Data Preprocessing EDA and Feature Engineering

Zhenyu Zhang

Section 1: Data Understanding

Objectives:

- Understand the structure and purpose of key MIMIC-III tables.
- Identify relationships between tables to facilitate feature extraction.
- Review variable definitions and units of measurement.

Steps:

1.1 **Overview of all Tables**:

- Focus on the following pre-processed tables in 'mimic_data' folder: - `admissions.csv` : Admission details, including admission and discharge times. - `antibiotics.csv` : Antibiotic usage data. - `bloodculture.csv` : Results of blood culture tests. - `gcs_hourly.csv` : Glasgow Coma Score records. - `icd9_diag.csv` : ICD-9 diagnostic codes for patient conditions. - `icustays.csv` : ICU stay details (e.g., admission, discharge times). - `labs_hourly.csv` : Hourly laboratory results. - `output_hourly.csv` : Fluid output data. - `patients.csv` : Demographics and mortality information. - `pt_icu_outcome.csv` : Patient outcomes (e.g., mortality) per ICU stay. - `pt_stay_hr.csv` : Hourly records of ICU stays. - `pt_weight.csv` : Patient weight records. - `pv_mechvent.csv` : Mechanical ventilation data. - `transfers.csv` : Information on patient transfers within the hospital. - `vasopressors.csv` : Administration of vasopressors. - `vitals_hourly.csv` : Hourly vital sign measurements.
- Use `data.table` for efficient loading of large datasets.

1.2 **Relationships Between Tables**:

- Key relationships include:
  - `subject_id` : Links `patients` , `admissions` , and `icustays` .
  - `hadm_id` : Links `admissions` and `icustays` .
  - `icustay_id` : Links ICU-specific data (e.g., `vitals_hourly` , `labs_hourly` , etc.).
  - Other tables (e.g., `antibiotics` , `bloodculture`) use these IDs to connect to patient-specific data.

1.3 **Initial Summarisation**:

- Explore each table:
  - Number of rows and columns.
  - Key variables and their data type.
  - Missing data percentages.

```
##
## ### Summary of Table: admissions ###
## Number of rows: 58976
## Number of columns: 19
## Column names and data types:
##                      Data_Type Missing_Count Missing_Pct.
## row_id                 integer             0         0.00
## subject_id             integer             0         0.00
## hadm_id                integer             0         0.00
## admittime              POSIXct             0         0.00
## dischtime              POSIXct             0         0.00
## deathtime              POSIXct         53122        90.07
## admission_type       character             0         0.00
## admission_location   character             0         0.00
## discharge_location   character             0         0.00
## insurance            character             0         0.00
## language             character             0         0.00
## religion             character             0         0.00
## marital_status       character             0         0.00
## ethnicity            character             0         0.00
## edregtime              POSIXct         28099        47.64
## edouttime              POSIXct         28099        47.64
## diagnosis            character             0         0.00
## hospital_expire_flag   integer             0         0.00
## has_chartevents_data   integer             0         0.00
##
## ---
##
## ### Summary of Table: antibiotics ###
## Number of rows: 164927
## Number of columns: 16
## Column names and data types:
##                   Data_Type Missing_Count Missing_Pct.
## icustay_id          integer            74         0.04
## starttime           POSIXct            24         0.01
## endtime             POSIXct             7         0.00
## amount              numeric             7         0.00
## amountuom         character             0         0.00
## rate                logical        164927       100.00
## rateuom             logical        164927       100.00
## ordercategoryname character             0         0.00
## patientweight       numeric            31         0.02
## totalamount         integer          8315         5.04
## totalamountuom    character             0         0.00
## statusdescription character             0         0.00
## label             character             0         0.00
## abbreviation      character             0         0.00
## antibiotic          integer             0         0.00
## dbsource          character             0         0.00
##
## ---
##
## ### Summary of Table: bloodculture ###
## Number of rows: 632506
## Number of columns: 10
```

```
## Column names and data types:
##                      Data_Type Missing_Count Missing_Pct.
## hadm_id                integer             0         0.00
## icustay_id             integer        156798        24.79
## dy                     integer        156798        24.79
## hr                     integer        187726        29.68
## charttime              POSIXct         41791         6.61
## chartdate              POSIXct             0         0.00
## org_name             character             0         0.00
## positiveculture        integer             0         0.00
## ab_name              character             0         0.00
## antibioticresistance character             0         0.00
##
## ---
##
## ### Summary of Table: gcs_hourly ###
## Number of rows: 1515342
## Number of columns: 7
## Column names and data types:
##               Data_Type Missing_Count Missing_Pct.
## icustay_id      integer             0         0.00
## hr              integer             0         0.00
## gcs             integer             0         0.00
## gcseyes         integer          1444         0.10
## gcsmotor        integer          3768         0.25
## gcsverbal       integer          3568         0.24
## endotrachflag   integer             0         0.00
##
## ---
##
## ### Summary of Table: icd9_diag ###
## Number of rows: 651047
## Number of columns: 7
## Column names and data types:
##             Data_Type Missing_Count Missing_Pct.
## row_id        integer             0         0.00
## subject_id    integer             0         0.00
## hadm_id       integer             0         0.00
## seq_num       integer            47         0.01
## icd9_code   character             0         0.00
## short_title character             0         0.00
## long_title  character             0         0.00
##
## ---
##
## ### Summary of Table: icustays ###
## Number of rows: 61532
## Number of columns: 12
## Column names and data types:
##                Data_Type Missing_Count Missing_Pct.
## row_id           integer             0         0.00
## subject_id       integer             0         0.00
## hadm_id          integer             0         0.00
## icustay_id       integer             0         0.00
## dbsource       character             0         0.00
## first_careunit character             0         0.00
```

```
## last_careunit  character            0         0.00
## first_wardid     integer            0         0.00
## last_wardid      integer            0         0.00
## intime           POSIXct            0         0.00
## outtime          POSIXct           10         0.02
## los              numeric           10         0.02
##
## ---
##
## ### Summary of Table: labs_hourly ###
## Number of rows: 928195
## Number of columns: 22
## Column names and data types:
##                        Data_Type Missing_Count Missing_Pct.
## icustay_id               integer             0         0.00
## hr                       integer             0         0.00
## neutrophil               numeric        853757        91.98
## creactiveprotein         numeric        926422        99.81
## whitebloodcell           numeric        576380        62.10
## partialpressureo2        numeric        473825        51.05
## bicarbonate              numeric        534726        57.61
## lactate                  numeric        763646        82.27
## troponin                 numeric        886778        95.54
## bloodureanitrogen        numeric        549233        59.17
## creatinine               numeric        547907        59.03
## alaninetransaminase      numeric        836577        90.13
## aspartatetransaminase    numeric        836642        90.14
## hemoglobin               numeric        514069        55.38
## intnormalisedratio       numeric        675513        72.78
## platelets                numeric        559952        60.33
## albumin                  numeric        866292        93.33
## chloride                 numeric        496072        53.44
## glucose                  numeric        401175        43.22
## sodium                   numeric        524290        56.48
## bilirubin                numeric        816141        87.93
## hematocrit               numeric        487396        52.51
##
## ---
##
## ### Summary of Table: output_hourly ###
## Number of rows: 3325543
## Number of columns: 3
## Column names and data types:
##             Data_Type Missing_Count Missing_Pct.
## icustay_id    integer             0         0.00
## hr            integer             0         0.00
## urineoutput   numeric         11920         0.36
##
## ---
##
## ### Summary of Table: patients ###
## Number of rows: 46520
## Number of columns: 8
## Column names and data types:
##             Data_Type Missing_Count Missing_Pct.
## row_id        integer             0         0.00
```

```
## subject_id    integer             0          0.00
## gender      character             0          0.00
## dob           POSIXct             0          0.00
## dod           POSIXct         30761         66.12
## dod_hosp      POSIXct         36546         78.56
## dod_ssn       POSIXct         33142         71.24
## expire_flag   integer             0          0.00
##
## ---
##
## ### Summary of Table: pt_icu_outcome ###
## Number of rows: 61533
## Number of columns: 17
## Column names and data types:
##                       Data_Type Missing_Count Missing_Pct.
## row_id                  integer             0         0.00
## subject_id              integer             0         0.00
## dob                     POSIXct             0         0.00
## hadm_id                 integer             0         0.00
## admittime               POSIXct         12348        20.07
## dischtime               POSIXct         12348        20.07
## icustay_id              integer             0         0.00
## age_years               numeric             0         0.00
## intime                  POSIXct             0         0.00
## outtime                 POSIXct            10         0.02
## los                     numeric            10         0.02
## hosp_deathtime          POSIXct         59256        96.30
## icu_expire_flag         integer             0         0.00
## hospital_expire_flag    integer         12348        20.07
## dod                     POSIXct         37341        60.68
## expire_flag             integer             0         0.00
## ttd_days                integer         37341        60.68
##
## ---
```

```
##
## ### Summary of Table: pt_stay_hr ###
## Number of rows: 3687586
## Number of columns: 9
## Column names and data types:
##            Data_Type Missing_Count Missing_Pct.
## icustay_id   integer             0         0.00
## hadm_id      integer             0         0.00
## subject_id   integer             0         0.00
## intime       POSIXct             0         0.00
## outtime      POSIXct             0         0.00
## starttime    POSIXct             0         0.00
## endtime      POSIXct             0         0.00
## hr           integer             0         0.00
## dy           integer          1398         0.04
##
## ---
##
## ### Summary of Table: pt_weight ###
## Number of rows: 396241
## Number of columns: 11
## Column names and data types:
##                  Data_Type Missing_Count Missing_Pct.
## icustay_id         integer             0         0.00
## dy                 integer             0         0.00
## starttime          POSIXct             0         0.00
## endtime            POSIXct             0         0.00
## admissionweight    numeric        331567        83.68
## dailyweight        numeric        241485        60.94
## previousweight     numeric        312334        78.82
## echoweight         numeric        262244        66.18
## avg_weight_naive   numeric         20263         5.11
## min_weight         numeric         20263         5.11
## max_weight         numeric         20263         5.11
##
## ---
```

```
##
## ### Summary of Table: pv_mechvent ###
## Number of rows: 694958
## Number of columns: 21
## Column names and data types:
##                    Data_Type Missing_Count Missing_Pct.
## icustay_id           integer             0         0.00
## charttime            POSIXct            76         0.01
## starttime            POSIXct             0         0.00
## endtime              POSIXct             0         0.00
## duration_hours       numeric             0         0.00
## ventnum              integer             0         0.00
## minutevolume         numeric        233141        33.55
## settidalvolume       numeric        461276        66.37
## obstidalvolume       numeric        314290        45.22
## sponttidalvolume     numeric        439527        63.25
## setpeep              numeric         21446         3.09
## totalpeep            numeric        669702        96.37
## pressurehighaprv     integer        694415        99.92
## pressurelowaprv      numeric        694428        99.92
## timehighaprv         numeric        694419        99.92
## timelowaprv          numeric        694423        99.92
## meanairwaypressure   numeric        236821        34.08
## peakinsppressure     numeric        322901        46.46
## neginspforce         numeric        693945        99.85
## insptime             numeric        572792        82.42
## plateaupressure      numeric        564736        81.26
##
## ---
##
## ### Summary of Table: transfers ###
## Number of rows: 261897
## Number of columns: 13
## Column names and data types:
##               Data_Type Missing_Count Missing_Pct.
## row_id          integer             0         0.00
## subject_id      integer             0         0.00
## hadm_id         integer             0         0.00
## icustay_id      integer        174176        66.51
## dbsource      character             0         0.00
## eventtype     character             0         0.00
## prev_careunit character             0         0.00
## curr_careunit character             0         0.00
## prev_wardid     integer         58933        22.50
## curr_wardid     integer         58943        22.51
## intime          POSIXct            24         0.01
## outtime         POSIXct         58976        22.52
## los             numeric         58976        22.52
##
## ---
##
## ### Summary of Table: vasopressors ###
## Number of rows: 314964
## Number of columns: 11
## Column names and data types:
```

```
##                         Data_Type Missing_Count Missing_Pct.
## icustay_id                integer           834         0.26
## starttime                 POSIXct             0         0.00
## endtime                   POSIXct        231185        73.40
## norepinephrine_rate       numeric        231436        73.48
## norepinephrine_amount     numeric        237721        75.48
## epinephrine_rate          numeric        259713        82.46
## epinephrine_amount        numeric        278082        88.29
## dopamine_rate             numeric        181993        57.78
## dopamine_amount           numeric        227051        72.09
## dobutamine_rate           numeric        272450        86.50
## dobutamine_amount         numeric        286648        91.01
##
## ---
##
## ### Summary of Table: vitals_hourly ###
## Number of rows: 7292362
## Number of columns: 11
## Column names and data types:
##                       Data_Type Missing_Count Missing_Pct.
## icustay_id              integer             0         0.00
## hr                      integer             0         0.00
## spo2                    numeric       1972385        27.05
## fio2                    numeric       6341965        86.97
## temperature             numeric       5726438        78.53
## resprate                numeric       2550173        34.97
## heartrate               numeric        872012        11.96
## sysbp                   numeric       2647890        36.31
## diasbp                  numeric       2648668        36.32
## glucose                 numeric       6144343        84.26
## meanarterialpressure    numeric       2631818        36.09
##
## ---
```

```
##
## All CSV tables have been successfully loaded and summarized!
```

```
## Loaded tables:
##  admissions, antibiotics, bloodculture, gcs_hourly, icd9_diag, icustays, labs_hourly, output
_hourly, patients, pt_icu_outcome, pt_stay_hr, pt_weight, pv_mechvent, transfers, vasopressors,
vitals_hourly
```

1. **admissions**

- **Rows**: 58,976 | **Columns**: 19
- **Key Missingness**:
  - `deathtime` : 90.07% missing. Relevant for mortality analysis but likely reflects non-deceased patients.
  - Minimal missingness for core variables like `admittime` , `dischtime` , and demographic details.
- **Observation**: High-quality foundational data with minimal issues, apart from `deathtime` .

2. **antibiotics**

- **Rows**: 164,927 | **Columns**: 16
- **Key Missingness**:

- `rate` and `rateuom` : Both 100% missing, suggesting these variables can be dropped.
- `totalamount` : 5.04% missing.

- **Observation**: Useful for understanding antibiotic administration, though some variables appear irrelevant.

3. bloodculture

- **Rows**: 632,506 | **Columns**: 10
- **Key Missingness**:
  - `icustay_id` : 24.79% missing, significant for ICU-related analyses.
  - `hr` : 29.68% missing.
- **Observation**: Moderate missingness for key ICU identifiers may limit linking with other tables.

4. gcs_hourly

- **Rows**: 1,515,342 | **Columns**: 7
- **Key Missingness**:
  - `gcseyes` , `gcsmotor` , and `gcsverbal` : <0.3% missing, indicating good data quality.
- **Observation**: Reliable source for Glasgow Coma Scale (GCS) data with low missingness.

5. icd9_diag

- **Rows**: 651,047 | **Columns**: 7
- **Key Missingness**:
  - Minimal issues, with <0.01% missing in `seq_num` .
- **Observation**: High-quality diagnosis data, ready for analysis.

6. icustays

- **Rows**: 61,532 | **Columns**: 12
- **Key Missingness**:
  - `outtime` and `los` : Both 0.02% missing.
- **Observation**: Reliable ICU stay details with minimal issues.

7. labs_hourly

- **Rows**: 928,195 | **Columns**: 22
- **Key Missingness**:
  - Many variables exceed 90% missingness, including `creactiveprotein (99.81%)` and `alaninetransaminase (90.13%)` .
  - Core variables like `neutrophil` (91.98%) also have high missingness.
- **Observation**: Key lab data but requires careful selection and imputation due to widespread missingness.

8. output_hourly

- **Rows**: 3,325,543 | **Columns**: 3
- **Key Missingness**:
  - `urineoutput` : 0.36% missing.
- **Observation**: High-quality output data with negligible issues.

9. patients

- **Rows**: 46,520 | **Columns**: 8
- **Key Missingness**:
  - Mortality-related fields (`dod_hosp` , `dod_ssn`) have >70% missingness.
  - Demographic fields like `gender` and `dob` are complete.
- **Observation**: Core patient demographics are robust, but mortality data requires handling.

10. pt_icu_outcome

- **Rows**: 61,533 | **Columns**: 17
- **Key Missingness**:
  - Critical fields like `hosp_deathtime (96.30%)` and `dod (60.68%)` have very high missingness.
- **Observation**: ICU outcomes are incomplete for most patients.

11. pt_stay_hr

- **Rows**: 3,687,586 | **Columns**: 9
- **Key Missingness**:
  - Minimal, with `dy` missing 0.04%.
- **Observation**: Comprehensive hourly stay data with excellent quality.

12. pt_weight

- **Rows**: 396,241 | **Columns**: 11
- **Key Missingness**:
  - Most weight-related fields exceed 60% missingness, e.g., `admissionweight (83.68%)` .
- **Observation**: Data quality for weight variables is poor, limiting analysis.

13. pv_mechvent

- **Rows**: 694,958 | **Columns**: 21
- **Key Missingness**:
  - Ventilation parameters like `pressurehighaprv` and `timelowaprv` exceed 99% missingness.
  - `minutevolume` : 33.55% missing.
- **Observation**: Highly sparse data, with only a few useful variables.

14. transfers

- **Rows**: 261,897 | **Columns**: 13
- **Key Missingness**:
  - `icustay_id` : 66.51% missing.
  - Ward identifiers (`prev_wardid` and `curr_wardid`) have ~22.5% missingness.
- **Observation**: Transfer details are partially incomplete, limiting their utility.

15. vasopressors

- **Rows**: 314,964 | **Columns**: 11
- **Key Missingness**:
  - Missingness ranges from 57.78% (`dopamine_rate`) to 91.01% (`dobutamine_amount`).
- **Observation**: Sparse data for vasopressor administration, with limited reliable variables.

16. vitals_hourly

- **Rows**: 7,292,362 | **Columns**: 11
- **Key Missingness**:
  - Vital signs like `fio2` and `temperature` exceed 75% missingness.
  - Core variables like `heartrate` and `spo2` are ~10-30% missing.
- **Observation**: Rich time-series data but requires significant preprocessing.

1.4 Key Observations from Data:

1.4.1 High Missingness in Time-Series Data (`vitals_hourly` and `labs_hourly`):

- Many variables in `labs_hourly` and `vitals_hourly` exceed **70% missingness**.
- Some variables (`creactiveprotein` , `alaninetransaminase`) in `labs_hourly` have almost **complete missingness**, making them unsuitable for imputation or analysis.

1.4.2 **Time Discrepancies in `hr` Across Tables**:

- `vitals_hourly` : `hr` starts at 1 (post-ICU admission) and increments hourly.
- `labs_hourly` : `hr` includes negative values for pre-ICU measurements.
- Different intervals or irregular sampling times make direct alignment across tables challenging.

1.4.3 **Key Tables for the First 24 Hours**:

- `pt_stay_hr` : Provides a comprehensive hourly structure for ICU stays and can act as a unifying table for `hr` alignment.
- `vitals_hourly` and `labs_hourly` : Crucial for predictive modeling but need proper filtering for the first 24 hours.

1.4.4 **Predictive Modeling Needs**:

- Accurate prediction of mortality requires reliable features extracted from the **first 24 hours**.
- Time-sensitive modeling approaches (e.g., LSTMs, GRUs) need continuous time-series data, while tree-based models (e.g., XGBoost) can use aggregated features.

1.5 Consideration for following analysis

Ensure high-quality data is used while minimizing the impact of missingness on analyses. #### 1.4.1 **Prioritize Tables with Low Missingness**: - `admissions` , `patients` , `icustays` , and `gcs_hourly` are the most reliable tables for initial analysis.

1.4.2 **Handle High Missingness Strategically**:

- For tables like `labs_hourly` and `vitals_hourly` , consider using imputation, variable selection, or excluding highly sparse variables.

1.4.3 **Focus on Core Time-Series Data**:

- `vitals_hourly` and `output_hourly` provide crucial insights into patient conditions, despite moderate missingness.

1.4.4 **Exclude Variables with >90% Missingness**:

- # Tables like `vasopressors` and `pv_mechvent` have several variables with near-complete missingness, which may not have value.

Section 2: Data Preprocessing

Objectives:

- Prepare the dataset for analysis by filtering, merging, and handling missing data.
- Ensure consistency and completeness in the preprocessed data.

Define the study population:

- Focus on ICU patients aged between **18 and 89 years**, Aligns with the MIMIC-III age shifting policy for HIPAA compliance and avoids pediatric and super-elderly populations.
- Retain only the **first ICU admission** per patient to ensure independence of observations and avoids over representation of specific patients.
- Ensure the ICU stay duration is **≥ 24 hours** for providing sufficient data for meaningful feature extraction.

- Add **Weekend/Weekday Flag**, which directly supports Aim 2, investigating mortality association with admission timing.

Validate inclusion and exclusion criteria

- Exclude records missing critical demographic variables like gender, age, or ICU admission/discharge times.
- Align with project aims to predict mortality using data from the **first 24 hours of ICU stay** (for predictors) but not restrict mortality outcomes to the same timeframe.

Steps:

2.1 **Merge Patients, Admissions, and ICU Stays**

- **Objective**: Combine demographic, admission, and ICU stay data into a cohesive dataset for initial filtering.
- **Why**: `patients` , `admissions` , and `icustays` tables provide core demographic and hospitalization data that form the backbone of our analysis.
- **How**:
    - Merge `patients` and `admissions` using the key `subject_id` to align patient demographics with their hospital admissions.
    - Merge the resulting dataset with `icustays` using the keys `subject_id` and `hadm_id` to include ICU-specific stay details.

2.2 **Retain the First ICU Admission per Patient**

- **Objective**: Ensure that each patient contributes only their first ICU admission to the analysis.
- **Why**: Retaining only the first ICU admission avoids over-representation of patients with multiple ICU stays and ensures independence of observations.
- **How**:
    - Sort the data by `subject_id` and `admittime` .
    - Use `.SD[1]` to retain the first ICU stay for each `subject_id` .

2.3 **Filter by Age (18 ≤ Age ≤ 89)**

- **Objective**: Focus on adult patients while excluding pediatric and super-elderly populations.
- **Why**: The MIMIC-III dataset masks ages above 89 due to HIPAA compliance, making exact age unknown for these patients.
- **How**:
    - Calculate patient age at admission as the difference between `admittime` and `dob` .
    - Retain records where age is between 18 and 89.

2.4 **Filter ICU Stays Lasting ≥ 24 Hours**

- **Objective**: Exclude ICU stays shorter than 24 hours to ensure sufficient data for analysis.
- **Why**: Short ICU stays may not provide enough information for meaningful predictive modeling.
- **How**:
    - The ‘los’ variable in the icustays table already represents the length of stay in days.
    - Convert it to hours (los_hours = los * 24) for consistency.
    - Retain records where `los_hours` is 24 or more.

2.5 **Add Weekend Admission Flag**

- **Objective**: Identify ICU admissions occurring on weekends to address Aim 2 of the project.
- **Why**:
    - Investigate whether weekend ICU admissions are associated with higher mortality rates.

    - Weekend admissions could differ in outcomes due to variations in staffing, resource availability, or other factors.
- **How**:
    - Add a new column `intime_weekdays` to display the day of the week (e.g., “Monday”, “Saturday”).
    - Use this column to create a boolean flag `is_weekend_admission` , which is set to `TRUE` for admissions occurring on “Saturday” or “Sunday”.
    - Save the updated dataset to include these new columns for downstream analysis.

2.6 **Save Intermediate Filtered Data**

- **Objective**: Save the filtered dataset for reproducibility and debugging purposes.
- **Why**: Provides a checkpoint to avoid repeating prior filtering steps if further processing needs adjustments.
- **How**:
    - Save the filtered dataset (`filtered_data`) as an RDS file using `saveRDS` .

2.7 **Merge Time-Series Data into** `pt_stay_hr`

- **Objective**: Combine hourly clinical measurements into the base time-series structure of `pt_stay_hr` .
- **Why**: Hourly data from `vitals_hourly` , `labs_hourly` , `gcs_hourly` , and `output_hourly` provide critical features for predictive modeling.
- **How**:
    - Sequentially left join `vitals_hourly` , `labs_hourly` , `gcs_hourly` , and `output_hourly` to `pt_stay_hr` using `icustay_id` and `hr` .

2.8 **Filter Time-Series Data to First 24 Hours**

- **Objective**: Retain only the data corresponding to the first 24 hours of ICU stay.
- **Why**: Aligns with the project requirement to use the first 24 hours of ICU data for prediction while not limiting outcomes to the same timeframe.
- **How**:
    - Filter records where the `hr` column is less than or equal to 24.

2.9 **Save Processed Time-Series Data**

- **Objective**: Save the merged and filtered time-series data for further analysis.
- **Why**: Provides a checkpoint for reproducibility and supports efficient debugging.
- **How**:
    - Save the processed time-series dataset as an RDS file.

2.10 **Merge Filtered Time-Series Data with** `filtered_data`

- **Objective**: Combine the filtered demographic and admission data with time-series data for the first 24 hours.
- **Why**: Integrates all relevant information into a single dataset for subsequent analysis and model building.
- **How**:
    - Left join the time-series data with `filtered_data` using `icustay_id` .

2.11 **Save Final Master Dataset**

- **Objective**: Save the fully preprocessed dataset for predictive modeling and hypothesis testing.
- **Why**: Ensures the final dataset is ready for downstream tasks and avoids repetition of preprocessing steps.
- **How**:
    - Save the final dataset (`master_data`) as an RDS file using `saveRDS` .

```
##     hadm_id         icustay_id       MDROs
##  Min.   :100001   Min.   :    -1   Mode :logical
##  1st Qu.:125063   1st Qu.:    -1   FALSE:67737
##  Median :149996   Median :228138   TRUE :318
##  Mean   :149982   Mean   :174167
##  3rd Qu.:174895   3rd Qu.:264167
##  Max.   :199999   Max.   :299998
```

```
## No duplicates exist for hadm_id.
```

```
## [1] "C"
```

```
## Final filtered data saved with 36522 rows and 40 columns.
```

```
## Filtered Time-Series data saved with 568037 rows and 44 columns.
```

```
## Updated Filtering and Merging Steps Completed.
```

```
## Filtered Dataset Rows: 36522
```

```
## Master Dataset Rows: 375552
```

```
## Master Dataset Columns: 83
```

Section 3: Exploratory Data Analysis (EDA)

Objectives:

- Understand the structure and relationships in the filtered data.
- Identify trends, distributions, and potential outliers.
- Evaluate key predictors and their correlations with the target variable (mortality).

Steps:

3.1 Basic Descriptive Statistics:

- **Objective**: Summarize the dataset to understand its structure and identify potential issues.
- **Why**:
  - Ensure numerical and categorical variables are within expected ranges.
  - Identify missing values that may need handling during modeling.
- **How**:
  - Calculate summary statistics for numerical variables (mean, median, standard deviation, min, max).
  - Tabulate categorical variables (frequency and proportions).
  - Summarize missing values for all variables to identify those requiring imputation or exclusion.
  - Stratify statistics by mortality (`EXPIRE_FLAG`) to detect differences between survivors and non-survivors.

3.2 Target Variable Analysis:

- **Objective**: Understand the distribution of the target variable and its relationship with key features.
- **Why**:
  - Explore the prevalence of mortality (`EXPIRE_FLAG`).
  - Analyze survival times for additional insights.
- **How**:
  - Visualize the distribution of mortality (`EXPIRE_FLAG`) as proportions or counts.
  - Use histograms and bar plots to compare mortality trends across age groups, gender, and ICU types.
  - Explore survival times using Kaplan-Meier curves or other survival analysis techniques.

3.3 Key Predictor Exploration:

- **Objective**: Investigate the distribution and predictive power of key clinical variables.
- **Why**:
  - Determine whether predictors show significant differences across mortality outcomes.
  - Identify potential predictive patterns or outliers in vital signs and lab results.
- **How**:
  - Use boxplots and density plots to visualize distributions of vital signs and lab values.
  - Focus on predictors with lower percentages of missing values to ensure robust analysis.
  - Stratify by `EXPIRE_FLAG` to compare trends between survivors and non-survivors.

3.4 Correlation Analysis:

- **Objective**:
  - Focus on numerical variables in `master_data` .
  - Address potential multicollinearity by identifying highly correlated variables (> 0.8 or < -0.8).
- **Why**:
  - Identify groups of correlated variables to avoid redundancy in modeling.
  - Highlight potential key predictors.
- **How**:
  - Compute a correlation matrix for numerical variables using complete cases.
  - Visualize correlations using a heatmap with hierarchical clustering to reveal relationships.

3.5 Demographics and ICU Characteristics:

- **Objective**: Explore the relationships between patient demographics, ICU characteristics, and mortality outcomes.
- **Why**:
  - Assess the impact of variables such as age, gender, and ICU type on mortality.
  - Investigate potential differences in outcomes between weekend and weekday admissions.
- **How**:
  - Analyze mortality rates across demographic groups (age, gender, ethnicity).
  - Visualize age distribution and compare across survival groups.
  - Visualize the distribution of ICU types (`first_careunit`) and their association with mortality.
  - Assess the impact of weekend (`is_weekend_admission`) vs. weekday admissions.
  - Perform t-tests for continuous variables (e.g., age).
  - Use chi-squared tests for categorical variables (e.g., gender, ICU types).

Insights of step 3.1 results

```
##
## ### Summary of Table: final_filtered_data ###
## Number of rows: 36522
## Number of columns: 40
##
## Column names and data types:
##                          Data_Type Missing_Count Missing_Pct.
## hadm_id                    integer             0         0.00
## icustay_id                 integer             0         0.00
## subject_id                 integer             0         0.00
## row_id                     integer             0         0.00
## dob.x                      POSIXct             0         0.00
## admittime.x                POSIXct          5860        16.05
## dischtime.x                POSIXct          5860        16.05
## age_years                  numeric             0         0.00
## intime                     POSIXct             0         0.00
## outtime                    POSIXct             2         0.01
## los                        numeric             2         0.01
## hosp_deathtime             POSIXct         35249        96.51
## icu_expire_flag            integer             0         0.00
## hospital_expire_flag.x     integer          5860        16.05
## dod                        POSIXct         22343        61.18
## expire_flag.x              integer             0         0.00
## ttd_days                   integer         22343        61.18
## first_careunit           character             0         0.00
## last_careunit            character             0         0.00
## first_wardid               integer             0         0.00
## last_wardid                integer             0         0.00
## insurance                character             0         0.00
## language                 character             0         0.00
## religion                 character             0         0.00
## marital_status           character             0         0.00
## ethnicity                character             0         0.00
## admission_type           character             0         0.00
## admission_location       character             0         0.00
## hospital_expire_flag.y     integer             0         0.00
## admittime.y                POSIXct             0         0.00
## dischtime.y                POSIXct             0         0.00
## deathtime                  POSIXct         32570        89.18
## icd9_code                character             0         0.00
## intime_weekdays          character             0         0.00
## is_weekend_admission       logical             0         0.00
## gender                   character             0         0.00
## dob.y                      POSIXct             0         0.00
## expire_flag.y              integer             0         0.00
## avg_weight_naive           numeric          2312         6.33
## MDROs                      logical          6859        18.78
##
## ---
```

```
##
## ### Summary of Table: master_data ###
## Number of rows: 375552
## Number of columns: 83
##
## Column names and data types:
##                         Data_Type Missing_Count Missing_Pct.
## icustay_id                integer             0         0.00
## hadm_id.x                 integer             0         0.00
## subject_id.x              integer             0         0.00
## row_id                    integer             0         0.00
## dob.x                     POSIXct             0         0.00
## admittime.x               POSIXct         57789        15.39
## dischtime.x               POSIXct         57789        15.39
## age_years                 numeric             0         0.00
## intime.x                  POSIXct             0         0.00
## outtime.x                 POSIXct             2         0.00
## los                       numeric             2         0.00
## hosp_deathtime            POSIXct        362480        96.52
## icu_expire_flag           integer             0         0.00
## hospital_expire_flag.x    integer         57789        15.39
## dod                       POSIXct        231193        61.56
## expire_flag.x             integer             0         0.00
## ttd_days                  integer        231193        61.56
## first_careunit          character             0         0.00
## last_careunit           character             0         0.00
## first_wardid              integer             0         0.00
## last_wardid               integer             0         0.00
## insurance               character             0         0.00
## language                character             0         0.00
## religion                character             0         0.00
## marital_status          character             0         0.00
## ethnicity               character             0         0.00
## admission_type          character             0         0.00
## admission_location      character             0         0.00
## hospital_expire_flag.y    integer             0         0.00
## admittime.y               POSIXct             0         0.00
## dischtime.y               POSIXct             0         0.00
## deathtime                 POSIXct        336408        89.58
## icd9_code               character             0         0.00
## intime_weekdays         character             0         0.00
## is_weekend_admission      logical             0         0.00
## gender                  character             0         0.00
## dob.y                     POSIXct             0         0.00
## expire_flag.y             integer             0         0.00
## avg_weight_naive          numeric         21751         5.79
## MDROs                     logical         68089        18.13
## hr                        integer         21246         5.66
## hadm_id.y                 integer         21246         5.66
## subject_id.y              integer         21246         5.66
## intime.y                  POSIXct         21246         5.66
## outtime.y                 POSIXct         21246         5.66
## starttime                 POSIXct         21246         5.66
## endtime                   POSIXct         21246         5.66
## dy                        integer         21582         5.75
```

```
## spo2                     numeric       53170        14.16
## fio2                     numeric      356903        95.03
## temperature              numeric      256856        68.39
## resprate                 numeric       60883        16.21
## heartrate                numeric       55848        14.87
## sysbp                    numeric       65010        17.31
## diasbp                   numeric       65063        17.32
## glucose.x                numeric      286346        76.25
## meanarterialpressure     numeric       64372        17.14
## neutrophil               numeric      372230        99.12
## creactiveprotein         numeric      375428        99.97
## whitebloodcell           numeric      351211        93.52
## partialpressureo2        numeric      338361        90.10
## bicarbonate              numeric      349620        93.09
## lactate                  numeric      360710        96.05
## troponin                 numeric      370982        98.78
## bloodureanitrogen        numeric      349457        93.05
## creatinine               numeric      349354        93.02
## alaninetransaminase      numeric      369411        98.36
## aspartatetransaminase    numeric      369409        98.36
## hemoglobin               numeric      345076        91.89
## intnormalisedratio       numeric      355970        94.79
## platelets                numeric      349277        93.00
## albumin                  numeric      371438        98.90
## chloride                 numeric      346295        92.21
## glucose.y                numeric      333377        88.77
## sodium                   numeric      349594        93.09
## bilirubin                numeric      369437        98.37
## hematocrit               numeric      341138        90.84
## gcs                      integer      260984        69.49
## gcseyes                  integer      261112        69.53
## gcsmotor                 integer      261276        69.57
## gcsverbal                integer      261288        69.57
## endotrachflag            integer      260984        69.49
## urineoutput              numeric      151427        40.32
##
## ---
```

Summary of final_filtered_data

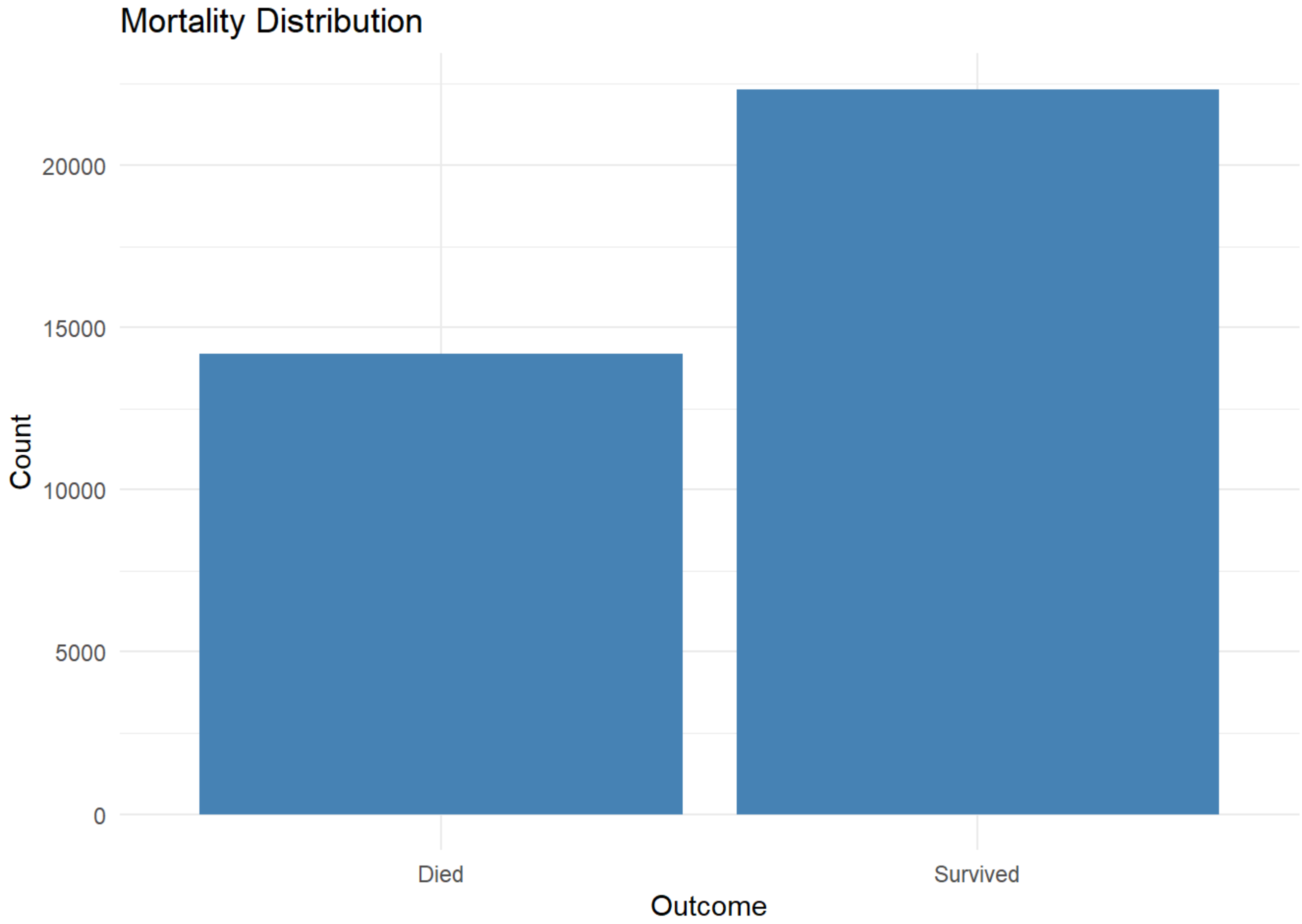
1. **Overall Dataset Shape**:
    - **Rows**: 36,522
    - **Columns**: 40
2. **Key Observations**:
    - Variables like `dod` and `ttd_days` have a significant percentage of missing values (61.18%).
    - `deathtime` and `hosp_deathtime` has 89.18% anf 96.51%missing values, indicating most records lack time-of-death information.
    - Most categorical fields have no missing values.
    - All numeric values (`los` , `age` , `icu_los_hours`) are complete and ready for analysis.

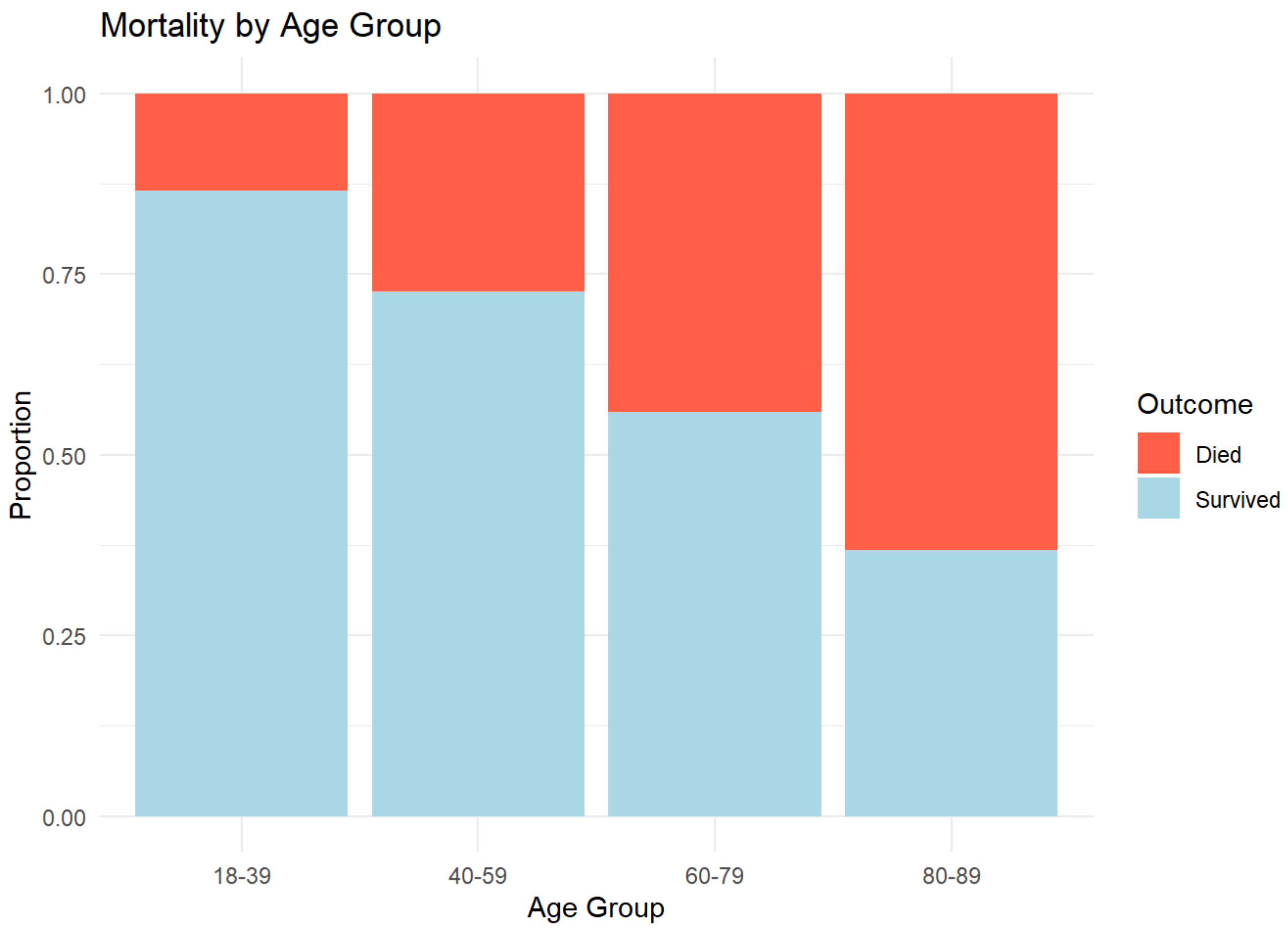
Summary of `master_data`

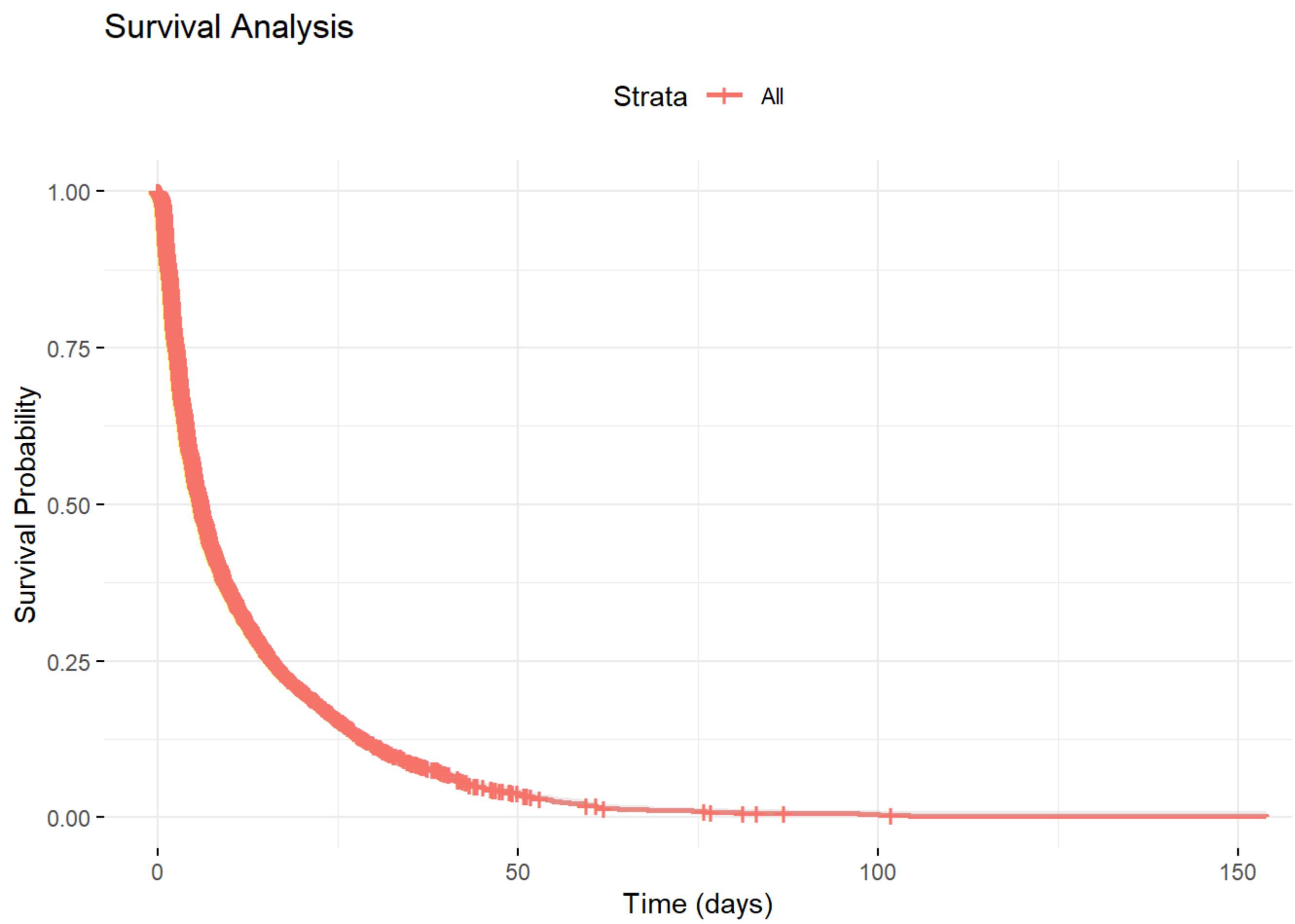
1. **Overall Dataset Shape**:
    - **Rows**: 375,552
    - **Columns**: 83
2. **Key Observations**:

- High missingness in many clinical variables (`creactiveprotein` , `alaninetransaminase` , etc.), with some exceeding 90%.
- Critical time-series variables (`spo2` , `temperature` , etc.) also show significant missingness, requiring imputation or exclusion strategies.
- Demographic and admission-related variables (`gender` , `ethnicity` , `admittime`) are fully populated, which is good for initial analysis.

Observation and insights of step 3.2 results







Mortality Distribution

- **Observation**:
  - A larger proportion of ICU patients survived, as shown by the taller bar labeled “Survived.”
  - A smaller proportion of the patients did not survive (“Died”).
- **Insight**:
  - The dataset is imbalanced, with a majority of the patients surviving. This imbalance could influence predictive modeling, requiring techniques like balancing the dataset or using metrics robust to class imbalance (e.g., F1 score, AUC).

Mortality by Age Group

- **Observation**:
  - The mortality rate increases with age.
  - In the age group 18–39, the proportion of patients who died is minimal compared to those who survived.
  - In the 80–89 age group, a significant proportion of patients did not survive, nearly matching or exceeding the survivors.
- **Insight**:
  - Age is a critical factor influencing ICU outcomes, with older patients at a much higher risk of mortality.
  - Predictive models should incorporate age as a key feature, potentially treating it as a non-linear variable to capture this trend.

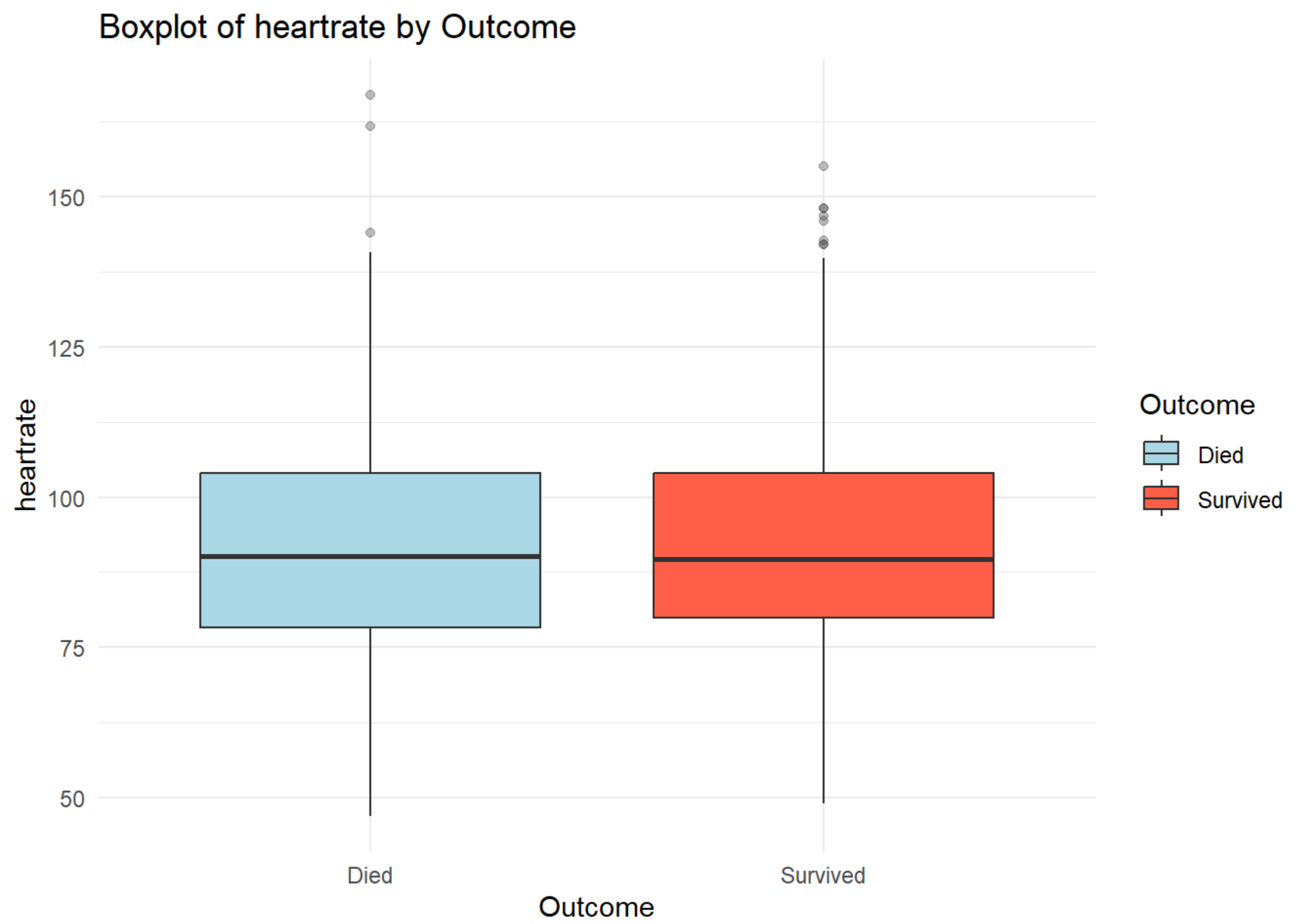
Survival Analysis

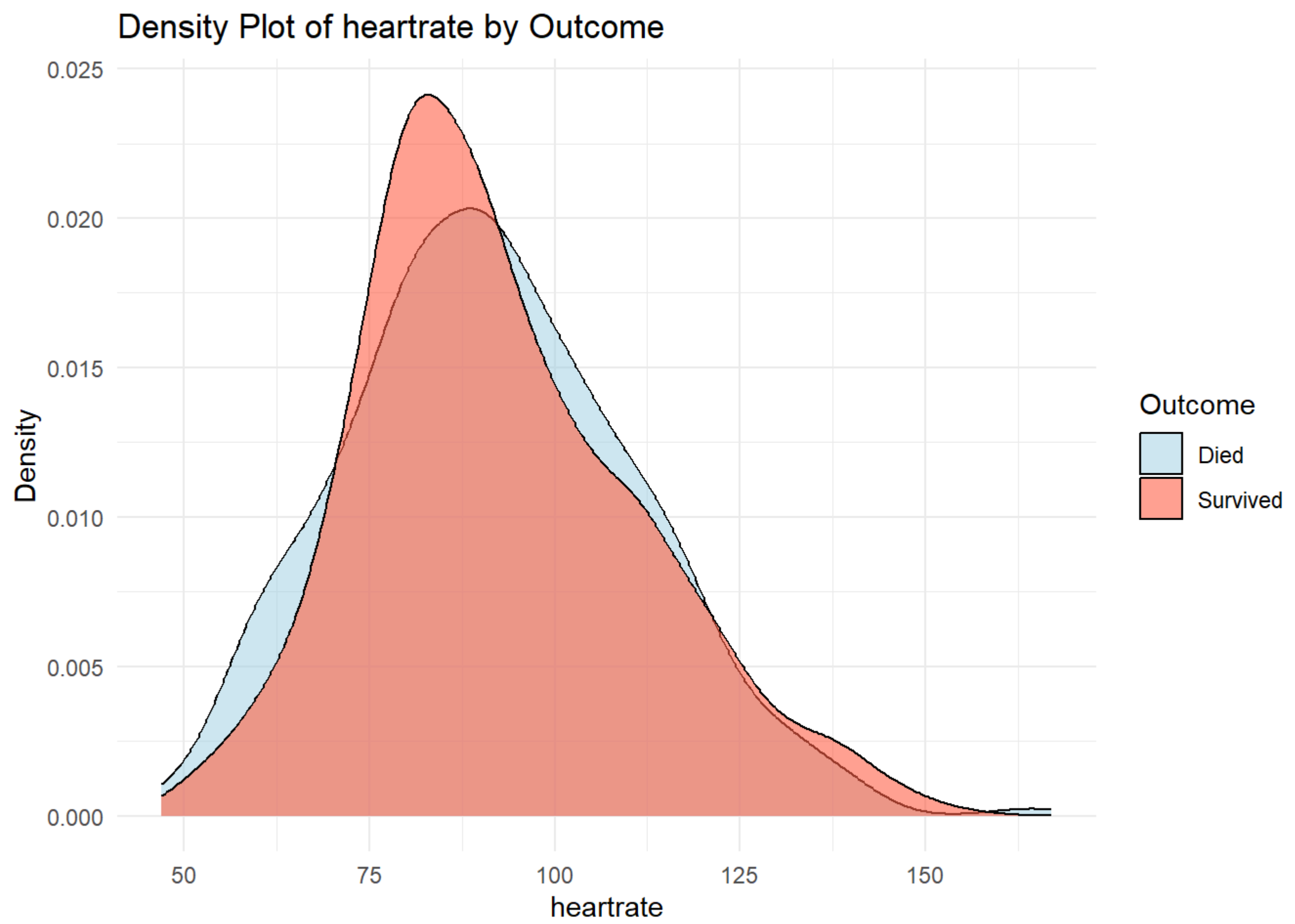
- **Observation**:
  - The survival probability drops steeply during the initial days of ICU stay and gradually levels off as time progresses.
  - The steep decline indicates that the first few days in the ICU are critical for patient survival.

- **Insight**:
    - This suggests that immediate and intensive care during the initial period is crucial for improving survival rates.
    - The leveling off of survival probability after a certain point may indicate a higher likelihood of recovery or stabilization for longer-staying patients.
    - Survival analysis supports the hypothesis that time-dependent features and early intervention are vital for predicting mortality.

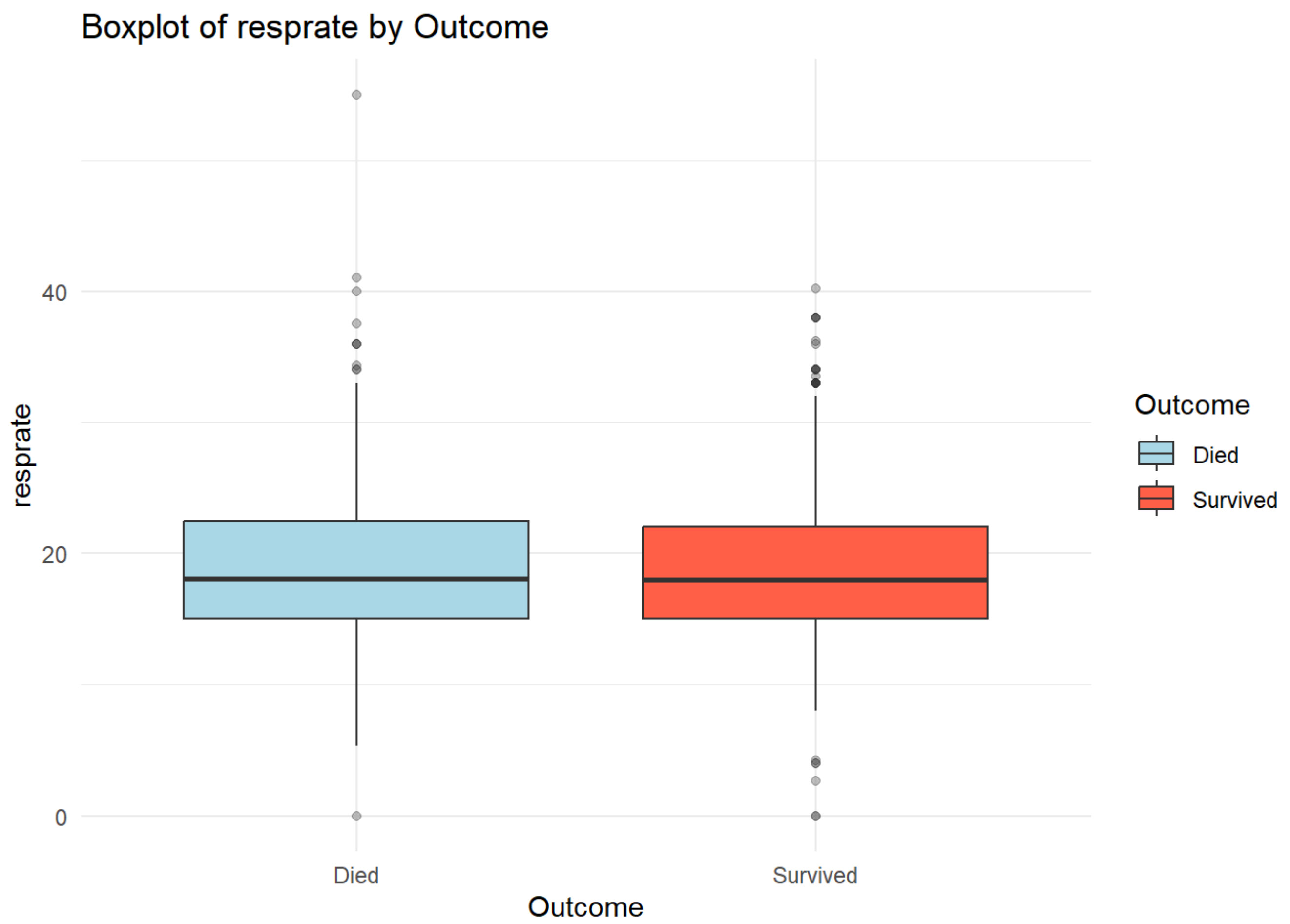
Insights of step 3.3 results

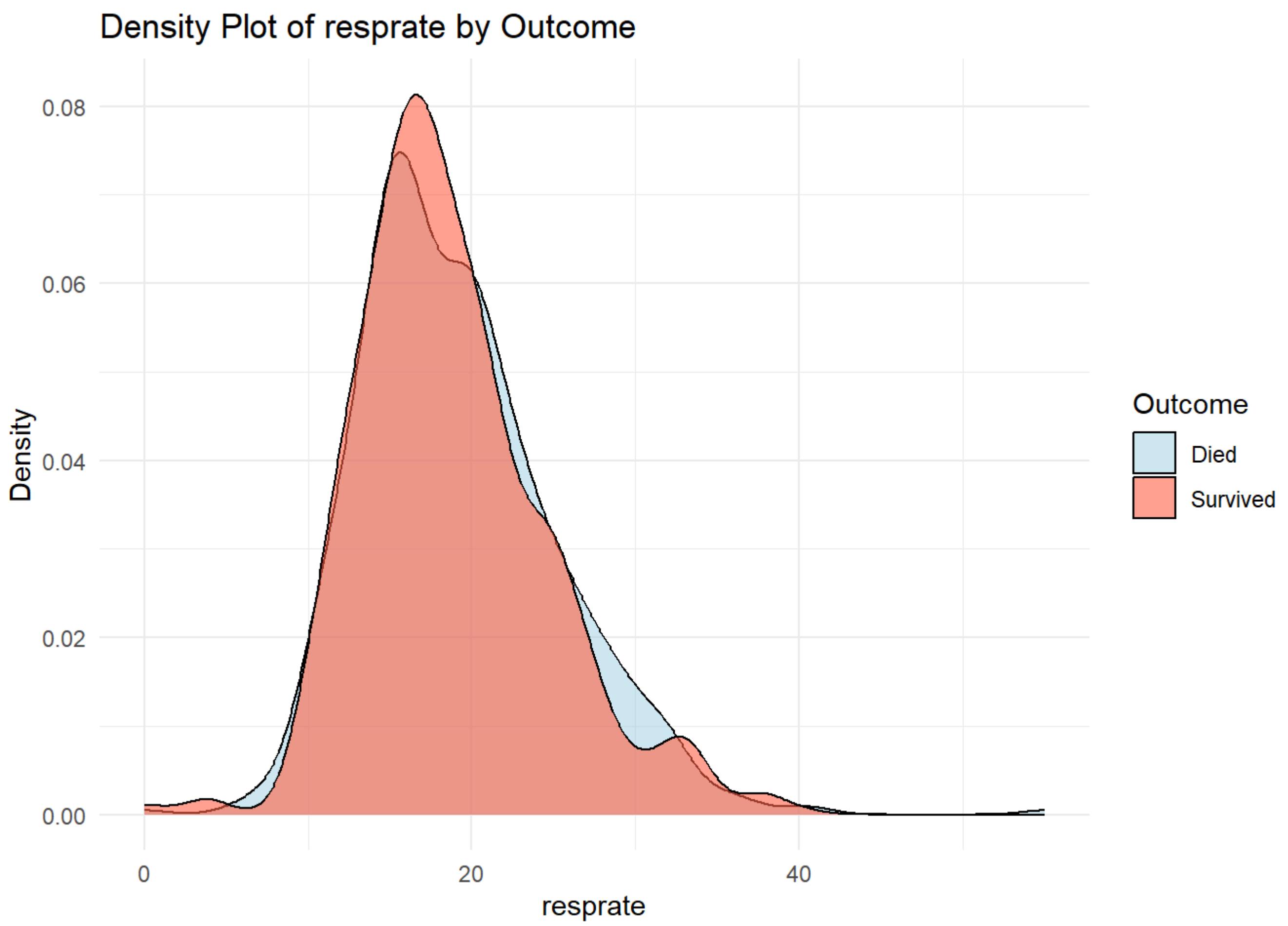
```
##
## --- Visualizations for heartrate ---
```



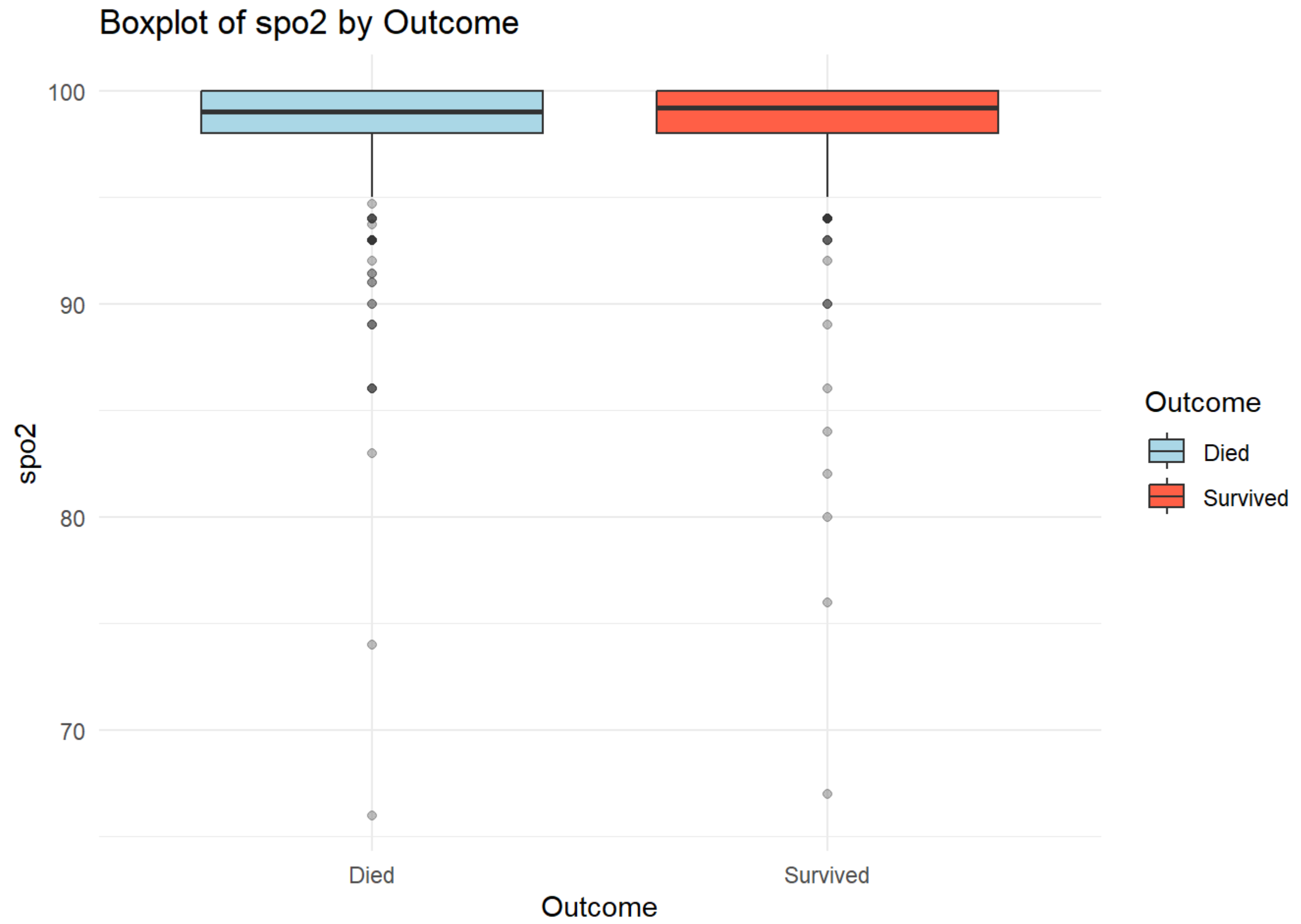


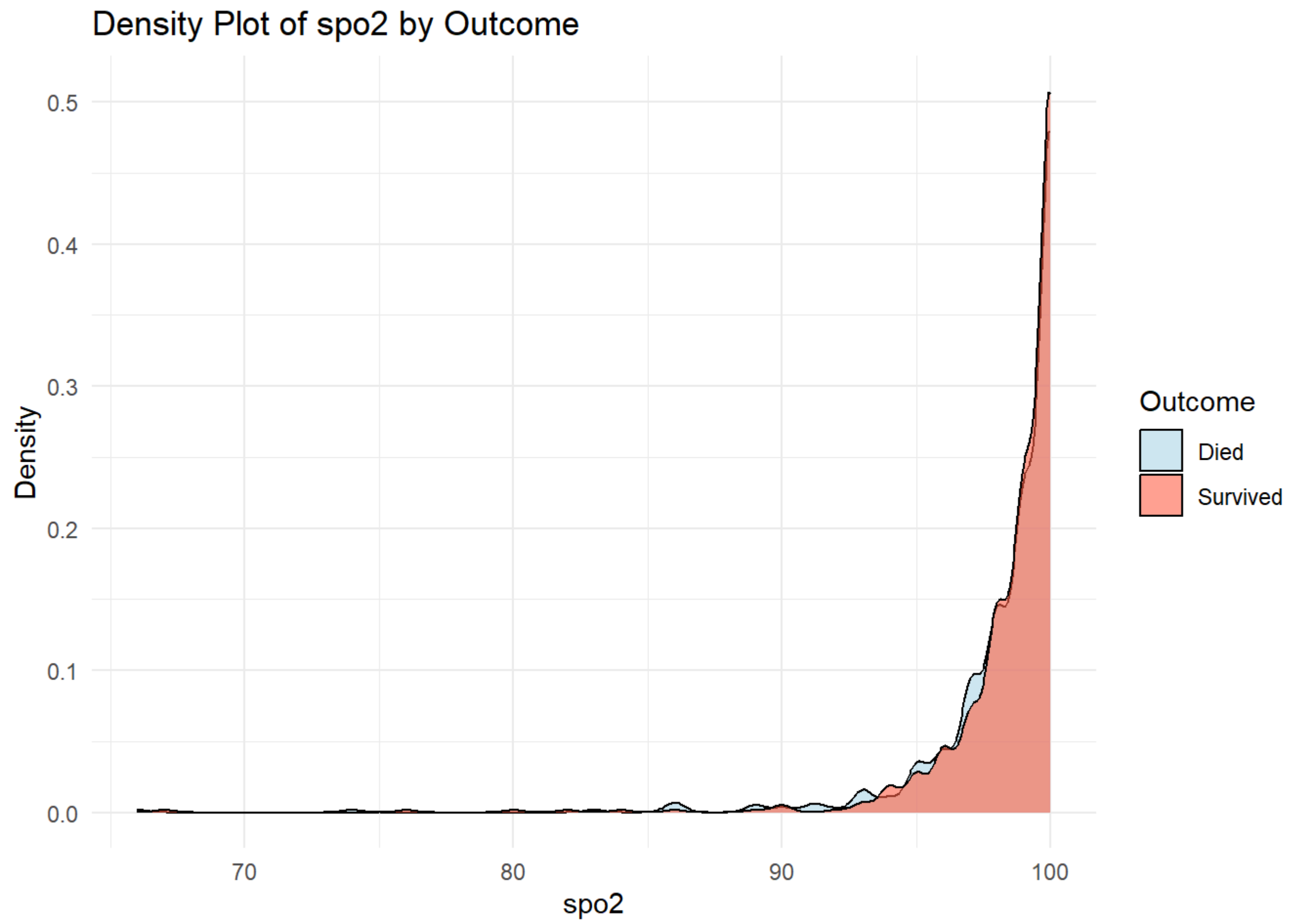

```
##
## --- Visualizations for resprate ---
```



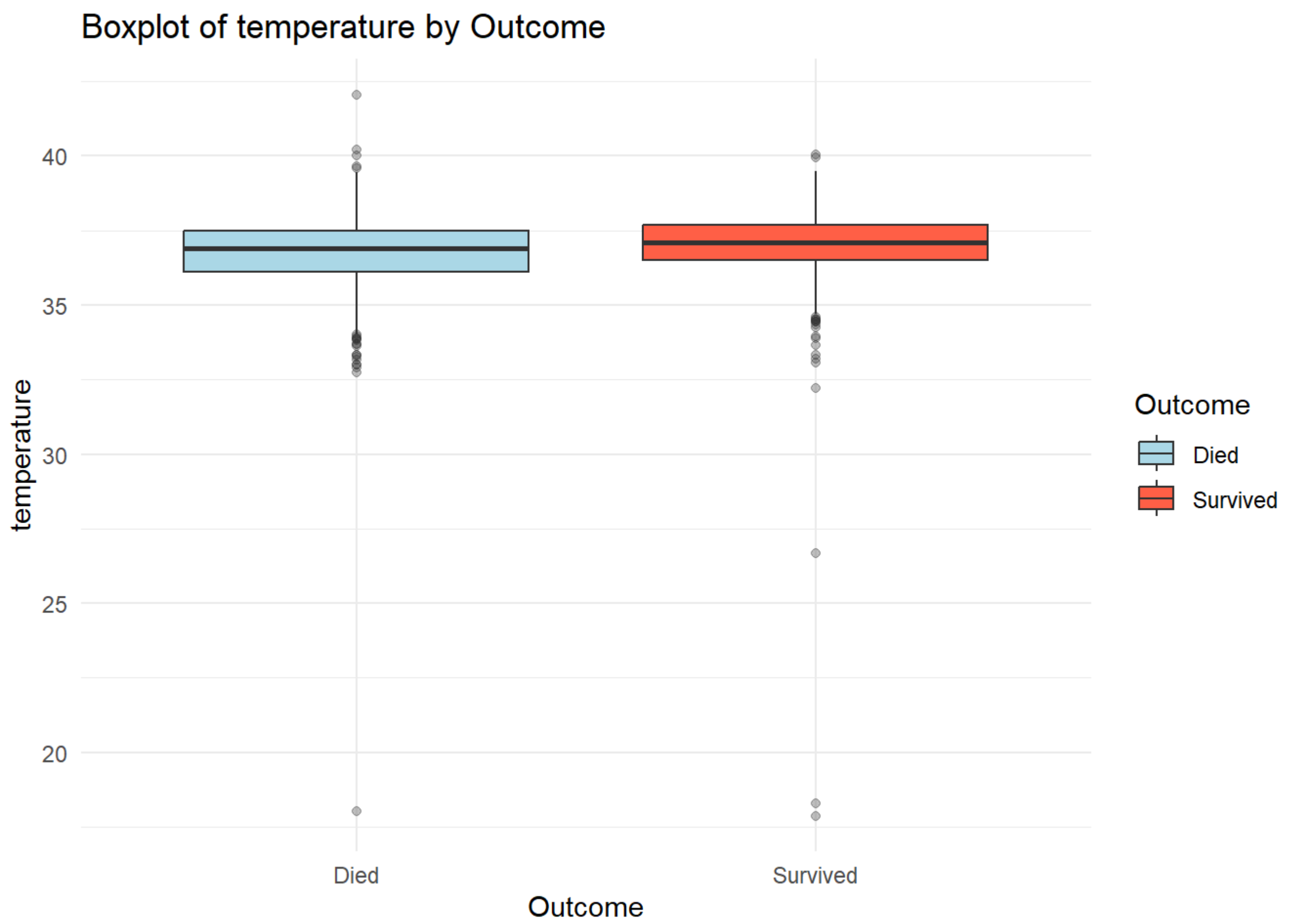


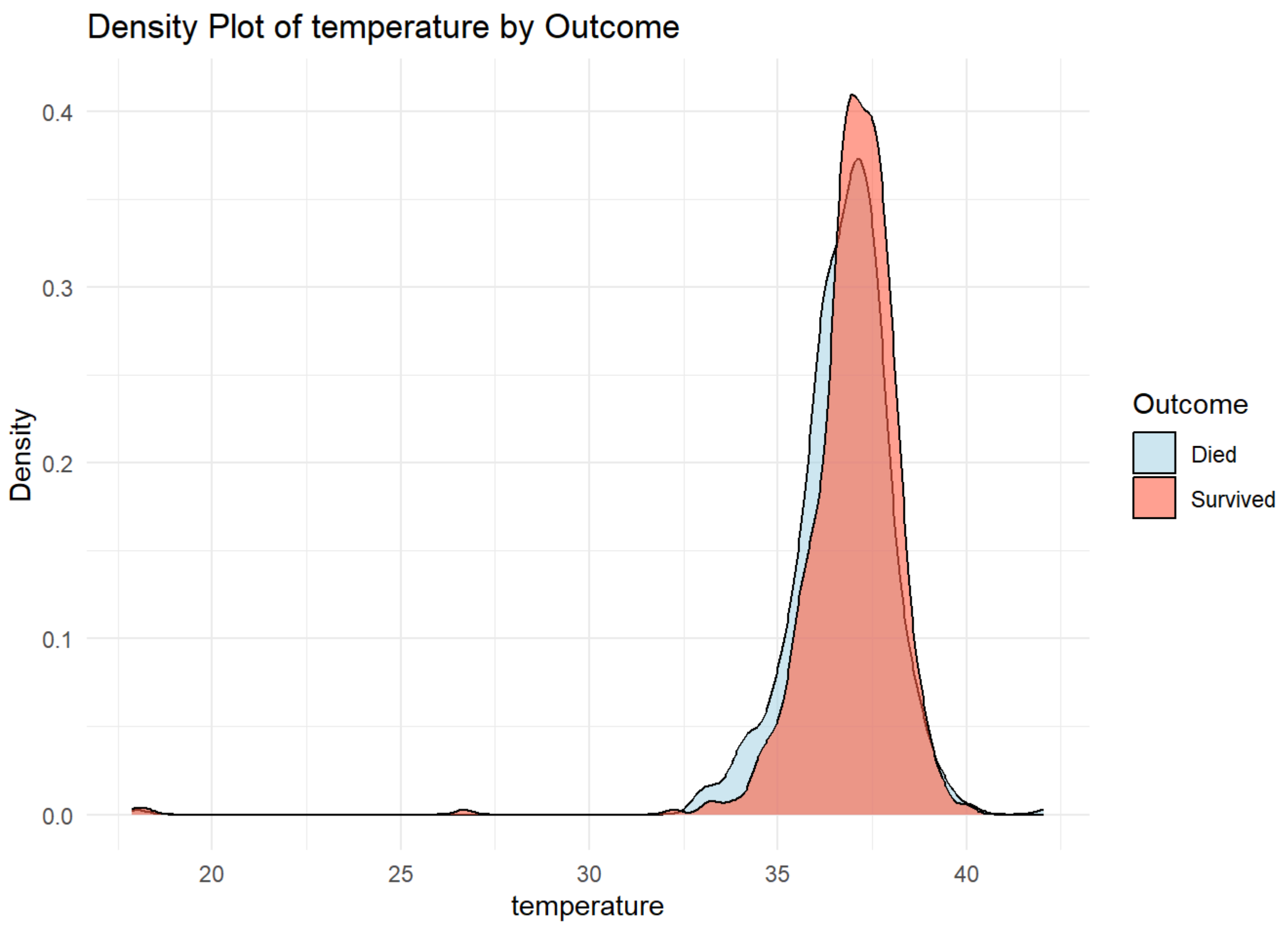

```
##
## --- Visualizations for spo2 ---
```



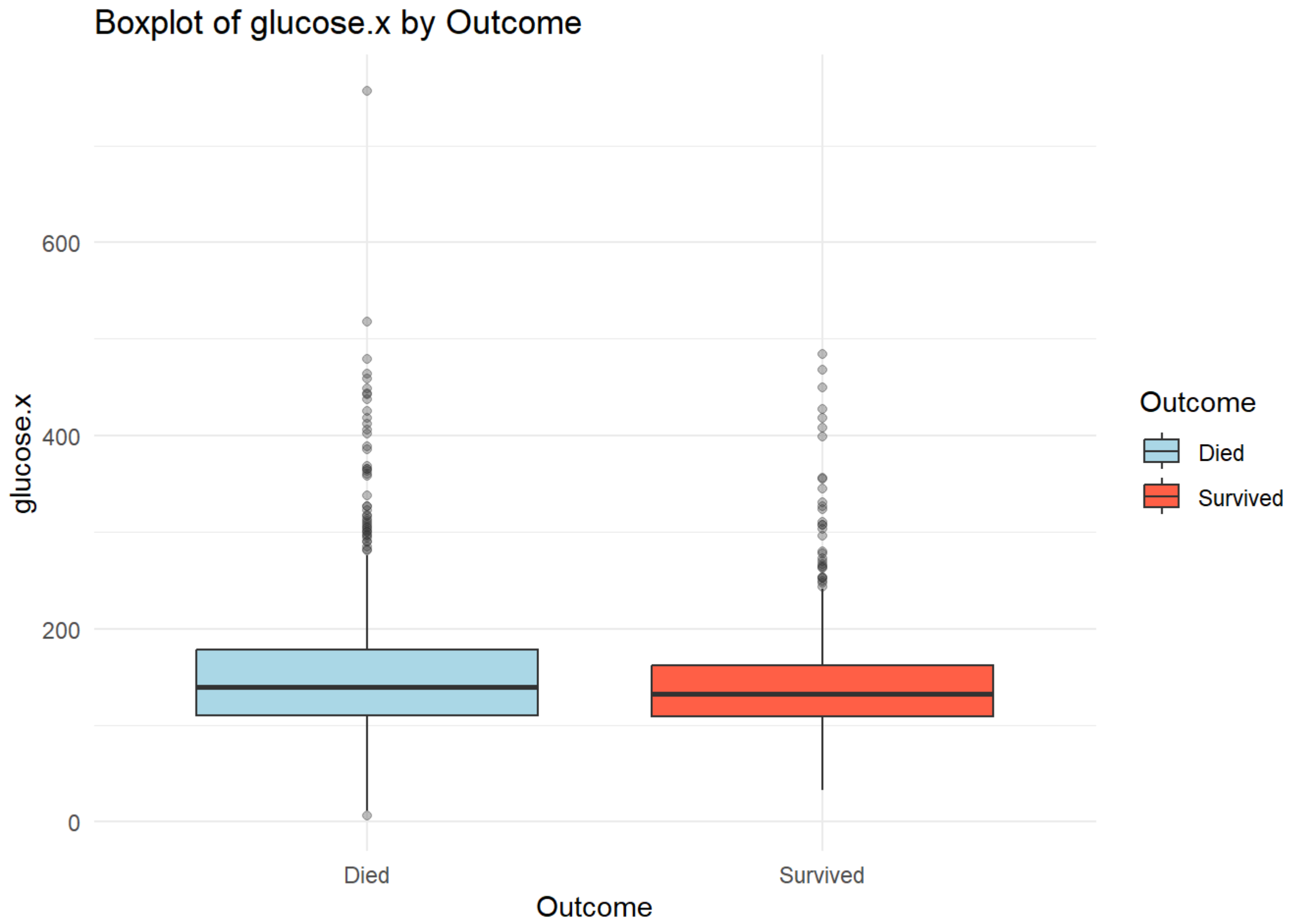


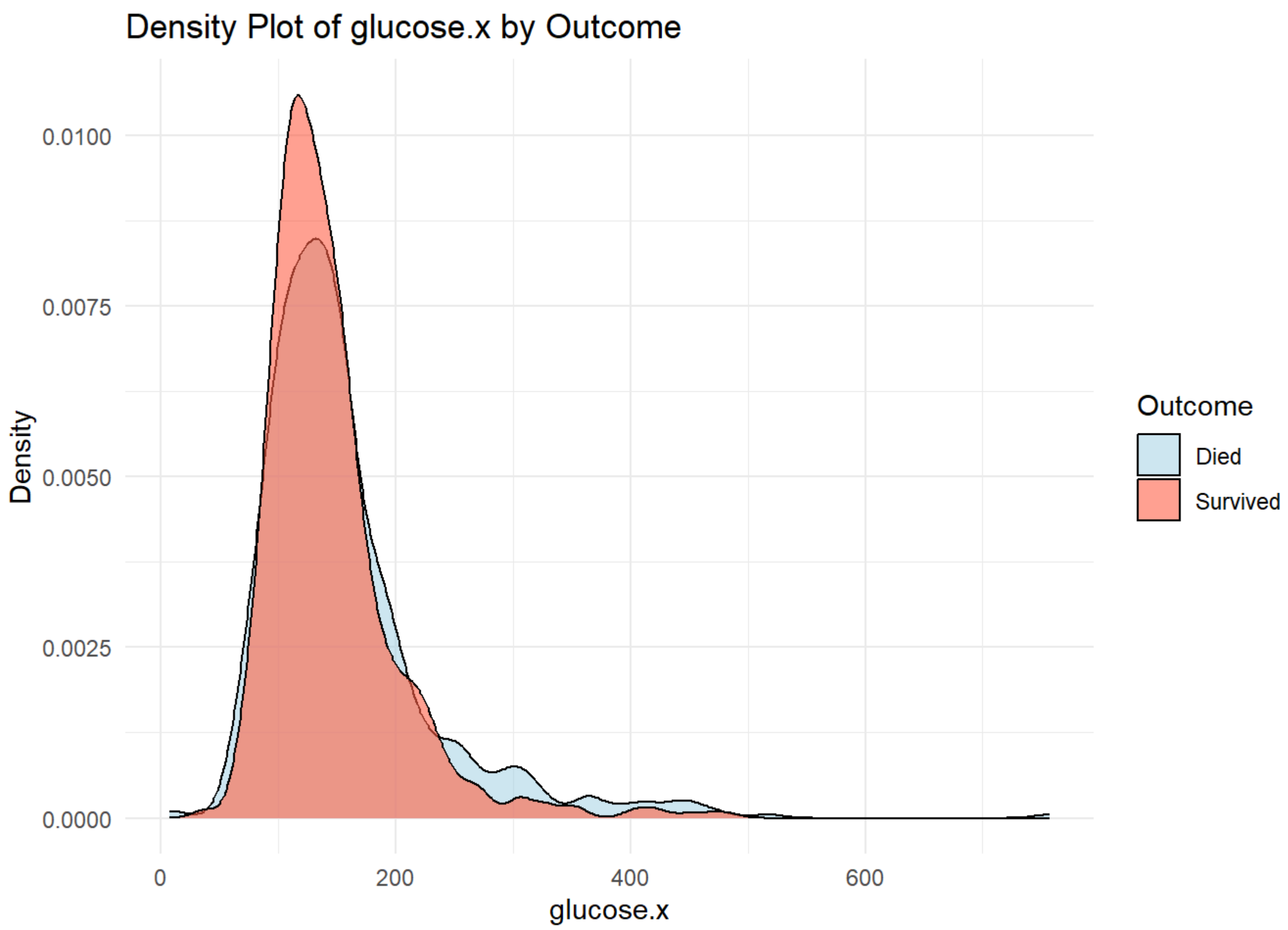

```
##
## --- Visualizations for temperature ---
```



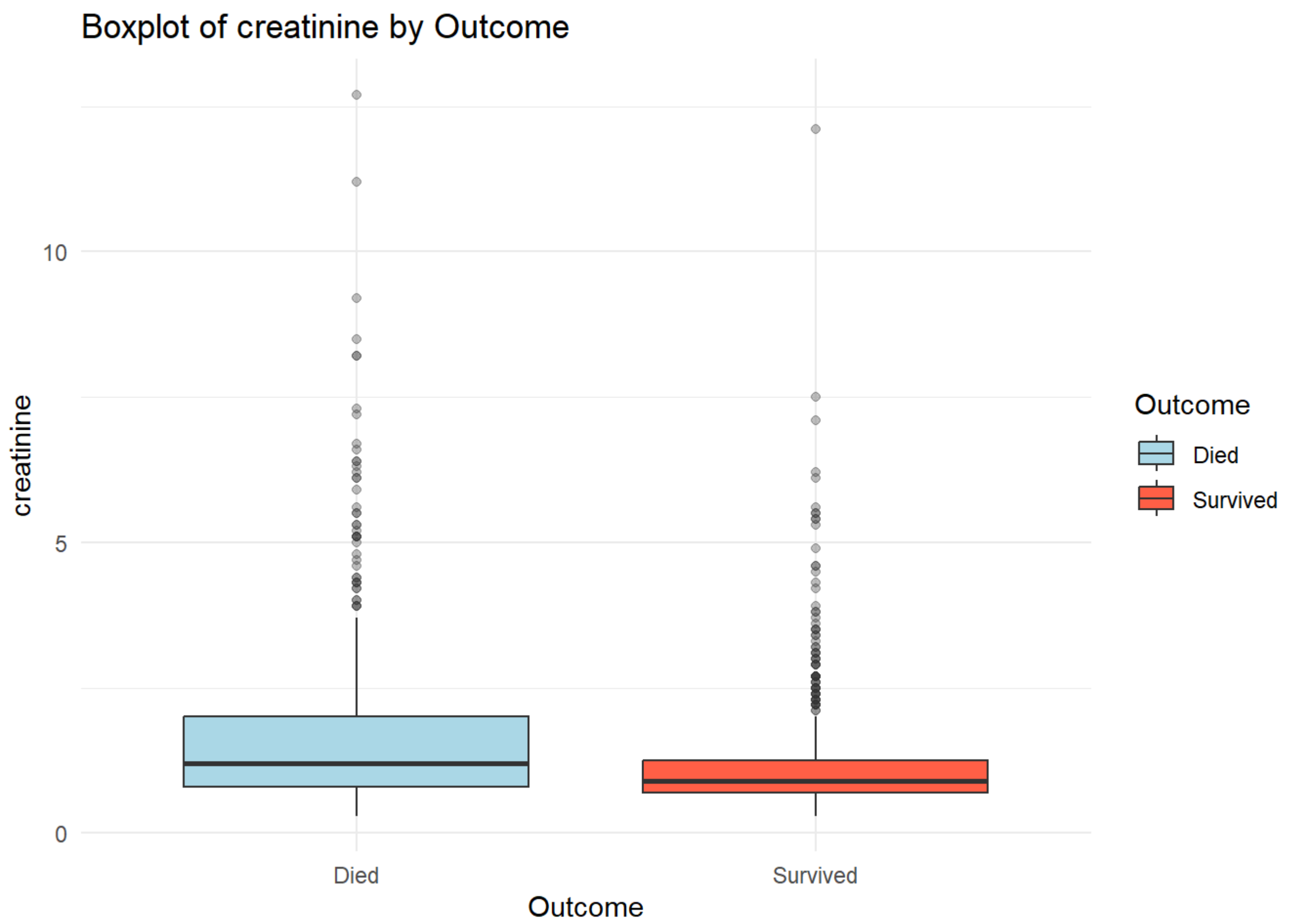


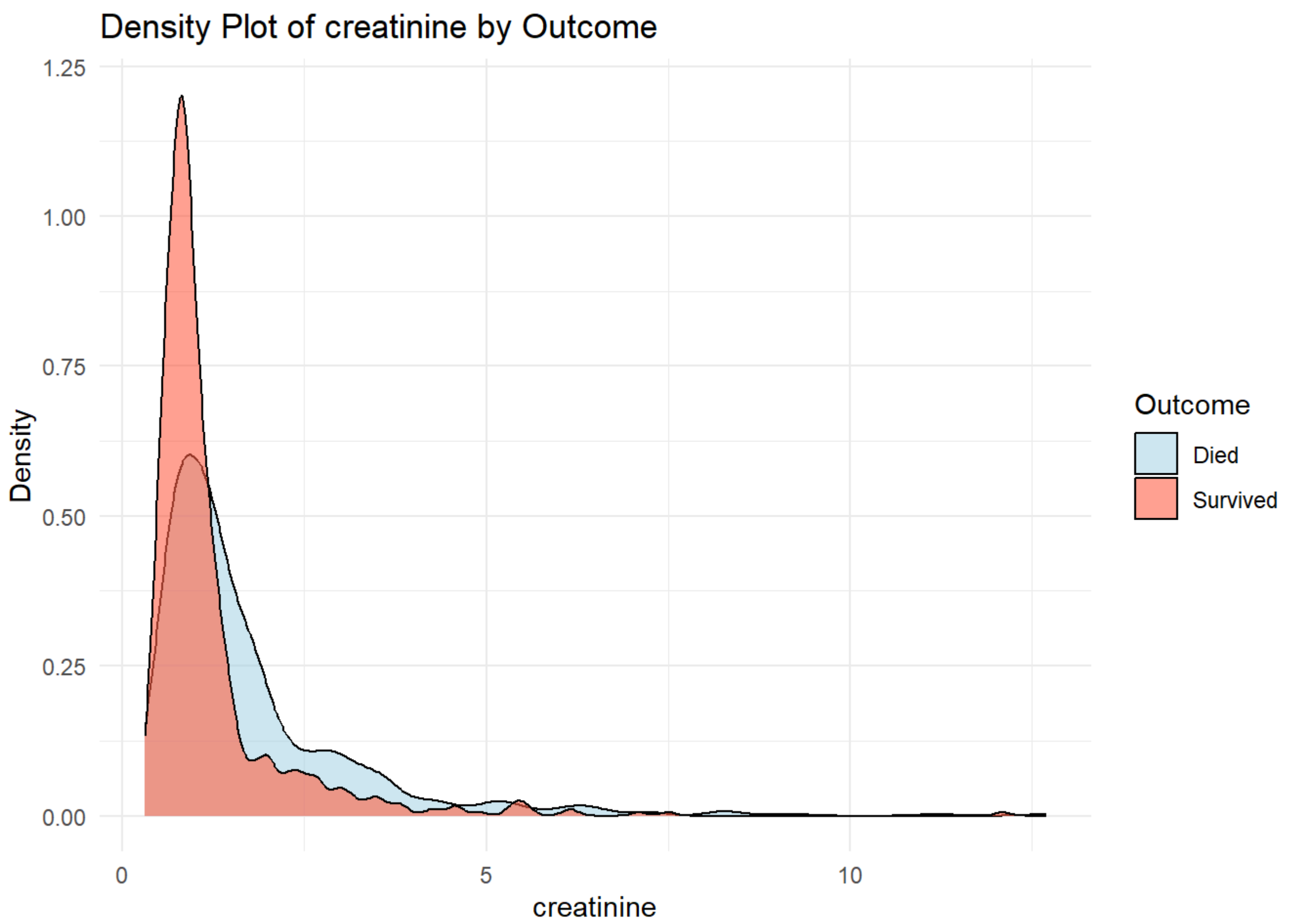

```
##
## --- Visualizations for glucose.x ---
```



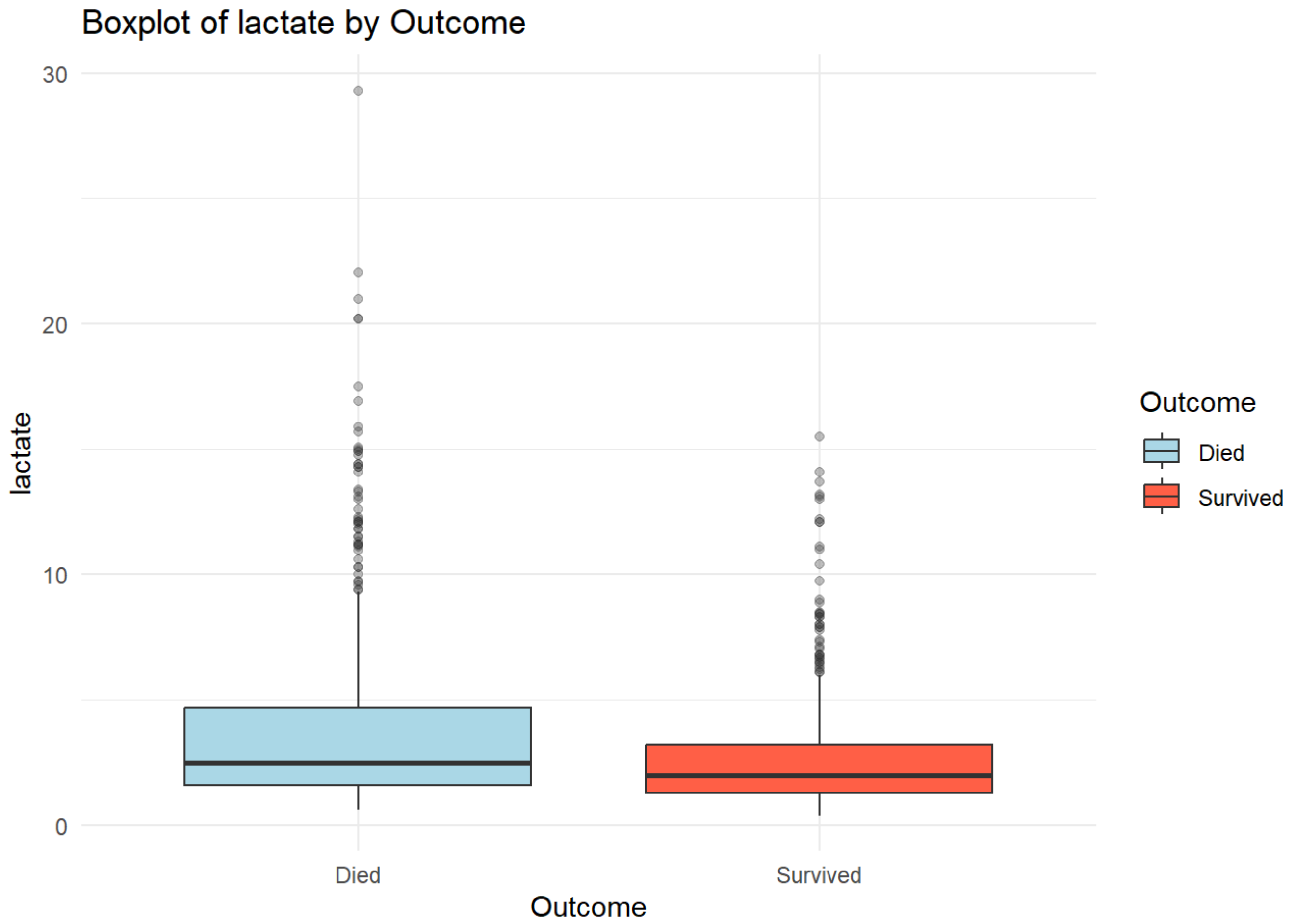


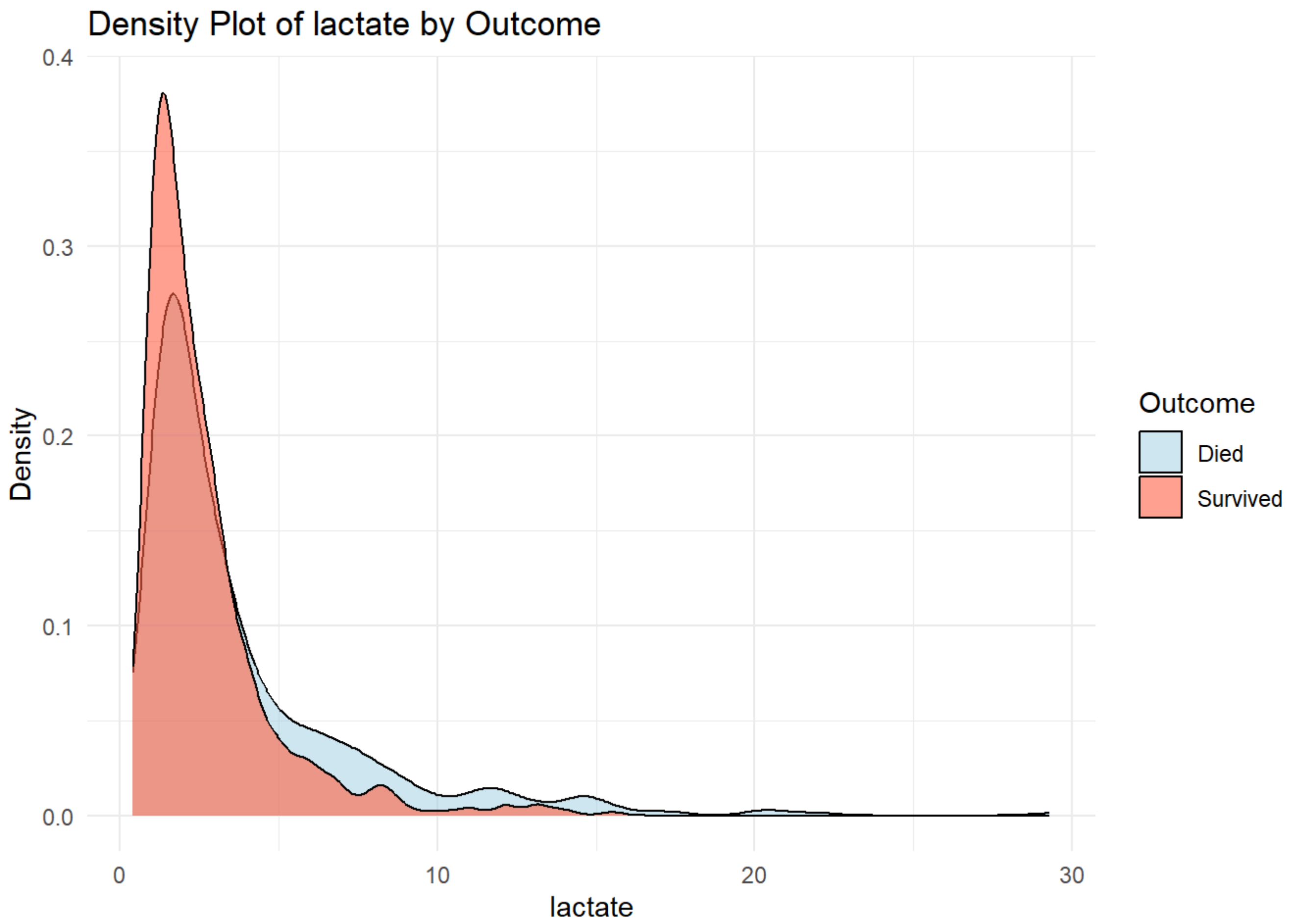

```
##
## --- Visualizations for creatinine ---
```



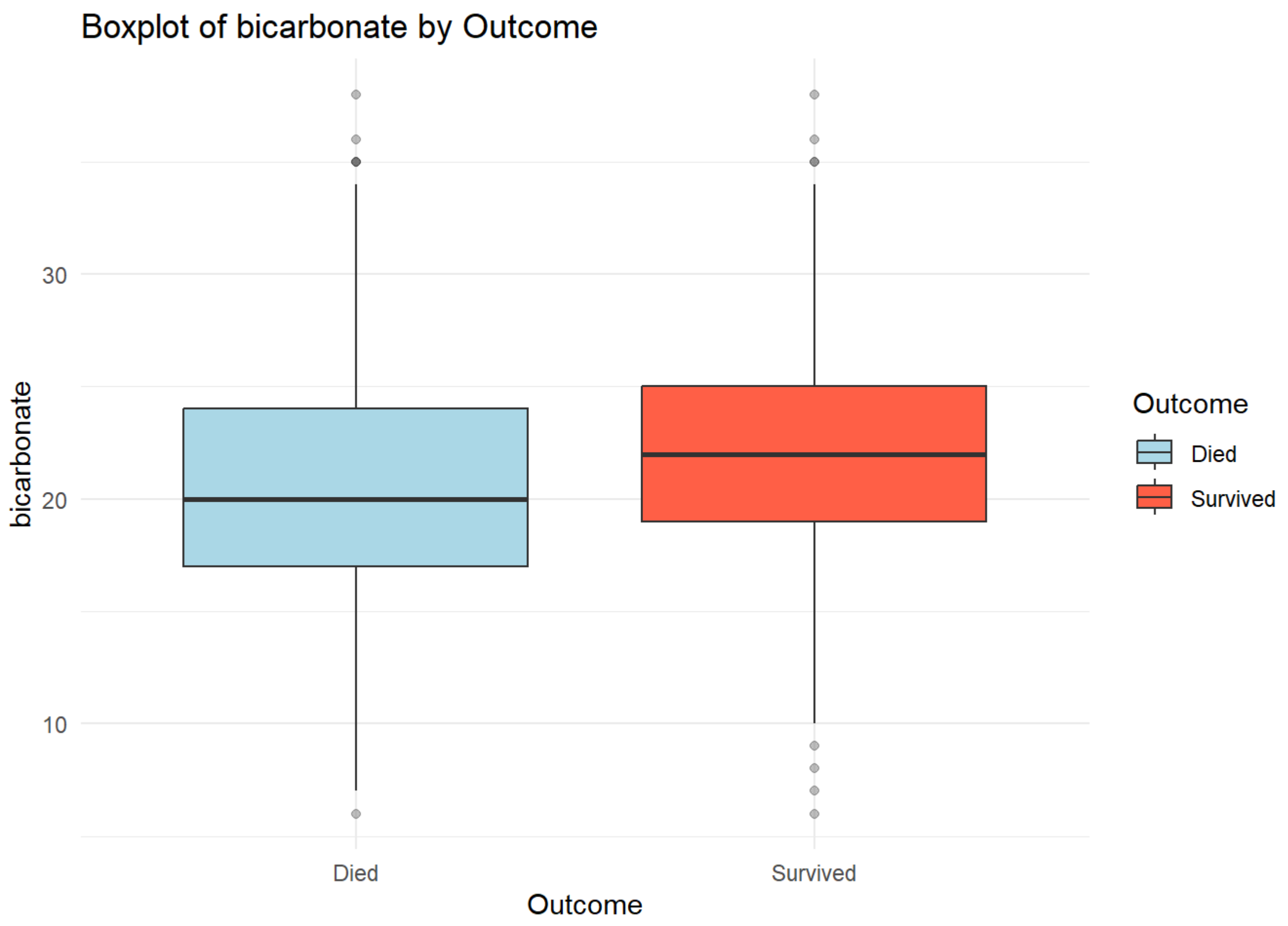



```
##
## --- Visualizations for lactate ---
```

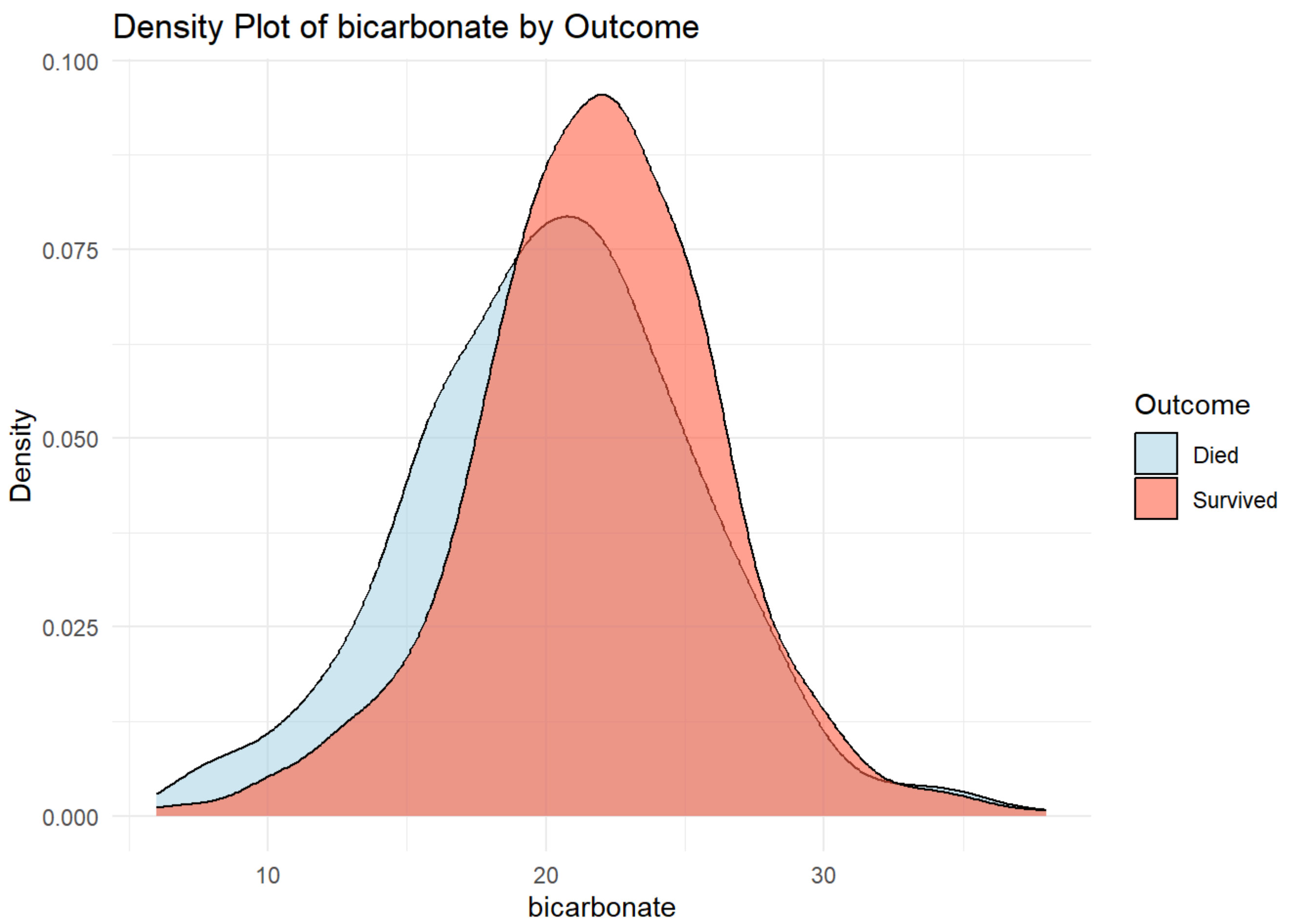





```
##
## --- Visualizations for bicarbonate ---
```



Boxplot of bicarbonate by Outcome
bicarbonate
30
20
10
Died
Survived
Outcome
Outcome
Died
Survived



Density Plot of bicarbonate by Outcome
0.100
0.075
0.050
0.025
0.000
Density
10
20
30
bicarbonate
Outcome
Died
Survived

```
##                       Variable       P_Value Mean_Survived  Mean_Died
## mean in group 01    resprate  0.000000e+00     18.387521  19.325925
## mean in group 05  creatinine 1.166878e-134      1.182105   1.660277
## mean in group 04   glucose.x 8.100044e-126    139.132221 149.799808
## mean in group 06     lactate 1.881095e-114      2.477553   3.483804
## mean in group 0    heartrate  9.081887e-83     84.640268  85.901786
## mean in group 07 bicarbonate  1.564196e-27     23.507164  22.833644
## mean in group 03 temperature  8.359552e-14     37.038217  36.903344
## mean in group 02        spo2  3.004859e-01    102.239292  97.155565
```

```
##                Variable       P_Value
## 2       first_careunit  0.000000e+00
## 3      intime_weekdays 1.183170e-142
## 1               gender  3.479749e-83
## 4 is_weekend_admission  1.573501e-69
```

Statistical Test Results:

- Variables like `resprate` , `creatinine` , `glucose.x` , and `lactate` show strong statistical significance (p-values close to 0), indicating a clear difference between survivors and non-survivors.
- Less significant variables such as `spo2` (p-value ~0.30) may not contribute significantly to outcome prediction.

Categorical Variable Analysis:

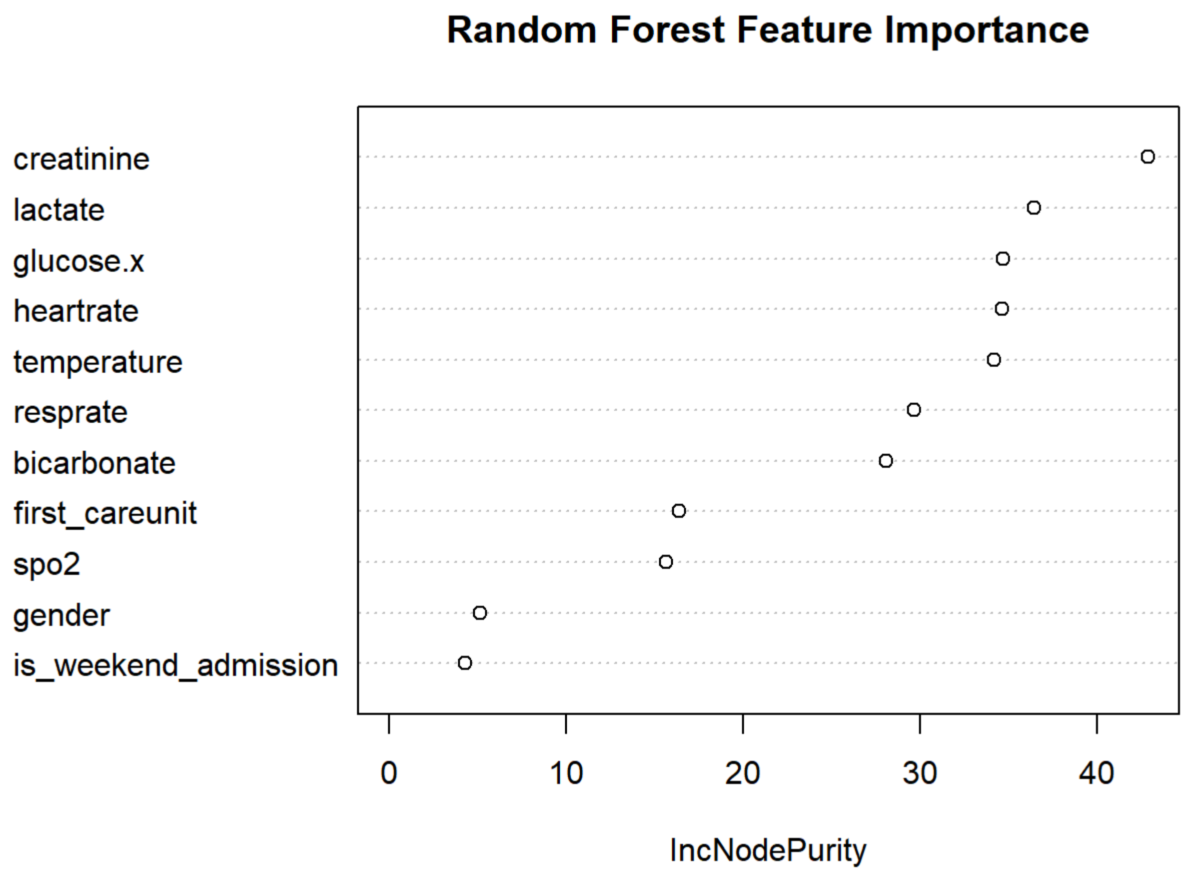
- Categorical predictors like `first_careunit` and `intime_weekdays` exhibit highly significant associations with mortality (p-values ~0), suggesting these are strong predictors.
- The variable `is_weekend_admission` shows weaker significance but is still worth considering due to its contextual relevance.

```
##
## Call:
## glm(formula = expire_flag.x ~ heartrate + resprate + spo2 + temperature +
##     glucose.x + creatinine + lactate + bicarbonate + gender +
##     first_careunit + is_weekend_admission, family = "binomial",
##     data = master_data)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               2.774997   3.001960   0.924  0.35528
## heartrate                -0.007743   0.003437  -2.253  0.02427 *
## resprate                 -0.013222   0.011494  -1.150  0.24998
## spo2                     -0.013253   0.023265  -0.570  0.56891
## temperature              -0.024385   0.043306  -0.563  0.57338
## glucose.x                -0.000257   0.001016  -0.253  0.80027
## creatinine                0.303653   0.062077   4.892 1.00e-06 ***
## lactate                   0.152457   0.028754   5.302 1.15e-07 ***
## bicarbonate              -0.009335   0.014587  -0.640  0.52220
## genderM                  -0.130711   0.124169  -1.053  0.29249
## first_careunitCSRU       -0.664066   0.257079  -2.583  0.00979 **
## first_careunitMICU       -0.022545   0.245116  -0.092  0.92672
## first_careunitSICU        0.164567   0.256174   0.642  0.52061
## first_careunitTSICU      -0.519692   0.254471  -2.042  0.04113 *
## is_weekend_admissionTRUE -0.018291   0.146443  -0.125  0.90060
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1758.7  on 1268  degrees of freedom
## Residual deviance: 1617.6  on 1254  degrees of freedom
##   (因为不存在，374283个观察量被删除了)
## AIC: 1647.6
##
## Number of Fisher Scoring iterations: 4
```

Logistic Regression:

- Predictors like `creatinine` , `lactate` , and certain ICU units (`first_careunitCSRU` and `first_careunitTSICU`) are statistically significant with strong effects.
- Variables such as `gender` and `is_weekend_admission` are not significant, indicating limited predictive value for mortality.

```
##                      IncNodePurity
## heartrate                34.650075
## resprate                 29.705450
## spo2                     15.685075
## temperature              34.217176
## glucose.x                34.698600
## creatinine               42.932079
## lactate                  36.439048
## bicarbonate              28.124834
## gender                    5.158534
## first_careunit           16.383881
## is_weekend_admission      4.293866
```

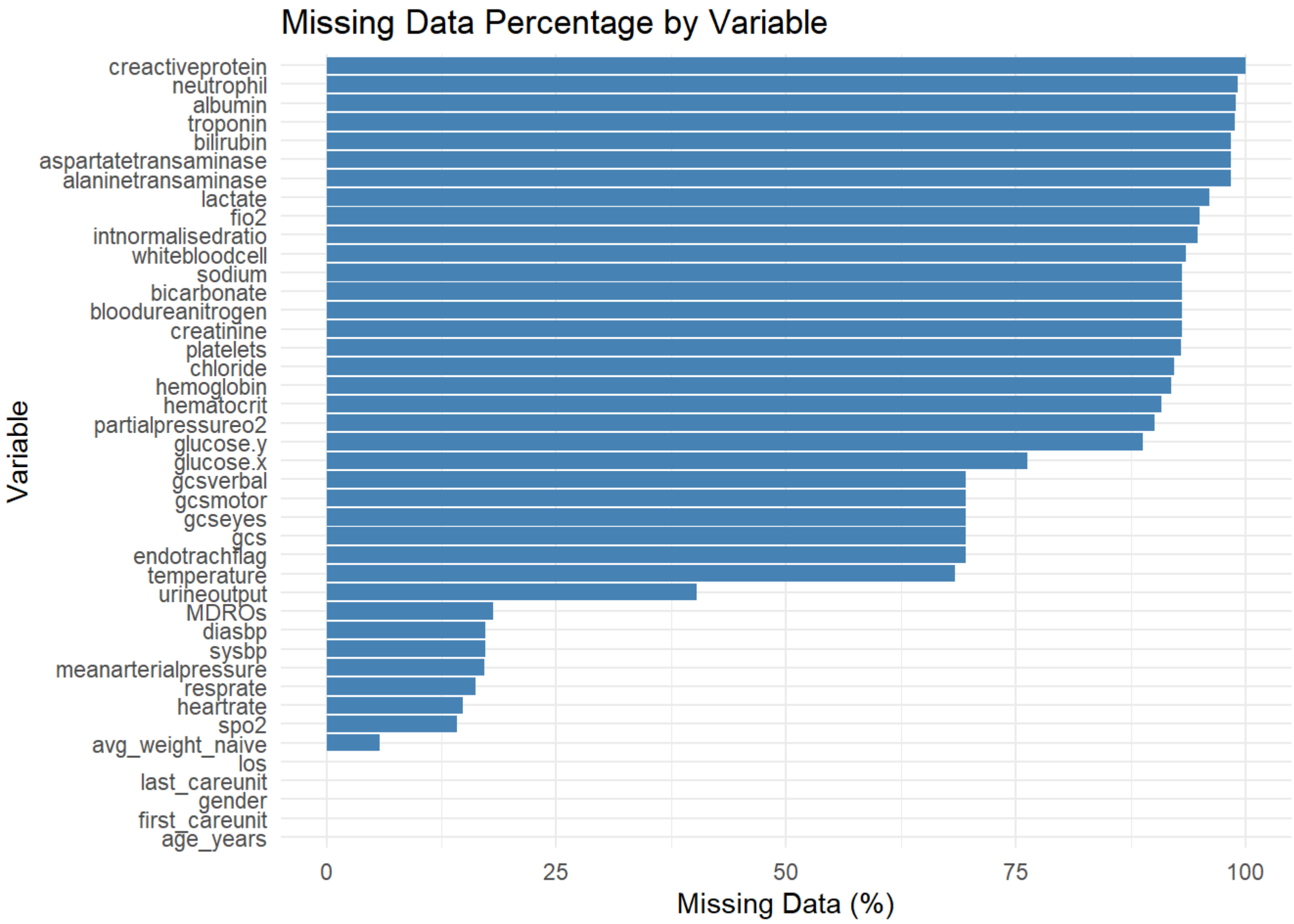


Random Forest Feature Importance: - Top predictors include `creatinine`, `lactate`, `glucose.x`, and `temperature`, aligning with both t-test and logistic regression results. - Categorical variables like `first_careunit` also play a significant role, reaffirming their importance.

Observation and insights of step 3.4 results

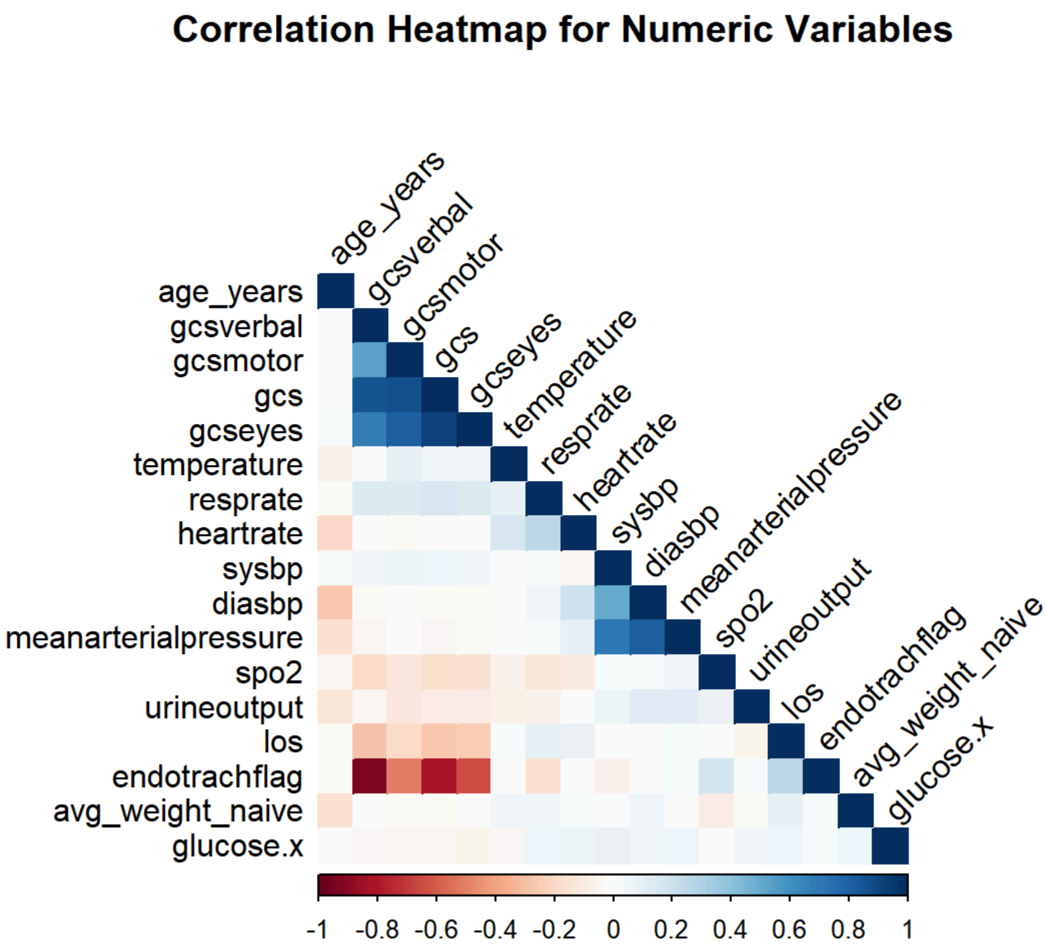
```
## ### Missing Data Summary for Clinically Important Variables ###
```

```
##                                      Variable Missing_Count Missing_Pct
## los                                       los             2        0.00
## age_years                           age_years             0        0.00
## avg_weight_naive             avg_weight_naive         21751        5.79
## spo2                                     spo2         53170       14.16
## fio2                                     fio2        356903       95.03
## temperature                       temperature        256856       68.39
## resprate                             resprate         60883       16.21
## heartrate                           heartrate         55848       14.87
## sysbp                                   sysbp         65010       17.31
## diasbp                                 diasbp         65063       17.32
## glucose.x                           glucose.x        286346       76.25
## meanarterialpressure   meanarterialpressure         64372       17.14
## neutrophil                         neutrophil        372230       99.12
## creactiveprotein             creactiveprotein        375428       99.97
## whitebloodcell                 whitebloodcell        351211       93.52
## partialpressureo2           partialpressureo2        338361       90.10
## bicarbonate                       bicarbonate        349620       93.09
## lactate                               lactate        360710       96.05
## troponin                             troponin        370982       98.78
## bloodureanitrogen           bloodureanitrogen        349457       93.05
## creatinine                         creatinine        349354       93.02
## alaninetransaminase       alaninetransaminase        369411       98.36
## aspartatetransaminase aspartatetransaminase        369409       98.36
## hemoglobin                         hemoglobin        345076       91.89
## intnormalisedratio         intnormalisedratio        355970       94.79
## platelets                           platelets        349277       93.00
## albumin                               albumin        371438       98.90
## chloride                             chloride        346295       92.21
## glucose.y                           glucose.y        333377       88.77
## sodium                                 sodium        349594       93.09
## bilirubin                           bilirubin        369437       98.37
## hematocrit                         hematocrit        341138       90.84
## urineoutput                       urineoutput        151427       40.32
## gcs                                       gcs        260984       69.49
## gcseyes                               gcseyes        261112       69.53
## gcsmotor                             gcsmotor        261276       69.57
## gcsverbal                           gcsverbal        261288       69.57
## MDROs                                   MDROs         68089       18.13
## endotrachflag                   endotrachflag        260984       69.49
## first_careunit                 first_careunit             0        0.00
## last_careunit                   last_careunit             0        0.00
## gender                                 gender             0        0.00
```



Missing Data Analysis

- Missing data percentages are clearly calculated and visualized.
- Variables with significant missingness, such as `creactiveprotein` (99.97%) and `neutrophil` (99.12%), highlight potential candidates for exclusion or imputation.



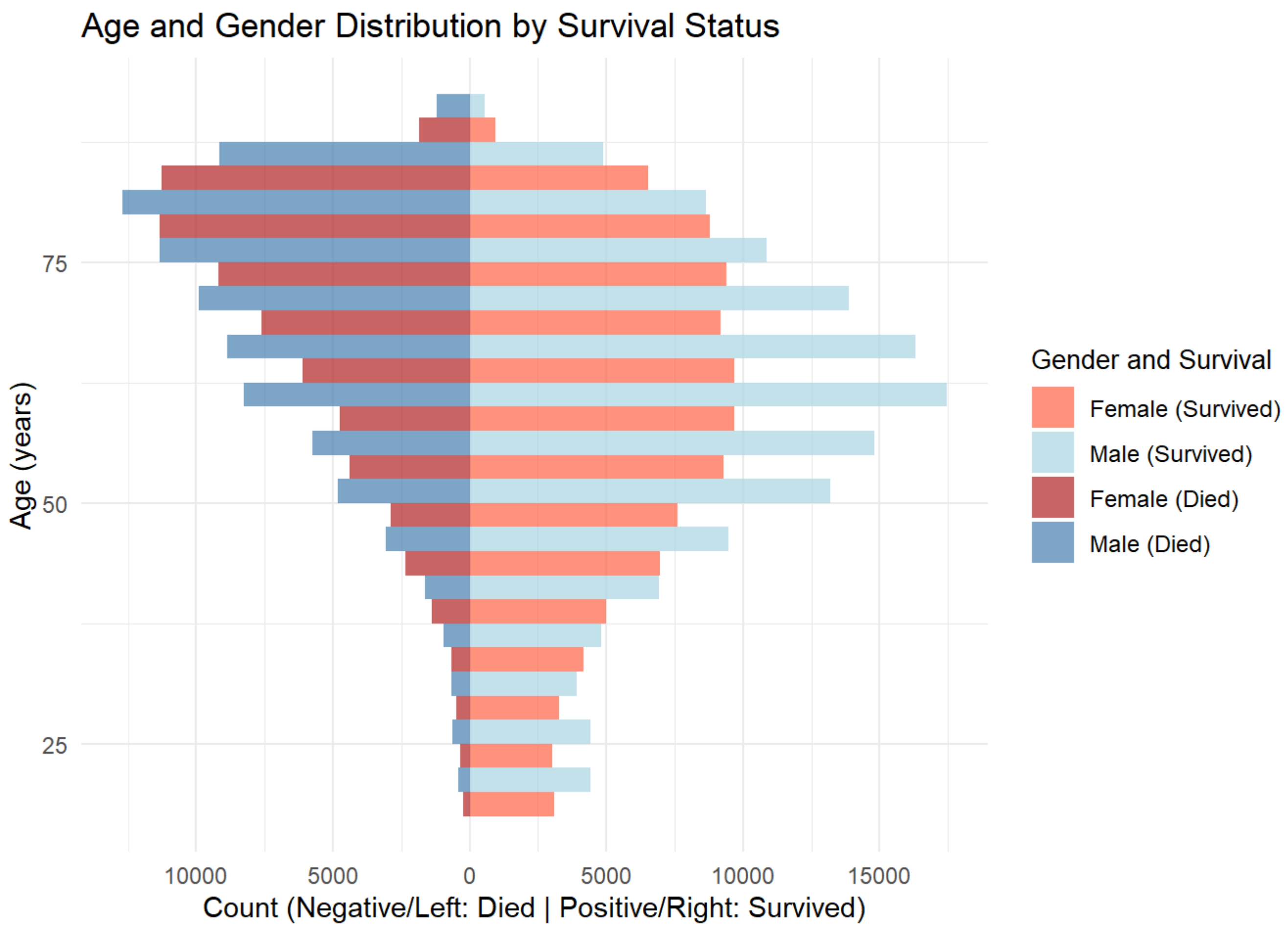
Correlation Analysis - The correlation heatmap effectively visualizes relationships among numeric variables. - Significant correlations, such as `gcs` and its subcomponents (`gcseyes` , `gcsmotor` , `gcsverbal`) and blood pressure (`sysbp` , `disabp` , `meanarterialpressure`), are expected due to their clinical relationships. - Variables like `gcs` and its subcomponents might exhibit multicollinearity. Consider removing highly correlated variables before regression or model training. - The negative correlation of `gcsverbal` and `endotrachflag` makes sense as intubation often impairs verbal response.

```
##
## Highly Correlated Numeric Variable Pairs:
##          Var1                 Var2 Correlation
## 1      diasbp meanarterialpressure   0.8194698
## 2         gcs              gcseyes   0.9188379
## 3         gcs             gcsmotor   0.8780389
## 4     gcseyes             gcsmotor   0.8150518
## 5         gcs            gcsverbal   0.8593701
## 6         gcs        endotrachflag  -0.8035365
## 7   gcsverbal        endotrachflag  -0.9378105
```

```
## Data filtering and correlation analysis completed.
```

If highly correlated numeric variable pairs found， use Principal Component Analysis (PCA) or select one representative variable from each correlated group to avoid redundancy in predictive modeling.

Insights of step 3.5 results




```
##
## T-Test for Age by Mortality Status:
```

```
##
##  Welch Two Sample t-test
##
## data:  age_years by expire_flag.x
## t = -207.58, df = 345654, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
0
## 95 percent confidence interval:
##  -10.71817 -10.51766
## sample estimates:
## mean in group 0 mean in group 1
##        57.94767        68.56558
```

```
##
## Gender Distribution by Mortality:
```

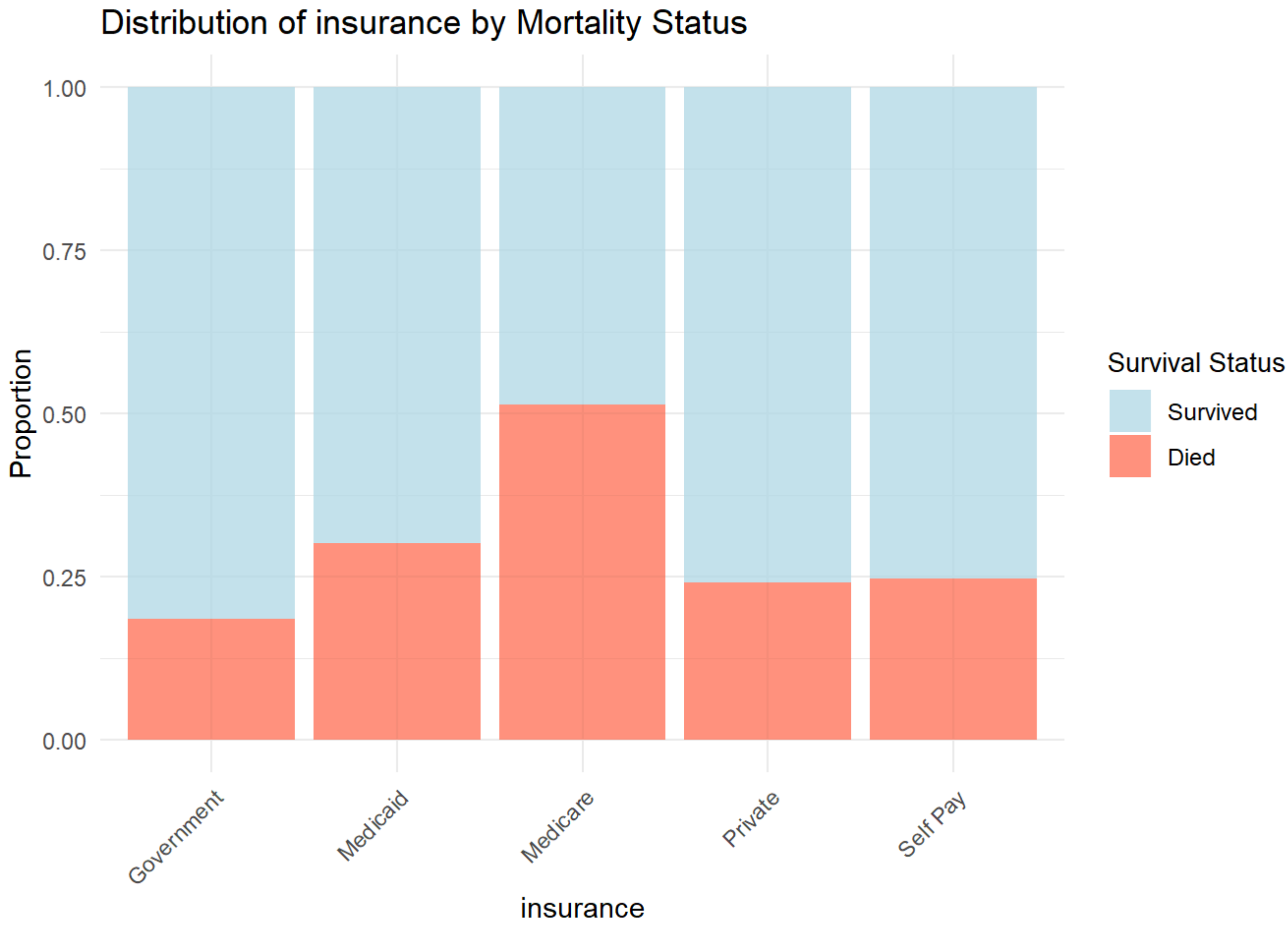
```
##
##          0      1
##   F  96636  64974
##   M 134557  79385
```

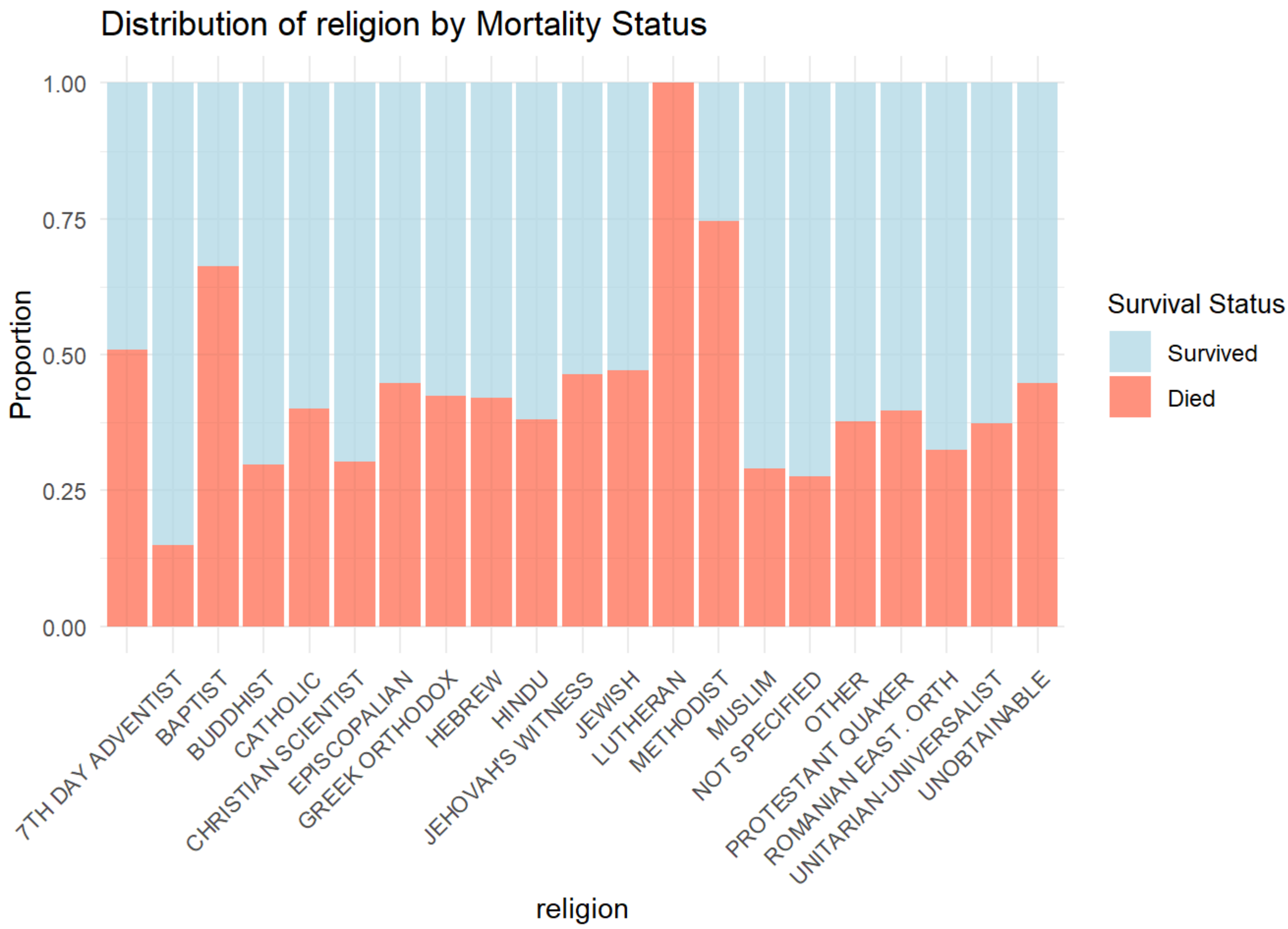
```
##
## Chi-Squared Test for Gender by Mortality:
```

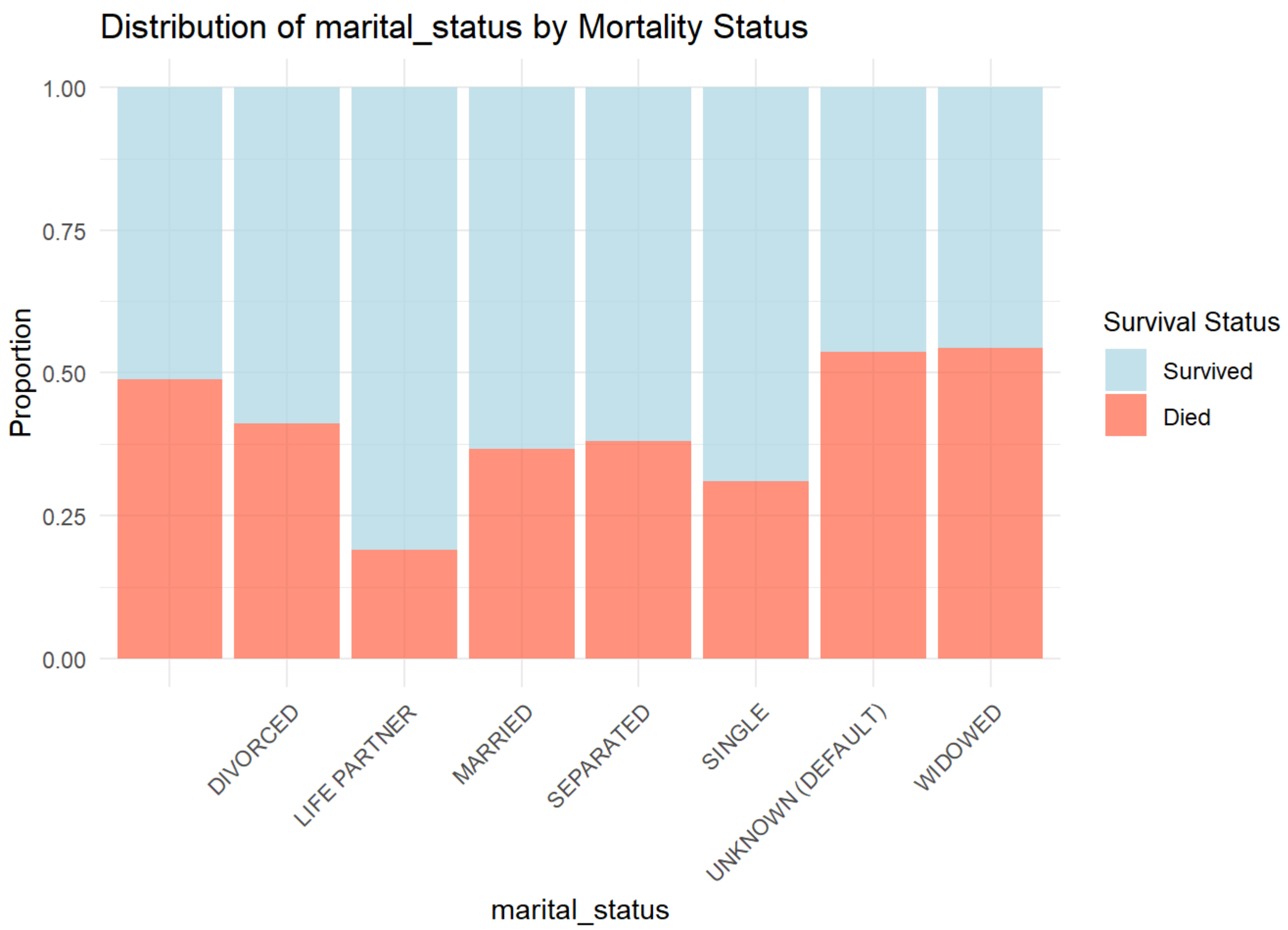
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  gender_table
## X-squared = 373.36, df = 1, p-value < 2.2e-16
```

Age and gender Distribution:

1. **Visualization:**
   - Mortality is higher among older age groups for both genders, with non-survivor bars dominating at higher ages.
   - Males have a slightly higher proportion of survivors in the younger age groups compared to females.
   - The overlap in the middle age range indicates similar mortality rates between genders for these age categories.
2. **Statistical Test:**
   - The Welch Two Sample t-test reveals a statistically significant difference in the mean age of survivors (57.95 years) and non-survivors (68.57 years), with a p-value < 2.2e-16. This highlights that age is a critical factor associated with mortality.
   - The mortality is slightly higher among males (59.02% of non-survivors) compared to females (40.98% of non-survivors).
   - The Chi-squared test confirms a significant association between gender and mortality (p-value < 2.2e-16). However, the effect size would need further exploration to assess its clinical relevance.


Distribution of insurance by Mortality Status
1.00
0.75
0.50
0.25
0.00
Proportion
Government
Medicaid
Medicare
Private
Self Pay
insurance
Survival Status
Survived
Died


Distribution of religion by Mortality Status
1.00
0.75
0.50
0.25
0.00
Proportion
7TH DAY ADVENTIST
BAPTIST
BUDDHIST
CATHOLIC
CHRISTIAN SCIENTIST
EPISCOPALIAN
GREEK ORTHODOX
HEBREW
HINDU
JEHOVAH'S WITNESS
JEWISH
LUTHERAN
METHODIST
MUSLIM
NOT SPECIFIED
OTHER
PROTESTANT QUAKER
ROMANIAN EAST. ORTH
UNITARIAN-UNIVERSALIST
UNOBTAINABLE
religion
Survival Status
Survived
Died




```
## Language Distribution:
```

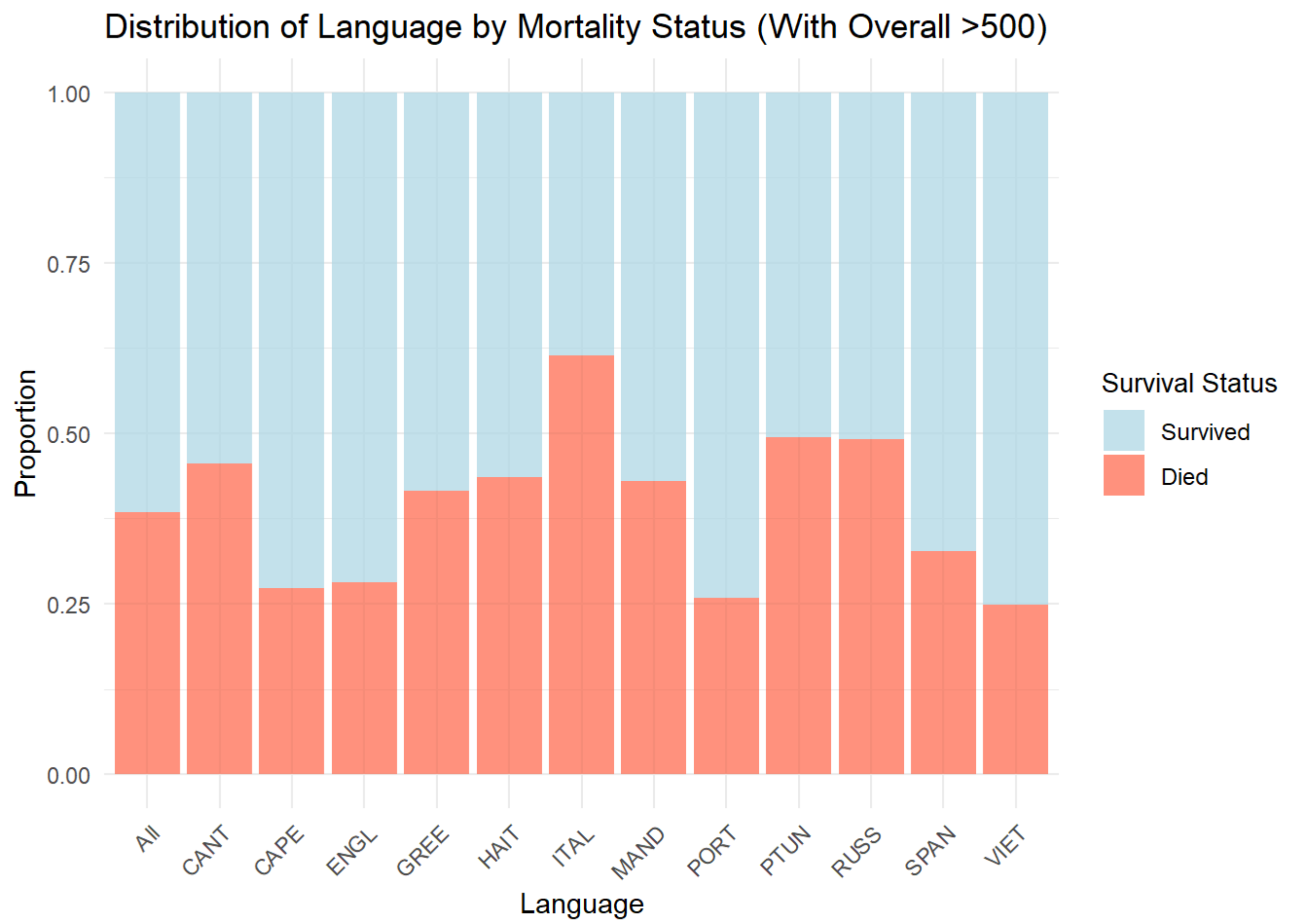
```
##     Category  Count   Percentage
##       <char>  <int>        <num>
##  1:     ENGL 203603 5.421433e+01
##  2:          140543 3.742305e+01
##  3:     SPAN   8601 2.290229e+00
##  4:     PTUN   4700 1.251491e+00
##  5:     RUSS   4154 1.106105e+00
##  6:     CANT   2666 7.098884e-01
##  7:     PORT   2638 7.024327e-01
##  8:     CAPE   1690 4.500043e-01
##  9:     MAND   1041 2.771920e-01
## 10:     HAIT   1027 2.734641e-01
## 11:     VIET    755 2.010374e-01
## 12:     ITAL    665 1.770727e-01
## 13:     GREE    520 1.384628e-01
## 14:     ARAB    303 8.068124e-02
## 15:     AMER    269 7.162790e-02
## 16:     PERS    267 7.109535e-02
## 17:     HIND    224 5.964554e-02
## 18:     CAMB    184 4.899455e-02
## 19:     POLI    157 4.180513e-02
## 20:     KORE    155 4.127258e-02
## 21:     *BEN    122 3.248551e-02
## 22:     ETHI    121 3.221924e-02
## 23:     FREN    104 2.769257e-02
## 24:     ALBA     98 2.609492e-02
## 25:     LAOT     98 2.609492e-02
## 26:     THAI     77 2.050315e-02
## 27:     *ARM     76 2.023688e-02
## 28:     *GUJ     50 1.331374e-02
## 29:     JAPA     49 1.304746e-02
## 30:     *BUL     49 1.304746e-02
## 31:     SOMA     28 7.455692e-03
## 32:     *URD     27 7.189417e-03
## 33:     *DUT     25 6.656868e-03
## 34:     TURK     25 6.656868e-03
## 35:     *FAR     24 6.390593e-03
## 36:     *TEL     24 6.390593e-03
## 37:     *NEP     24 6.390593e-03
## 38:     TAGA     24 6.390593e-03
## 39:     *TOI     24 6.390593e-03
## 40:     *KHM     24 6.390593e-03
## 41:     *PUN     24 6.390593e-03
## 42:     ** T     24 6.390593e-03
## 43:     *MAN     24 6.390593e-03
## 44:     *PHI     24 6.390593e-03
## 45:     * BE     24 6.390593e-03
## 46:     *TOY     24 6.390593e-03
## 47:     *YOR     24 6.390593e-03
## 48:     *YID     24 6.390593e-03
## 49:     *ARA     24 6.390593e-03
## 50:     BENG     24 6.390593e-03
## 51:     * FU     24 6.390593e-03
## 52:     *IBO      5 1.331374e-03
## 53:     *CDI      4 1.065099e-03
```

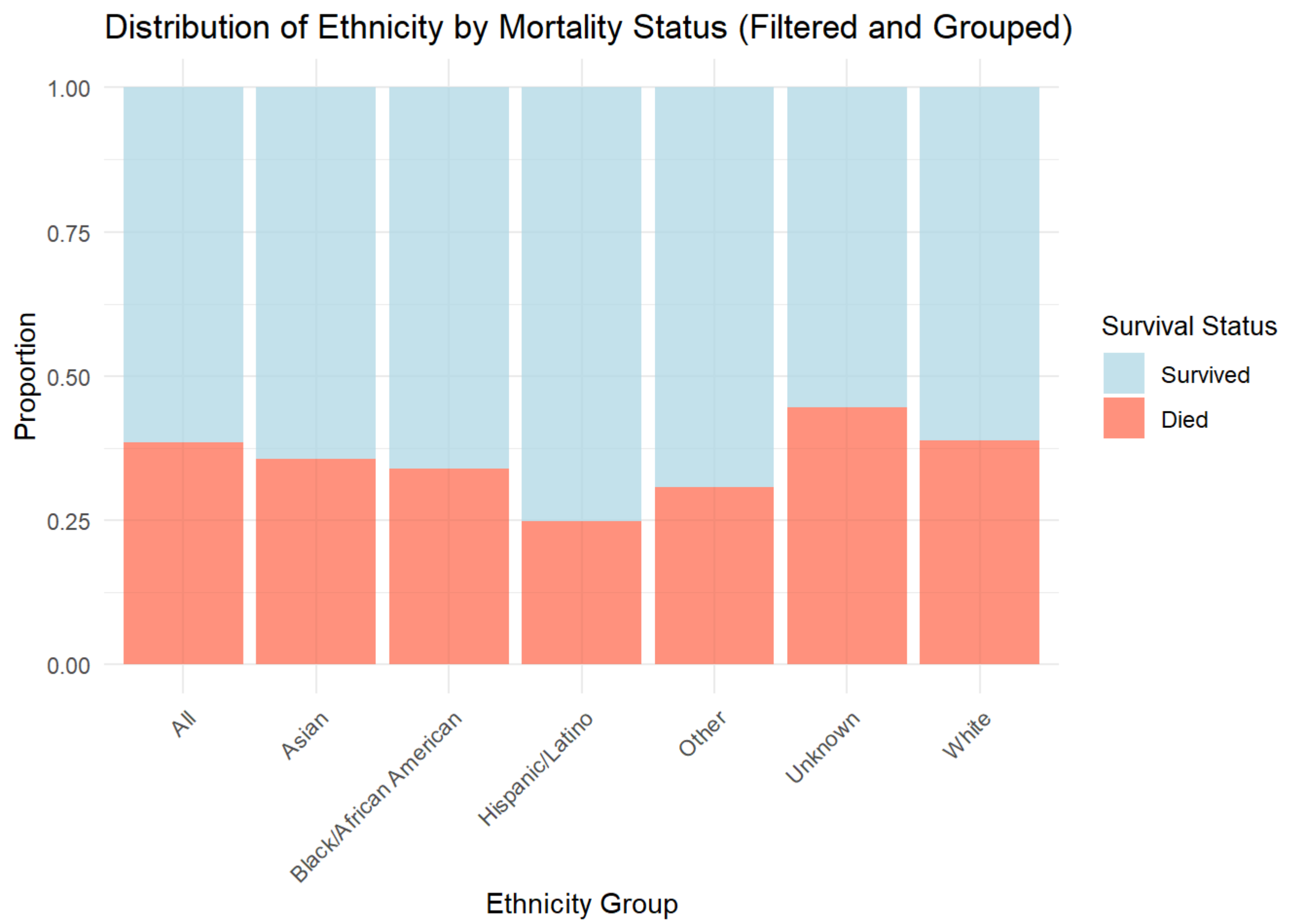
```
## 54:      *HUN      2 5.325494e-04
## 55:      *BUR      2 5.325494e-04
## 56:      URDU      2 5.325494e-04
## 57:      **TO      2 5.325494e-04
## 58:      *AMH      2 5.325494e-04
## 59:      *LEB      2 5.325494e-04
## 60:      *PER      1 2.662747e-04
## 61:      **SH      1 2.662747e-04
## 62:      *SPA      1 2.662747e-04
## 63:      *FIL      1 2.662747e-04
## 64:      *BOS      1 2.662747e-04
## 65:      *ROM      1 2.662747e-04
## 66:      *MOR      1 2.662747e-04
## 67:      SERB      1 2.662747e-04
## 68:      *CAN      1 2.662747e-04
## 69:      *DEA      1 2.662747e-04
## 70:      *FUL      1 2.662747e-04
## 71:      *TAM      1 2.662747e-04
##     Category  Count   Percentage
```

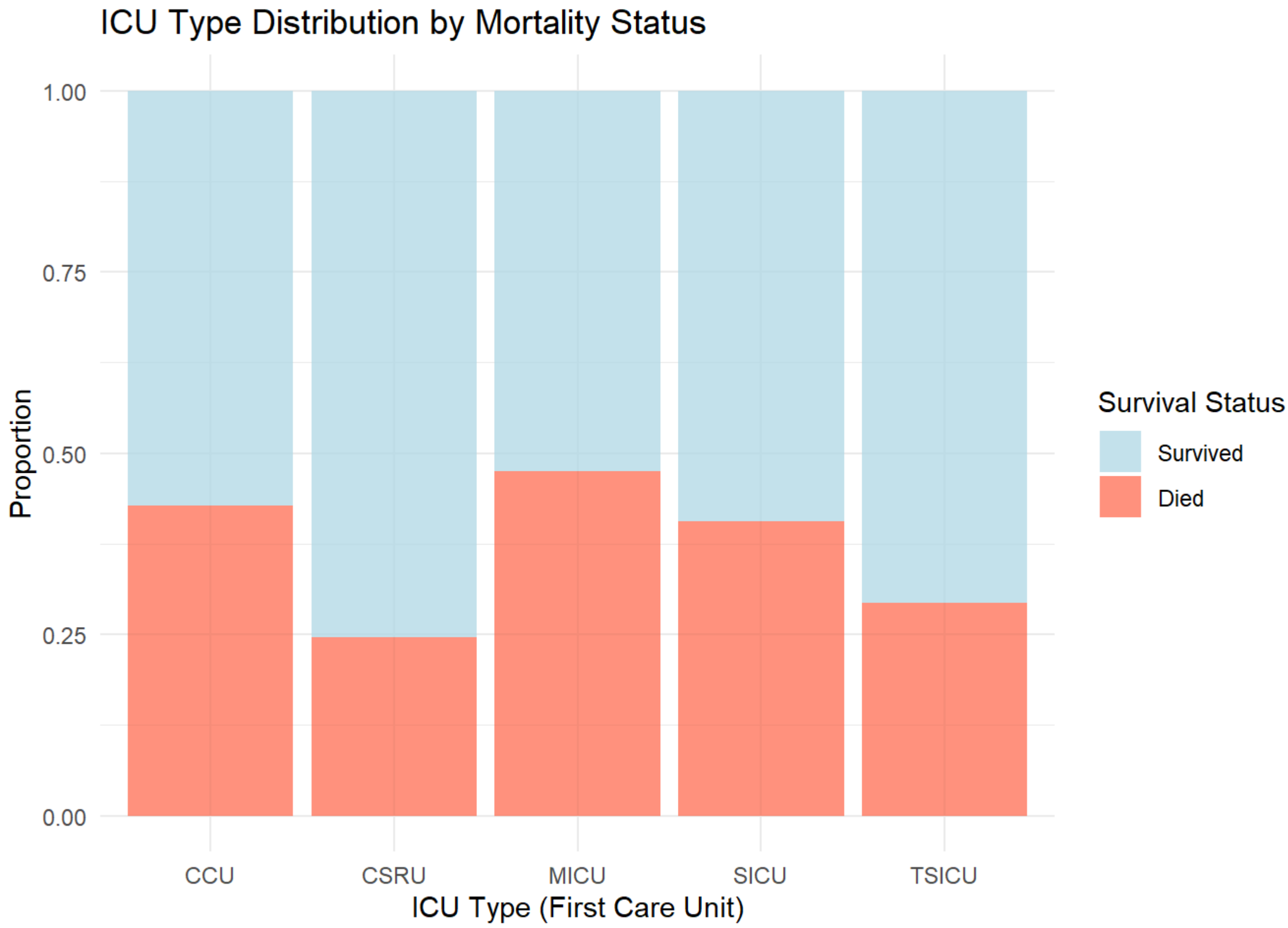
```
##
## Ethnicity Distribution:
```

```
##                                                       Category  Count
##                                                         <char>  <int>
##  1:                                                      WHITE 261734
##  2:                                      UNKNOWN/NOT SPECIFIED  37445
##  3:                                     BLACK/AFRICAN AMERICAN  27384
##  4:                                         HISPANIC OR LATINO   9905
##  5:                                                      OTHER   9199
##  6:                                           UNABLE TO OBTAIN   7514
##  7:                                                      ASIAN   5655
##  8:                                  PATIENT DECLINED TO ANSWER   4142
##  9:                                            ASIAN - CHINESE   1685
## 10:                             HISPANIC/LATINO - PUERTO RICAN   1578
## 11:                                         BLACK/CAPE VERDEAN   1313
## 12:                                       MULTI RACE ETHNICITY    837
## 13:                                            WHITE - RUSSIAN    813
## 14:                               HISPANIC/LATINO - DOMINICAN    664
## 15:                                               BLACK/HAITIAN    660
## 16:                                     WHITE - OTHER EUROPEAN    634
## 17:                                          WHITE - BRAZILIAN    520
## 18:                                       ASIAN - ASIAN INDIAN    419
## 19:                                         ASIAN - VIETNAMESE    374
## 20:                                                 PORTUGUESE    331
## 21:                                              BLACK/AFRICAN    301
## 22:                                            MIDDLE EASTERN    300
## 23:                               HISPANIC/LATINO - GUATEMALAN    281
## 24:                                   WHITE - EASTERN EUROPEAN    204
## 25:                                    HISPANIC/LATINO - CUBAN    181
## 26:                                              ASIAN - OTHER    148
## 27:                                           ASIAN - FILIPINO    128
## 28:                             AMERICAN INDIAN/ALASKA NATIVE    127
## 29:                               HISPANIC/LATINO - SALVADORAN    127
## 30:                                          ASIAN - CAMBODIAN    125
## 31:                                  HISPANIC/LATINO - MEXICAN    124
## 32:                                HISPANIC/LATINO - COLOMBIAN    123
## 33:                                           CARIBBEAN ISLAND    117
## 34:                                             ASIAN - KOREAN    102
## 35:                  NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER     80
## 36:                 HISPANIC/LATINO - CENTRAL AMERICAN (OTHER)     76
## 37:                                             SOUTH AMERICAN     74
## 38:                                           ASIAN - JAPANESE     51
## 39:                                               ASIAN - THAI     26
## 40:                                 HISPANIC/LATINO - HONDURAN     26
## 41: AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGNIZED TRIBE     25
##                                                       Category  Count
##        Percentage
##             <num>
##  1: 69.693145024
##  2:  9.970656527
##  3:  7.291666667
##  4:  2.637451005
##  5:  2.449461060
##  6:  2.000788173
##  7:  1.505783487
##  8:  1.102909850
##  9:  0.448672887
```

```
## 10:   0.420181493
## 11:   0.349618695
## 12:   0.222871933
## 13:   0.216481339
## 14:   0.176806408
## 15:   0.175741309
## 16:   0.168818166
## 17:   0.138462849
## 18:   0.111569104
## 19:   0.099586742
## 20:   0.088136929
## 21:   0.080148688
## 22:   0.079882413
## 23:   0.074823194
## 24:   0.054320041
## 25:   0.048195723
## 26:   0.039408657
## 27:   0.034083163
## 28:   0.033816888
## 29:   0.033816888
## 30:   0.033284339
## 31:   0.033018064
## 32:   0.032751789
## 33:   0.031154141
## 34:   0.027160020
## 35:   0.021301977
## 36:   0.020236878
## 37:   0.019704329
## 38:   0.013580010
## 39:   0.006923142
## 40:   0.006923142
## 41:   0.006656868
##         Percentage
```


Distribution of Language by Mortality Status (With Overall >500)
1.00
0.75
0.50
0.25
0.00
Proportion
All
CANT
CAPE
ENGL
GREE
HAIT
ITAL
MAND
PORT
PTUN
RUSS
SPAN
VIET
Language
Survival Status
Survived
Died


Distribution of Ethnicity by Mortality Status (Filtered and Grouped)
1.00
0.75
0.50
0.25
0.00
Proportion
All
Asian
Black/African American
Hispanic/Latino
Other
Unknown
White
Ethnicity Group
Survival Status
Survived
Died




```
##
## ICU Type Distribution by Mortality:
```

```
##
##              0     1
##   CCU   31674 23622
##   CSRU  58908 19283
##   MICU  67127 60734
##   SICU  38033 25961
##   TSICU 35451 14759
```

```
##
## Chi-Squared Test for ICU Type by Mortality:
```

```
##
##  Pearson's Chi-squared test
##
## data:  icu_table
## X-squared = 12995, df = 4, p-value < 2.2e-16
```

ICU Characteristics:

1. **Visualization Insights**:

- **CCU (Coronary Care Unit)**: Patients in CCU have a moderate mortality rate, reflecting the focused care for cardiac conditions.

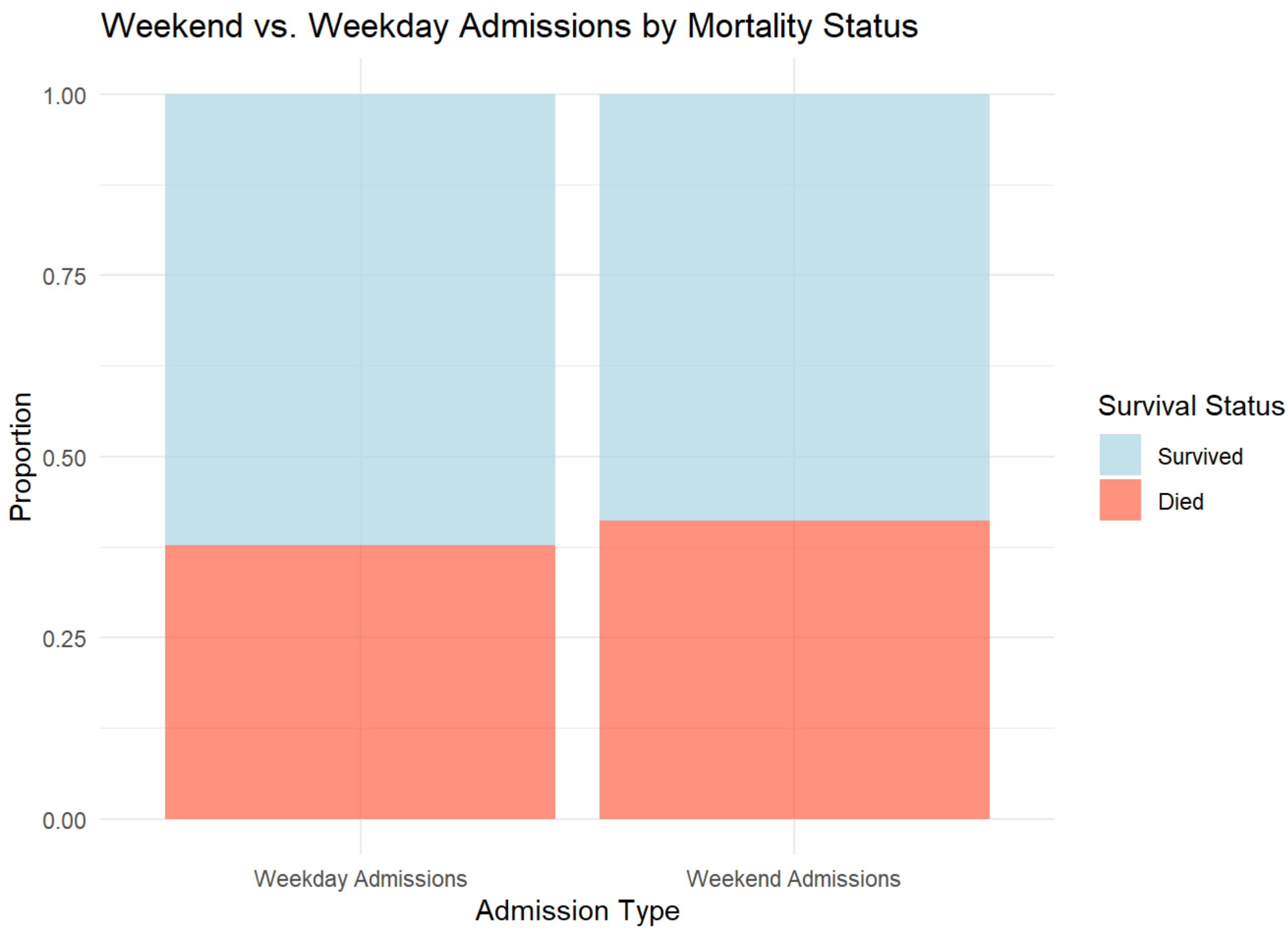
- **CSRU (Cardiac Surgery Recovery Unit)**: This unit has one of the lowest mortality rates, likely due to the controlled recovery environment post-surgery.
- **MICU (Medical Intensive Care Unit)**: The highest mortality rate is observed here, which is expected given its focus on managing critical medical conditions.
- **SICU (Surgical Intensive Care Unit)**: Mortality rates are moderate, possibly related to complex post-surgical care cases.
- **TSICU (Trauma/Surgical Intensive Care Unit)**: The lowest mortality rate suggests effective care for trauma/surgical emergencies.

2. **Statistical Analysis**:

- The chi-squared test for ICU type by mortality yields a **highly significant result** (p-value < 2.2e-16), indicating a strong association between ICU type and mortality status.

3. **Clinical Interpretation**:

- The results suggest that ICU type is a critical factor influencing patient outcomes.
- MICU patients, likely being the most critically or complexity ill, exhibit a substantially higher mortality rate.
- Further analysis could explore patient characteristics (e.g., age, comorbidities) within each ICU type to better understand these differences.




```
##
## Weekend Admission Distribution by Mortality:
```

```
##
##              0      1
##   FALSE 185322 112254
##   TRUE   45871  32105
```

```
##
## Chi-Squared Test for Weekend Admission by Mortality:
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  weekend_table
## X-squared = 310.65, df = 1, p-value < 2.2e-16
```

Weekend vs. Weekday Admissions:

1. **Visualization Insights**:

- **Weekday Admissions (FALSE)**: Higher total admissions compared to weekends, with a slightly lower proportion of mortality.
- **Weekend Admissions (TRUE)**: Lower total admissions, with a slightly higher mortality proportion compared to weekdays.

2. **Statistical Analysis**:

- The chi-squared test (p-value < 2.2e-16) confirms a statistically significant association between weekend admissions and mortality. This suggests a potential difference in outcomes based on the day of admission.

3. **Clinical Interpretation**:

- The slightly higher mortality proportion for weekend admissions may reflect differences in resource availability, staffing, or severity of cases during weekends. Further investigation into staffing levels, patient profiles, and care processes during weekends is recommended.
- Hospitals could consider optimizing weekend staffing and resources to ensure consistent care quality throughout the week.

Section 4: Feature Engineering

Objectives:

- Create meaningful features to enhance the predictive power of the dataset.
- Transform raw time-series data into aggregated features suitable for machine learning models.

Steps:

1. **Aggregate Time-Series Data**:

- Identify key vitals/lab variables and define clinical thresholds for abnormal values.
- Generate flags for abnormal values (e.g., heart rate > 120 bpm, lactate > 4 mmol/L).

2. **Feature Transformation**:

- Standardization: Apply z-score normalization to continuous variables (e.g., age, vitals, labs).
- Categorical Encoding: Convert categorical variables (e.g., ICU type, gender) to one-hot encoding.

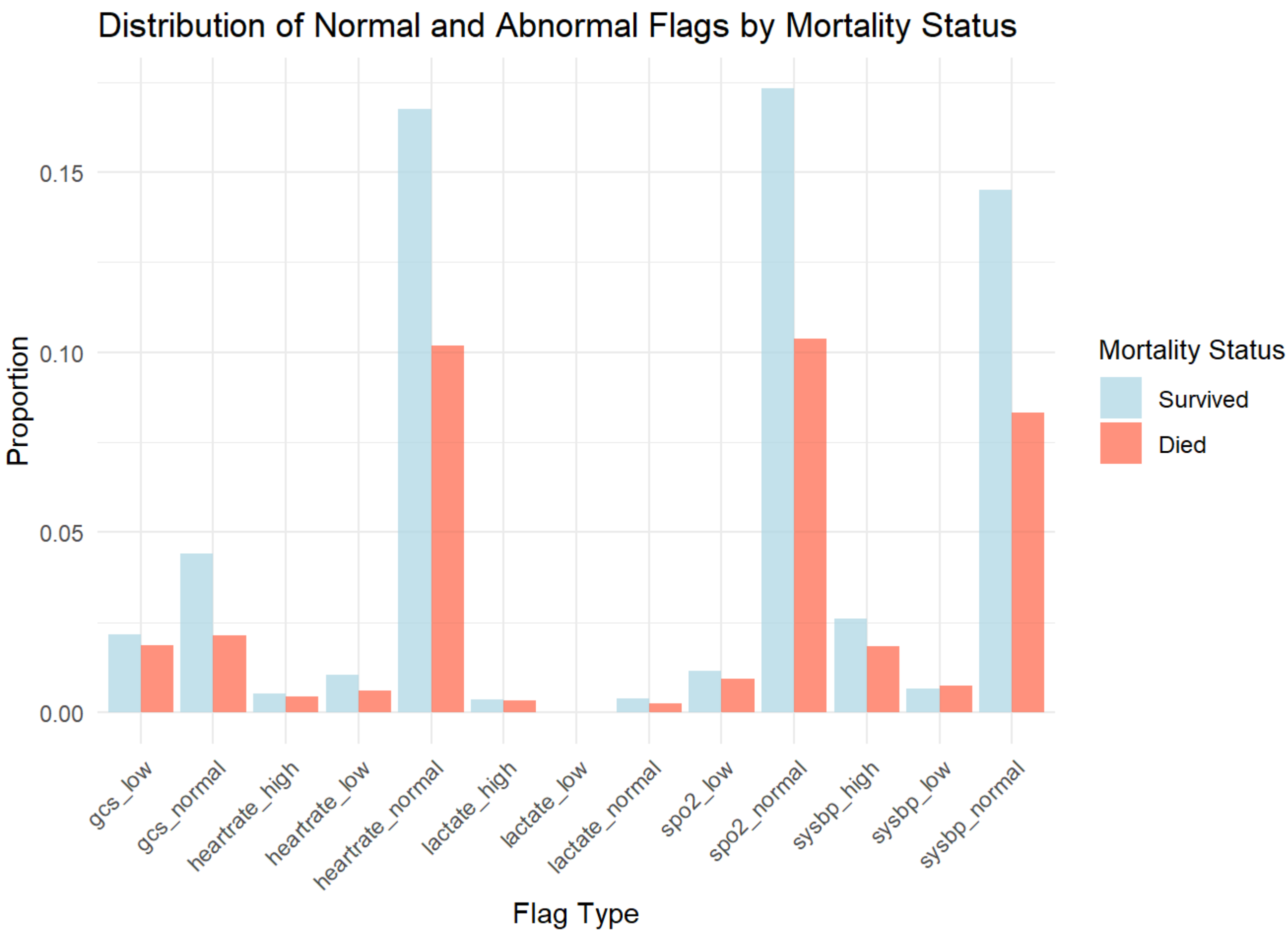
3. **Interaction and Derived Features**:

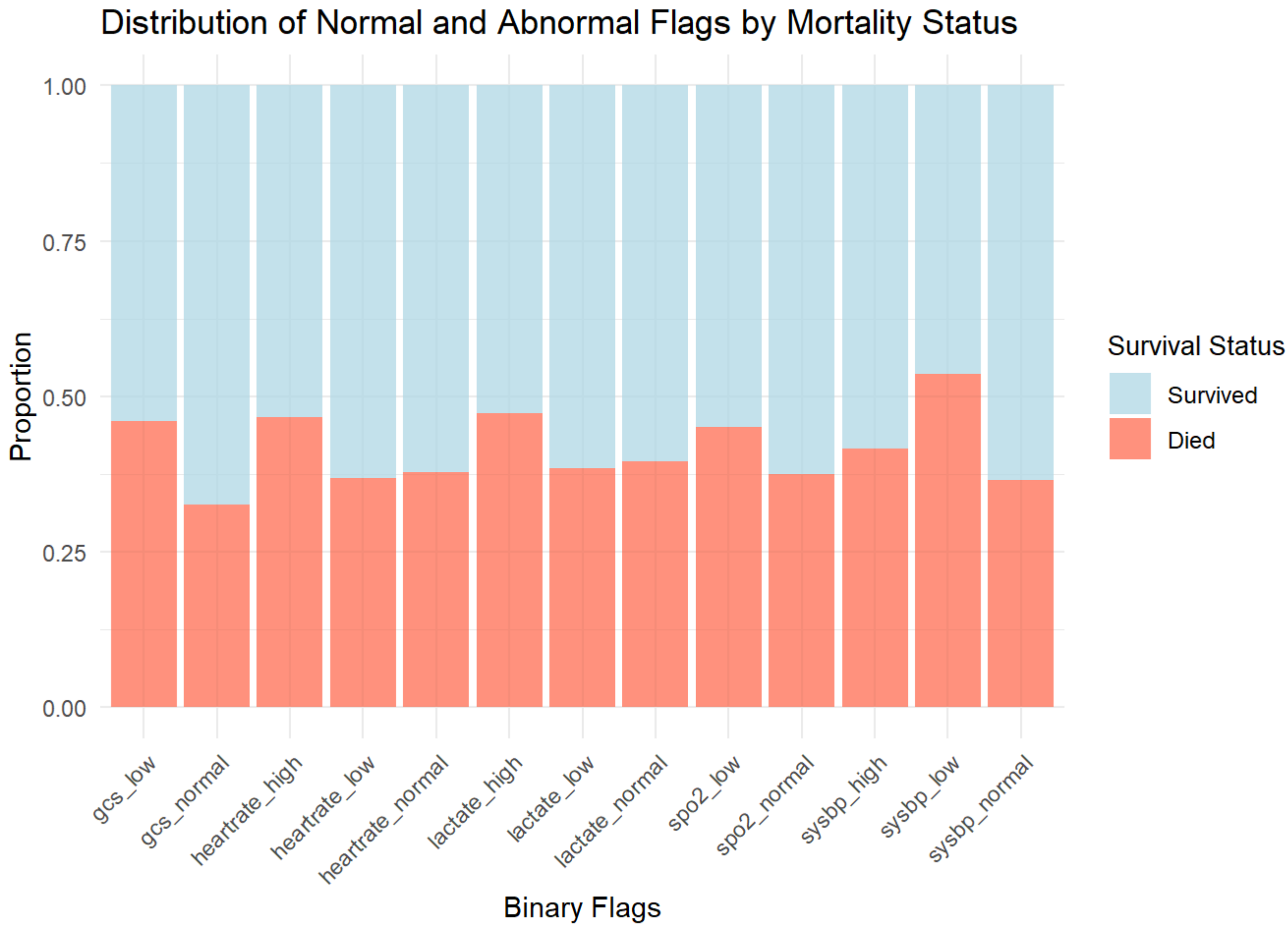
- Interaction Terms: Create interaction terms for meaningful combinations (e.g., age × ICU type, age × lactate).

- Calculate clinically relevant ratios, such as:
    - Systolic/diastolic blood pressure (sysbp/diasbp).
    - BUN/creatinine ratio.
- Cumulative Measures: Add aggregate features like total urine output in the first 24 hours.

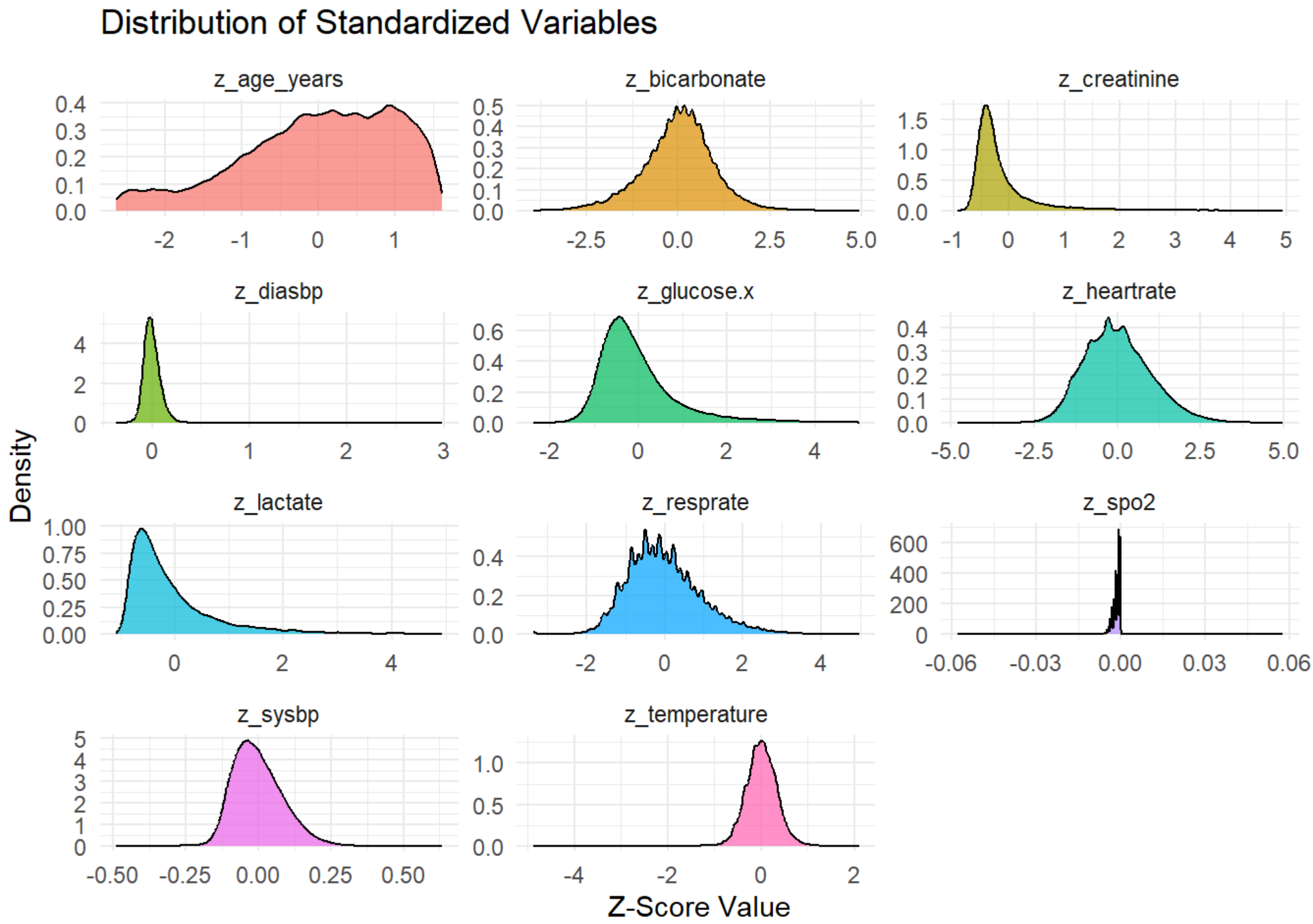
```
## [1] "Summary of Normal and Abnormal Flags:"
```

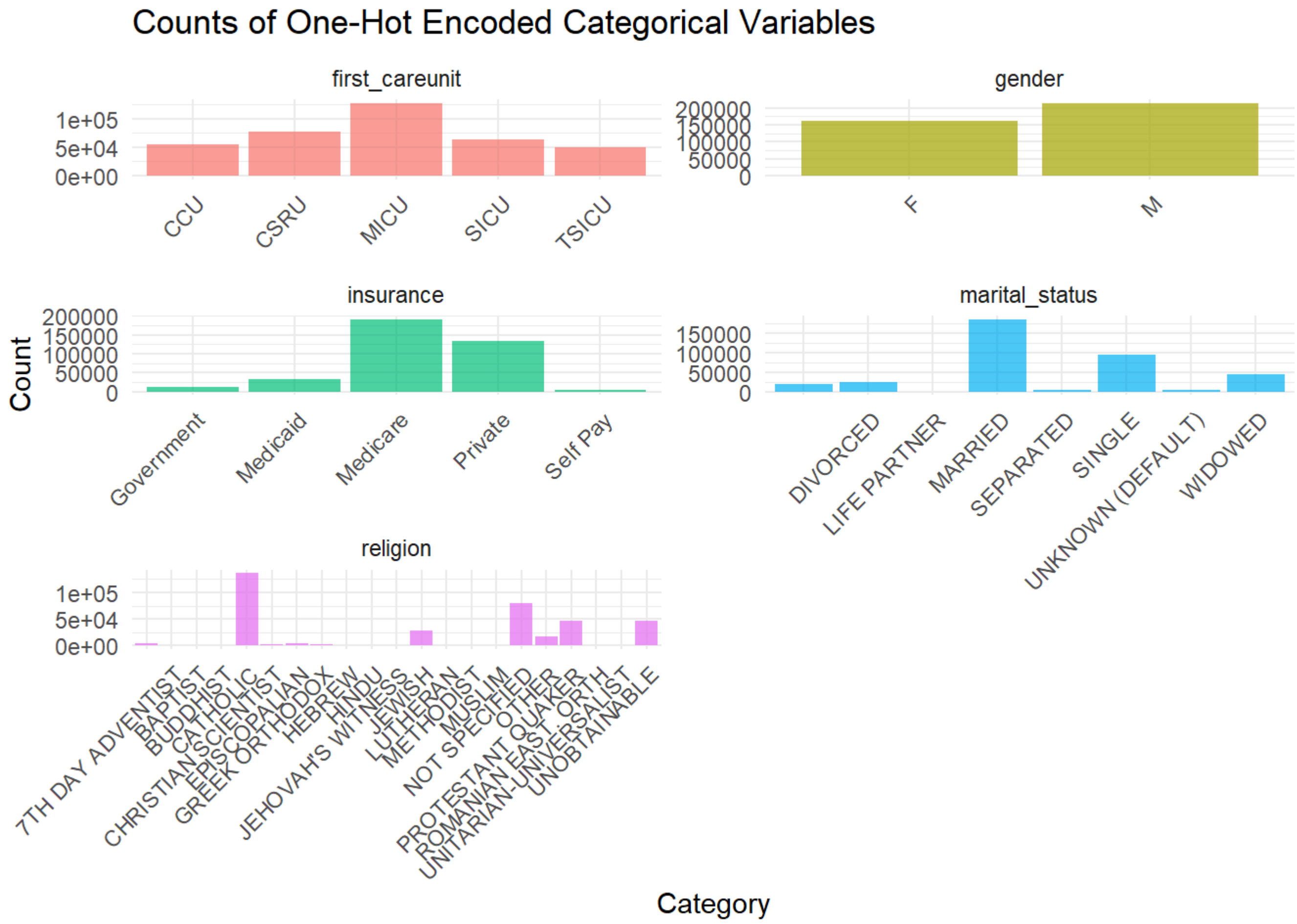
```
##   heartrate_normal lactate_normal spo2_normal sysbp_normal gcs_normal
## 1           291424           7170      299699       246987      70942
##   heartrate_high lactate_high sysbp_high heartrate_low lactate_low spo2_low
## 1          10503         7659      48259         17777          13    22683
##   sysbp_low gcs_low
## 1     15296   43626
```

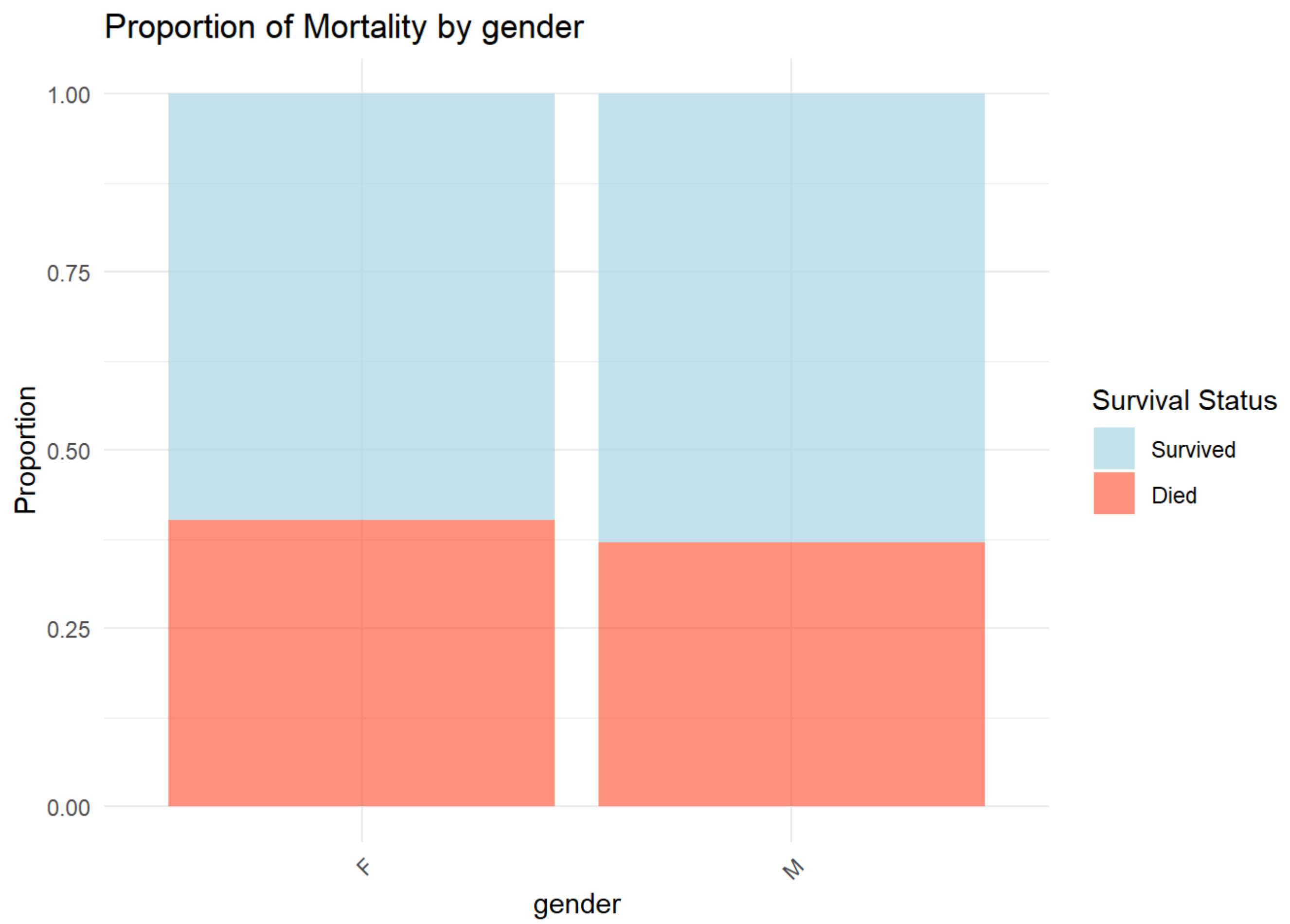


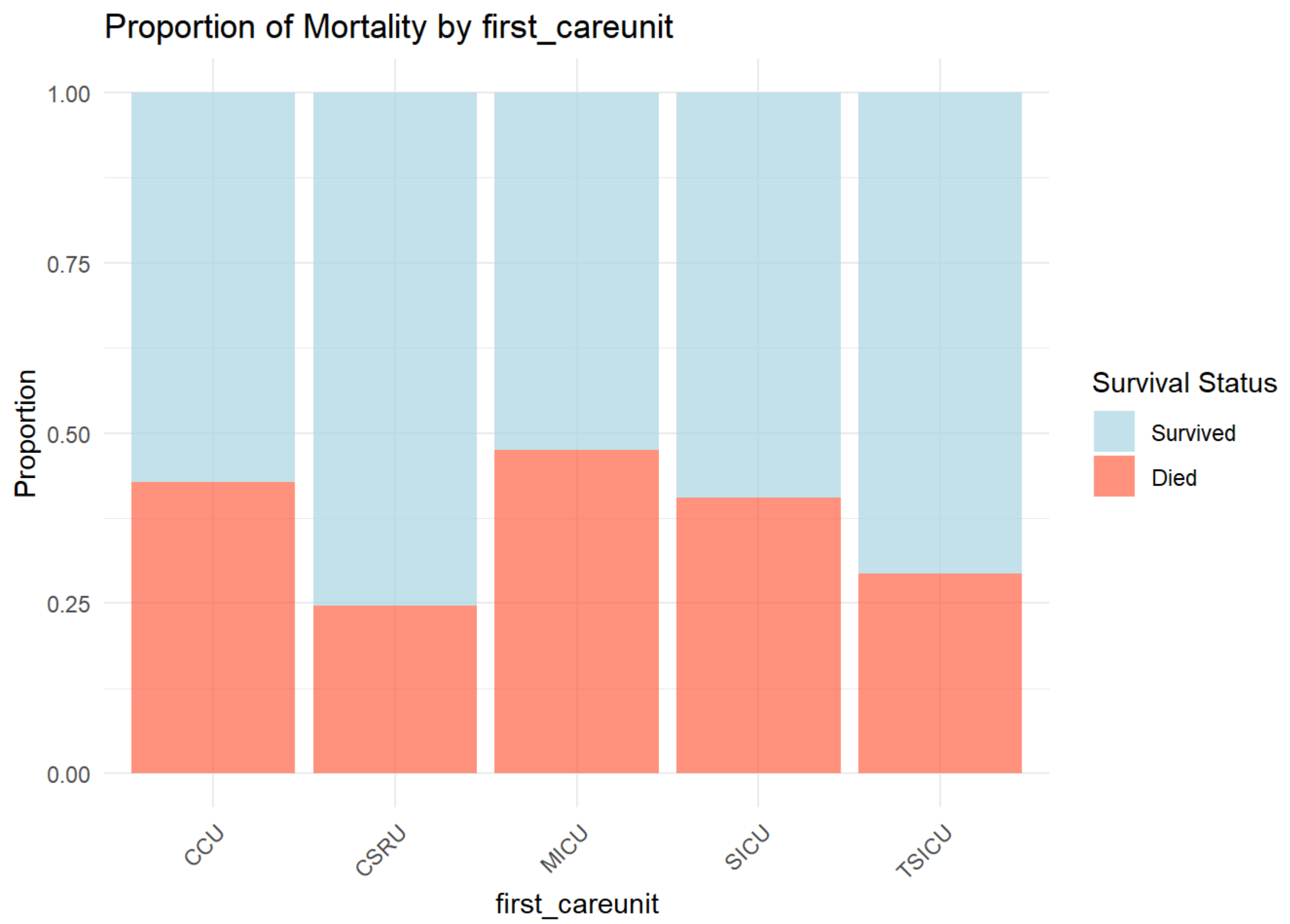


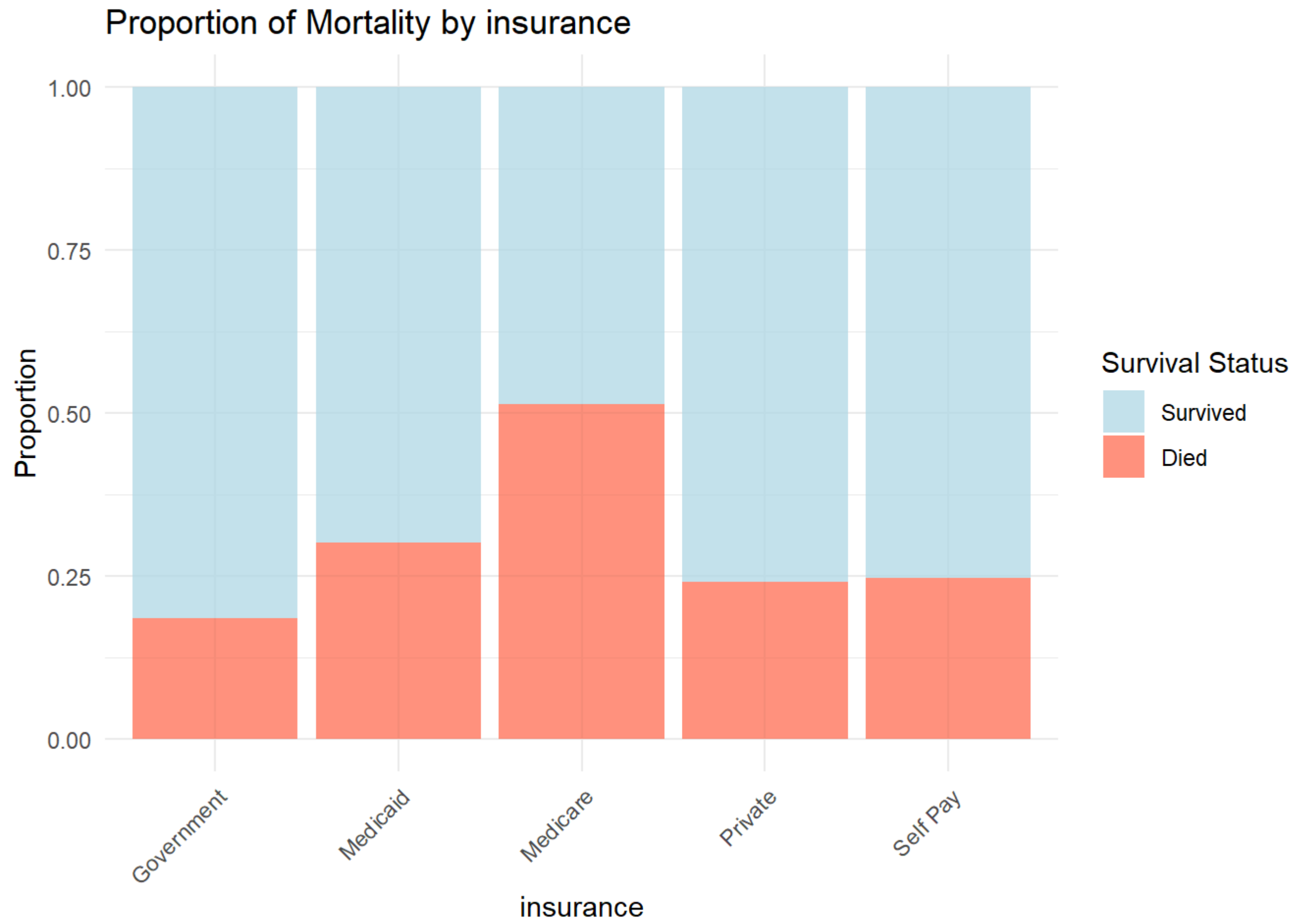

```
## Feature transformation completed: Standardized continuous variables and one-hot encoded cate
gorical variables.
```

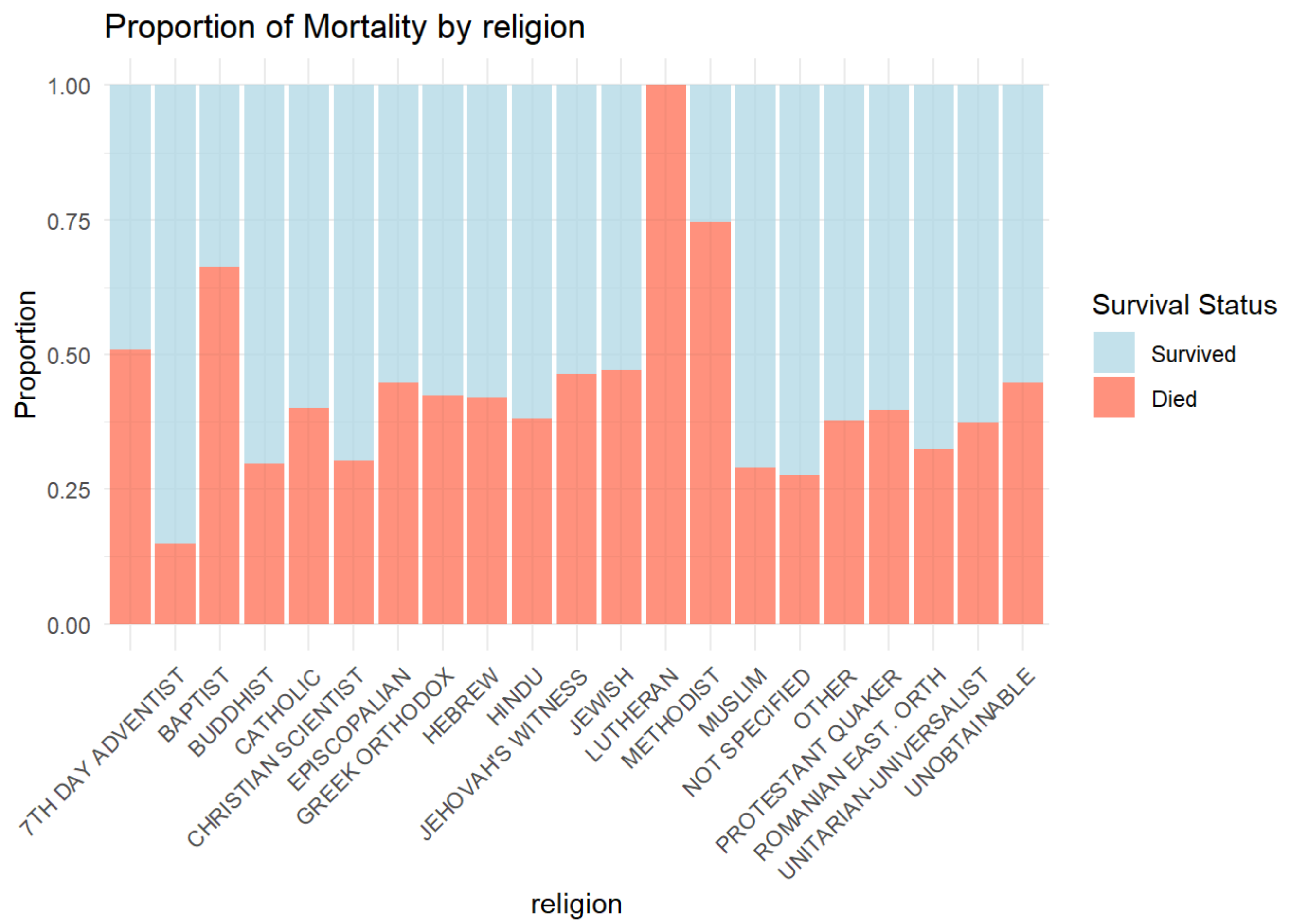

Distribution of Standardized Variables
z_age_years
z_bicarbonate
z_creatinine
z_diasbp
z_glucose.x
z_heartrate
z_lactate
z_resprate
z_spo2
z_sysbp
z_temperature
Density
Z-Score Value

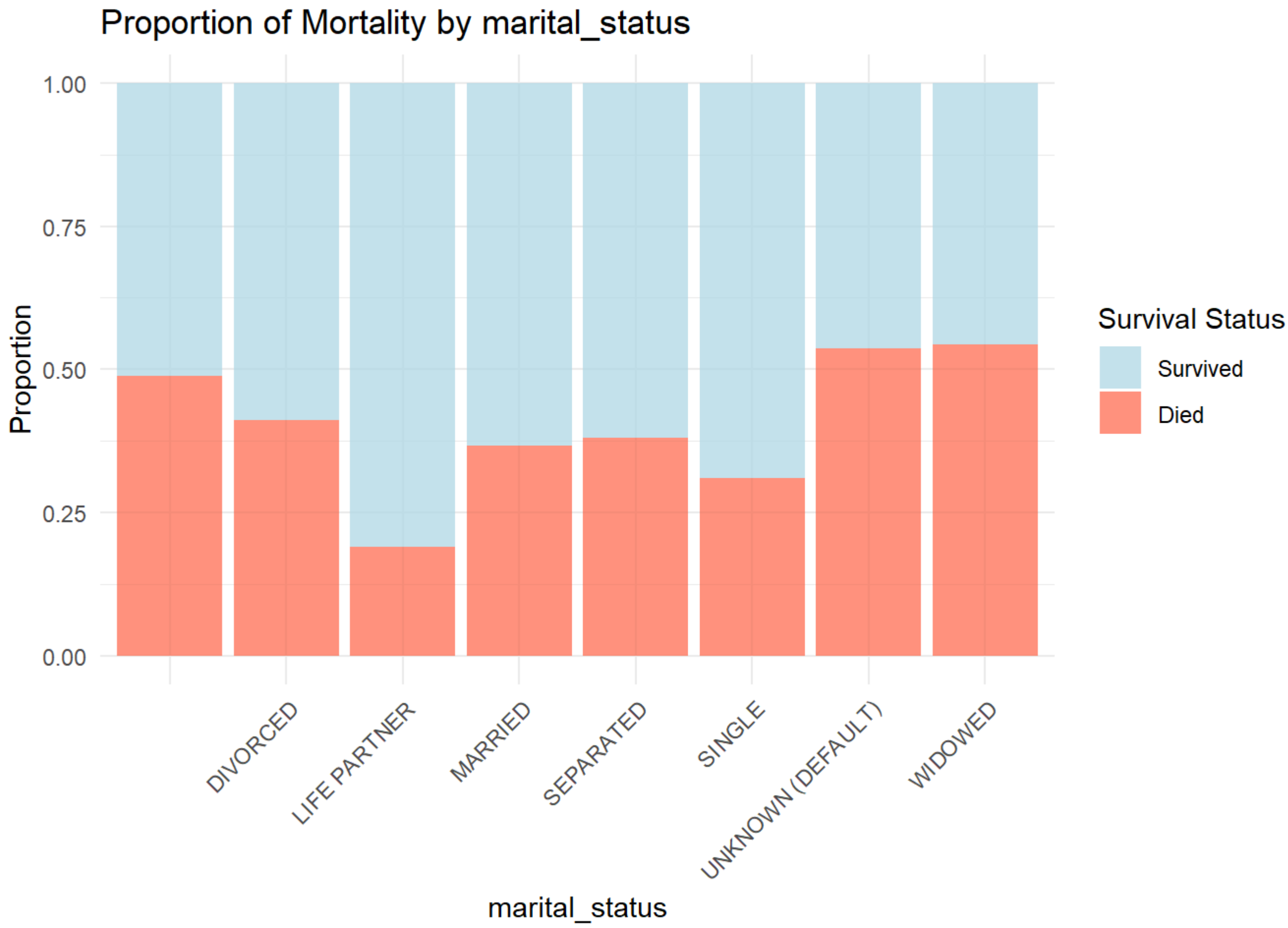

Counts of One-Hot Encoded Categorical Variables
first_careunit
CCU
CSRU
MICU
SICU
TSICU
gender
F
M
insurance
Government
Medicaid
Medicare
Private
Self Pay
marital_status
DIVORCED
LIFE PARTNER
MARRIED
SEPARATED
SINGLE
UNKNOWN (DEFAULT)
WIDOWED
religion
7TH DAY ADVENTIST
BAPTIST
BUDDHIST
CATHOLIC
CHRISTIAN SCIENTIST
EPISCOPALIAN
GREEK ORTHODOX
HEBREW
HINDU
JEHOVAH'S WITNESS
JEWISH
LUTHERAN
METHODIST
MUSLIM
NOT SPECIFIED
OTHER
PROTESTANT QUAKER
ROMANIAN EAST. ORTH
UNITARIAN-UNIVERSALIST
UNOBTAINABLE
Count
Category

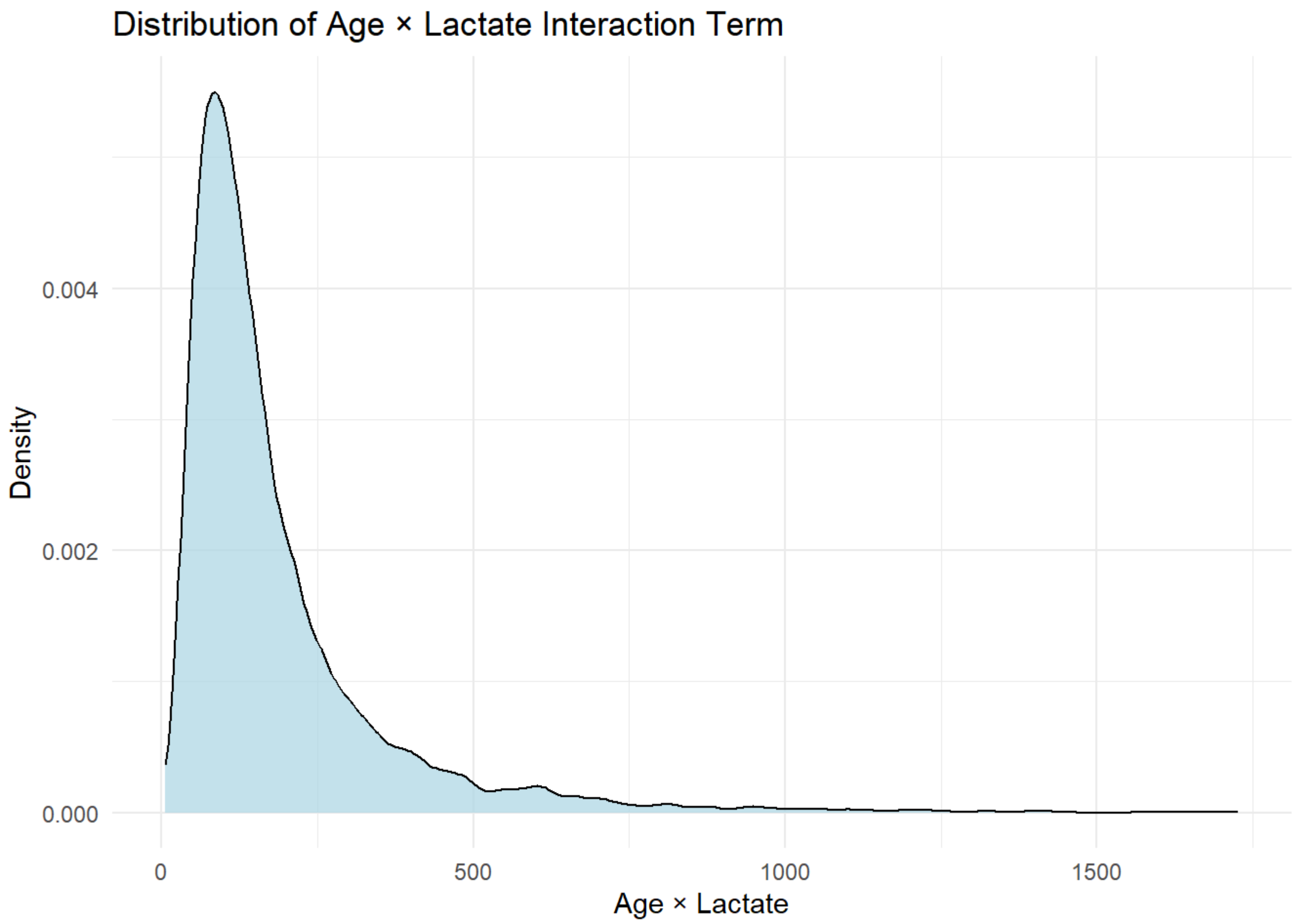

Proportion of Mortality by gender
Proportion
1.00
0.75
0.50
0.25
0.00
F
M
gender
Survival Status
Survived
Died

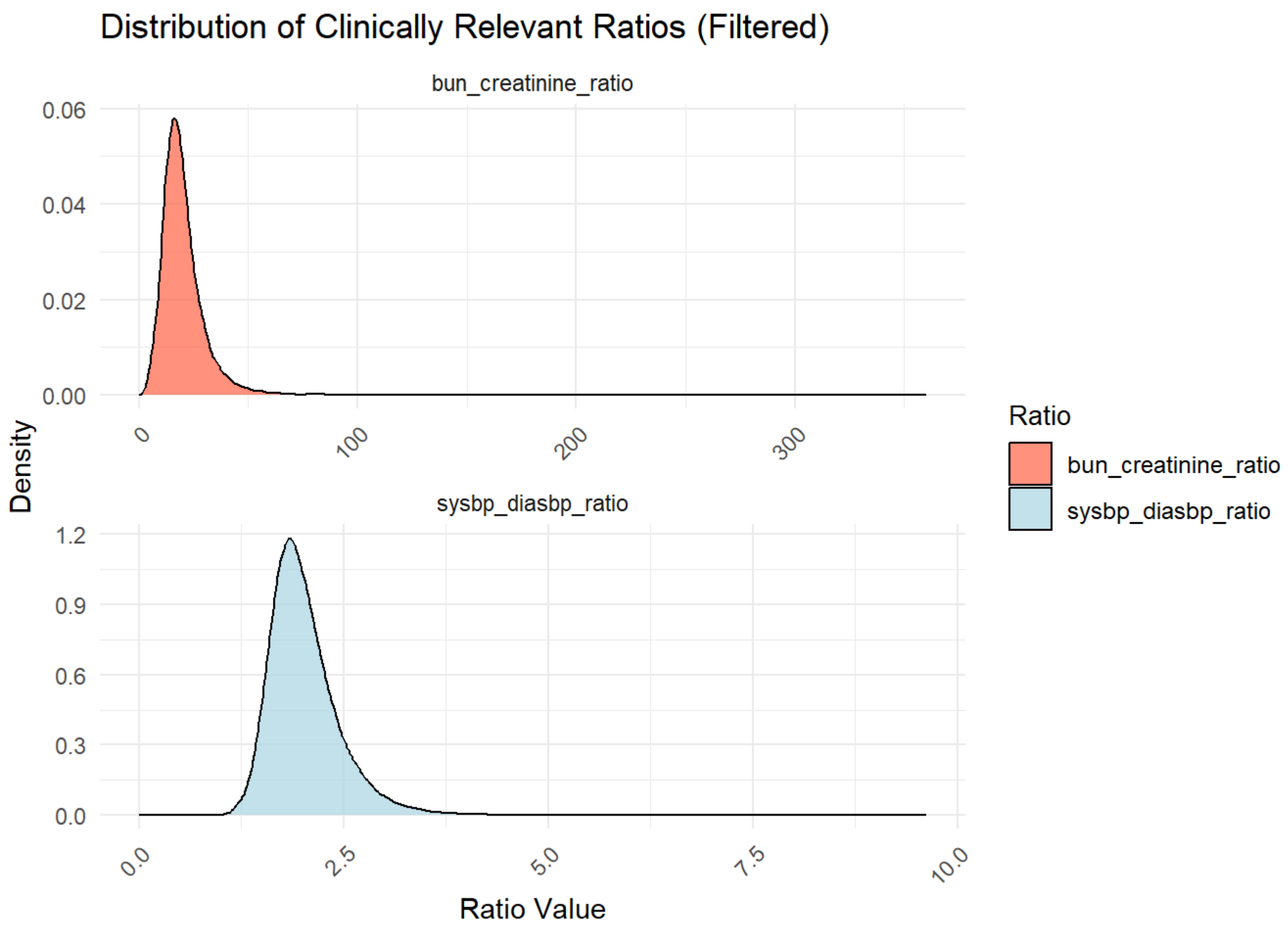

Proportion of Mortality by first_careunit
Proportion
1.00
0.75
0.50
0.25
0.00
CCU
CSRU
MICU
SICU
TSICU
first_careunit
Survival Status
Survived
Died


Proportion of Mortality by insurance
Proportion
1.00
0.75
0.50
0.25
0.00
Government
Medicaid
Medicare
Private
Self Pay
insurance
Survival Status
Survived
Died

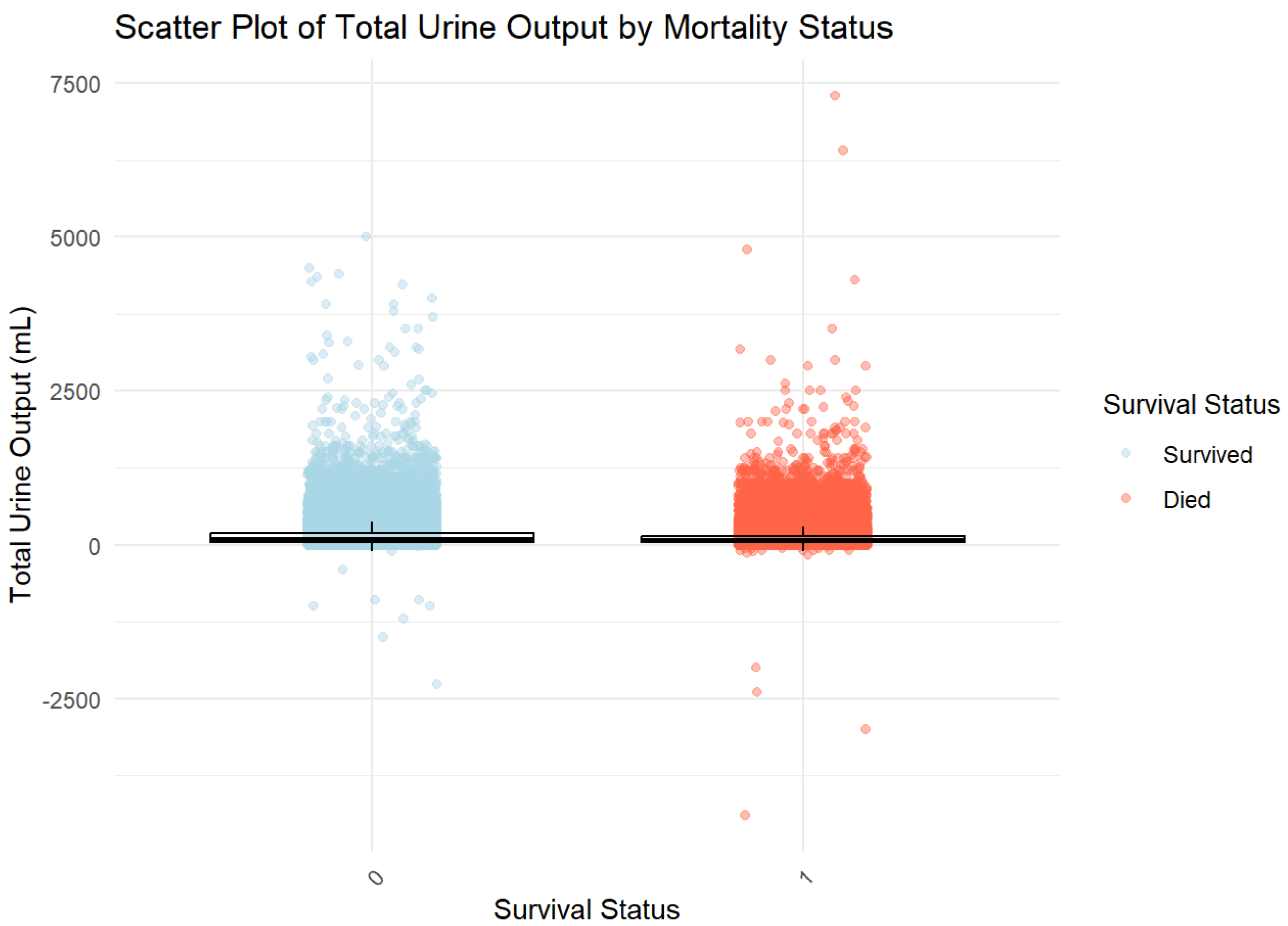

Proportion of Mortality by religion
Proportion
1.00
0.75
0.50
0.25
0.00
7TH DAY ADVENTIST
BAPTIST
BUDDHIST
CATHOLIC
CHRISTIAN SCIENTIST
EPISCOPALIAN
GREEK ORTHODOX
HEBREW
HINDU
JEHOVAH'S WITNESS
JEWISH
LUTHERAN
METHODIST
MUSLIM
NOT SPECIFIED
OTHER
PROTESTANT QUAKER
ROMANIAN EAST. ORTH
UNITARIAN-UNIVERSALIST
UNOBTAINABLE
religion
Survival Status
Survived
Died


Proportion of Mortality by marital_status
1.00
0.75
0.50
0.25
0.00
Proportion
DIVORCED
LIFE PARTNER
MARRIED
SEPARATED
SINGLE
UNKNOWN (DEFAULT)
WIDOWED
marital_status
Survival Status
Survived
Died


Distribution of Age × Lactate Interaction Term
0.004
0.002
0.000
Density
0
500
1000
1500
Age × Lactate



Distribution of Clinically Relevant Ratios (Filtered)
bun_creatinine_ratio
0.06
0.04
0.02
0.00
0
100
200
300
Density
sysbp_diasbp_ratio
1.2
0.9
0.6
0.3
0.0
0.0
2.5
5.0
7.5
10.0
Ratio Value
Ratio
bun_creatinine_ratio
sysbp_diasbp_ratio



Scatter Plot of Total Urine Output by Mortality Status
7500
5000
2500
0
-2500
Total Urine Output (mL)
0
1
Survival Status
Survival Status
Survived
Died

Section 5: Save All Processed Datasets to CSV Format