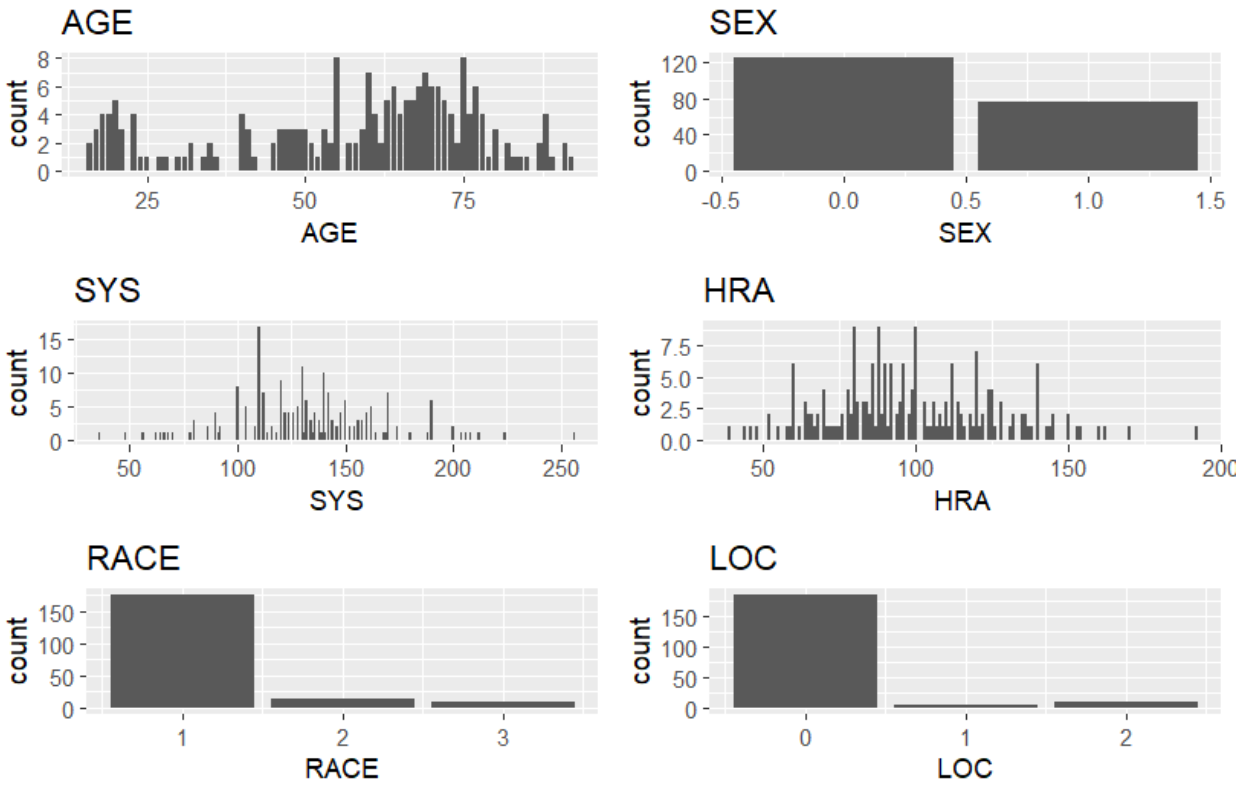
1. Initial analysis shows that all variables are in integer form.
   STA will be our target variable which is whether the patient survived or died.
   The structure consists of 21 variables and 200 datapoints.
   For our target variable according to the data dictionary 0 = Lived and 1 = Died
   The outcome of these patients results in a 75% survival rate with 160 patients.

   The figure below shows correlations between STA and each variable.

| AGE | SEX | RACE | SER | CAN | CRN | INF | CPR |
|---|---|---|---|---|---|---|---|
| 0.1894579 | 0.02060214 | -0.05057217 | -0.1854549 | 1.990527e-20 | 0.1790495 | 0.1823492 | 0.2231003 |
| SYS | HRA | PRE | TYP | FRA | PO2 | PH | PCO |
| -0.2046723 | 0.0317609 | 0.035007 | 0.2435801 | -6.817649e-20 | 0.0829361 | 0.07098647 | -4.941026e-21 |
| BIC | CRE | LOC | | | | | |
| 0.0949158 | 0.1720618 | 0.4370867 | | | | | |

   We can see level of consciousness has the strongest correlation with vital status and systolic blood pressure at ICU admission has the strongest negative correlation.We




   We can see distributions are all over the place with AGE showing a bimodal distribution with two peaks. SYS is right skewed, as well as HRA, RACE, and LOC.

   These will have to be addressed before model building can begin.

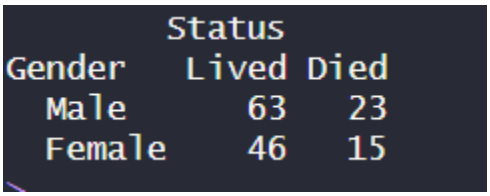
   Recalling our discussion of problems in inequity in healthcare, it's important that our drop strategy incorporates a race blind approach. This is not a guarantee that the model we build will work right off the bat, but data needs to be adjusted to ensure that race does not influence the predictions too much.

Our drop conditions are as follows.

1. Only those who entered due to emergency

   This is the only one needed, as elective will make it too confusing on why someone would have died. This could introduce malpractice instead of a death caused by an accident or incident.

2.



| Gender | Status | |
|---|---|---|
| | Lived | Died |
| Male | 63 | 23 |
| Female | 46 | 15 |

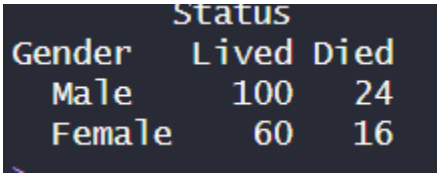
Probability of survival Males: 73.3%

Probability of survival Females: 75.4%

Total survival: 74.1%

The above is after applying the drop conditions where only those admitted for emergency reasons.

Below is without any drop conditions.



| Gender | Status | |
|---|---|---|
| | Lived | Died |
| Male | 100 | 24 |
| Female | 60 | 16 |

Probability of survival Males: 80.6%

Probability of survival Females: 79%

Total survival: 80%

It's obvious that emergencies led to more deaths as opposed to the total population which includes elective operations.

3.

| Type_of_Admission | Status | |
|---|---|---|
| | Lived | Died |
| Elective | 51 | 2 |
| Emergency | 109 | 38 |

Probability of survival Elective: 96.2%

Probability of survival Emergency: 74%

Odds Ratio between Elective and Emergency of survival: 8.8899

This high odds ratio shows that you are almost 9 times more likely to survive if your visit is elective in nature. This is not surprising as elective procedures are not emergencies by their very nature.

4.

a. Logistic Regression model of STA (Y) using AGE (X)

$$[\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 \times \text{AGE}]$$

Where P is the probability of survival

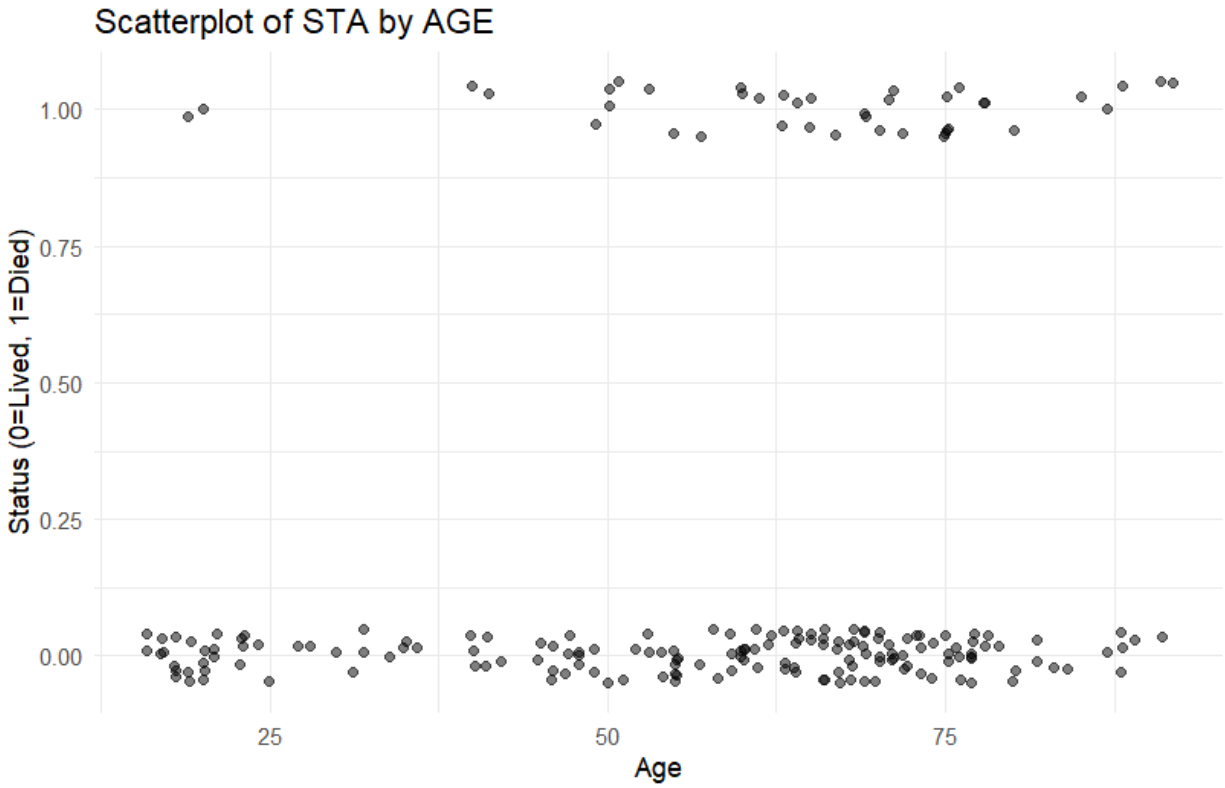
B0 is the intercept term

Beta1 acts as the coefficient for AGE

And ln is the natural logarithm.

$$[\text{logit}(P) = \beta_0 + \beta_1 \times \text{AGE}]$$

b.



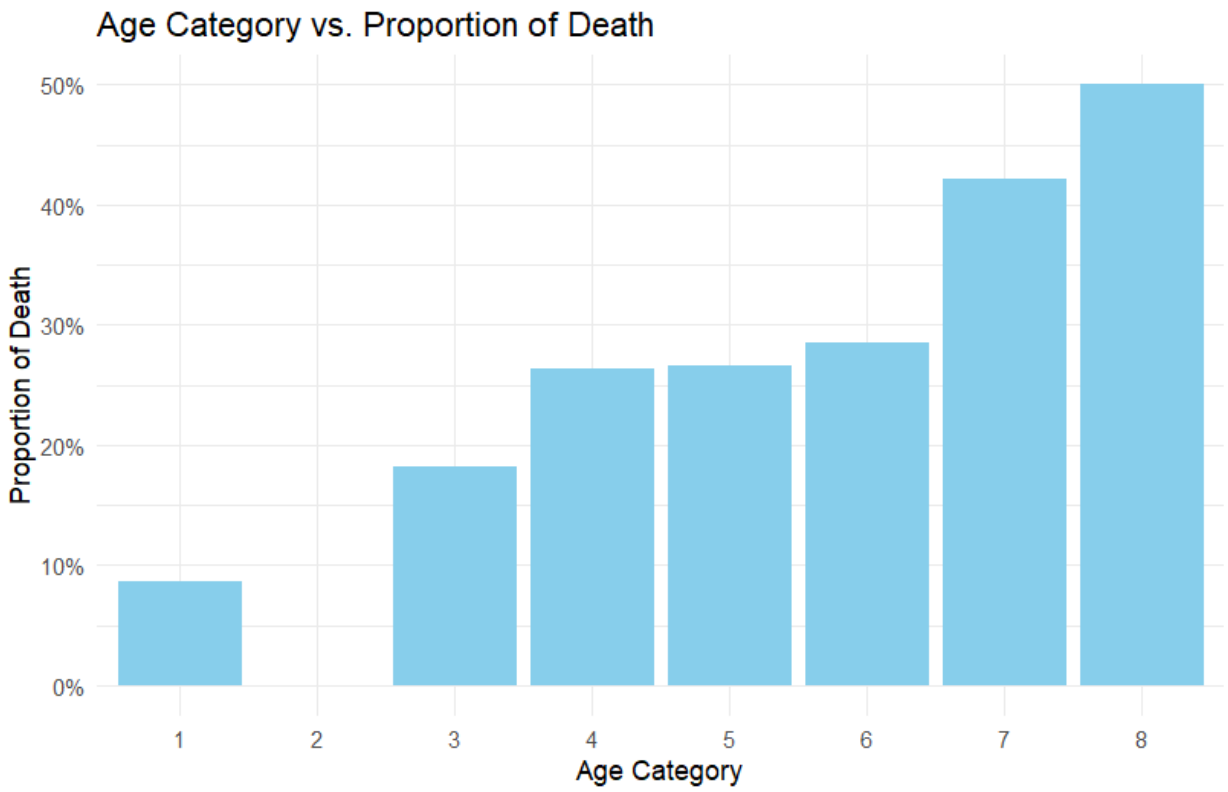

There is a slight correlation between Age and survival with higher density of deaths starting at age 50.

c.

Counts of Age Groups

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| 23 | 7 | 11 | 19 | 30 | 28 | 19 | 10 | 0 |

| AGE_CAT | ProportionOfSurvival |
|---|---|
| 1 | 0.08695652 |
| 2 | 0.00000000 |
| 3 | 0.18181818 |
| 4 | 0.26315789 |
| 5 | 0.26666667 |
| 6 | 0.28571429 |
| 7 | 0.42105263 |
| 8 | 0.50000000 |




The left skewed data backs up common sense that those who are older have a higher percent chance of death when visiting for emergency rooms.

d.

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1851  -0.8331  -0.5943   1.1627   2.2202

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.00757    0.69533  -4.325 1.52e-05 ***
AGE          0.03325    0.01076   3.089  0.00201 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 168.02  on 146  degrees of freedom
Residual deviance: 156.81  on 145  degrees of freedom
AIC: 160.81

Number of Fisher Scoring iterations: 4
```

When AGE is 0, the log odds of surviving are -3, however since AGE cannot be zero, we cannot be certain of this.

For every additional year of AGE, the log odds of survival increase by 0.03325

e.

Being that AGE's P-value is well below the cutoff of 0.05 we must reject the null hypothesis and accept the alternative hypothesis that AGE does affect outcome of survival. There is sufficient evidence in the data to suggest that AGE has a significant impact on the log odds of survival.

$$[H_0 : \beta_{\text{AGE}} = 0]$$

$$[H_A : \beta_{\text{AGE}} \neq 0]$$

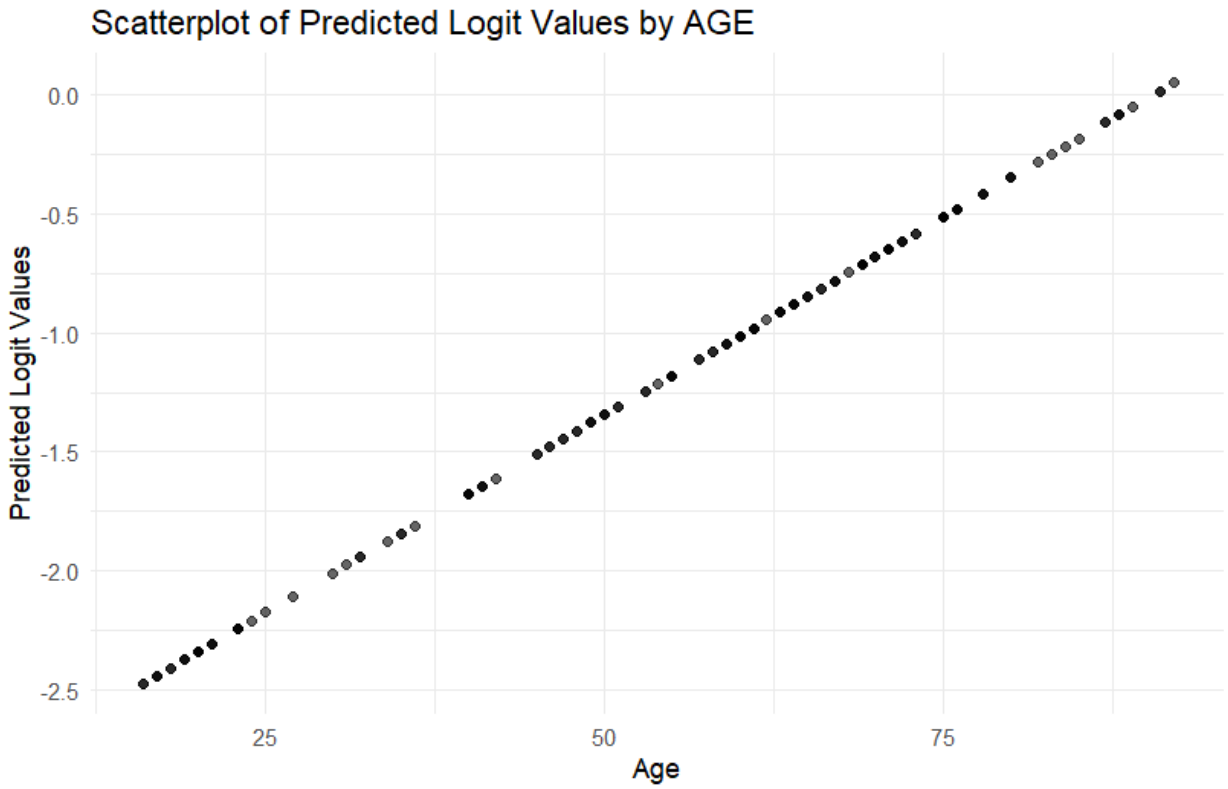
f.

AIC value for the logistic model: 160.8115

BIC value for the logistic model: 166.7923

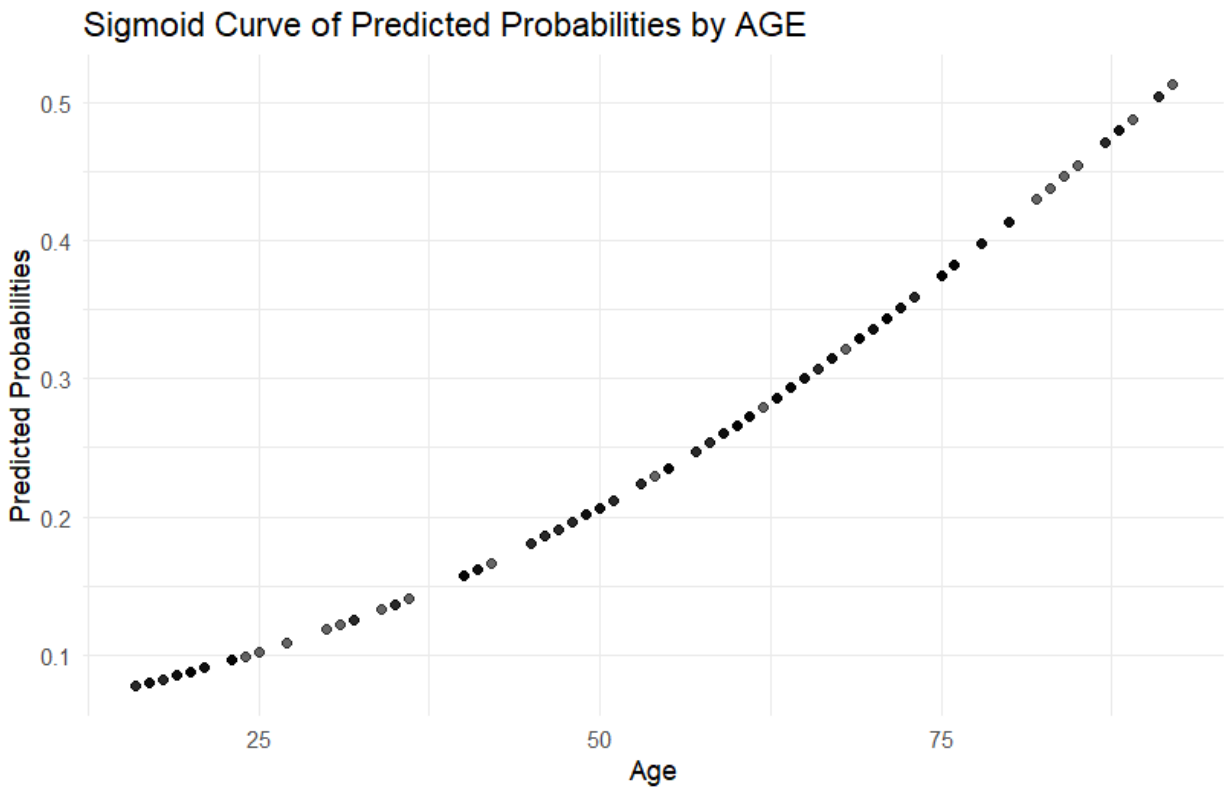
Deviance value for the logistic model: 156.8115

g.




The model is consistent with the formula. As AGE increases, the logit values or log-odds of survival increase linearly. This is not sigmoid shaped because we have not yet applied a function to determine probabilities of survival versus death.

h.




Something didn't work as expected. I applied the function correctly to create new predicted probabilities using plogis() which gets the inverse of the logit function. However, applying it has caused only a slight sigmoid shape, and is not representative of the higher percentage of deaths. The predicted probabilities should go higher than just about 0.5.

i.

Running Age 36 through the simulation gives me a chance of survival of 85%. This seems reasonable since the drop conditions were for emergencies. It still may be a bit low, but with only 200 observations, and only a handful of those at my age, it's almost impossible to say. Many more would be needed to gauge the accuracy of this model.

As of right now, the model is not complete. We only have one variable against which to build a model which is age. I'm sure quite a bit of variance is contained in this, but we have not used any of the other variables.

5.

My steps would include going back to the EDA phase and making sure all distributions are normal. Already, we've seen both left and right skewed variables that are throwing our data out of whack. It's possible to apply a log function to this data to bring them more in line.

We could try applying One Hot Encoding as many of these are integers, but in reality categorical variables and the model could be weighing 0 and 1 differently.