Data Visualization with the Titanic Data

Karen Mazidi

Data exploration in R with the Titanic data set.

Graphical Parameters: read more here

Colors in R Graphs: read more here

Load the data

Load the data, changing certain columns to factors.

```
df <- read.csv("data/titanic.csv", na.strings="NA", header=TRUE)
str(df)
```

```
## 'data.frame':    1309 obs. of  14 variables:
##  $ pclass   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ survived : int  1 1 0 0 0 1 1 0 1 0 ...
##  $ name     : chr  "Allen, Miss. Elisabeth Walton" "Allison, Master. Hudson Trevor" "Allison, Miss. H
##  $ sex      : chr  "female" "male" "female" "male" ...
##  $ age      : num  29 0.917 2 30 25 ...
##  $ sibsp    : int  0 1 1 1 1 0 1 0 2 0 ...
##  $ parch    : int  0 2 2 2 2 0 0 0 0 0 ...
##  $ ticket   : chr  "24160" "113781" "113781" "113781" ...
##  $ fare     : num  211 152 152 152 152 ...
##  $ cabin    : chr  "B5" "C22 C26" "C22 C26" "C22 C26" ...
##  $ embarked : chr  "S" "S" "S" "S" ...
##  $ boat     : chr  "2" "11" "" "" ...
##  $ body     : int  NA NA NA 135 NA NA NA NA NA 22 ...
##  $ home.dest: chr  "St Louis, MO" "Montreal, PQ / Chesterville, ON" "Montreal, PQ / Chesterville, ON
```

```
df$survived <- as.factor(df$survived)
df$pclass <- as.factor(df$pclass)
df$sex <- factor(df$sex, levels=c("male", "female"))
```

Plotting X and Y

In this notebook we look at plotting two variables, which gives us four cases:

- both X and Y are qualitative
- X is qualitative, Y is quantitative
- X is quantitative, Y is qualitative
- X and Y ar both quantitative

X,Y both Qualitative

If X and Y are both qualitative data, mosaic and association plots are helpful.

Mosaic and association plot examples

Using the vcd (visualizing categorical data) package.

First, a mosaic example. We want to plot survived and pclass. The mosaic() function wants the first argument to be a table or formula, so we surround the subsetted data frame with table(). SHADE=TRUE gives you a color graph, FALSE gives you a greyscale graph.

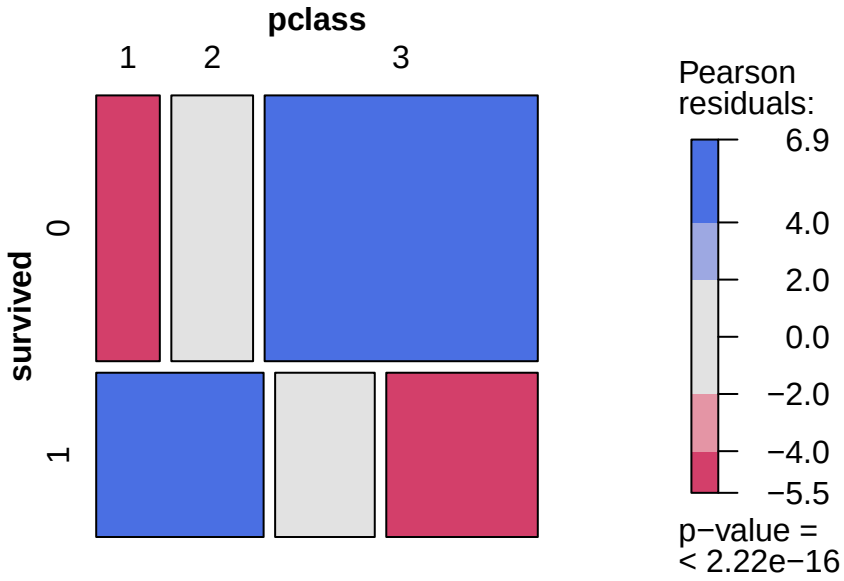
The mosaic plot shows each group in tiles. The area of the tiles is proportional to its counts in the data.

The legend indicates the Pearson residuals. The "null" model would consider an even distribution into the cells but clearly we don't have that case here. The blue indicates we have more observations than expected, the red indicates fewer than expected, and gray is about what is expected given a null hypothesis. We didn't have to specify legend=TRUE because that is the default.

```
library(vcd)
```

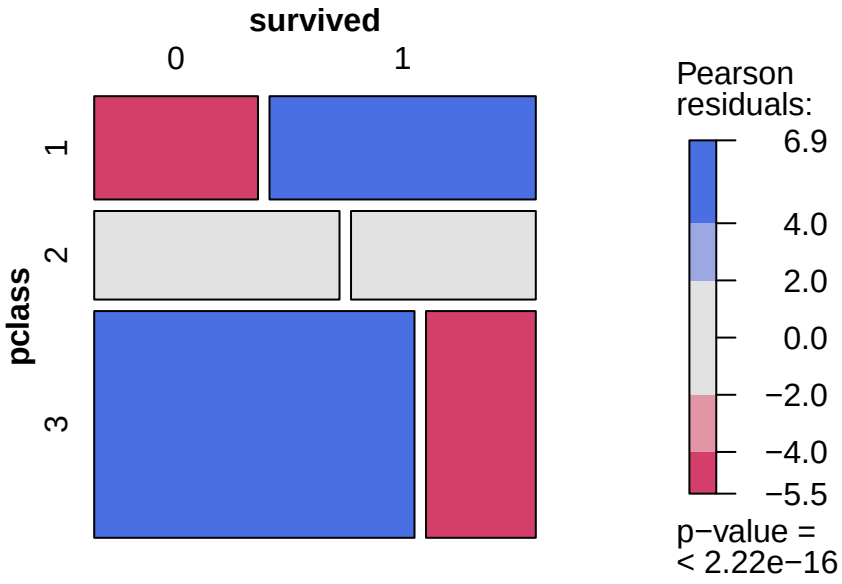
```
## Loading required package: grid
```

```
mosaic(table(df[,c(2,1)]),  shade=TRUE, legend=TRUE)
```



We get the same information if we reverse columns 1 and 2. The graph flips around.

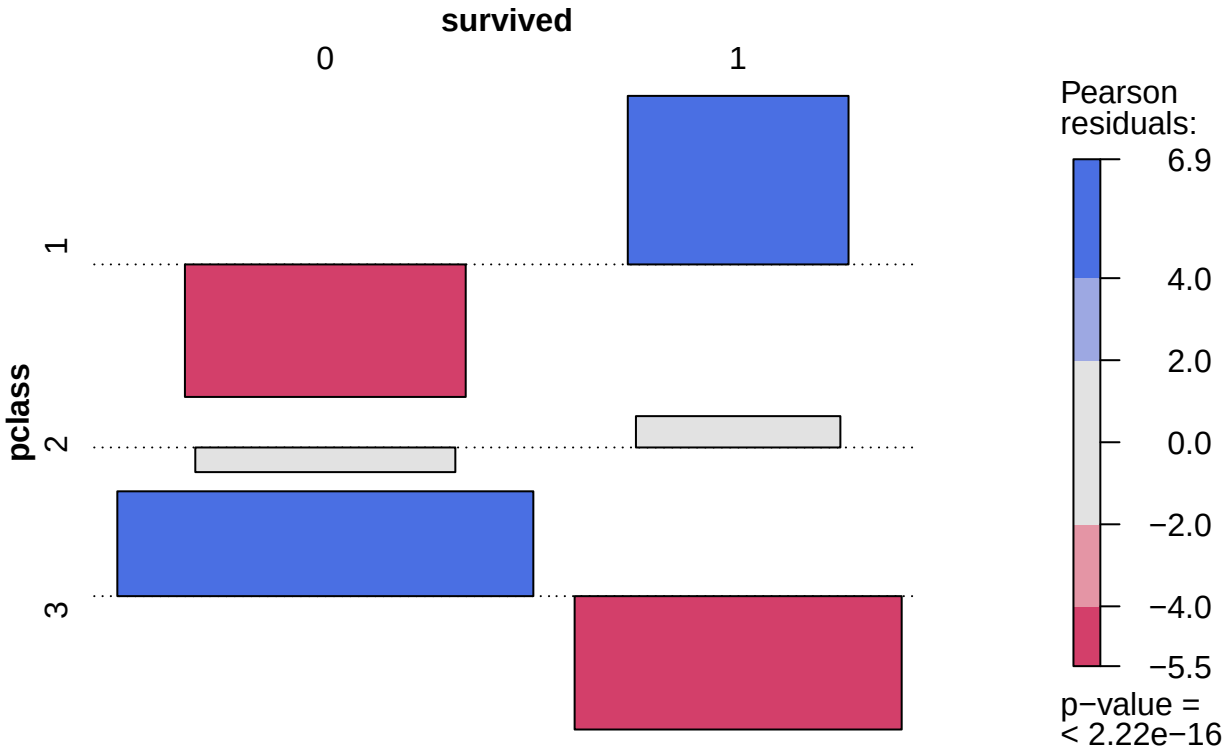
```
mosaic(table(df[,c(1,2)]),  shade=TRUE, legend=TRUE)
```



An association plot visualizes the residuals of an independence model. Each tile has an area that is proportional to the difference in observed and expected frequencies. The dotted line is the baseline. Tiles above the line have a frequency greater than what was expected, those below have a frequency below what was expected.

In the plot below, pclass 1 survived more than expected, pclass 3 less than expected.

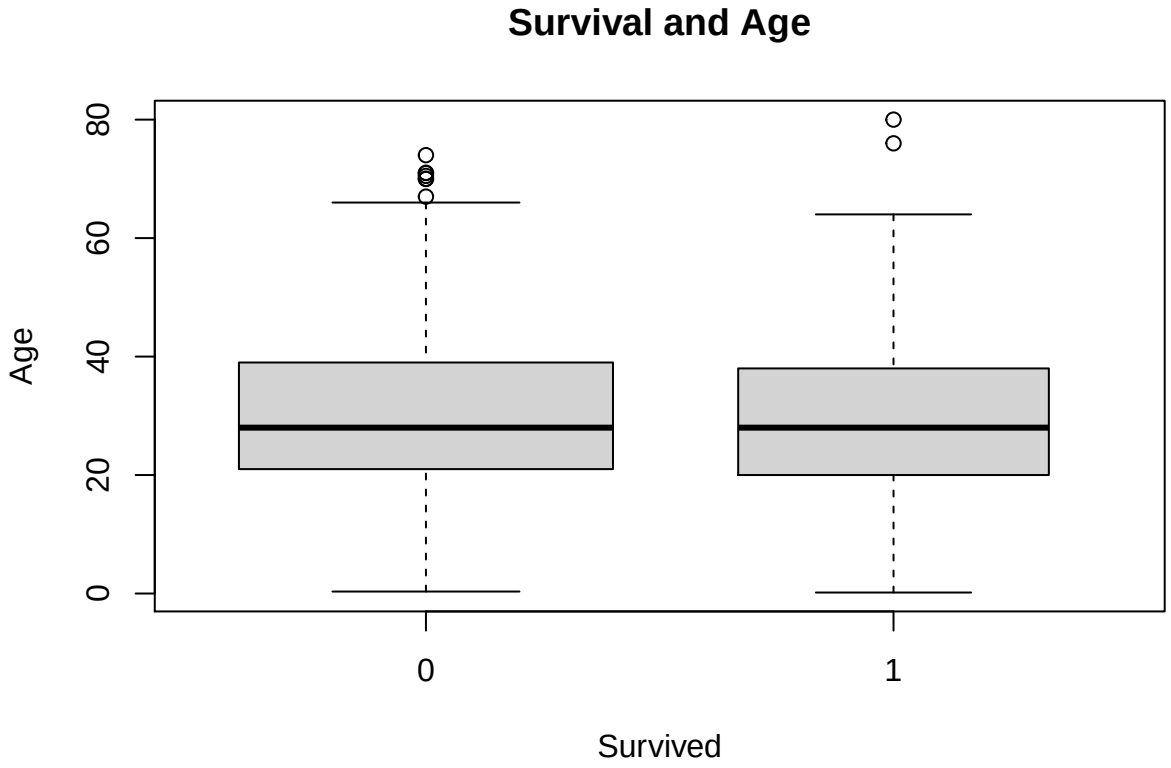
```
assoc(table(df[,c(1,2)]), shade=TRUE)
```



X is Qualitative, Y is Quantitative

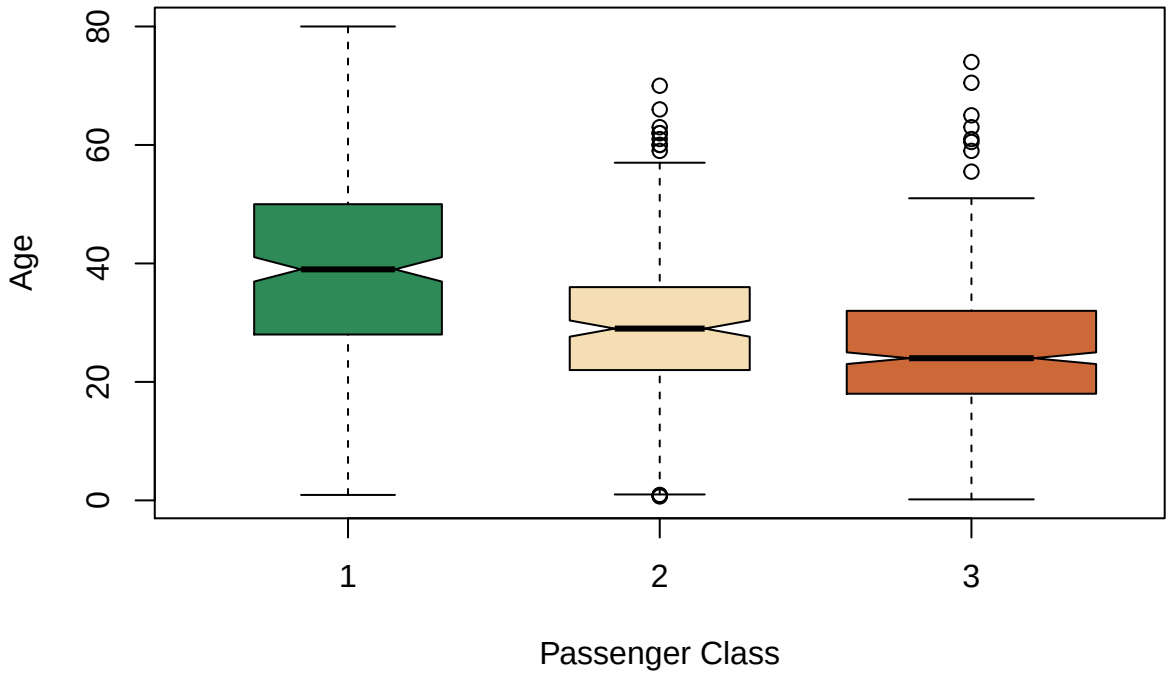
When X is qualitative (a factor), and Y is quantitative, box plots are good choices.

```
plot(df$survived, df$age, varwidth=TRUE, main="Survival and Age", xlab="Survived", ylab="Age")
# the following creates an identical plot
boxplot(df$age~df$survived, varwidth=TRUE, main="Survival and Age", xlab="Survived", ylab="Age")
```



Notches at the median can be added with the notch=TRUE parameter. If the notches do not overlap, then it is likely that medians differ.

```
plot(df$age~df$pclass, varwidth=TRUE, notch=TRUE, xlab="Passenger Class", ylab="Age", col=c("seagreen",
```



Note: You can also create violin plots with package vioplot. Violin plots are a combination of a boxplot and a kernel density plot. This plot does not like NAs so we remove them.

```
library(vioplot)
```

```
## Loading required package: sm
```

```
## Warning in fun(libname, pkgname): couldn't connect to display ":0"
```

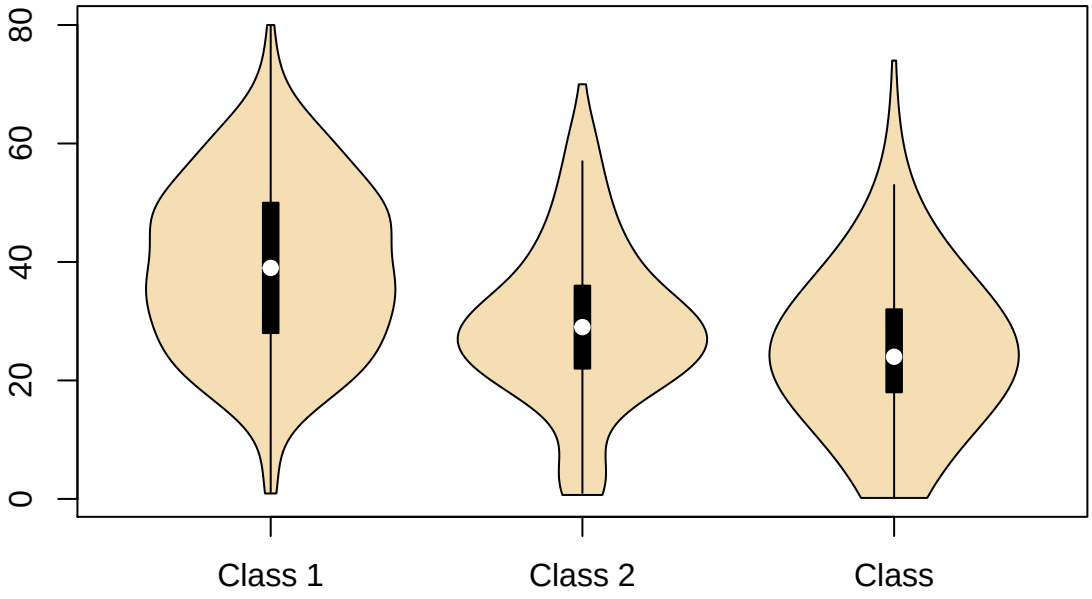
```
## Package 'sm', version 2.2-5.6: type help(sm) for summary information
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

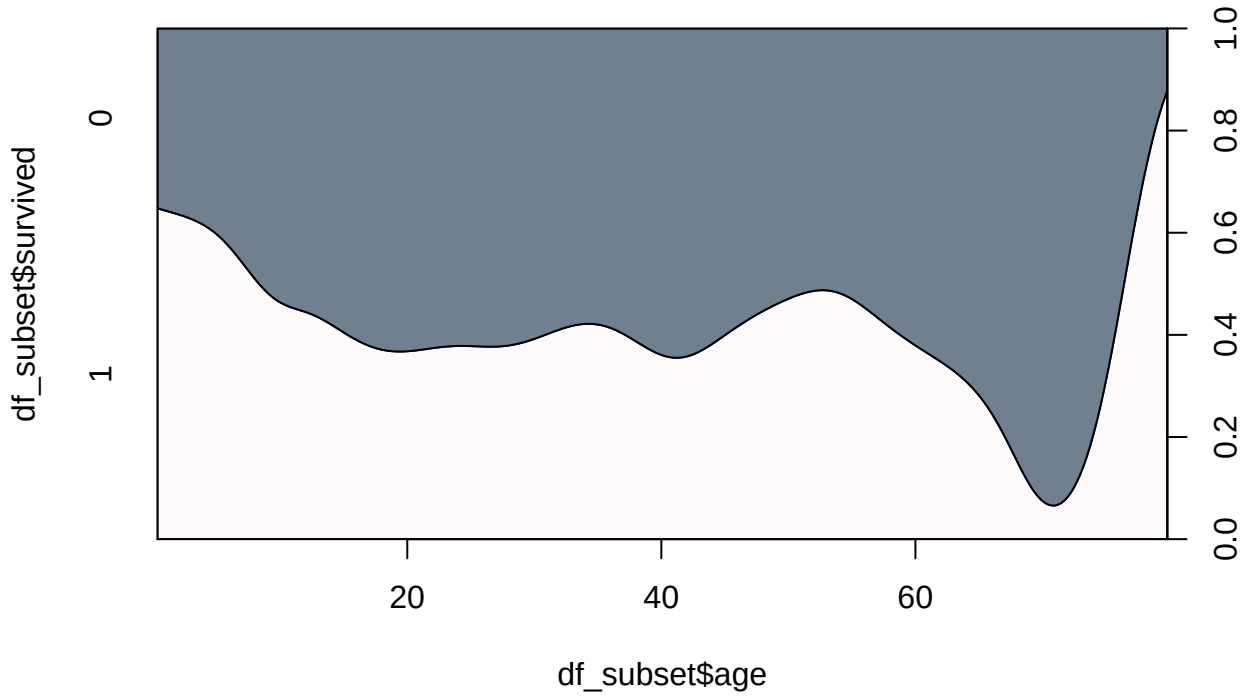
```
df_subset <- df[,c(1,2,5)]
df_subset <- df_subset[complete.cases(df_subset),]
x1 <- df_subset$age[df_subset$pclass==1]
x2 <- df_subset$age[df_subset$pclass==2]
x3 <- df_subset$age[df_subset$pclass==3]
vioplot(x1, x2, x3, col="wheat", names=c("Class 1", "Class 2", "Class"))
```



X is Quantitative, Y is Qualitative

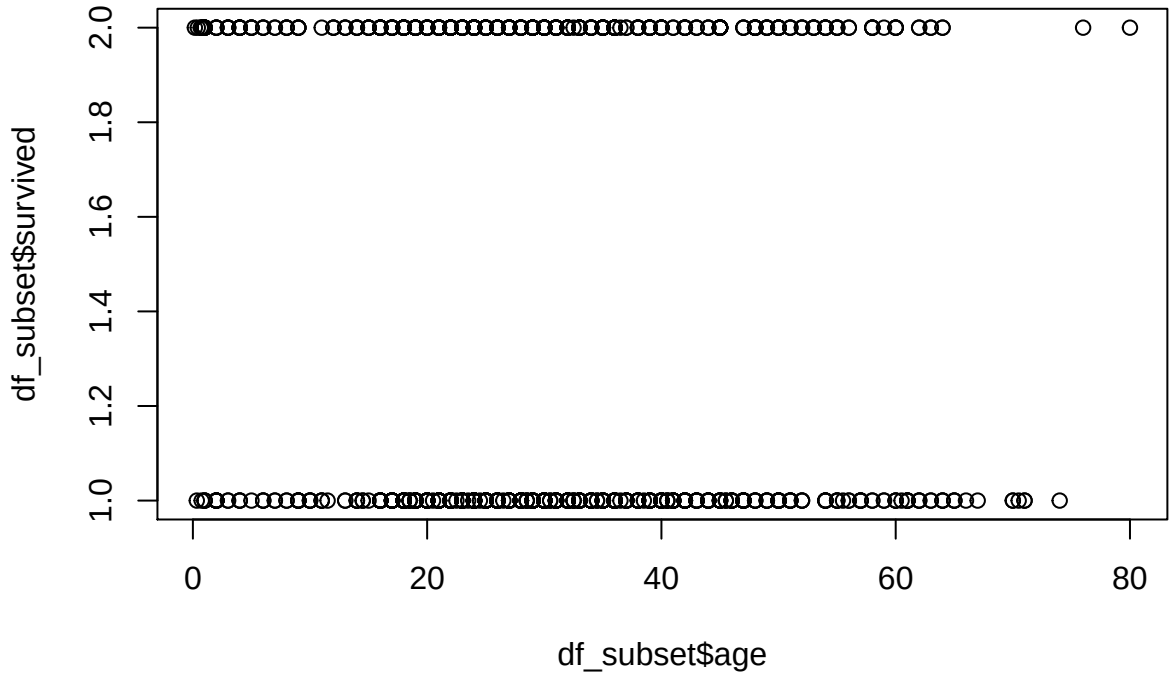
When X is quantitative and Y is qualitative, a conditional density plot can be used. The following plot shows how survived changes over the various ages.

```
cdplot(df_subset$age, df_subset$survived, col=c("snow", "slategray"))
```



The following is not informative because X is quantitative and Y is qualitative.

```
plot(df_subset$age, df_subset$survived)
```



X,Y both Quantitative

If X and Y are both quantitative, scatter plots are recommended. Here we have crosses for the points in blue, 75% of the usual size. We would have to dig further into the Titanic data to understand this chart. Why do so many passengers seem to have a fare of 0? And why did a few passengers pay 500? Perhaps the 500 fares paid for several people and the 0 fares reflect passengers whose fares were paid by a spouse or parent or adult

child? Further investigation is required to understand this.

```
plot(df$age, df$fare, pch='+', cex=0.75, col="blue", xlab="Age", ylab="Fare")
```