

CMDA-3654

Homework 8

Eduardo Salvador

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.

```
#Loading libraries
library(tidyverse)
library(vcd)

#Creating dataframe
coalminerdf<-as.data.frame(CoalMiners)
coalminerdf
```

	Breathlessness	Wheeze	Age	Freq
1	B	W	20-24	9
2	NoB	W	20-24	95
3	B	NoW	20-24	7
4	NoB	NoW	20-24	1841
5	B	W	25-29	23
6	NoB	W	25-29	105
7	B	NoW	25-29	9
8	NoB	NoW	25-29	1654
9	B	W	30-34	54
10	NoB	W	30-34	177
11	B	NoW	30-34	19
12	NoB	NoW	30-34	1863
13	B	W	35-39	121
14	NoB	W	35-39	257
15	B	NoW	35-39	48
16	NoB	NoW	35-39	2357
17	B	W	40-44	169
18	NoB	W	40-44	273
19	B	NoW	40-44	54
20	NoB	NoW	40-44	1778
21	B	W	45-49	269
22	NoB	W	45-49	324
23	B	NoW	45-49	88
24	NoB	NoW	45-49	1712
25	B	W	50-54	404
26	NoB	W	50-54	245
27	B	NoW	50-54	117
28	NoB	NoW	50-54	1324
29	B	W	55-59	406
30	NoB	W	55-59	225
31	B	NoW	55-59	152
32	NoB	NoW	55-59	967
33	B	W	60-64	372
34	NoB	W	60-64	132
35	B	NoW	60-64	106
36	NoB	NoW	60-64	526

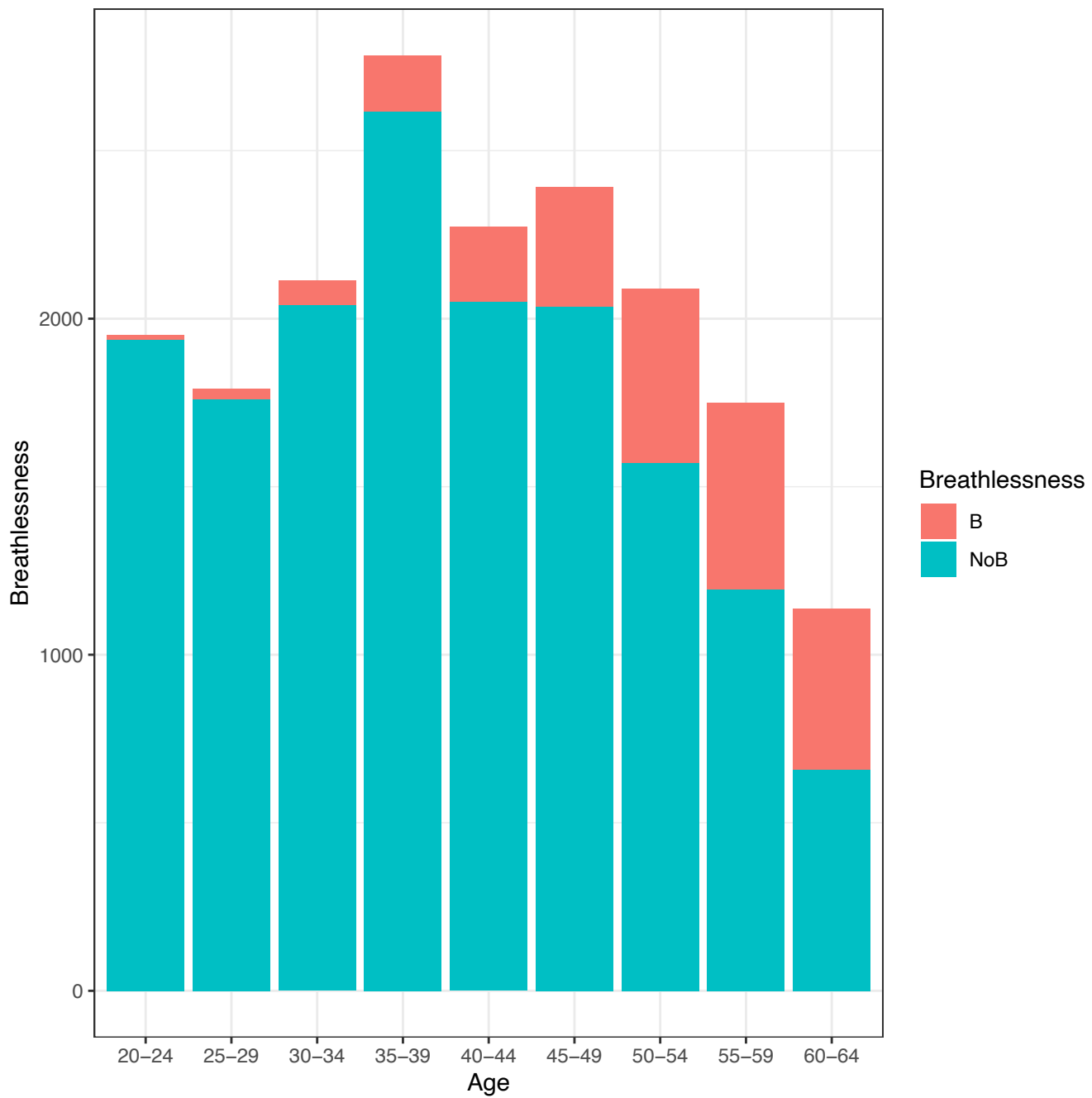
- b. Flatten the 3-way table so that we can see everything in a single large table.

```
#flattening table to see everything in a single large table with ftable
ftable(coalminerdf)
```

			Freq	7	9	19	23	48	54	88	95	105	106	117	121	132	152	169	177	225	245	257	269	273	324
Breathlessness	Wheeze	Age																							
B	W	20-24	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		25-29	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		30-34	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		35-39	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
		40-44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
		45-49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
		50-54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		55-59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		60-64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	NoW	20-24	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		25-29	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		30-34	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		35-39	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		40-44	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		45-49	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		50-54	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
		55-59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
		60-64	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
NoB	W	20-24	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		25-29	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		30-34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
		35-39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
		40-44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
		45-49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
		50-54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
		55-59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
		60-64	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	NoW	20-24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		25-29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		30-34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		35-39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		40-44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		45-49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		50-54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		55-59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		60-64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- 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.

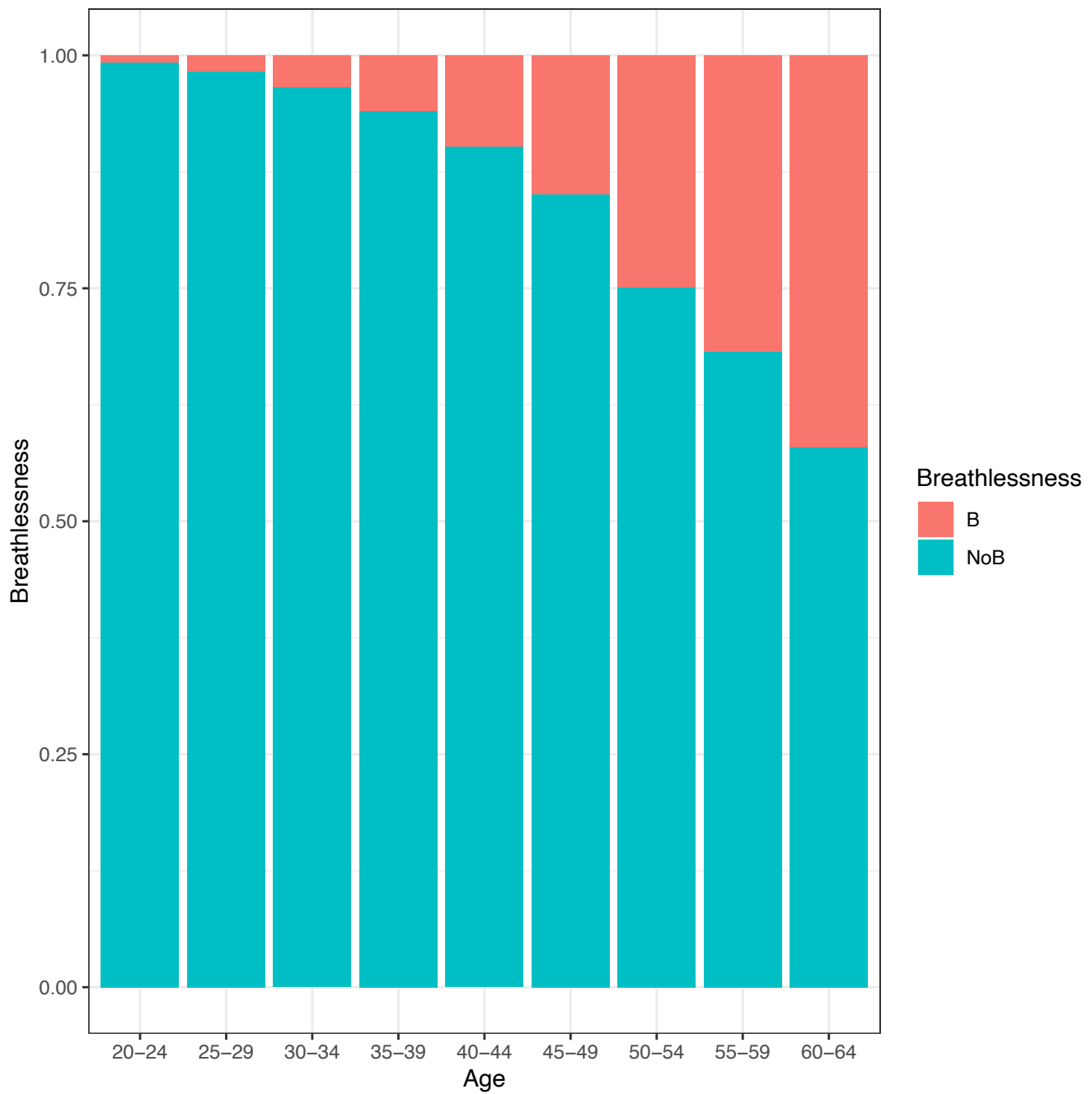
```
#Barplot using ggplot
ggplot(data=coalminerdf, aes(x=Age,y=Freq,fill=Breathlessness))+labs(x="Age",y="Breathlessness",tittle="Breath
```



- 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).

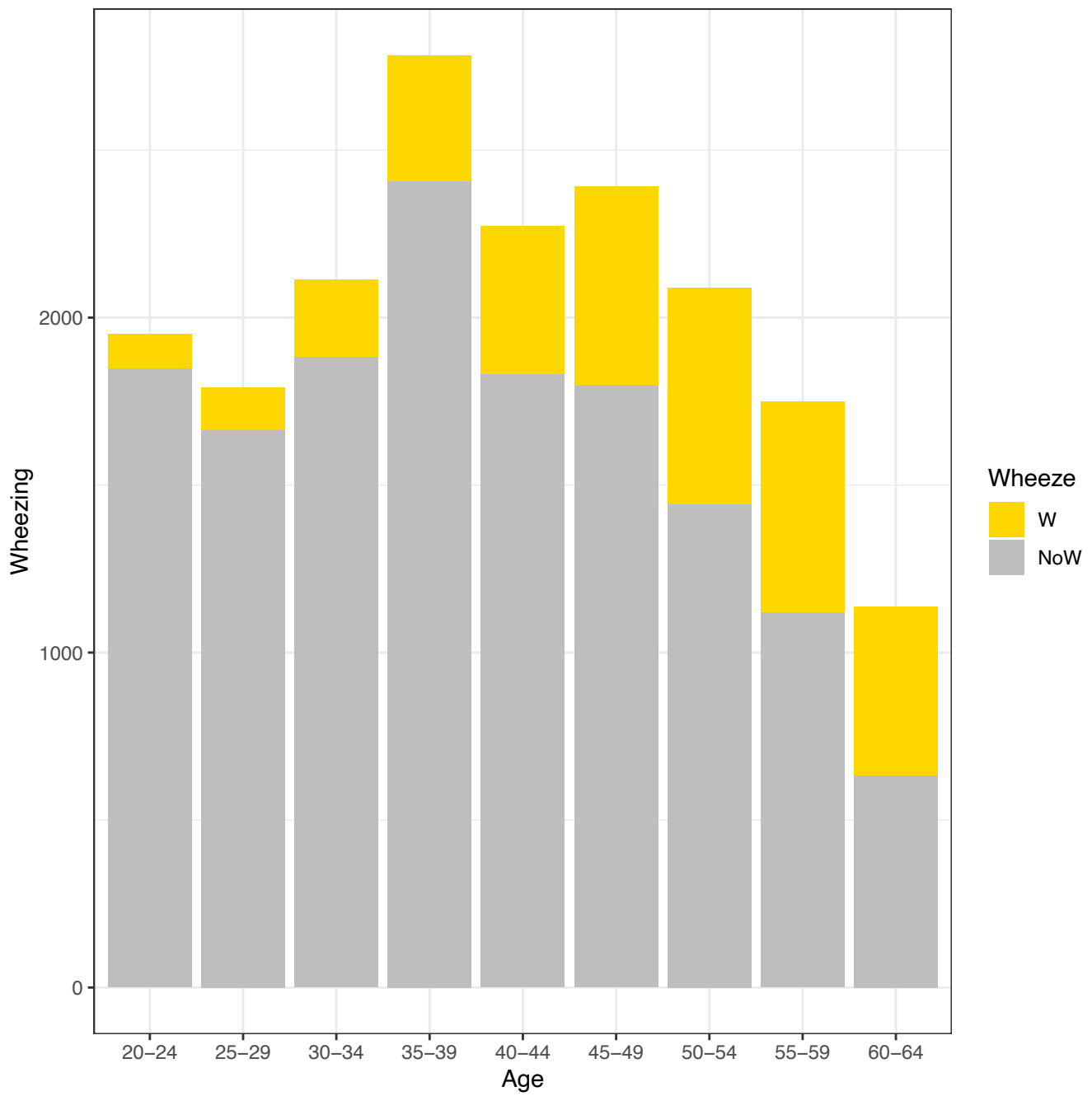
#Creating relative frequency plot using ggplot

```
ggplot(data=coalminerdf, aes(x=Age,y=Freq,fill=Breathlessness))+labs(x="Age",y="Breathlessness",title="Relati
```



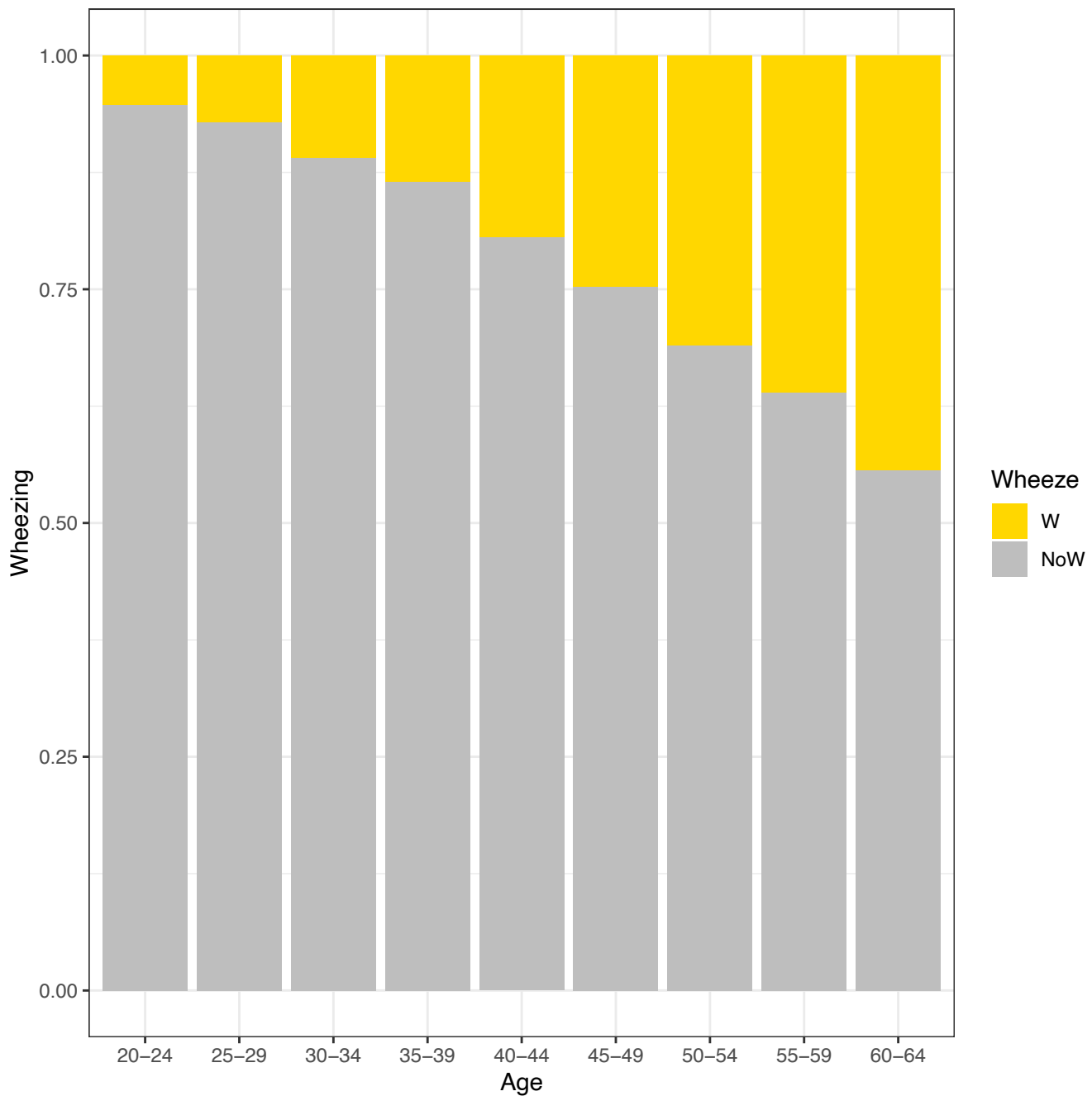
d. Repeat the above steps but this time with **Age** group on the x-axis and **Wheezing** on the y-axis with the different outcomes of **Wheezing** having different colors.

```
#Barplot using ggplot
ggplot(data=coalminerdf, aes(x=Age,y=Freq,fill=Wheeze))+labs(x="Age",y="Wheezing",title="Wheezing by Age")+ t.
```



#Creating relative frequency plot using ggplot

ggplot(data=coalminerdf, aes(x=Age,y=Freq,fill=Wheeze))+labs(x="Age",y="Wheezing",tittle="Relative Frequency W



- 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.

??????

#Creating 3 groups by age

```
coalminerdf$Career[coalminerdf$Age=="20-24"]<-"Early"
```

Contacting Delphi...the oracle is unavailable.

We apologize for any inconvenience.

```
coalminerdf$Career[coalminerdf$Age=="25-29"]<-"Early"
```

```
coalminerdf$Career[coalminerdf$Age=="30-34"]<-"Early"
```

```
coalminerdf$Career[coalminerdf$Age=="35-39"]<-"Middle"
coalminerdf$Career[coalminerdf$Age=="40-44"]<-"Middle"
coalminerdf$Career[coalminerdf$Age=="45-49"]<-"Middle"

coalminerdf$Career[coalminerdf$Age=="50-54"]<-"Late"
coalminerdf$Career[coalminerdf$Age=="55-59"]<-"Late"
coalminerdf$Career[coalminerdf$Age=="60-64"]<-"Late"
```

```
#Creating 3 tables
```

```
xtabs(Freq~Wheeze+Breathlessness+Career,data=coalminerdf)
```

```
, , Career = Early
```

		Breathlessness	
Wheeze		B	NoB
W		77	282
NoW		28	3517

```
, , 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:

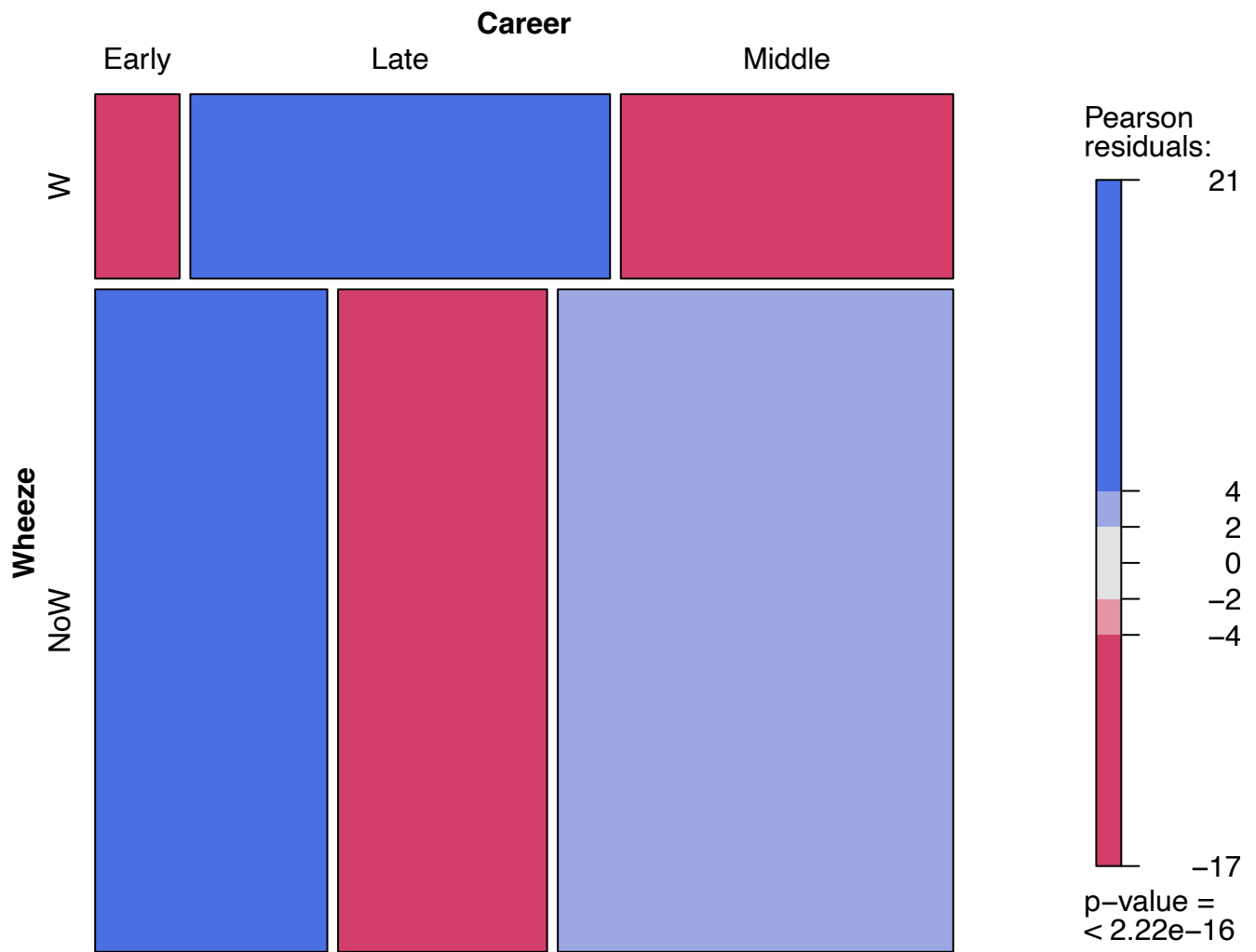
i. Wheeze versus Career

```
#Creating variable for Wheeze vs Carrear
```

```
wc<-xtabs(Freq~Wheeze+Career,data=coalminerdf)
```

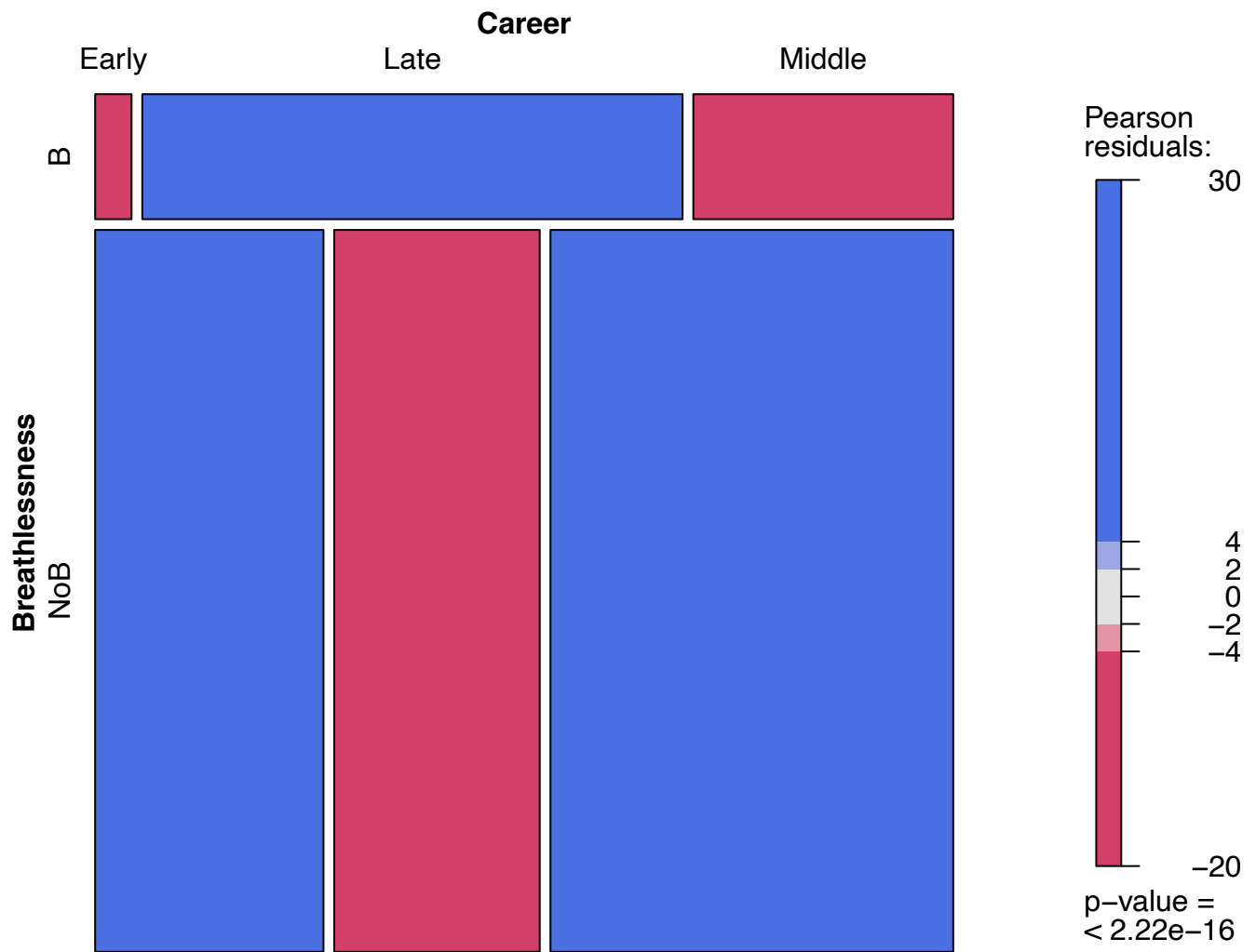
```
#Mosaic Plot
```

```
mosaic(wc,shade = T)
```

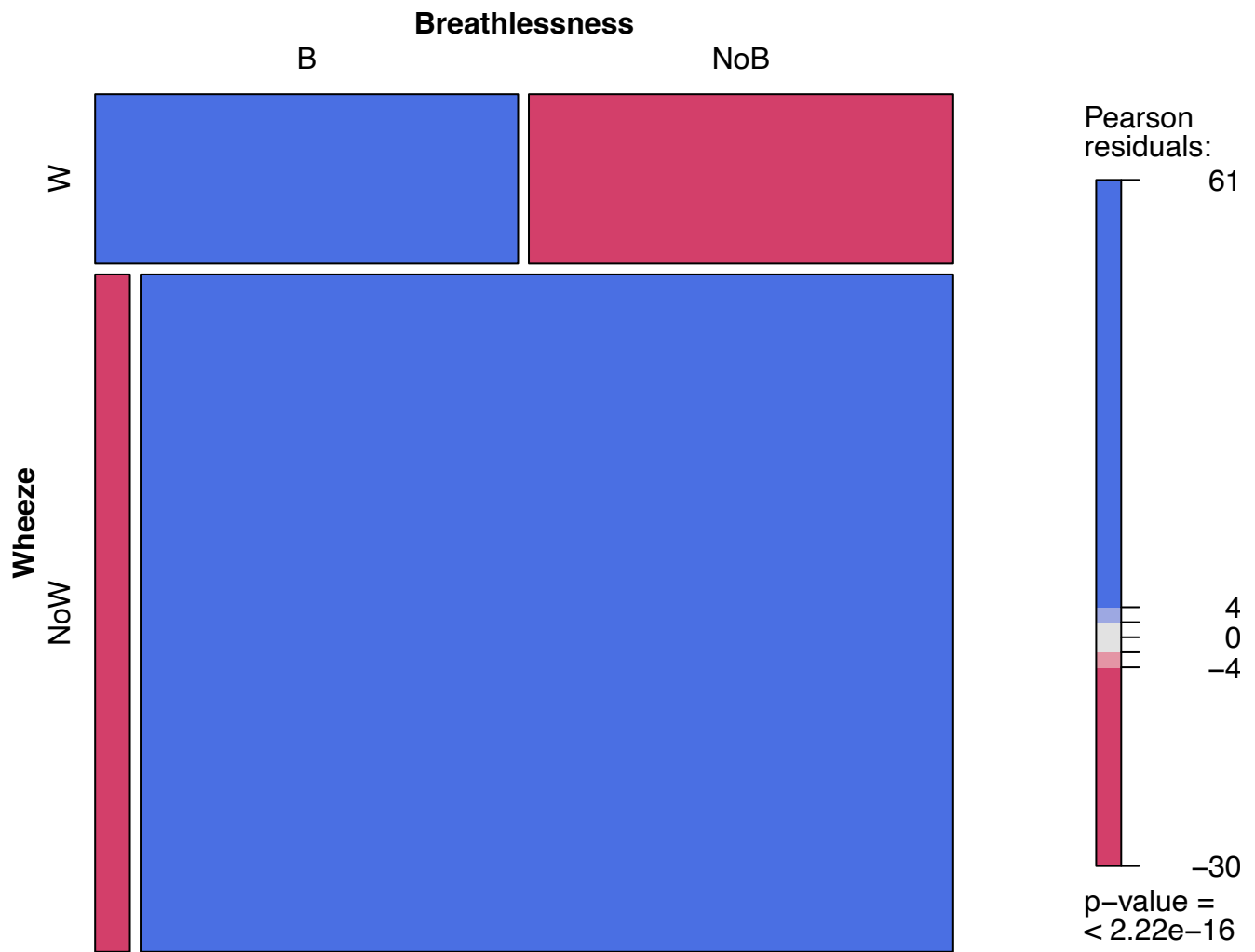
ii. 'Breathlessness' versus 'Career'

```
#Creating variable for Breathlessness vs Carrear
bc<-xtabs(Freq~Breathlessness+Career,data=coalminerdf)
#Mosaic Plot
mosaic(bc,shade = T)
```



iii. 'Wheeze' versus 'Breathlessness'

```
#Creating variable for Wheeze vs Breathlessness
wb<-xtabs(Freq~Wheeze+Breathlessness,data=coalminerdf)
#Mosaic Plot
mosaic(wb,shade = T)
```



iv. Comment on the results.

#There is no correlation between breathlessness and wheeze late in career.

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.

#Chi-square using chiq-test function
chisq.test(wc)

Pearson's Chi-squared test

data: wc
X-squared = 976.05, df = 2, p-value < 2.2e-16

```
chisq.test(bc)
```

Pearson's Chi-squared test

```
data: bc
```

```
X-squared = 1663.6, df = 2, p-value < 2.2e-16
```

```
chisq.test(wb,correct=F)
```

Pearson's Chi-squared test

```
data: wb
```

```
X-squared = 5336.8, df = 1, p-value < 2.2e-16
```

```
#Because the p-value is less than 0.05 we can say there is correlation between features
```

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.

```
#Reading csv file
```

```
dfacceptance<-read.csv("/Users/eduardosalvador/Desktop/FINAL\ Spring\ Semester\ 2021/CMDA\ /Assignments/HW8/ac
```

```
#Reading first 5 rows
```

```
head(dfacceptance)
```

	Vocation	Gender	Accepted
1	Cosmetology	Female	No
2	Plumbing	Male	Yes
3	Welding	Male	No
4	Nursing	Female	No
5	Welding	Male	No
6	Plumbing	Male	Yes

```
#Summerizing into 3d array
```

```
byVoc<-table(dfacceptance$Gender,dfacceptance$Accepted,dfacceptance$Vocation)
```

```
byVoc
```

```
, , = Cosmetology
```

	No	Yes
Female	515	40
Male	582	36

```
, , = Nursing
```

	No	Yes
Female	404	217
Male	462	229

, , = Plumbing

	No	Yes
Female	31	148
Male	519	848

, , = Welding

	No	Yes
Female	13	28
Male	343	585

```
#Displaing data using contigency table
Cosmetology<-byVoc[,,"Cosmetology"]
Cosmetology
```

	No	Yes
Female	515	40
Male	582	36

```
Nursing<-byVoc[,,"Nursing"]
Nursing
```

	No	Yes
Female	404	217
Male	462	229

```
Plumbing<-byVoc[,,"Plumbing"]
Plumbing
```

	No	Yes
Female	31	148
Male	519	848

```
Welding<-byVoc[,,"Welding"]
Welding
```

	No	Yes
Female	13	28
Male	343	585

```
#Flatting tables
ftable(Cosmetology)
```

	No	Yes
Female	515	40
Male	582	36

```
ftable(Nursing)
```

```
      No Yes
```

```
Female 404 217
```

```
Male   462 229
```

```
ftable(Plumbing)
```

```
      No Yes
```

```
Female  31 148
```

```
Male   519 848
```

```
ftable(Welding)
```

```
      No Yes
```

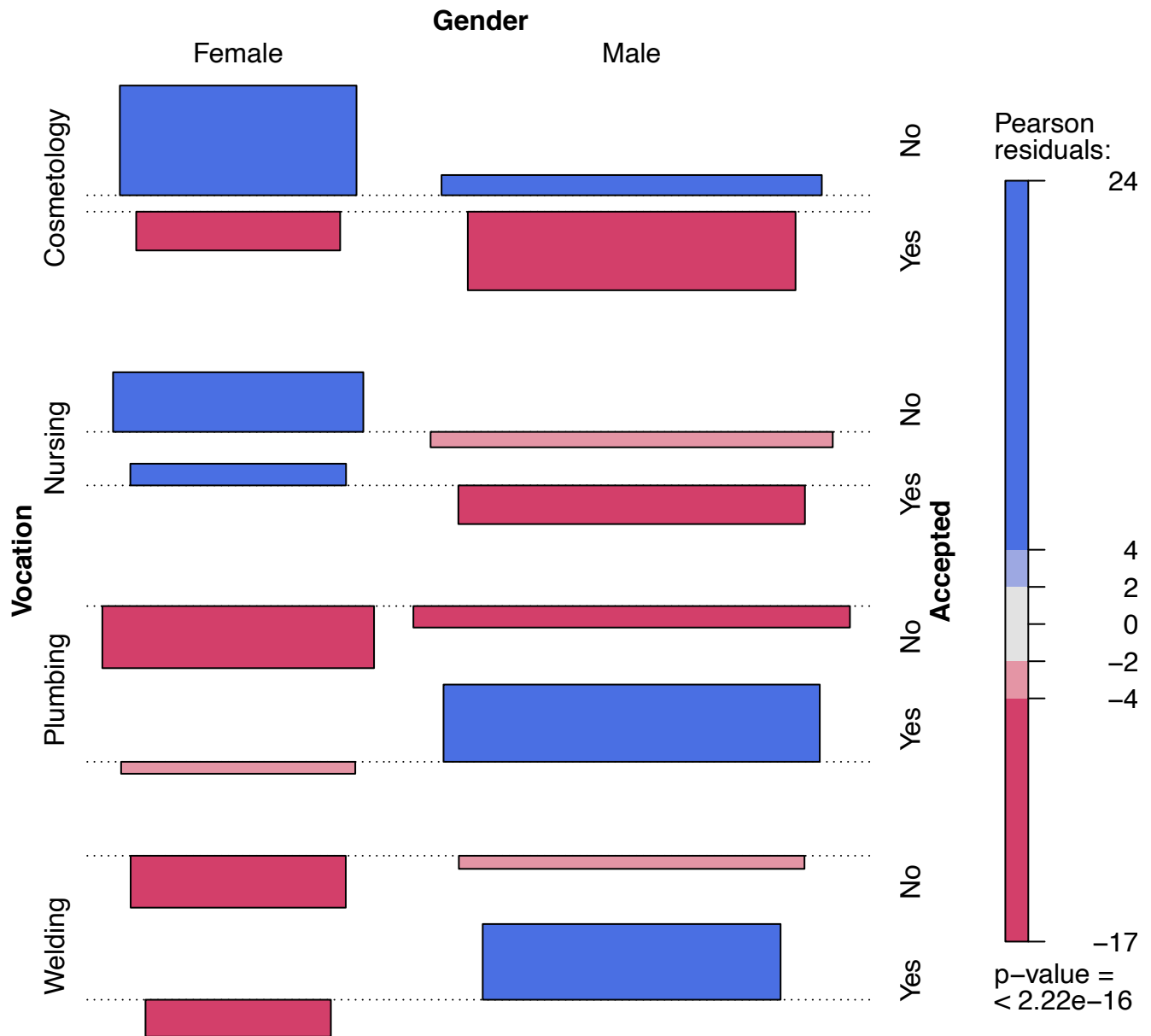
```
Female  13  28
```

```
Male   343 585
```

- 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.

```
#Association plot
```

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



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

#Using `chisq.test` function to carry out chi-square test of independence for every Vocation

```
chisq.test(Cosmetology)
```

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

```
data: Cosmetology
X-squared = 0.70766, df = 1, p-value = 0.4002
```

```
chisq.test(Nursing)
```

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

```
data: Nursing
X-squared = 0.39703, df = 1, p-value = 0.5286
```

```
chisq.test(Plumbing)
```

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

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

```
chisq.test(Welding)
```

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

```
data: Welding
X-squared = 0.26768, df = 1, p-value = 0.6049
```

```
#Because the p-value is greater than 0.05 this means that there could be a lack
#of association between Gender and Acceptance in Cosmetology, Nursing and Welding but,
#for Plumbing since the p-value is less than 0.05, there is association.
```

- 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`.

```
#Creating xtabs variable to ignore Vocation
IgVoc<-xtabs(~Gender+Accepted,data=dfacceptance)
IgVoc
```

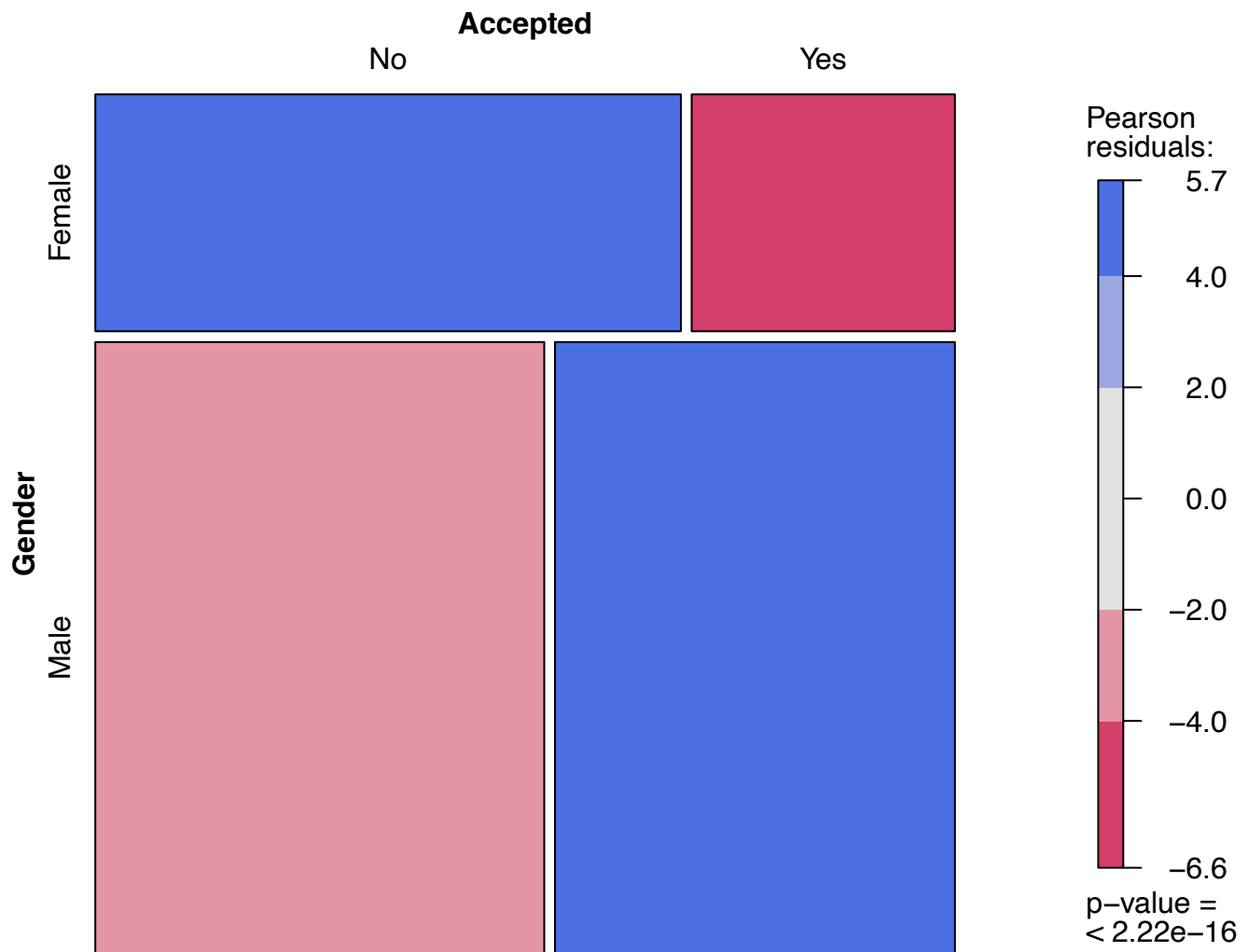
	Accepted	
Gender	No	Yes
Female	963	433
Male	1906	1698

```
#Doing the chi-square test while ignoring Vocation
chisq.test(IgVoc,correct = F )
```

Pearson's Chi-squared test

```
data: IgVoc
X-squared = 106.62, df = 1, p-value < 2.2e-16
```

```
#Creating mosaic plot using mosaic function
mosaic(IgVoc,shade = T)
```

#Because the p-value is less than 0.05, there is association between Gender and Acceptance ignoring Vocation.

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

```
#Producing a common(weighted) odds ratio using mantelhaen.test() from the stats package
library(stats)
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
```

```
#Because the p=value is less than 0.05 we can say that there is indeed association between gender and
#acceptance doing the cmh chi-square test
```

- 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?

```
#It is hard to say because for b-e there is association in all of the results but, for c there is
#no association so my conclusion would be that there is a high chance of association between acceptance and 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?

```
#Creating variables based on success rate
male<-c(prop.table(Cosmetology,2)[4],prop.table(Nursing,2)[4],prop.table(Plumbing,2)[4],prop.table(Welding,2)[4])
female<-c(prop.table(Cosmetology,2)[2],prop.table(Nursing,2)[2],prop.table(Plumbing,2)[2],prop.table(Welding,2)[2])

#Cbinding male and female success rates
combination<-cbind(male,female)
rownames(combination)<-c("Cosmetology","Nursing","Plumbing","Welding")
combination
```

	male	female
Cosmetology	0.4736842	0.5305378
Nursing	0.5134529	0.5334873
Plumbing	0.8514056	0.9436364
Welding	0.9543230	0.9634831

```
#My empirical conclusion looking at the success rate for male and female applicants in each Vocation is that
#there is no gender bias whatsoever since the percentages for male and female are not too far apart.
```

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?

```
#Loading and summarizing data to find the amount of transactions and most frequent item
library(arules)
data("Groceries",package = "arules")
summary(Groceries)
```

transactions as itemMatrix in sparse format with
 9835 rows (elements/itemsets/transactions) and
 169 columns (items) and a density of 0.02609146

most frequent items:

whole milk	other vegetables	rolls/buns	soda
2513	1903	1809	1715
yogurt	(Other)		
1372	34055		

element (itemset/transaction) length distribution:
 sizes

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
2159	1643	1299	1005	855	645	545	438	350	246	182	117	78	77	55	46
17	18	19	20	21	22	23	24	26	27	28	29	32			
29	14	14	9	11	4	6	1	1	1	1	3	1			

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	2.000	3.000	4.409	6.000	32.000

includes extended item information - examples:

	labels	level2	level1
1	frankfurter	sausage meat	and sausage
2	sausage	sausage meat	and sausage
3	liver loaf	sausage meat	and sausage

#There are 9835 rows of transactions and the whole milk is the most frequent item with 2513 times appearance.

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

#Looking at the summary table the percentage of transactions involving 20 or more items

are 12/29 which is around 41%.

#On average there were 339.13793103448 items involved per transaction by summing all the items

#divided by the number of transactions from summary

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?

#Finding all rules with rule function

#rules(Groceries,parameter=list(supp=0.01,conf=0.5))

Problem 4 [10 pts Extra Credit]

Continue working with the data in problem 3.

- 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.
- 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.