Poject-3

Group 1

2022-04-27

#Loading required R-packages

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(rsvg)
```

```
## Linking to librsvg 2.48.4
```

```
library(ggimage)
```

```
## Loading required package: ggplot2
```

```
library(ggplot2)
library(tibble)
library(cvms)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
library(arm)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:olsrr':
##
##     cement
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

```
##
## arm (Version 1.12-2, built: 2021-10-15)
```

```
## Working directory is /Users/charleskolozsvary/Documents/Collegiate/Spring_2022/MATH_4
56/essays/project_2/R-env/project-2
```

#Load and glimpse data

```
df <- read.csv("~/Documents/Collegiate/Spring_2022/MATH_456/essays/project_2/R-env/proje
ct-2/proj-3-data/age_known_titatic.csv")
glimpse(df)
```

```
## Rows: 755
## Columns: 5
## $ Name     <chr> "\"Allen, Miss Elisabeth Walton\"", "\"Allison, Miss Helen Lo…
## $ PClass   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ Age      <dbl> 29.00, 2.00, 30.00, 25.00, 0.92, 47.00, 63.00, 39.00, 58.00, …
## $ Sex      <int> 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1…
## $ Survived <int> 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1…
```

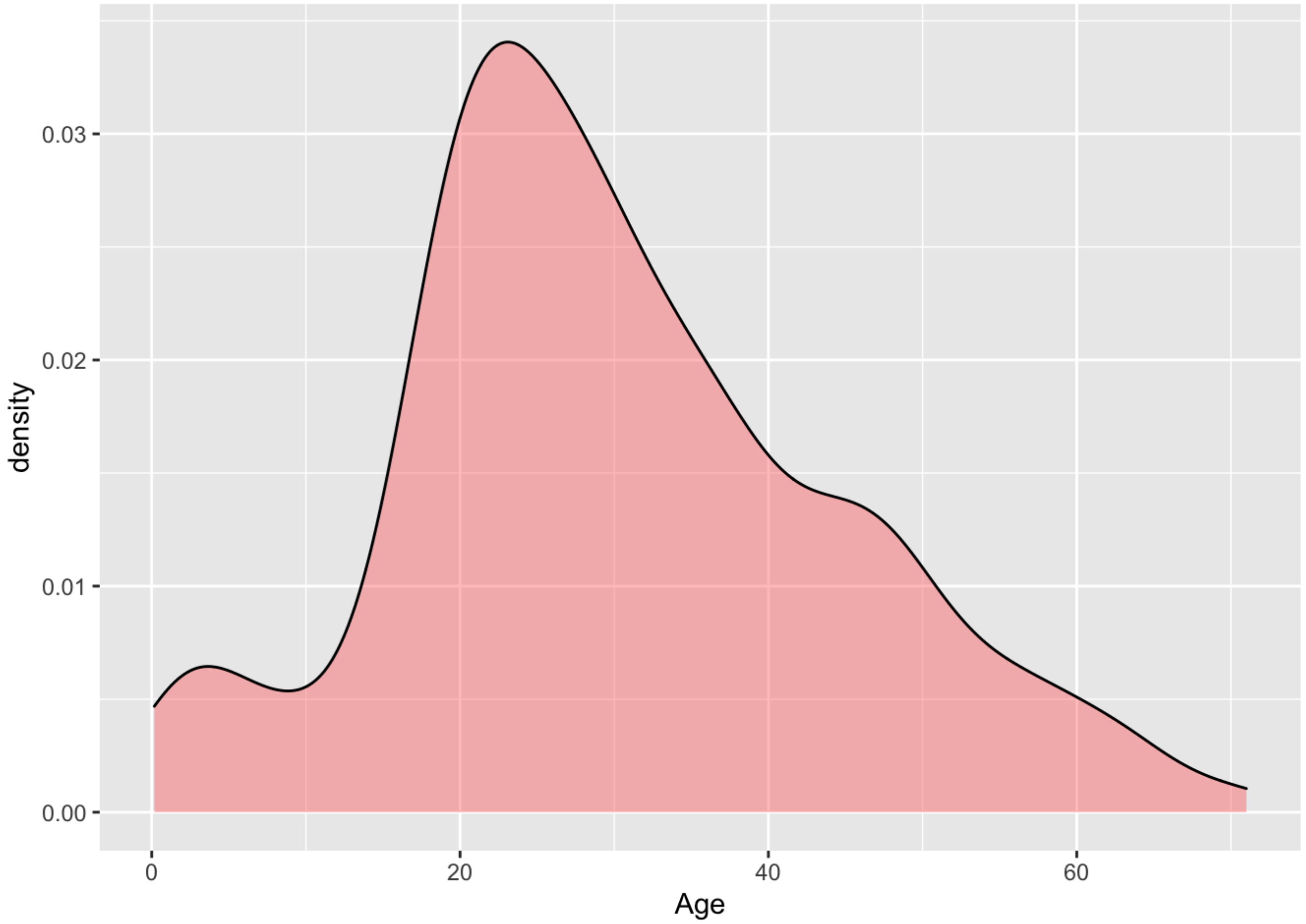
#Check Distribution of Continuous Variables

```
continuous <-select_if(df, is.numeric)
summary(continuous)
```

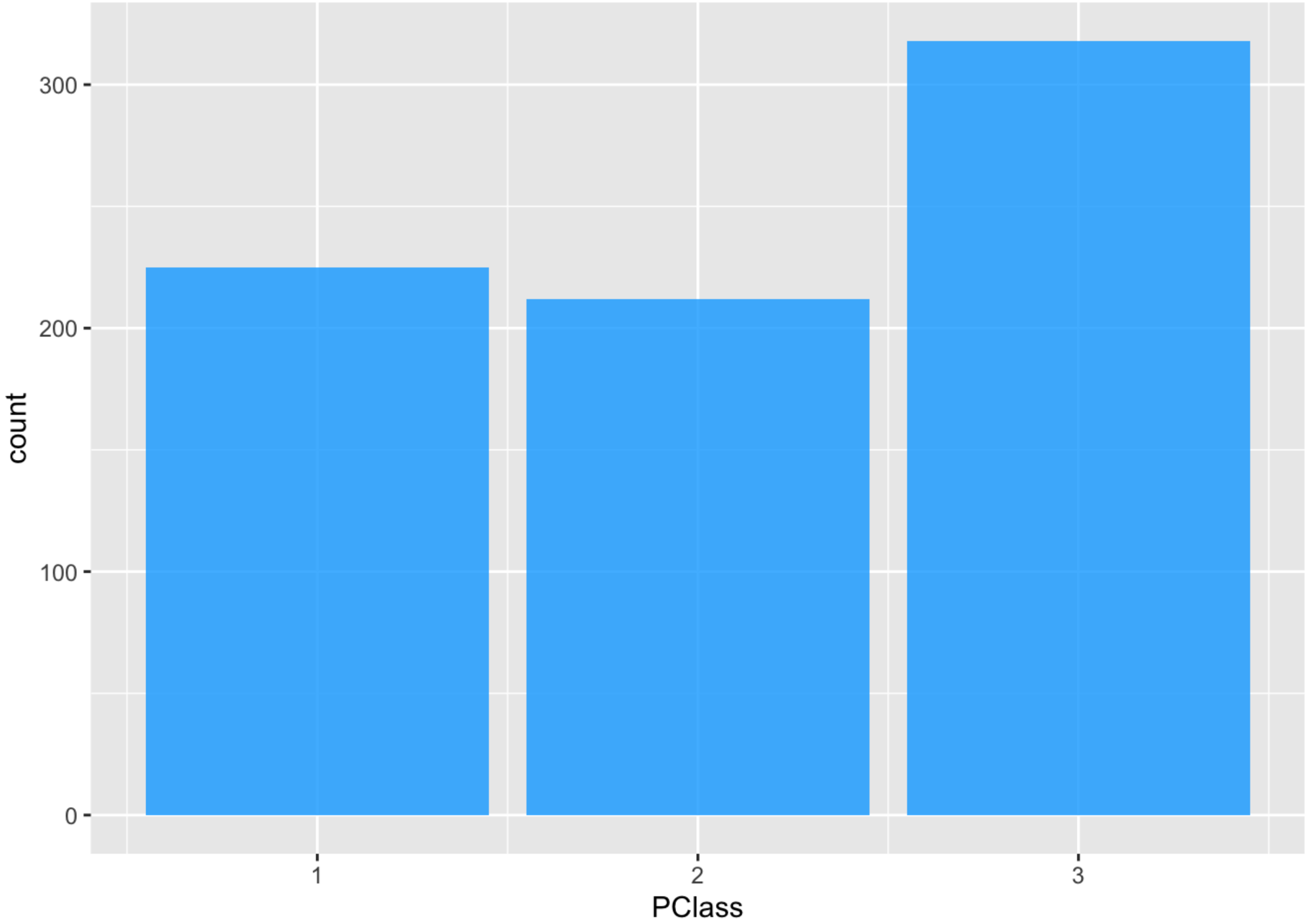
```
##      PClass           Age             Sex            Survived
##  Min.   :1.000   Min.   : 0.17   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :2.000   Median :28.00   Median :0.0000   Median :0.0000
##  Mean   :2.123   Mean   :30.38   Mean   :0.3801   Mean   :0.4132
##  3rd Qu.:3.000   3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :71.00   Max.   :1.0000   Max.   :1.0000
```

#Data inspection

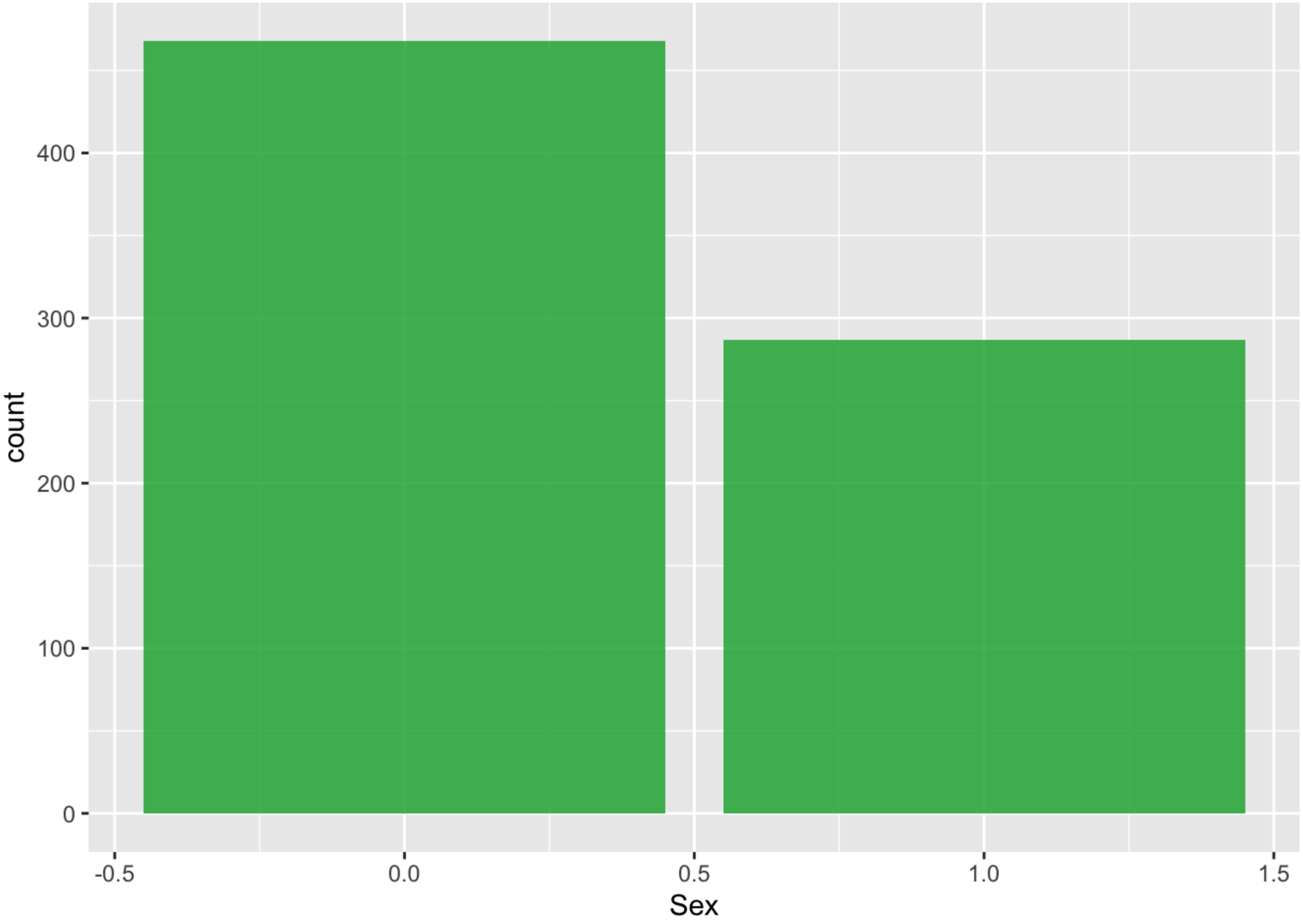
```
ggplot(continuous, aes(x = Age))+geom_density(alpha = .4, fill = "#FF6666")
```



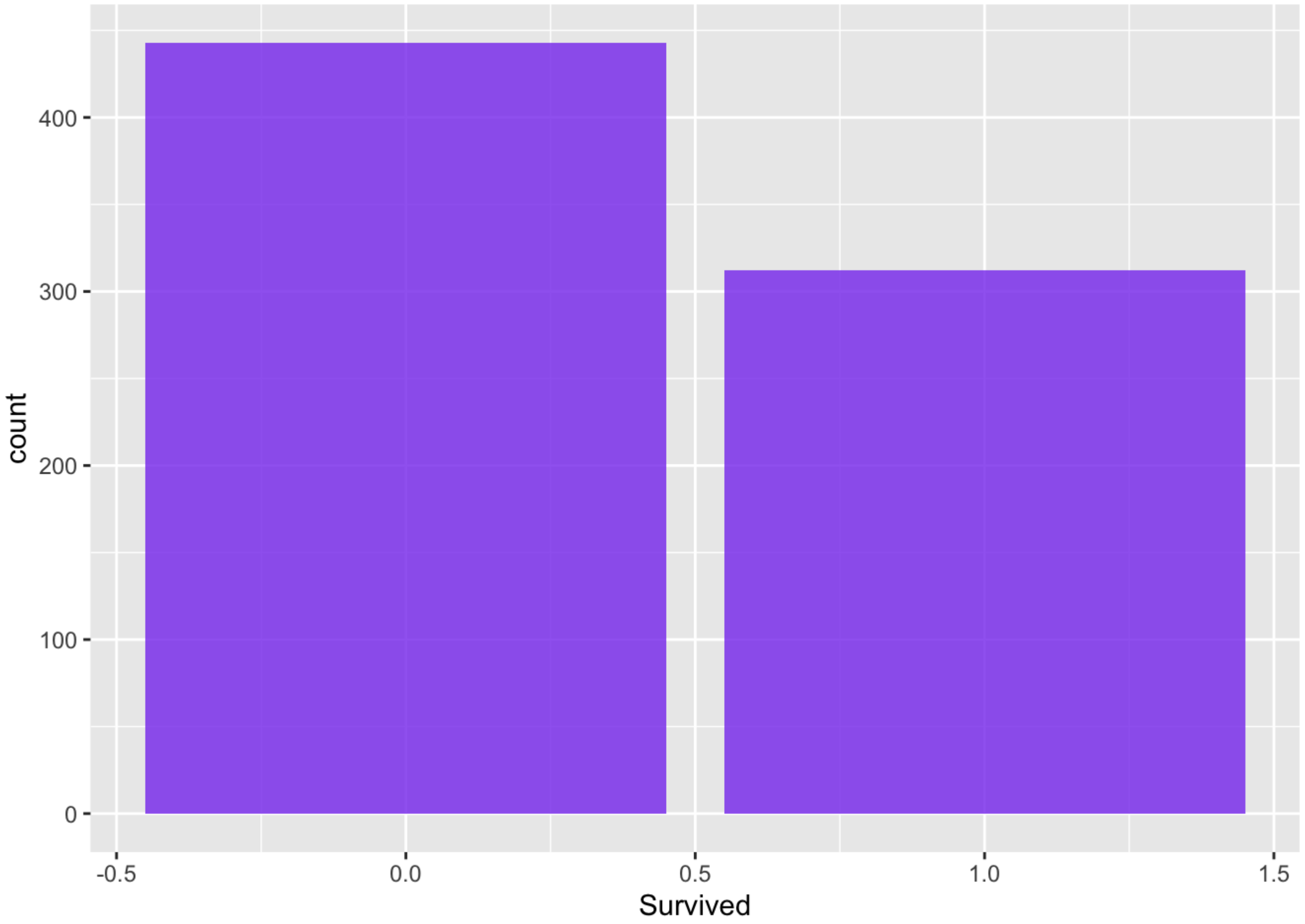
```
ggplot(continuous, aes(x = PClass))+geom_bar(alpha = 0.8, fill = "#00AAFF")
```



```
ggplot(continuous, aes(x = Sex))+geom_bar(alpha = 0.8, fill = "#01AA33")
```



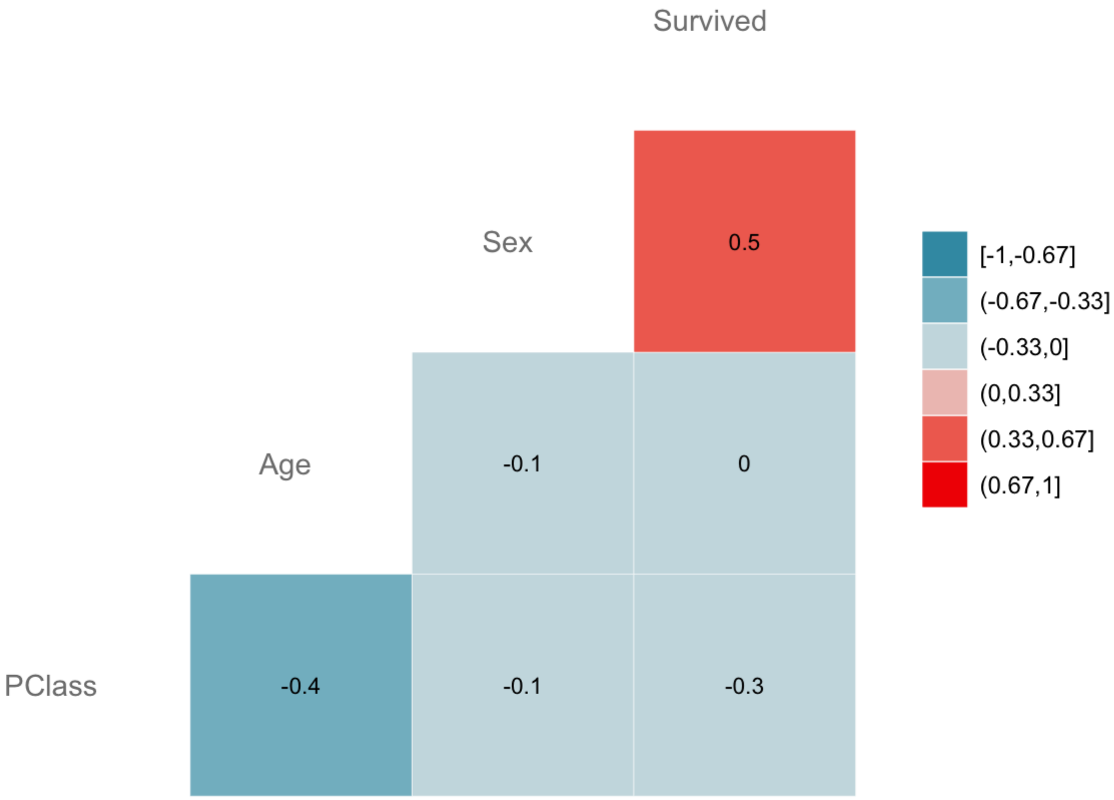
```
ggplot(continuous, aes(x = Survived))+geom_bar(alpha = 0.8, fill = "#8845EE")
```



#Visualize Correlation

```
# Convert data to numeric if not already
corr <- df
# Plot the graphg
    ggcorr(corr,
    method = c("pairwise", "spearman"),
    nbreaks = 6,
    hjust = 0.8,
    label = TRUE,
    label_size = 3,
    color = "grey50")
```

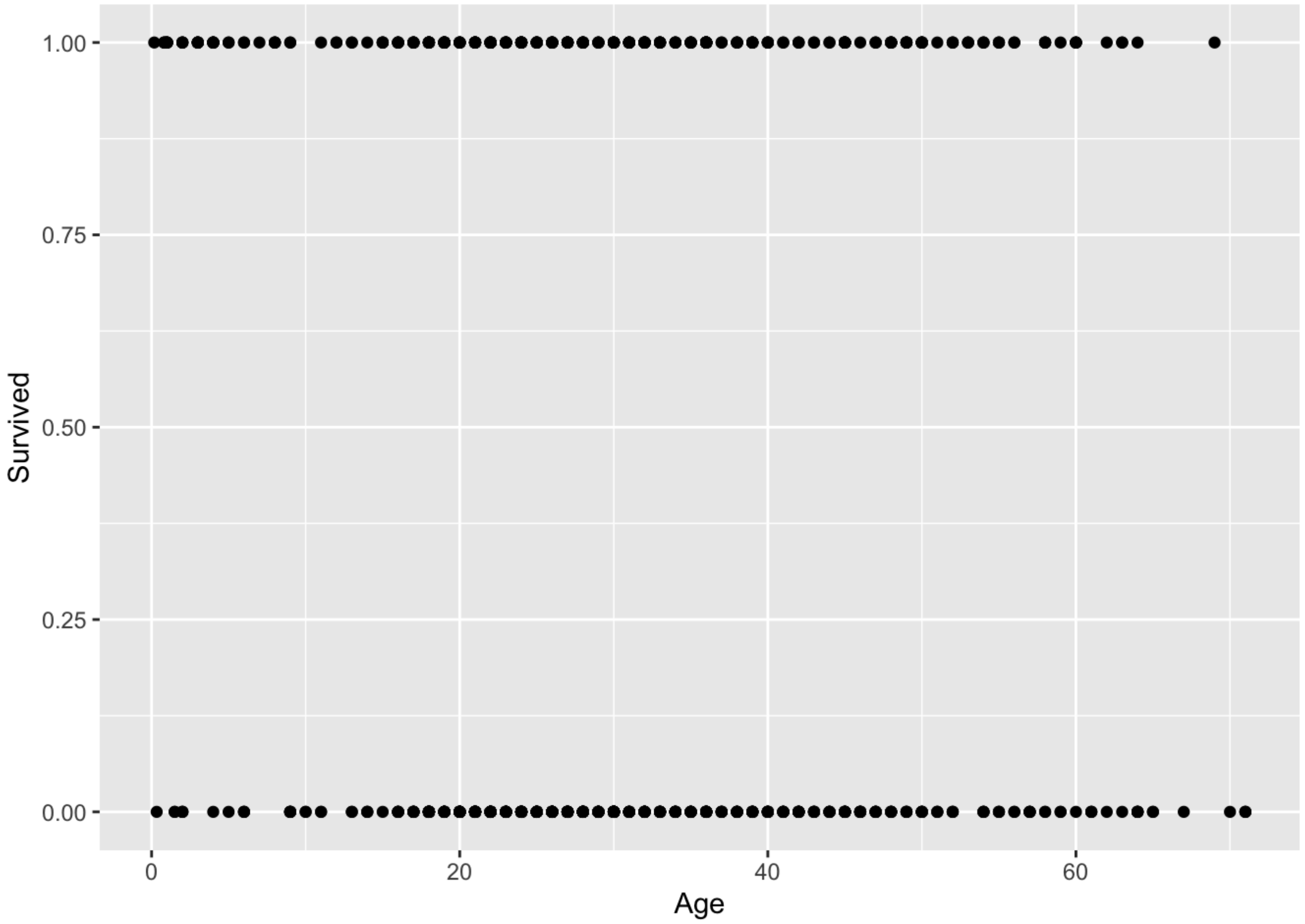
```
## Warning in ggcorr(corr, method = c("pairwise", "spearman"), nbreaks = 6, : data
## in column(s) 'Name' are not numeric and were ignored
```



#Age

Versus Survival

```
ggplot(df, aes(x = Age, y = Survived))+geom_point()
```



#Split Data into Train and Test

```
## 75% of the sample size
set.seed(1234)
smp_size <- floor(0.75 * nrow(df))
train_ind <- sample(seq_len(nrow(df)), size = smp_size)
train <- df[train_ind, ]
test <- df[-train_ind, ]
```

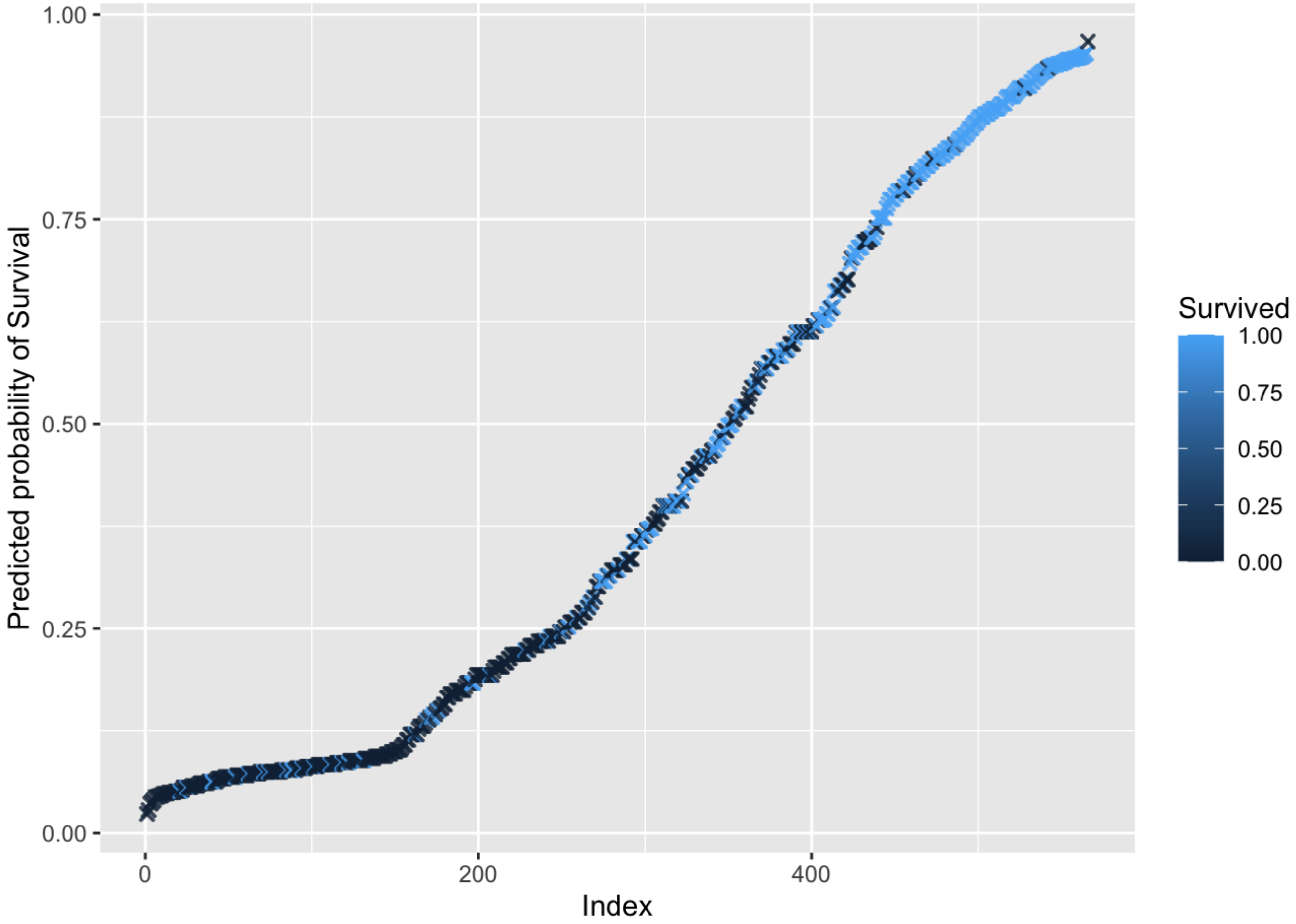
#Build Model

```
model_with_age <- glm(Survived ~ Age + PClass + Sex, data = train, family = 'binomial')
model_without_age <- glm(Survived ~ PClass + Sex, data = train, family = 'binomial')
```

#Outcome of prediction

```
predicted.data <- data.frame(
  probability.of.Survived = model_with_age$fitted.values,
  Survived = train$Survived
)
predicted.data <- predicted.data[
  order(predicted.data$probability.of.Survived, decreasing = FALSE),]
predicted.data$rank <- 1:nrow(predicted.data)

ggplot(data = predicted.data, aes(x = rank, y=probability.of.Survived))+
  geom_point(aes(color = Survived), alpha = 0.8, shape = 4, stroke = 1)+
  xlab("Index")+
  ylab("Predicted probability of Survival")
```

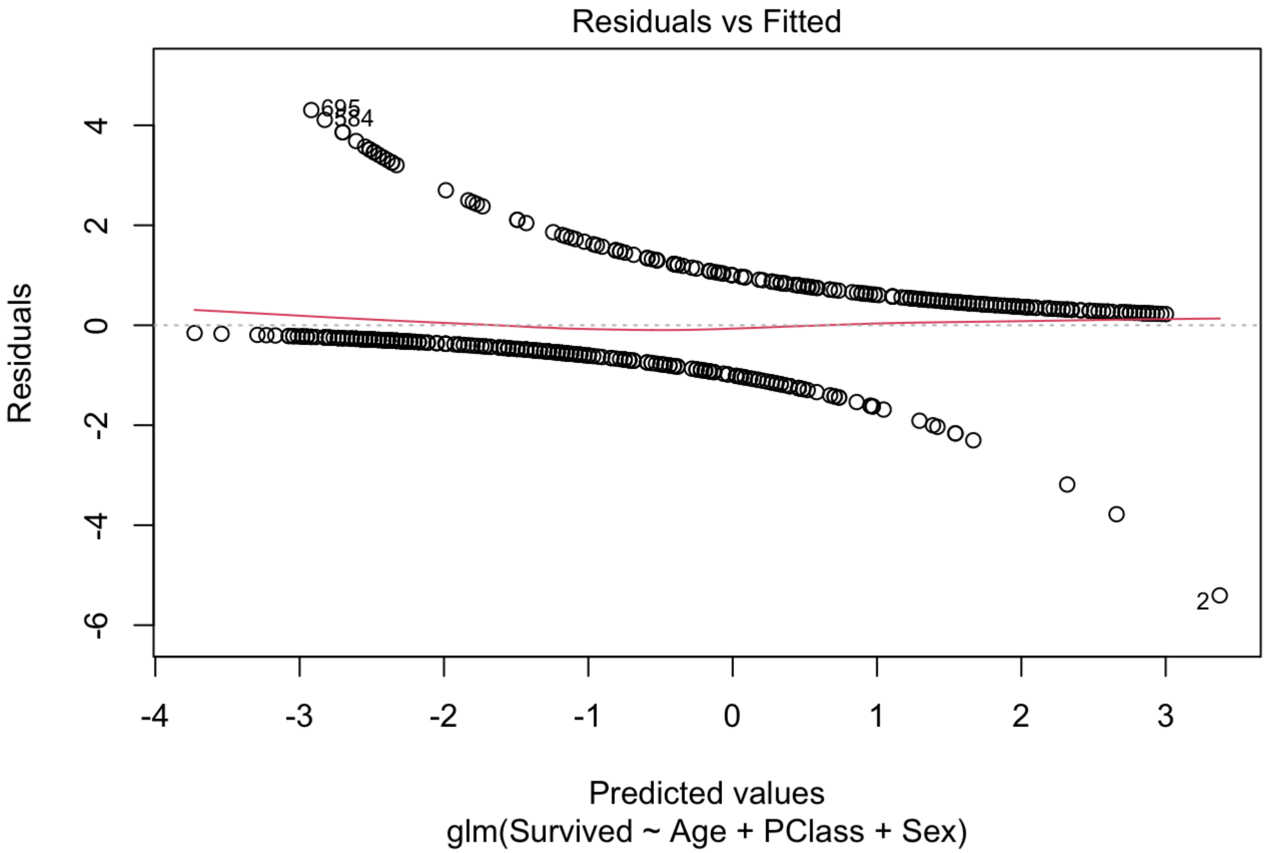


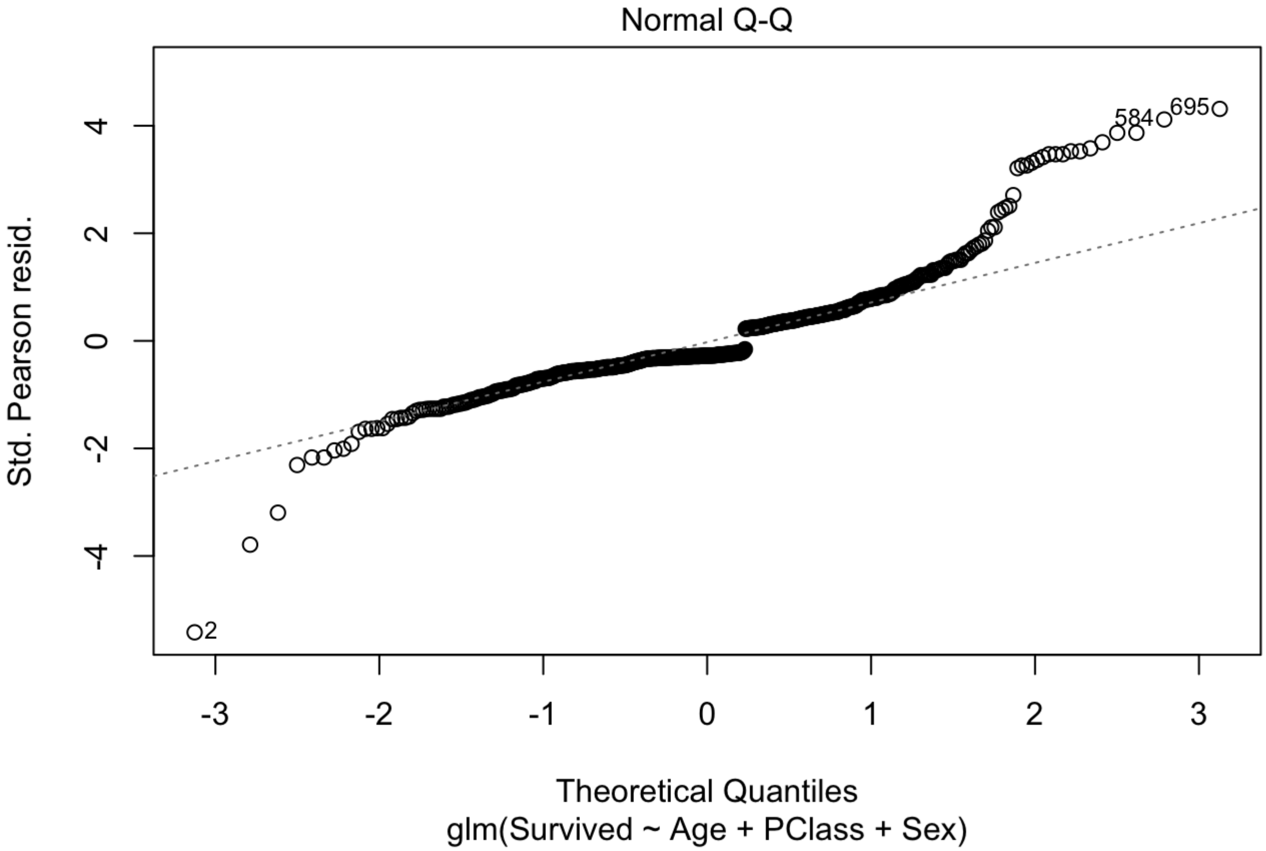
#Model Summary with age

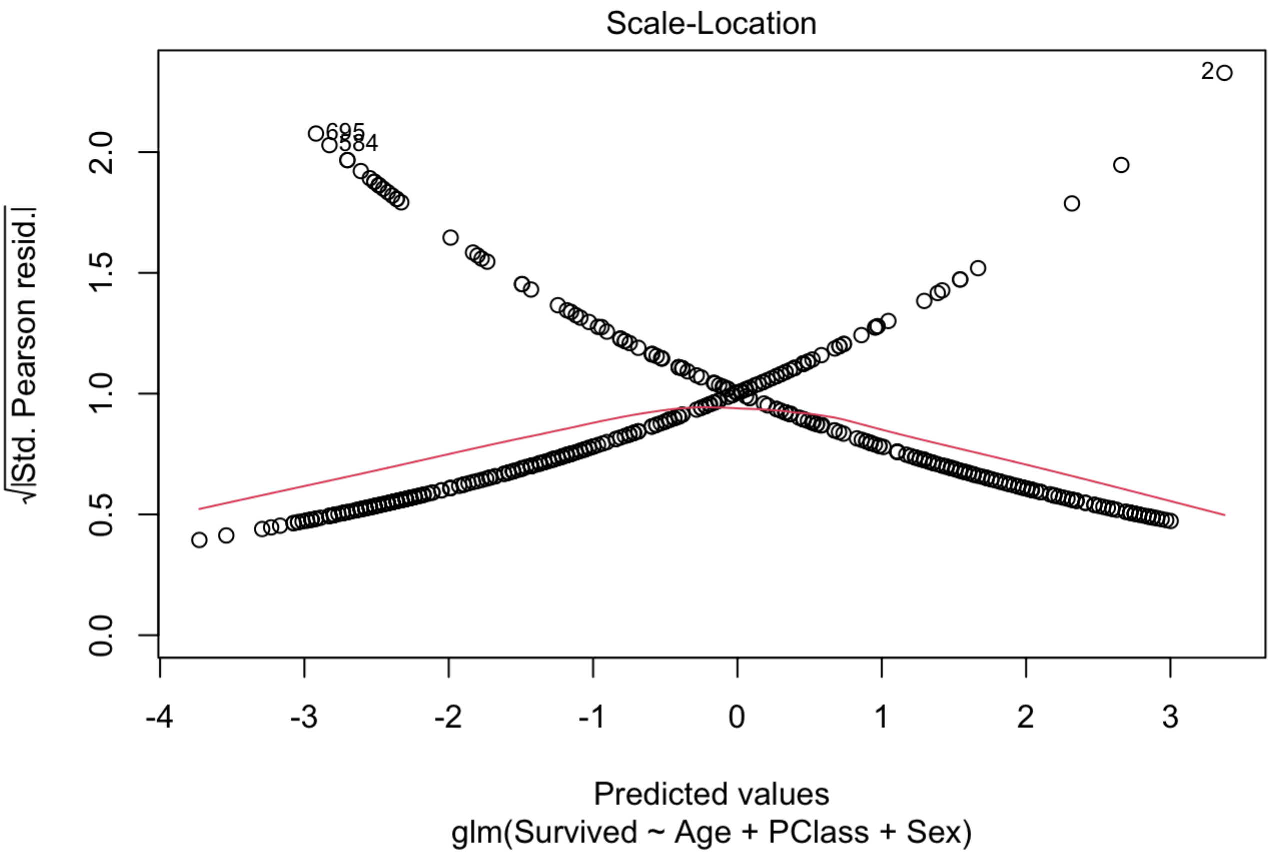
```
summary(model_with_age)
```

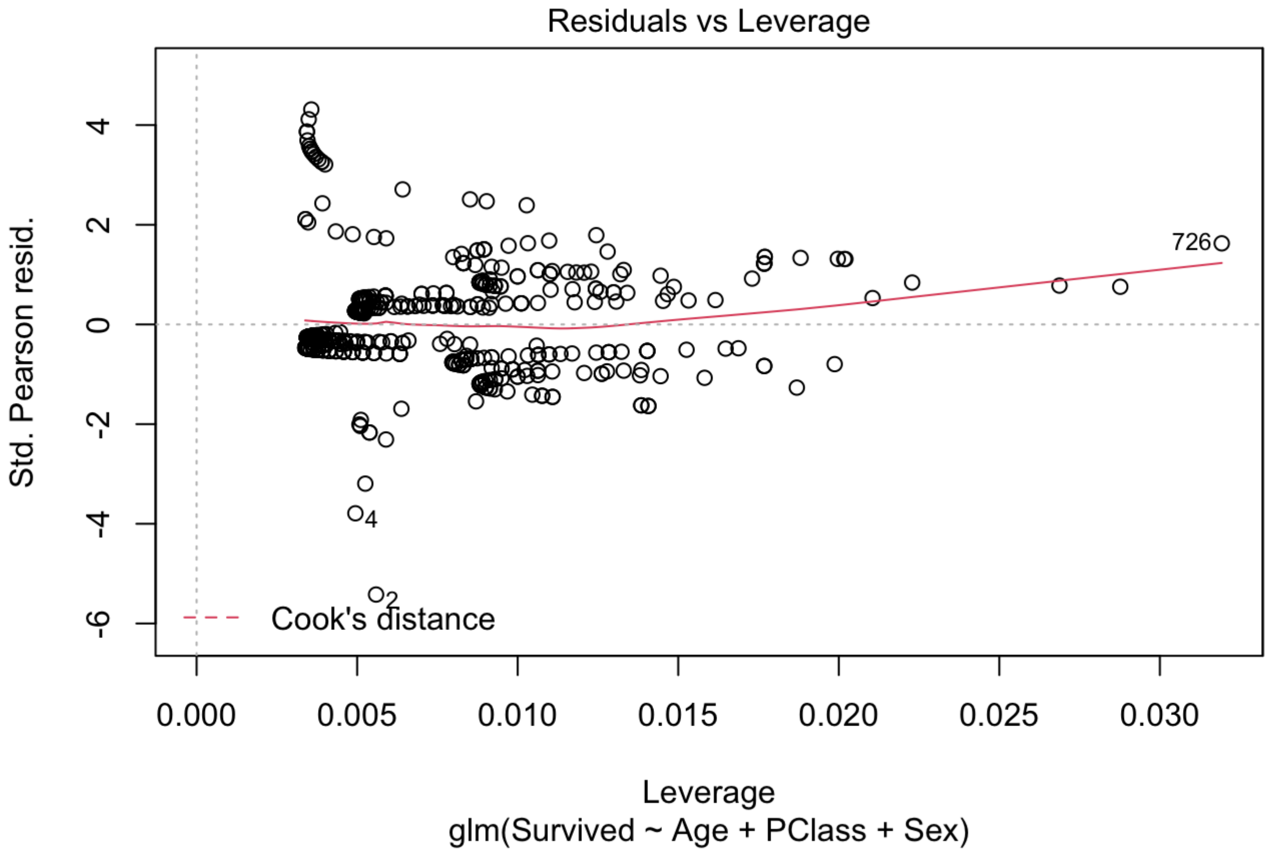
```
##
## Call:
## glm(formula = Survived ~ Age + PClass + Sex, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6108  -0.6925  -0.3826   0.6318   2.4381
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.922073   0.505566   3.802 0.000144 ***
## Age         -0.031084   0.008695  -3.575 0.000351 ***
## PClass      -1.209805   0.159045  -7.607 2.81e-14 ***
## Sex          2.724316   0.235016  11.592  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 765.42  on 565  degrees of freedom
## Residual deviance: 513.09  on 562  degrees of freedom
## AIC: 521.09
##
## Number of Fisher Scoring iterations: 5
```

```
plot(model_with_age)
```


Residuals vs Fitted
Residuals
4
2
0
-2
-4
-6
-4
-3
-2
-1
0
1
2
3
695
584
2
Predicted values
glm(Survived ~ Age + PClass + Sex)


Normal Q-Q
Std. Pearson resid.
4
2
0
-2
-4
-3
-2
-1
0
1
2
3
584
695
2
Theoretical Quantiles
glm(Survived ~ Age + PClass + Sex)



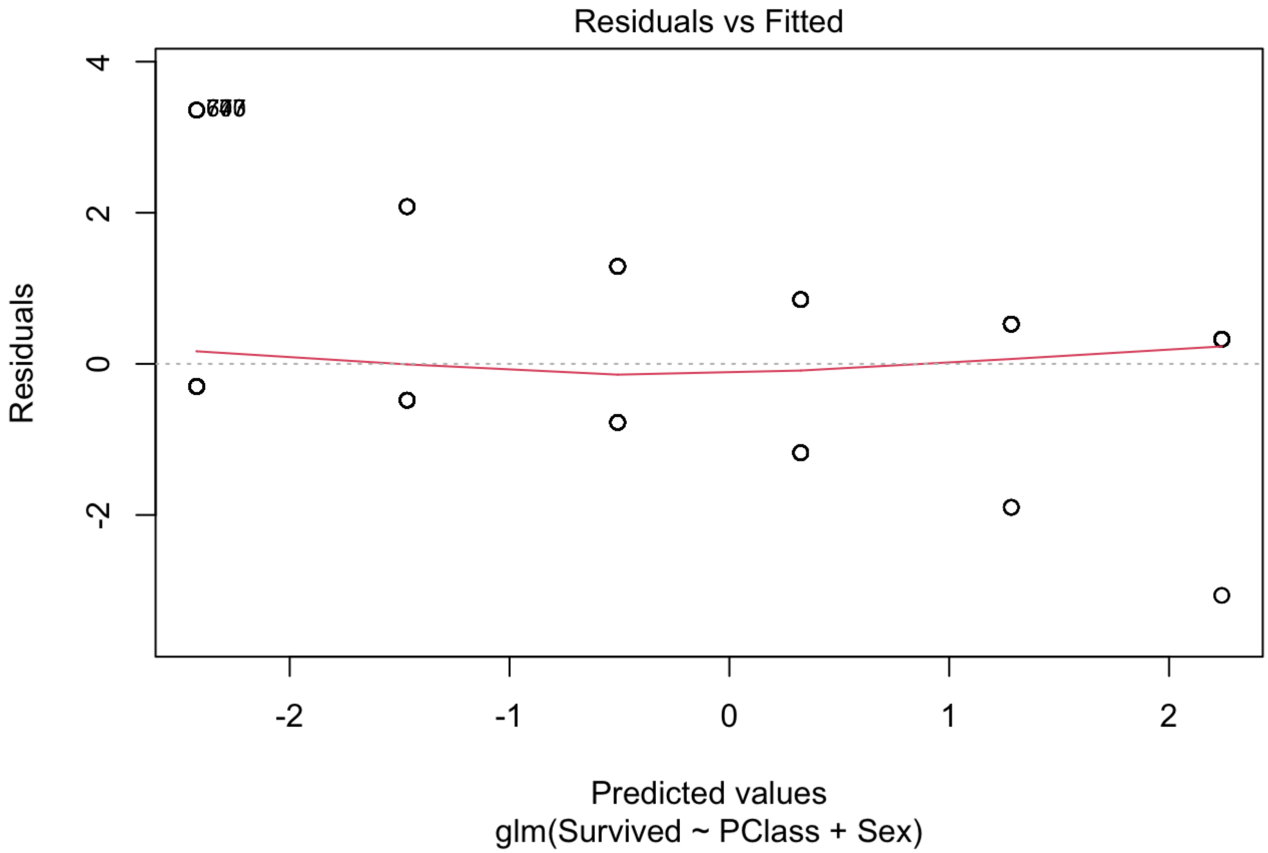


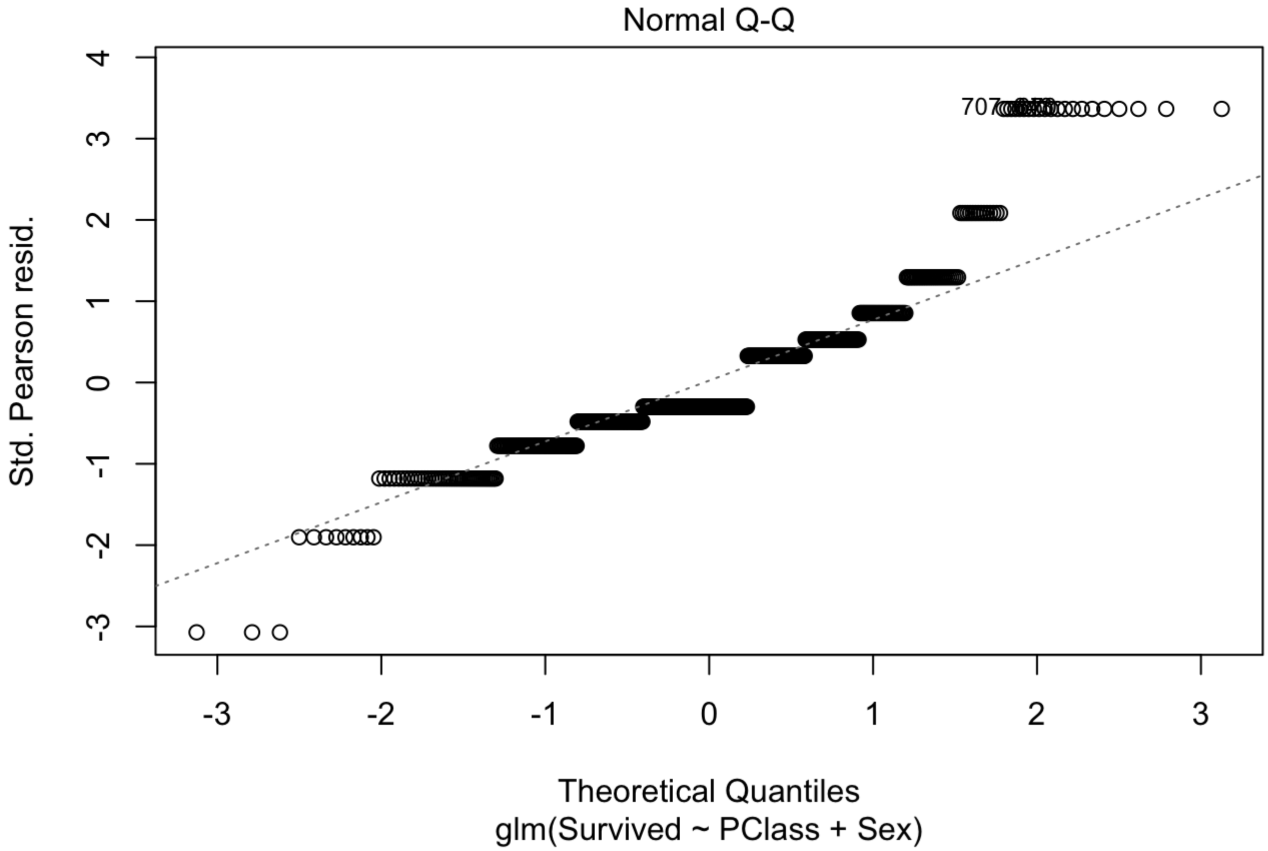
#Model Summary without age

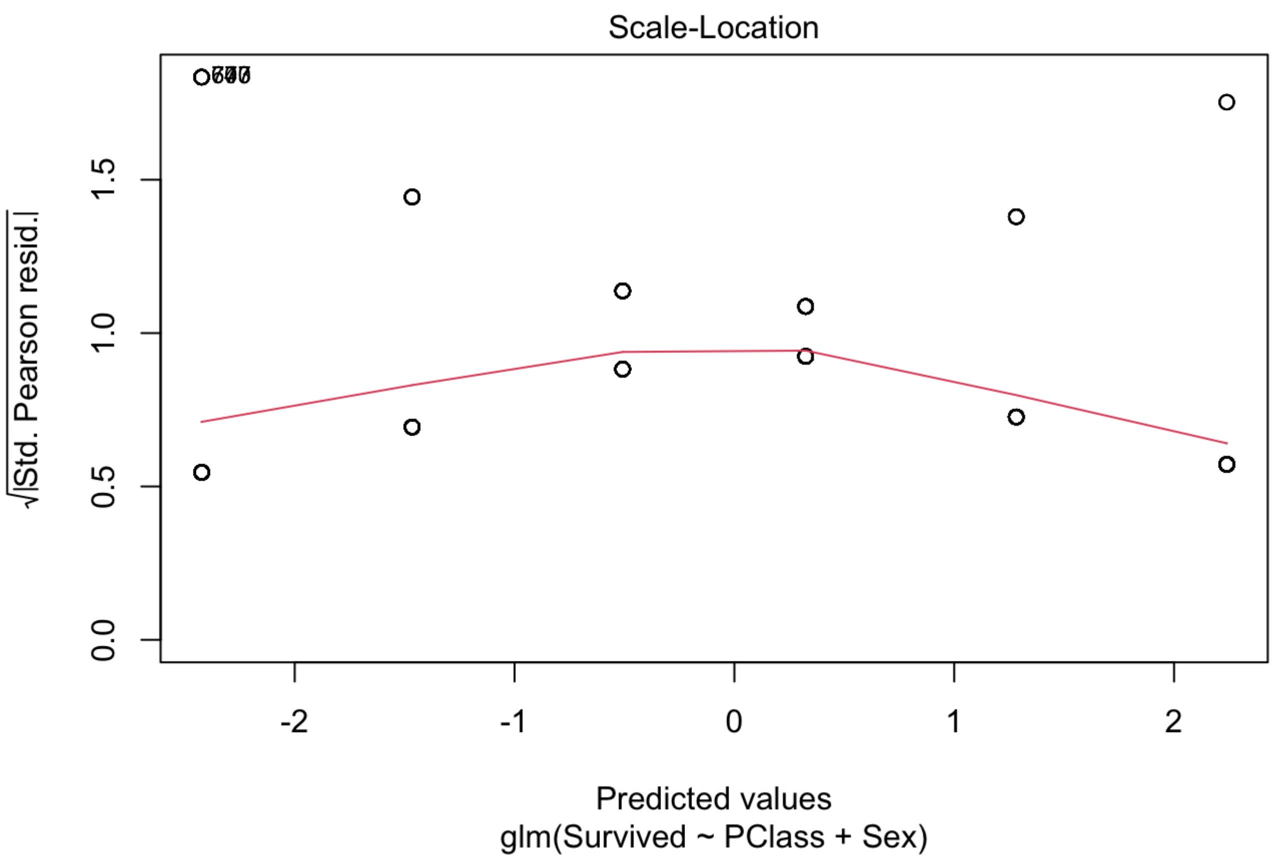
```
summary(model_without_age)
```

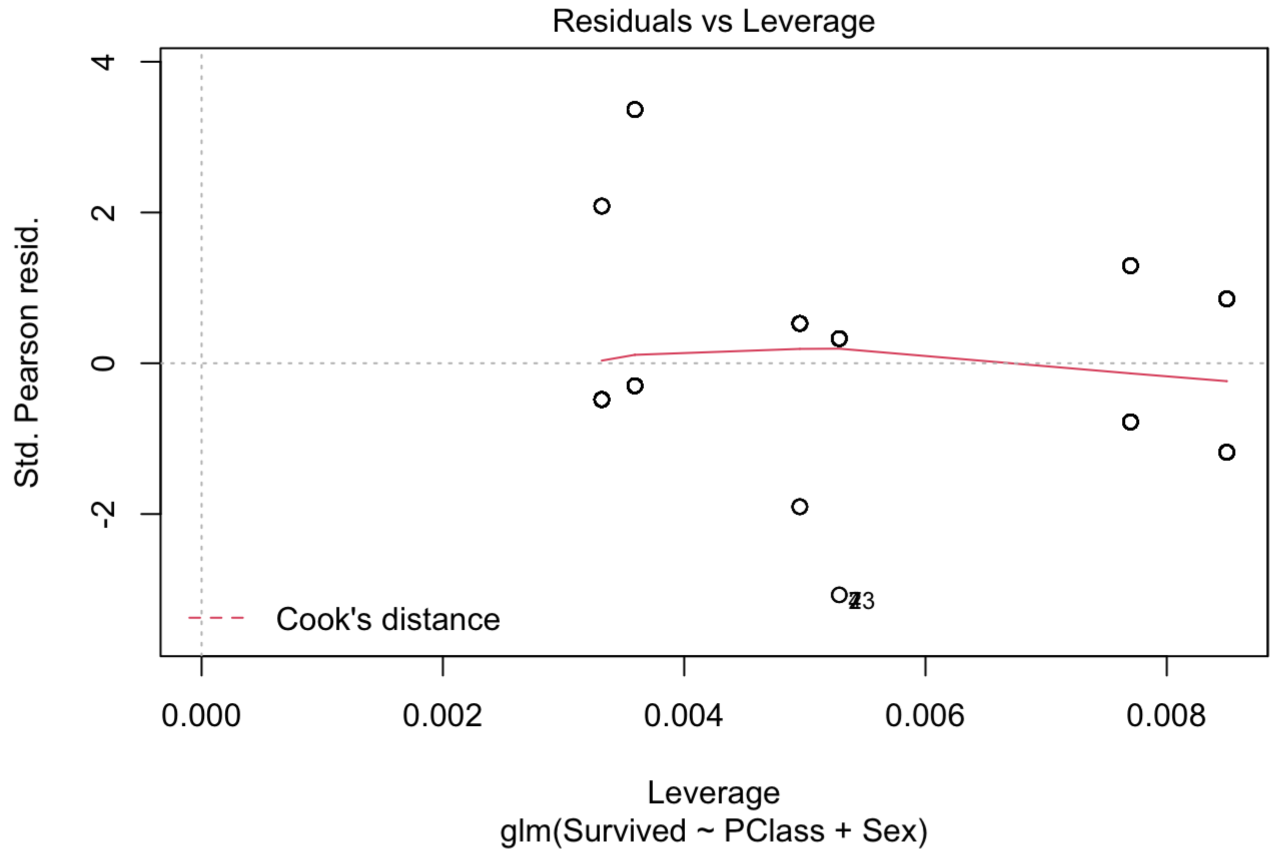
```
##
## Call:
## glm(formula = Survived ~ PClass + Sex, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1638  -0.6446  -0.4121   0.6998   2.2398
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.4495     0.2850   1.577    0.115
## PClass       -0.9577     0.1366  -7.010 2.38e-12 ***
## Sex           2.7479     0.2314  11.874  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 765.42  on 565  degrees of freedom
## Residual deviance: 526.42  on 563  degrees of freedom
## AIC: 532.42
##
## Number of Fisher Scoring iterations: 4
```

```
plot(model_without_age)
```


Residuals vs Fitted
Residuals
4
2
0
-2
-2
-1
0
1
2
Predicted values
glm(Survived ~ PClass + Sex)

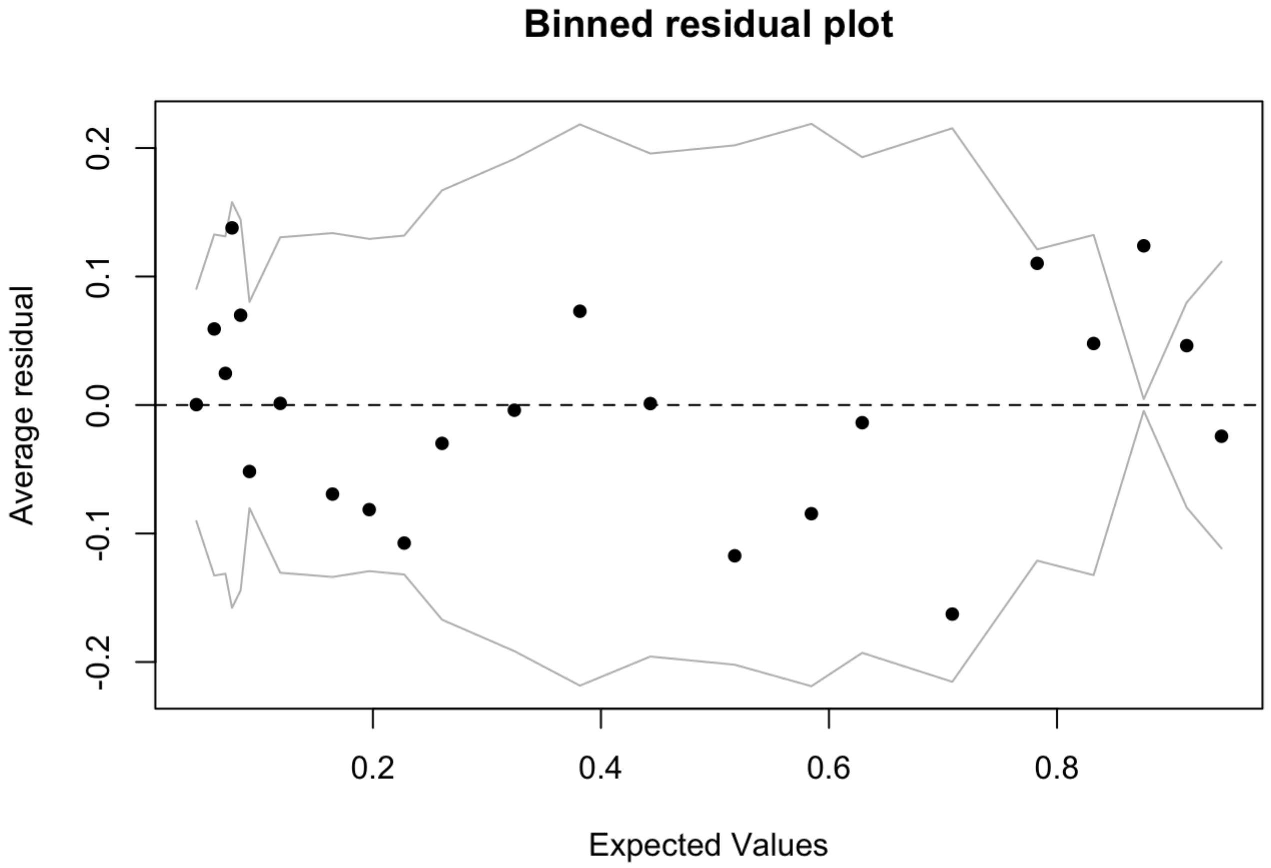

Normal Q-Q
Std. Pearson resid.
707
4
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
Theoretical Quantiles
glm(Survived ~ PClass + Sex)





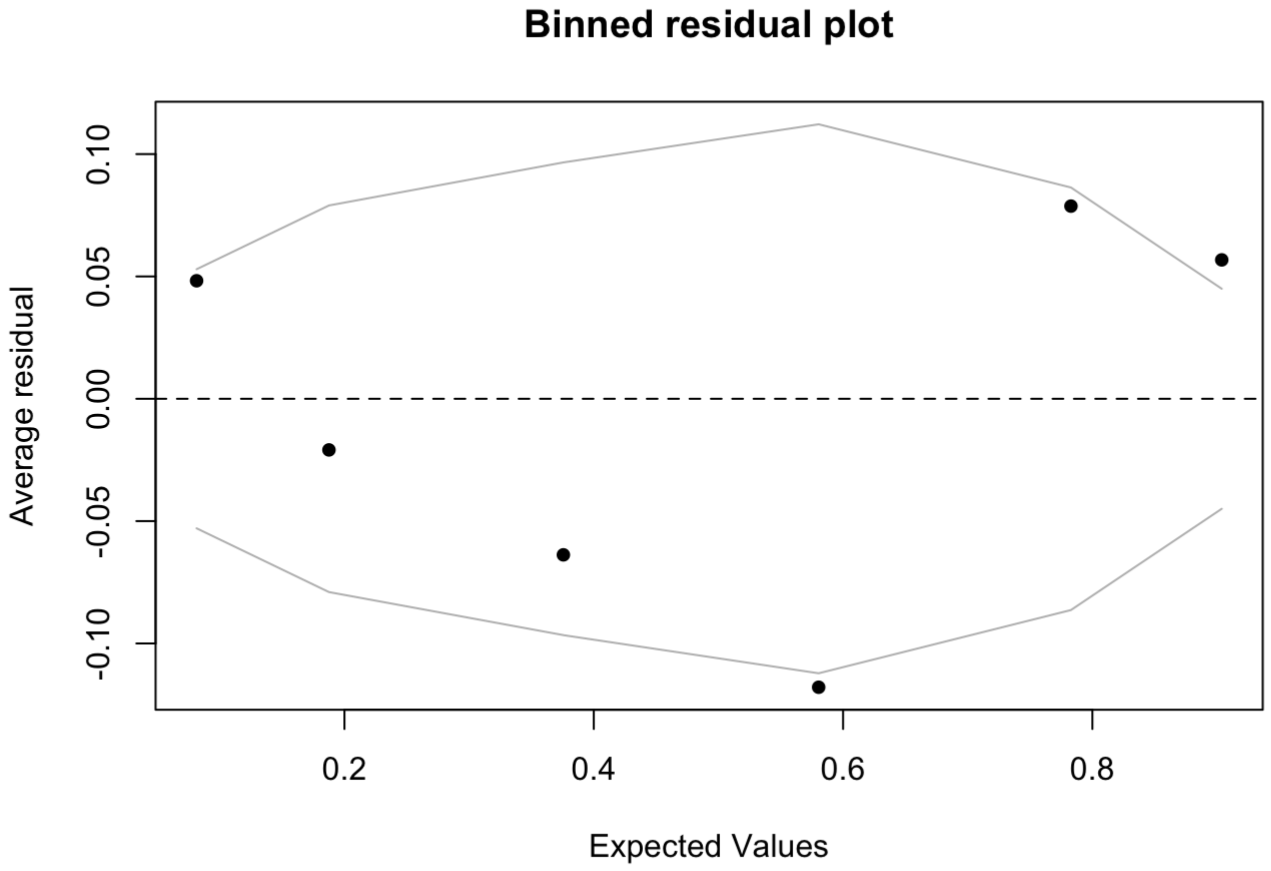
#Binned residual plot for model with age

```
#Majority of data points fall within standard error bands
binnedplot(fitted(model_with_age),
           residuals(model_with_age, type = "response"),
           nclass = NULL,
           xlab = "Expected Values",
           ylab = "Average residual",
           main = "Binned residual plot",
           cex.pts = 0.8,
           col.pts = 1,
           col.int = "gray")
```



#Binnedplot for model without age

```
#Majority of data points fall within standard error bands
binnedplot(fitted(model_without_age),
           residuals(model_without_age, type = "response"),
           nclass = NULL,
           xlab = "Expected Values",
           ylab = "Average residual",
           main = "Binned residual plot",
           cex.pts = 0.8,
           col.pts = 1,
           col.int = "gray")
```



#Pseudo R^2

```
ll.null <- model_with_age$null.deviance/-2
ll.proposed <- model_with_age$deviance/-2
(ll.null - ll.proposed) / ll.null
```

```
## [1] 0.3296625
```

#P-value of R^2

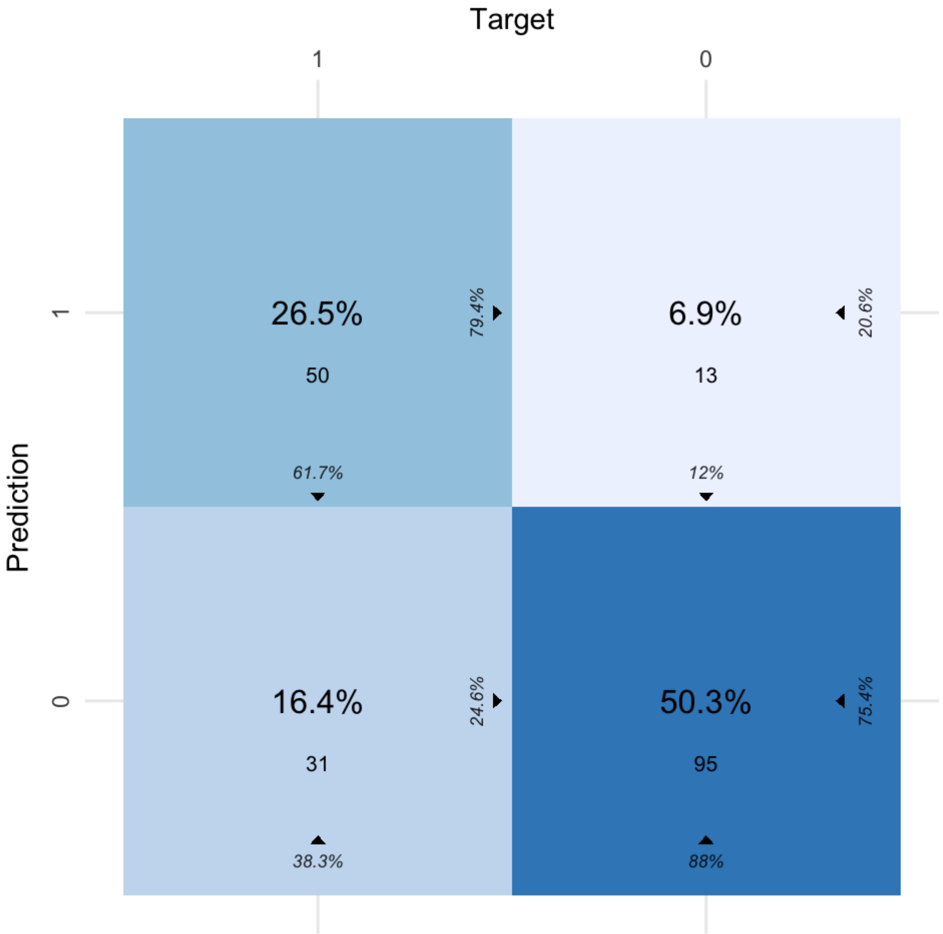
```
1-pchisq(2*(ll.proposed - ll.null), df = (length(model_with_age$coefficients)-1))
```

```
## [1] 0
```

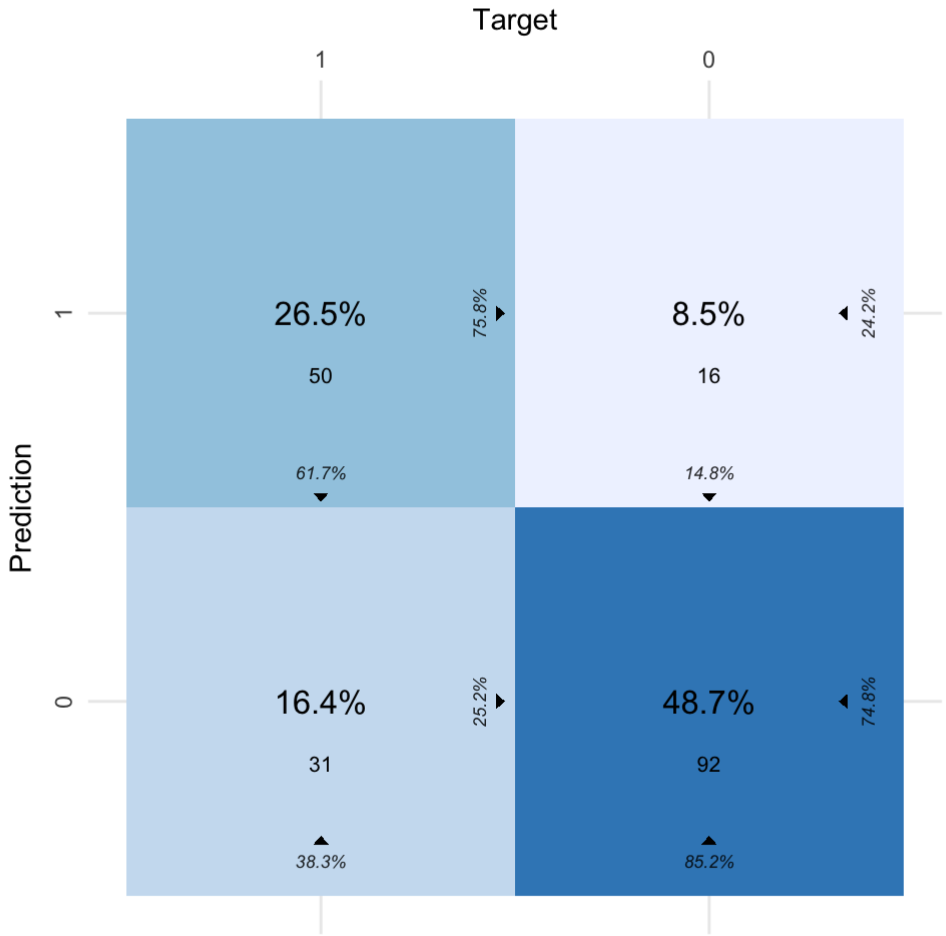
```
#so we know our R^2 is accurate
```

#Plotting Confusion Matrices

```
predict.1 <- predict(model_with_age, test, type = 'response')
#simple table
table_mat.1 <- table(test$Survived, predict.1 > 0.5)
# confusion matrix
f.1 <- tibble("target" = test$Survived,
            "prediction" = ifelse(predict.1 > 0.5, 1, 0))
f.1_table <- table(f.1)
cfm.1 <- as_tibble(f.1_table)
plot_confusion_matrix(cfm.1, target_col = "target", prediction_col = "prediction", count
s_col = "n")
```



```
predict.2 <- predict(model_without_age, test, type = 'response')
#simple table
table_mat.2 <- table(test$Survived, predict.2 > 0.5)
# confusion matrix
f.2 <- tibble("target" = test$Survived,
            "prediction" = ifelse(predict.2 > 0.5, 1, 0))
f.2_table <- table(f.2)
cfm.2 <- as_tibble(f.2_table)
plot_confusion_matrix(cfm.2, target_col = "target", prediction_col = "prediction", count
s_col = "n")
```



#Accuracy Test

```
accuracy_Test.1 <- sum(diag(table_mat.1)) / sum(table_mat.1)
accuracy_Test.1
```

```
## [1] 0.7671958
```

```
accuracy_Test.2 <- sum(diag(table_mat.2)) / sum(table_mat.2)
accuracy_Test.2
```

```
## [1] 0.7513228
```

#Precision Vs Recall

```
precision <- function(matrix) {
    # True positive
    tp <- matrix[2, 2]
    # false positive
    fp <- matrix[1, 2]
    return (tp / (tp + fp))
}
recall <- function(matrix) {
# true positive
    tp <- matrix[2, 2]# false positive
    fn <- matrix[2, 1]
    return (tp / (tp + fn))
}
prec.1 <- precision(table_mat.1)
prec.1
```

```
## [1] 0.7936508
```

```
prec.2 <- precision(table_mat.2)
prec.2
```

```
## [1] 0.7575758
```

```
rec.1 <- recall(table_mat.1)
rec.1
```

```
## [1] 0.617284
```

```
rec.2 <- recall(table_mat.2)
rec.2
```

```
## [1] 0.617284
```

#Harmonic Mean of precision and recall

```
f1 <- 2 * ((prec.1 * rec.1) / (prec.1 + rec.1))
f1
```

```
## [1] 0.6944444
```

```
f1.other <- 2 * ((prec.2 * rec.2) / (prec.2 + rec.2))
f1.other
```

```
## [1] 0.6802721
```

#The ROC curve

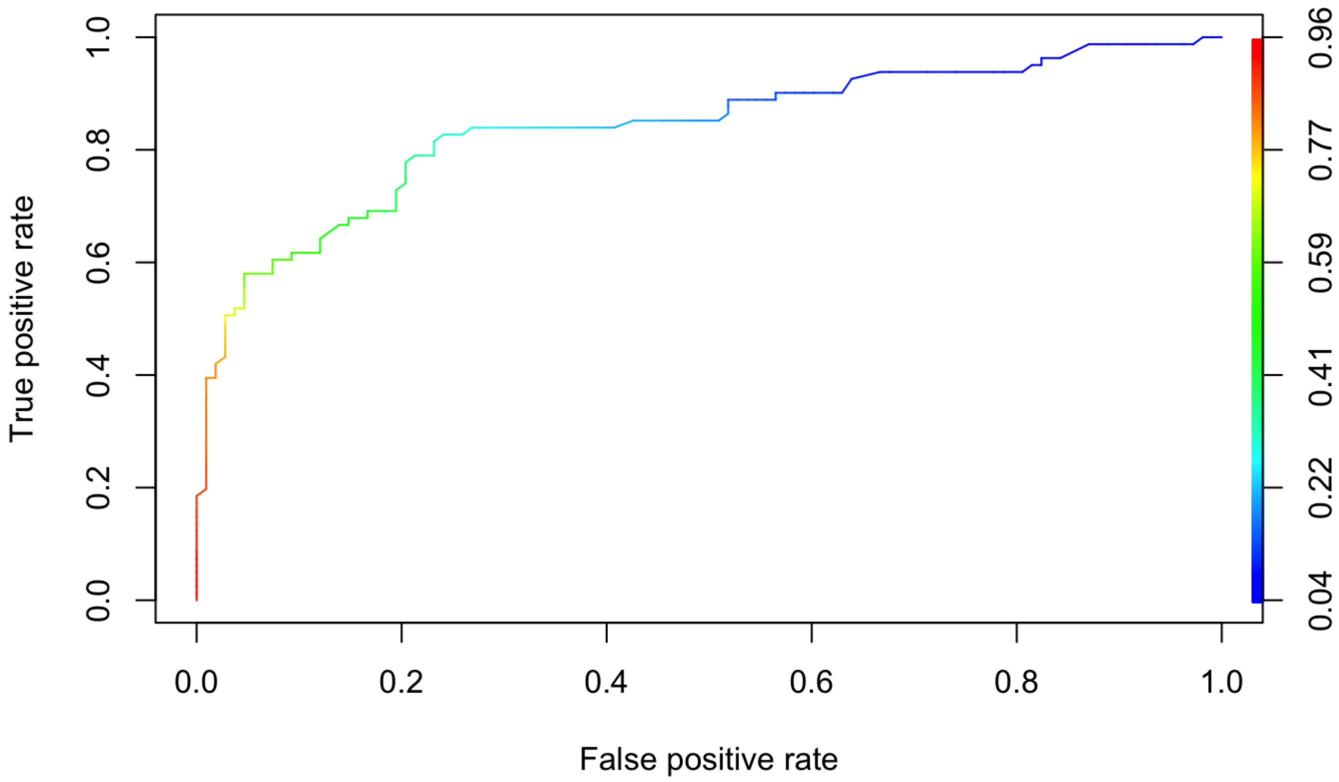
```
library(ROCR)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
ROCRpred <- prediction(predict.1, test$Survived)
ROCRperf <- performance(ROCRpred, 'tpr', 'fpr')
plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2, 1.7))
```



```
auc_ROCR <- performance(ROCRpred, measure = "auc")
auc_ROCR <- auc_ROCR@y.values[[1]]
auc_ROCR
```

```
## [1] 0.8389346
```