CMDA-3654

Homework 8

Michael La Vance

Due as a .pdf upload

Problem 1: [35] Tests of association

Load the `CoalMiners` data from the `vcd` library in R.

a. Convert the 3-way table into a data frame with 36 rows and 4 columns.

```
coal <- CoalMiners

coal <- as.data.frame(ftable(coal))
```

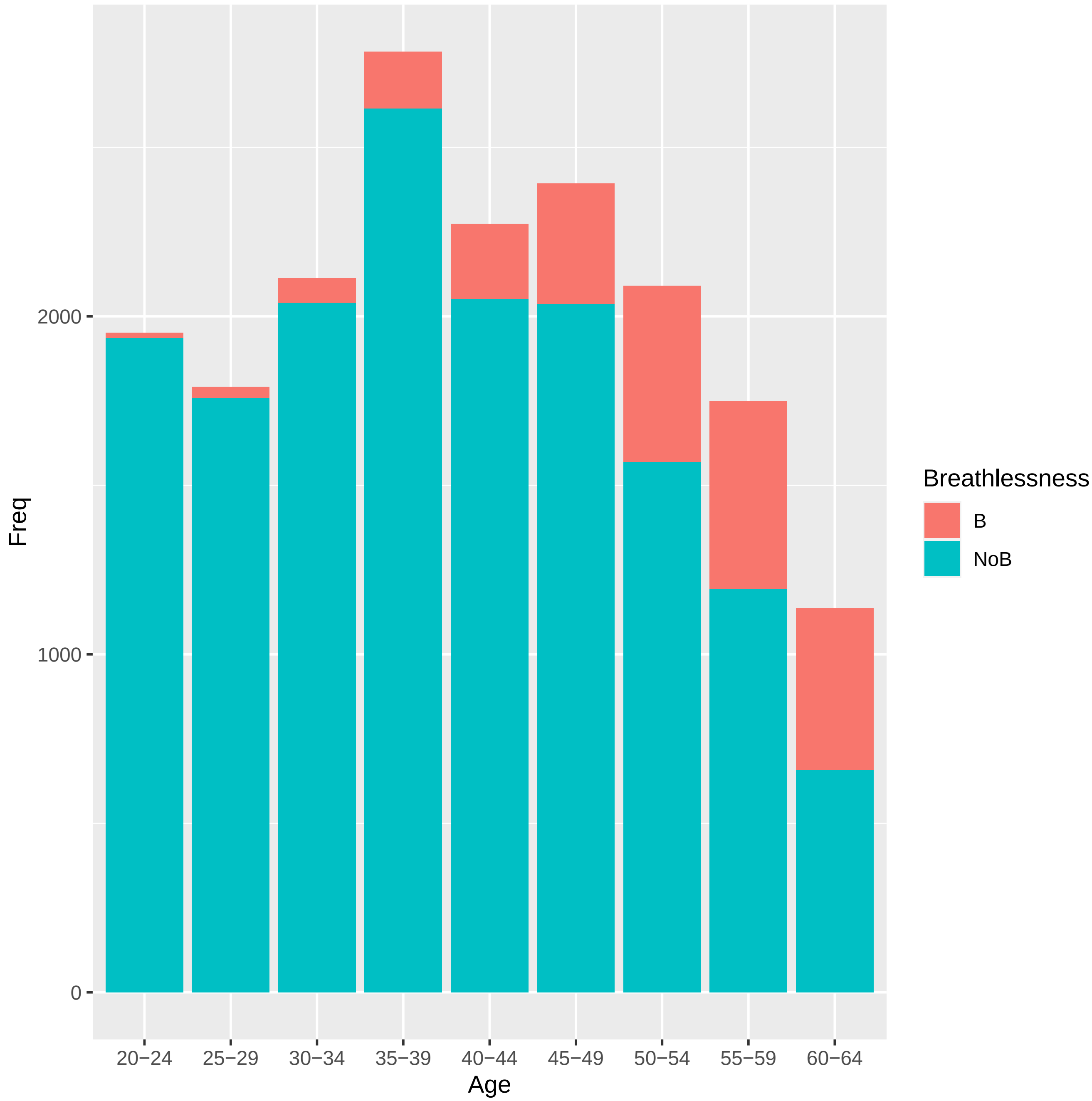
b. Flatten the 3-way table so that we can see everything in a single large table.

```
ftable(CoalMiners)
```

```
                             Age 20-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64
Breathlessness Wheeze
B              W                     9    23    54   121   169   269   404   406   372
               NoW                   7     9    19    48    54    88   117   152   106
NoB            W                    95   105   177   257   273   324   245   225   132
               NoW                1841  1654  1863  2357  1778  1712  1324   967   526
```

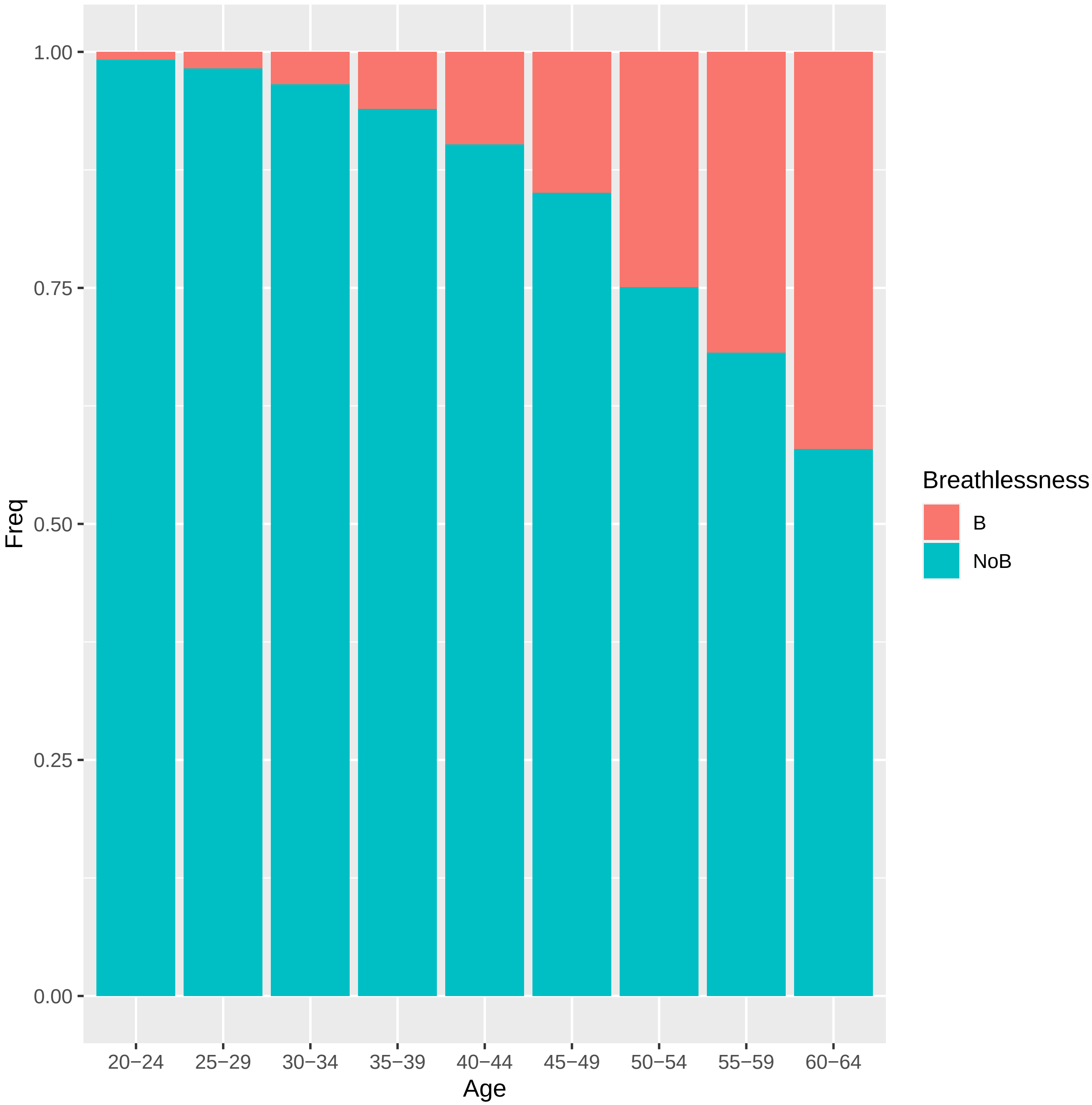
c. Construct a stacked barplot with `Age` group on the x-axis and `Breathlessness` on the y-axis with the different outcomes of `Breathlessness` having different colors.

```
ggplot(data = coal, aes(fill = Breathlessness, y = Freq, x = Age)) + geom_bar(position = "stack", stat = "iden
```



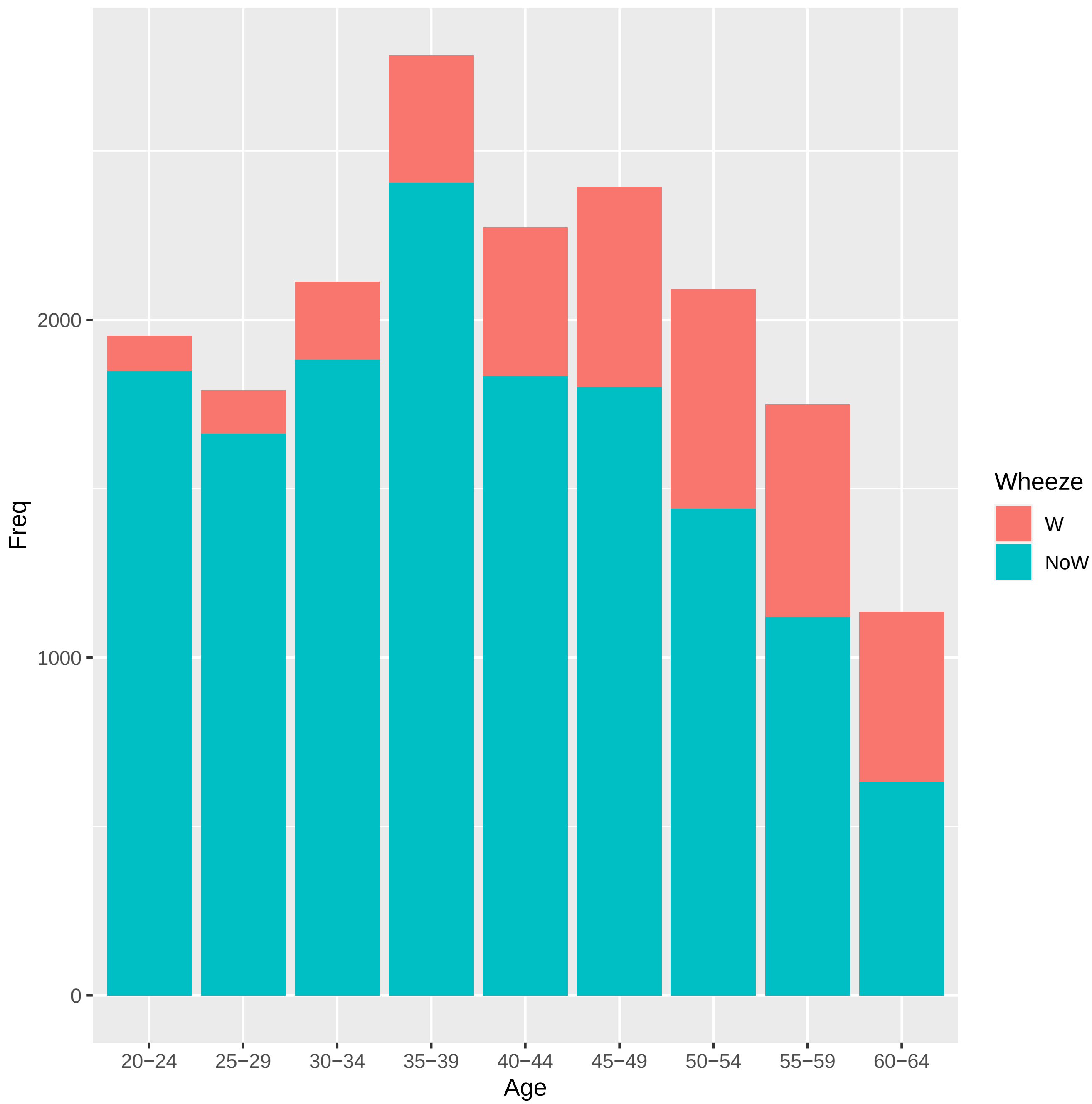
- The above plot is clearly an absolute frequency barplot. Remake the plot, this time using the relative frequencies (there are many ways to do this, do whatever seems easiest).

```
ggplot(data = coal, aes(fill = Breathlessness, y = Freq, x = Age)) + geom_bar(position = "fill", stat = "ident
```

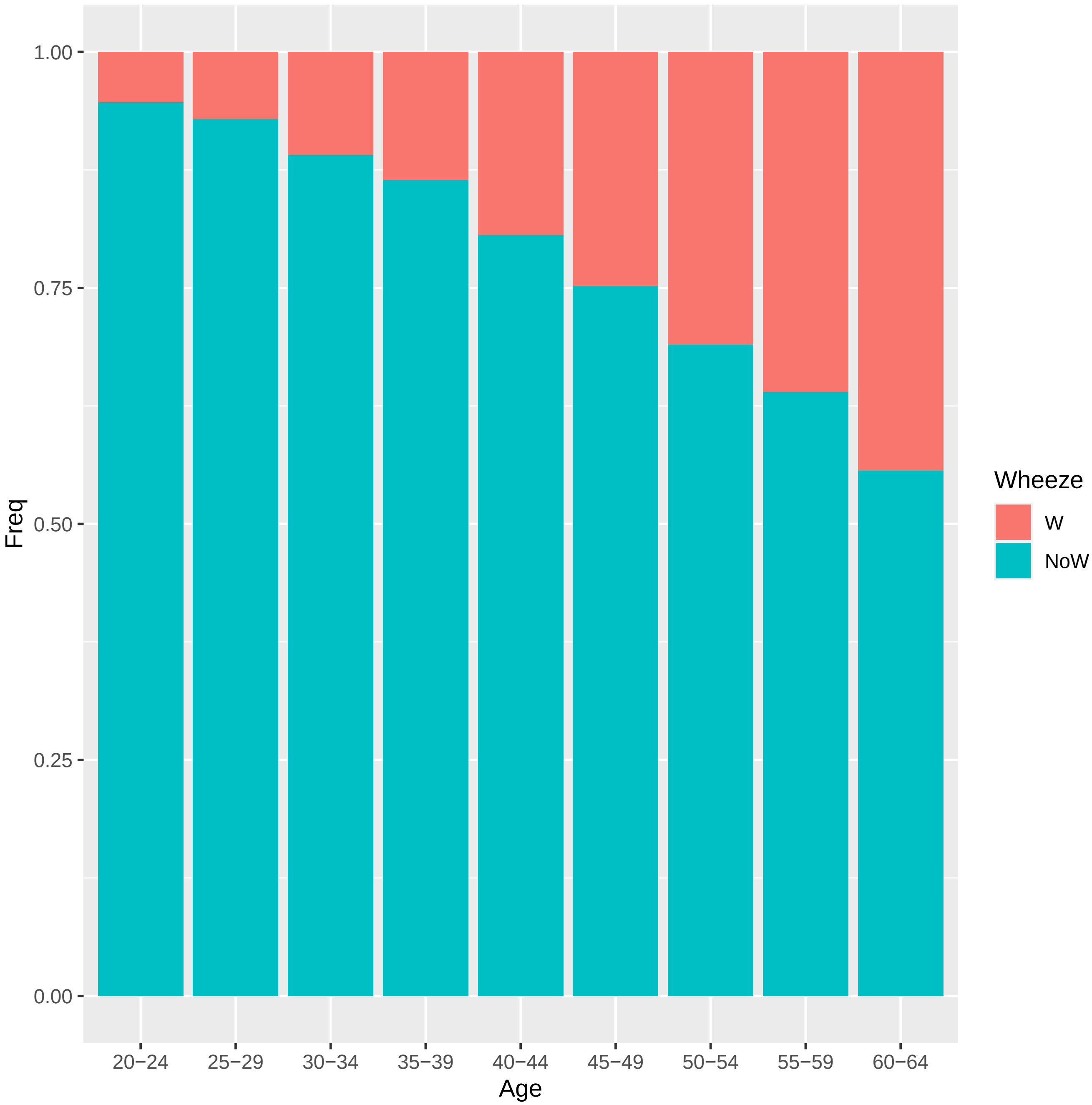


d. Repeat the above steps but this time with `Age` group on the x-axis and `Wheezing` on the y-axis with the different outcomes of `Wheezing` having different colors.

```
ggplot(data = coal, aes(fill = Wheeze, y = Freq, x = Age)
       ) + geom_bar(position = "stack", stat = "identity")
```




```
ggplot(data = coal, aes(fill = Wheeze, y = Freq, x = Age)
       ) + geom_bar(position = "fill", stat = "identity")
```



e. Add a new column with the feature named "`Career`" to your data frame where you will recode the ages into the following three groups: "`Early`" = 20-34, "`Middle`" = 35-49, and "`Late`" = 50 - 64. These groups will reflect where people tend to be if they started their career at the age of 20 and stayed employed, i.e. Early Career, Middle Career, Late Career.

- Construct a 3-way table for `Wheezing` Symptoms and `Breathlessness` Symptoms for the three `Career` levels. Each two-way table slice should be `Wheezing` versus `Breathlessness`.

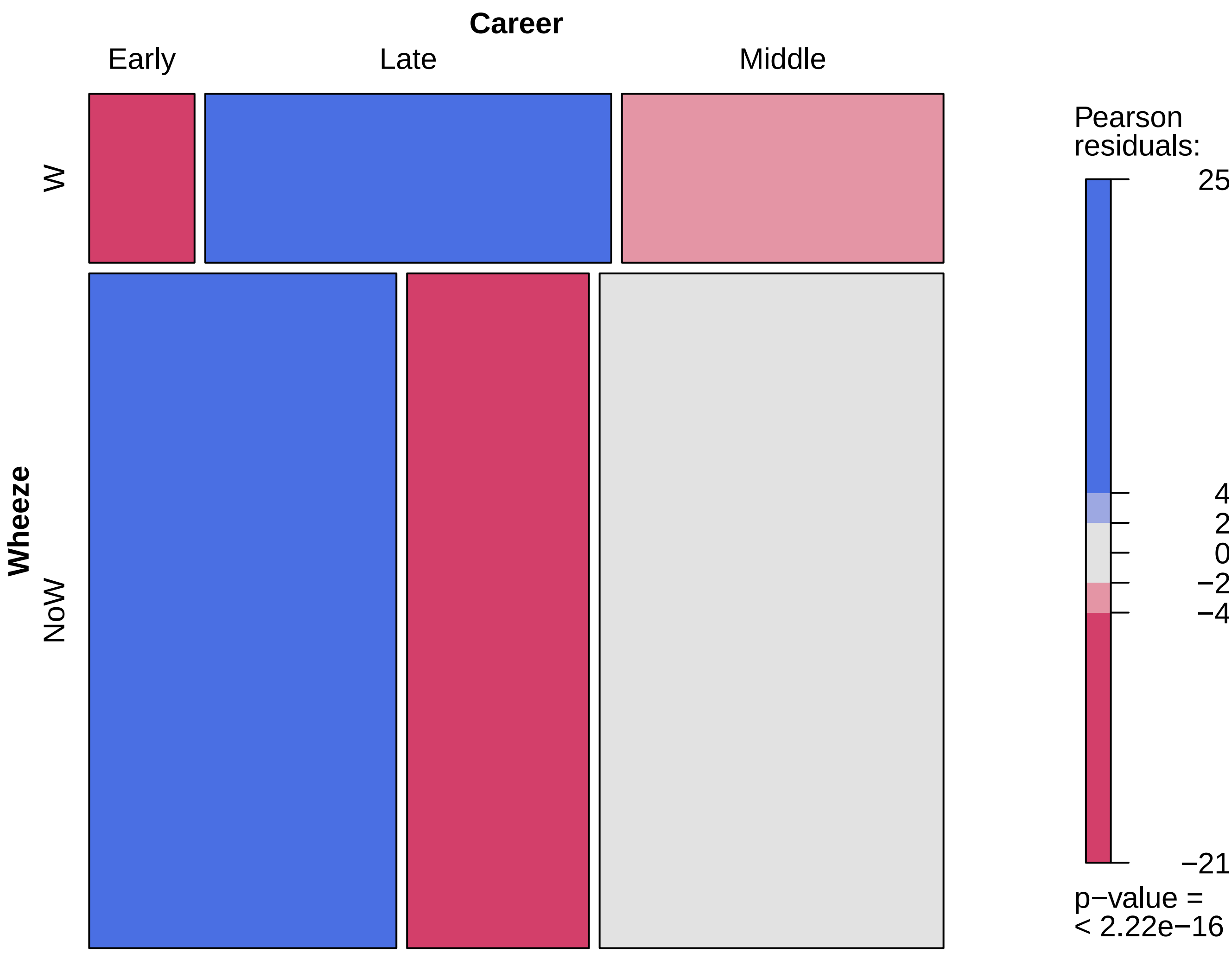
```
Career <- 1:nrow(coal)

temp.lut <- c("20-24" = "Early", "25-29" = "Early", "30-34" = "Early", "35-39" =
               "Middle", "40-44" = "Middle", "45-49" = "Middle", "50-54" =
               "Late", "55-59" = "Late", "60-64" = "Late")
Career <- temp.lut[coal$Age]
coal <- data.frame(coal,Career)
rm(Career)
rm(temp.lut)

cl.tbl <- xtabs(Freq ~Wheeze + Breathlessness + Career, data = coal)
```

```
cl.tbl
```

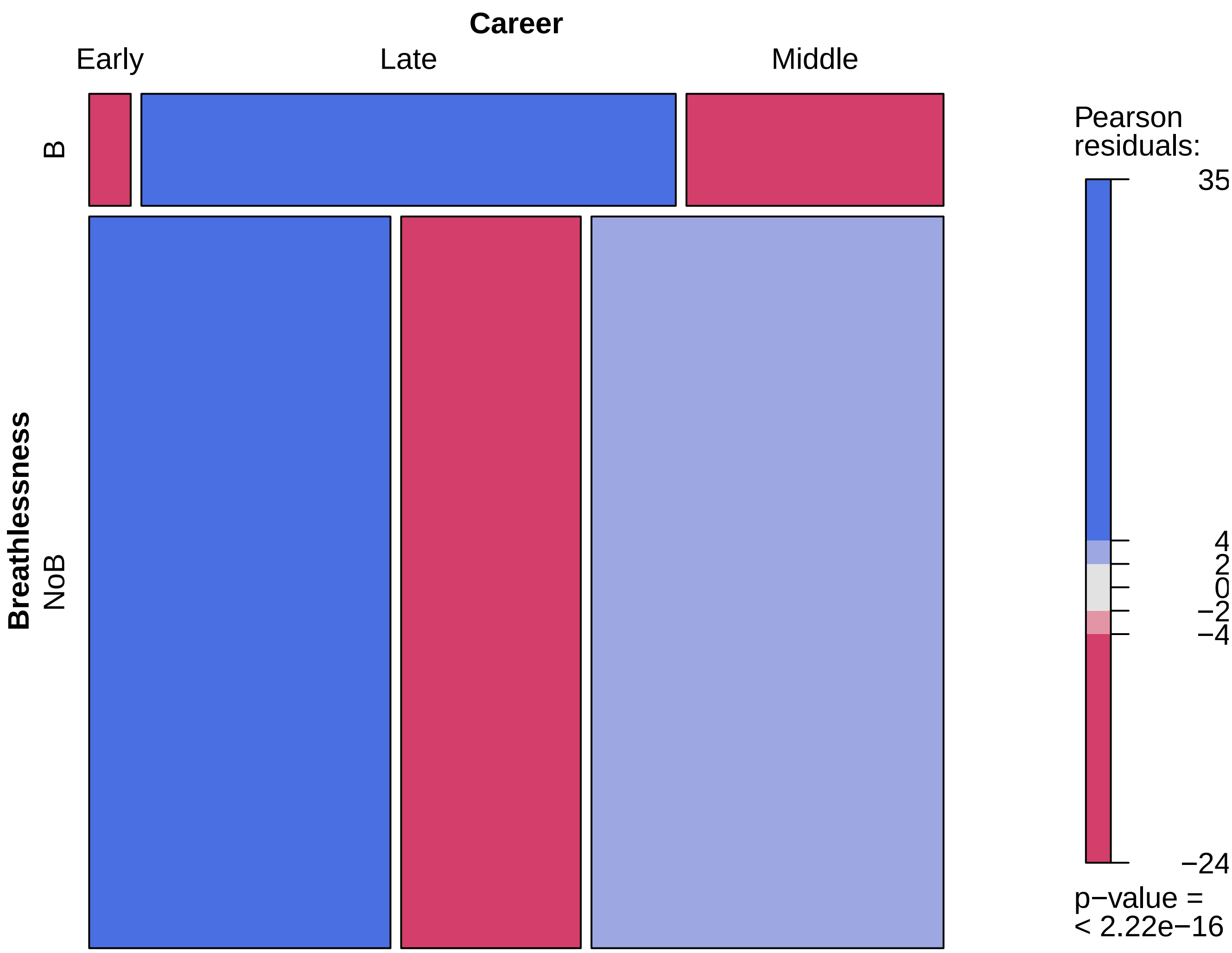
```
, , Career = Early

      Breathlessness
Wheeze    B  NoB
   W     86  377
   NoW   35 5358

, , Career = Late

      Breathlessness
Wheeze    B  NoB
   W   1182  602
   NoW  375 2817

, , Career = Middle

      Breathlessness
Wheeze    B  NoB
   W    559  854
   NoW  190 5847
```

f. Make a mosaic plot (use `shade = T` and the `mosaic()` function from the `vcd` library) for each of the following pair of features:
   i. `Wheeze` versus `Career`
   ii. `Breathlessness` versus `Career`
   iii. `Wheeze` versus `Breathlessness`
   iv. Comment on the results.

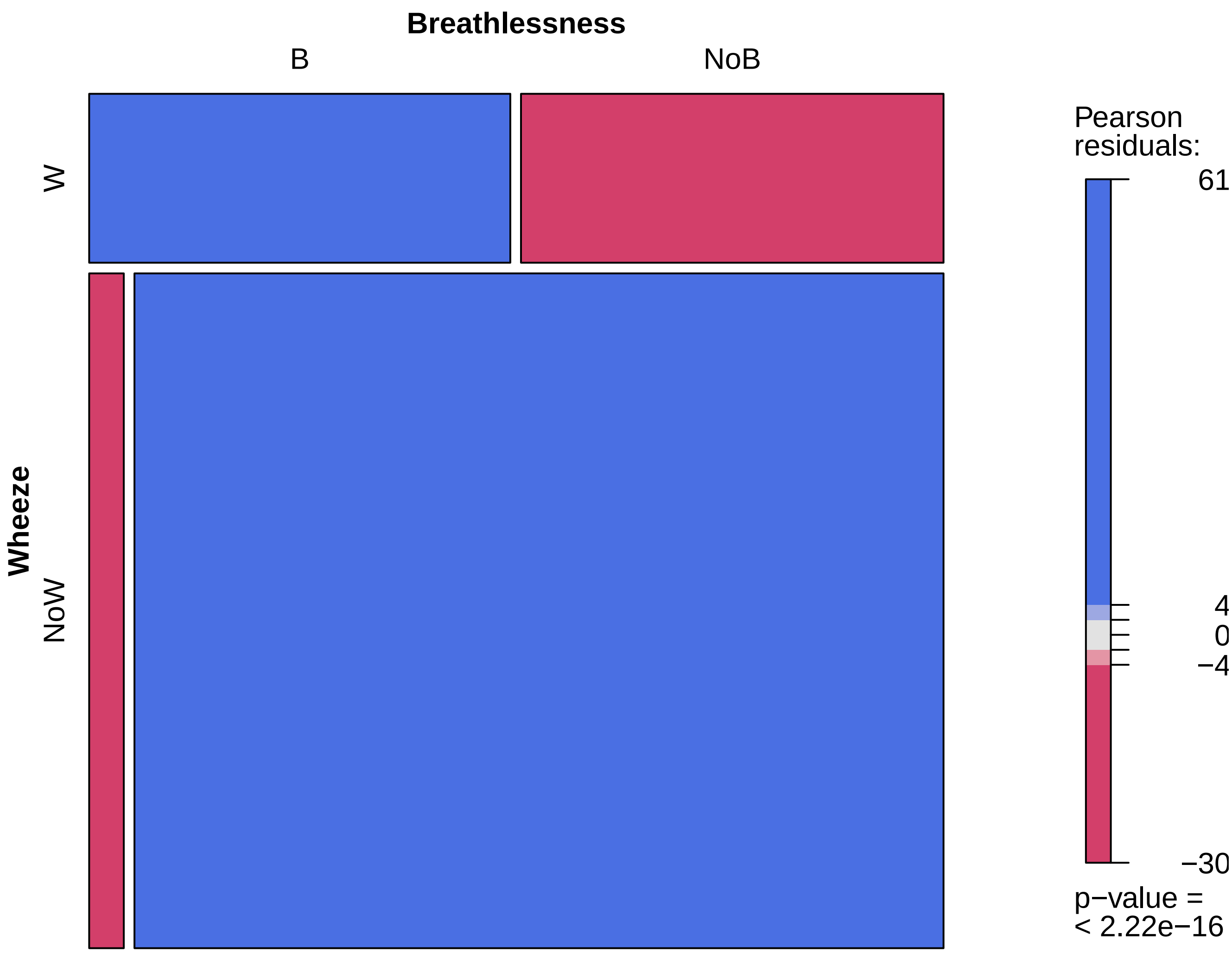
```
mosaic( ~ Wheeze + Career , data = coal, shade = T)
```




```
mosaic( Freq ~ Breathlessness + Career , data = coal, shade = T)
```




```
mosaic( Freq ~ Wheeze + Breathlessness , data = coal, shade = T)
```



g. Consider the 3-way table you constructed in part (e). There are three features: `Breathlessness`, `Wheezing`, and `Career`. **For each pair of features**, carry out a chi-square test of independence and report whether there is association between features.

```
#Creating groups
nocareer <- margin.table(cl.tbl, c(1,2))

nobreath <- margin.table(cl.tbl, c(1,3))

nowheeze <- margin.table(cl.tbl, c(2,3))

chisq.test(nocareer)

    Pearson's Chi-squared test with Yates' continuity correction

data:  nocareer
X-squared = 5332.9, df = 1, p-value < 2.2e-16

chisq.test(nobreath)
```

```
    Pearson's Chi-squared test

data:  nobreath
X-squared = 1320.8, df = 2, p-value < 2.2e-16
```

```
chisq.test(nowheeze)
```

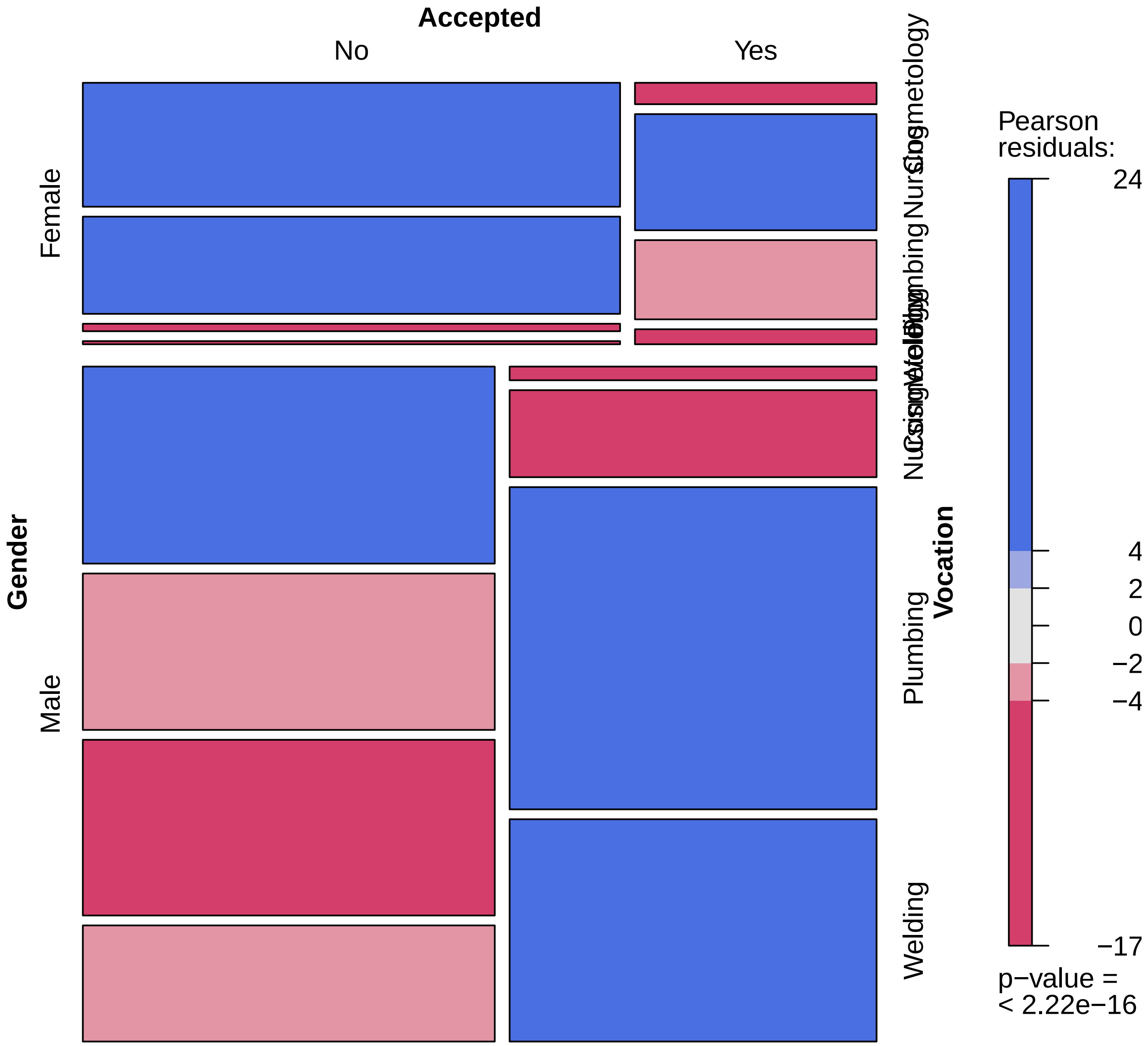
```
    Pearson's Chi-squared test

data:  nowheeze
X-squared = 2108.9, df = 2, p-value < 2.2e-16
```

Problem 2 [35 pts] Tests of association.

A random sample of 5,000 high school students who have applied for vocational training has been collected which contains their `Gender` and `Acceptance` into the program. The data is contained in `acceptance.csv`.

a. After reading in the data, summarize the data into a 3D array of the counts (name this `byVoc` table) where the 3rd dimension corresponds to the `Vocation.` Display this output in the 3D format. Additionally display the data using a flat contingency table.

```
accept <- read.csv("acceptance.csv")
byVoc <-xtabs(~ Gender + Accepted + Vocation, data = accept)
mosaic(byVoc, shade = T)
```

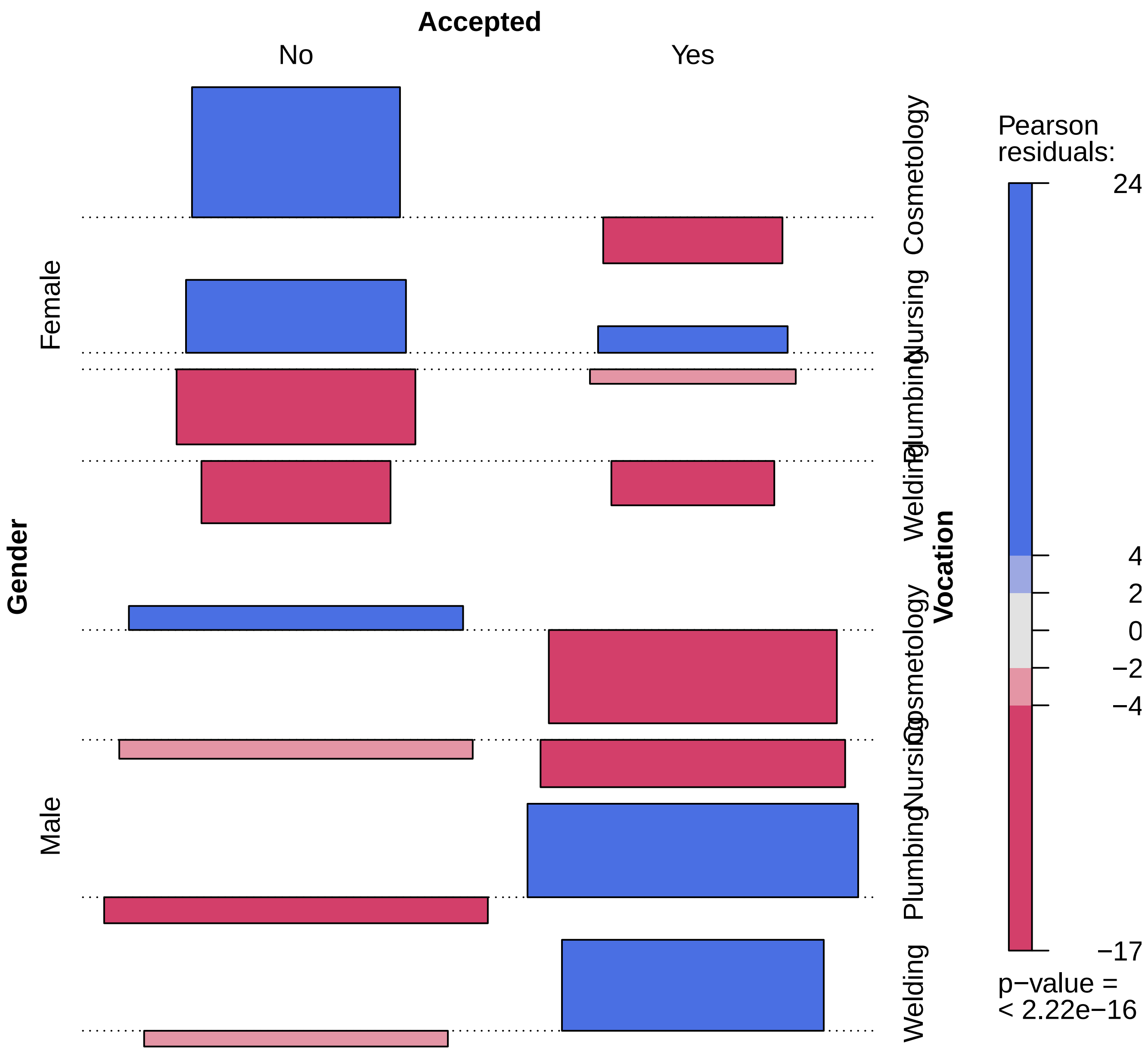



```
ftable(byVoc)
```

```
                Vocation Cosmetology Nursing Plumbing Welding
Gender Accepted
Female No                        515     404       31      13
       Yes                        40     217      148      28
Male   No                        582     462      519     343
       Yes                        36     229      848     585
```

b. Construct an association plot using `assoc()` from the `vcd` library, use `shade = T` for the three features: `Accepted`, `Vocation`, and `Gender`. Comment on any patterns that you see.

```
assoc(byVoc, shade = T)
```



Males are being accepted into plumbing and welding more than expected, and Women are being denied to cosmetology and nursing more than expected.

c. For each Vocation, carry out a chi-square test of independence and report whether there is association between Gender and Acceptance.

```
chisq.test(byVoc[ , ,"Cosmetology"])

    Pearson's Chi-squared test with Yates' continuity correction

data:  byVoc[, , "Cosmetology"]
X-squared = 0.70766, df = 1, p-value = 0.4002
```

```
chisq.test(byVoc[ , ,"Plumbing"])

    Pearson's Chi-squared test with Yates' continuity correction

data:  byVoc[, , "Plumbing"]
X-squared = 28.548, df = 1, p-value = 9.142e-08
```

```r
chisq.test(byVoc[ , ,"Welding"])
```

```
	Pearson's Chi-squared test with Yates' continuity correction

data:  byVoc[, , "Welding"]
X-squared = 0.26768, df = 1, p-value = 0.6049
```

```r
chisq.test(byVoc[ , ,"Nursing"])
```

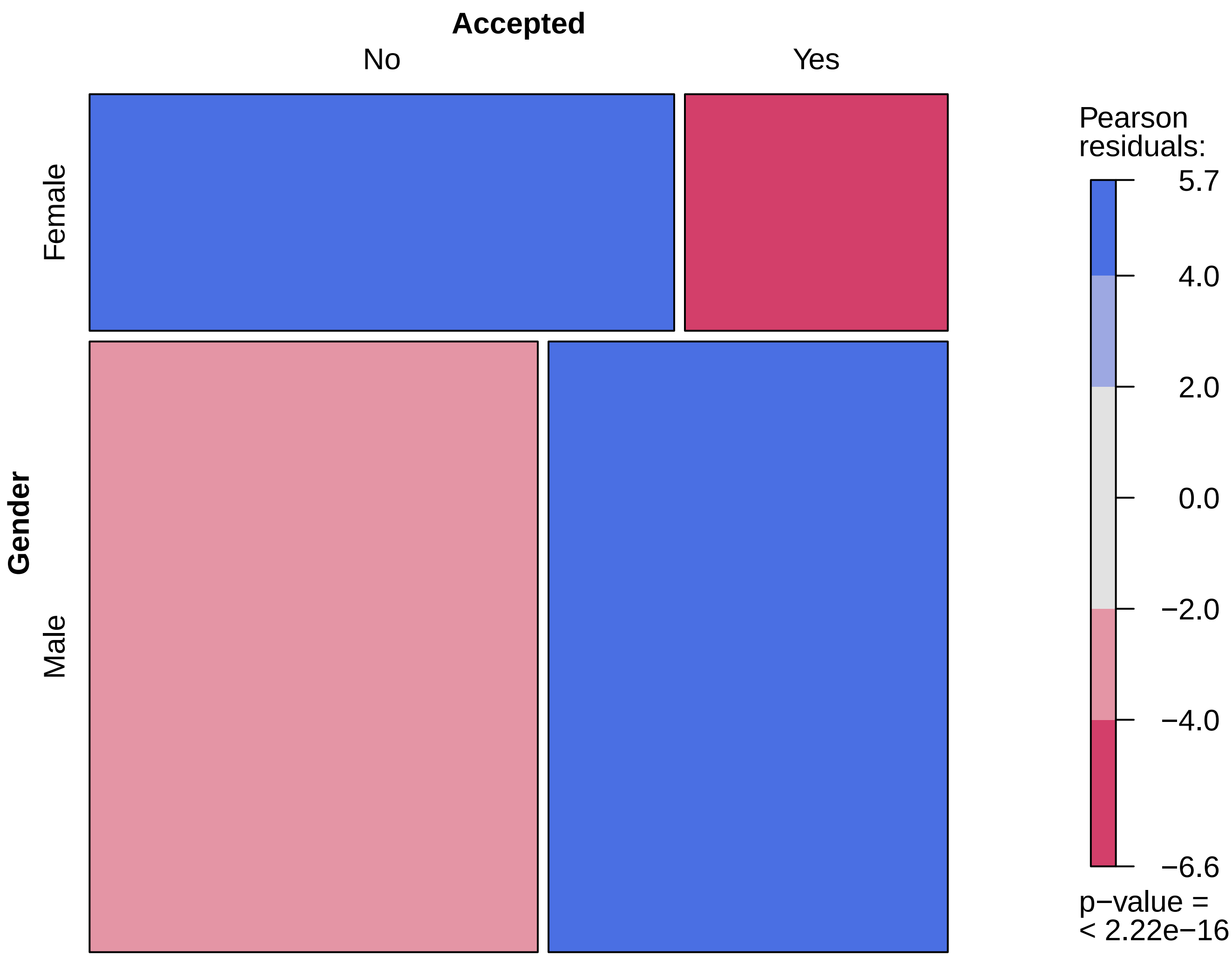
```
	Pearson's Chi-squared test with Yates' continuity correction

data:  byVoc[, , "Nursing"]
X-squared = 0.39703, df = 1, p-value = 0.5286
```

d. Ignoring Vocation, carry out a single chi-square test of independence for the whole data and report whether there is association between `Gender` and `Acceptance`. Additionally provide a mosaic plot with `shade = T`.

```r
gen.accept <- margin.table(byVoc, c(1,2))

chisq.test(gen.accept)
```

```
	Pearson's Chi-squared test with Yates' continuity correction

data:  gen.accept
X-squared = 105.96, df = 1, p-value < 2.2e-16
```

```r
mosaic(gen.accept, shade = T)
```



e. Carry out a **CMH chi-square test** and report whether there is association between Gender and Acceptance taking into account the different vocations.

```
mantelhaen.test(byVoc)

    Mantel-Haenszel chi-squared test with continuity correction

data:  byVoc
Mantel-Haenszel X-squared = 14.289, df = 1, p-value = 0.0001568
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.6003882 0.8474693
sample estimates:
common odds ratio
        0.7133096
```

f. Is there any conflict between the results obtained in parts (b-e), and c? What is your final conclusion regarding association between `Gender` and `Acceptance`? The CMH chi-squared tests states that there is reasonable suspicion that the acceptance rates by gender are not equal across the 4 vocations looked at. Section C suggested that only plumbing had different acceptance rates by gender.

g. Construct a summary matrix with success rates for male and female applicants in each `Vocation`. Also calculate the overall success rate (i.e., ignoring department) of male and female candidates. From these numbers (without referring to statistical tests) what is your empirical conclusion—do you think there is gender bias in admissions? Why or why not?

```
apply(byVoc, 3, odds.ratio)
```

```
Cosmetology     Nursing    Plumbing     Welding 
  0.7963918   0.9228160   0.3422382   0.7918576 
```

```
margin.table(byVoc, c(1,2)) %>% prop.table
```

```
        Accepted
Gender       No    Yes
  Female 0.1926 0.0866
  Male   0.3812 0.3396
```

I do think there is a gender bias in admissions, but not to the extent that this data would suggest. I think this is due to the particular vocations picked, as I have only seen male plumbers and welders, only female cosmetologists, and an overwhelming majority of female nurses.

Problem 3 [30 pts] Market Basket Analysis.

Load the `Groceries` transactions database from the `arules` package in R (you will need to do `data("Groceries", package = "arules")` this time around). Answer the following questions:

a. How many transactions and items are there in this database? What is the most frequent item and how many times was it bought?

```
data("Groceries", package = "arules")

summary(Groceries)
```

```
      Length        Class         Mode 
        9835 transactions           S4 
```

The Most Frequent item purchased is Milk, it was bought 2513 times over a one month timespan.

b. What percentage of transactions involved 20 or more items? On average, how many items were involved per transaction?

There were on average 4.4 items purchased per transaction, .39% of purchases had over 20 items.

c. Find all rules with support > 1% and confidence > 50%. How many such rules are there? Which of these rules has the highest confidence and highest support? Report the support, confidence, and lift of this rule. What are the interpretations of these numbers?

```
rules <- apriori(Groceries, parameter = list(supp=0.01, conf=0.50))
```

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.5    0.1    1 none FALSE            TRUE       5    0.01      1
 maxlen target  ext
     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 98 

set item appearances ...[0 item(s)] done [0.00s].
```

```
set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
sorting and recoding items ... [88 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [15 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

```
length(rules)
```

```
[1] 15
```

```
inspect(rules)
```

```
     lhs                                   rhs                support
[1]  {curd,yogurt}                      => {whole milk}       0.01006609
[2]  {other vegetables,butter}          => {whole milk}       0.01148958
[3]  {other vegetables,domestic eggs}   => {whole milk}       0.01230300
[4]  {yogurt,whipped/sour cream}        => {whole milk}       0.01087951
[5]  {other vegetables,whipped/sour cream} => {whole milk}    0.01464159
[6]  {pip fruit,other vegetables}       => {whole milk}       0.01352313
[7]  {citrus fruit,root vegetables}     => {other vegetables} 0.01037112
[8]  {tropical fruit,root vegetables}   => {other vegetables} 0.01230300
[9]  {tropical fruit,root vegetables}   => {whole milk}       0.01199797
[10] {tropical fruit,yogurt}            => {whole milk}       0.01514997
[11] {root vegetables,yogurt}           => {other vegetables} 0.01291307
[12] {root vegetables,yogurt}           => {whole milk}       0.01453991
[13] {root vegetables,rolls/buns}       => {other vegetables} 0.01220132
[14] {root vegetables,rolls/buns}       => {whole milk}       0.01270971
[15] {other vegetables,yogurt}          => {whole milk}       0.02226741
     confidence coverage   lift     count
[1]  0.5823529  0.01728521 2.279125  99
[2]  0.5736041  0.02003050 2.244885 113
[3]  0.5525114  0.02226741 2.162336 121
[4]  0.5245098  0.02074225 2.052747 107
[5]  0.5070423  0.02887646 1.984385 144
[6]  0.5175097  0.02613116 2.025351 133
[7]  0.5862069  0.01769192 3.029608 102
[8]  0.5845411  0.02104728 3.020999 121
[9]  0.5700483  0.02104728 2.230969 118
[10] 0.5173611  0.02928317 2.024770 149
[11] 0.5000000  0.02582613 2.584078 127
[12] 0.5629921  0.02582613 2.203354 143
[13] 0.5020921  0.02430097 2.594890 120
[14] 0.5230126  0.02430097 2.046888 125
[15] 0.5128806  0.04341637 2.007235 219
```

Rule # 8 has the highest support and confidence levels, support = 0.01230300, confidence = 0.5845411, lift = 3.020999. about 1.23 % of purchases had tropical fruit and root vegetables. About 58.5% of transactions that had tropical fruit and root vegetables also had other vegetables. If you know that tropical fruit and root vegetables were in the transaction, it is 3 times more likely that there were other vegetables purchased.

Problem 4 [10 pts Extra Credit]

Continue working with the data in problem 3.

a. Which items do “whole milk” lead to? Find all rules with support > 1%, confidence > 20%, and “whole milk” on the left hand side. Report these rules.

b. Which items lead to “whole milk”? Find all rules with support > 1%, confidence > 20%, and “whole milk” on the right hand side. Report these rules.