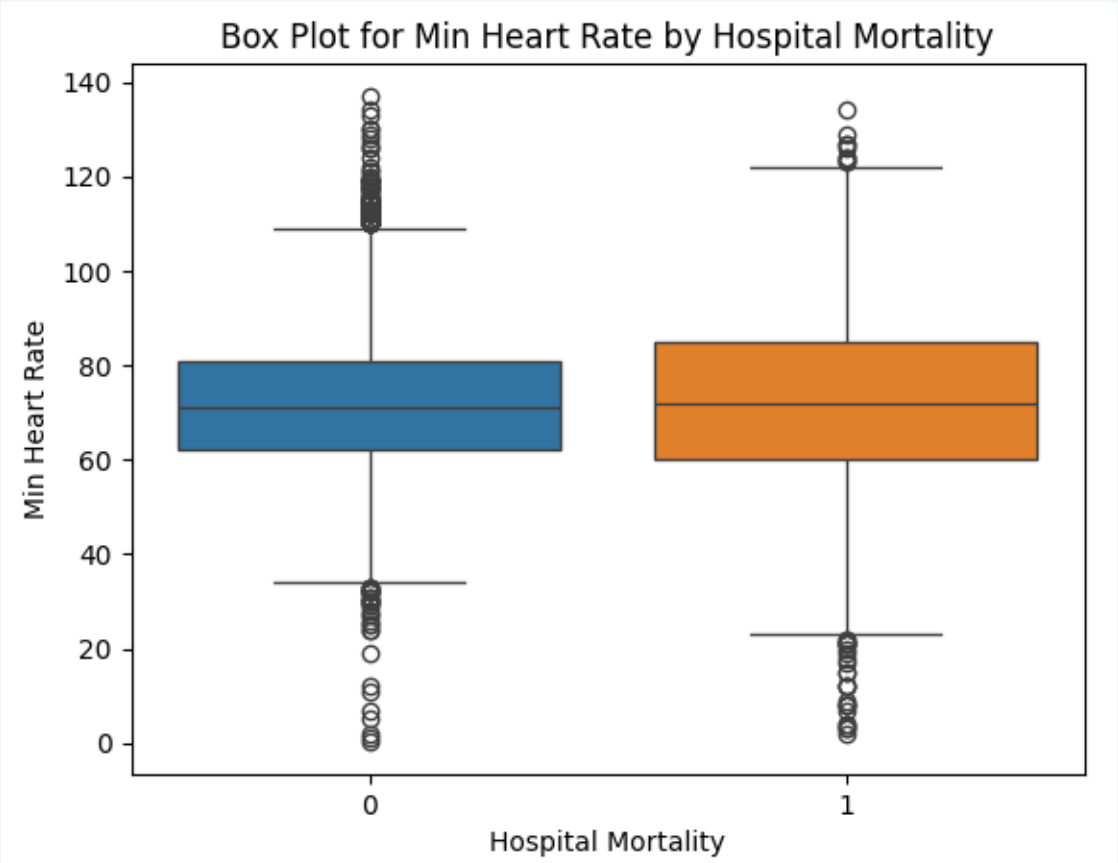
Percentage of Patients in Each Category of Neurologic Dysfunction by Hospital Mortality

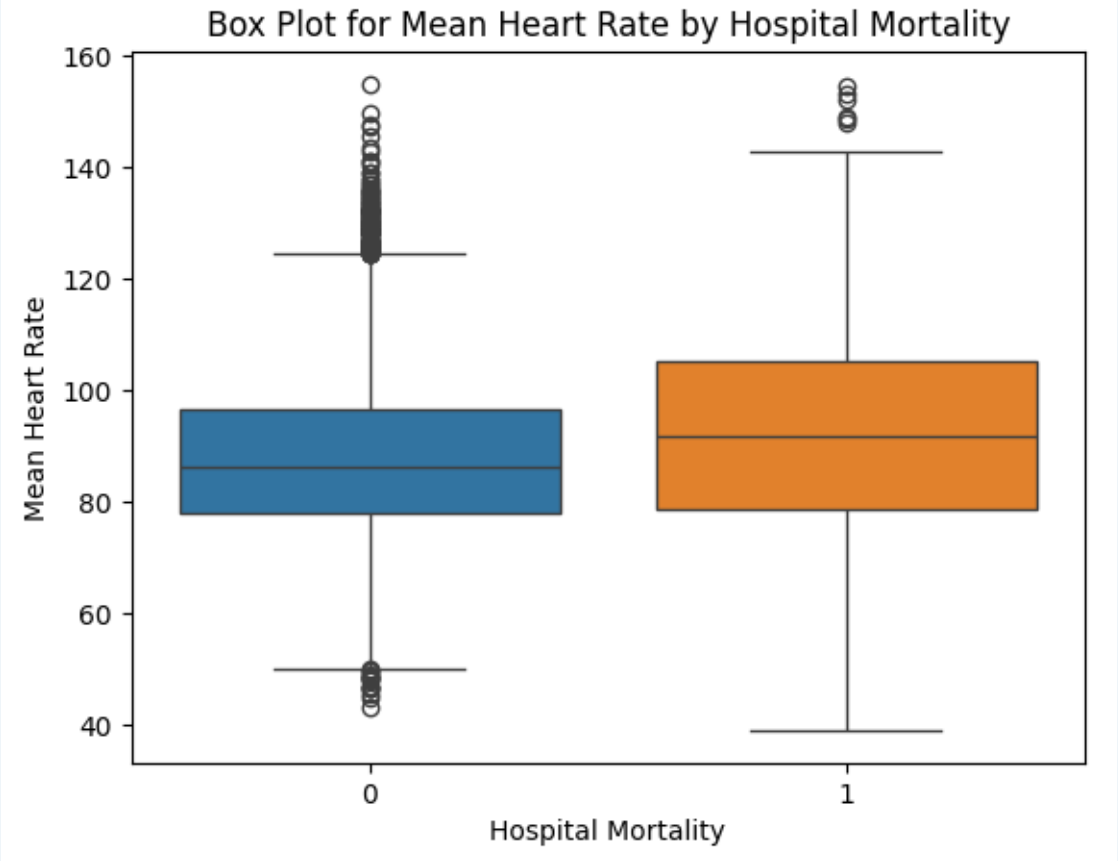


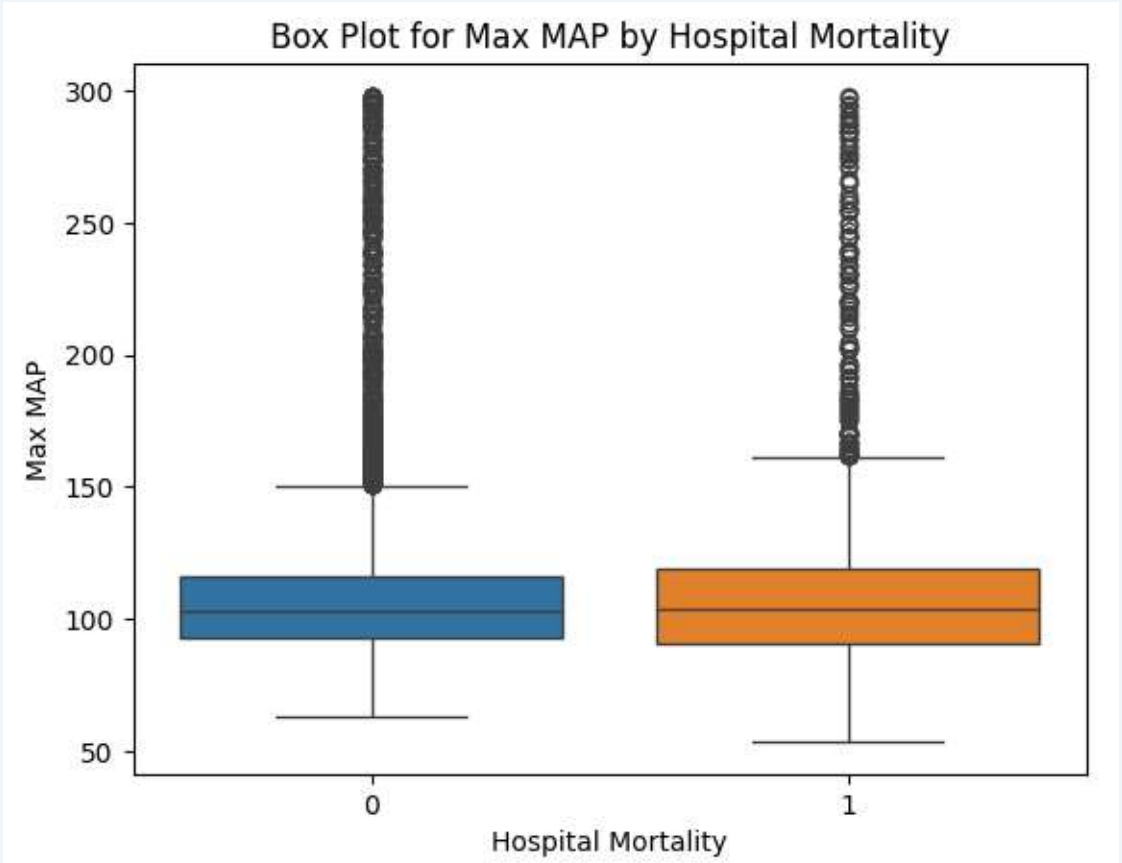
Box Plots

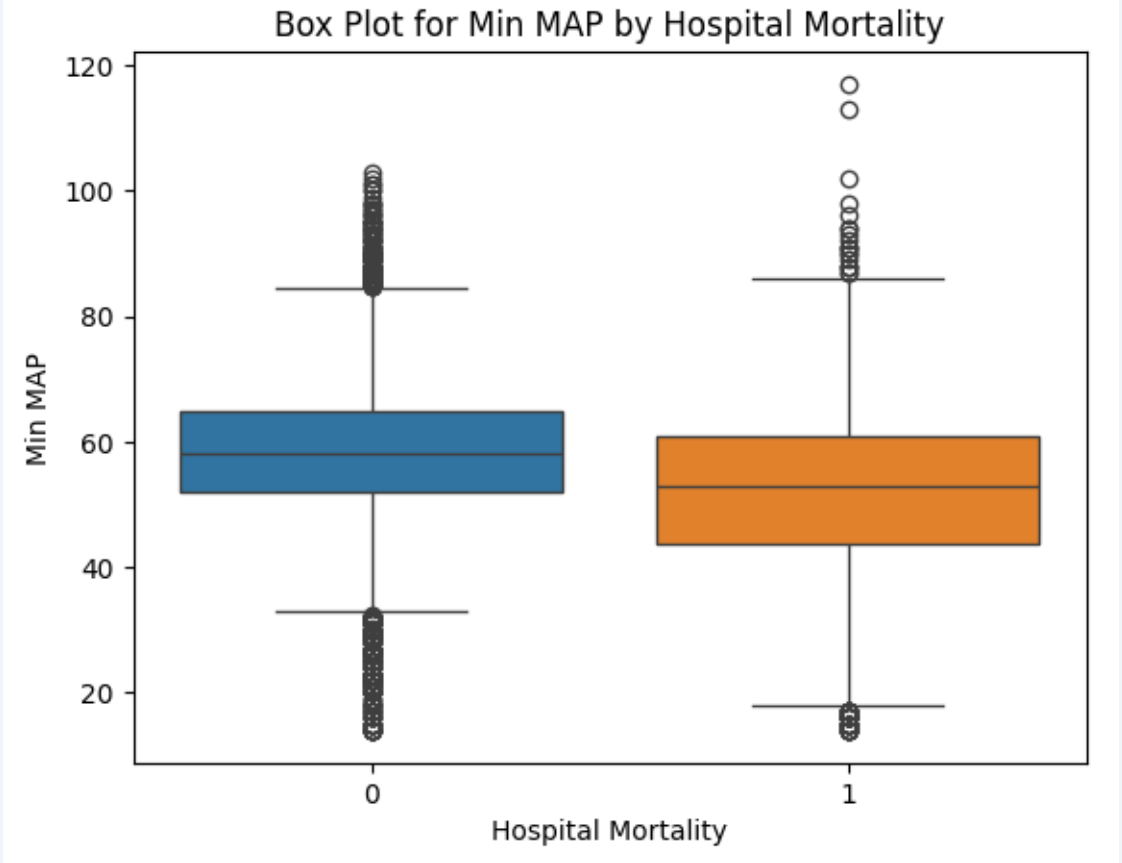

Box Plot for Age by Hospital Mortality
Age
90
80
70
60
50
40
30
20
0
1
Hospital Mortality

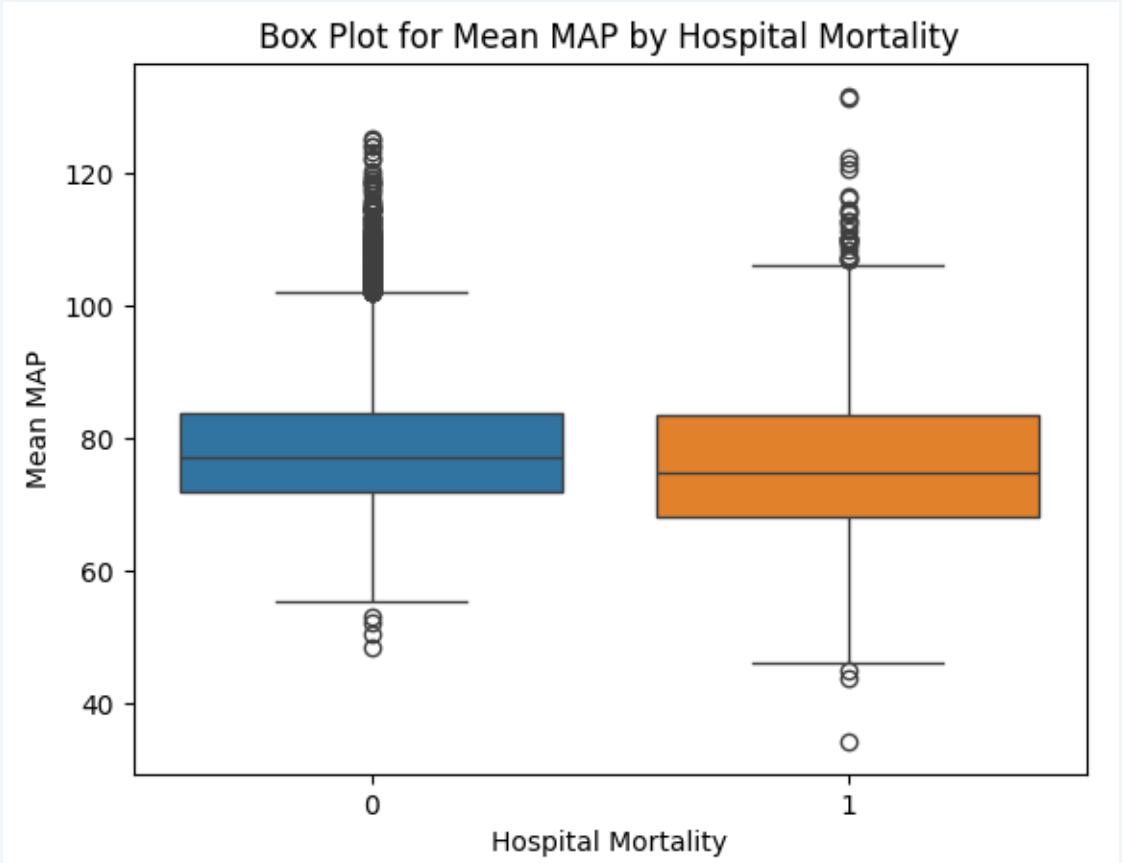

Box Plot for Max Heart Rate by Hospital Mortality
Max Heart Rate
250
200
150
100
50
0
1
Hospital Mortality

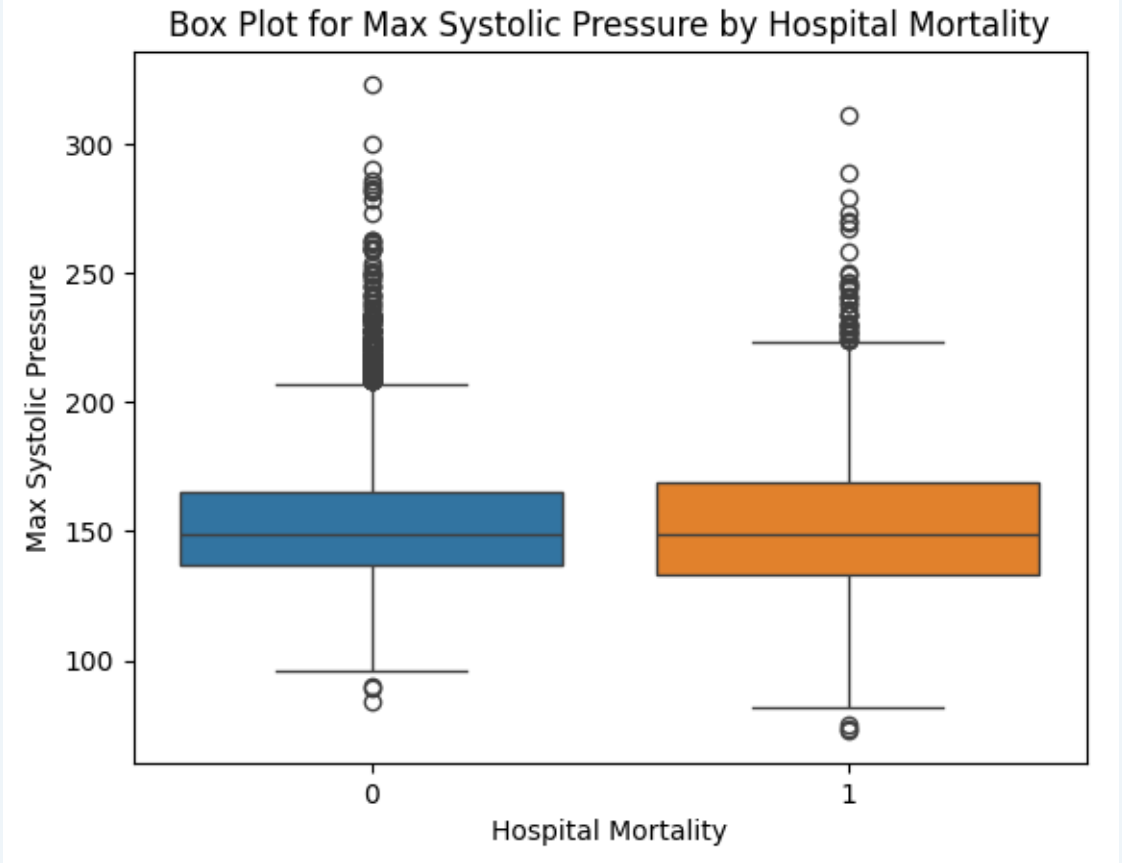

Box Plot for Min Heart Rate by Hospital Mortality
Min Heart Rate
140
120
100
80
60
40
20
0
0
1
Hospital Mortality

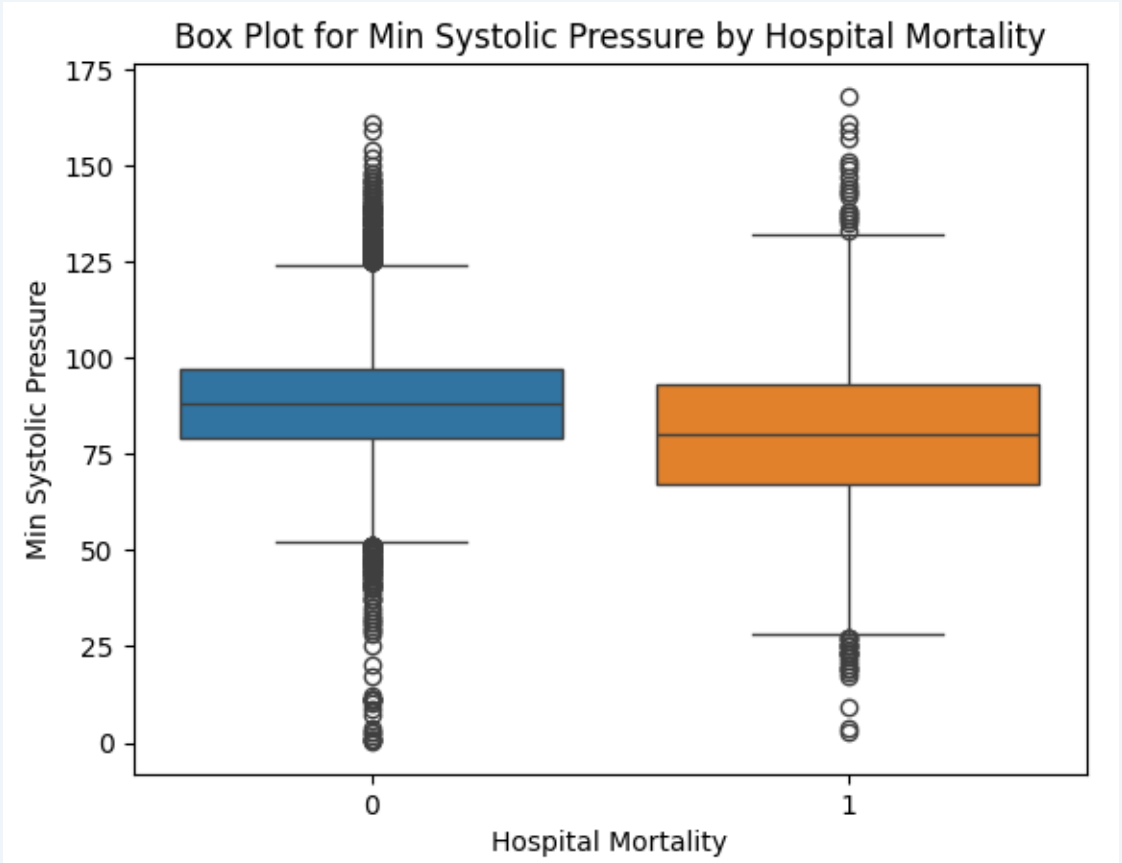

Box Plot for Mean Heart Rate by Hospital Mortality
Mean Heart Rate
160
140
120
100
80
60
40
0
1
Hospital Mortality

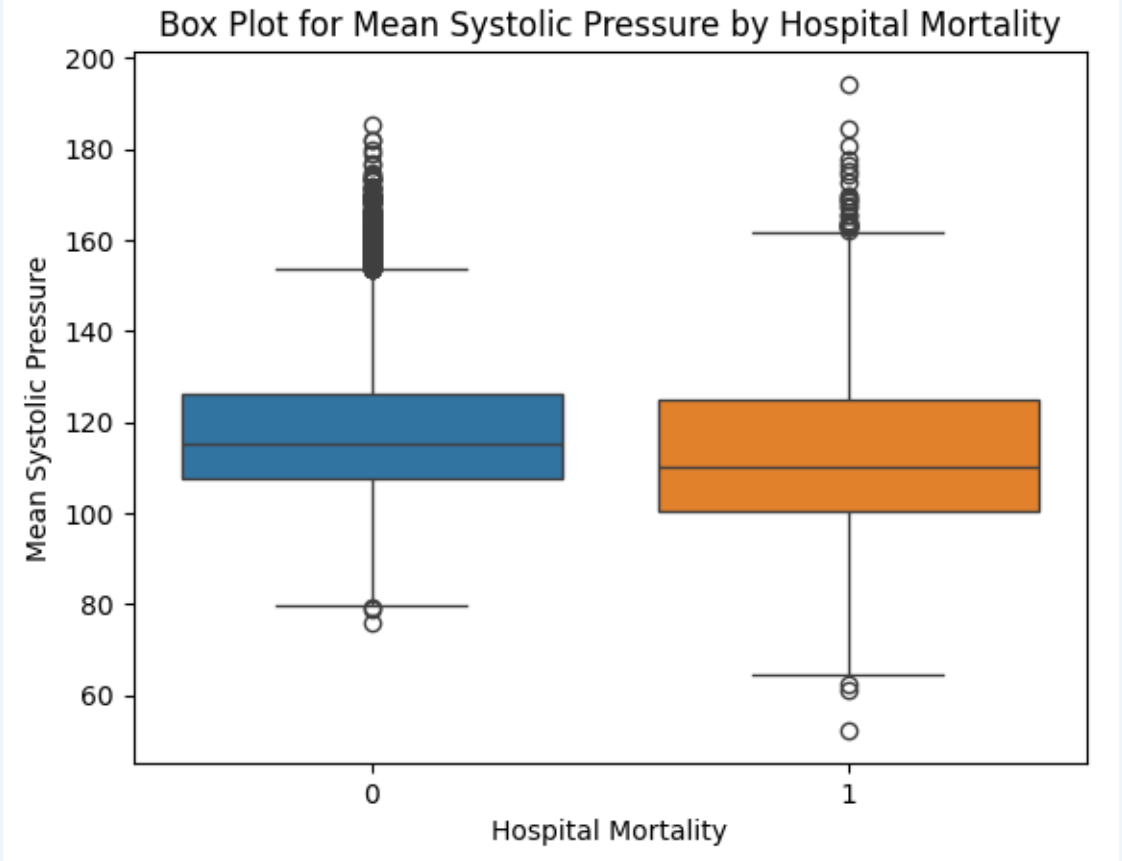

Box Plot for Max MAP by Hospital Mortality
Max MAP
300
250
200
150
100
50
0
1
Hospital Mortality

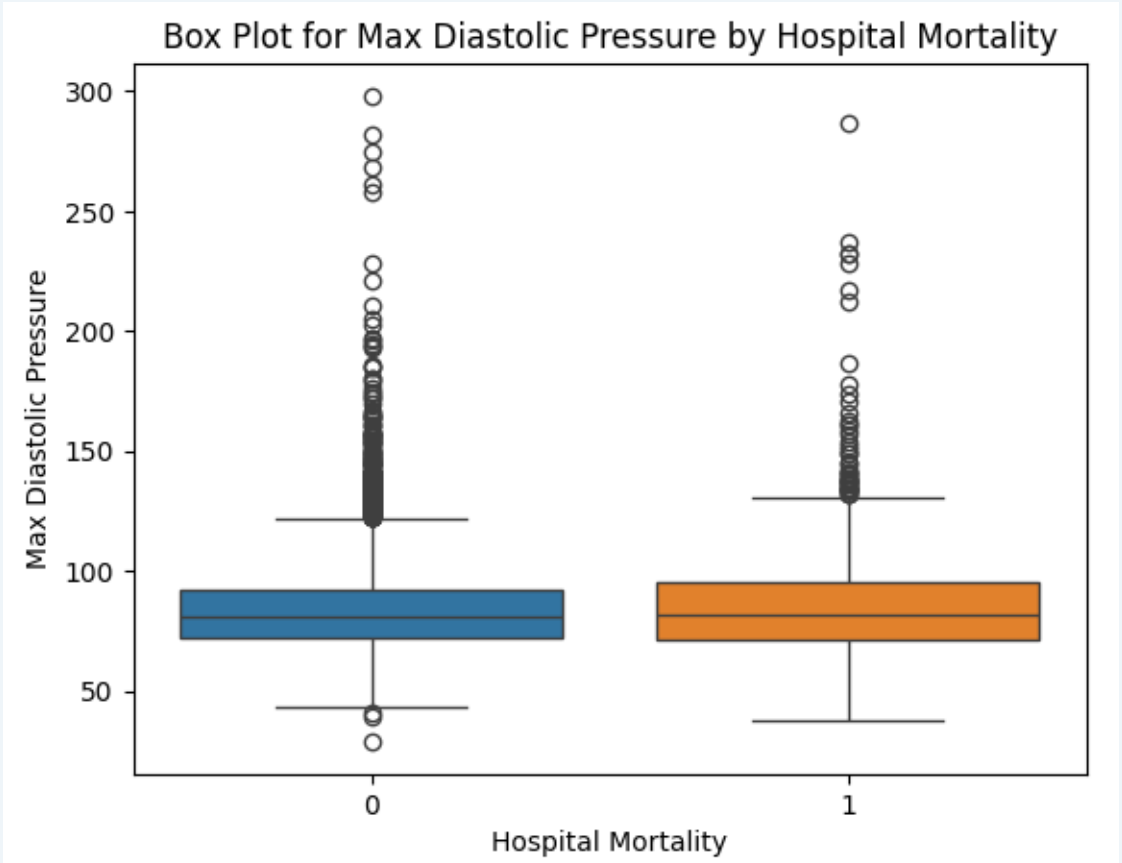

Box Plot for Min MAP by Hospital Mortality
Min MAP
120
100
80
60
40
20
0
1
Hospital Mortality

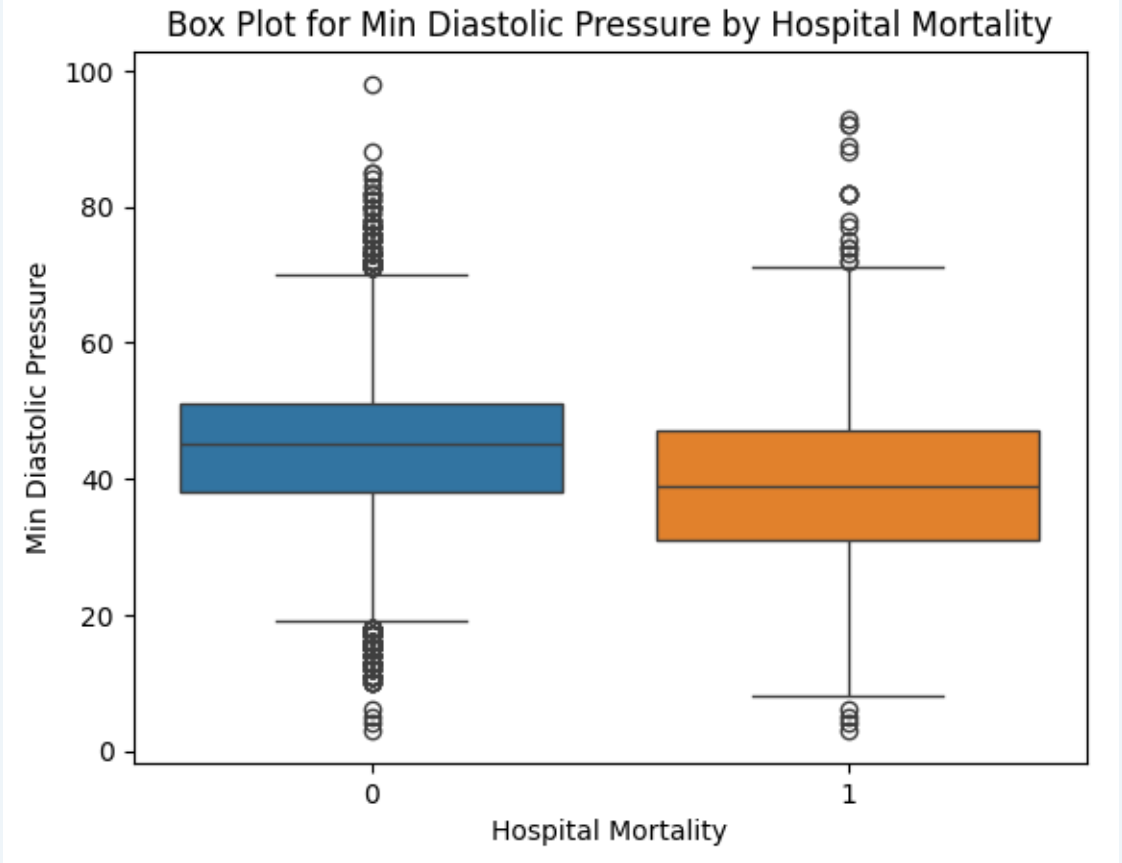

Box Plot for Mean MAP by Hospital Mortality
Mean MAP
120
100
80
60
40
0
1
Hospital Mortality

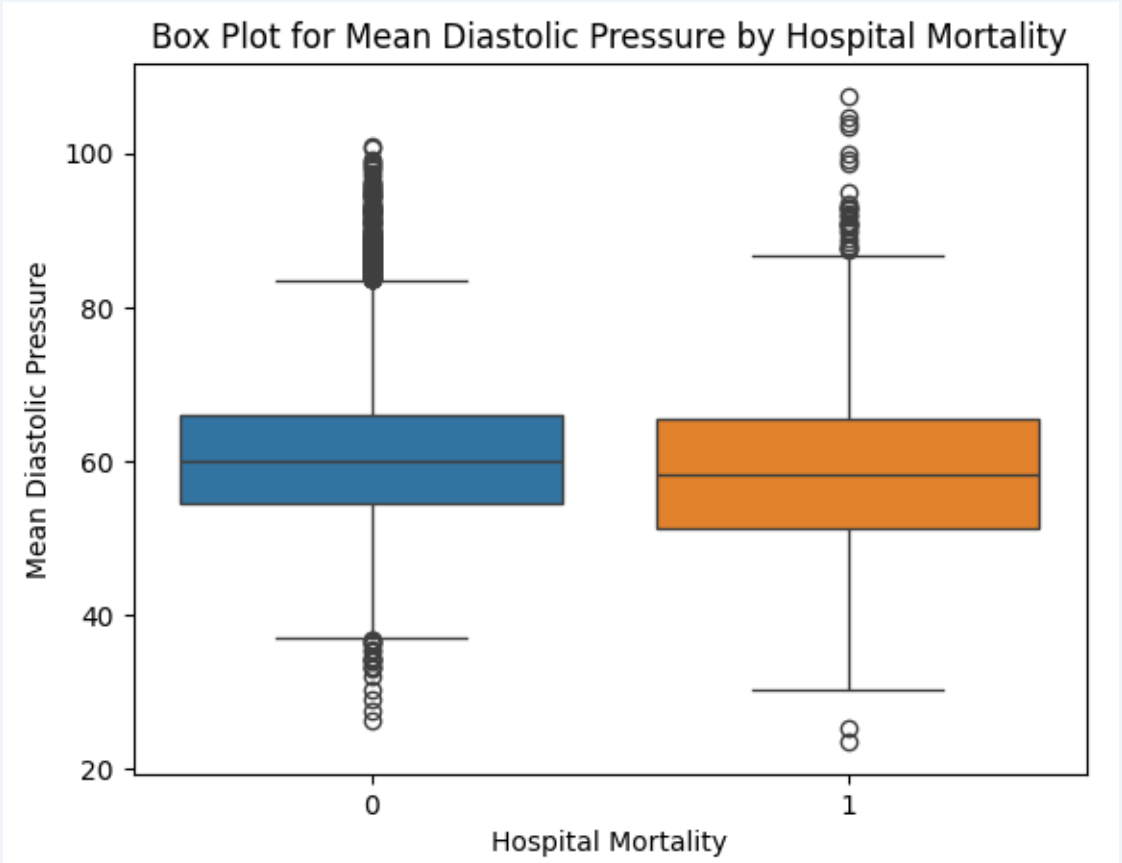

Box Plot for Max Systolic Pressure by Hospital Mortality
Max Systolic Pressure
300
250
200
150
100
0
1
Hospital Mortality

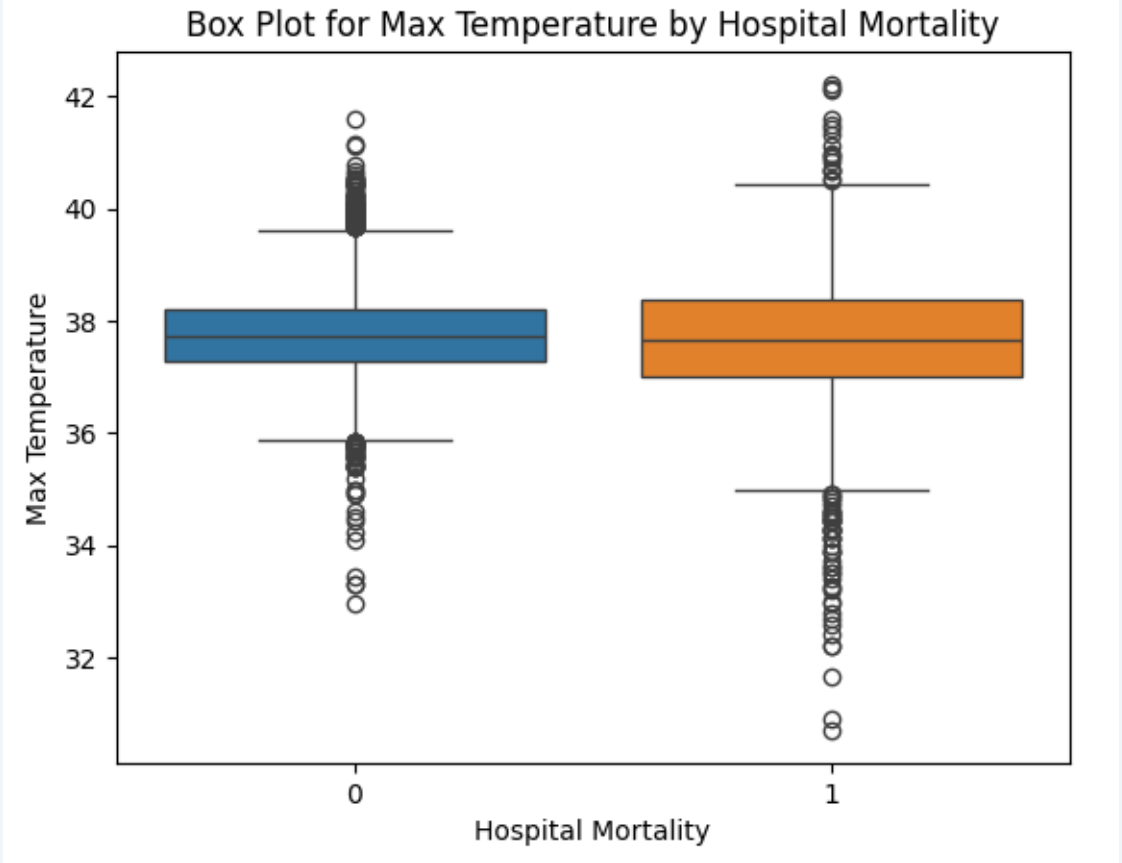

Box Plot for Min Systolic Pressure by Hospital Mortality
Min Systolic Pressure
175
150
125
100
75
50
25
0
0
1
Hospital Mortality

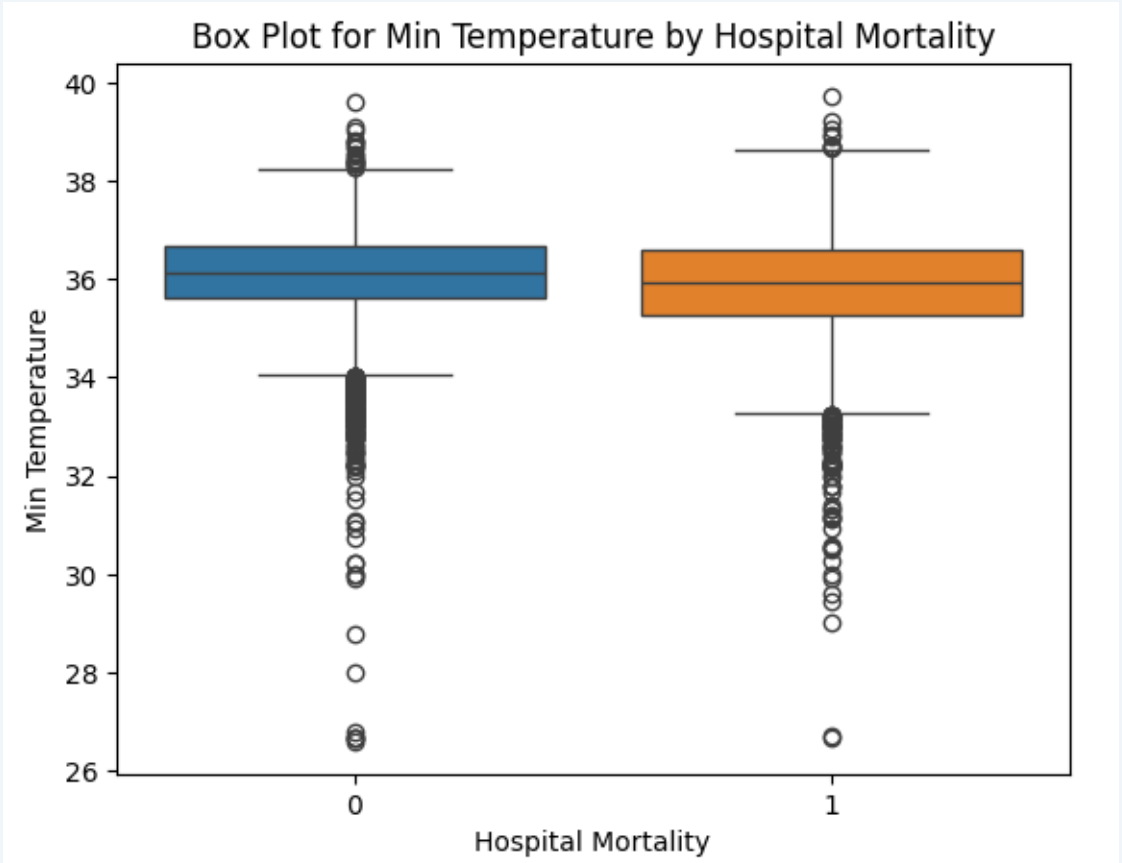

Box Plot for Mean Systolic Pressure by Hospital Mortality
Mean Systolic Pressure
200
180
160
140
120
100
80
60
0
1
Hospital Mortality

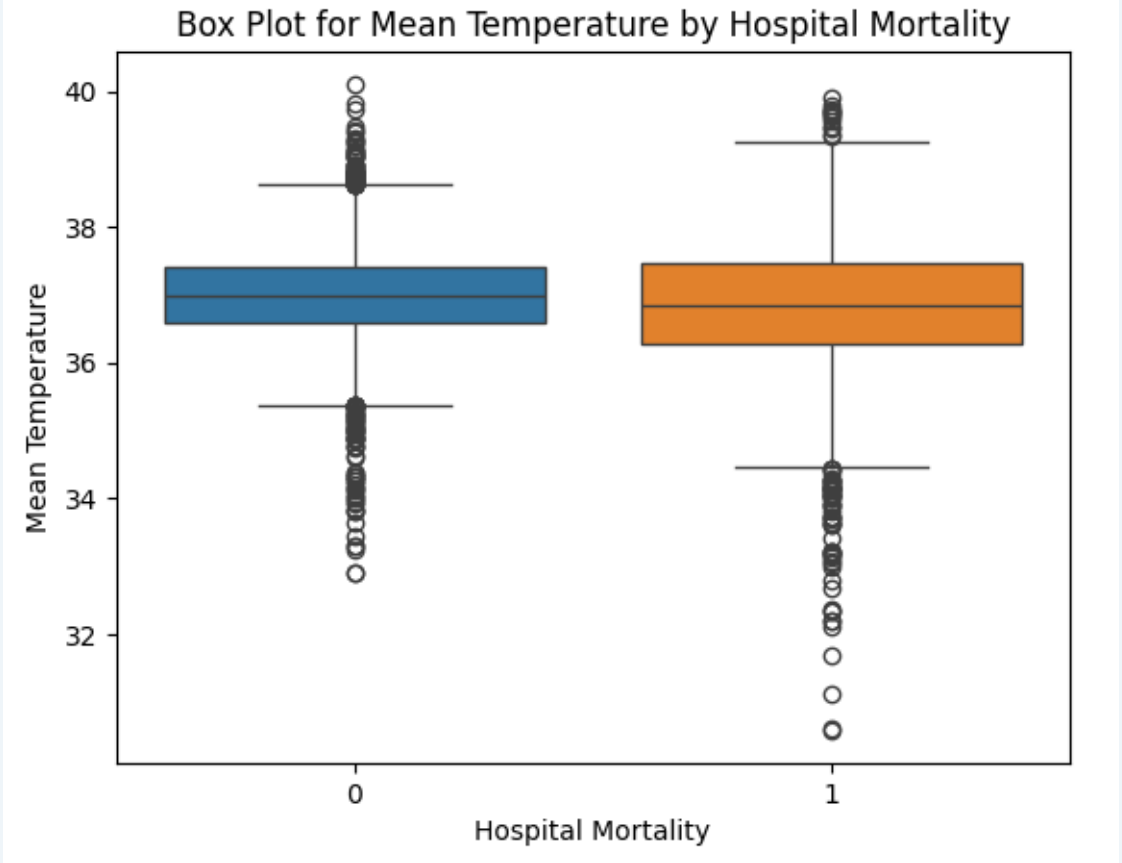

Box Plot for Max Diastolic Pressure by Hospital Mortality
Max Diastolic Pressure
300
250
200
150
100
50
0
1
Hospital Mortality

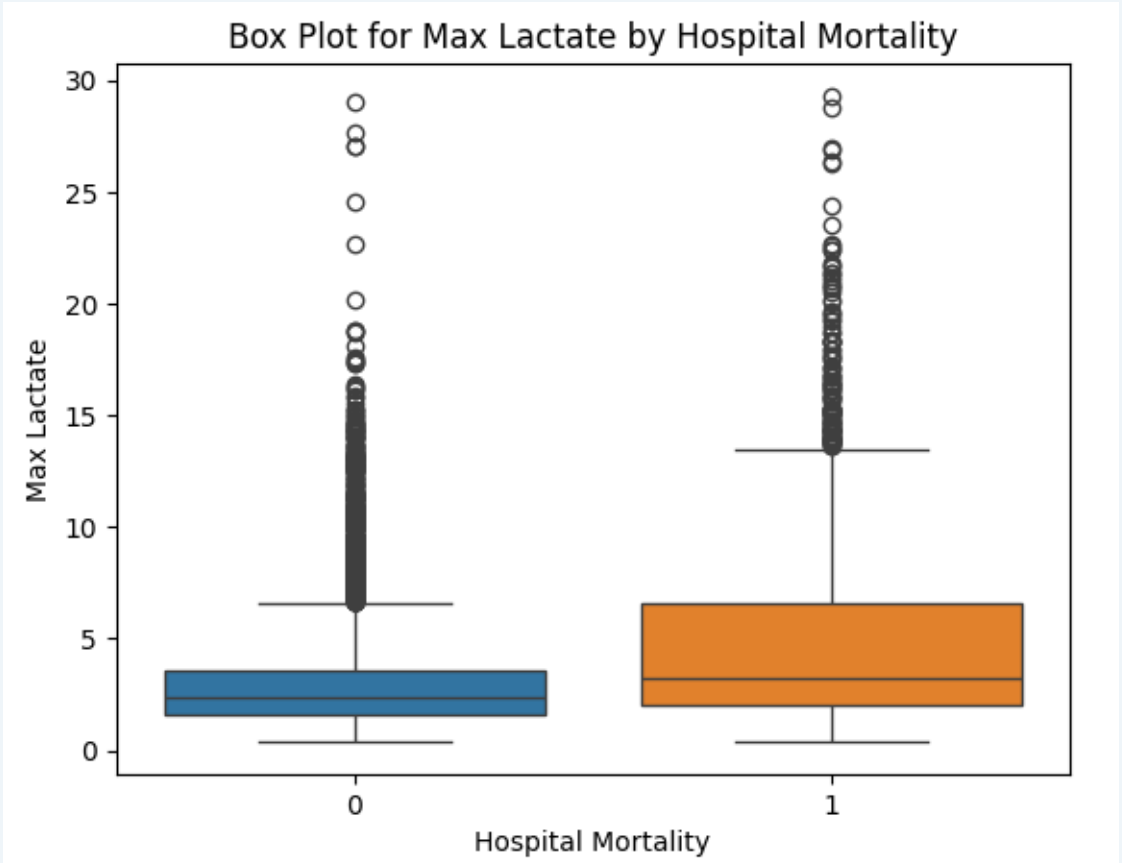

Box Plot for Min Diastolic Pressure by Hospital Mortality
Min Diastolic Pressure
100
80
60
40
20
0
0
1
Hospital Mortality

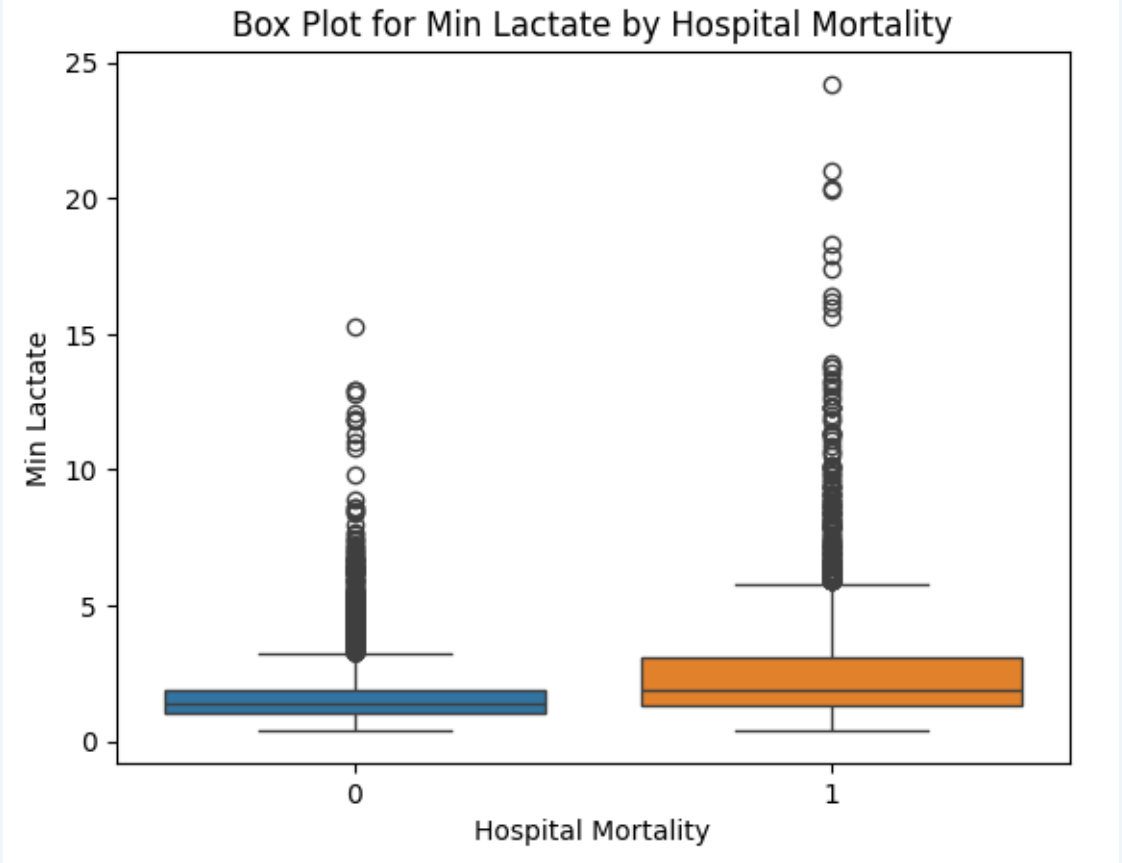

Box Plot for Mean Diastolic Pressure by Hospital Mortality
Mean Diastolic Pressure
100
80
60
40
20
0
1
Hospital Mortality

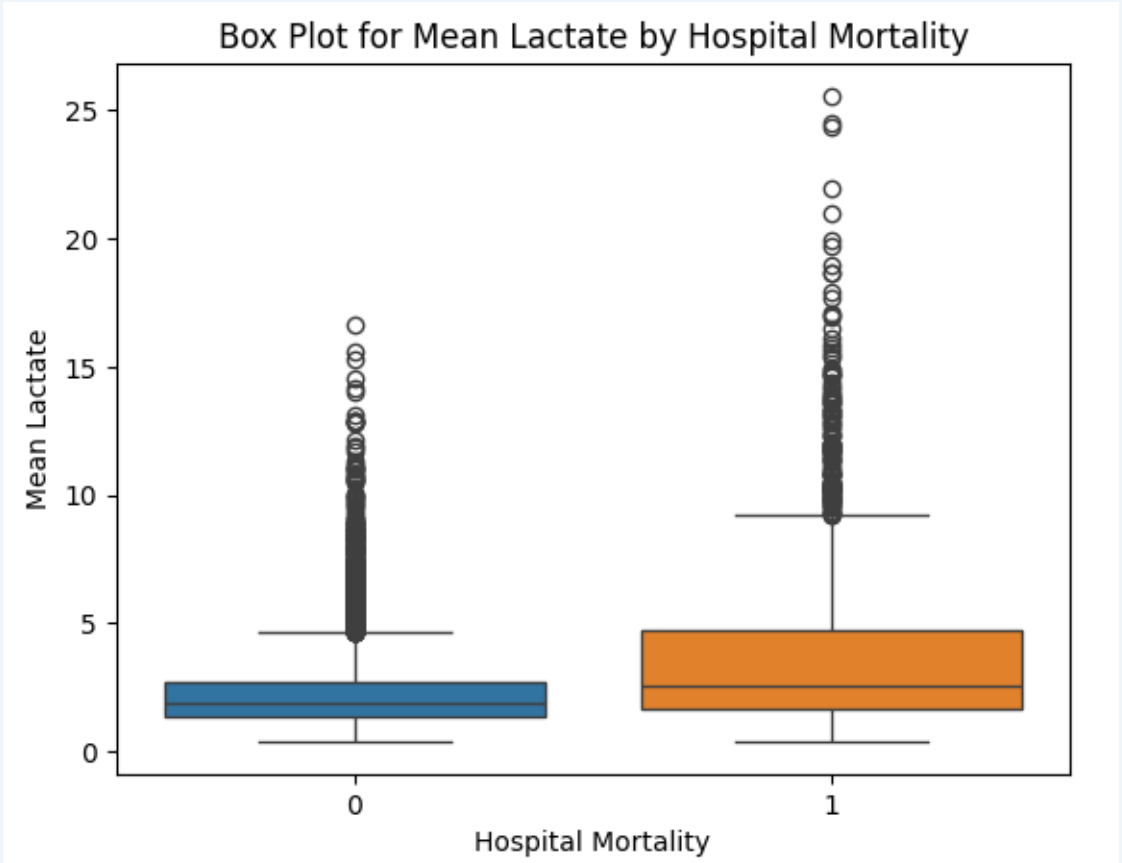

Box Plot for Max Temperature by Hospital Mortality
Max Temperature
42
40
38
36
34
32
0
1
Hospital Mortality

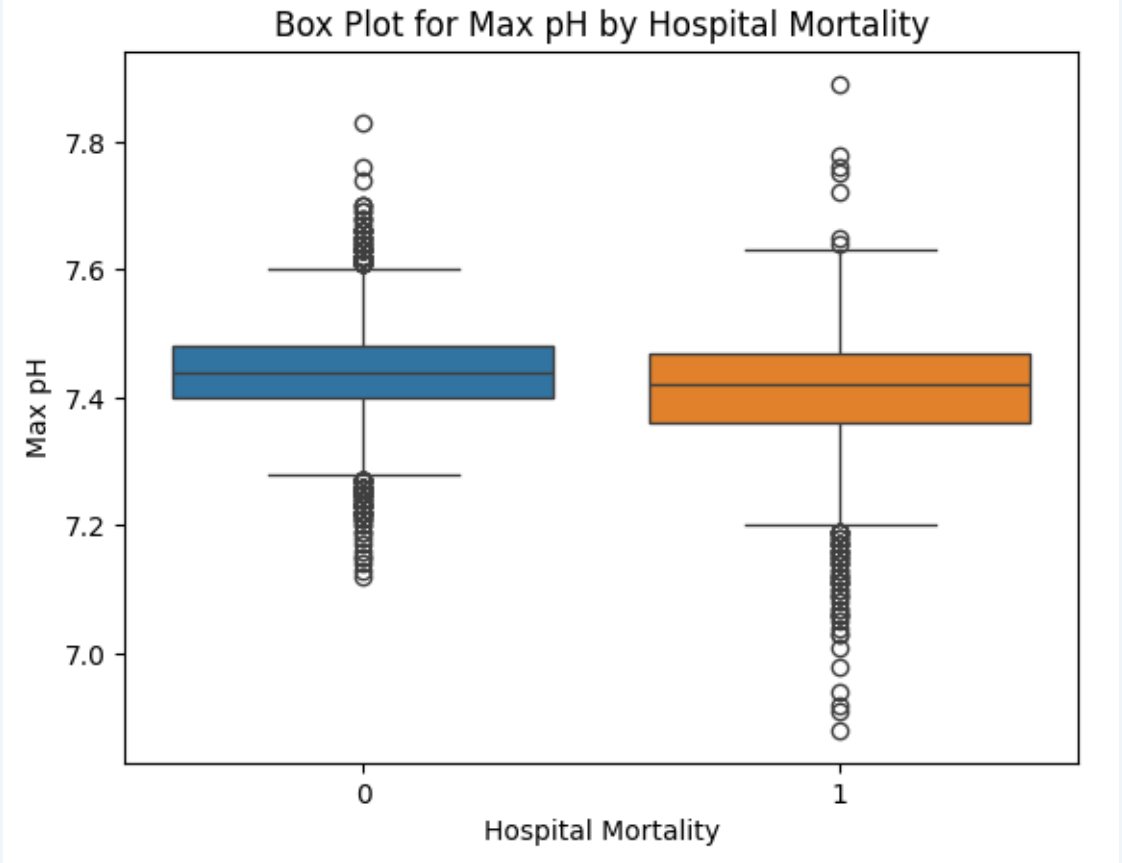

Box Plot for Min Temperature by Hospital Mortality
Min Temperature
40
38
36
34
32
30
28
26
0
1
Hospital Mortality

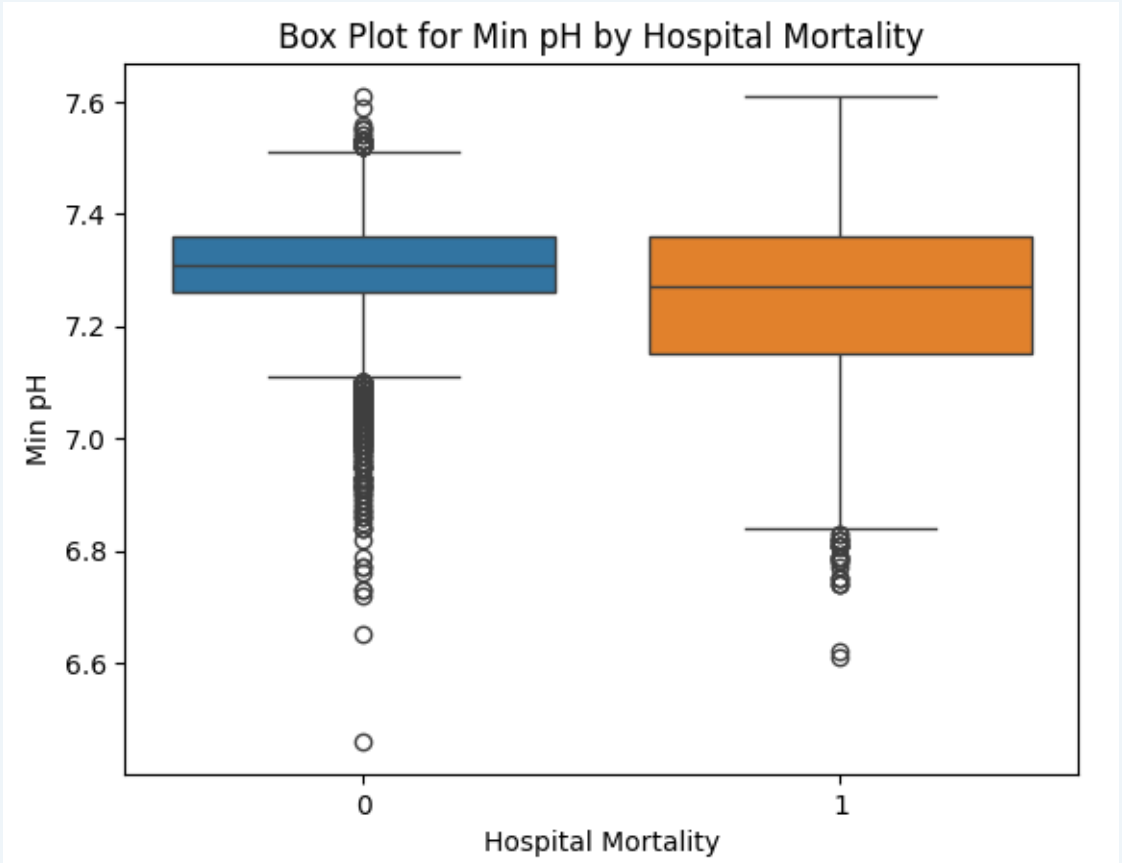

Box Plot for Mean Temperature by Hospital Mortality
Mean Temperature
40
38
36
34
32
0
1
Hospital Mortality

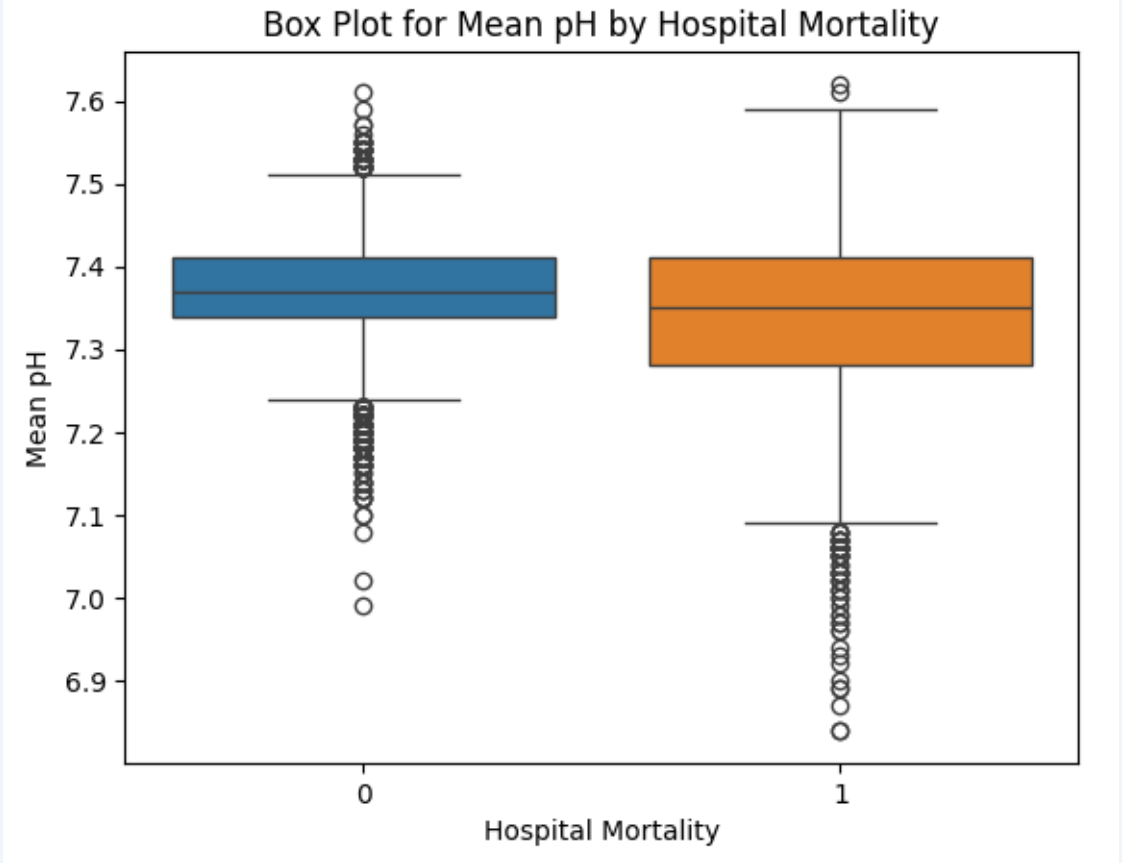

Box Plot for Max Lactate by Hospital Mortality
Max Lactate
30
25
20
15
10
5
0
0
1
Hospital Mortality

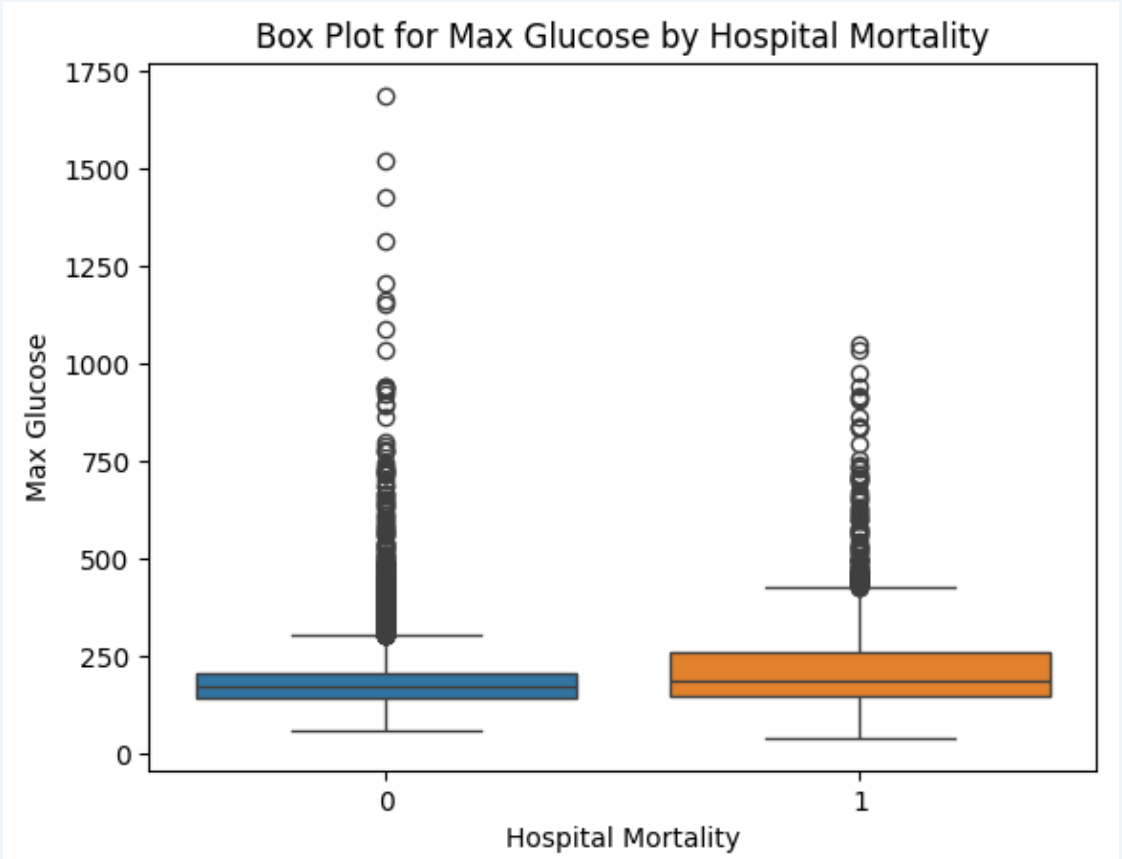

Box Plot for Min Lactate by Hospital Mortality
Min Lactate
25
20
15
10
5
0
0
1
Hospital Mortality

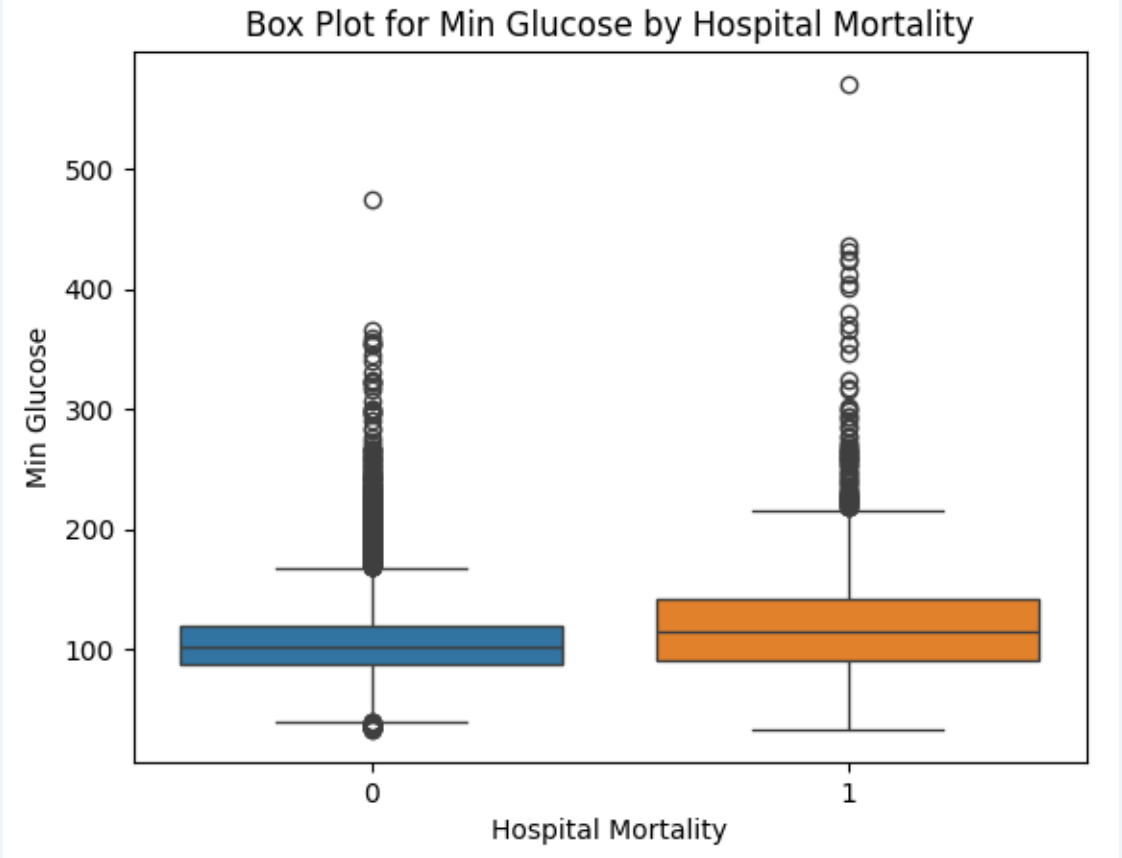

Box Plot for Mean Lactate by Hospital Mortality
Mean Lactate
25
20
15
10
5
0
0
1
Hospital Mortality

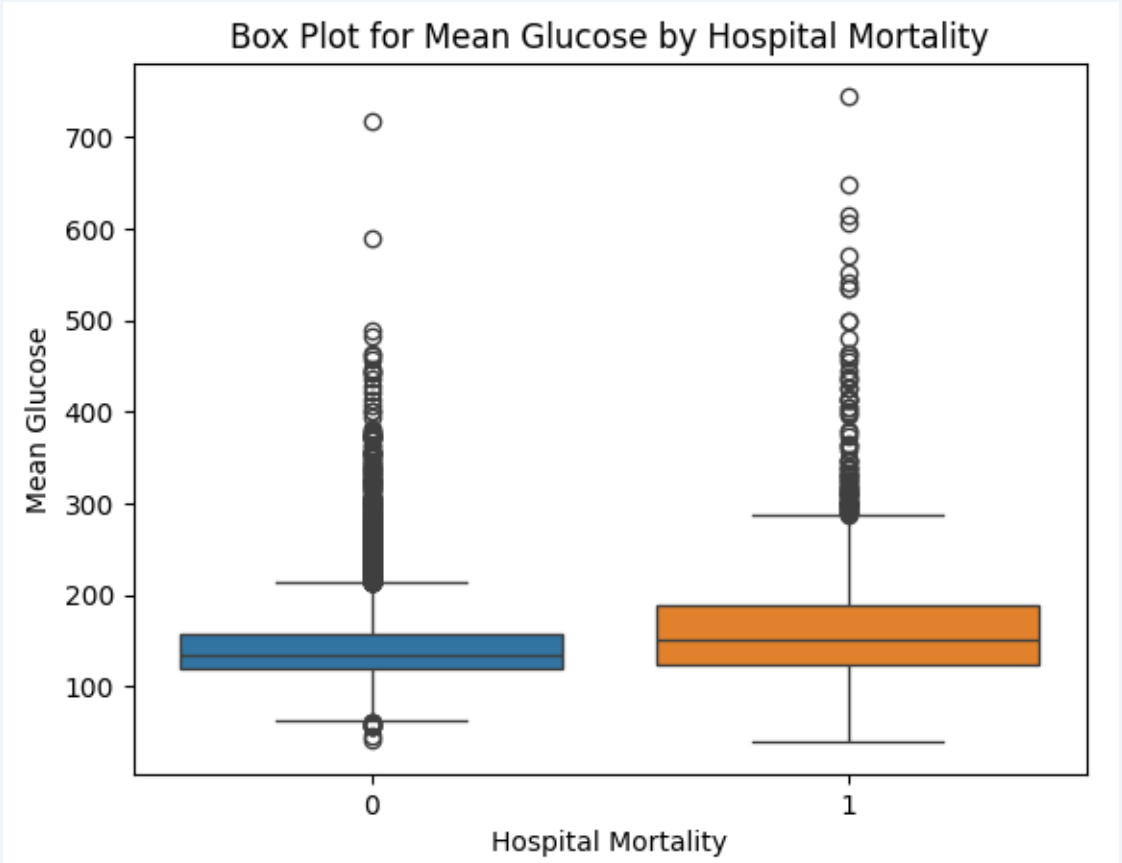

Box Plot for Max pH by Hospital Mortality
Max pH
7.8
7.6
7.4
7.2
7.0
0
1
Hospital Mortality

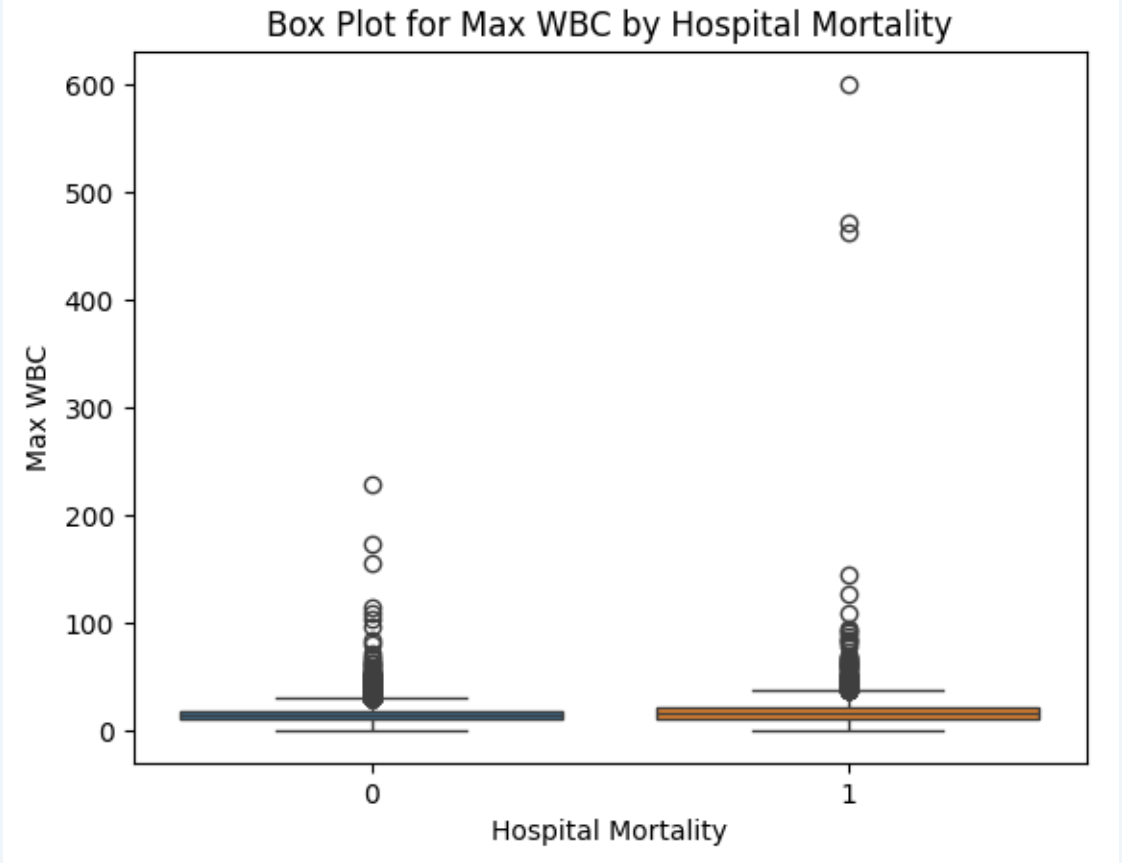

Box Plot for Min pH by Hospital Mortality
Min pH
7.6
7.4
7.2
7.0
6.8
6.6
0
1
Hospital Mortality

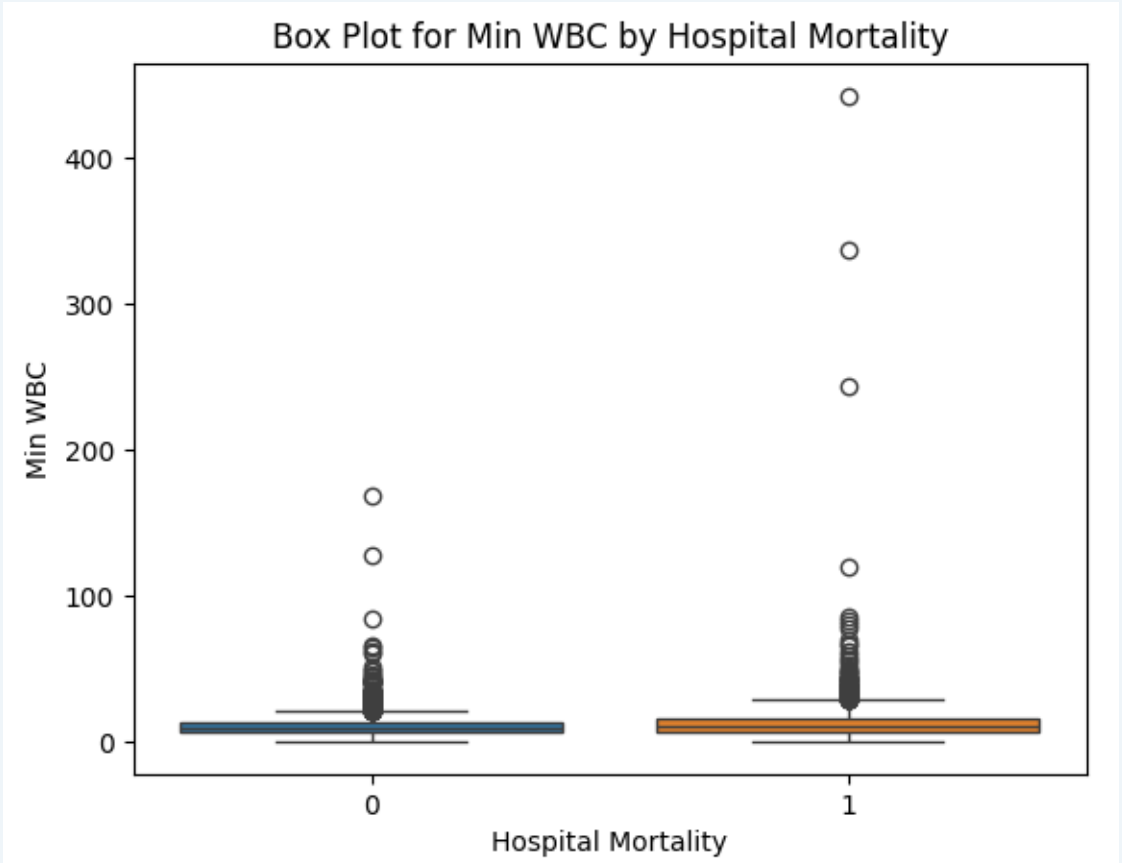

Box Plot for Mean pH by Hospital Mortality
Mean pH
7.6
7.5
7.4
7.3
7.2
7.1
7.0
6.9
0
1
Hospital Mortality

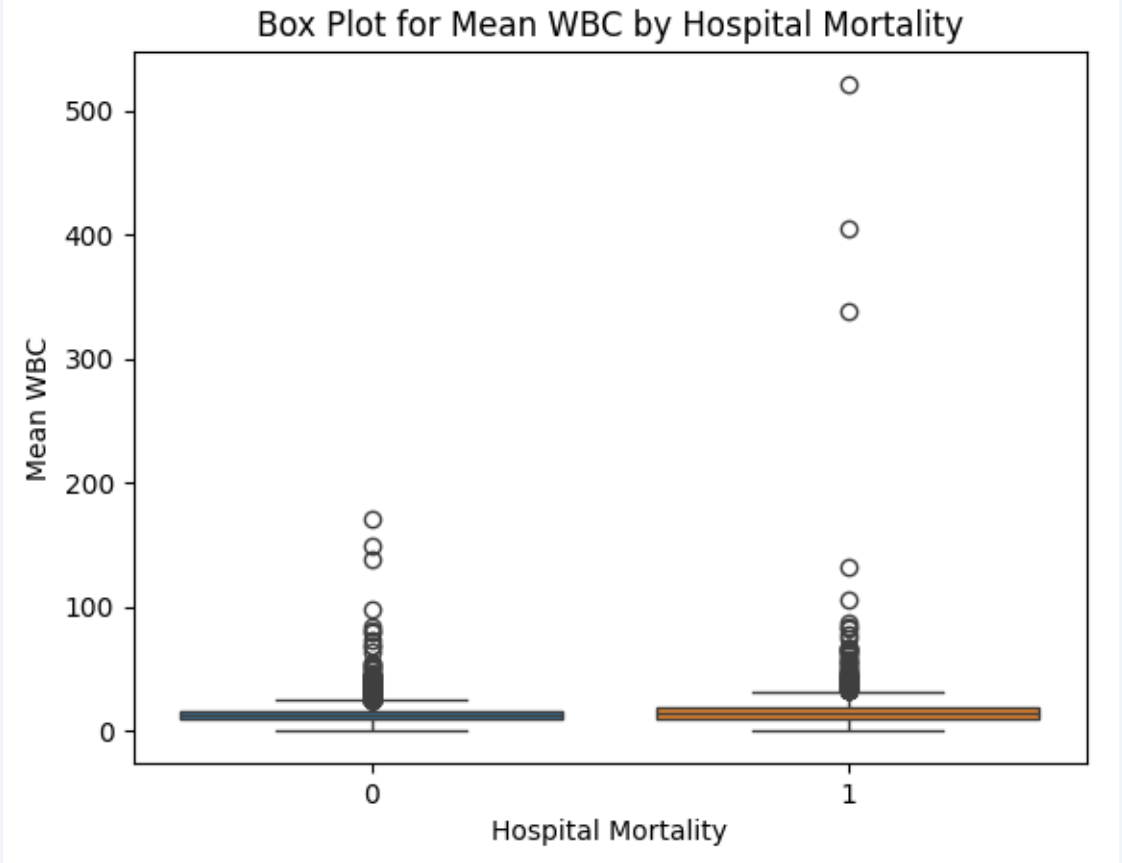

Box Plot for Max Glucose by Hospital Mortality
Max Glucose
1750
1500
1250
1000
750
500
250
0
0
1
Hospital Mortality

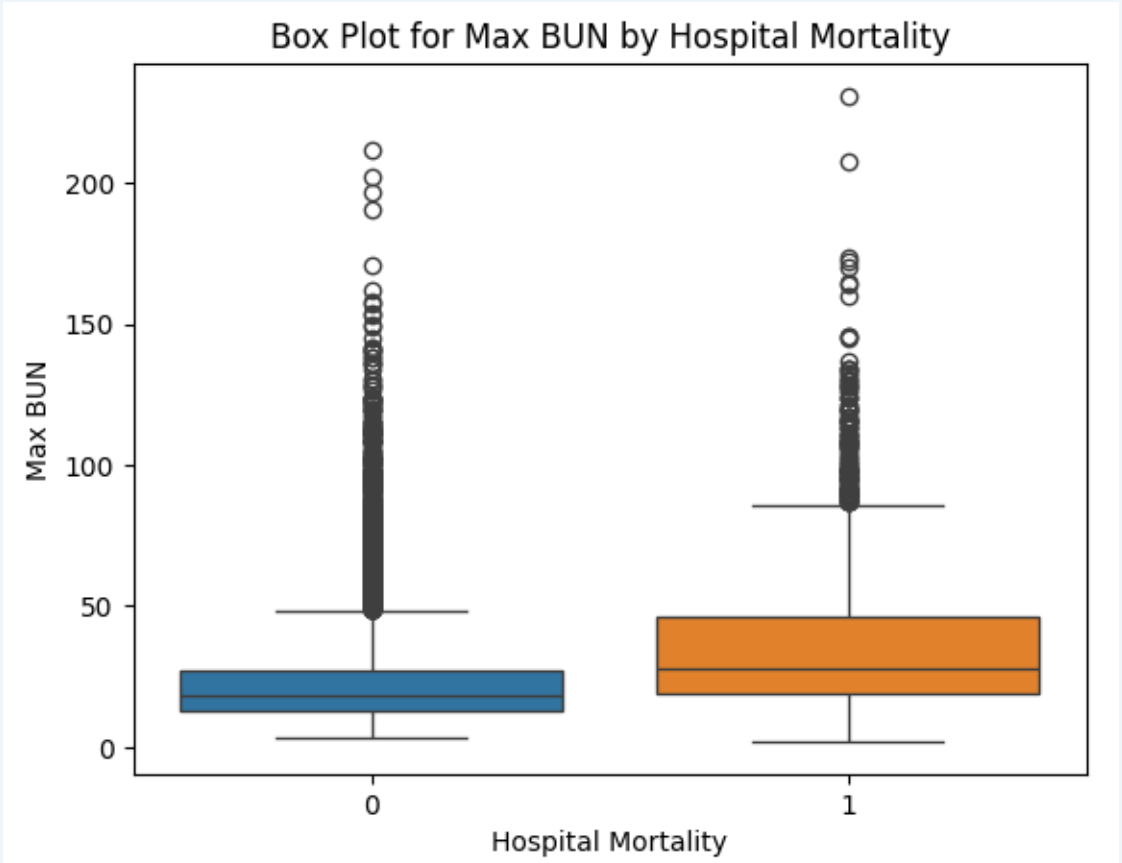

Box Plot for Min Glucose by Hospital Mortality
Min Glucose
500
400
300
200
100
0
1
Hospital Mortality

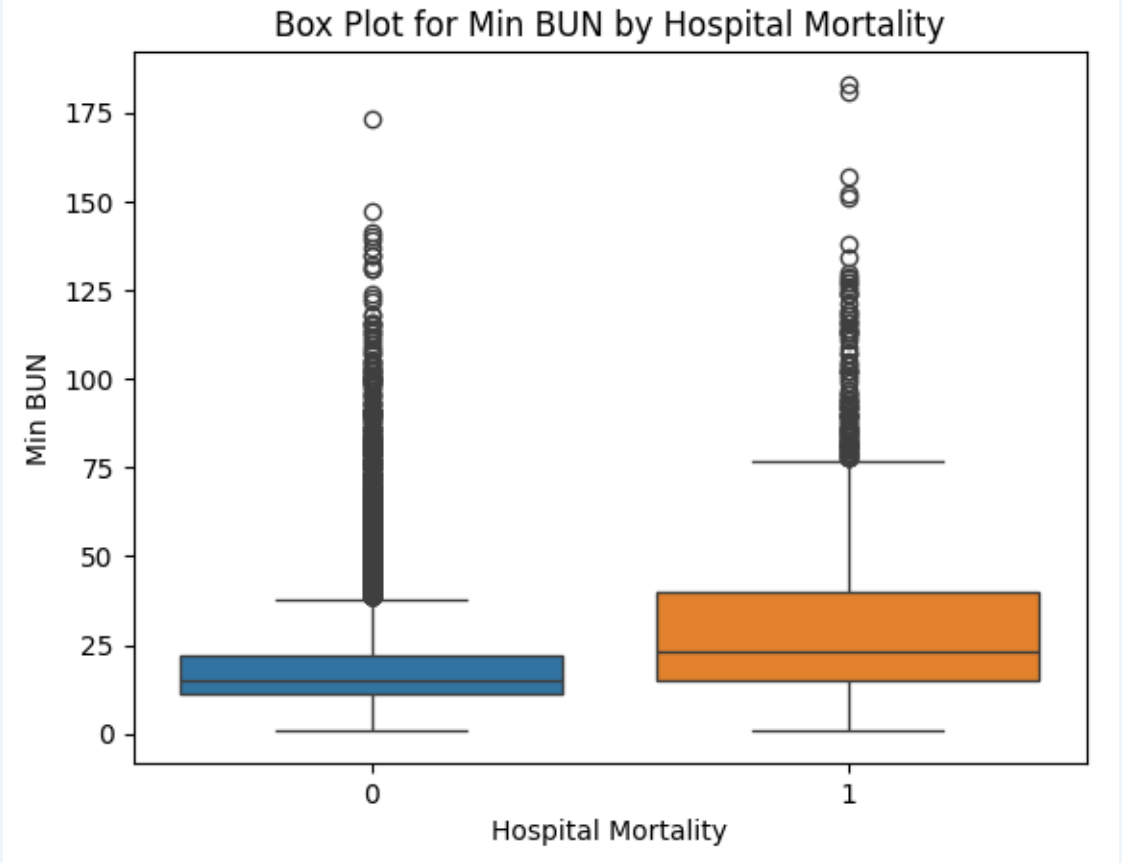

Box Plot for Mean Glucose by Hospital Mortality
Mean Glucose
700
600
500
400
300
200
100
0
1
Hospital Mortality

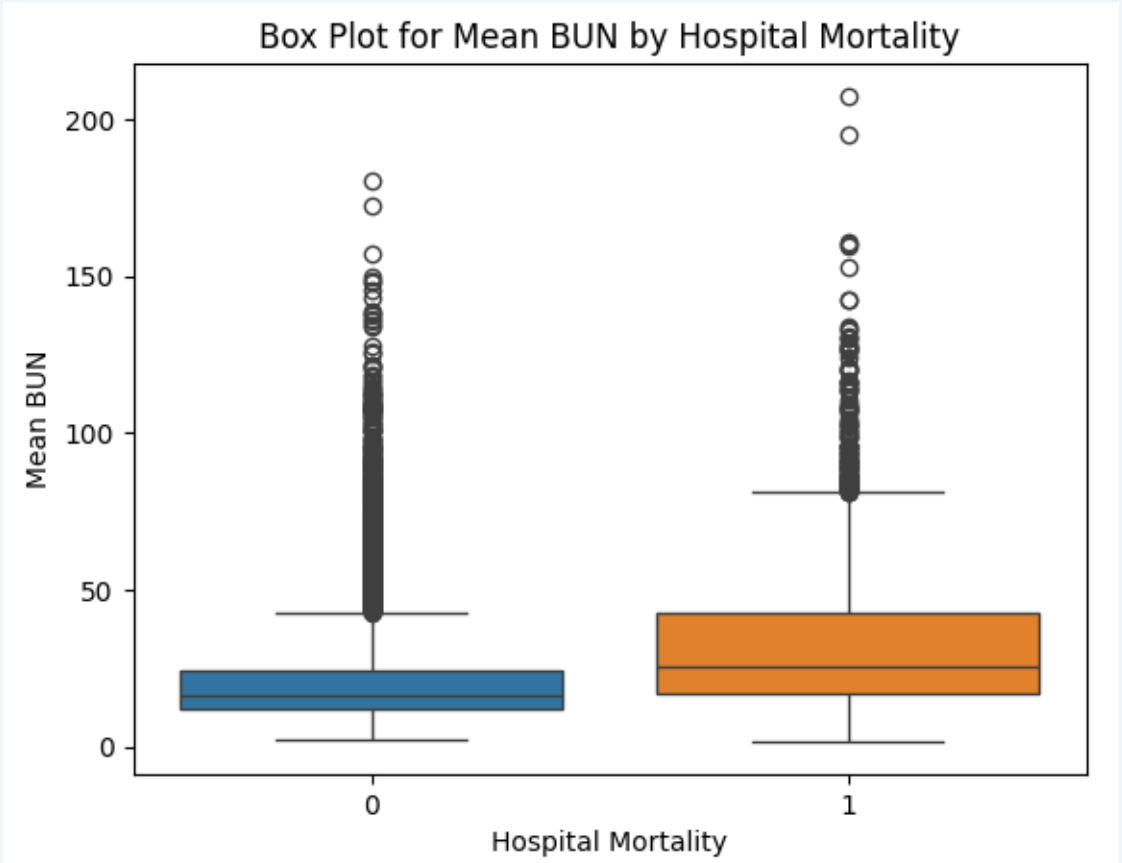

Box Plot for Max WBC by Hospital Mortality
Max WBC
600
500
400
300
200
100
0
0
1
Hospital Mortality

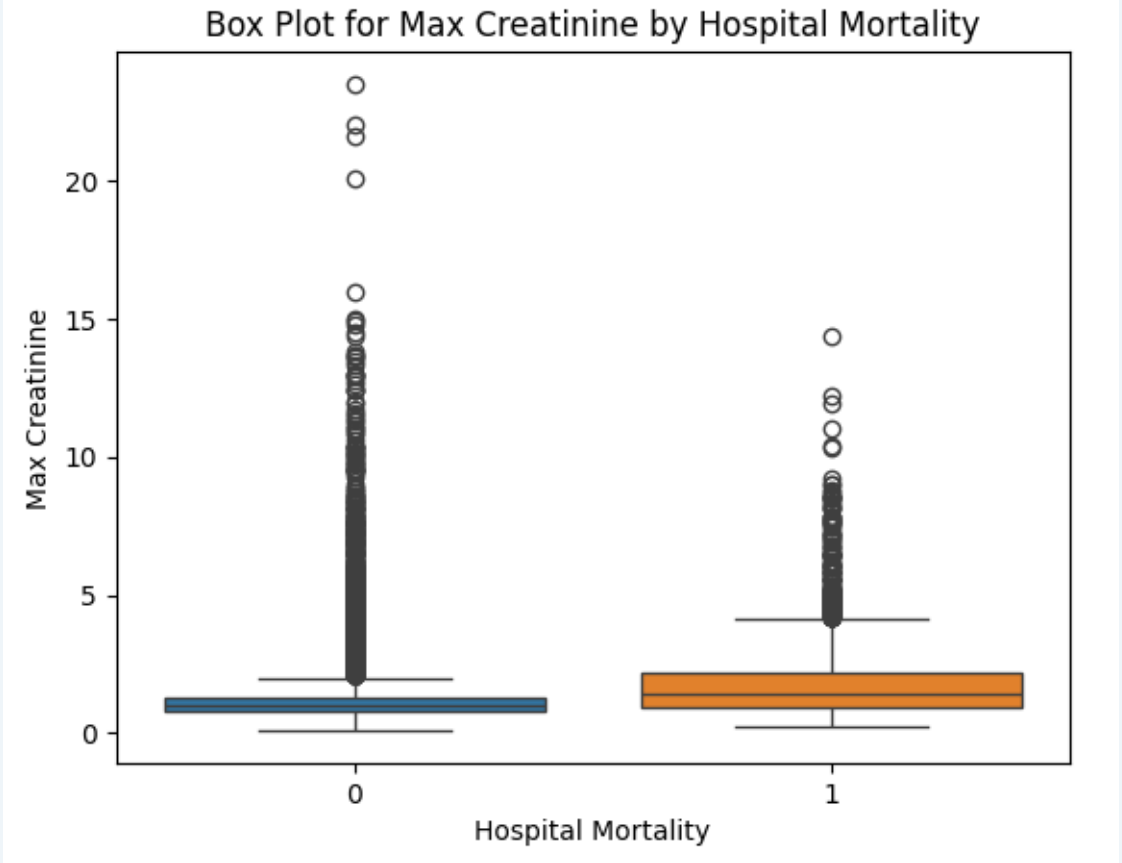

Box Plot for Min WBC by Hospital Mortality
Min WBC
400
300
200
100
0
0
1
Hospital Mortality

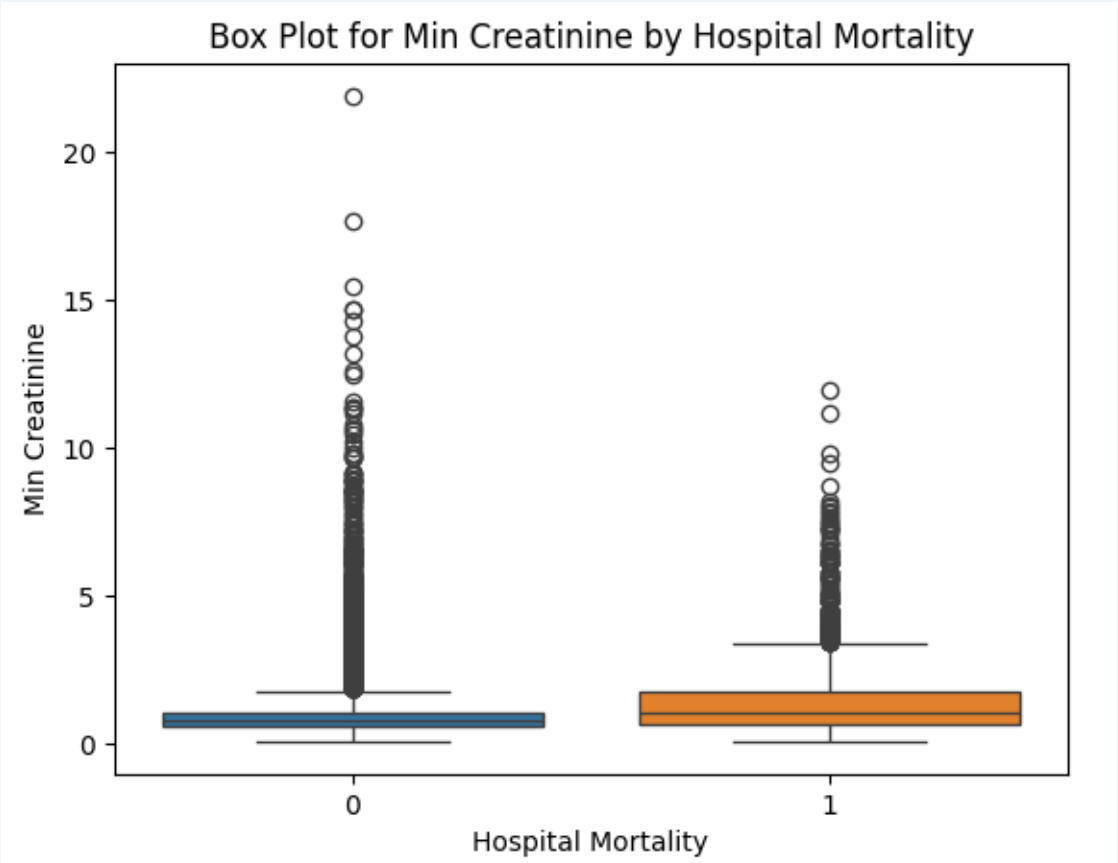

Box Plot for Mean WBC by Hospital Mortality
Mean WBC
500
400
300
200
100
0
0
1
Hospital Mortality

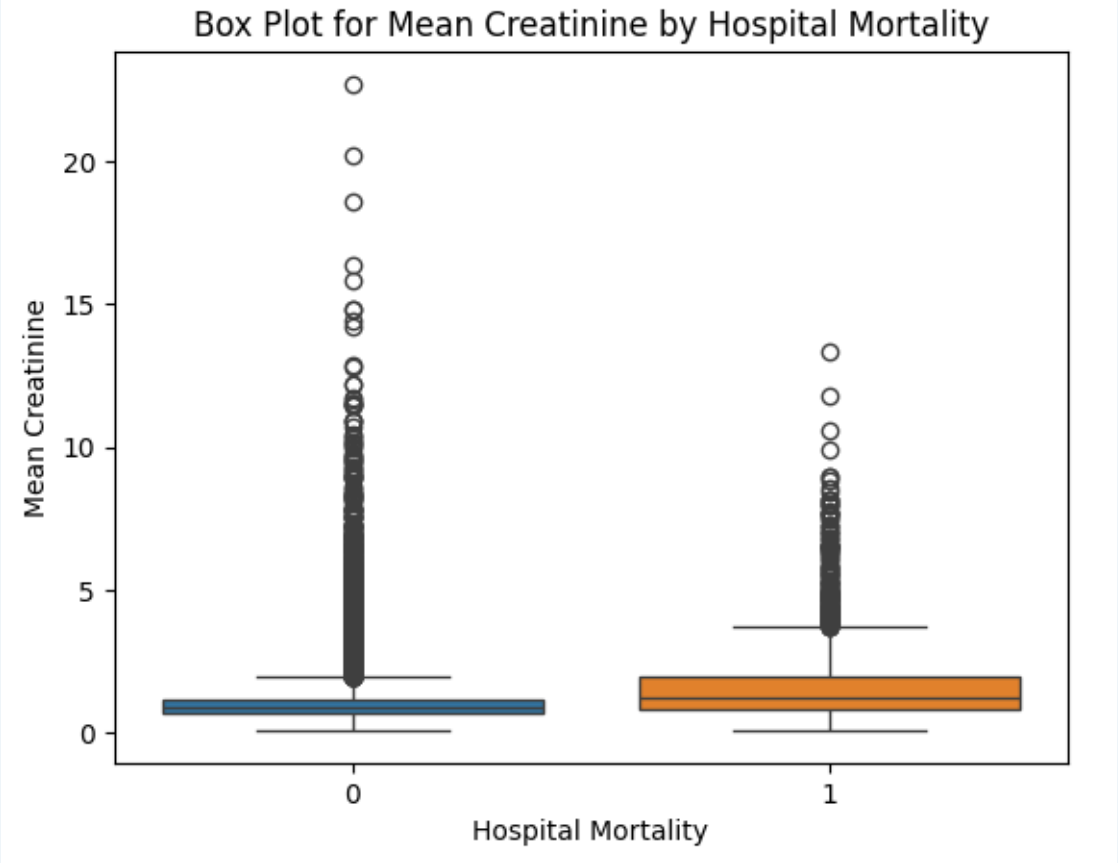

Box Plot for Max BUN by Hospital Mortality
Max BUN
200
150
100
50
0
0
1
Hospital Mortality

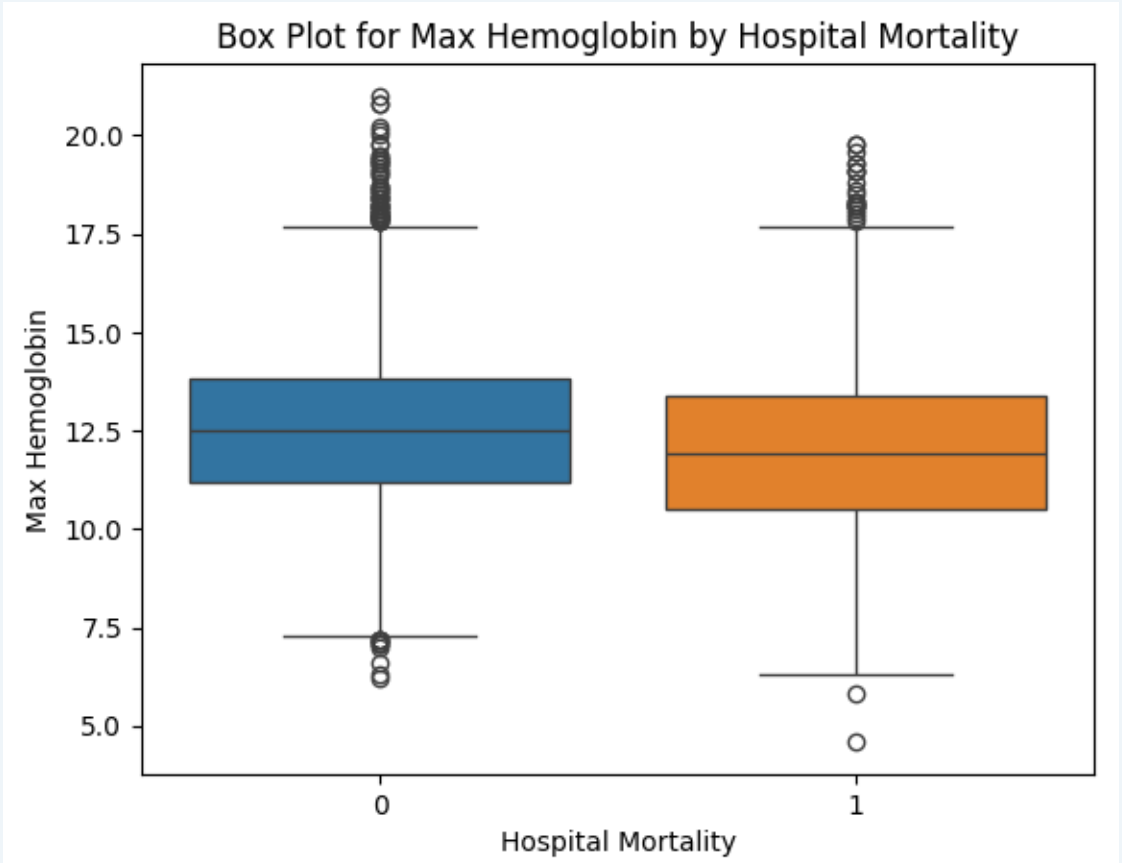

Box Plot for Min BUN by Hospital Mortality
Min BUN
175
150
125
100
75
50
25
0
0
1
Hospital Mortality

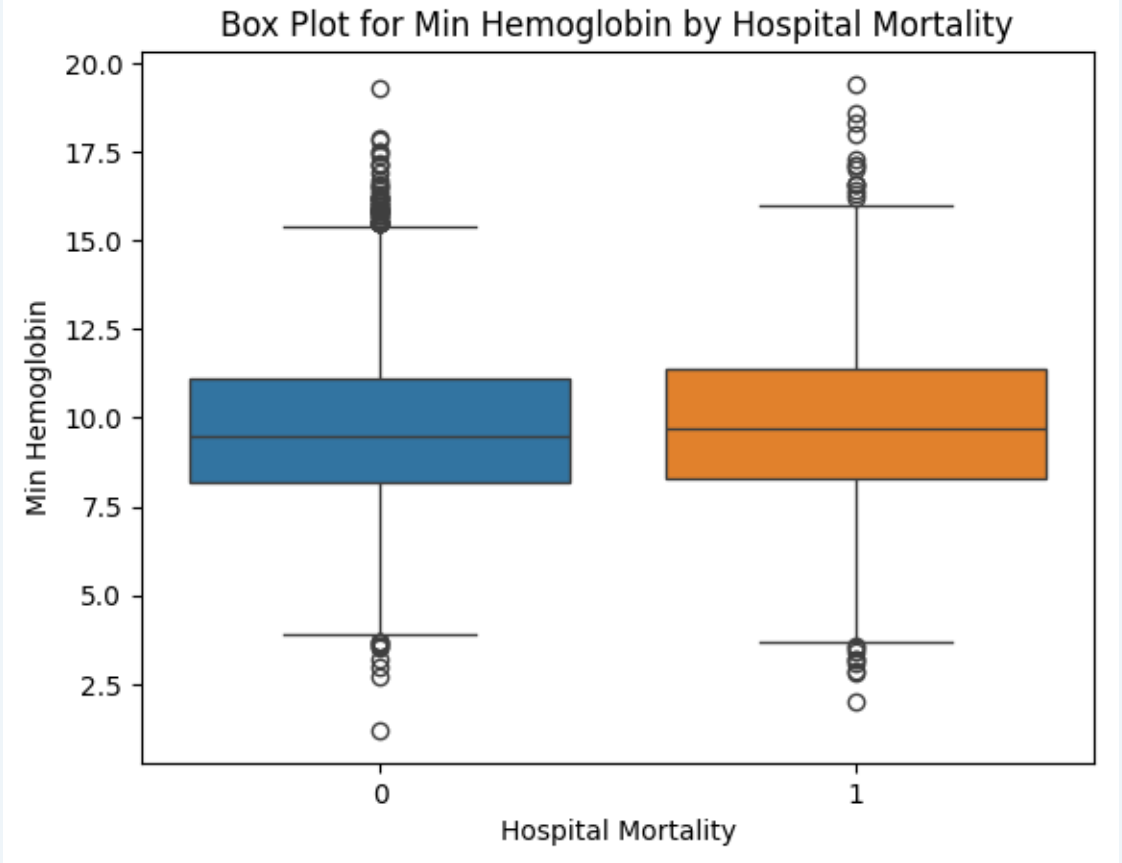

Box Plot for Mean BUN by Hospital Mortality
Mean BUN
200
150
100
50
0
0
1
Hospital Mortality

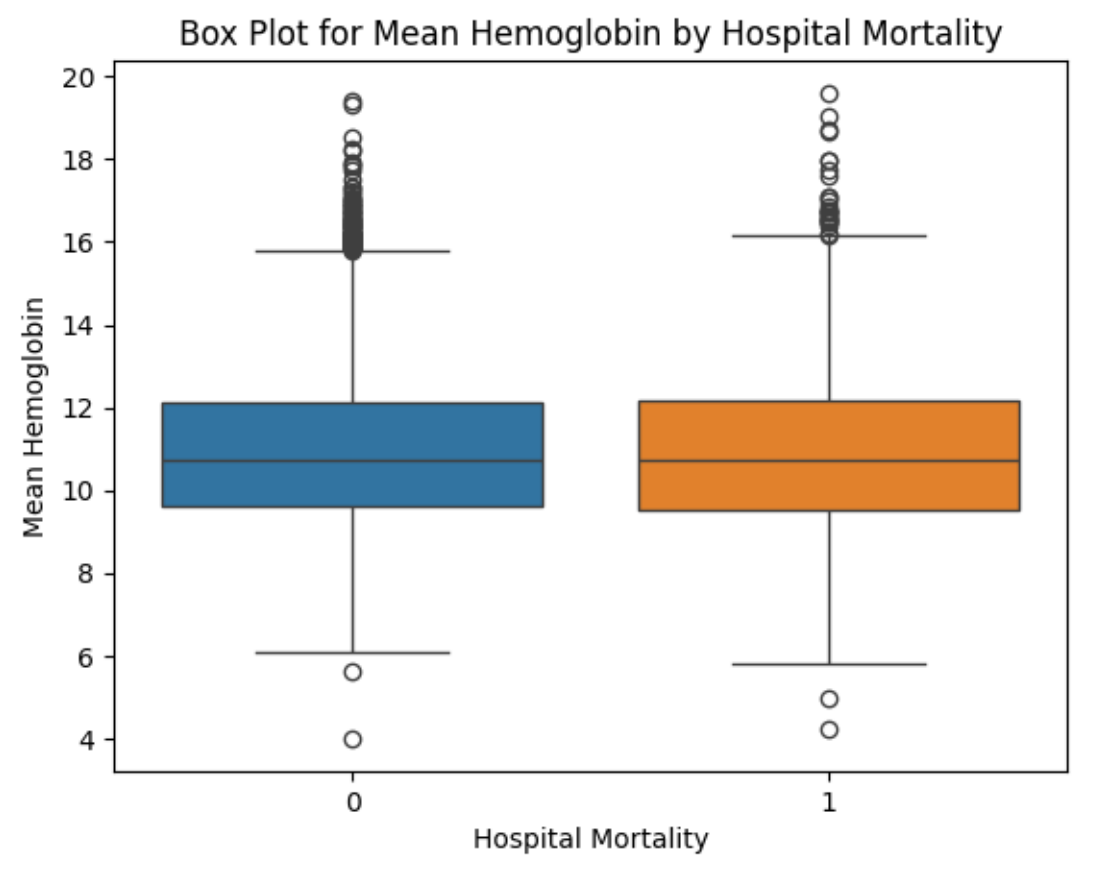

Box Plot for Max Creatinine by Hospital Mortality
Max Creatinine
20
15
10
5
0
0
1
Hospital Mortality


Box Plot for Min Creatinine by Hospital Mortality
Min Creatinine
0
5
10
15
20
0
1
Hospital Mortality


Box Plot for Mean Creatinine by Hospital Mortality
Mean Creatinine
0
5
10
15
20
0
1
Hospital Mortality

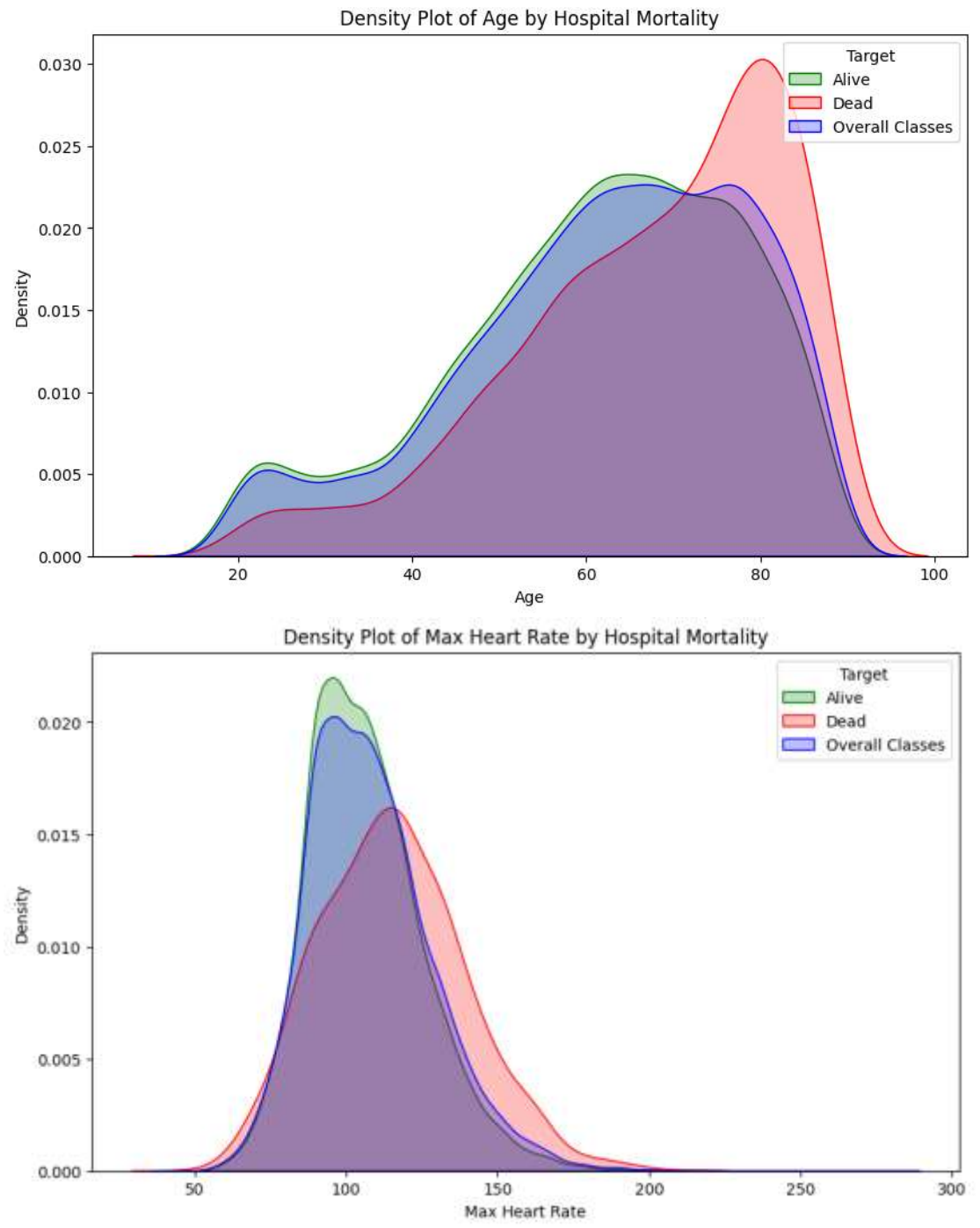

Box Plot for Max Hemoglobin by Hospital Mortality
Max Hemoglobin
20.0
17.5
15.0
12.5
10.0
7.5
5.0
0
1
Hospital Mortality

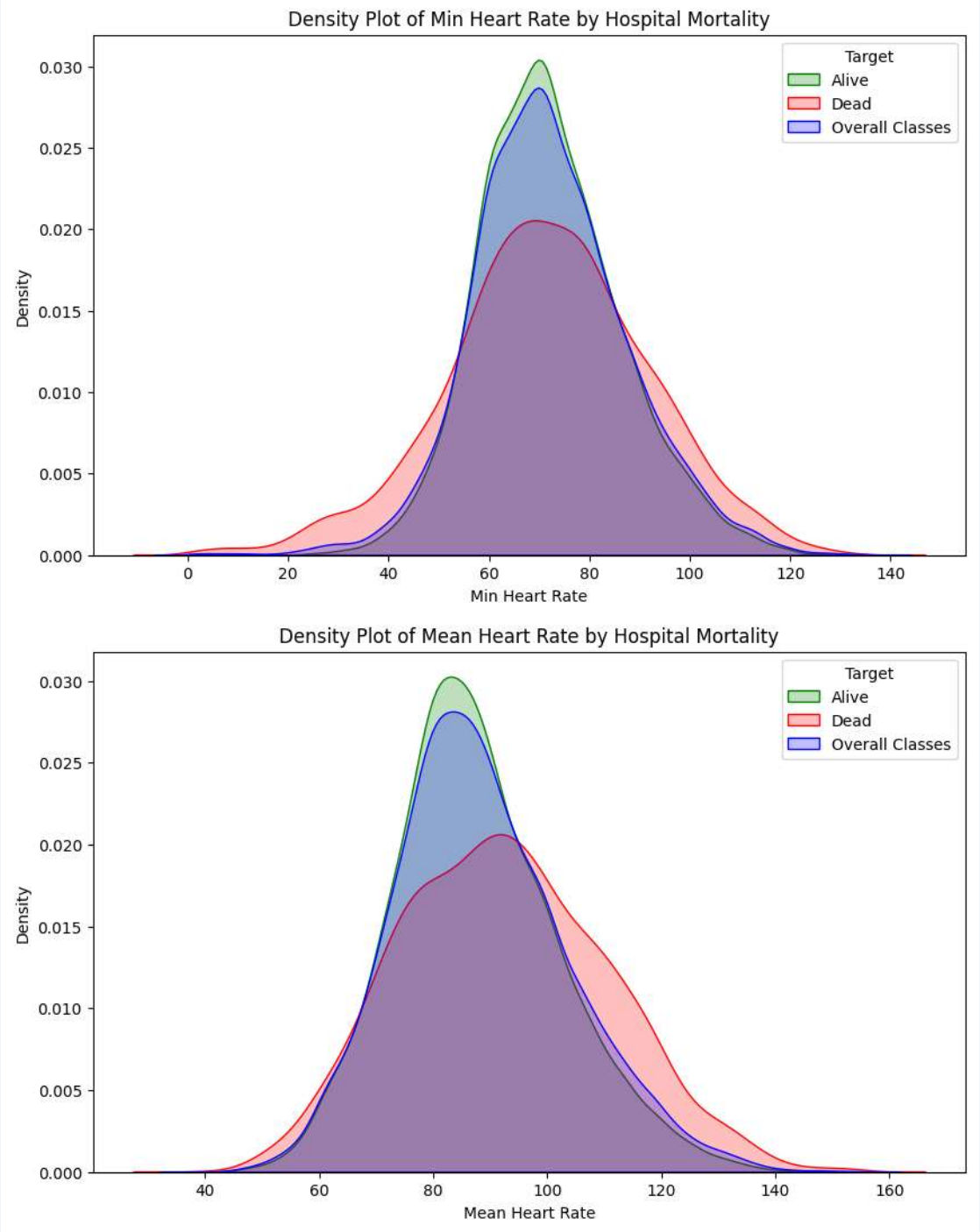

Box Plot for Min Hemoglobin by Hospital Mortality
Min Hemoglobin
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
1
Hospital Mortality

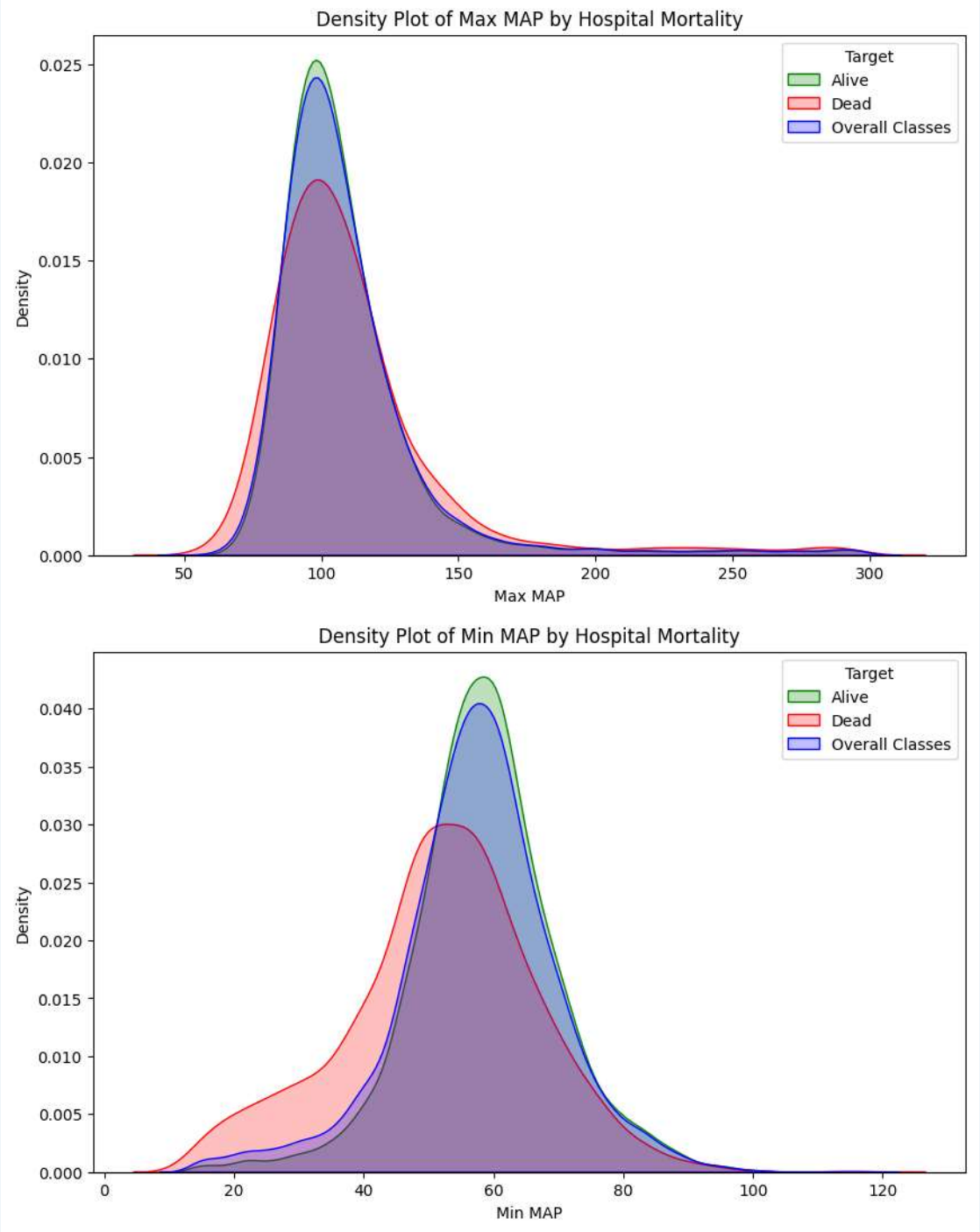

Box Plot for Mean Hemoglobin by Hospital Mortality
Mean Hemoglobin
20
18
16
14
12
10
8
6
4
0
1
Hospital Mortality

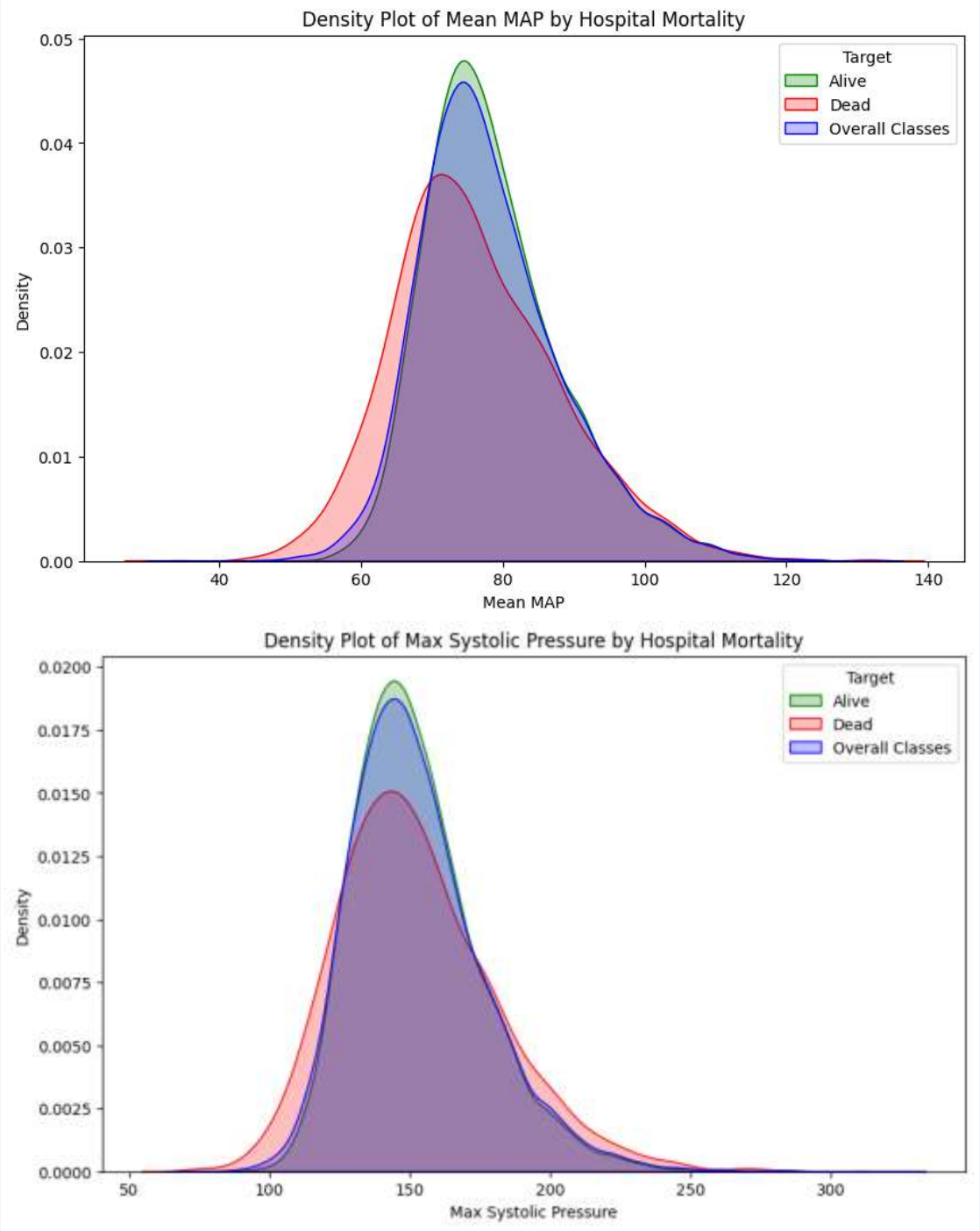
Visualization of Continuous Variable

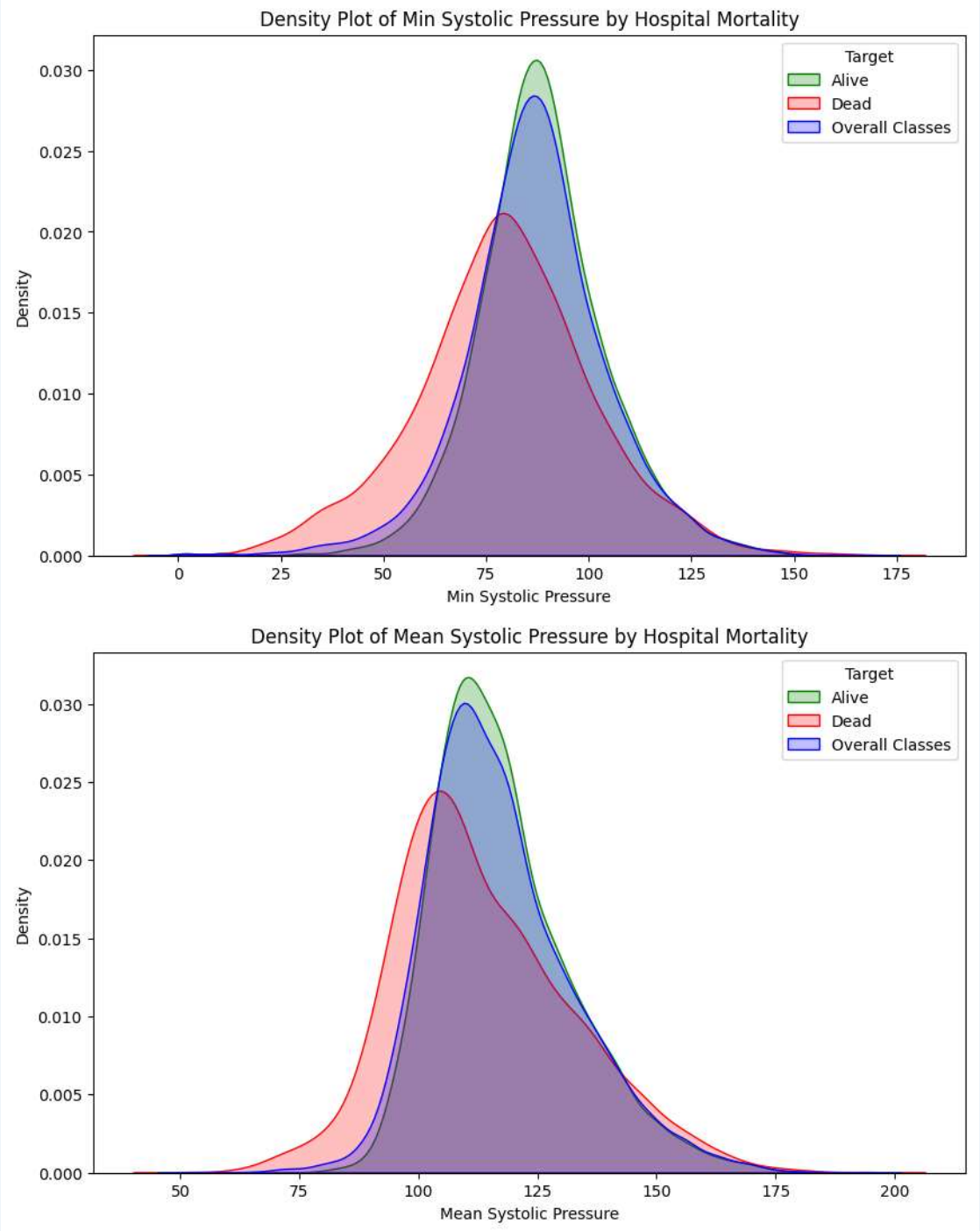
Density plots

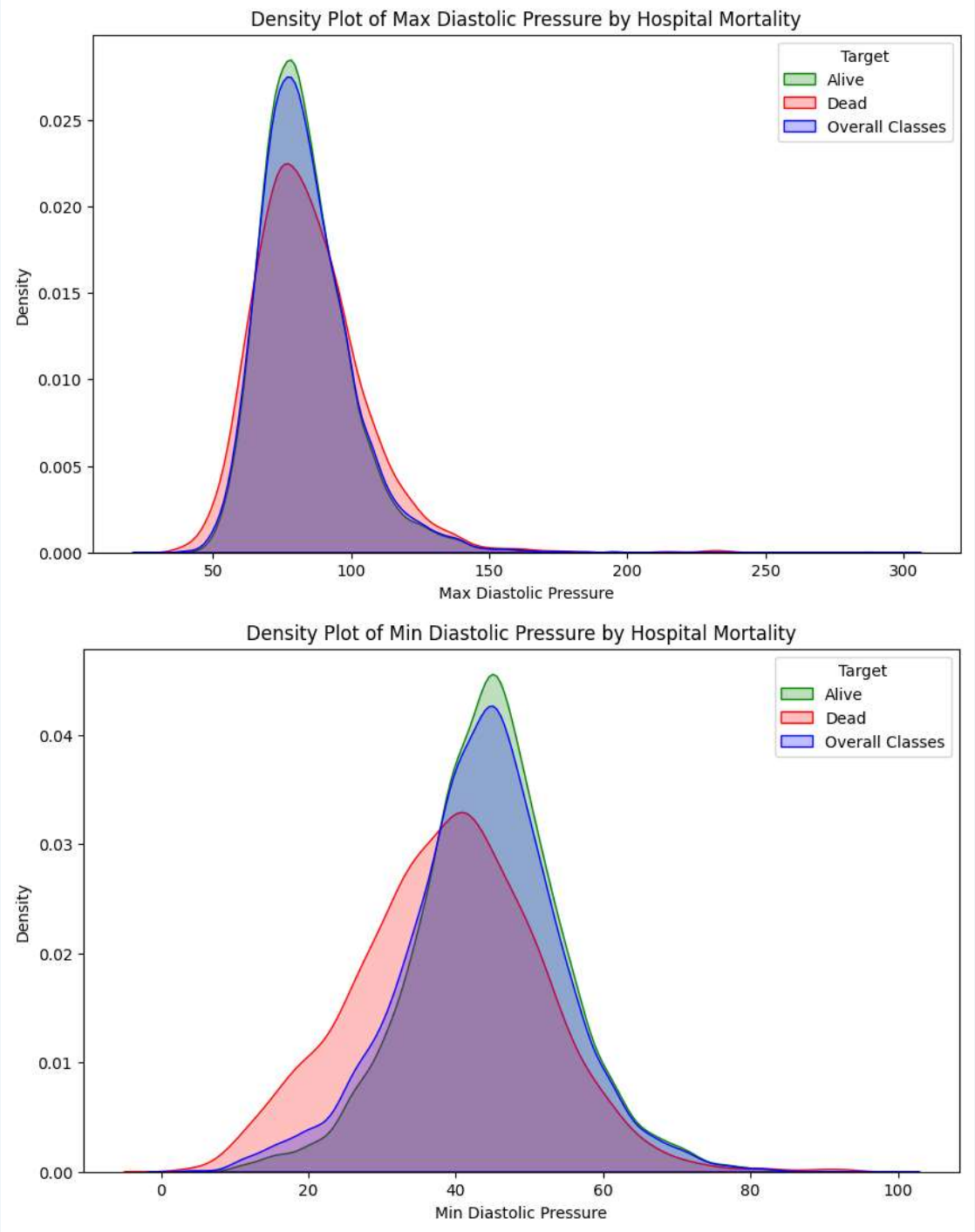


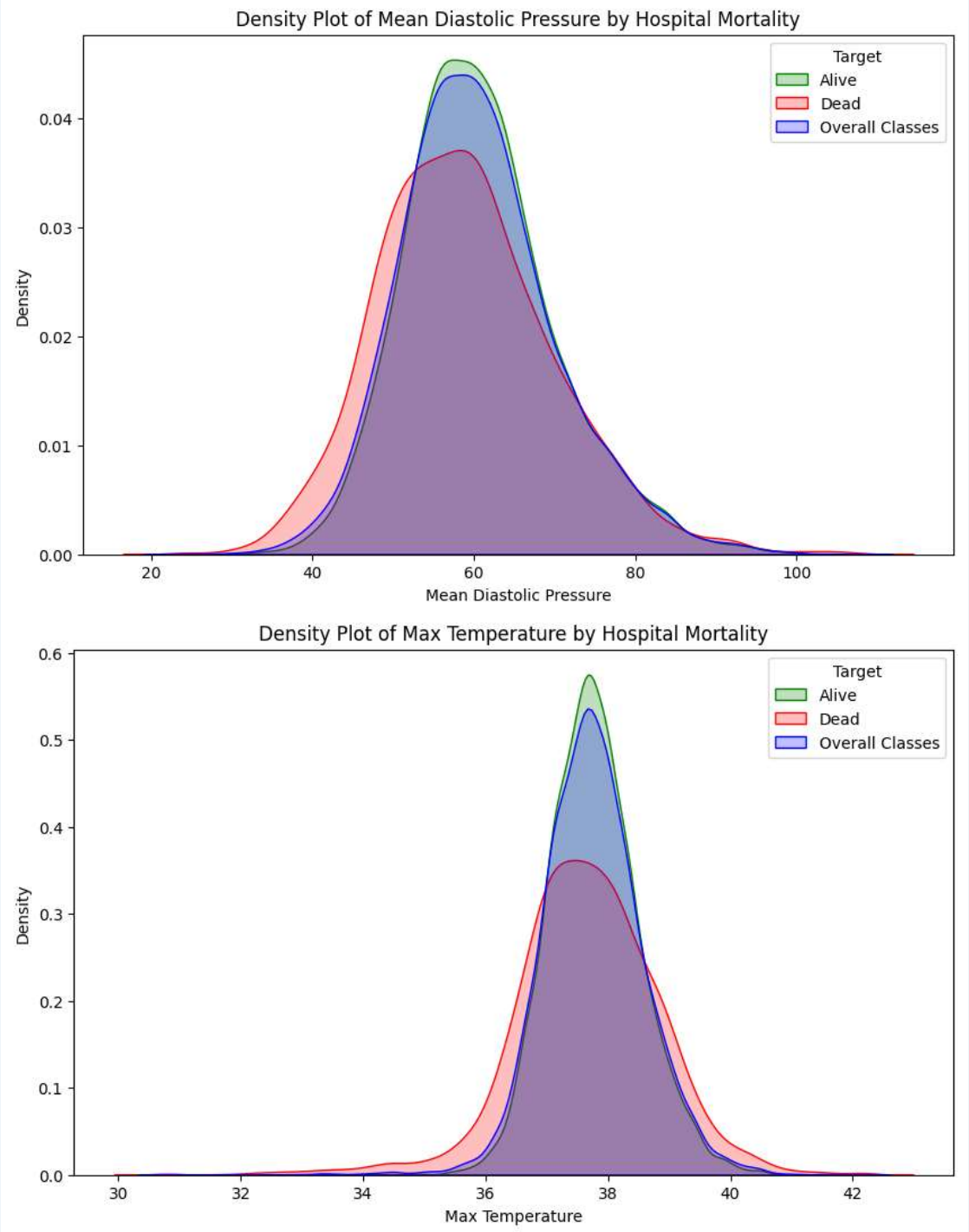

Density Plot of Min Heart Rate by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
140
Min Heart Rate
Density Plot of Mean Heart Rate by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.030
0.025
0.020
0.015
0.010
0.005
0.000
40
60
80
100
120
140
160
Mean Heart Rate

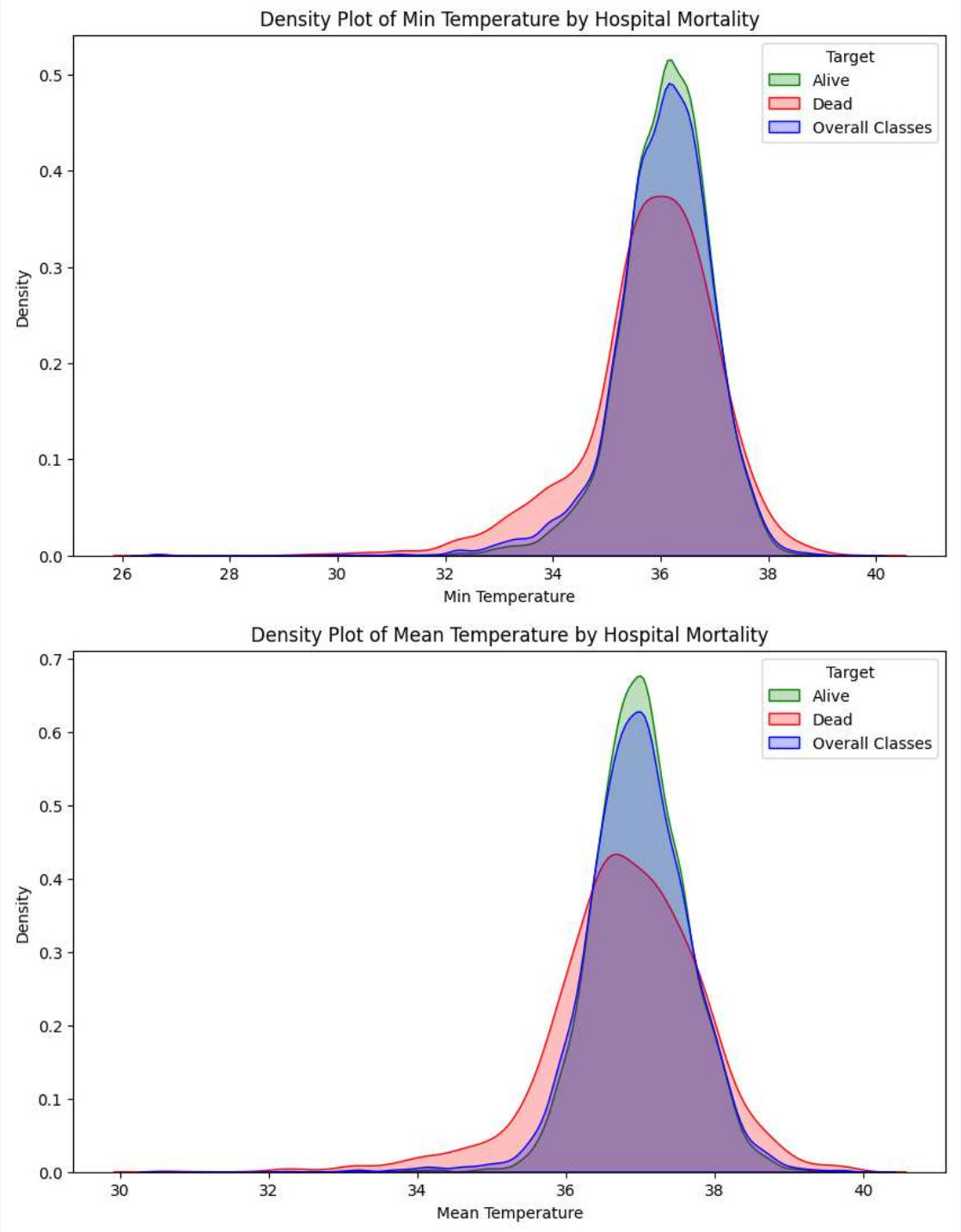

Density Plot of Max MAP by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.025
0.020
0.015
0.010
0.005
0.000
50
100
150
200
250
300
Max MAP
Density Plot of Min MAP by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
Min MAP

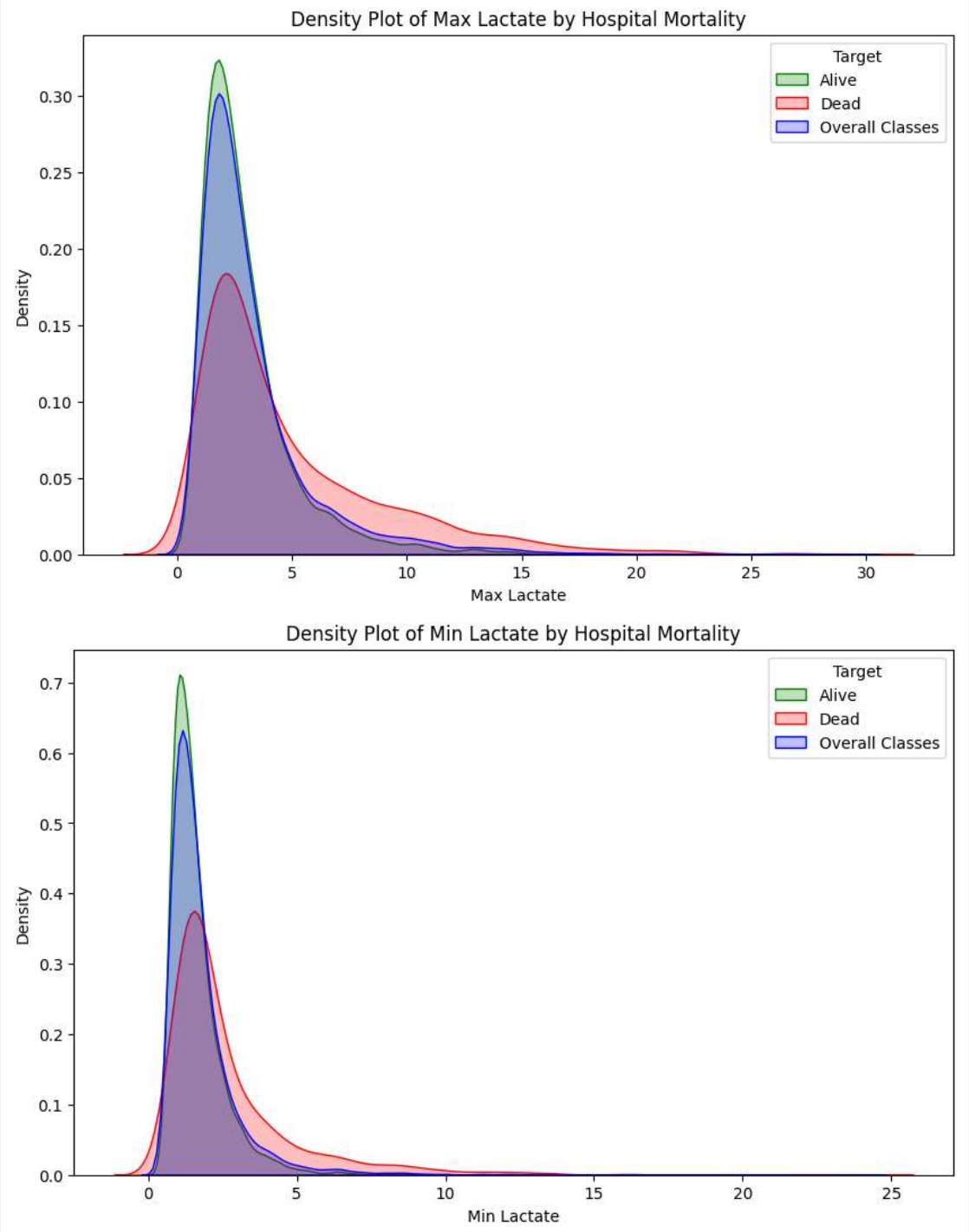

Density Plot of Mean MAP by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
Mean MAP
0.00
0.01
0.02
0.03
0.04
0.05
40
60
80
100
120
140
Density Plot of Max Systolic Pressure by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
Max Systolic Pressure
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0.0200
50
100
150
200
250
300

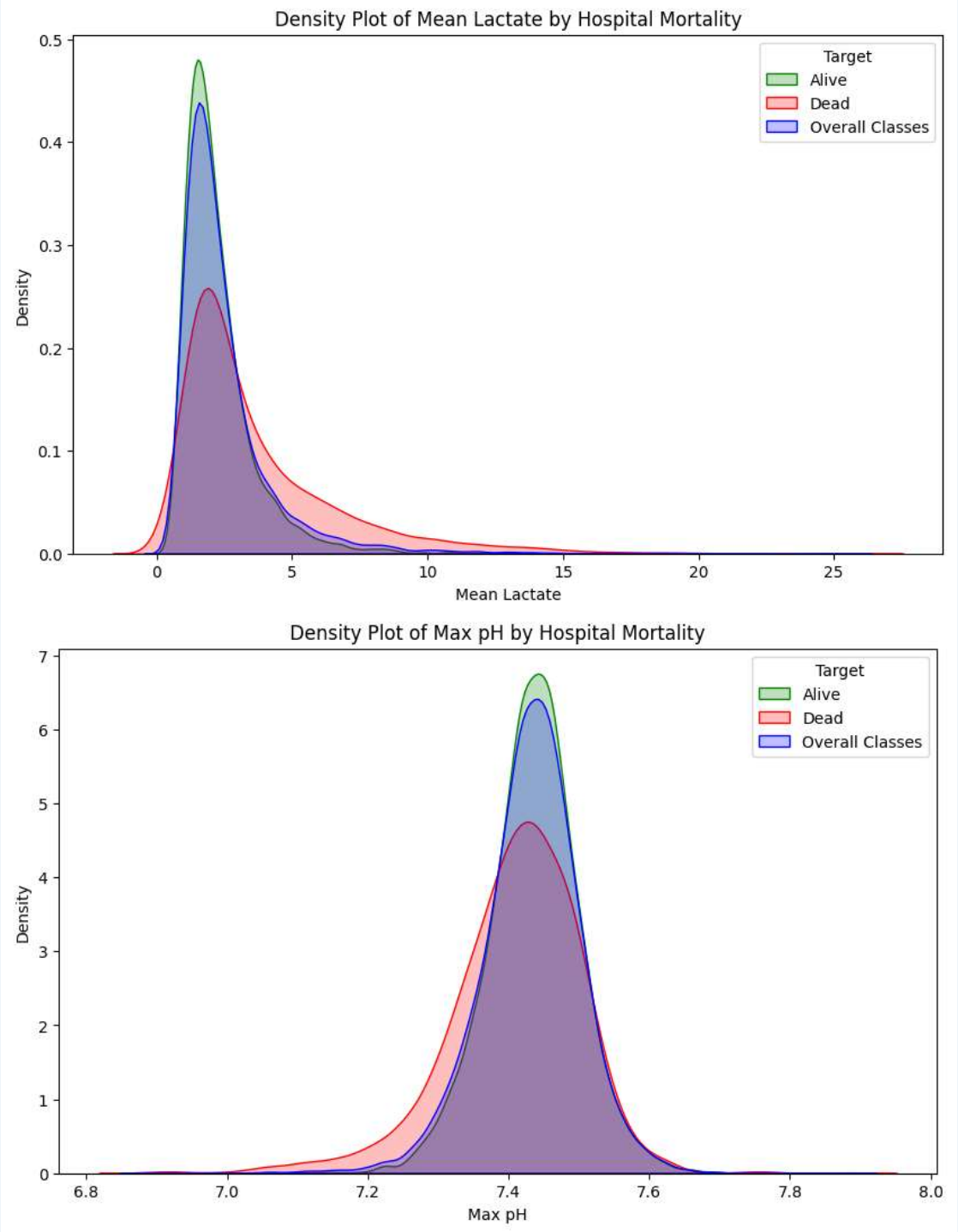

Density Plot of Min Systolic Pressure by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
25
50
75
100
125
150
175
Min Systolic Pressure
Density Plot of Mean Systolic Pressure by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.030
0.025
0.020
0.015
0.010
0.005
0.000
50
75
100
125
150
175
200
Mean Systolic Pressure

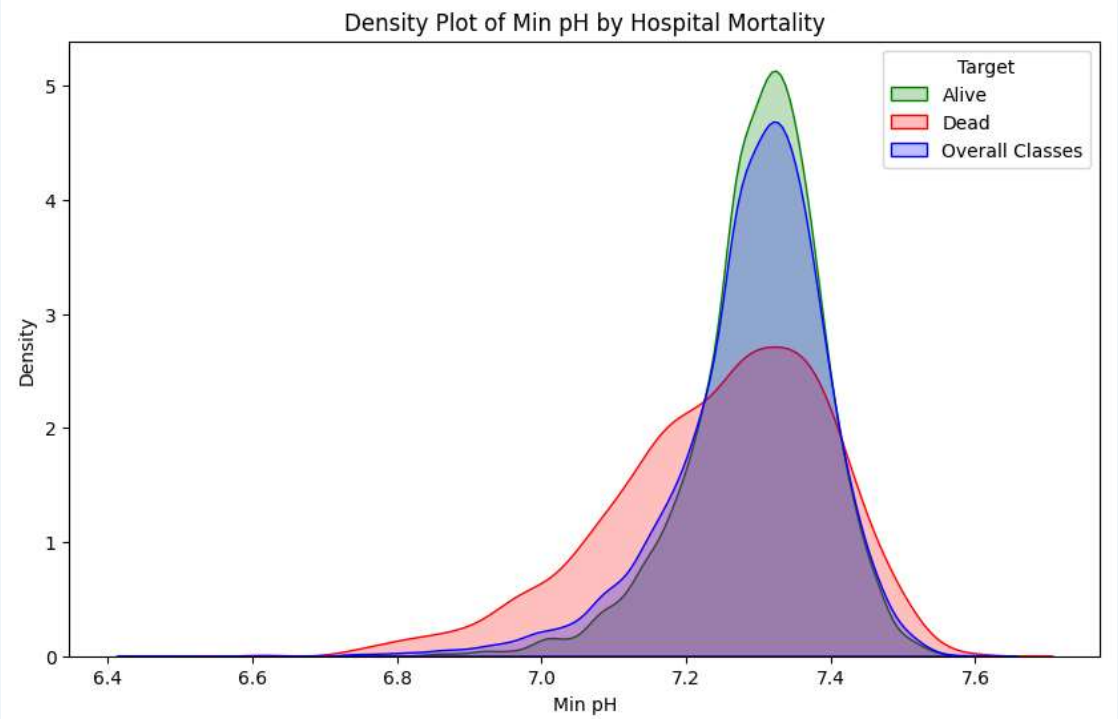

Density Plot of Max Diastolic Pressure by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.025
0.020
0.015
0.010
0.005
0.000
50
100
150
200
250
300
Max Diastolic Pressure
Density Plot of Min Diastolic Pressure by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.04
0.03
0.02
0.01
0.00
0
20
40
60
80
100
Min Diastolic Pressure

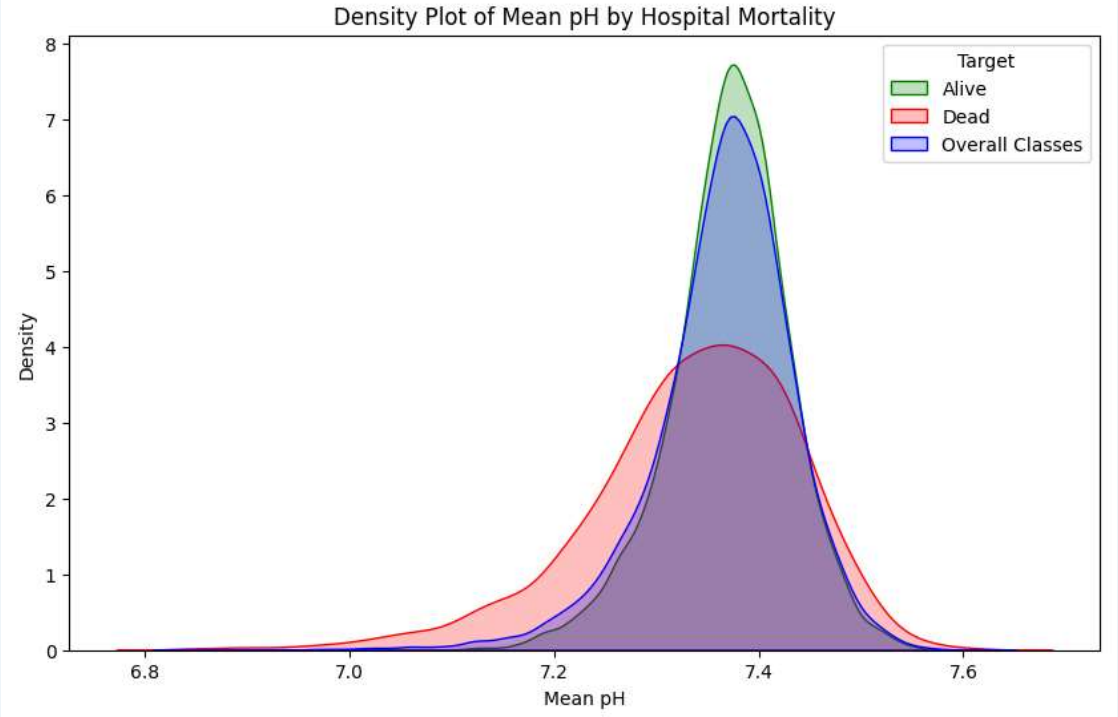

Density Plot of Mean Diastolic Pressure by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.00
0.01
0.02
0.03
0.04
20
40
60
80
100
Mean Diastolic Pressure
Density Plot of Max Temperature by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
30
32
34
36
38
40
42
Max Temperature

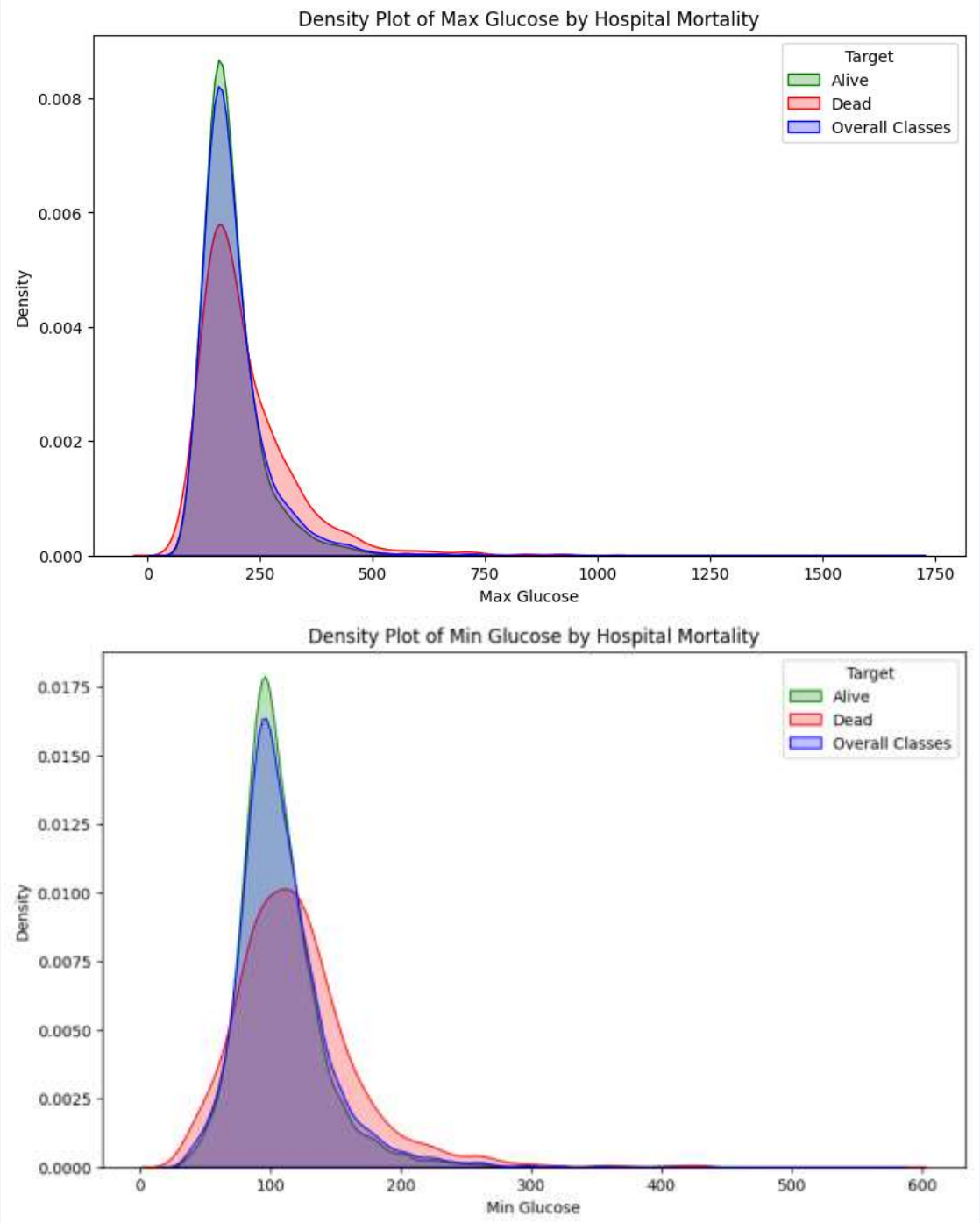

Density Plot of Min Temperature by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
Min Temperature
0.0
0.1
0.2
0.3
0.4
0.5
26
28
30
32
34
36
38
40
Density Plot of Mean Temperature by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
Mean Temperature
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
30
32
34
36
38
40

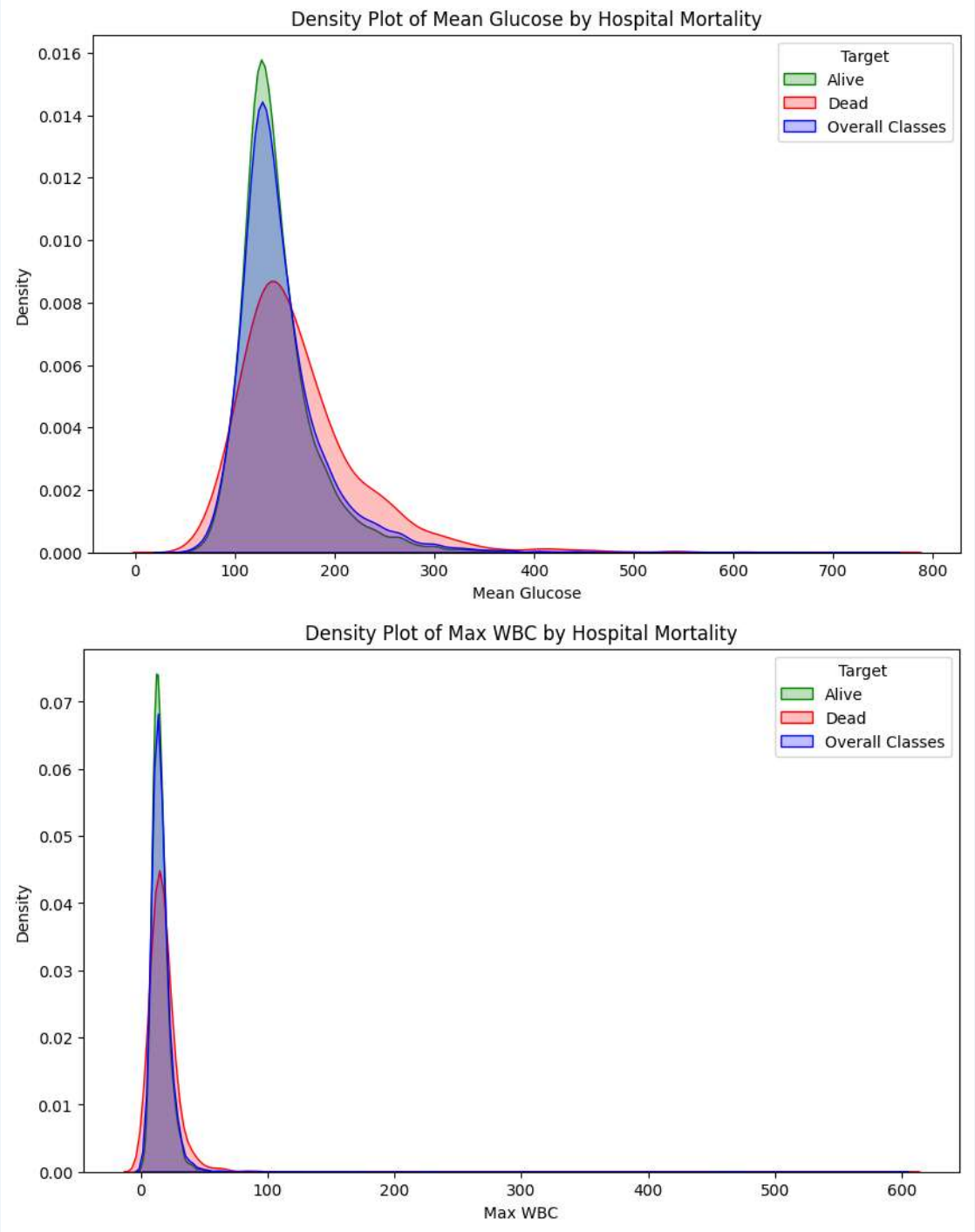

Density Plot of Max Lactate by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
5
10
15
20
25
30
Max Lactate
Density Plot of Min Lactate by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
5
10
15
20
25
Min Lactate

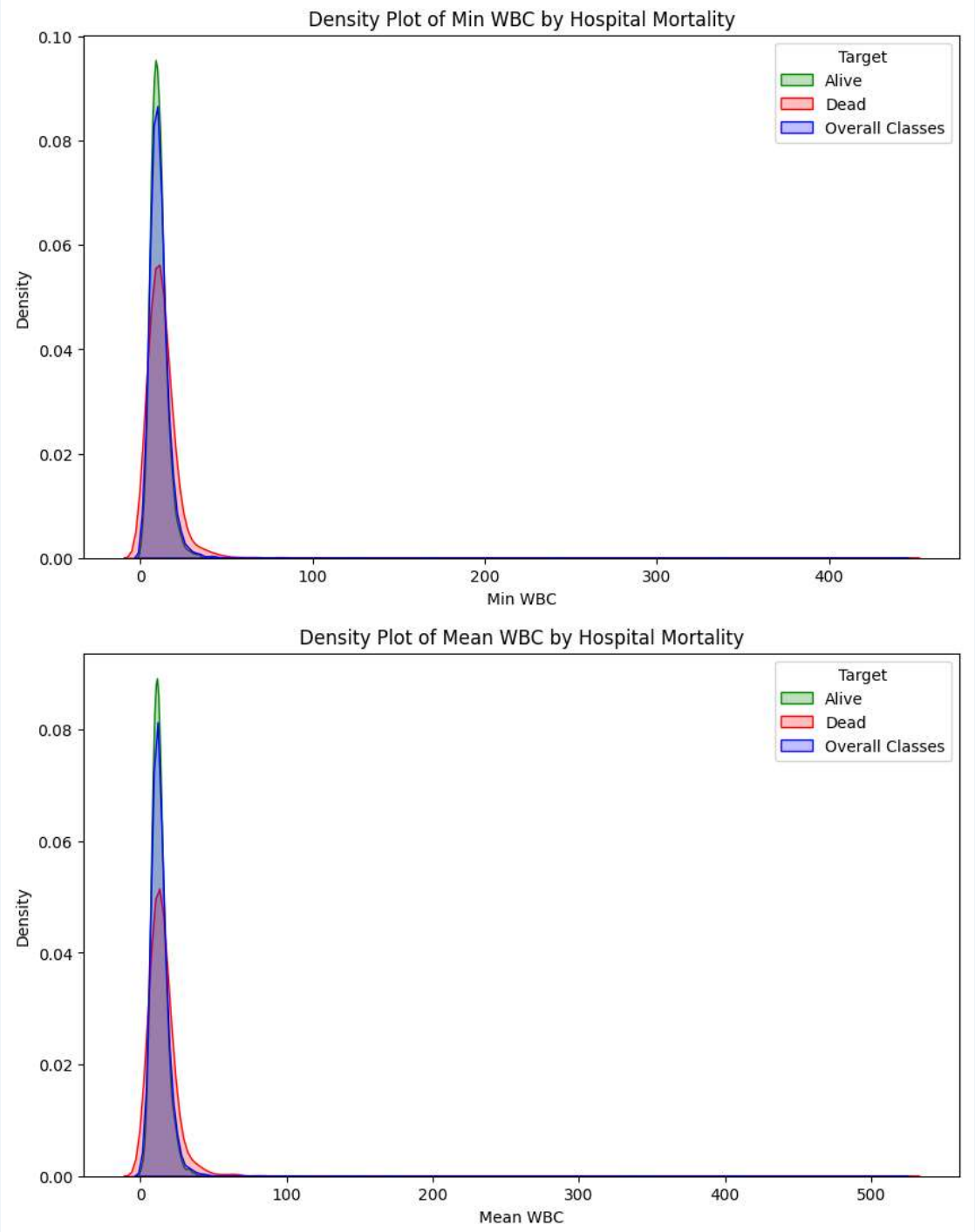

Density Plot of Mean Lactate by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
Mean Lactate
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
20
25
Density Plot of Max pH by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
Max pH
0
1
2
3
4
5
6
7
6.8
7.0
7.2
7.4
7.6
7.8
8.0

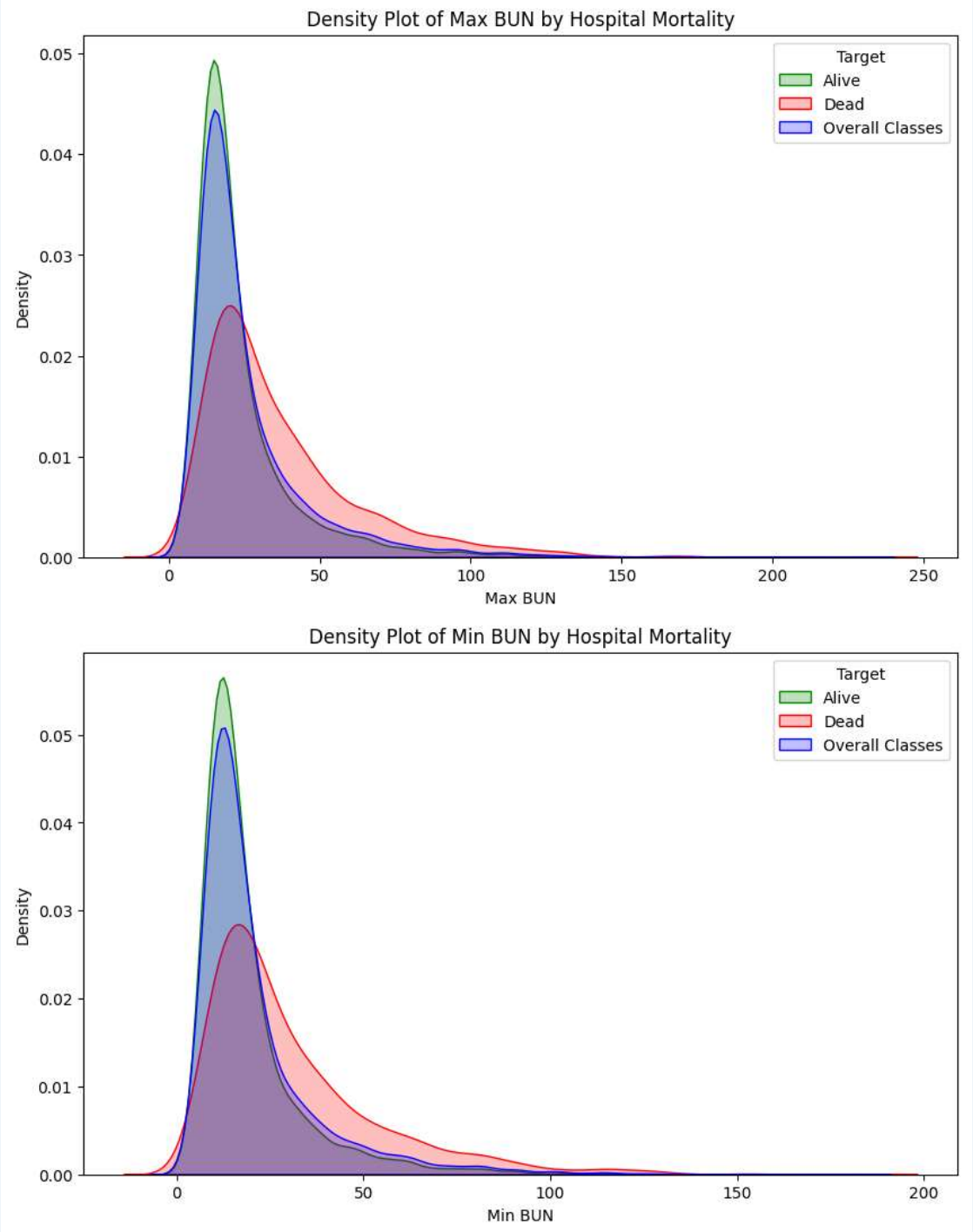

Density Plot of Min pH by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
Min pH
0
1
2
3
4
5
6.4
6.6
6.8
7.0
7.2
7.4
7.6

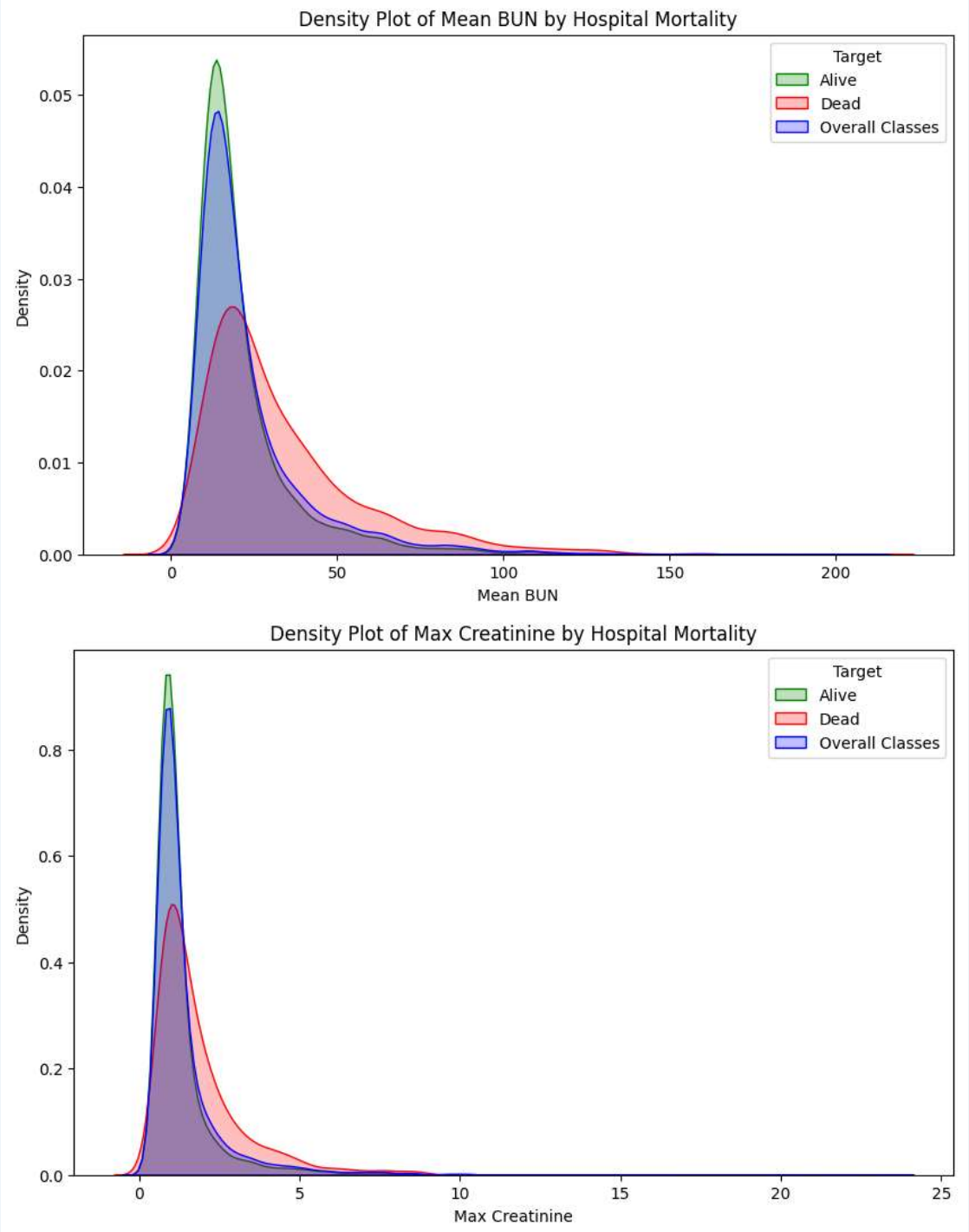

Density Plot of Mean pH by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
Mean pH
0
1
2
3
4
5
6
7
8
6.8
7.0
7.2
7.4
7.6

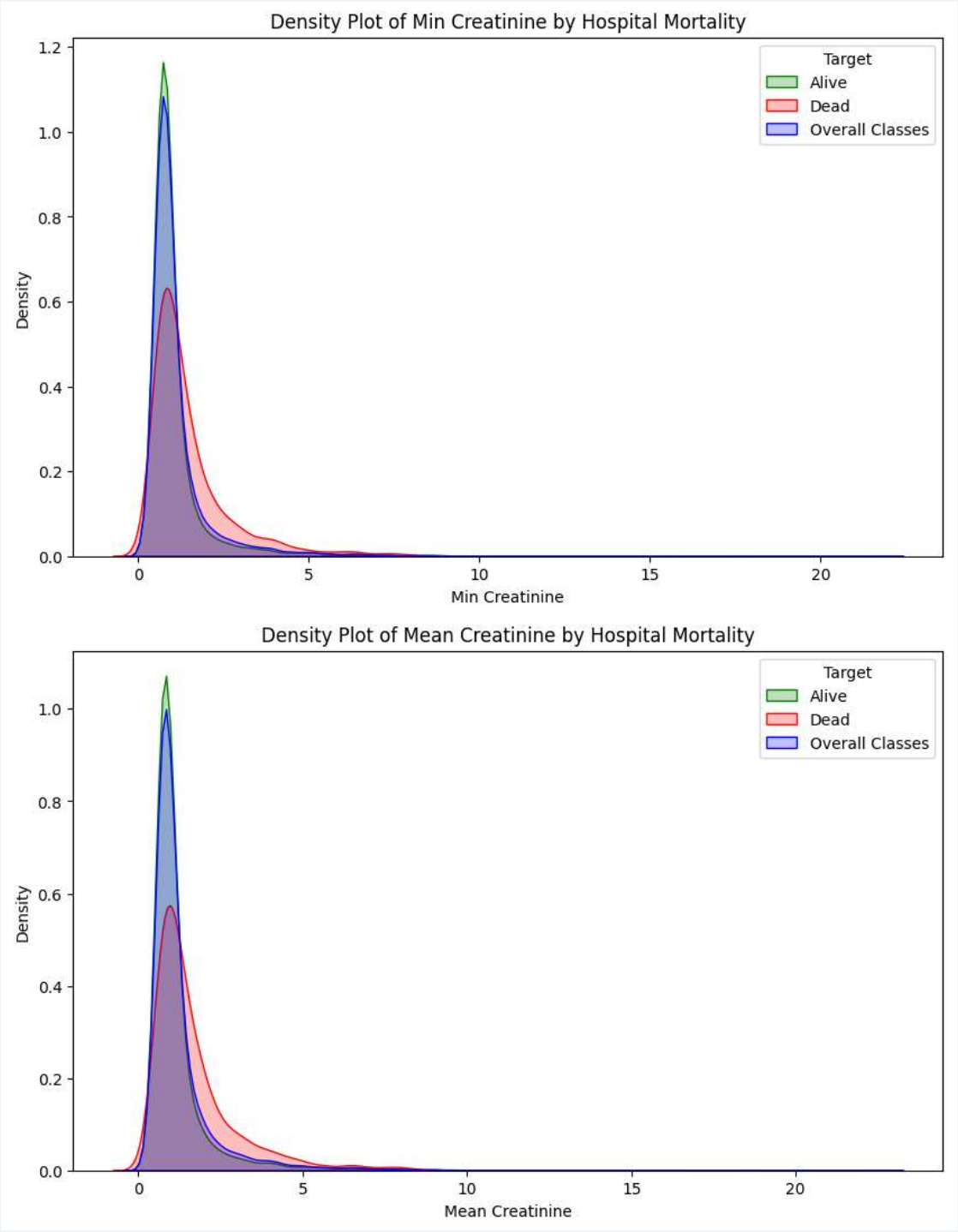

Density Plot of Max Glucose by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.000
0.002
0.004
0.006
0.008
0
250
500
750
1000
1250
1500
1750
Max Glucose
Density Plot of Min Glucose by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0
100
200
300
400
500
600
Min Glucose

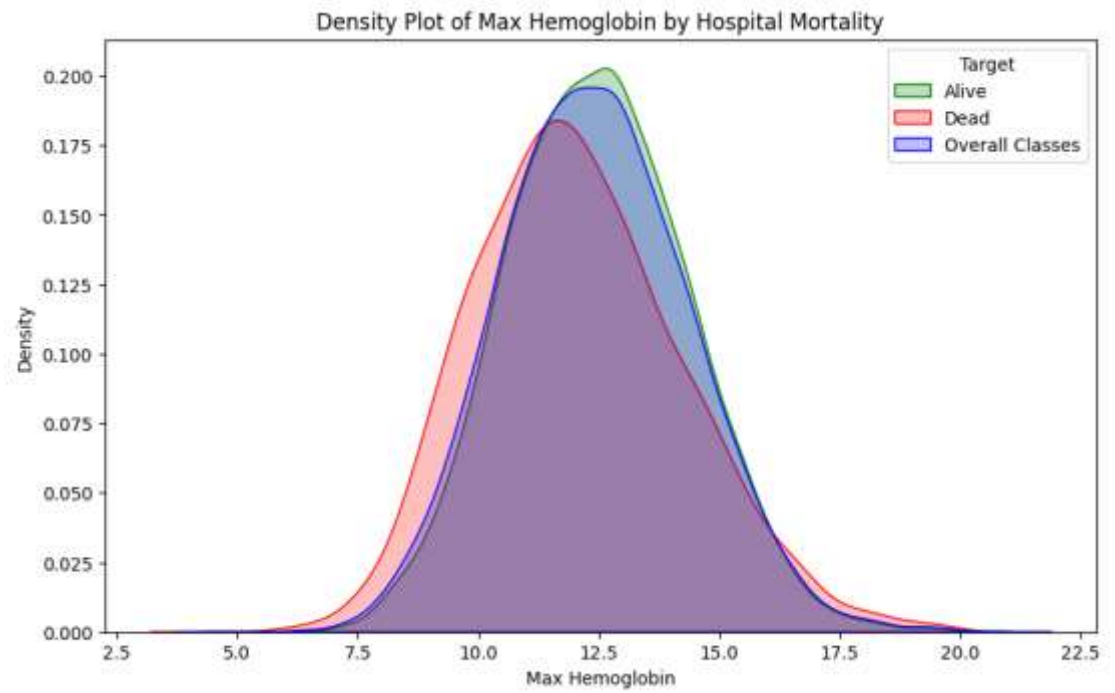

Density Plot of Mean Glucose by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
100
200
300
400
500
600
700
800
Mean Glucose
Density Plot of Max WBC by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
100
200
300
400
500
600
Max WBC

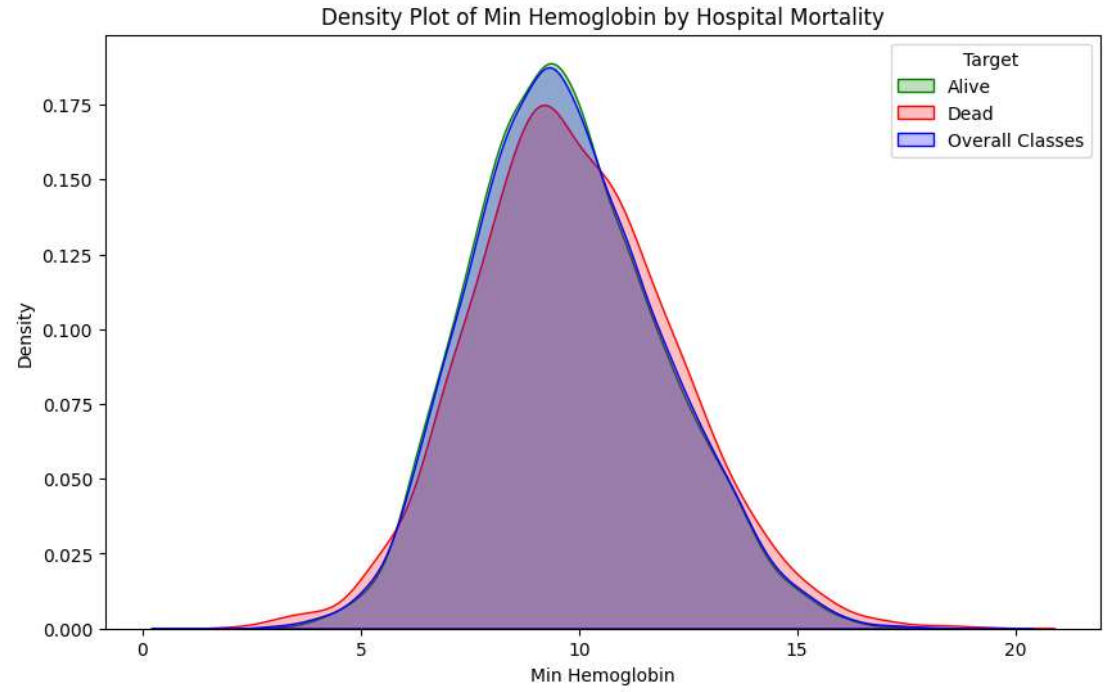

Density Plot of Min WBC by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.10
0.08
0.06
0.04
0.02
0.00
0
100
200
300
400
Min WBC
Density Plot of Mean WBC by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.08
0.06
0.04
0.02
0.00
0
100
200
300
400
500
Mean WBC

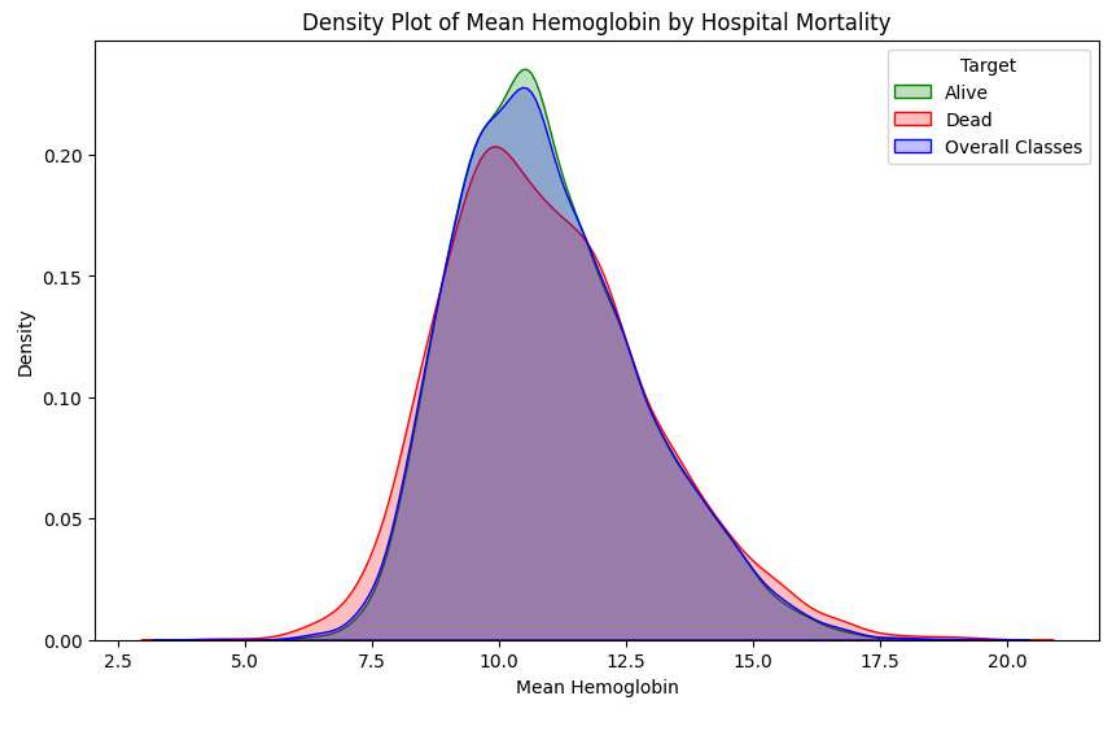

Density Plot of Max BUN by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.05
0.04
0.03
0.02
0.01
0.00
0
50
100
150
200
250
Max BUN
Density Plot of Min BUN by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.05
0.04
0.03
0.02
0.01
0.00
0
50
100
150
200
Min BUN


Density Plot of Mean BUN by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.00
0.01
0.02
0.03
0.04
0.05
0
50
100
150
200
Mean BUN
Density Plot of Max Creatinine by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.0
0.2
0.4
0.6
0.8
0
5
10
15
20
25
Max Creatinine


Density Plot of Min Creatinine by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
Min Creatinine
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
Density Plot of Mean Creatinine by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
Mean Creatinine
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20


Density Plot of Max Hemoglobin by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
Max Hemoglobin


Density Plot of Min Hemoglobin by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
5
10
15
20
Min Hemoglobin


Density Plot of Mean Hemoglobin by Hospital Mortality
Target
Alive
Dead
Overall Classes
Density
0.00
0.05
0.10
0.15
0.20
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Mean Hemoglobin