Homework 3 HDS 823, Multiple Correspondence Analysis - Titanic.

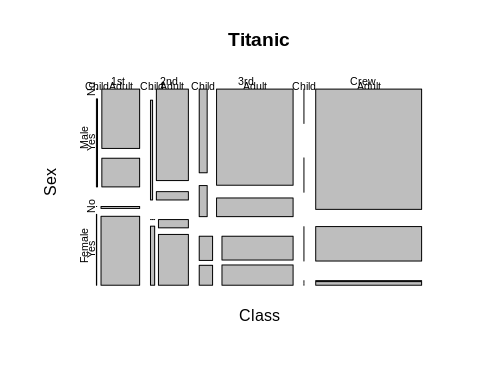
This analysis performed in R, using the FactoMineR library.

Kyle P. Rasku RN BSN

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

In RStudio, you can execute each chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

plot(Titanic)



Titanic Data Set - Modified for MCA by Nick Buonomia Perform Multiple Correspondence Analysis of Class, Age & Sex with Supplemental Variable: Survived

library("FactoMineR")  
library("factoextra")

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

titanic <- read.csv("titanic\_MCA.csv")  
titanic.active <- titanic[1:2200, 2:5]  
head(titanic.active[, 1:4], 10)

## Class Age Sex Survived  
## 1 first adult male yes  
## 2 first adult male yes  
## 3 first adult male yes  
## 4 first adult male yes  
## 5 first adult male yes  
## 6 first adult male yes  
## 7 first adult male yes  
## 8 first adult male yes  
## 9 first adult male yes  
## 10 first adult male yes

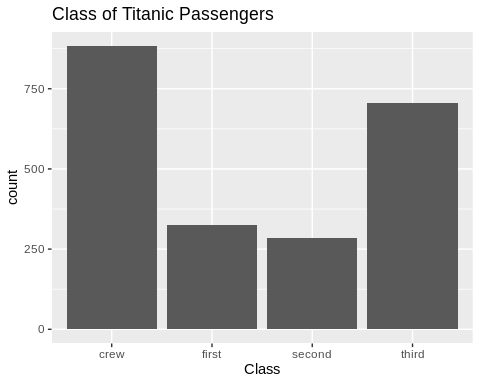
Active Rows for MCA are set. Provide data summary of active rows:

summary(titanic.active)[, 1:4]

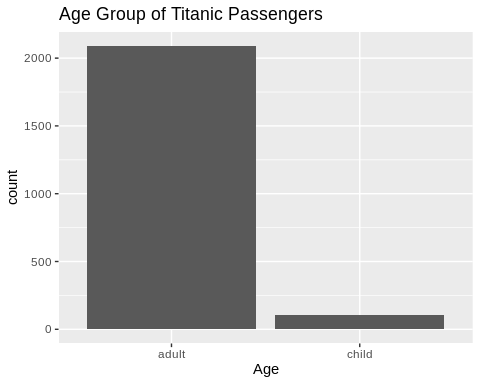
## Class Age Sex Survived   
## Length:2200 Length:2200 Length:2200 Length:2200   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character

This data set has 2,200 rows and all variables are categorical. Plot frequencies:

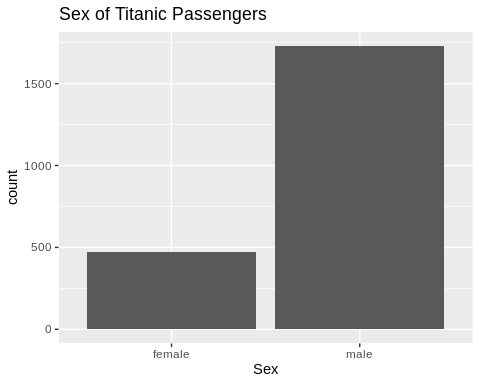
library(ggplot2)  
ggplot(titanic.active, aes(x = Class)) + geom\_bar() + labs(title = "Class of Titanic Passengers")



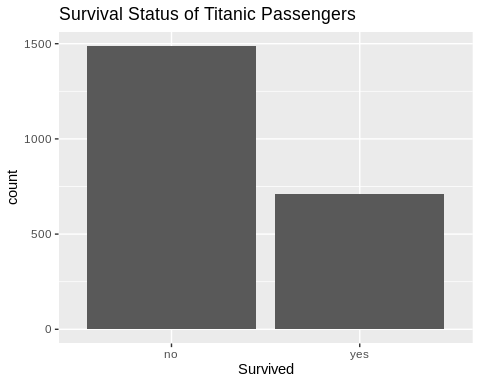
ggplot(titanic.active, aes(x = Age)) + geom\_bar() + labs(title = "Age Group of Titanic Passengers")



ggplot(titanic.active, aes(x = Sex)) + geom\_bar() + labs(title = "Sex of Titanic Passengers")



ggplot(titanic.active, aes(x = Survived)) + geom\_bar() + labs(title = "Survival Status of Titanic Passengers")



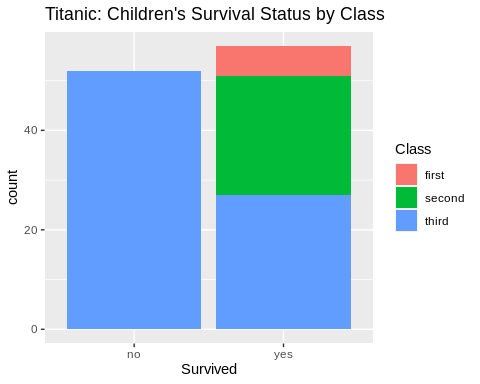
Looking more closely at Children:

children <- subset(titanic, Age=="child")  
nrow(children)

## [1] 109

There are 109 children out of 2200 people, or 5% of all the passengers. How many Children Survived by Class?

ggplot(data = children, aes(x = Survived, fill = Class)) +  
 geom\_bar() + labs(title = "Titanic: Children's Survival Status by Class")



library(sqldf)

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

num\_died <- sqldf('SELECT count(\*) from children where Survived="no"')  
num\_third\_survived <- sqldf('SELECT count(\*) from children where Survived="yes" and Class="third"')  
num\_second\_survived <- sqldf('SELECT count(\*) from children where Survived="yes" and Class="second"')  
num\_first\_survived <- sqldf('SELECT count(\*) from children where Survived="yes" and Class="first"')  
num\_third\_children = num\_died + num\_third\_survived

Number of First Class Children on Board: 6, Survived: 6 Number of Second Class Children on Board: 24, Survived: 24 Number of Third Class Children on Board: 79, Survived: 27 (34.18%)

While there were fewer children on board with parents holding First and Second Class tickets, and many more with parents holding Third Class tickets, 100% of the children with parents holding First and Second class tickets survived the sinking, while only about 34% of the children with parents holding Third Class tickets survived the sinking.

# Perform MCA, only active variables: Class, Sex and Age  
# The SUPPLEMENTAL variable is "Survived"  
# Use the "Burt" method for calculation (per Nick)  
res.mca <- MCA(titanic.active, quali.sup=4, method="burt", graph = FALSE)  
res.mca

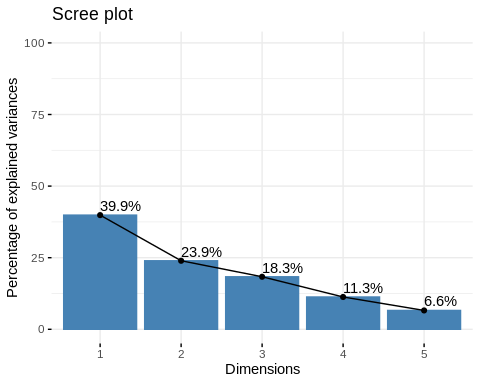
## \*\*Results of the Multiple Correspondence Analysis (MCA)\*\*  
## The analysis was performed on 2200 individuals, described by 4 variables  
## \*The results are available in the following objects:  
##   
## name description   
## 1 "$eig" "eigenvalues"   
## 2 "$var" "results for the variables"   
## 3 "$var$coord" "coord. of the categories"   
## 4 "$var$cos2" "cos2 for the categories"   
## 5 "$var$contrib" "contributions of the categories"   
## 6 "$var$v.test" "v-test for the categories"   
## 7 "$ind" "results for the individuals"   
## 8 "$ind$coord" "coord. for the individuals"   
## 9 "$ind$cos2" "cos2 for the individuals"   
## 10 "$ind$contrib" "contributions of the individuals"   
## 11 "$quali.sup" "results for the supplementary categorical variables"  
## 12 "$quali.sup$coord" "coord. for the supplementary categories"   
## 13 "$quali.sup$cos2" "cos2 for the supplementary categories"   
## 14 "$quali.sup$v.test" "v-test for the supplementary categories"   
## 15 "$call" "intermediate results"   
## 16 "$call$marge.col" "weights of columns"   
## 17 "$call$marge.li" "weights of rows"

Intermediate Results:

# Commented out - this does include the Burt Tables, but it is about 100 pages long.  
#res.mca$call

The proportions of variance retained by the different dimensions, extracted to Eigenvalues:

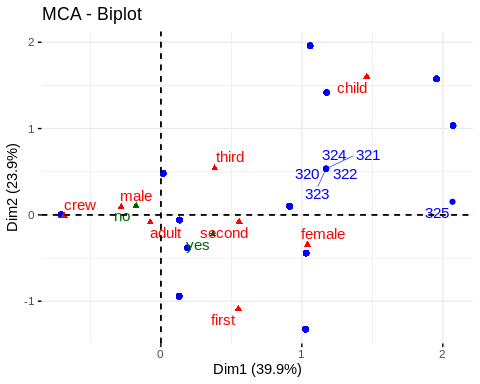
eig.val <- get\_eigenvalue(res.mca)  
fviz\_screeplot(res.mca, addlabels = TRUE, ylim = c(0, 99))



Using the Burt method, 39.5% of the variance is explained by the 1st Dimension, and 23.9% is explained by the 2nd. An additional 18.3% is explained by a 3rd. This improves variance account in Eigenvectors 1 and 2 by about 10%. 100% of the variance is explained by 5 Dimensions. Therefore, if we could plot the data in 3 Dimensions, we’d be able to see about 82% of the variance. Unfortunately, with MCA two dimensions must be visualized at a time. A visualization of Dimensions 1 and 2 or 1 and 3 will cover about 64% of the variance. A visualization of 2 and 3 would cover 42%.

# This should be quite messy, with 2200 individuals...  
# There were so many overlapping data points that the library chose not to show them  
fviz\_mca\_biplot(res.mca,   
 repel = TRUE, # Avoid text overlapping (slow if many point)  
 ggtheme = theme\_minimal())

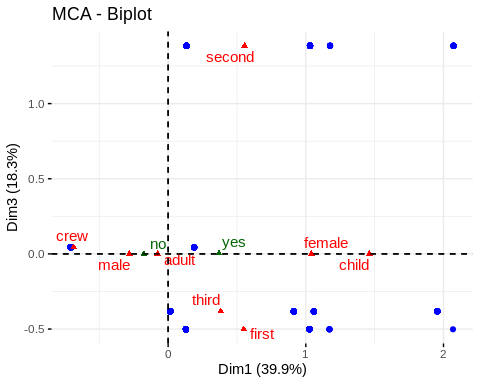
## Warning: ggrepel: 2194 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps



It looks like Dim 1 is quite like a survival continuum with children and women on the right, and males and crew members on the left. It takes a minute to see what Dim 2 is about, but it looks like Age - with children at the top and older people at the bottom.

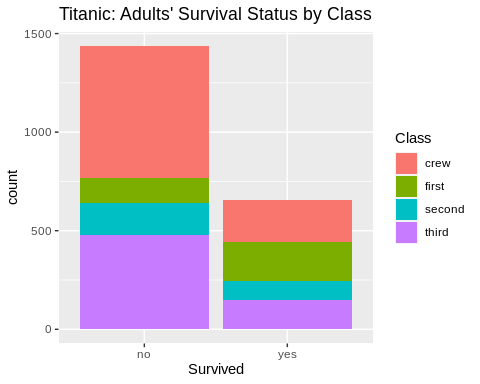
# Now let's look at Dimensions 1 and 3  
fviz\_mca\_biplot(res.mca,   
 axes = c(1,3),  
 repel = TRUE, # Avoid text overlapping (slow if many point)  
 ggtheme = theme\_minimal())

## Warning: ggrepel: 2200 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps

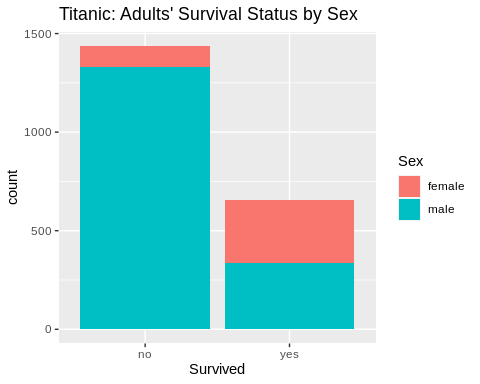


Hmm…ok. So Here we have the crew and children at polar opposite extremes. And on Dim3, first and third class are at the bottom of Dim3 while second class if very far away near the top.

adults <- subset(titanic, Age=="adult")  
ggplot(data = adults, aes(x = Survived, fill = Class)) +  
 geom\_bar() + labs(title = "Titanic: Adults' Survival Status by Class")



ggplot(data = adults, aes(x = Survived, fill = Sex)) +  
 geom\_bar() + labs(title = "Titanic: Adults' Survival Status by Sex")



num\_adults\_died <- sqldf('SELECT count(\*) from adults where Survived="no"')  
num\_adults\_lived <- sqldf('SELECT count(\*) from adults where Survived="yes"')  
num\_tadults\_survived <- sqldf('SELECT count(\*) from adults where Survived="yes" and Class="third"')  
num\_tadults\_died <- sqldf('SELECT count(\*) from adults where Survived="no" and Class="third"')  
num\_sadults\_survived <- sqldf('SELECT count(\*) from adults where Survived="yes" and Class="second"')  
num\_sadults\_died <- sqldf('SELECT count(\*) from adults where Survived="no" and Class="second"')  
num\_fadults\_survived <- sqldf('SELECT count(\*) from adults where Survived="yes" and Class="first"')  
num\_fadults\_died <- sqldf('SELECT count(\*) from adults where Survived="no" and Class="first"')  
num\_tadults = num\_tadults\_survived + num\_tadults\_died  
num\_sadults = num\_sadults\_survived + num\_sadults\_died  
num\_fadults = num\_fadults\_survived + num\_fadults\_died

There were 2092 adults on board Titanic. 1438 adults died (68.74%), 654 (31.26%) survived. There were 319 adults in first class. 197 of them (62%) survived. There were 261 adults in second class. 94 of them (36%) survived. There were 627 adults in third class. 151 of them (24%) survived.

num\_ffemales <- sqldf('SELECT count(\*) from adults where Sex="female" and Class="first"')  
num\_ffemales\_survived <- sqldf('SELECT count(\*) from adults where Sex="female" and Class="first" and Survived="yes"')  
num\_sfemales <- sqldf('SELECT count(\*) from adults where Sex="female" and Class="second"')  
num\_sfemales\_survived <- sqldf('SELECT count(\*) from adults where Sex="female" and Class="second" and Survived="yes"')  
num\_tfemales <- sqldf('SELECT count(\*) from adults where Sex="female" and Class="third"')  
num\_tfemales\_survived <- sqldf('SELECT count(\*) from adults where Sex="female" and Class="third" and Survived="yes"')  
num\_cfemales <- sqldf('SELECT count(\*) from adults where Sex="female" and Class="crew"')  
num\_cfemales\_survived <- sqldf('SELECT count(\*) from adults where Sex="female" and Class="crew" and Survived="yes"')

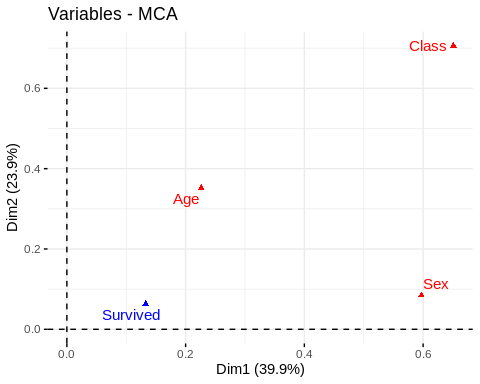
There were 144 women traveling first class, and 140 of them survived the sinking (97%). As we know from before, 100% of the first class children survived (but there were only 6, according to this data set). There were 93 women traveling second class, and 80 of them survived (86%). Again, all the second class children survived. There were 165 women traveling third class, and 76 of them survived (46%). And 34% of their children (calculated earlier). While the crew fared poorly in general in terms of survival, there were only 23 women on the crew, and 20 of them survived (87%).

# Looking at the coordinates, cos2 and contribution of variable categories  
var <- get\_mca\_var(res.mca)  
var

## Multiple Correspondence Analysis Results for variables  
## ===================================================  
## Name Description   
## 1 "$coord" "Coordinates for categories"   
## 2 "$cos2" "Cos2 for categories"   
## 3 "$contrib" "contributions of categories"

The correlation between variables and principal dimensions:

fviz\_mca\_var(res.mca, choice = "mca.cor",   
 repel = TRUE, # Avoid text overlapping (slow)  
 ggtheme = theme\_minimal())



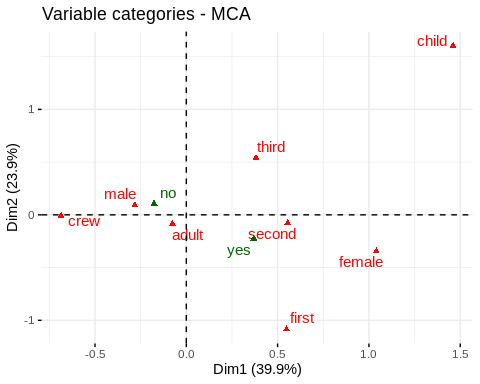
Class contributes strongly to both Dim 1 and 2. Sex contributes strongly to Dim 1 but NOT Dim 2. Age is a moderate contributor to both Dim 1 and Dim 2 and Survived is more of a contributor to Dim 1 than 2.

# Category / Value Coordinates  
round(var$coord, 2)

## Dim 1 Dim 2 Dim 3 Dim 4 Dim 5  
## crew -0.69 -0.01 0.04 0.17 0.29  
## first 0.55 -1.09 -0.50 0.44 -0.34  
## second 0.55 -0.08 1.38 -0.10 -0.24  
## third 0.38 0.54 -0.38 -0.38 -0.11  
## adult -0.08 -0.08 0.00 -0.08 0.00  
## child 1.46 1.60 0.00 1.45 0.08  
## female 1.04 -0.34 0.00 -0.16 0.46  
## male -0.28 0.09 0.00 0.04 -0.13

Ok, based on this, it looks like Dim 1 is about first and third class, children and women. Dim 2 is about third class and children. Dim 3 is about second class. Dim 4 is another share of first class and children, but also crew. And Dim 5 is predominantly crew. If wanting to know how a particular class fared in terms of survival, this would be important to note. You may not be able to look at Dim 1 and 2 and draw conclusions about the second class passengers for example.

# Visualizing Category / Value Coordinates  
fviz\_mca\_var(res.mca,   
 repel = TRUE, # Avoid text overlapping (slow)  
 ggtheme = theme\_minimal())



Variable categories with a similar profile are grouped together. (Example: Male and Adult) Negatively correlated variable categories are positioned on opposite sides of the plot origin (opposed quadrants). (Crew vs. Child) The distance between category points and the origin measures the **quality** of the variable category on the factor map. Category points that are away from the origin are *well represented* on the factor map. Adults are undifferentiated. Non-survival, adults and second class are less well represented here. Children, first class and females better represented.

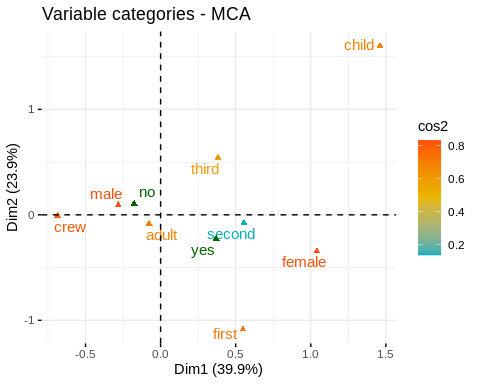
Degrees of association between variable categories and axes:

head(var$cos2, 10)

## Dim 1 Dim 2 Dim 3 Dim 4 Dim 5  
## crew 0.8045008 0.0001163271 3.251087e-03 0.050969692 0.1411620606  
## first 0.1478079 0.5780531707 1.235559e-01 0.095184209 0.0553988237  
## second 0.1339680 0.0026007050 8.343042e-01 0.003940801 0.0251863344  
## third 0.1965379 0.3951209036 1.974787e-01 0.195201720 0.0156607115  
## adult 0.3130468 0.3770377740 6.117649e-28 0.308988289 0.0009271487  
## child 0.3130468 0.3770377740 8.214726e-28 0.308988289 0.0009271487  
## female 0.7508468 0.0821925641 1.616615e-30 0.017879783 0.1490808967  
## male 0.7508468 0.0821925641 6.034649e-30 0.017879783 0.1490808967

Dim 1 is representing the crew, a small equalish amt of all 3 classes, about equal amounts of adults & children, males & females. Dim 2 is representing first class, and a good chunk of third class. Also some more adults and children. Dim 3 is a hodgepodge with lion’s share of second class, but there’s not much there (large negative exp values). Dim 4 is a bit more of third class, more of the variance of adults & children. Dim 5 is the rest of the crew and males & females.

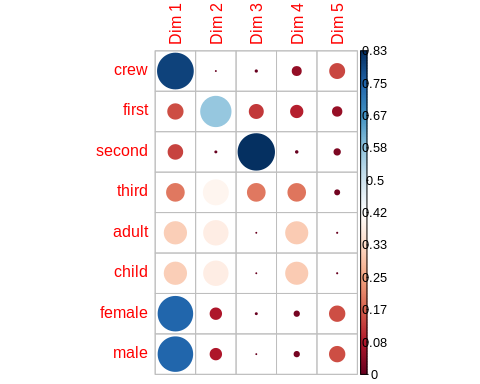
# Color by cos2 values: quality on the factor map  
fviz\_mca\_var(res.mca, col.var = "cos2",  
 gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),   
 repel = TRUE, # Avoid text overlapping  
 ggtheme = theme\_minimal())

 The lower quality of the Second class and - to a smaller extent - Third class data points is reiterated above.

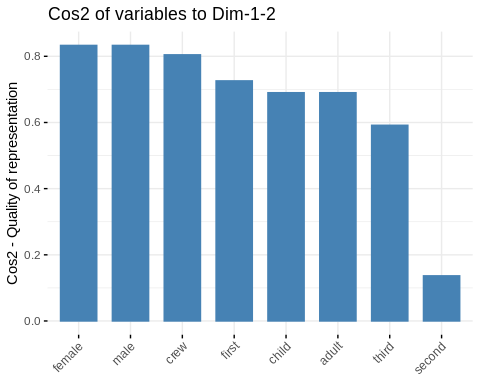
library("corrplot")

## corrplot 0.90 loaded

corrplot(var$cos2, is.corr=FALSE)

 Another view of this with corrplot library. If we want to get a good look at survival vv. Second class, we may need to look beyond Dims 1 & 2.

# Cos2 of variable categories on Dim.1 and Dim.2  
fviz\_cos2(res.mca, choice = "var", axes = 1:2)

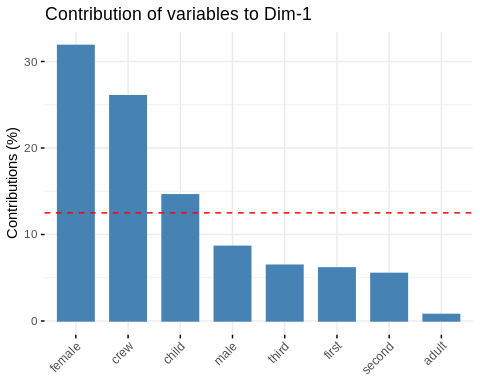


# Contributions of Variable Categories to the Dimensions  
head(round(var$contrib,2), 10)

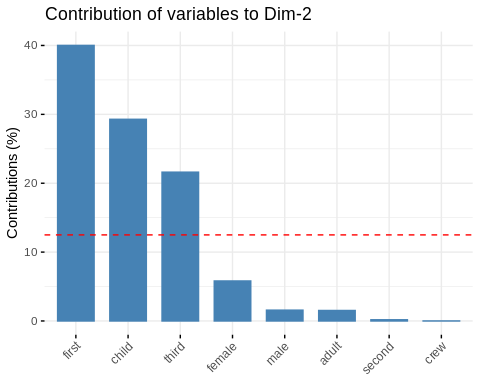
## Dim 1 Dim 2 Dim 3 Dim 4 Dim 5  
## crew 26.05 0.01 0.23 5.83 27.69  
## first 6.14 40.01 11.16 13.98 13.94  
## second 5.50 0.18 74.52 0.57 6.27  
## third 6.45 21.61 14.09 22.65 3.11  
## adult 0.76 1.53 0.00 2.65 0.01  
## child 14.59 29.28 0.00 50.91 0.26  
## female 31.86 5.81 0.00 2.68 38.32  
## male 8.63 1.57 0.00 0.73 10.38

The variables that contribute MOST to Dims 1 and 2 explain most of the variation in the data set (53%) - crew, child, female, first & third.

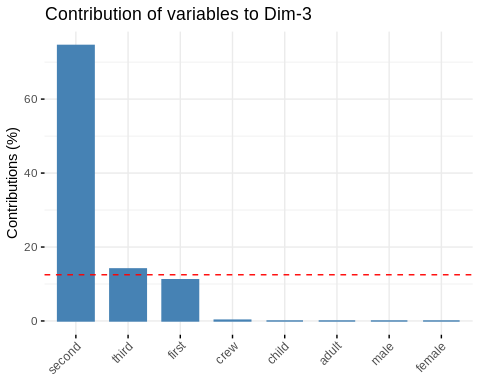
# Contributions of rows to dimension 1  
fviz\_contrib(res.mca, choice = "var", axes = 1, top = 15)



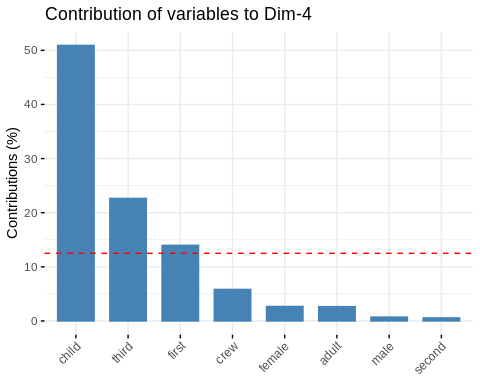
# Contributions of rows to dimension 2  
fviz\_contrib(res.mca, choice = "var", axes = 2, top = 15)



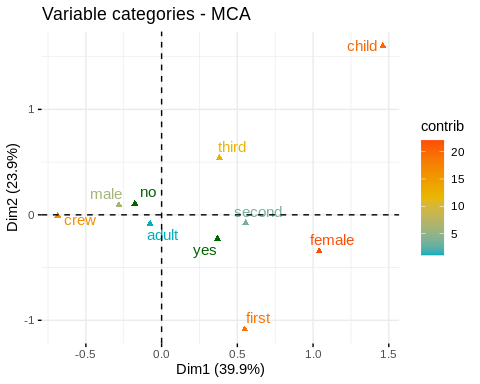
# Contributions of rows to dimension 3  
fviz\_contrib(res.mca, choice = "var", axes = 3, top = 15)



# Contributions of rows to dimension 4  
fviz\_contrib(res.mca, choice = "var", axes = 4, top = 15)

 The red dashed line on the graphs above indicates the *expected average value*, if the contributions were uniform.

fviz\_mca\_var(res.mca, col.var = "contrib",  
 gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),   
 repel = TRUE, # avoid text overlapping (slow)  
 ggtheme = theme\_minimal()  
 )



Graphing Individuals

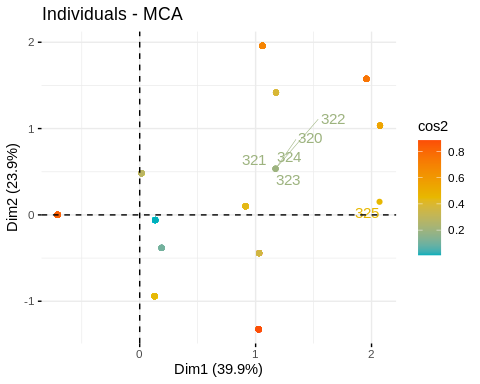
ind <- get\_mca\_ind(res.mca)  
ind

## Multiple Correspondence Analysis Results for individuals  
## ===================================================  
## Name Description   
## 1 "$coord" "Coordinates for the individuals"   
## 2 "$cos2" "Cos2 for the individuals"   
## 3 "$contrib" "contributions of the individuals"

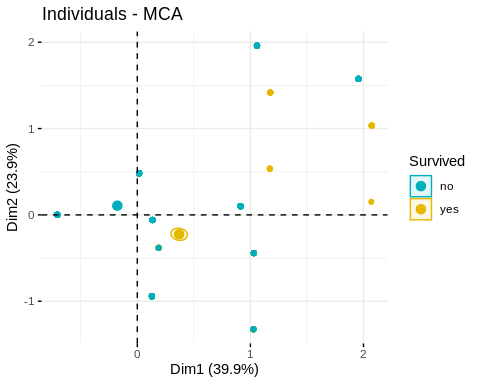
Quality and Contribution of Individuals

fviz\_mca\_ind(res.mca, col.ind = "cos2",   
 gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),  
 repel = TRUE, # Avoid text overlapping (slow if many points)  
 ggtheme = theme\_minimal())

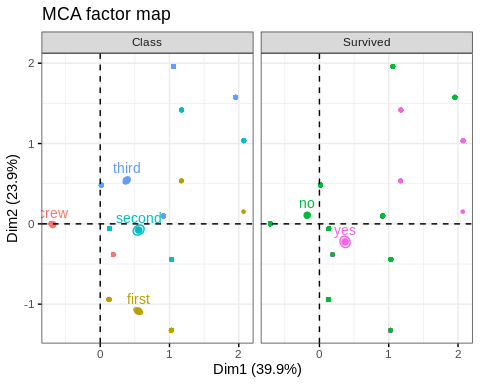
## Warning: ggrepel: 2194 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps



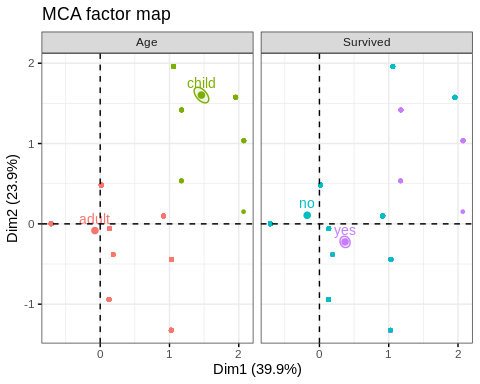
# Apply SURVIVED as Grouping Variable  
fviz\_mca\_ind(res.mca,   
 label = "none", # hide individual labels  
 habillage = "Survived", # color by groups   
 palette = c("#00AFBB", "#E7B800"),  
 addEllipses = TRUE, ellipse.type = "confidence",  
 ggtheme = theme\_minimal())



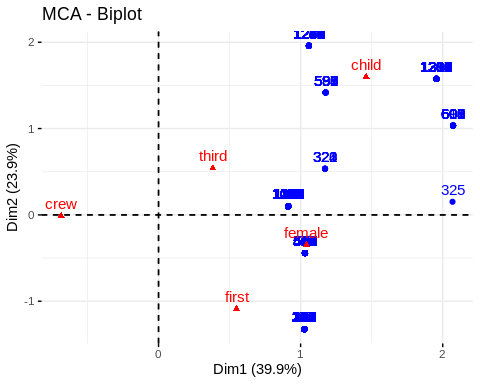
fviz\_ellipses(res.mca, c("Survived", "Class"),  
 geom = "point")



fviz\_ellipses(res.mca, c("Survived", "Age"),  
 geom = "point")



# top 500 contributing individuals and top 5 variable categories  
fviz\_mca\_biplot(res.mca, select.ind = list(contrib = 500),   
 select.var = list(contrib = 5),  
 ggtheme = theme\_minimal())



All in all, the MCA concludes clearly that Men, Crew members and Adults in general were less likely to survive, while Females, Children and First class passengers were much more likely.