Task 1

Familiarize yourself with the codes for each of the variables. The response variable (Y) for this analysis will be the Status variable (STA). Conduct a basic exploratory data analysis to familiarize yourself with the data and the potential predictive relationships here. What is the population of interest for this problem? Do we need dropdown conditions of any kind?

```
> 'install.packages("ggplot2")
+ install.packages("gridExtra")
+ install.packages("dplyr")'
[1] "install.packages(\"ggplot2\")\ninstall.packages(\"gridExtra\")\ninstall.packages(\"dplyr\")"
>
>
> library(readxl)
> library(ggplot2)
> library(gridExtra)
> library(dplyr)
>
> # Explore relationship of STA with continuous variables
> # Read and Explore Data
>
> icu <- read_excel("icu.xlsx")
>
> # View the dataset
> View(icu)
>

 > # Structure of the dataset
 > cat("Structure of the dataset:\n")
Structure of the dataset:
 > str(icu)
tibble [200 × 21] (S3: tbl_df/tbl/data.frame)
  $ ID  : num [1:200] 4 8 12 14 27 28 32 38 40 41 ...
  $ STA : num [1:200] 1 0 0 0 1 0 0 0 0 0 ...
  $ AGE : num [1:200] 87 27 59 77 76 54 87 69 63 30 ...
  $ SEX : num [1:200] 1 1 0 0 1 0 1 0 0 1 ...
  $ RACE: num [1:200] 1 1 1 1 1 1 1 1 1 1 ...
  $ SER : num [1:200] 1 0 0 1 1 0 1 0 1 0 ...
  $ CAN : num [1:200] 0 0 0 0 0 0 0 0 0 0 ...
  $ CRN : num [1:200] 0 0 0 0 0 0 0 0 0 0 ...
  $ INF : num [1:200] 1 1 0 0 1 1 1 1 0 0 ...
  $ CPR : num [1:200] 0 0 0 0 0 0 0 0 0 0 ...
  $ SYS : num [1:200] 80 142 112 100 128 142 110 110 104 144 ...
  $ HRA : num [1:200] 96 88 80 70 90 103 154 132 66 110 ...
  $ PRE : num [1:200] 0 0 1 0 1 0 1 0 0 0 ...
  $ TYP : num [1:200] 1 1 1 0 1 1 1 1 0 1 ...
  $ FRA : num [1:200] 1 0 0 0 0 1 0 0 0 0 ...
  $ PO2 : num [1:200] 1 0 0 0 0 0 0 1 0 0 ...
  $ PH  : num [1:200] 1 0 0 0 0 0 0 0 0 0 ...
  $ PCO : num [1:200] 1 0 0 0 0 0 0 0 0 0 ...
  $ BIC : num [1:200] 0 0 0 0 0 0 0 1 0 0 ...
  $ CRE : num [1:200] 0 0 0 0 0 0 0 0 0 0 ...
  $ LOC : num [1:200] 0 0 0 0 0 0 0 0 0 0 ...
 >
 > # Dimensions
 > cat("Dimensions of the dataset: ", dim(icu), "\n")
Dimensions of the dataset:  200 21
```

```
> # Print the structure of the file
> str(icu)
tibble [200 × 21] (S3: tbl_df/tbl/data.frame)
 $ ID  : num [1:200] 4 8 12 14 27 28 32 38 40 41 ...
 $ STA : num [1:200] 1 0 0 0 1 0 0 0 0 0 ...
 $ AGE : num [1:200] 87 27 59 77 76 54 87 69 63 30 ...
 $ SEX : num [1:200] 1 1 0 0 1 0 1 0 0 1 ...
 $ RACE: num [1:200] 1 1 1 1 1 1 1 1 1 1 ...
 $ SER : num [1:200] 1 0 0 1 1 0 1 0 1 0 ...
 $ CAN : num [1:200] 0 0 0 0 0 0 0 0 0 0 ...
 $ CRN : num [1:200] 0 0 0 0 0 0 0 0 0 0 ...
 $ INF : num [1:200] 1 1 0 0 1 1 1 1 0 0 ...
 $ CPR : num [1:200] 0 0 0 0 0 0 0 0 0 0 ...
 $ SYS : num [1:200] 80 142 112 100 128 142 110 110 104 144 ...
 $ HRA : num [1:200] 96 88 80 70 90 103 154 132 66 110 ...
 $ PRE : num [1:200] 0 0 1 0 1 0 1 0 0 0 ...
 $ TYP : num [1:200] 1 1 1 0 1 1 1 1 0 1 ...
 $ FRA : num [1:200] 1 0 0 0 0 1 0 0 0 0 ...
 $ PO2 : num [1:200] 1 0 0 0 0 0 0 1 0 0 ...
 $ PH  : num [1:200] 1 0 0 0 0 0 0 0 0 0 ...
 $ PCO : num [1:200] 1 0 0 0 0 0 0 0 0 0 ...
 $ BIC : num [1:200] 0 0 0 0 0 0 0 1 0 0 ...
 $ CRE : num [1:200] 0 0 0 0 0 0 0 0 0 0 ...
 $ LOC : num [1:200] 0 0 0 0 0 0 0 0 0 0 ...
>
> # Print the fist 6 rows of the dataset
> head(icu)
# A tibble: 6 × 21
     ID   STA   AGE   SEX  RACE   SER   CAN   CRN   INF   CPR   SYS   HRA   PRE   TYP   FRA  PO2    PH   PCO   BIC   CRE   LOC
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1     4     1    87     1     1     1     0     0     1     0    80    96     0     1     1     1     1     1     0     0     0
2     8     0    27     1     1     0     0     0     1     0   142    88     0     1     0     0     0     0     0     0     0
3    12     0    59     0     1     0     0     0     0     0   112    80     1     1     0     0     0     0     0     0     0
4    14     0    77     0     1     1     0     0     0     0   100    70     0     0     0     0     0     0     0     0     0
5    27     1    76     1     1     1     0     0     1     0   128    90     1     1     0     0     0     0     0     0     0
6    28     0    54     0     1     0     0     0     1     0   142   103     0     1     1     0     0     0     0     0     0
>
```

```
> # Print the names of the columns
> names(icu)
 [1] "ID"   "STA"  "AGE"  "SEX"  "RACE" "SER"  "CAN"  "CRN"  "INF"  "CPR"  "SYS"  "HRA"  "PRE"  "TYP"  "FRA"  "PO2"  "PH"   "PCO"
[19] "BIC"  "CRE"  "LOC"
>
> # Get the names of the columns with missing values
> columns_with_missing <- names(icu)[colSums(is.na(icu)) > 0]
> print(columns_with_missing)
character(0)
>
> # Summary statistics
> summary(icu)
       ID              STA              AGE              SEX              RACE             SER              CAN              CRN
 Min.   :  4.0   Min.   :0.0   Min.   :16.00   Min.   :0.00   Min.   :1.000   Min.   :0.000   Min.   :0.0   Min.   :0.000
 1st Qu.:210.2   1st Qu.:0.0   1st Qu.:46.75   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0   1st Qu.:0.000
 Median :412.5   Median :0.0   Median :63.00   Median :0.00   Median :1.000   Median :1.000   Median :0.0   Median :0.000
 Mean   :444.8   Mean   :0.2   Mean   :57.55   Mean   :0.38   Mean   :1.175   Mean   :0.535   Mean   :0.1   Mean   :0.095
      INF              CPR              SYS              HRA              PRE              TYP              FRA              PO2
 Min.   :0.00   Min.   :0.000   Min.   : 36.0   Min.   : 39.00   Min.   :0.00   Min.   :0.000   Min.   :0.000   Min.   :0.00
 1st Qu.:0.00   1st Qu.:0.000   1st Qu.:110.0   1st Qu.: 80.00   1st Qu.:0.00   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.00
 Median :0.00   Median :0.000   Median :130.0   Median : 96.00   Median :0.00   Median :1.000   Median :0.000   Median :0.00
 Mean   :0.42   Mean   :0.065   Mean   :132.3   Mean   : 98.92   Mean   :0.15   Mean   :0.735   Mean   :0.075   Mean   :0.08
       PH              PCO              BIC              CRE              LOC
 Min.   :0.000   Min.   :0.0   Min.   :0.000   Min.   :0.00   Min.   :0.000
 1st Qu.:0.000   1st Qu.:0.0   1st Qu.:0.000   1st Qu.:0.00   1st Qu.:0.000
 Median :0.000   Median :0.0   Median :0.000   Median :0.00   Median :0.000
 Mean   :0.065   Mean   :0.1   Mean   :0.075   Mean   :0.05   Mean   :0.125
 [ reached getOption("max.print") -- omitted 2 rows ]
>
```

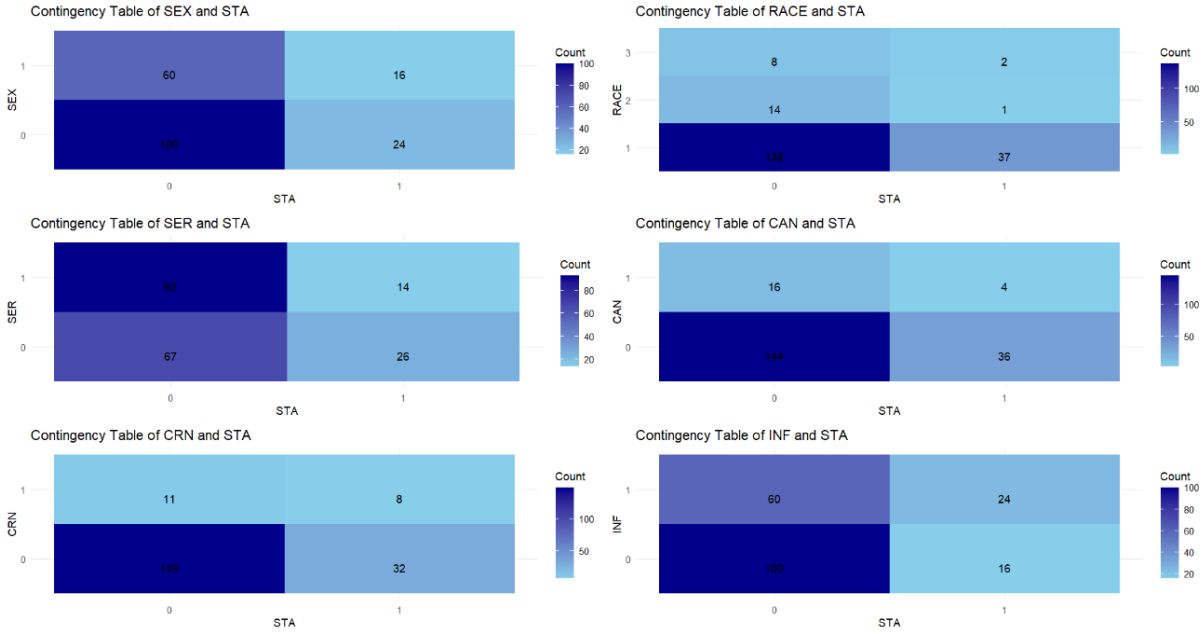
```
> # Identify column types
> continuous_cols <- names(icu)[sapply(icu, function(col) is.numeric(col) && length(unique(col)) > 10)]
> discrete_cols <- names(icu)[sapply(icu, function(col) is.numeric(col) && length(unique(col)) <= 10)]
>
>
> print(paste("Continuous columns:", paste(continuous_cols, collapse = ", ")))
[1] "Continuous columns: ID, AGE, SYS, HRA"
> print(paste("Discrete columns:", paste(discrete_cols, collapse = ", ")))
[1] "Discrete columns: STA, SEX, RACE, SER, CAN, CRN, INF, CPR, PRE, TYP, FRA, PO2, PH, PCO, BIC, CRE, LOC"
>
>
> # Remove ID from the continuous columns
> continuous_cols <- setdiff(continuous_cols, "ID")
>
>
> # Remove STA from the continuous columns
> discrete_cols <- setdiff(discrete_cols, "STA")
>
```

- ID: Identification Code for each patient.
- STA: Vital Status
  - 0 = Lived
  - 1 = Died
- AGE: Age of the patient in years.
- SEX: Gender of the patient
  - 0 = Male
  - 1 = Female
- RACE: Race of the patient
  - 1 = White
  - 2 = Black
  - 3 = Other
- SER: Service at ICU Admission

  - 0 = Medical
  - 1 = Surgical
- CAN: Indicates if cancer is part of the present problem
  - 0 = No
  - 1 = Yes
- CRN: History of Chronic Renal Failure
  - 0 = No
  - 1 = Yes
- INF: Infection Probable at ICU Admission
  - 0 = No
  - 1 = Yes
- CPR: CPR Prior to ICU Admission
  - 0 = No
  - 1 = Yes
- SYS: Systolic Blood Pressure at ICU Admission (in mm Hg).
- HRA: Heart Rate at ICU Admission (in beats/min).
- PRE: Previous Admission to an ICU within the last 6 months
  - 0 = No
  - 1 = Yes
- TYP: Type of Admission
  - 0 = Elective
  - 1 = Emergency
- FRA: Indicates if there's a long bone, multiple, neck, single area, or hip fracture
  - 0 = No
  - 1 = Yes
- PO2: PO2 from Initial Blood Gases
  - 0 = >60
  - 1 = <= 60
- PH: PH from Initial Blood Gases
  - 0 = >= 7.25
  - 1 = < 7.25
- PCO: PCO2 from Initial Blood Gases
  - 0 = <=45
  - 1 > 45
- BIC: Bicarbonate from Initial Blood Gases
  - 0 = >=18
  - 1 <18
- CRE: Creatinine from Initial Blood Gases
  - 0 = <=2
  - 1 >2

- LOC: Level of Consciousness at ICU Admission
  - 0 = No Coma or Deep Stupor
  - 1 = Deep Stupor
  - 2 = Coma

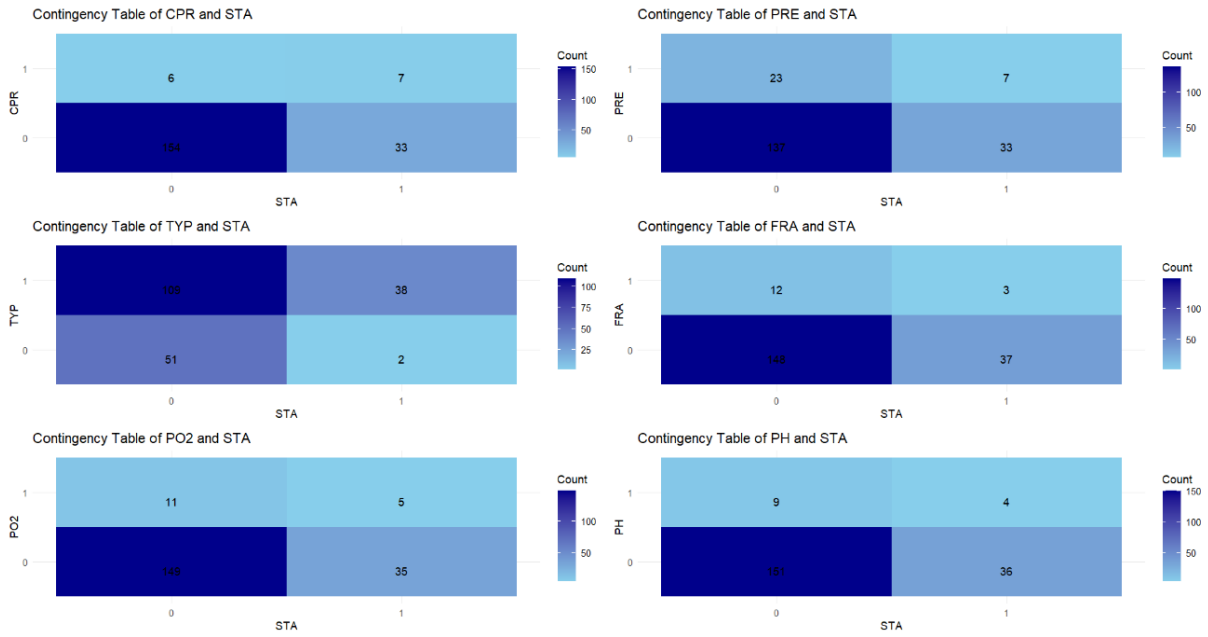
```
> ##############################################################################
> # Explore relationship of STA with discrete variables
> 
> 
> # Function to plot a Contingency Table with Heatmap
> plot_contingency_heatmap <- function(var) {
+ 
+   # Create a contingency table
+   contingency_table <- table(icu$STA, icu[[var]])
+ 
+   # Convert the contingency table into a data frame for plotting
+   plot_data <- as.data.frame(as.table(contingency_table))
+ 
+   # Create the heatmap plot
+   p <- ggplot(plot_data, aes_string(x = "Var1", y = "Var2", fill = "Freq")) +
+     geom_tile() +
+     geom_text(aes(label = sprintf("%d", Freq)), vjust = 1.5) +
+     scale_fill_gradient(low = "skyblue", high = "darkblue") +
+     labs(title = paste("Contingency Table of", var, "and STA"),
+          x = "STA",
+          y = var,
+          fill = "Count") +
+     theme_minimal()
+ 
+   return(p)
+ }
> 
> # List to hold the plots
> plots_list <- lapply(discrete_cols, plot_contingency_heatmap)
> 
> # Group every 6 plots into separate lists for panels
> plot_groups <- split(plots_list, ceiling(seq_along(plots_list)/6))
> 
> # Display each panel
> for(panel in plot_groups) {
+   do.call(grid.arrange, c(panel, ncol = 2))
+ }
```



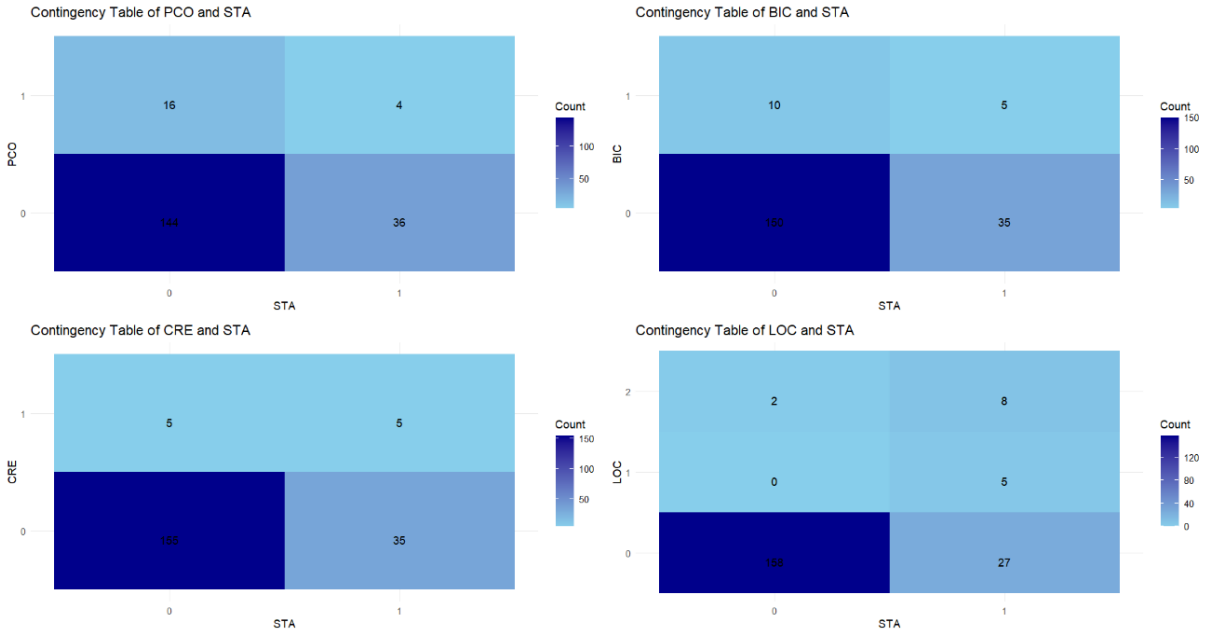
Interpretation:

1. Contingency Table of SEX and STA:
   - A majority of individuals categorized as "SEX = 0" fall under "STA = 0" (100 of 124 males survived).
   - For those labeled "SEX = 1", a clear majority (60 of 76 females survived) also fall under "STA = 0".
2. Contingency Table of RACE and STA:
   - 138 (RACE=1 and STA=0) of 175 White people survived.
   - 14 (RACE=2 and STA=0) of 15 Black people survived.
   - 8 (RACE=3 and STA=0) of 10 Other Race people survived.
3. Contingency Table of SER and STA:
   - 67 (SER=0 and STA=0) of 93 people for Medical Service survived.
   - 93 (SER=1 and STA=0) of 117 people for Surgical Service survived.
4. Contingency Table of CAN and STA:
   - 144 (CAN=0 and STA=0) of 180 people with ‘Cancer Not Part of Problem’ survived.
   - 16 (CAN=1 and STA=0) of 20 people with ‘Cancer Part of Problem’ survived.

5. Contingency Table of CRN and STA:
    - 149 (CRN=0 and STA=0) of 181 people with ‘No History of Chronic Renal Failure’ survived.
    - 11 (CRN=1 and STA=0) of 19 people with ‘History of Chronic Renal Failure’ survived.
6. Contingency Table of INF and STA:
    - 100 (INF=0 and STA=0) of 116 people with ‘Infection Not Probable at time of Admission’ survived.
    - 60 (INF=1 and STA=0) of 84 people with ‘Infection Probable at time of Admission’ survived.

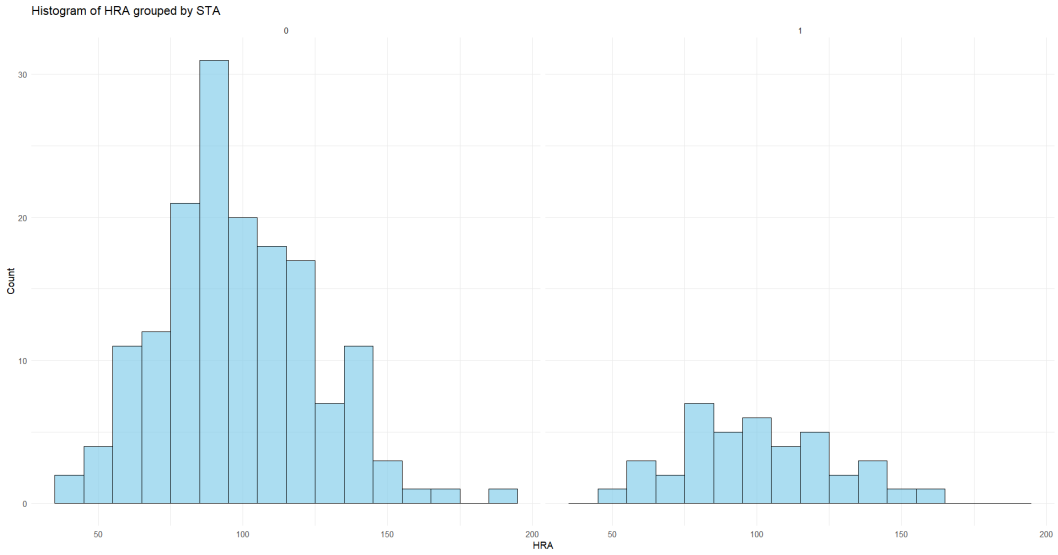


7. Contingency Table of CPR and STA:
    - 154 (CPR=0 and STA=0) of 188 people with ‘No CPR Prior to Admission’ survived.
    - 6 (CPR=1 and STA=0) of 13 people with ‘CPR Prior to Admission’ survived.
8. Contingency Table of PRE and STA:
    - 137 (PRE=0 and STA=0) of 170 people with ‘No Previous Admission to ICU in 6 months’ survived.
    - 23 (PRE=1 and STA=0) of 30 people with ‘Previous Admission to ICU in 6 months’ survived.
9. Contingency Table of TYP and STA:
    - 51 (TYP=0 and STA=0) of 53 people with ‘Elective Admission’ survived.
    - 109 (TYP=1 and STA=0) of 147 people with ‘Emergency Admission’ survived.
10. Contingency Table of FRA and STA:
    - 148 (FRA=0 and STA=0) of 185 people with ‘No Fracture’ survived.
    - 12 (FRA=1 and STA=0) of 15 people with ‘Fracture’ survived.
11. Contingency Table of PO2 and STA:
    - 149 (PO2=0 and STA=0) of 184 people with ‘PO2 from Initial Blood Gases ≥ 60’ survived.
    - 11 (PO2=1 and STA=0) of 16 people with ‘PO2 from Initial Blood Gases < 60’ survived.
12. Contingency Table of PH and STA:
    - 151 (PH=0 and STA=0) of 187 people with ‘PH from Initial Blood Gases ≥ 7.25’ survived.
    - 9 (PH=1 and STA=0) of 13 people with ‘PH from Initial Blood Gases < 7.25’ survived.



13. Contingency Table of PCO and STA:
    - 144 (PCO=0 and STA=0) of 180 people with ‘PCO2 from Initial Blood Gases ≤ 45’ survived.
    - 16 (PCO=1 and STA=0) of 4 people with ‘PCO2 from Initial Blood Gases > 45’ survived.
14. Contingency Table of BIC and STA:
    - 150 (BIC=0 and STA=0) of 185 people with ‘Bicarbonate from Initial Blood Gases ≥ 18’ survived.
    - 10 (BIC=1 and STA=0) of 15 people with ‘Bicarbonate from Initial Blood Gases < 18’ survived.
15. Contingency Table of CRE and STA:
    - 155 (CRE=0 and STA=0) of 190 people with ‘Creatinine from Initial Blood Gases ≤ 2’ survived.
    - 5 (CRE=1 and STA=0) of 5 people with ‘Creatinine from Initial Blood Gases > 2’ survived.
16. Contingency Table of LOC and STA:
    - 158 (LOC=0 and STA=0) of 185 people ‘Not in Coma’ survived.
    - 0 (LOC=1 and STA=0) of 5 people ‘In Deep Stupor’ survived.
    - 2 (LOC=2 and STA=0) of 10 people ‘In Coma’ survived.

```
> ################################################################################
> # Explore relationship of STA with continuous variables
>
> # Loop through continuous columns and plot histograms
> for(var in continuous_cols){
+   p <- ggplot(icu, aes_string(x = var)) +
+     geom_histogram(binwidth=10, fill="skyblue", color="black", alpha=0.7) + # adjust binwidth as needed
+     facet_wrap(~STA) +
+     labs(title = paste("Histogram of", var, "grouped by STA"), x = var, y = "Count") +
+     theme_minimal()
+   print(p)
+ }
>
```

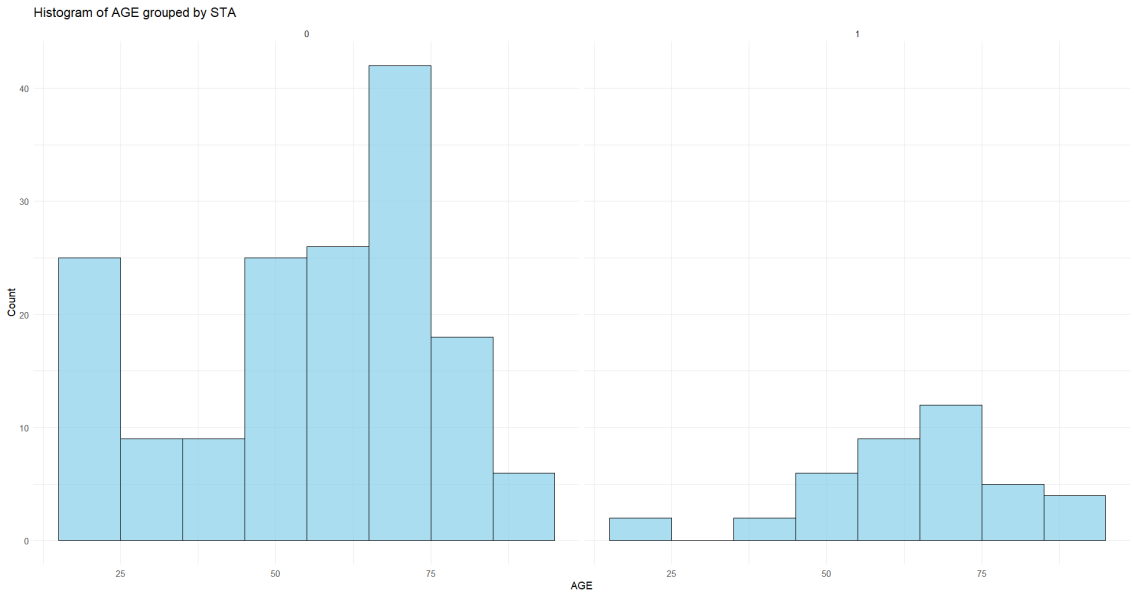


Group 0(Survived):

- The distribution for this group has a prominent peak around the 'Heart Rate at time of Admission to ICU (HRA)' value of 80-100.
- The majority of the data for Group 0 falls within the range of roughly 60 to 140.
- The distribution shows a rapid decline as the HRA values increase beyond 140 or decrease below 80. This indicates that such values are less common for this group.

Group 1(Died):

- The distribution for Group 1 is much flatter compared to Group 0, meaning there's a more even spread of data across different HRA values.
- The HRA values for Group 1 mainly range between 50 to 160, but with no sharp peak like Group 0.

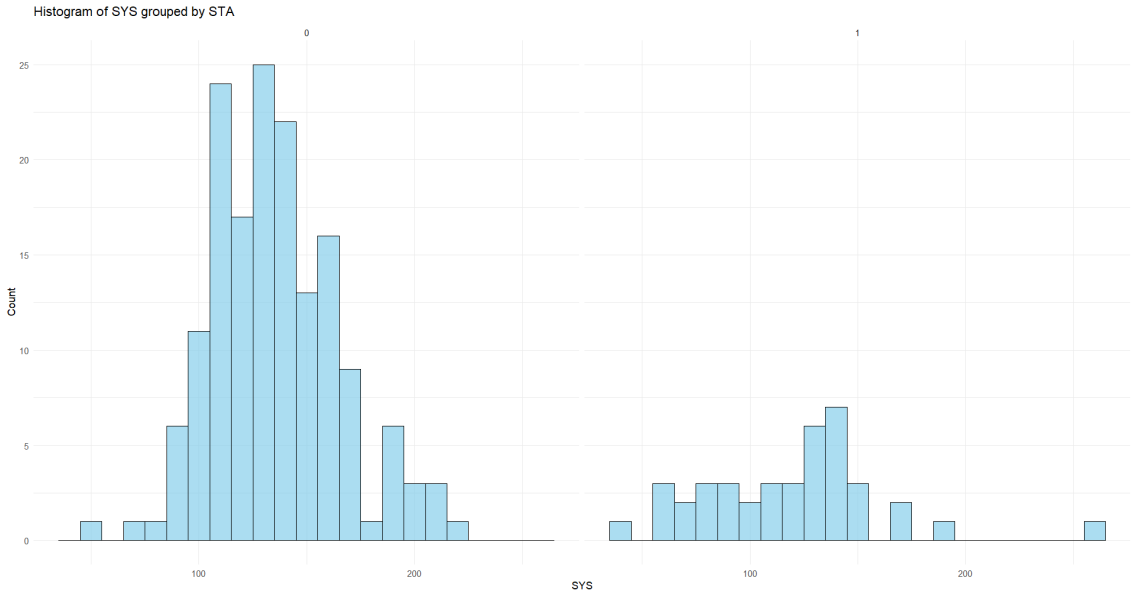


Group 0(Survived):

- The age distribution for this group has its highest peak around the age of 65-75.
- There is a notable dip in the age range of 25-45, indicating fewer individuals in this age bracket for Group 0.
- The majority of the data for Group 0 falls within the age range of 50 to 75, with a gradual decline as the age goes beyond 70 or drops below 50, except for a spike for ages less than 25.

Group 1(Died):

- The distribution for Group 1 is more scattered compared to Group 0, with peaks and valleys at the same intervals, but these peaks and valleys are much shorted than for Group 0.
- The most prominent peak for Group 1 is around the age range of 65-75.
- There is a visible dip or fewer counts in the age range of 25-45, similar to Group 0.



Group 0 (Survived):

- The distribution for this group is concentrated around the 'Systolic Blood Pressure mm Hg at ICU Admission (SYS)' value of approximately 100-170.
- The highest peak in Group 0 occurs around the SYS value of 140.
- The SYS values for this group predominantly fall within the range of 90 to 180.

Group 1(Died):

- The distribution for Group 1 is more scattered compared to Group 0.
- The most significant peak for Group 1 is around the SYS value of 140-150.
- SYS values for Group 1 predominantly fall within the range of 60 to 150. While there are instances of higher SYS values, they are relatively rare and less frequent than in Group 0.

```
> ##############################################################################
>
> # Count levels for each discrete variable
> level_counts <- lapply(discrete_cols, function(var) {
+   data_frame(Variable = var,
+              CountOfLevels = length(unique(icu[[var]])))
+ })
Warning message:
`data_frame()` was deprecated in tibble 1.1.0.
i Please use `tibble()` instead.
This warning is displayed once every 8 hours.
Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
>
> # Bind all data frames together
> level_counts_df <- bind_rows(level_counts)
>
> # Print the result
> print(level_counts_df)
# A tibble: 16 x 2
   Variable CountOfLevels
   <chr>            <int>
 1 SEX                  2
 2 RACE                 3
 3 SER                  2
 4 CAN                  2
 5 CRN                  2
 6 INF                  2
 7 CPR                  2
 8 PRE                  2
 9 TYP                  2
10 FRA                  2
11 PO2                  2
12 PH                   2
13 PCO                  2
14 BIC                  2
15 CRE                  2
16 LOC                  3
```

```
> # Loop through discrete columns and print the frequency of each level
> for(col in discrete_cols) {
+   cat("Frequency for", col, ":\n")
+   print(table(icu[[col]]))
+   cat("\n")
+ }
Frequency for SEX :

  0   1
124  76

Frequency for RACE :

  1   2   3
175  15  10

Frequency for SER :

  0   1
 93 107

Frequency for CAN :

  0   1
180  20

Frequency for CRN :

  0   1
181  19

Frequency for INF :

  0   1
116  84

Frequency for CPR :

  0   1
187  13

Frequency for PRE :

  0   1
170  30

Frequency for TYP :

  0   1
 53 147

Frequency for FRA :

  0   1
185  15

Frequency for PO2 :

  0   1
184  16

Frequency for PH :

  0   1
187  13

Frequency for PCO :

  0   1
180  20

Frequency for BIC :

  0   1
185  15
```

```
Frequency for CRE :

  0   1
190  10

Frequency for LOC :

  0   1   2
185   5  10
```

```
> ###############################################################################
> # Drop records with number of records < 10 for any level of a discrete variable
> # Function to check if any level has less than 10 records
> drop_rows <- function(data, col) {
+   tbl <- table(data[[col]])
+   drop_levels <- names(tbl[tbl < 10])
+   if (length(drop_levels) > 0) {
+     return(!(data[[col]] %in% drop_levels))
+   }
+   return(rep(TRUE, nrow(data)))
+ }
>
> # Apply the function for each discrete column
> for (col in discrete_cols) {
+   icu <- icu[drop_rows(icu, col), ]
+ }
>
> dim(icu)
[1] 195  21
>
```

Crop Down Condition: We have tried to drop all the records whose number of records less than 10 (i.e., 5% of total records), but only 5 records with Level of Consciousness (LOC = 1) at 'Deep Stupor' met the criterion and were dropped.

Population of Interest: Intensive Care Unit (ICU) Patients:

We now have a total of 195 records, with 21 columns, of which the first column is ID, and the Response Variable is Vital Status (STS) is our Y. We have a total of 19 independent variables.

The study focuses on a cohort of patients admitted to the Intensive Care Unit (ICU). The data set captures a range of variables that provide insights into the patients' demographics, medical history, and clinical parameters during their ICU admission.

Demographic Information:

- Identification Code (ID): A unique identifier assigned to each patient.
- Age (AGE): The age of the patients in years.
- Sex (SEX): Patients are categorized as Male or Female.
- Race (RACE): Patients are grouped into three racial categories White, Black, or Other.

Clinical Information:
Service at ICU Admission: This variable distinguishes between patients admitted for medical reasons versus those admitted for surgical reasons.
Present Problem and Medical History:

- Vital Status (STA)
- Cancer as a present problem (CAN)
- History of chronic renal failure (CRN)
- History of previous ICU admissions within the last six months (PRE)
- Infection probable at the time of ICU admission (INF)
- Cardiopulmonary resuscitation (CPR) prior to ICU admission (CPR)
- Type of ICU admission, categorized as elective or emergency (TYP)
- Presence of fractures, specified as long bone, multiple, neck, single area, or hip fracture (FRA)

Clinical Parameters at ICU Admission:

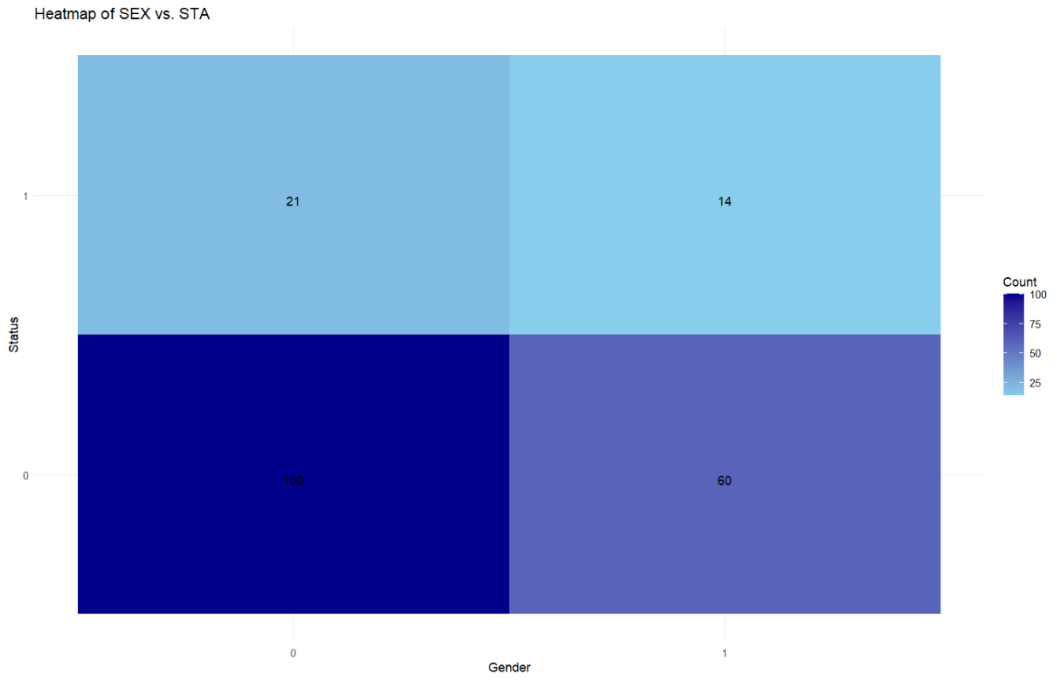
- Blood Pressure Measurements: Systolic blood pressure at admission. (SYS)
- Heart Rate: Measured in beats per minute. (HRA)
- Blood Gas Measurements: These provide insights into a patient's respiratory and metabolic status. They include:
    - PO2: Partial pressure of oxygen
    - PH: A measure of the acidity or alkalinity of the blood
    - PCO2: Partial pressure of carbon dioxide
    - BIC: Bicarbonate (A measure related to the buffering capacity of the blood)
    - CRE: Creatinine (A renal function parameter)
- Level of Consciousness: This variable evaluates the neurological status of the patient, categorized into three levels – no coma or deep stupor, deep stupor, and coma. (LOC)

In summary, the population of interest in this study comprises ICU patients, and the data set aims to shed light on a wide range of demographic and clinical variables that might influence their outcomes during their ICU stay. The richness of these variables can provide valuable insights for clinicians and researchers aiming to improve patient care and outcomes in the ICU setting.

Task 2

Obtain a 2x2 contingency table that relates gender (SEX) to Status (STA). Determine the odds and the probabilities of survival among males and females. Then compute the odds ratio of survival that compares males to females. Does anything seem interesting here?

```
> # Create the contingency table
> table_sex_sta <- as.data.frame(table(icu$SEX, icu$STA))
> names(table_sex_sta) <- c("SEX", "STA", "Count")
>
> # Plot the heatmap
> p <- ggplot(table_sex_sta, aes(x = SEX, y = STA, fill = Count)) +
+   geom_tile() +
+   geom_text(aes(label = Count), vjust = 1) +
+   scale_fill_gradient(low = "skyblue", high = "darkblue") +
+   labs(title = "Heatmap of SEX vs. STA", x = "Gender", y = "Status") +
+   theme_minimal()
>
> print(p)
>
```



Vital Status (STA)

- 0 = Lived
- 1= Died

Sex (SEX)

- 0 = Male
- 1 = Female

1. Probabilities of Survival Among Males and Females:
   - P_{male} is the probability that a male survives, and $P_{female}$is the probability that a female survives.
   - From the heatmap:
     - Number of males that survived (SEX = 0, STA = 0) = 100
     - Total number of males = 21 (not survived) + 100 (survived) = 121
     - Number of females that survived (SEX = 1, STA = 0) = 60
     - Total number of females = 14 (not survived) + 60 (survived) = 74
     - $P_{male} = \frac{100}{121} = 0.8264$
     - $P_{female} = \frac{60}{74} = 0.8108$
2. Odds of Survival Among Males and Females:
   - Odds of survival for males $= \frac{P_{male}}{1- P_{male}} = \frac{0.8264}{1-0.8264} = 4.760$
   - Odds of survival for females $= \frac{P_{female}}{1- P_{female}} = \frac{0.8108}{1-0.8108} = 4.285$

3. Odds Ratio:
   - $OR = \frac{Odds\ of\ Survival\ of\ Males}{Odds\ of\ Survival\ of\ Feamles} = \frac{4.760}{4.285} = 1.1108$

Interpretation:

1. Survival Rates by Gender: Despite the general perception that females might have better survival outcomes in certain medical situations, in this ICU dataset, the difference in survival probabilities between males and females is minimal. Both genders have a survival probability over 80%, indicating a low mortality rate in this ICU setting.

2. Odds Ratio Close to 1: An odds ratio close to 1 suggests that there's little difference in the odds of survival between the two groups being compared (in this case, males vs. females). The odds ratio of 1.1108 indicates that males have roughly 111% of the odds of survival compared to females. In practical terms, this difference might not be clinically significant, especially without knowing the margin of error or the confidence interval around this estimate.

3. Low Mortality Rate: The majority of patients in the ICU, regardless of gender, survived. This might indicate a not particularly severe patient population.

4. Dataset Composition: There are more males (121) than females (74) in this dataset. This uneven distribution might influence the reliability of comparative statistics and could reflect specific population demographics, admission criteria, or other external factors.

5. Implications for Further Analysis:

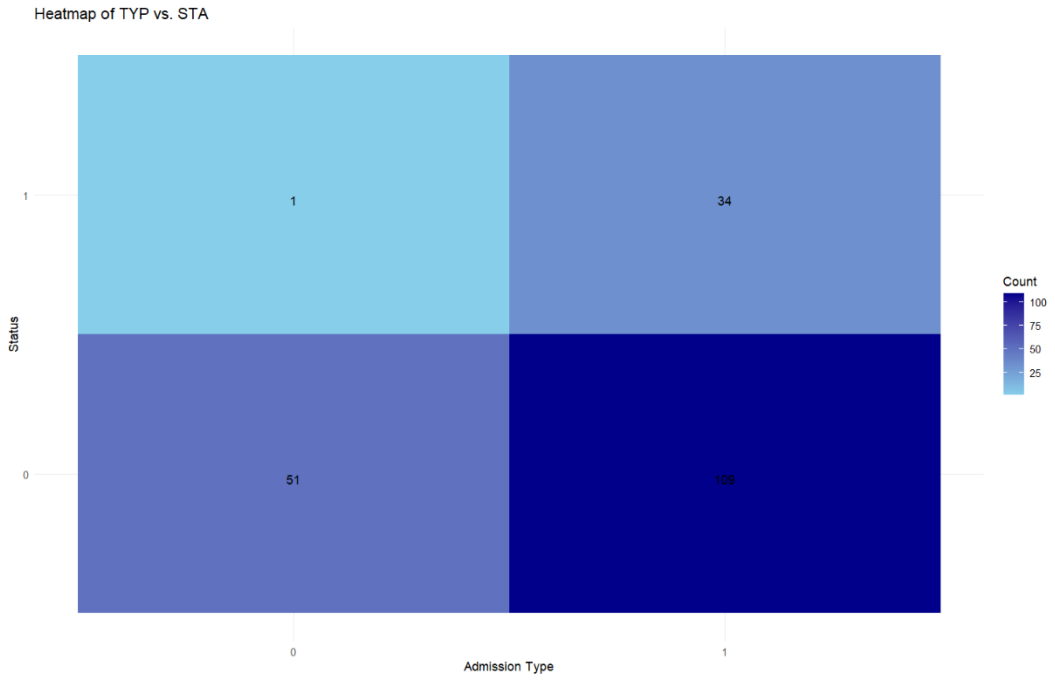
- It would be essential to conduct hypothesis testing (like a chi-squared test) to determine if the observed differences in survival between males and females are statistically significant.
- It would also be interesting to investigate other variables' effects on survival, either individually or in combination (multivariate analysis). For instance, does the type of service at ICU admission (medical vs. surgical) influence survival differently for males and females?

In conclusion, while the odds ratio provides a quantified comparison of survival odds between males and females, the real-world implications of these findings would require a deeper dive, considering clinical significance, statistical significance, and other contextual factors.

Task 3

Obtain a 2x2 contingency table that relates Type of Admission (TYP) to Status (STA). Again, determine the odds and probabilities of survival among the different Types of Admission. Then compute and interpret the odds ratio of survival that compare them.

```
> # Create the contingency table
> table_typ_sta <- as.data.frame(table(icu$TYP, icu$STA))
> names(table_typ_sta) <- c("TYP", "STA", "Count")
>
> # Plot the heatmap
> p <- ggplot(table_typ_sta, aes(x = TYP, y = STA, fill = Count)) +
+   geom_tile() +
+   geom_text(aes(label = Count), vjust = 1) +
+   scale_fill_gradient(low = "skyblue", high = "darkblue") +
+   labs(title = "Heatmap of TYP vs. STA", x = "Admission Type", y = "Status") +
+   theme_minimal()
>
> print(p)
```



Vital Status (STA)

- 0 = Lived
- 1 = Died

Type of Admission (TYP)

- 0 = Elective
- 1 = Emergency

4. Probabilities of Survival Among Elective and Emergency:
   - P_{elec} is the probability that an elective admission survives, and P_{emer} is the probability that an emergency admission survives.

- From the heatmap:
    - Number of elective admissions that survived (TYP = 0, STA = 0) = 51
    - Total number of elective admissions = 51 (survived) + 1 (not survived) = 52
    - Number of emergency admissions that survived (TYP = 1, STA = 0) = 109
    - Total number of emergency admissions = 109 (survived) + 34 (not survived) = 143
    - $P_{elec} = \frac{51}{52} = 0.9808$
    - $P_{emer} = \frac{109}{143} = 0.7622$

5. Odds of Survival among Elective and Emergency Admissions:
    - Odds of survival for elective admissions $= \frac{P_{elec}}{1 - P_{elec}} = \frac{0.9808}{1-0.9808} = 51.083$
    - Odds of survival for emergency admissions $= \frac{P_{emer}}{1-P_{emer}} = \frac{0.7622}{1-0.7622} = 3.206$
6. Odds Ratio:
    - $OR = \frac{Odds\ of\ Survival\ of\ Elective\ Admissions}{Odds\ of\ Survival\ of\ Emergency\ Admissions} = \frac{51.083}{3.206} = 15.934$

Interpretation:

1. High Probability of Survival for Elective Admissions: The probability that someone admitted electively survives is notably high (98.08%). This might indicate that elective procedures or admissions, which are typically planned and non-urgent, carry lower risks or are prepared for in advance, leading to better patient outcomes.
2. Lower Survival Rate for Emergency Admissions: The probability that someone admitted in an emergency survives is lower (76.22%) compared to elective admissions. This can be expected as emergency admissions usually result from sudden and unforeseen circumstances, where patients might be in a more critical condition.
3. Odds Ratio: The odds ratio of 15.934 is particularly significant. An odds ratio greater than 1 indicates that the odds of survival are higher for the first group (in this case, elective admissions) than the second group (emergency admissions). The value of 15.934 implies that the odds of survival for patients admitted electively are almost 16 times higher than those admitted on an emergency basis. This large difference highlights the impact of the type of admission on survival outcomes.

Task 4

Suppose the patient's AGE is considered to be a key determinant of the patient's survival. With this information, complete the following:

a. Fit a logistic regression model to predict STA using the original continuous AGE variable. Report and interpret the coefficients for the model.

```
> # Fitting the logistic regression model
> model <- glm(STA ~ AGE, data=icu, family=binomial)
>
> # Displaying the summary of the model to see the coefficients and statistics
> summary(model)

Call:
glm(formula = STA ~ AGE, family = binomial, data = icu)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.29798    0.74926  -4.402 1.07e-05 ***
AGE          0.02913    0.01127   2.585  0.00975 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 183.54  on 194  degrees of freedom
Residual deviance: 175.70  on 193  degrees of freedom
AIC: 179.7

Number of Fisher Scoring iterations: 5
```

1. Coefficients:

- The `(Intercept)` has an estimate of -3.29798. This value represents the log odds of STA being 1 (presumably representing death) when AGE is zero. However, an AGE of zero is not meaningful in this context, so the intercept is not interpretable by itself.
- The `AGE` coefficient is 0.02913, indicating that for each additional year of AGE, the log odds of STA being 1 (death) increases by 0.02913, when all other variables in the model are held constant. This is assuming AGE is measured in years.

2. Standard Error (Std. Error):
   - The standard error of the intercept is 0.74926, and for AGE, it is 0.01127. These values measure the variability or precision of the coefficient estimates.

3. z-value:
   - The z-values are the ratios of the coefficients to their standard errors. For AGE, a z-value of 2.585 suggests that the coefficient is 2.585 standard deviations away from 0.

4. Pr(>|z|):
   - These are the p-values associated with the z-tests. A p-value of 1.07e-05 for the intercept and 0.00975 for AGE suggests that both coefficients are statistically significant (as indicated by the stars next to the p-values), meaning there is strong evidence against the null hypothesis of the coefficients being zero.

5. Dispersion Parameter:
   - For the binomial family, the dispersion parameter is taken to be 1 by default.

6. Deviance:
   - The null deviance is the goodness of fit measure for a model with only the intercept (no predictors), and it's 183.54 on 194 degrees of freedom.
   - The residual deviance is the goodness of fit for the model with predictors, in this case just AGE, and it's 175.70 on 193 degrees of freedom. The difference between the null and residual deviance shows that AGE does explain some of the variability in STA.
   - Lower deviance indicates a better fit of the model to the data.

10. Number of Fisher Scoring iterations:
   - This indicates the number of iterations the algorithm took to converge to the best-fit values. In this case, it took 5 iterations, which is within normal expectations.

In summary, AGE is a statistically significant predictor of STA in this logistic regression model, with a positive relationship between AGE and the log odds of STA being 1. The model appears to be a reasonable fit to the data, as evidenced by the decrease in deviance from the null model to the model including AGE and the significance of the AGE coefficient.

b. Write the equation for the logistic regression model of STA (Y) using AGE (X). Write the equation for the logit transformation of this logistic regression model.

The logistic regression model can be described by the following equation, where $P(Y = 1)$ represents the probability that the dependent variable Y (in this case, STA) equals 1 for a given value of X (in this case, AGE):

$$P(Y = 1) = \frac{e^{\beta 0 + \beta 1 X}}{1 + e^{\beta 0 + \beta 1 X}}$$

$$\beta 0(Intercept) = -3.29798$$

$$\beta 1(AGE) = 0.02913$$

So, the logistic regression model for STA using AGE is:

$$P(STA = 1) = \frac{e^{-3.29798 + 0.02913 * AGE}}{1 + e^{-3.29798 + 0.02913 * AGE}}$$

The logit transformation is given by:

$$\log\left(\frac{P(STA = 1)}{1 - P(STA = 1)}\right) = \beta 0 + \beta 1 X$$

Substituting our coefficients into this equation gives us the logit model for STA using AGE:

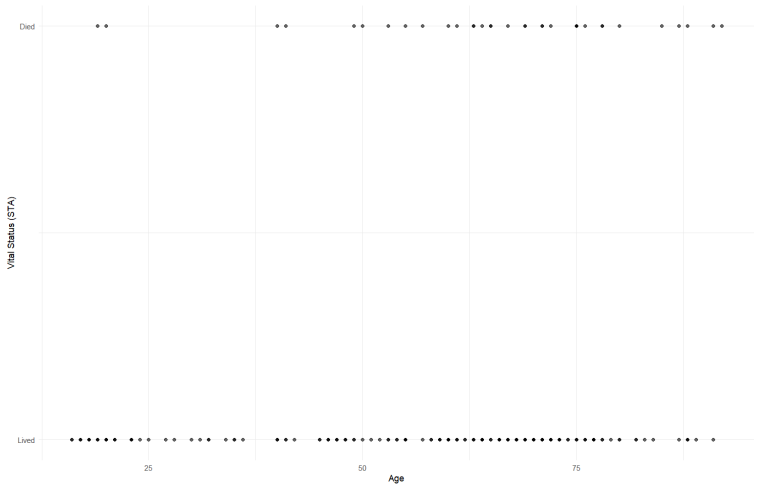
$$\log\left(\frac{P(STA = 1)}{1 - P(STA = 1)}\right) = -3.29798 + 0.02913 * AGE$$

This logit equation represents the linear relationship between the log-odds of the outcome and the continuous predictor AGE.

This equation gives us a model for calculating the probability of STA being 1 for any given AGE. Because the relationship is logarithmic, it's not a linear increase in probability; instead, the odds of STA (odds of death) increase by a factor of $e^{0.02913}$ (1.02956) for each additional year of age. This means that for every one-year increase in AGE, the odds of death increase by about 2.956%.

c. Make a scatterplot of STA (Y) by AGE(Y). Does Age seem to be a good discriminator between levels of STA?

```
> # Load the ggplot2 package
> library(ggplot2)
>
> # Create a scatterplot
> scatter_plot <- ggplot(icu, aes(x = AGE, y = STA)) +
+   geom_point(alpha = 0.6) +          # alpha is used for point transparency
+   labs(x = "Age", y = "Vital Status (STA)") +
+   theme_minimal() +
+   scale_y_continuous(labels = c("Lived", "Died"), breaks = c(0, 1))
>
> # Display the plot
> print(scatter_plot)
```



Age does appear to have some discriminatory power between the levels of STA, especially when considering older age groups, because:

1. Higher Mortality in Elderly: There's a noticeable concentration of points in the "Died" level as the age increases, particularly past the age of 50. This suggests that the likelihood of death (STA = 1) might be higher for older individuals.

2. Lower Mortality in Younger Population: Fewer points are observed in the "Died" level for ages below 50 compared to the "Lived" level, indicating that younger individuals might have a lower risk of death.
3. Overlap in Middle Ages: However, there's some overlap around middle ages (around 50), where both "Lived" and "Died" levels have data points. This indicates that age might not be as strong a discriminator in this age bracket.

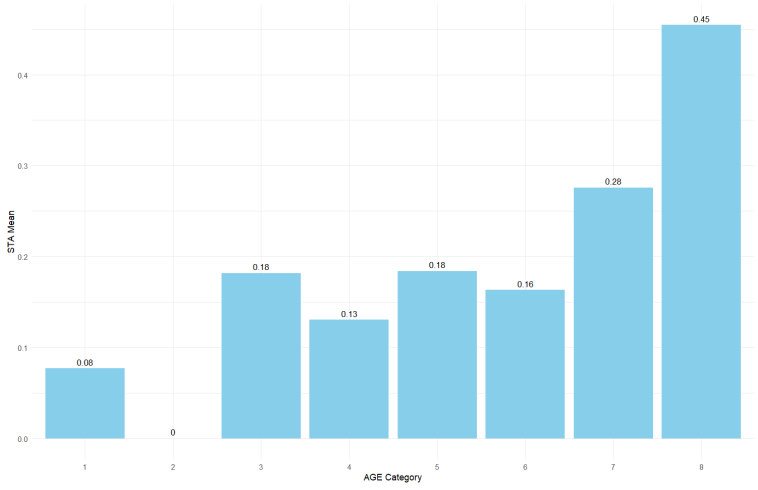
In conclusion, while age seems to have some discriminatory power, especially for the younger and older age groups, it may not be a perfect or sole discriminator between the levels of STA. Other factors and variables might also play a significant role in determining STA.

d. Construct a new categorical variable by discretizing AGE into the following intervals:

i. AGE_CAT = 1 if AGE is in the interval [15,24]
ii. AGE_CAT = 2 if AGE is in the interval [25,34]
iii. AGE_CAT = 3 if AGE is in the interval 3 = [35,44]
iv. AGE_CAT = 4 if AGE is in the interval 4 = [45,54]
v. AGE_CAT = 5 if AGE is in the interval 5 = [55,64]
vi. AGE_CAT = 6 if AGE is in the interval 6 = [65,74]
vii. AGE_CAT = 7 if AGE is in the interval 7 = [75,84]
viii. AGE_CAT = 8 if AGE is in the interval 8 = [85,94]
ix. AGE_CAT = 9 if AGE is in the interval 9 = 95 and over

Using this categorical variable, compute the STA mean (i.e. proportion) over subjects in the age interval. Plot these means versus the categorical variable.

```
> # Creating AGE_CAT variable
> icu <- icu %>%
+   mutate(AGE_CAT = case_when(
+     AGE >= 15 & AGE <= 24 ~ 1,
+     AGE >= 25 & AGE <= 34 ~ 2,
+     AGE >= 35 & AGE <= 44 ~ 3,
+     AGE >= 45 & AGE <= 54 ~ 4,
+     AGE >= 55 & AGE <= 64 ~ 5,
+     AGE >= 65 & AGE <= 74 ~ 6,
+     AGE >= 75 & AGE <= 84 ~ 7,
+     AGE >= 85 & AGE <= 94 ~ 8,
+     AGE >= 95         ~ 9
+   ))
>
> # Compute STA mean for each age category
> age_summary <- icu %>%
+   group_by(AGE_CAT) %>%
+   summarize(STA_mean = mean(STA, na.rm = TRUE))
>
> # Plotting
> ggplot(age_summary, aes(x = as.factor(AGE_CAT), y = STA_mean)) +
+   geom_bar(stat = "identity", fill = "skyblue") +
+   geom_text(aes(label = round(STA_mean, 2)), vjust = -0.5, size = 3.5) +
+   labs(x = "AGE Category", y = "STA Mean") +
+   theme_minimal()
```



Observations:

- There is a general trend where the mortality rate is relatively stable across the middle age categories (35-74) with slight variations.
- The highest mortality rates are observed in the oldest age groups, with nearly half of the individuals in the age bracket of 85-94 passing away. This is consistent with general expectations, as mortality rates typically increase with age.
- The 25-34 age group lacks data, which might indicate a lack of representation from this age group or a zero-mortality rate for them.

In conclusion, the data underscores that mortality rates tend to be higher in the oldest age groups, with the 85-94 age bracket showing the highest mortality rate among all categories.

e. Report and interpret all hypothesis test results. What do you conclude?

The hypothesis tests in the logistic regression model output relate to whether the coefficients for the intercept and the AGE variable are significantly different from zero. Here's the interpretation of the hypothesis test results:

1. Intercept (-3.29798):

- Estimate: The estimated coefficient for the intercept is -3.29798. This represents the log-odds of STA being 1 when AGE is zero. However, since AGE being zero is not meaningful in this context, the intercept here is not interpretable by itself.
- Standard Error (0.74926): This is the standard error of the estimate of the intercept.
- z value (-4.402): This is the test statistic, calculated as the estimate divided by the standard error. A negative value indicates that the estimate is below zero.
- P-value (1.07e-05): The probability of observing such an extreme test statistic under the null hypothesis that the intercept's true value is zero. A value of 1.07e-05 (which is less than 0.0001) is highly significant, meaning there is strong evidence against the null hypothesis.

2. AGE (0.02913):

- Estimate: The estimated coefficient for AGE is 0.02913. This represents the change in the log-odds of STA being 1 for a one-unit increase in AGE.
- Standard Error (0.01127): This is the standard error of the estimate of the AGE coefficient.
- z value (2.585): The test statistic for AGE. A positive value suggests that the estimate is above zero.
- P-value (0.00975): The probability of observing such an extreme test statistic under the null hypothesis that the AGE's true coefficient is zero. A p-value of 0.00975 (which is less than 0.01) indicates that the evidence is strong against the null hypothesis.

Interpretation:

- The intercept being significantly less than zero suggests that when AGE is at the reference level (zero in this model, but not interpretable), the log-odds of STA being 1 is negative, implying low odds of STA being 1.
- The AGE coefficient being significantly greater than zero suggests that as AGE increases, the log-odds of STA being 1 also increase. For each additional year of AGE, the odds of STA being 1 increase by a multiplicative factor of approximately 1.02956, or an increase in odds of about 2.956%.

Conclusion:

Given the significance of the AGE coefficient, we can conclude that there is a statistically significant association between AGE and the likelihood of STA being 1. As AGE increases, the probability of STA (dying) also increases, after controlling for other variables in the model.

f. Report the AIC and BIC values. What is the value of the deviance for the fitted model?

```
Call:
glm(formula = STA ~ AGE, family = binomial, data = icu)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.29798    0.74926  -4.402 1.07e-05 ***
AGE          0.02913    0.01127   2.585  0.00975 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 183.54  on 194  degrees of freedom
Residual deviance: 175.70  on 193  degrees of freedom
AIC: 179.7

Number of Fisher Scoring iterations: 5

> model <- glm(formula = STA ~ AGE, family = binomial, data = icu)
> BIC(model)
[1] 186.2442
```

The AIC (Akaike Information Criterion) value for the model is 179.7, and the BIC (Bayesian Information Criterion) value is 186.2442

Regarding deviance, there are two types reported:

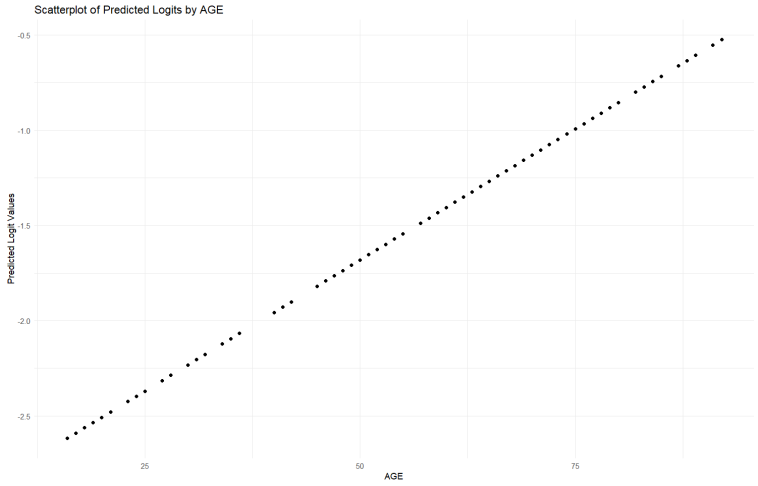
- Null deviance: 183.54 on 194 degrees of freedom
- Residual deviance: 175.70 on 193 degrees of freedom

The null deviance represents the difference between a model with only the intercept (no predictors) and the saturated model (a theoretical model with a perfect fit). The residual deviance represents the difference between the model (with the AGE predictor) and the saturated model.

The lower the deviance, the better the model fits the data. In the case, the residual deviance is lower than the null deviance, which indicates that the model with the AGE predictor fits the data better than a model with no predictors.

g. Use the fitted model to predict logit values for each record in the dataset. Save the logits to your analysis file. Then make a scatterplot of the predicted logits(Y) by AGE (X). Discuss the scatterplot.

```
> ## Task 4g
>
> # Computing tghe Logit Values
> logit_values <- predict(model, type = "link")
>
> # Saving the values in the dataframe
> icu$logit_values <- predict(model, type = "link")
>
>
> # Create a scatterplot
> ggplot(icu, aes(x = AGE, y = logit_values)) +
+   geom_point() +
+   labs(x = "AGE", y = "Predicted Logit Values",
+        title = "Scatterplot of Predicted Logits by AGE") +
+   theme_minimal()
```



Interpretation:

1. Linear Trend: The plot showcases a clear linear trend. As the age increases, the predicted logit values also increase. This aligns with the positive coefficient for `AGE` from the logistic regression model, suggesting that as age increases, the log odds of the event (in this case, death as indicated by `STA = 1`) also increase.

2. Range of AGE: The data points start from a young age, somewhere around early 20s, and extend up to late 70s or early 80s.

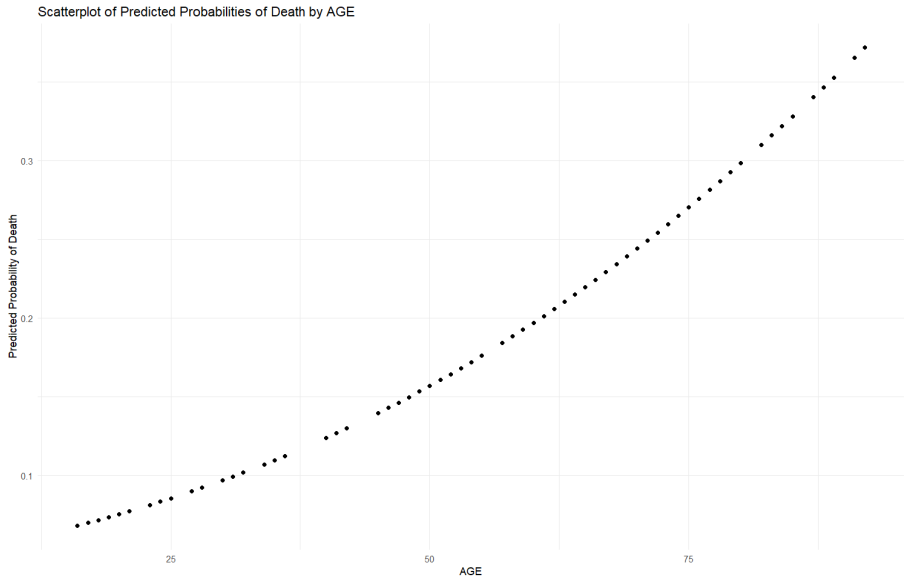
3. Predicted Logit Values: The range of logit values is from about -2.5 to -0.5. A more negative logit value indicates lower odds of the event occurring, while a value closer to zero suggests higher odds. As seen in the plot, younger individuals have more negative logit values, indicating lower odds of death, while older individuals have logit values closer to zero, indicating higher odds.

4. Data Distribution: The data points seem to be evenly distributed across different age groups, without any obvious gaps or clusters. This evenly spread data provides a more generalizable model.

5. Uniformity: The points lie along a straight line without much deviation, indicating that age is a strong predictor for the outcome in this dataset. There isn't much scatter around the line, suggesting low variability in the predictions across the age range, possibly because AGE is the only factor in the model.

h. Write a line or two or three of R-code to compute the probabilities of survival (pi) from the logits. Save the predicted probabilities to your analysis file. Then make a scatterplot of the predicted probabilities (Y) by AGE (X). Do you see the typical 'S' shaped logistic curve? If possible, overlay the raw data of Y=STA on top of your predicted values of probability of Survival.

```
> ## Task 4h
>
>
> # Compute the probabilities from the logits
> icu$probability_death <- exp(icu$logit_values) / (1 + exp(icu$logit_values))
>
> # Scatterplot of the predicted probabilities by AGE
> ggplot(icu, aes(x = AGE, y = probability_death)) +
+   geom_point() +
+   labs(x = "AGE", y = "Predicted Probability of Death",
+        title = "Scatterplot of Predicted Probabilities of Death by AGE") +
+   theme_minimal()
```



Linear Relationship with Age: There's a clear positive relationship between age and the predicted probability of death. As age increases, the probability of death also increases. This relationship appears linear for the majority of the age range, with a steeper incline observed for older ages.

Lower Risk for Younger Individuals: For ages roughly below 50, the predicted probabilities of death are relatively low, all below 0.2. This suggests that younger individuals are less likely to die, according to the model's predictions.

Increased Risk with Age: Starting from age 50 and onwards, there's a more pronounced increase in the predicted probability of death. By the age of 75, the probability nears 0.3, which is a significant increase compared to younger ages. This could go higher and flatter, if we had more datapoints for ages between 80 and 100, and vice-versa if we had more datapoints for ages between 0 and 20.

Steep Increase for Older Ages: The curve gets steeper as age increases, particularly after age 65 or so. This indicates that for each additional year of age, the increase in the probability of death becomes larger. This might suggest an accelerating risk as one gets into the later stages of life.

Data Distribution: The even distribution of data points across ages indicates that the dataset likely has a good representation of various age groups. This is important for the reliability of the model's predictions across different age categories.

In summary, the scatterplot illustrates that as individuals age, their predicted risk of death increases. The model behind this data seems to suggest a particularly pronounced risk for older individuals, especially those above 65 years of age.

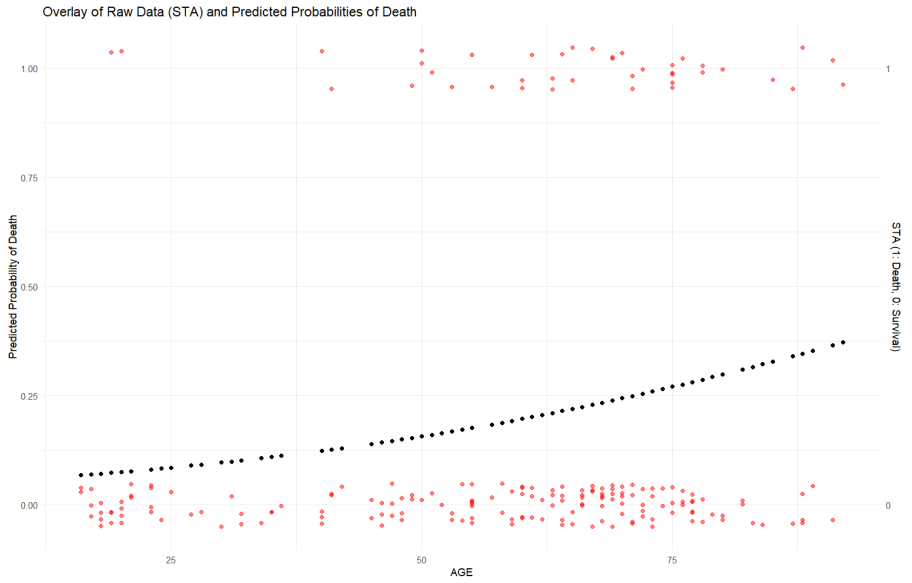
The expected S-shaped curve (Sigmoid curve) is often associated with logistic regression, especially when predicting probabilities. However, this plot does not exhibit this S-shape, possibly because:

Limited Range of Independent Variable: The S-shape is more pronounced when the independent variable (in this case, AGE) covers the full range where the changes in the probability occur. If the dataset is restricted to certain ages, one might only capture a segment of the curve, and this is what seems to be happening here. This model has been built with only 195 datapoints, which could possibly mean that the model is not robust enough.

Transformations and Interactions: The relationship between age and the probability of death might not be purely logistic. There could be transformations of the age variable (e.g., polynomial terms) or interaction terms included in the model that adjust the shape of the curve.

Interaction or Polynomial Terms: The logistic regression model might include interaction terms or polynomial terms. For instance, adding squared or cubic terms of the predictor variable could modify the shape of the curve.

```
> # Adding the origianl STA values to the plot
>
> ggplot(icu, aes(x = AGE, y = probability_death)) +
+   # Predicted probability of death
+   geom_point(aes(y = probability_death)) +
+
+   # Raw data: Jitter is added for the STA values to better visualize overlapping points
+   geom_jitter(aes(y = STA), color = "red", position = position_jitter(width = 0, height = 0.05), al
pha = 0.5) +
+
+   labs(x = "AGE", y = "Predicted Probability of Death", title = "Overlay of Raw Data (STA) and Pred
icted Probabilities of Death") +
+   scale_y_continuous(
+     "Predicted Probability of Death",
+     sec.axis = sec_axis(~., name = "STA (1: Death, 0: Survival)", breaks = c(0,1))
+   ) +
+   theme_minimal()
```



STA Distribution:

- For younger ages (less than 50), most of the red dots (actual outcomes) are clustered at the bottom, showing a higher number of survivals.
- For older ages (greater than 50), there's an increase in the number of deaths (red dots closer to 1), although survivals are still observed.

Model Fit: The black dotted line doesn't perfectly fit all the red dots, which is expected. No model will capture every nuance in the data. The fit appears reasonably good for the younger ages but does not capture some of the variability in outcomes for the older ages. It's also evident that for ages 50 and above, the spread of actual outcomes (red dots) is wider, showing more variability.

Outliers: There are a few instances, especially in the younger age group, where despite a low predicted probability of death, deaths (STA = 1) have occurred. Similarly, in the older age group, despite higher predicted probabilities, survivals (STA = 0) are observed.

In summary, this graph provides a visual representation of how the logistic regression model's predictions align with the actual outcomes based on age. It suggests that while age is a significant predictor, other factors not included in this model might also play a crucial role in determining the outcome.

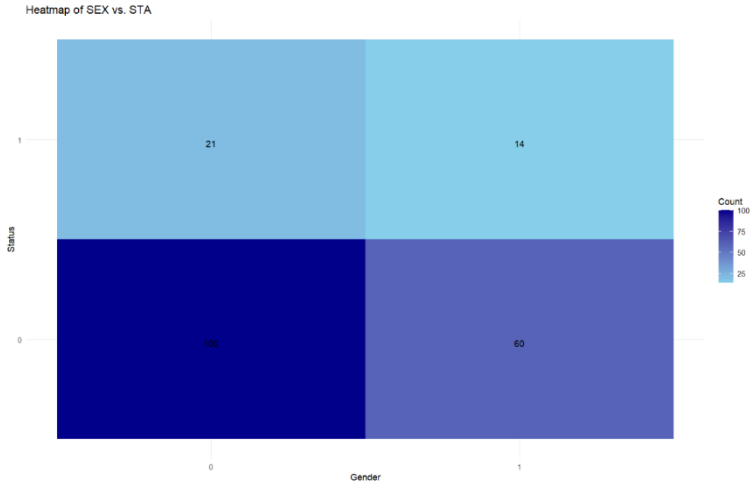
i. Use the logistic model you developed to predict the probability of survival for someone your age. Is this prediction consistent with what you see in the scatterplot above? Does this seem like a

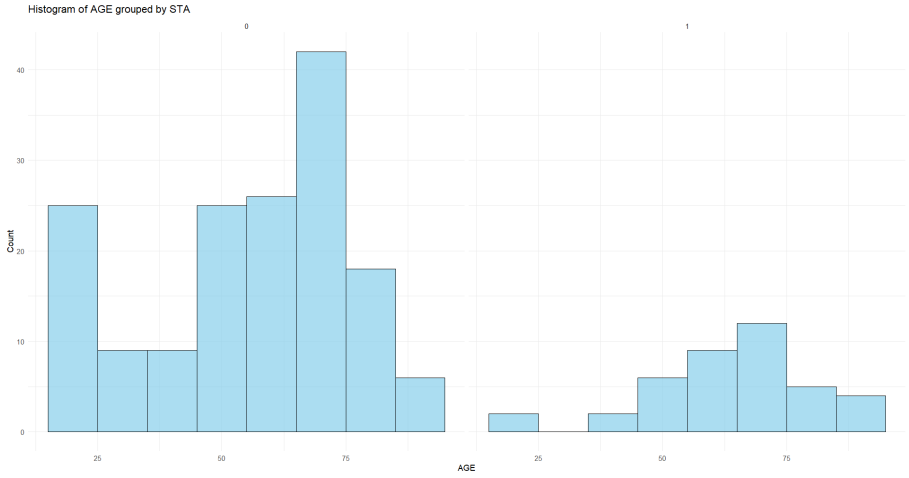
reasonable prediction given what you observed in Tasks 1 and 2? Do we have the correct model yet?

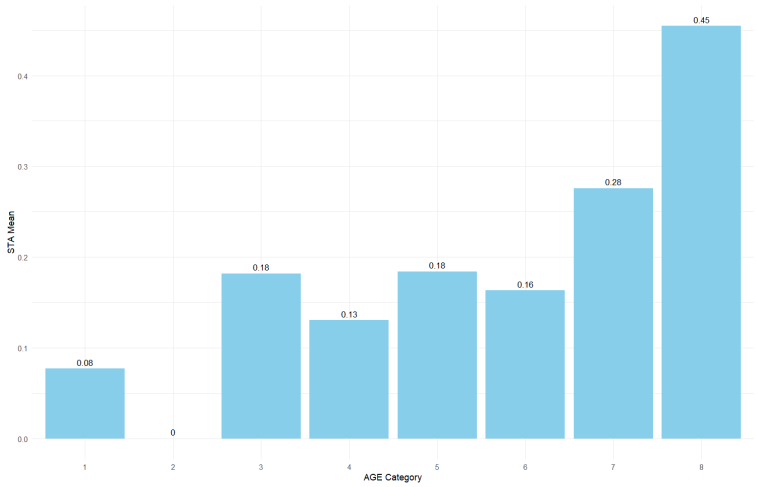
```
> # Assuming the logistic regression model is named 'logit_model'
> # Fit a logistic regression model (if you haven't done so already)
> logit_model <- glm(STA ~ AGE, data=icu, family=binomial)
>
> # Predict probability of death for someone 35 years old
> new_data <- data.frame(AGE = 35)
> predicted_prob <- predict(logit_model, newdata=new_data, type="response")
> prob_survival = 1 - predicted_prob
>
> cat("Probablity of Survival:", prob_survival)
Probablity of Survival: 0.8903696
```

Yes, the probability of survival (0.89) is reflected as probability of death (~0.1) in the scatterplot in Task 4h.

In Task 1 and 2, we plotted these plots:







For a 35-year-old male:

Heatmap of SEX vs. STA: We see that a substantial number of males (SEX=0) survived (STA=0). Only 21 males had STA=1, indicating death. 82.64% of the males survived.

Histogram of AGE grouped by STA: For the age group around 35, the number of survivors (STA=0) is higher than the number of people with STA=1. In fact, the STA=1 bar for number of people aged around 35 shows less than 5 people dead.

Bar Chart for Probability of Death for Age Groups: Only 18% of the people in the Age-Category 3 (35-44) died, and no-one in the Age-Category 2 died. The chance of survival of a 35-year-old is greater than 82%.

Given these plots, an 89% probability of survival for a 35-year-old male seems reasonable and consistent.

Task 5

Given what you have learned from this modeling endeavor so far, what are the next steps for our analysis? What is your recommended plan for the next phase of modeling?

The current model only considers age as a predictor. Although age is an important factor, it's likely not the only one affecting the survival outcome. We should consider including other predictors to improve the model's accuracy.

Recommendations and Next Steps:

Dataset:

- To yield more definitive and statistically significant outcomes, it's imperative to leverage a substantially larger dataset. With just 200 records, we are limited in our capacity to construct a robust predictive model or derive credible insights. Expanding our dataset will not only enhance the accuracy and reliability of our model but also allow for a more comprehensive analysis, accounting for a broader spectrum of scenarios and variations.

Categorical Variables:

- Convert categorical predictors into dummy or one-hot encoded variables before adding them to the regression model, ensuring that multicollinearity isn't introduced.

Multivariate Logistic Regression:

- Incorporate multiple predictors into the logistic regression model to enhance its predictive power. For instance, glm(STA ~ AGE + SEX + RACE + SER + CAN + CRN + INF + CPR + SYS + HRA + PRE + TYP + FRA + PO2 + PH + PCO + BIC + CRE + LOC, family = binomial, data = icu)

Interaction Terms:

- Introduce interaction terms between some predictors. For example, the interaction between age and chronic diseases (like renal failure) could be significant.

Feature Selection:

- Given the number of predictors, it's crucial to identify the most significant ones. Techniques like backward elimination, forward selection, or using regularization methods (like Lasso or Ridge) can be beneficial.

Model Diagnostics:

- With the introduction of multiple predictors, it's even more essential to check for multicollinearity. Variance Inflation Factor (VIF) can be employed for this purpose.
- Conduct residual analysis, and evaluate the model's performance using ROC curves and confusion matrices.

Model Validation:

- As before, ensure model validation using methods like cross-validation or splitting the data into training and testing sets to check its performance.