Titanic Analysis Project

Jawad Bin Anwar, Kristina Chaikina, Zofia Samsel

2024-01-26

Final project of the course Statistics.

This project concentrated on the Titanic dataset. It aims to create Naive Bayes Classifier using Rstudio.

Data exploration and description

Exercise 1

Download the training data titanic_train.Rdata

```r
#at first loading the data
load("/Users/kristina/Documents/repos/UPC/Statistics/titanic-statistics/data/titanic_train.Rdata")

#dataset is labeled as "train"

#checking the variables
str(train)
```

```
## 'data.frame':    594 obs. of  12 variables:
##  $ PassengerId: int  707 706 566 244 825 754 751 649 463 438 ...
##  $ Survived   : int  1 0 0 0 0 0 1 0 0 1 ...
##  $ Pclass     : int  2 2 3 3 3 3 2 3 1 2 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 439 561 205 505 635 420 859 872 285 69
##  $ Sex        : Factor w/ 2 levels "female","male": 1 2 2 2 2 2 1 2 2 1 ...
##  $ Age        : num  45 39 24 22 2 23 4 NA 47 24 ...
##  $ SibSp      : int  0 0 2 0 4 0 1 0 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 1 0 0 3 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 109 171 520 659 250 352 235 621 7 238 ...
##  $ Fare       : num  13.5 26 24.15 7.12 39.69 ...
##  $ Cabin      : Factor w/ 147 levels "A10","A14","A16",..: NA NA NA NA NA NA NA NA 134 NA ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 3 3 3 3 3 3 3 3 3 3 ...
```

```r
# Checking data types
var_types <- sapply(train, class)
var_types
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##   "integer"   "integer"   "integer"    "factor"    "factor"   "numeric"
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##   "integer"   "integer"    "factor"   "numeric"    "factor"    "factor"
```

Discussion: We can see that there are a total of 12 variables

Exercise 2

Explore the data frame : — How many observations and variables are there ?

```r
# Check the dimensions of the titanic frame
dim(train)
```

```
## [1] 594  12
```

There are 594 observations and 12 variables

— What are the qualitative variables ? And the quantitative ones ?

```r
#classifying into qualitative variables and quantitative variables
qualitative_vars <- names(var_types[var_types %in% c("factor", "character")])
quantitative_vars <- names(var_types[var_types %in% c("integer", "numeric")])

#how many data points we have (like dataframe rows)
n_observations <- nrow(train)

#how many variables we have (like dataframe columns)
n_variables <- ncol(train)

#concatenating and printing
cat("Total data points:", n_observations, "\n")
```

```
## Total data points: 594
```

```r
cat("Total variables:", n_variables, "\n")
```

```
## Total variables: 12
```

```r
#qualitative variables
cat("\nQualitative variables:\n")
```

```
## 
## Qualitative variables:
```

```r
print(qualitative_vars)
```

```
## [1] "Name"     "Sex"      "Ticket"   "Cabin"    "Embarked"
```

```r
#quantitative variables
cat("\nQuantitative variables:\n")
```

```
## 
## Quantitative variables:
```

```
print(quantitative_vars)
```

```
## [1] "PassengerId" "Survived"    "Pclass"      "Age"         "SibSp"
## [6] "Parch"       "Fare"
```

Discussion: Even though "Pclass" and "Survived" are classified as quantitative variables, such is not the case as they are encoded to denote different categories. For instance, Pclass has three categories each of which are denoted by numbers 1, 2 & 3. Again, the "Survived" variable has two values, 0 denotes not survived and 1 means survived. \n Therefore even though they are classified as quantitative variable, in reality they are treated as categorical variables.

Exercise 3

Describe S, Sx, P and A using the most appropriate descriptive statistical tools, including both summary and graphical statistics.

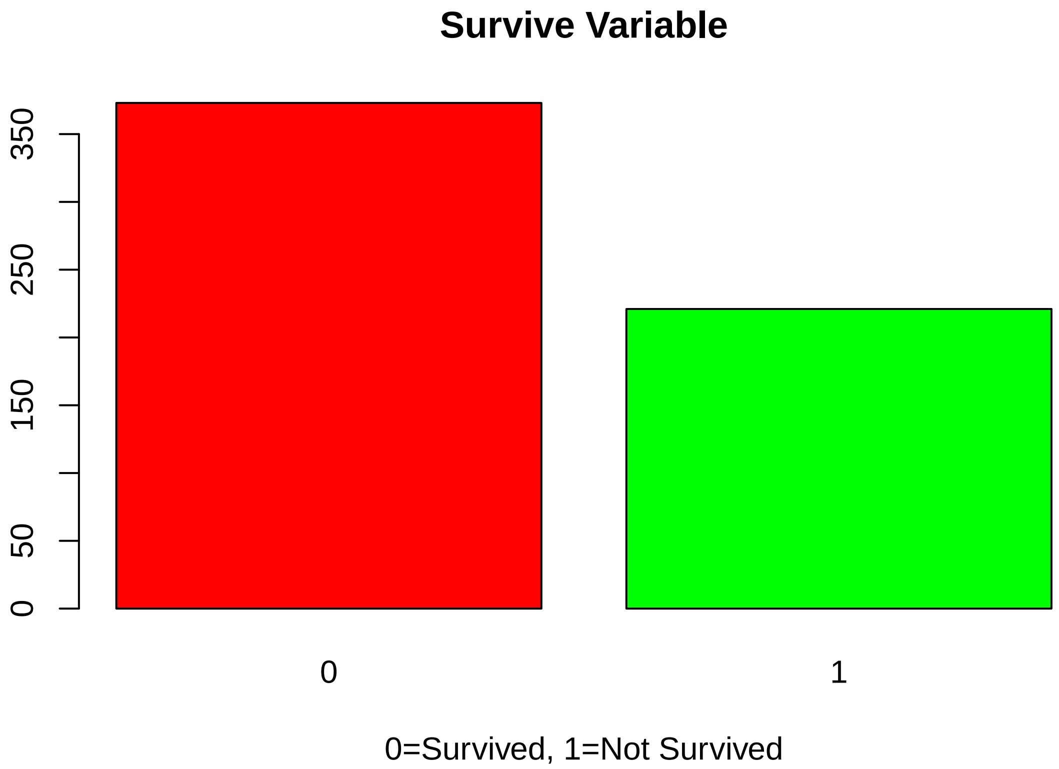
```
#Survived variable
train_survived <- na.omit(train$Survived)
summary(train_survived)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3721  1.0000  1.0000
```

```
table(train_survived)
```

```
## train_survived
##   0   1
## 373 221
```

```
barplot(table(train_survived), main="Survive Variable", xlab="0=Survived, 1=Not Survived", col=c("red",
```



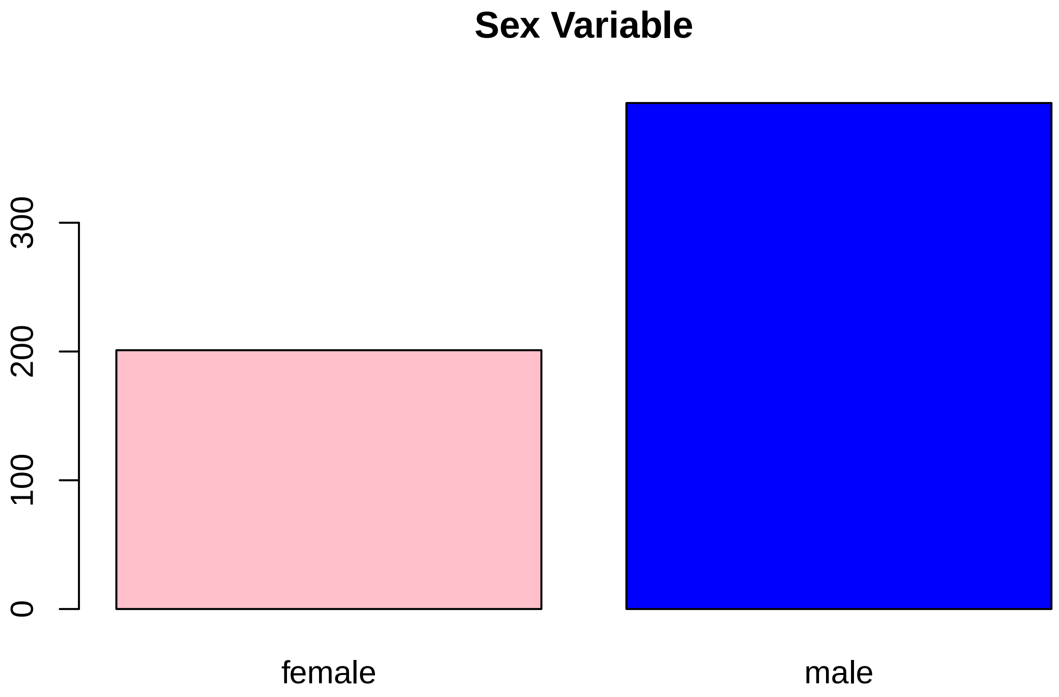
```
#Sex variable
train_sex <- na.omit(train$Sex)
summary(train_sex)
```

```
## female   male
##    201    393
```

```
table(train_sex)
```

```
## train_sex
## female   male
##    201    393
```

```
barplot(table(train_sex), main="Sex Variable", col=c("pink", "blue"))
```



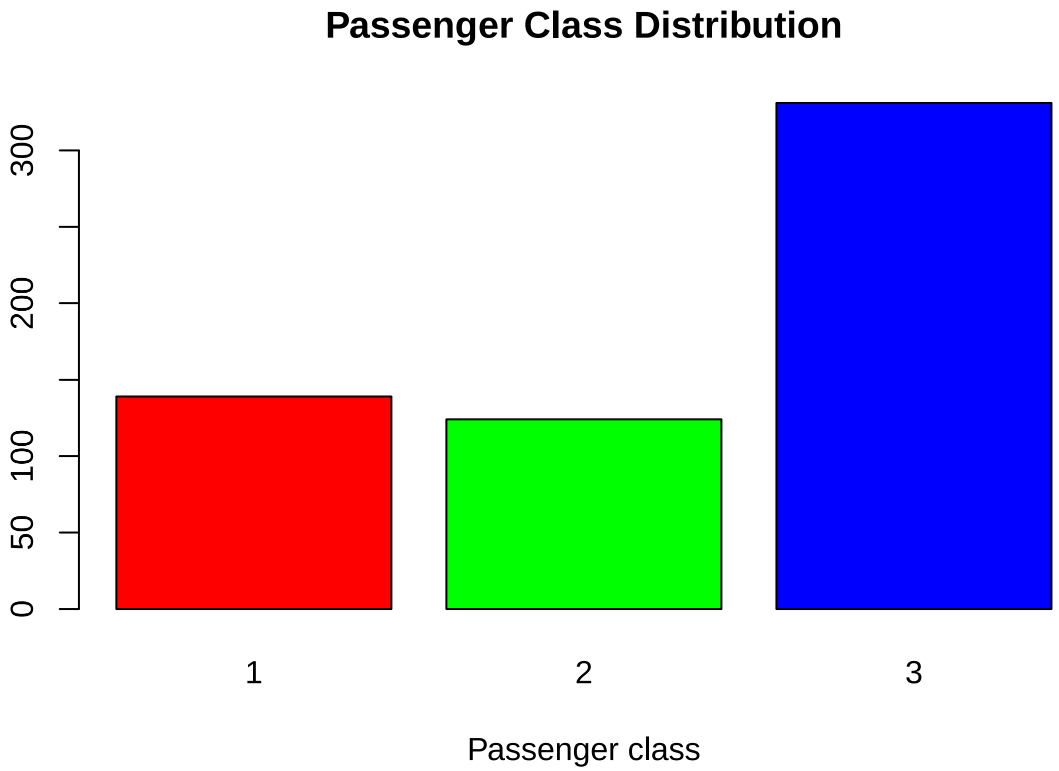
```
#Pclass variable
train_pclass <- na.omit(train$Pclass)
summary(train_pclass)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.323   3.000   3.000
```

```
table(train_pclass)
```

```
## train_pclass
##   1   2   3
## 139 124 331
```

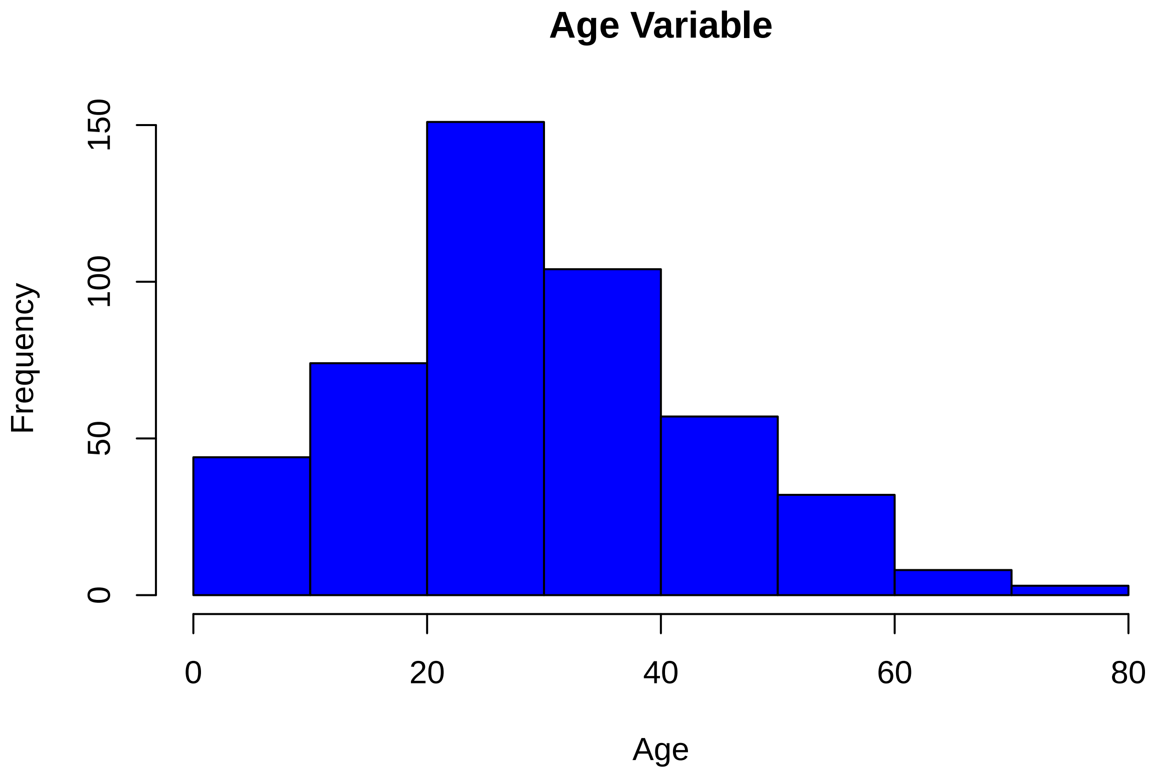
```
barplot(table(train_pclass), main="Passenger Class Distribution", xlab="Passenger class", col=c("red",
```



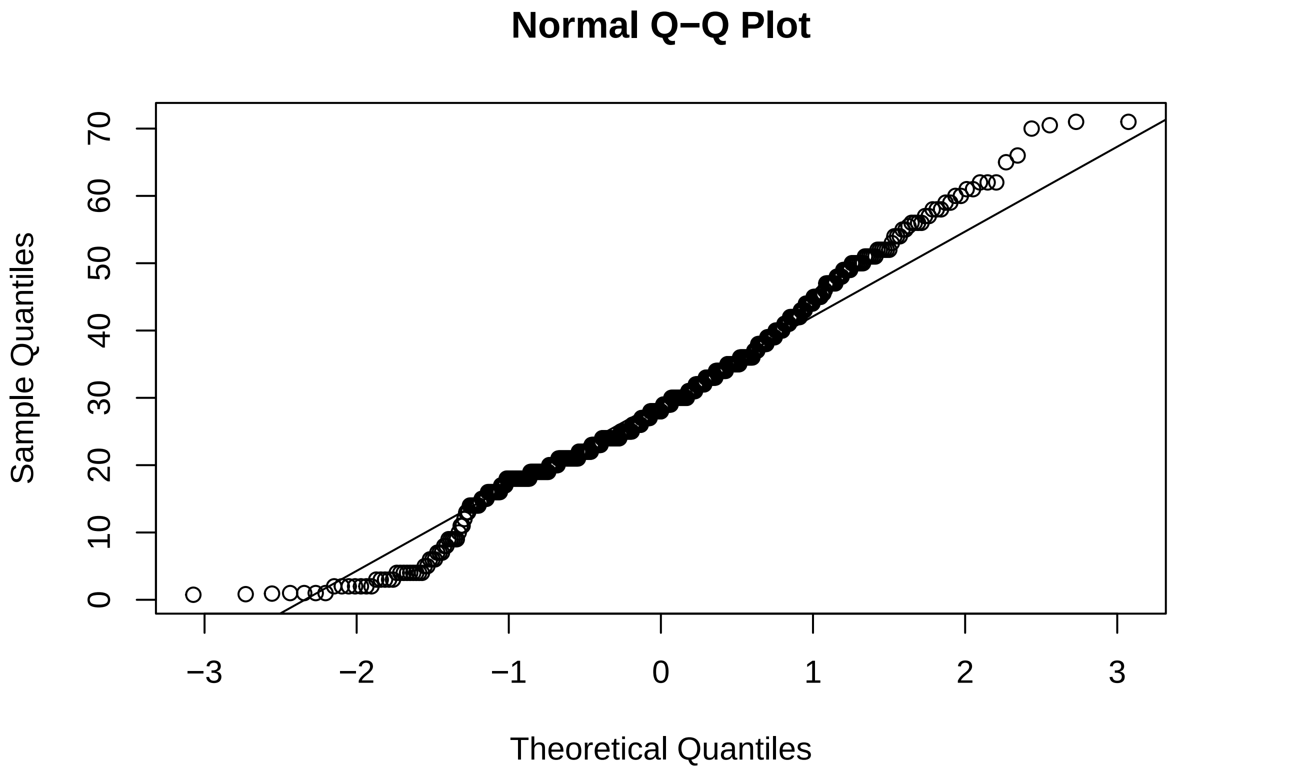
```
#Age varialbe
train_age <- na.omit(train$Age)
summary(train_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.75   21.00   28.00   29.58   38.00   71.00
```

```
hist(train_age, main="Age Variable", xlab="Age", col="blue", border="black")
```



```
qqnorm(train_age);qqline(train_age)
```



Discussion: For each categories, at first the "NaN" values were removed. From the barplot and summary of the "Survived" variable, we can observe that there were 288 people who did not survive and 185 people who managed to survive. \n Secondly, we can observe that there were 173 female passengers and 300 male passengers. \n Thirdly, there were 118 passengers in passenger class 1, 119 passengers in passenger class 2 and 236 passengers in passenger class 3. \n Finally, the histogram reveals that the age groups are roughly normally distributed with a median age of 28 and standard deviation of 14.36. We also plotted the qqplot and from the plot, we can see that it almost aligns with the theoretical distribution.

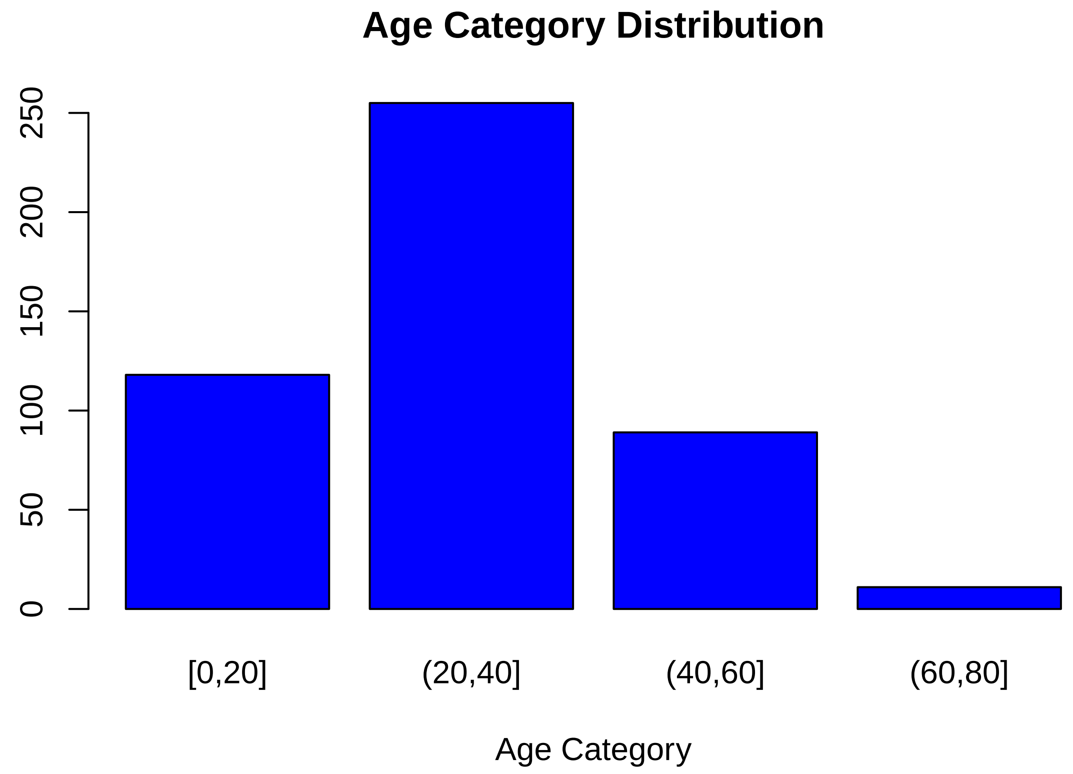
Exercise 4

We build a new variable **cAge** that categorizes **Age** by the following age categories : (0, 20], (20, 40], (40, 60] and (60, 80] years.

```
# Defining age categories
age_breaks <- c(0,  20,  40,  60,  80)
train_cAge <- cut(train_age, breaks = age_breaks, include.lowest = TRUE)
cAge <- train_cAge
summary(train_cAge)
```

```
##  [0,20] (20,40] (40,60] (60,80]
##     118     255      89      11
```

```
barplot(table(train_cAge), main="Age Category Distribution", xlab="Age Category", col="blue")
```



Discussion: The age categories are divided into 4. From the bar plot, we can see that most of the passengers are in the age range of 20 years to 40 years old.

Exercise 5

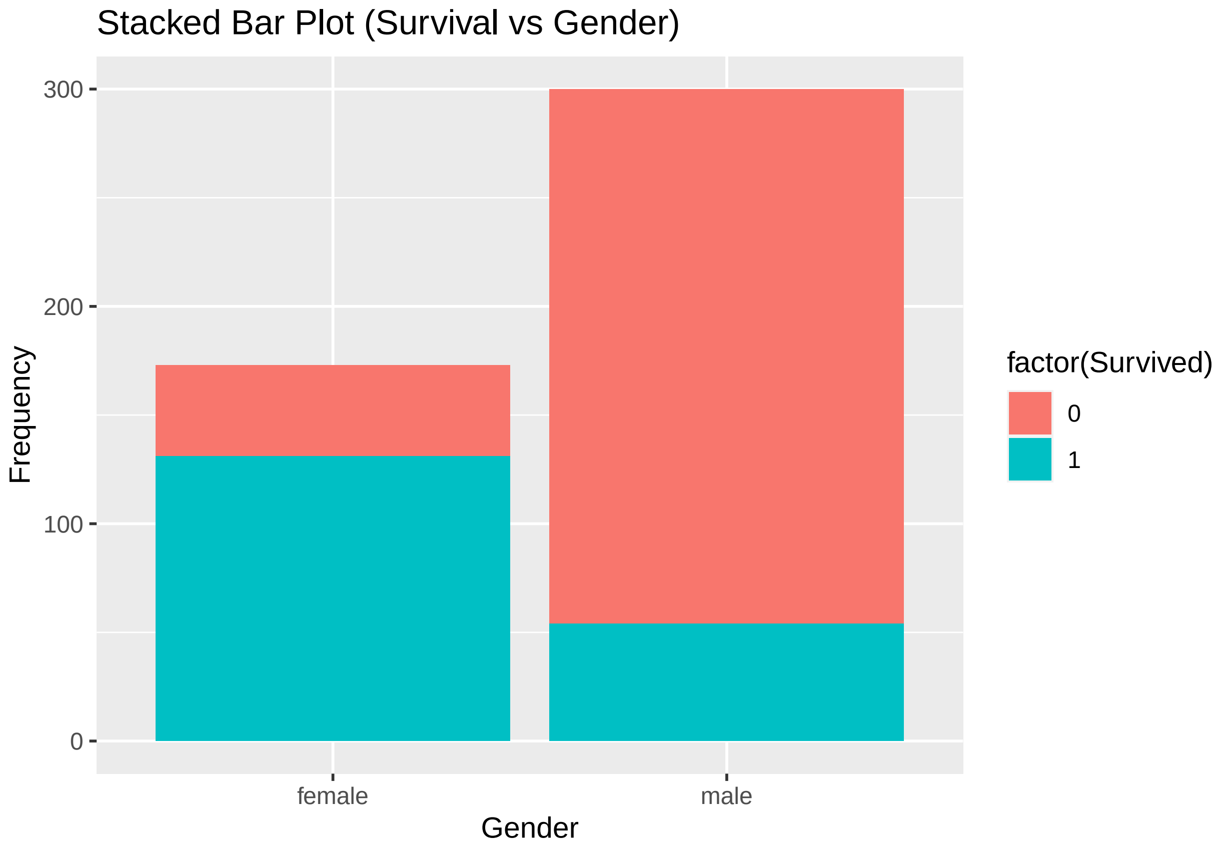
By using appropriate summary statistics and graphical tools describe the links between : — Sx and S — P and S — A and S — cA and S.

```
# Choosing the necessary variables
train <- train[, c('Age', 'Sex', 'Survived', 'Pclass')]

# Cleaning
train = na.omit(train)

# Defining variables
S <- train$Survived
Sx <- train$Sex
P <- train$Pclass
A <- train$Age

library(ggplot2)

#Survival vs Gender
table(train$Sex, train$Survived)
```

```
##
```

```
##           0   1
##   female  42 131
##   male   246  54
```

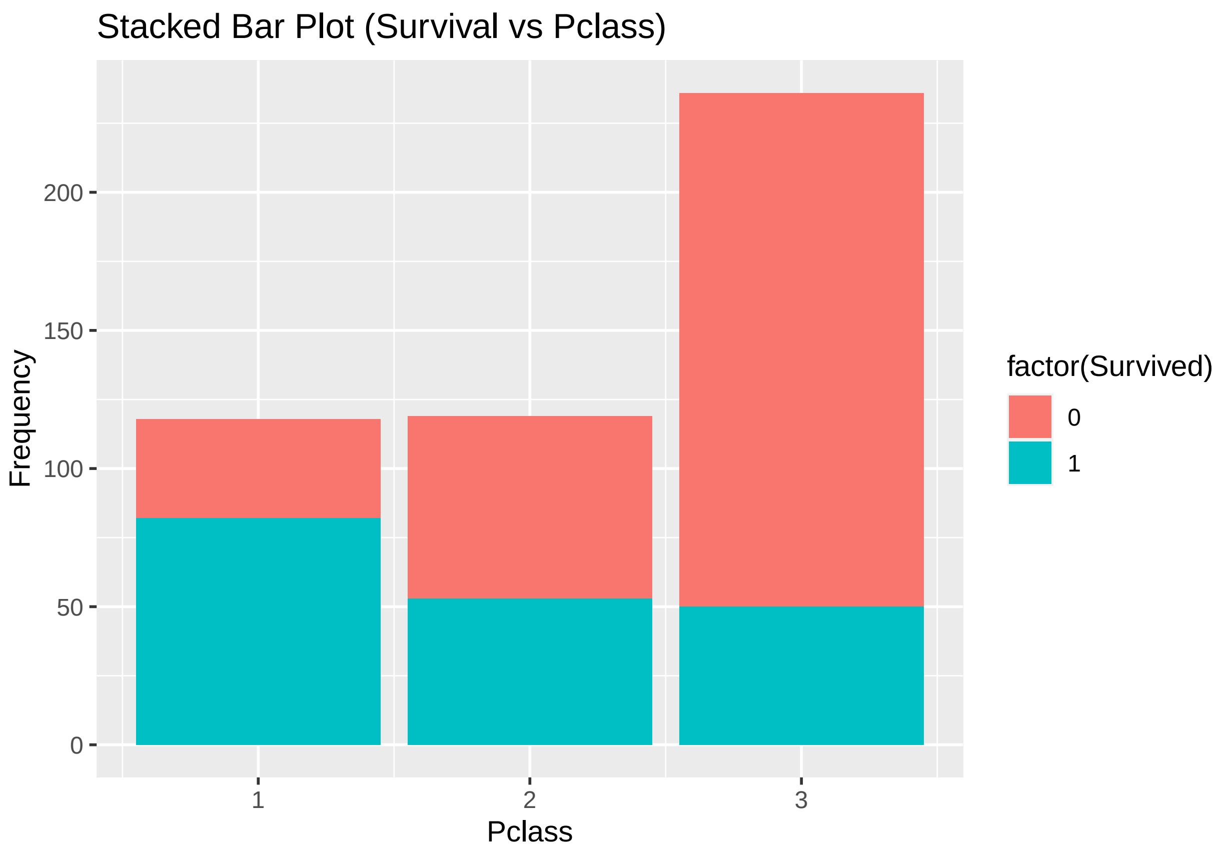
```
ggplot(train, aes(x = Sex, fill = factor(Survived))) +
  geom_bar(position = "stack") +
  labs(title = "Stacked Bar Plot (Survival vs Gender)", x = "Gender", y = "Frequency")
```



```
#Survival vs Pclass
table(train$Pclass, train$Survived)
```

```
##
##        0   1
##   1  36  82
##   2  66  53
##   3 186  50
```

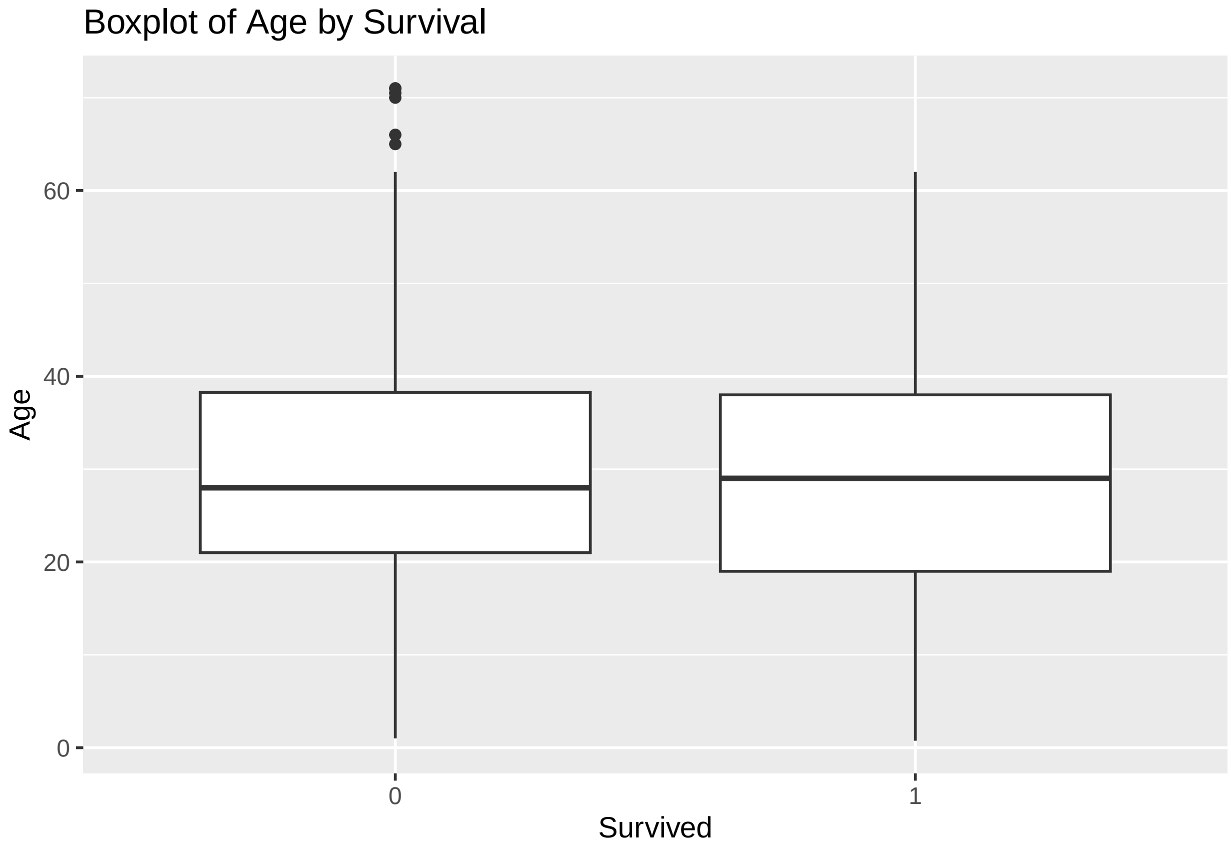
```
ggplot(train, aes(x = Pclass, fill = factor(Survived))) +
  geom_bar(position = "stack") +
  labs(title = "Stacked Bar Plot (Survival vs Pclass)", x = "Pclass", y = "Frequency")
```



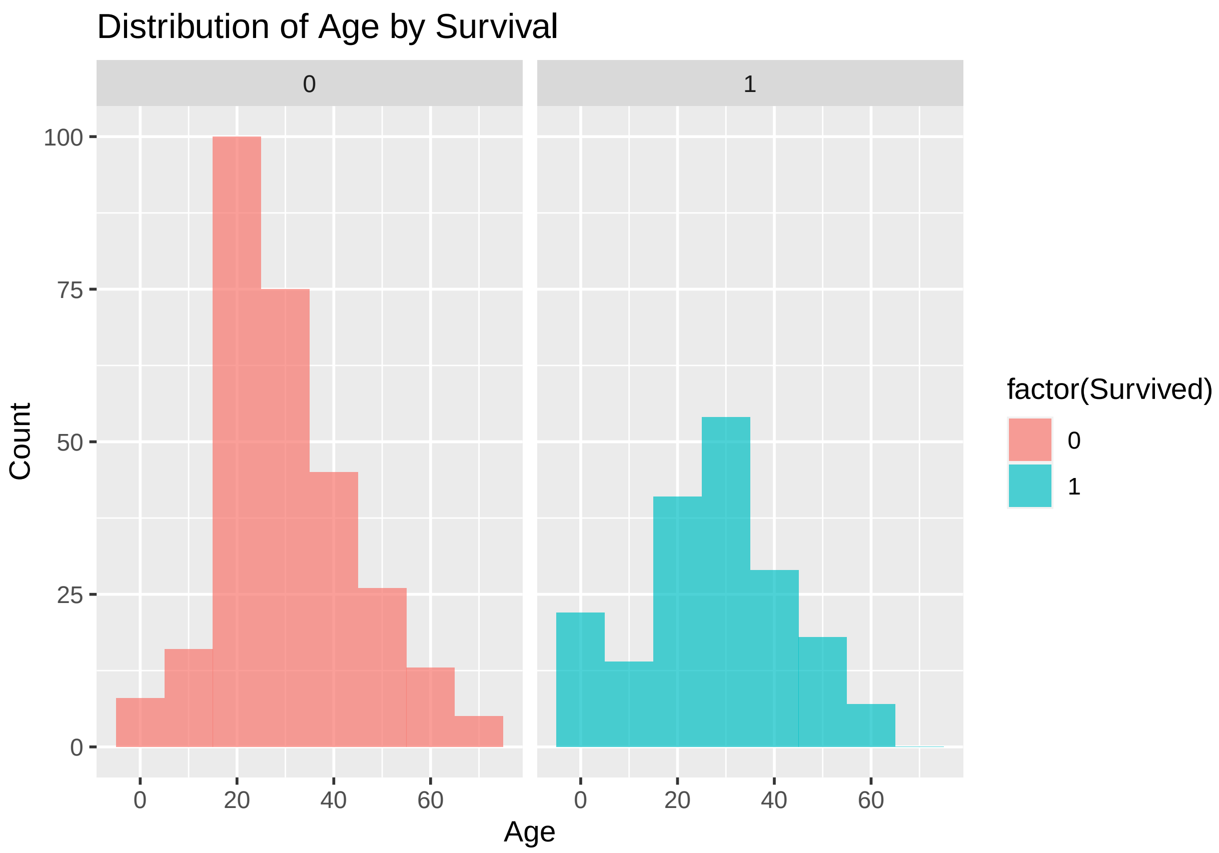
```
#Survival vs Age (Boxplot)
aggregate(Age ~ Survived, data = train, sd)
```

```
##   Survived      Age
## 1        0 13.89919
## 2        1 15.01455
```

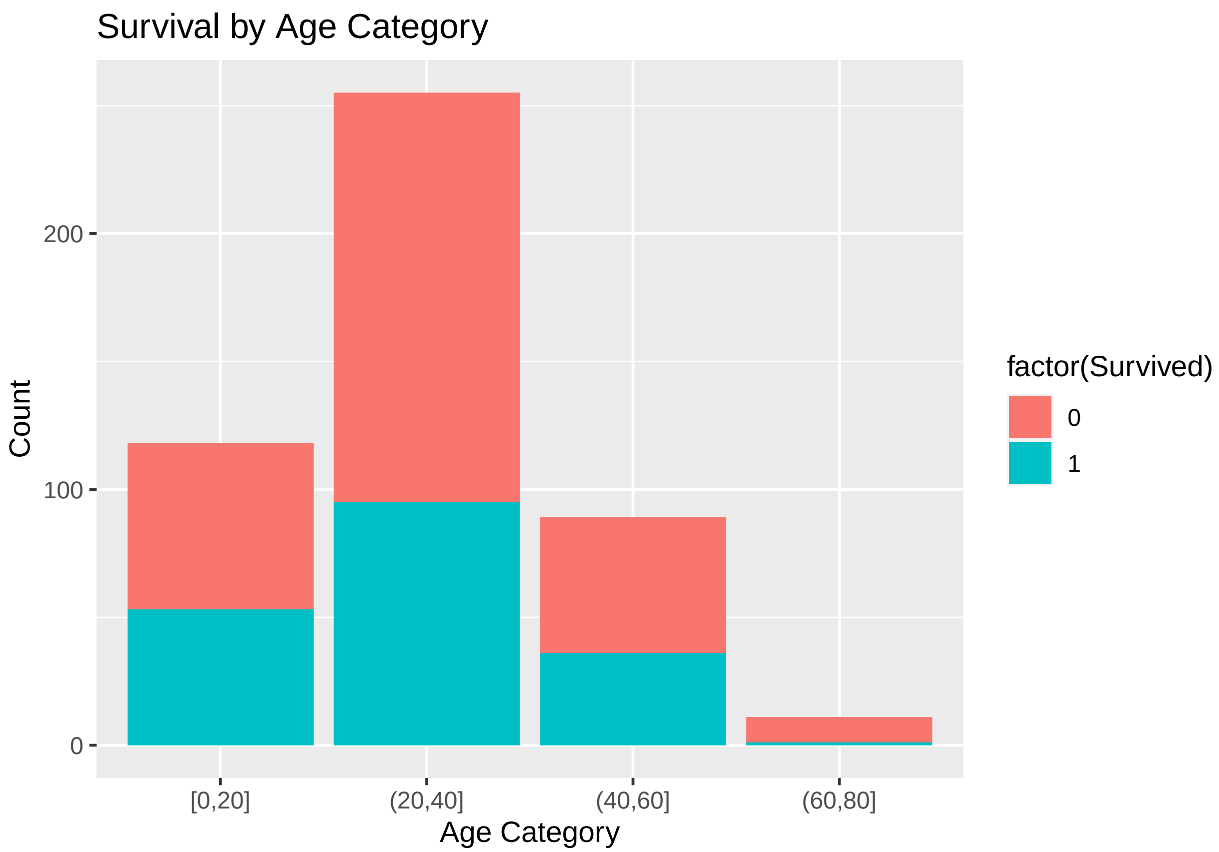
```
ggplot(train, aes(x = factor(Survived), y = Age)) +
  geom_boxplot() +
  labs(title = "Boxplot of Age by Survival", x = "Survived", y = "Age")
```



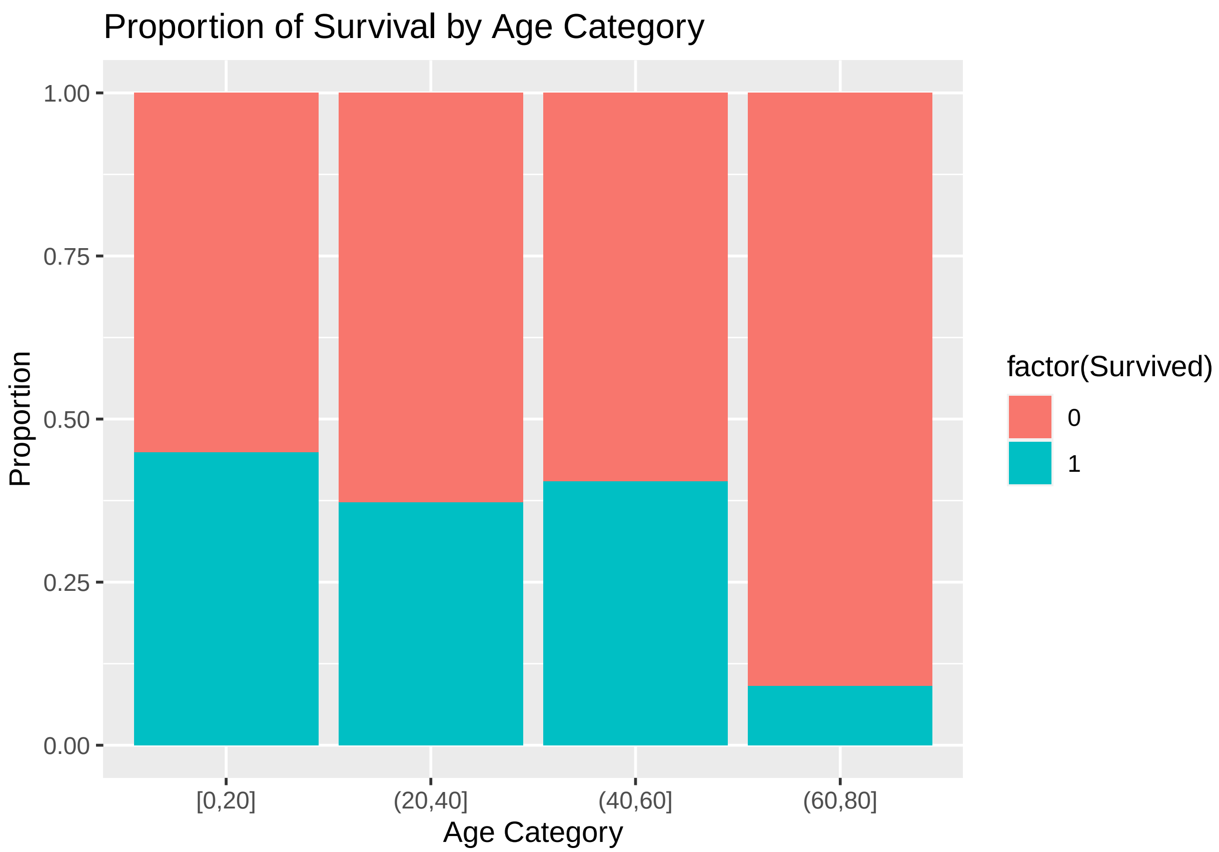
```r
# Survival vs Age (Faceted Histogram)
ggplot(train, aes(x = Age, fill = factor(Survived))) +
  geom_histogram(binwidth = 10, position = "identity", alpha = 0.7) +
  facet_grid(~Survived) +
  labs(title = "Distribution of Age by Survival", x = "Age", y = "Count")
```



```r
#Survival vs cAge (Stacked barplot)
ggplot(train, aes(x = cAge, fill = factor(Survived))) +
  geom_bar(position = "stack") +
  labs(title = "Survival by Age Category", x = "Age Category", y = "Count")
```



```
#Survival vs cAge (Proportion barplot)
ggplot(train, aes(x = cAge, fill = factor(Survived))) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Survival by Age Category", x = "Age Category", y = "Proportion")
```



Discussion: We have resorted to different several different summary statistics to capture the bivariate relationships. \n 1. Gender vs Survived: Since both of the variables are categorical, we used a contingency table. For Visualization, we used a stacked bar plot. From the stacked bar plot, we could observe that female passengers had more survival than male passengers. 2. Passenger class vs Survived: Since both of the variables are categorical, we used a contingency table. For Visualization, we used a stacked bar plot. \n 3. Age vs Survived: Since one variable is quantitative (Age) and another is categorical (Survived) we used aggregate fuunction on the basis of standard deviation. We found that passengers who did not survive had a standard deviation of 13.9 and people who did survive had a standard distribution of 15. Also for visualization, we showed the boxplot categorized on gender. We also showed the histogram categorized on gender. \n 4. cAge vs Survived: Since cAge has been transformed into a categorical variable from the quantitative variable Age, and we are comparing it against another categorical variable (Survived), we used a contingency table. Also for visualization, we used a stacked bar plot. But the stacked bar plot was hard to interpret. Then we used a proportion barplot. The proportion barplot revealed that age group (0,20] had the highest proportion of survival and age group (60,80] had the lowest proportion of survival.

Exercise 6

For P (Pclass):

- Null Hypothesis (H0): There is no significant difference in the survival rates among different passenger classes (Pclass).
- Alternative Hypothesis (H1): Survival rates differ significantly between passenger classes.

For Sx (Sex):

- Null Hypothesis (H0): There is no significant difference in the survival rates between male and female passengers.

- Alternative Hypothesis (H1): Survival rates differ significantly between male and female passengers.

For cA (Categorized Age):

- Null Hypothesis (H0): There is no significant association between age categories and survival.
- Alternative Hypothesis (H1): Survival rates differ significantly among different age categories.

Hypothesis testing

Exercise 7

a) We want to check **if Sx and S are independent**. Sx and S are categorical variables, so to test their independence we firstly have to create *contingency table.*

```
# create contingency table of sex and survived variables
table_freq_sex_survived <- table(S, Sx)
rownames(table_freq_sex_survived) = c("Non-Survivors", "Survivors")
colnames(table_freq_sex_survived) = c("Female", "Male")
table_freq_sex_survived
```

```
##                Sx
## S                Female Male
##   Non-Survivors     42  246
##   Survivors        131   54
```

Secondary, we have to do the X^2 test.

```
chisq.test(table_freq_sex_survived)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_freq_sex_survived
## X-squared = 151.1, df = 1, p-value < 2.2e-16
```

A Pearson's Chi-squared test with Yates' continuity correction revealed a significant association between variables X^2 (1, N = 594) = 198.48, p <0.001. The low p-value (< .05) indicates a rejection of the null hypothesis, suggesting a substantial dependency between the categorical variables (Sex, Survived).

b) We want to check **if cA and S are independent**. cA and S are categorical variables, so to test their independence we firstly have to create *contingency table.*

```
table_freq_age_survived <- table(S, cAge)
rownames(table_freq_age_survived) = c("Non-Survivors", "Survivors")
table_freq_age_survived
```

```
##                cAge
## S                [0,20] (20,40] (40,60] (60,80]
##   Non-Survivors     65     160      53      10
##   Survivors         53      95      36       1
```

Secondary, we have to do the X^2 test.

```
chisq.test(table_freq_age_survived, simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  table_freq_age_survived
## X-squared = 6.2678, df = NA, p-value = 0.09195
```

A Pearson's Chi-squared test with Yates' continuity correction revealed a significant association between variables X^2 (3, N = 594) = 6.268, p = 0.08896. The high p-value (> .05) indicates that we cannot reject the null hypothesis, suggesting independence between the categorical variables (cAge, Survived).

c) We want to check **if P and S are independent**. P and S are categorical variables, so to test their independence we firstly have to create *contingency table*.

```
table_freq_pclass_survived <- table(S, P)
rownames(table_freq_pclass_survived) = c("Non-Survivors", "Survivors")
table_freq_pclass_survived
```

```
##                P
## S                 1   2   3
##   Non-Survivors  36  66 186
##   Survivors      82  53  50
```

Secondary, we have to do the X^2 test.

```
chisq.test(table_freq_pclass_survived)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_freq_pclass_survived
## X-squared = 79.044, df = 2, p-value < 2.2e-16
```

A Pearson's Chi-squared test with Yates' continuity correction revealed a significant association between variables X^2 (2, N = 594) = 79.044, p < 0.001. The low p-value (< .05) indicates a rejection of the null hypothesis, suggesting a substantial dependency between the categorical variables (PClass, Survived).

Survival prediction

Exercise 8

Calculate conditional probability for:

- P(S = 1|Sx = female) - the conditional probability of surviving (S = 1) given that the the person was a female (Sx = female)

```
train$cA <- cAge

probability_survival_female <- sum(table_freq_sex_survived[2, "Female"]) /
  sum(table_freq_sex_survived[, "Female"])

cat("P(S = 1|Sx = female) =", probability_survival_female , "\n")
```

```
## P(S = 1|Sx = female) = 0.7572254
```

- P(S = 1|Sx = male) - the conditional probability of surviving (S = 1) given that the the person was a male (Sx = male)

```
probability_survival_male <- table_freq_sex_survived[2, "Male"] /
  sum(table_freq_sex_survived[, "Male"])

cat("P(S = 1|Sx = male) =", probability_survival_male, "\n")
```

```
## P(S = 1|Sx = male) = 0.18
```

- P(S = 1|P = 1) - the conditional probability of surviving (S = 1) given that the the person was in the first class (P = 1)

```
probability_survival_1_class <- table_freq_pclass_survived[2, '1'] /
  sum(table_freq_pclass_survived[, '1'])

cat("P(S = 1|P = 1) =", probability_survival_1_class, "\n")
```

```
## P(S = 1|P = 1) = 0.6949153
```

- P(S = 1|P = 2) - the conditional probability of surviving (S = 1) given that the the person was in the second class (P = 2)

```
probability_survival_2_class <- table_freq_pclass_survived[2, '2'] /
  sum(table_freq_pclass_survived[, '2'])

cat("P(S = 1|P = 2) =", probability_survival_2_class, "\n")
```

```
## P(S = 1|P = 2) = 0.4453782
```

- P(S = 1|P = 3) - the conditional probability of surviving (S = 1) given that the the person was in the third class (P = 3)

```
probability_survival_3_class <- table_freq_pclass_survived[2, '3'] /
  sum(table_freq_pclass_survived[, '3'])

cat("P(S = 1|P = 3) =", probability_survival_3_class, "\n")
```

```
## P(S = 1|P = 3) = 0.2118644
```

- P(S = 1|cA = (0,20]) - the conditional probability of surviving (S = 1) given that the the person was between 0 to 20 years old (cA = (0,20]))

```r
probability_survival_young <- table_freq_age_survived[2, '[0,20]'] /
  sum(table_freq_age_survived[, '[0,20]'])

cat("P(S = 1|cA = [0,20]) =", probability_survival_young, "\n")
```

```
## P(S = 1|cA = [0,20]) = 0.4491525
```

- P(S = 1|cA = (20,40]) - the conditional probability of surviving (S = 1) given that the the person was between 20 to 30 years old (cA = (20,40]))

```r
probability_survival_adults <- table_freq_age_survived[2, '(20,40]'] /
  sum(table_freq_age_survived[, '(20,40]'])

cat("P(S = 1|cA = (20,40]) =", probability_survival_adults, "\n")
```

```
## P(S = 1|cA = (20,40]) = 0.372549
```

- P(S = 1|cA = (40,60]) - the conditional probability of surviving (S = 1) given that the the person was between 40 to 60 years old (cA = (40,60]))

```r
probability_survival_2_adults <- table_freq_age_survived[2, '(40,60]'] /
  sum(table_freq_age_survived[, '(40,60]'])

cat("P(S = 1|cA = (40,60]) =", probability_survival_2_adults, "\n")
```

```
## P(S = 1|cA = (40,60]) = 0.4044944
```

- P(S = 1|cA = (60,80]) - the conditional probability of surviving (S = 1) given that the the person was between 60 to 80 years old (cA = (60,80]))

```r
probability_survival_3_adults <- table_freq_age_survived[2, '(60,80]'] /
  sum(table_freq_age_survived[, '(60,80]'])

cat("P(S = 1|cA = (60,80]) =", probability_survival_3_adults, "\n")
```

```
## P(S = 1|cA = (60,80]) = 0.09090909
```

Here we have calculated the conditional probabilities of the variable “Survival” based on 3 other variables.
1. Probability that a female person survives: 0.7572254
2. Probability that a male person survives: 0.18
3. Probability that a class 1 passenger survives: 0.6949153
4. Probability that a class 2 passenger survives: 0.4453782
5. Probability that a class 3 passenger survives: 0.2118644
6. Probability that a person between age (0,20] survives: 0.4491525
7. Probability that a person between age (20,40] survives: 0.372549
8. Probability that a person between age (40,60] survives: 0.4044944
9. Probability that a person between age (60,80] survives: 0.09090909

Exercise 9

In this exercise we want to prove that if predictors S_x, P, and cA are independent conditionally on S, i.e., if for each i:

$$P(S_x, P, cA|S = i) = P(S_x|S = i)P(P|S = i)P(cA|S = i) \quad (1)$$

then

$$P(S = 1|S_x, P, cA) = \frac{P(S_x|S = 1)P(P|S = 1)P(cA|S = 1)P(S = 1)}{\sum_{i=0} P(S_x|S = i)P(P|S = i)P(cA|S = i)P(S = i)}.$$

Let's assume:

- predictors S_x, P, and cA are independent, so
$$P(S_x, P, cA|S = i) = P(S_x|S = i)P(P|S = i)P(cA|S = i) \quad (1.1)$$
- conditional probability is
$$P(S_x, P, cA) = P(S_x, P, cA|S = 0)P(P|S = 0) + P(S_x, P, cA|S = 1)P(cA|S = 1) \quad (1.2)$$
- Bayes Formula is
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1.3)$$

The logic of proof:

$$P(S = 1|S_x, P, cA) \overset{\text{using 1.3}}{=}$$

$$\overset{\text{using 1.3}}{=} \frac{P(S_x, P, cA|S = 1)P(S = 1)}{P(S_x, P, cA)} \overset{\text{using 1.1}}{=}$$

$$\overset{\text{using 1.1}}{=} \frac{P(S_x|S = 1)P(P|S = 1)P(cA|S = 1)P(S = 1)}{P(S_x, P, cA)} \overset{\text{using 1.2}}{=}$$

$$\overset{\text{using 1.2}}{=} \frac{P(S_x|S = 1)P(P|S = 1)P(cA|S = 1)P(S = 1)}{P(S_x|S = 0)P(P|S = 0)P(cA|S = 0)P(S = 0) + P(S_x|S = 1)P(P|S = 1)P(cA|S = 1)P(S = 1)} =$$

$$= \frac{P(S_x|S = 1)P(P|S = 1)P(cA|S = 1)P(S = 1)}{\sum_{i=0} P(S_x|S = i)P(P|S = i)P(cA|S = i)P(S = i)}$$

Which finish the proof.

Exercise 10

Building the conditional probability tables corresponding to: - $P(S_x|S)$ (table with 2 rows and 2 columns), - $P(P|S)$ (3 rows, 2 columns), - $P(cA|S)$ (4 rows, 2 columns). The conditional probability tables represent the likelihood of certain events based on the values of specific variables within the context of survival (S) in the Titanic dataset. The tables include probabilities such as $P(\text{Sex}|\text{Survived})$, $P(\text{Pclass}|\text{Survived})$, and $P(\text{AgeCategory}|\text{Survived})$, providing insights into how these variables relate to the probability of survival. In columns, 0 represents Non-Survivors and 1 represents Survivors.

```r
# P(Sx|S)
S_Sx <- prop.table(table(train$Sex, train$Survived), margin = 2)

S_Sx
```

```
##
##                  0         1
##   female 0.1458333 0.7081081
##   male   0.8541667 0.2918919
```

```r
# P(P|S)
S_P <- prop.table(table(train$Pclass, train$Survived), margin = 2)
S_P
```

```
##
##             0         1
##   1 0.1250000 0.4432432
##   2 0.2291667 0.2864865
##   3 0.6458333 0.2702703
```

```r
# P(cA|S)
# Categorize Age using cut()
S_cA <- prop.table(table(train$cA, train$Survived), margin = 2)
S_cA
```

```
##
##                     0           1
##   [0,20]  0.225694444 0.286486486
##   (20,40] 0.555555556 0.513513514
##   (40,60] 0.184027778 0.194594595
##   (60,80] 0.034722222 0.005405405
```

```r
# P(S)
S <- prop.table(table(train$Survived))
S
```

```
##
##         0         1
## 0.6088795 0.3911205
```

The numbers in the rows labeled with categories represent the probability of falling into a specific category given the survival status.

For example: - $P((0,20]|S=0)$ is the probability of being in the age category (0,20] given that the passenger did not survive. This is approximately 0.226.

For easier understanding, below you can see the reverse representation of these tables.

```r
# P(Sx|S)
S_Sx_reverse <- prop.table(table(train$Survived, train$Sex), margin = 2)
rownames(S_Sx_reverse) <- c('Not Survived', 'Survived')
S_Sx_reverse
```

```
##
##                   female      male
##   Not Survived 0.2427746 0.8200000
##   Survived     0.7572254 0.1800000
```

```
# P(P|S)
S_P_reverse <- prop.table(table(train$Survived, train$Pclass), margin = 2)
rownames(S_P_reverse) <- c('Not Survived', 'Survived')
S_P_reverse
```

```
##
##                        1         2         3
##   Not Survived 0.3050847 0.5546218 0.7881356
##   Survived     0.6949153 0.4453782 0.2118644
```

```
# P(cA|S)
S_cA_reverse <- prop.table(table(train$Survived, train$cA), margin = 2)
rownames(S_cA_reverse) <- c('Not Survived', 'Survived')
S_cA_reverse
```

```
##
##                    [0,20]    (20,40]    (40,60]    (60,80]
##   Not Survived 0.55084746 0.62745098 0.59550562 0.90909091
##   Survived     0.44915254 0.37254902 0.40449438 0.09090909
```

Some observations. 1. $P(\textbf{Sex}|\textbf{Survived})$**:** - The table shows that the probability of survival for females is significantly higher (0.688) compared to males (0.312).

2. $P(\textbf{Pclass}|\textbf{Survived})$**:**

    - The probability of survival increases with increasing passenger class. The passengers of the 1 class had the highest probability of survival, while the passengers of the 3 class had the lowest probability of survival.

3. $P(\textbf{AgeCategory}|\textbf{Survived})$**:**

    - While the oldest age category 60-80 exhibited the lowest probability of survival, the distinctions among other age groups were less pronounced. Notably, passengers in the 0-20 age category demonstrated the highest probability of survival.

These trends suggest that being female, belonging to a higher passenger class, and falling within 20-40 age categories were associated with higher probabilities of survival in the Titanic dataset.

Exercise 11

Function that gives as an outcome the probability of survival that corresponds to the input values of Sex, Pclass, cAge.

```
prob_prediction <- function(Sex, Pclass, cAge) {
  # P(S = 1|Sx, P, cA) = P(S = 1) * P(Sx|S) * P(P|S) * P(cA|S)
  # the probability of survival of a person with certain characteristics
  probability_1 <- S['1'] * S_Sx[Sex, '1'] * S_P[Pclass, '1'] * S_cA[cAge, '1']
  # P(S = 0|Sx, P, cA) = P(S = 0) * P(Sx|S) * P(P|S) * P(cA|S)
  # the probability of not survival of a person with certain characteristics
  probability_0 <- S['0'] * S_Sx[Sex, '0'] * S_P[Pclass, '0'] * S_cA[cAge, '0']

  # Normalize probabilities of survival
  probability_1_normalized <- probability_1 / (probability_1 + probability_0)

  return(probability_1_normalized)
}
```

For example, below you can see a probability of female, 3-class, 40-60 y.o. passenger surviving.

```
result <- prob_prediction(Sex = 'female', Pclass = '3', cAge = '(40,60]')
print(result)
```

```
##         1
## 0.5798702
```

Evaluating the classifier's performance

Exercise 12

Download and load the testing data titanic_test.Rdata

To load the data load function was used.

It will be used as a second sample of passengers to evaluate the quality of theprediction model that was trained on the training dataset. There are 66 rows in the test dataset.

```
# Load the titanic_test.Rdata titanic file

load('/Users/kristina/Documents/repos/UPC/Statistics/titanic-statistics/data/titanic_test.Rdata')

# Check the dimensions of the titanic frame
dim(test)
```

```
## [1] 66  4
```

```
head(test)
```

```
##    Survived Pclass    Sex    cAge
## 7         0      1   male (40,60]
## 24        1      1   male (20,40]
## 28        0      1   male  (0,20]
## 55        0      1   male (60,80]
## 63        0      1   male (40,60]
## 67        1      2 female (20,40]
```

In the test dataset we have 66 observations with 5 variables.

Exercise 13

Task: Using the function prob_prediction() to predict the survival probability of each passenger in test. The function 'prob_probability()' can by used to predict the survival probability of each passenger of the test set.

The prob_function() was apply on each of the element of the testset. The values were saved as a new column of the test dataframe.

```
# create a vector with predictions for the test set
test$Survival_Probability <- sapply(1:nrow(test), function(x) prob_prediction(Sex = test[x,'Sex'], Pcla

head(test)
```

```
##    Survived Pclass    Sex    cAge Survival_Probability
## 7         0      1   male (40,60]            0.4514755
## 24        1      1   male (20,40]            0.4184268
## 28        0      1   male  (0,20]            0.4969917
## 55        0      1   male (60,80]            0.1080782
## 63        0      1   male (40,60]            0.4514755
## 67        1      2 female (20,40]            0.7828032
```

Exercise 15

Task: Predict the survival status of each passenger in test with the Maximum a Posteriori Probability (MAP) rule : a passenger will be classified as a survivor if and only if her survival probability is > 0.5.

In this ex. the function survival_status() was created. The function takes a probability of the survival and check whether it is greater or lower then 0.5 (Maximum Posterior Probability) and based on categorize whether the person is a survivor or non-survivor.

```
survival_status <- function(prob_prediction) {
  # Function to determine survival status based on probability prediction
  # Rule: If the probability prediction is greater than 0.5, the person is
  # classified as survived
  if (prob_prediction > 0.5){
    status = 'survivor'
  }
  # Otherwise, they are classified as non-survived
  else {
    status = 'non-survivor'
  }
  return(status)
}
```

The function is applayed on the test set. The vector predicted_survival collect all the prediction.

```
# create a vector with predictions for the test set
predicted_survival <- sapply(1:nrow(test), function(x)
  survival_status(prob_prediction(Sex = test[x,'Sex'], Pclass = test[x,'Pclass'],
                                  cAge = test[x,'cAge'])))
```

For the better visibility the predicted categories are added to the dataframe as a new column Predicted_Survival.

```
#add the vector as a new column in test table
test$Predicted_Survival <- as.integer(sapply(1:nrow(test), function(x)
  predicted_survival[x] == 'survivor'))

head(test,5)
```

```
##    Survived Pclass  Sex    cAge Survival_Probability Predicted_Survival
## 7         0      1 male (40,60]            0.4514755                  0
## 24        1      1 male (20,40]            0.4184268                  0
## 28        0      1 male  (0,20]            0.4969917                  0
## 55        0      1 male (60,80]            0.1080782                  0
## 63        0      1 male (40,60]            0.4514755                  0
```

Based on the table we can see that the predicted values are aligning with the real survival information. However, there is an row, where the prediction is wrong. To test how well the model predicts, the accuracy was calcualated.

Exercise 15

Task: Compare the vector with the predictions obtained in the previous questions with the column containing the true survival status in the test dataset. More specifically, calculate the proportion of test passengers well classified. This performance measure is known as the accuracy of the classifier.

In this exercise the accuracy() function was created. It compare the prediction with the real information about the person survival. Based on it the accuracy of the model is calcualted.

```
accuracy <- function(test) {
  # Compare prediction with the true survival status for all rows in the set
  comparision <- as.integer(sapply(1:nrow(test), function(x)
  test[x,'Predicted_Survival'] == test[x,'Survived']))
  # return an average of sum the number or correct predictions
  return (sum(comparision)/nrow(test))
}

accuracy(test)
```

```
## [1] 0.7424242
```

Our model has an accuracy equal to 0.74.

Exercise 16

Task: As said in Q.9, the naive Bayes classifier assumes that equation (1) holds. But how reasonable is this assumption ? Test whether — Sx and cA — Sx and P — cA and P are independent in each stratum of S (i.e. first considering only passengers with S = 1 and then only passengers with S = 0).

In this exercise the assumption of independence of the variables is tested.

1) check if **Sx and cA are independent for the S=0**. Sx and cA are categorical variables, so to test their independence firstly have to create *contingency table*.

```r
train <- na.omit(train)
non_survived_passangers_train <- train[train$Survived == 0, ]

# create contingency table of sex and age variables where S = 0
table_freq_Sx_cA_non_s <- table(non_survived_passangers_train$Sex, non_survived_passangers_train$cA)
table_freq_Sx_cA_non_s
```

```
##
##          [0,20] (20,40] (40,60] (60,80]
##   female     17      19       6       0
##   male       48     141      47      10
```

Secondary, X^2 test was done. However, since the categories doesn't have a lot of elements, the approximation of the test may be poor.

```r
chisq.test(table_freq_Sx_cA_non_s, simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  table_freq_Sx_cA_non_s
## X-squared = 10.088, df = NA, p-value = 0.02449
```

A Pearson's Chi-squared test with correction revealed a significant association between variables X^2 (NA, N = 594) = 10.088, p = 0.02199. The low p-value (< .05) indicates a rejection of the null hypothesis, suggesting a substantial dependency between the categorical variables (Sex, cAge) with S=0.

2) check if **Sx and cA are independent for the S = 1**. Sx and cA are categorical variables, so to test their independence firstly have to create *contingency table*.

```r
train <- na.omit(train)
survived_passangers_train <- train[train$Survived == 1, ]

# create contingency table of sex and age variables where S = 0
table_freq_Sx_cA_s <- table(survived_passangers_train$Sex, survived_passangers_train$cA)
table_freq_Sx_cA_s
```

```
##
##          [0,20] (20,40] (40,60] (60,80]
##   female     37      71      23       0
##   male       16      24      13       1
```

Secondary, X^2 test was done. However, since the categories doesn't have a lot of elements, the approximation of the test may be poor.

```r
chisq.test(table_freq_Sx_cA_s, simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  table_freq_Sx_cA_s
## X-squared = 3.9945, df = NA, p-value = 0.2569
```

A Pearson's Chi-squared test with correction revealed a significant association between variables X^2 (NA, N = 594) = 3.99, p = 0.2584 The high p-value (> .05) indicates that we cannot reject the null hypothesis, suggesting a independence between the categorical variables (AgeC, Sex) with S=1.

3) check if **Sx and P are independent for the S=0**. Sx and P are categorical variables, so to test their independence firstly have to create *contingency table.*

```
# create contingency table of sex and age variables where S = 0
table_freq_Sx_P_non_s <- table(non_survived_passangers_train$Sex, non_survived_passangers_train$P)
table_freq_Sx_P_non_s
```

```
##
##            1   2   3
##   female   1   4  37
##   male    35  62 149
```

Secondary, X^2 test was done. However, since the categories doesn't have a lot of elements, the approximation of the test may be poor.

```
chisq.test(table_freq_Sx_P_non_s, simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  table_freq_Sx_P_non_s
## X-squared = 12.085, df = NA, p-value = 0.001999
```

A Pearson's Chi-squared test with correction revealed a significant association between variables X^2 (NA, N = 594) = 12.085, p = 0.004998. The low p-value (< .05) indicates a rejection of the null hypothesis, suggesting a substantial dependency between the categorical variables (Sex, P).

4) check if **Sx and P are independent for the S = 1**. Sx and P are categorical variables, so to test their independence firstly have to create *contingency table.*

```
# create contingency table of sex and age variables where S = 0
table_freq_Sx_P_s <- table(survived_passangers_train$Sex, survived_passangers_train$P)
table_freq_Sx_P_s
```

```
##
##           1  2  3
##   female 56 45 30
##   male   26  8 20
```

Secondary, X^2 test was done. However, since the categories doesn't have a lot of elements, the approximation of the test may be poor.

```
chisq.test(table_freq_Sx_P_s, simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  table_freq_Sx_P_s
## X-squared = 8.173, df = NA, p-value = 0.01749
```

A Pearson's Chi-squared test with correction revealed a significant association between variables X^2 (NA, N = 594) = 8.173, p = 0.01049 The hi p-value (< .05) indicates a rejection of the null hypothesis, suggesting dependency between the categorical variables (Sex, P) when S=1.

5) check if **P and cA are independent for the S=0**. Sx and P are categorical variables, so to test their independence firstly have to create *contingency table*.

```
# create contingency table of sex and age variables where S = 0
table_freq_P_cA_non_s <- table(non_survived_passangers_train$Pclass, non_survived_passangers_train$cA)
table_freq_P_cA_non_s
```

```
##
##      [0,20] (20,40] (40,60] (60,80]
##   1      1      14      15       6
##   2      7      45      12       2
##   3     57     101      26       2
```

Secondary, X^2 test was done. However, since the categories doesn't have a lot of elements, the approximation of the test may be poor.

```
chisq.test(table_freq_P_cA_non_s, simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  table_freq_P_cA_non_s
## X-squared = 53.287, df = NA, p-value = 0.0004998
```

A Pearson's Chi-squared test with correction revealed a significant association between variables X^2 (NA, N = 594) = 53.28, p = 0.0.00049. The low p-value (< .05) indicates a rejection of the null hypothesis, suggesting a substantial dependency between the categorical variables (Pclass, cAge) with S=0.

6) check if **P and cA are independent for the S = 1**. P and cA are categorical variables, so to test their independence firstly have to create *contingency table*.

```
# create contingency table of sex and age variables where S = 0
table_freq_P_cA_s <- table(survived_passangers_train$Pclass, survived_passangers_train$cA)
table_freq_P_cA_s
```

```
##
##     [0,20] (20,40] (40,60] (60,80]
##   1      8      47      27       0
##   2     16      29       7       1
##   3     29      19       2       0
```

Secondary, X^2 test was done. However, since the categories doesn't have a lot of elements, the approximation of the test may be poor.

```r
chisq.test(table_freq_P_cA_s, simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  table_freq_P_cA_s
## X-squared = 45.055, df = NA, p-value = 0.0004998
```

A Pearson's Chi-squared test with correction revealed a significant association between variables X^2 (NA, N = 594) = 45.055, p = 0.0004998 The high p-value (< .05) indicates that we can reject the null hypothesis, suggesting a dependence between the categorical variables (Pclass, Sex) with S=1.

Based on the tests we can say that the variables are not independent.