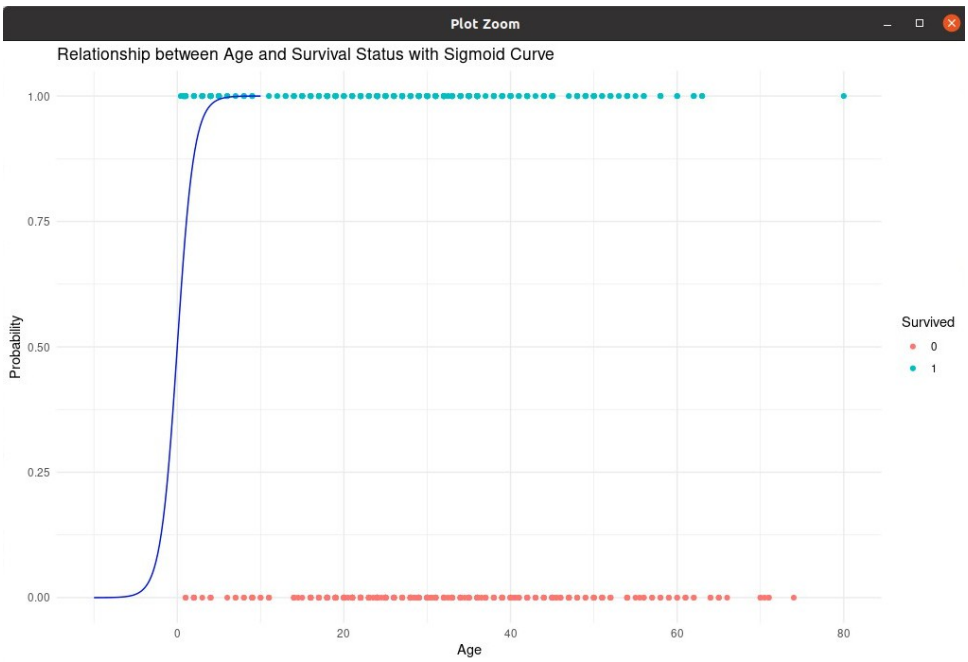
Program :-

```
library(caTools)
library(ggplot2)

# Load the dataset
data <- read.csv("/home/hadoop/Desktop/ajnaas/dataset/Titanic.csv")

# Check for missing values
sum(is.na(data))

# Remove rows with missing values
data <- na.omit(data)

# Convert categorical variables to factors
data$Sex <- factor(data$Sex)
data$Ticket <- factor(data$Ticket)
data$Cabin <- factor(data$Cabin)

# Split data into training and testing sets
set.seed(123)    # for reproducibility
split <- sample.split(data$Survived, SplitRatio = 0.7)
train_data <- subset(data, split == TRUE)
test_data <- subset(data, split == FALSE)

model <- glm(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare,
                    data = train_data, family = binomial)

# Summary of the logistic regression model
summary(model)

# Prediction
predictions <- predict(model, newdata = test_data, type = "response")
binary_predictions <- ifelse(predictions > 0.5, 1, 0)

# Calculate accuracy
accuracy <- mean(binary_predictions == test_data$Survived)
print(paste("Accuracy:", accuracy * 100))

# Plot sigmoid curve
x <- seq(-10, 10, length.out = 100)
sigmoid <- 1 / (1 + exp(-x))

# Plot data points and sigmoid curve using ggplot2
ggplot() +
  geom_point(data = data, aes(x = Age, y = Survived, color = factor(Survived))) +
  labs(title = "Relationship between Age and Survival Status",
       x = "Age",
       y = "Survived",
       color = "Survived") +
  theme_minimal() +
  geom_line(data = data.frame(x = x, sigmoid = sigmoid), aes(x = x, y = sigmoid),
            color = "blue", linetype = "solid") +
  labs(title = "Relationship between Age and Survival Status with Sigmoid Curve",
       x = "Age",
       y = "Probability")
```

Output :-


Plot Zoom
Relationship between Age and Survival Status with Sigmoid Curve
1.00
0.75
0.50
0.25
0.00
Probability
0
20
40
60
80
Age
Survived
0
1