SDA Group Submission Assignment Assign1

Group Gr18

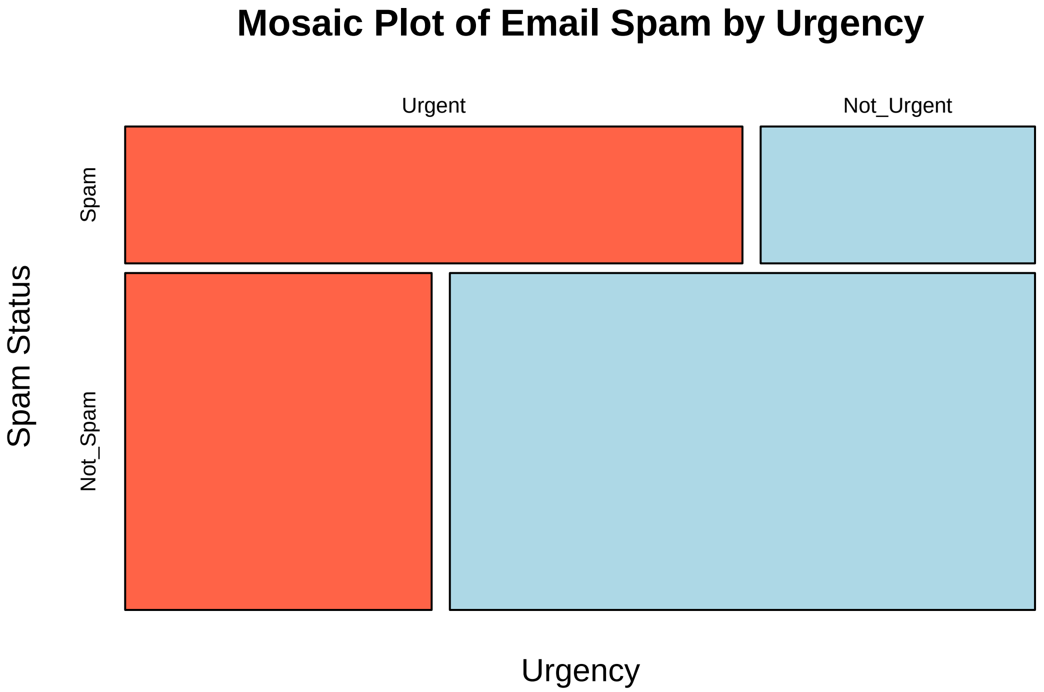
MengliFeng (2720589) and PepijnVanOostveen (2801582)

Exercise 1

a.

```
# Step 1: Create the contingency matrix
email_matrix <- matrix(c(9, 11, 4, 21), nrow = 2, byrow = FALSE,
                       dimnames = list(Spam_Status = c("Spam", "Not_Spam"),
                                       Urgency = c("Urgent", "Not_Urgent")))

# Step 2: Mosaic plot with custom colors
mosaicplot(email_matrix, color = c("tomato", "lightblue"),
           main = "Mosaic Plot of Email Spam by Urgency",
           xlab = "Urgency", ylab = "Spam Status",
           sort = 1:2, dir = c("h", "v"))
```



b.

P(Spam|Urgent) = 9/(9+4) = 0.692 this is computed by diving the number of spam labeled urgent and the total emails that are labeled as urgent.

c.

```
# Extract values
a <- email_matrix["Spam", "Urgent"]
b <- email_matrix["Not_Spam", "Urgent"]
c <- email_matrix["Spam", "Not_Urgent"]
d <- email_matrix["Not_Spam", "Not_Urgent"]

# Compute conditional probabilities
odds_ratio <- (a / b) / (c / d)
odds_ratio
```

```
## [1] 4.295455
```

d.

```
# Perform one-sided Fisher's exact test
fisher_result <- fisher.test(email_matrix, alternative = "greater")
fisher_result
```

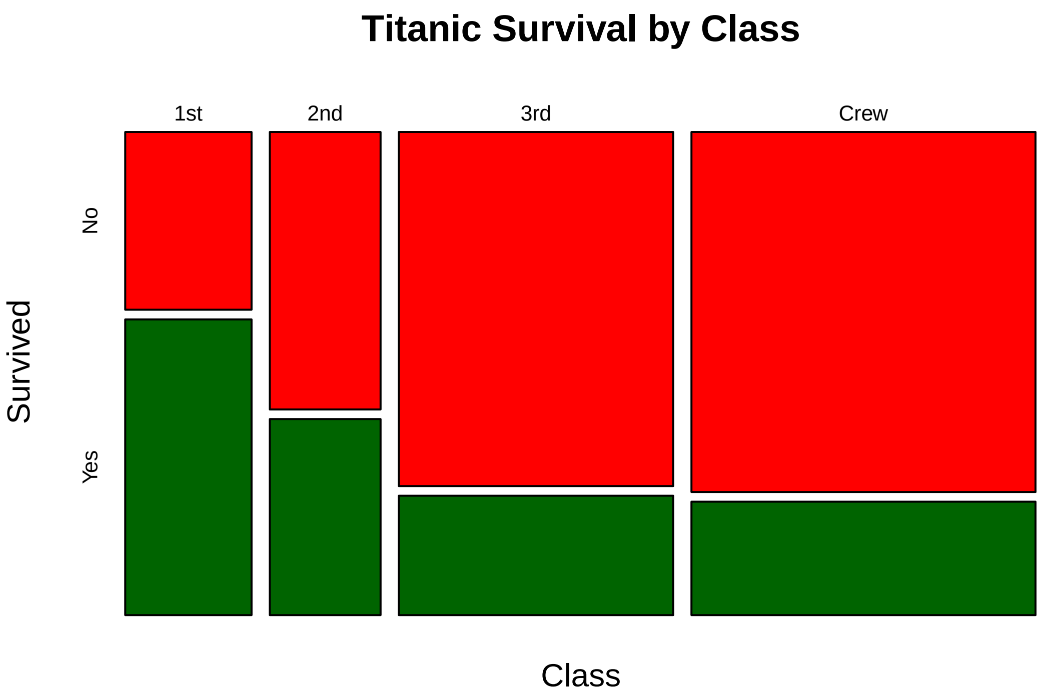
```
##
##  Fisher's Exact Test for Count Data
##
## data:  email_matrix
## p-value = 0.03566
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  1.107161      Inf
## sample estimates:
## odds ratio
##   4.148314
```

since we want to know if urgent means more likely spam (not just related), we use one-sided test, with alternative odds is greater than 1 (urgent emails have higher odds being spam). the null is rejected

Exercise 2

a.

```
# Create the table and then use it to make a mosaicplot
titanic_CS <- margin.table(Titanic, margin = c(1, 4))
mosaicplot(titanic_CS, main = "Titanic Survival by Class", color = c("red", "darkgreen"))
```



There is an obvious correlation between the class and chance of survival. So the wealthier you were (or more accurately the more you spent on your ticket) the likelier you were to survive with the crew being the least likely to survive.

b.

```
# Do the test
chisq.test(titanic_CS)
```

```
##
##  Pearson's Chi-squared test
##
## data:  titanic_CS
## X-squared = 190.4, df = 3, p-value < 2.2e-16
```

The p-value is less than 10^-15 it is far smaller than the significance level of 0.05 and thus the null is rejected as we expected. This means that the test had the same result as we had by just looking at the mosaic namely that there is a relationship between the class and the chance to survive.

c.

```
# Get the expected values
test_result <- chisq.test(titanic_CS)
test_result$expected
```

```
##       Survived
## Class        No       Yes
##   1st  220.0136 104.98637
##   2nd  192.9350  92.06497
##   3rd  477.9373 228.06270
##   Crew 599.1140 285.88596
```

```
# The real values
titanic_CS
```

```
##       Survived
## Class   No Yes
##   1st  122 203
##   2nd  167 118
##   3rd  528 178
##   Crew 673 212
```

The values that are the farthest apart from the expected values is the survival chance of 1st class, but I expected that so the thing that I found most striking is that there is almost no difference between 3rd class and the crew. I thought that knowing more about the ship would help the crew and that 3rd class had the lowest chances since 1st and 2nd class were situated better.

d.

```
# get the 3rd and crew rows from our table
third_crew <- titanic_CS[c("3rd", "Crew"), ]

# make it a matrix and then test as alternative greater since the first row is 3rd class
↪  and we want to know if first class has a better survival chance
survival_table <- as.matrix(third_crew)
fisher.test(survival_table, alternative = "greater")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  survival_table
## p-value = 0.7385
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.7657002       Inf
## sample estimates:
## odds ratio
##  0.9344464
```

The p-value is 0.7385 so the survival chance isn't significantly higher for 3rd class and thus the hypothesis that the chacnes are equal isn't rejected.