# Exploratory Data Analysis

**Upload the libraries:**

```
library(ggplot2)
library(patchwork)
library(psych)
```

**Upload the data:**

```
scales <- read.csv("scales_small_charlson.csv", header = FALSE,
    stringsAsFactors = FALSE)
lipids <- read.csv("lipids_small_charlson.csv", header = FALSE,
    stringsAsFactors = FALSE)
colnames(lipids) <- lipids[1, ]
lipids <- lipids[-1, ]
colnames(scales) <- scales[1, ]
scales <- scales[-1, ]
scales_lipids <- merge(scales, lipids, by = "MS ID")
```

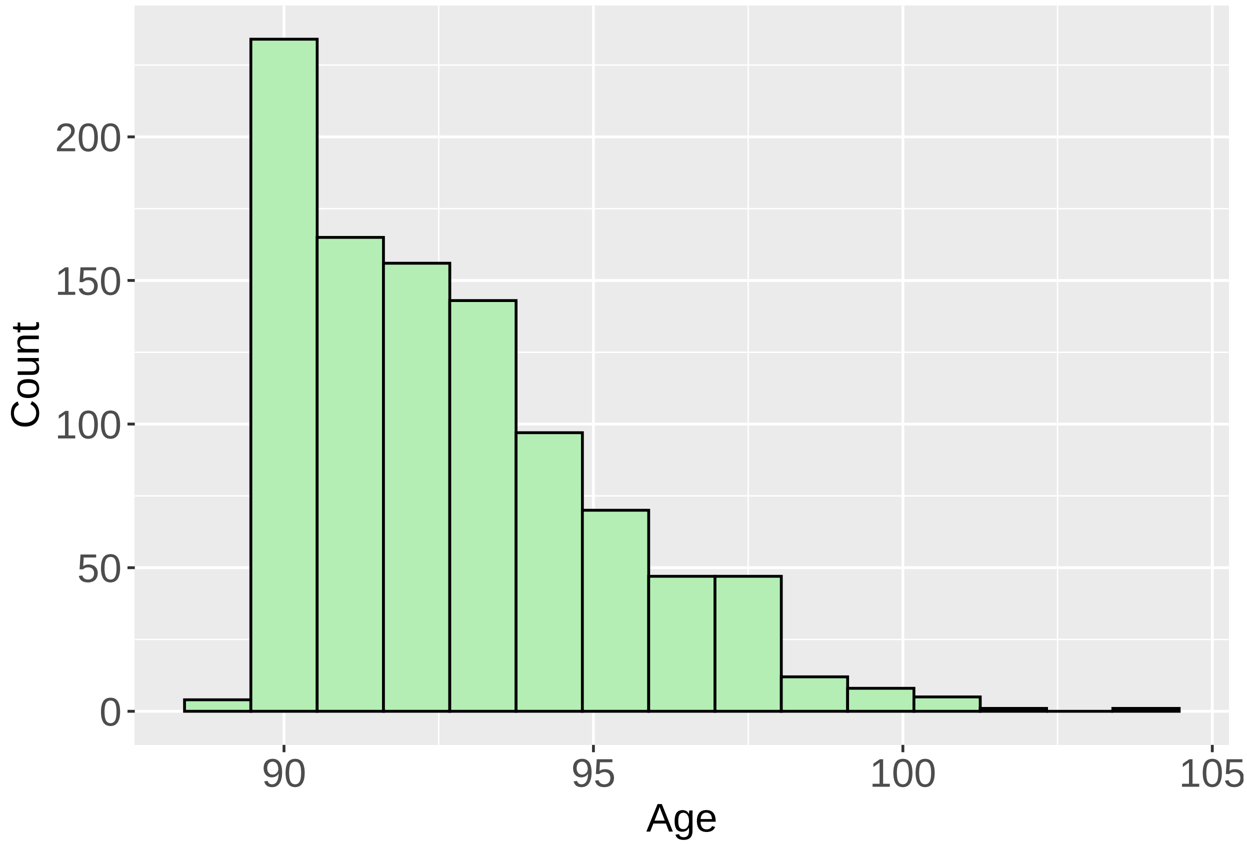
**Let's convert the data to the required type:**

```
columns_to_convert <- which(names(scales_lipids) != "sex" & names(scales_lipids) !=
    "MS ID")
scales_lipids[, columns_to_convert] <- lapply(scales_lipids[,
    columns_to_convert], as.numeric)
```

## Let's estimate the distribution of demographic factors in the data:

### Age distribution:

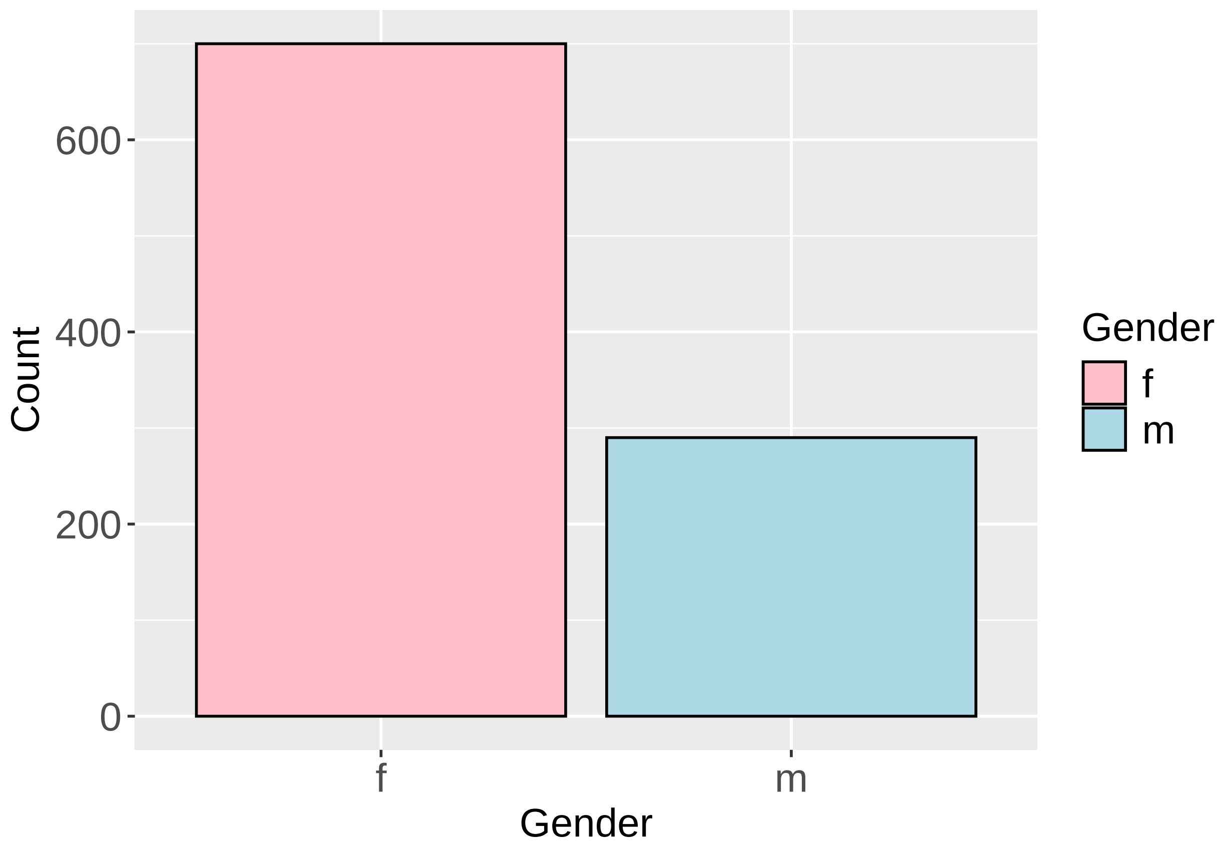
```
ggplot(data = scales_lipids, aes(x = age)) + geom_histogram(fill = "#B4EEB4",
    color = "black", bins = 15) + labs(x = "Age", y = "Count") +
    theme(axis.title.x = element_text(size = 15), axis.title.y = element_text(size = 15)) +
    theme(axis.text.x = element_text(size = 15), axis.text.y = element_text(size = 15))
```




**Most patients are between 90 and 95 years old.**

### Genders distribution:

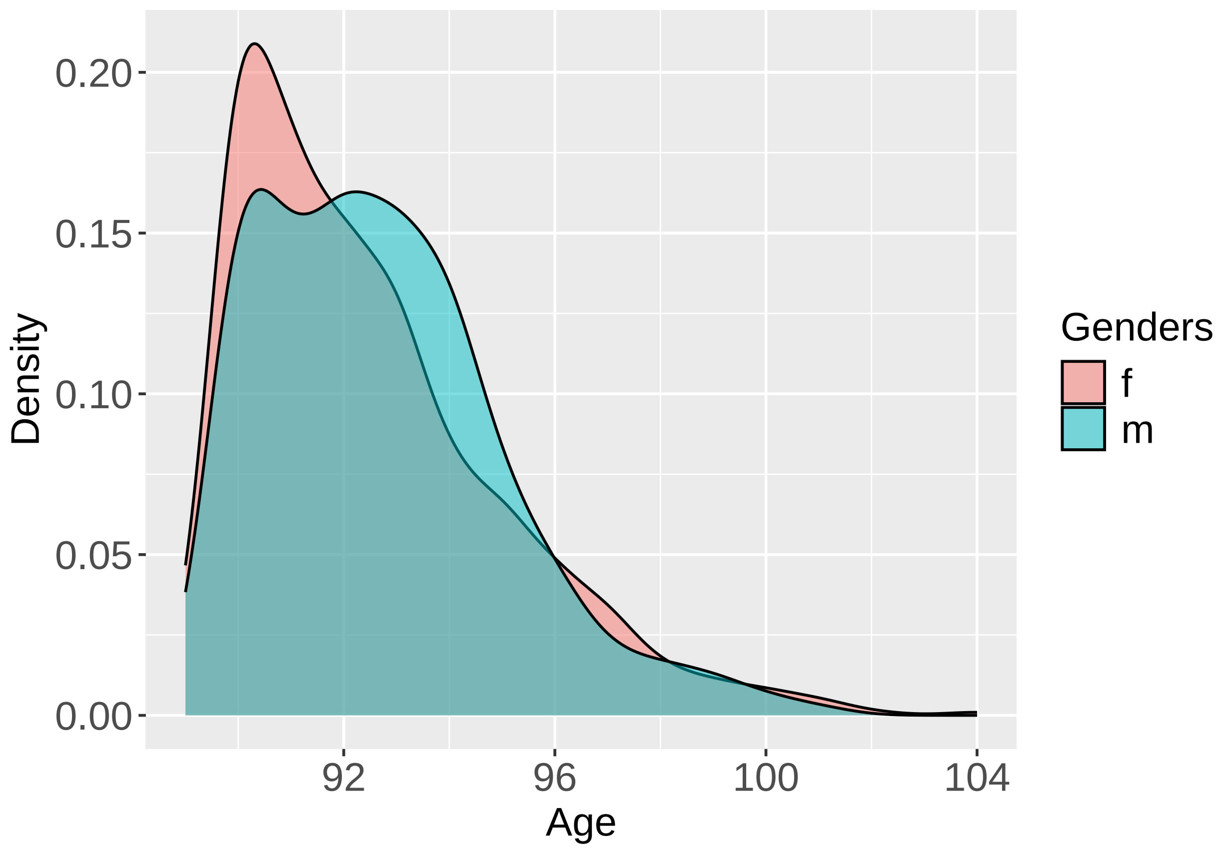
```
ggplot(scales_lipids, aes(x = factor(sex))) + geom_histogram(aes(fill = factor(sex)),
    color = "black", bins = 20, stat = "count") + scale_fill_manual(values = c(f = "pink",
    m = "lightblue")) + labs(x = "Gender", y = "Count", fill = "Gender") +
    theme(axis.title.x = element_text(size = 15), axis.title.y = element_text(size = 15),
        legend.text = element_text(size = 15), legend.title = element_text(size = 15)) +
    theme(axis.text.x = element_text(size = 15), axis.text.y = element_text(size = 15))
```




**Most patients are female.**

### Age distribution in different genders:

```
ggplot(data = scales_lipids, aes(x = age, fill = sex)) + geom_density(alpha = 0.5) +
    labs(x = "Age", y = "Density", fill = "Genders") + theme(axis.title.x = element_text(size = 15),
    axis.title.y = element_text(size = 15), legend.text = element_text(size = 15),
    legend.title = element_text(size = 15)) + theme(axis.text.x = element_text(size = 15),
    axis.text.y = element_text(size = 15))
```



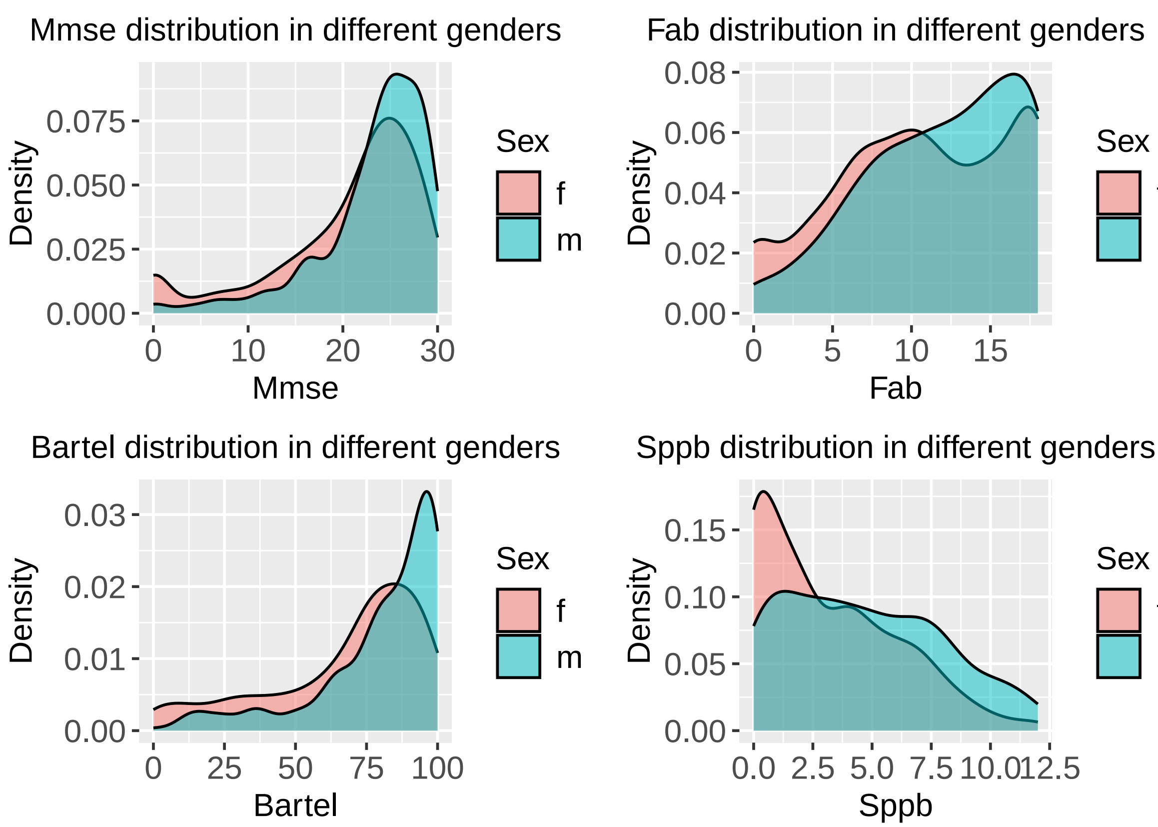

**There are differences in the distribution of age between the genders.**

## Let's estimate the distribution of scales in the data:

### Distribution of scales by gender:

```
p1 <- ggplot(data = scales_lipids, aes(x = mmse, fill = sex)) +
    geom_density(alpha = 0.5) + labs(x = "Mmse", y = "Density",
    fill = "Sex") + ggtitle("Mmse distribution in different genders") +
    theme(plot.title = element_text(hjust = 0.5, size = 12),
        axis.title.x = element_text(size = 12), axis.title.y = element_text(size = 12),
        legend.text = element_text(size = 12), legend.title = element_text(size = 12)) +
    theme(axis.text.x = element_text(size = 12), axis.text.y = element_text(size = 12))

p2 <- ggplot(data = scales_lipids, aes(x = fab, fill = sex)) +
    geom_density(alpha = 0.5) + labs(x = "Fab", y = "Density",
    fill = "Sex") + ggtitle("Fab distribution in different genders") +
    theme(plot.title = element_text(hjust = 0.5, size = 12),
        axis.title.x = element_text(size = 12), axis.title.y = element_text(size = 12),
        legend.text = element_text(size = 12), legend.title = element_text(size = 12)) +
    theme(axis.text.x = element_text(size = 12), axis.text.y = element_text(size = 12))

p3 <- ggplot(data = scales_lipids, aes(x = bartel, fill = sex)) +
```

```
    geom_density(alpha = 0.5) + labs(x = "Bartel", y = "Density",
    fill = "Sex") + ggtitle("Bartel distribution in different genders") +
    theme(plot.title = element_text(hjust = 0.5, size = 12),
        axis.title.x = element_text(size = 12), axis.title.y = element_text(size = 12),
        legend.text = element_text(size = 12), legend.title = element_text(size = 12)) +
    theme(axis.text.x = element_text(size = 12), axis.text.y = element_text(size = 12))

p4 <- ggplot(data = scales_lipids, aes(x = sppb, fill = sex)) +
    geom_density(alpha = 0.5) + labs(x = "Sppb", y = "Density",
    fill = "Sex") + ggtitle("Sppb distribution in different genders") +
    theme(plot.title = element_text(hjust = 0.5, size = 12),
        axis.title.x = element_text(size = 12), axis.title.y = element_text(size = 12),
        legend.text = element_text(size = 12), legend.title = element_text(size = 12)) +
    theme(axis.text.x = element_text(size = 12), axis.text.y = element_text(size = 12))

(p1 + p2)/(p3 + p4)
```

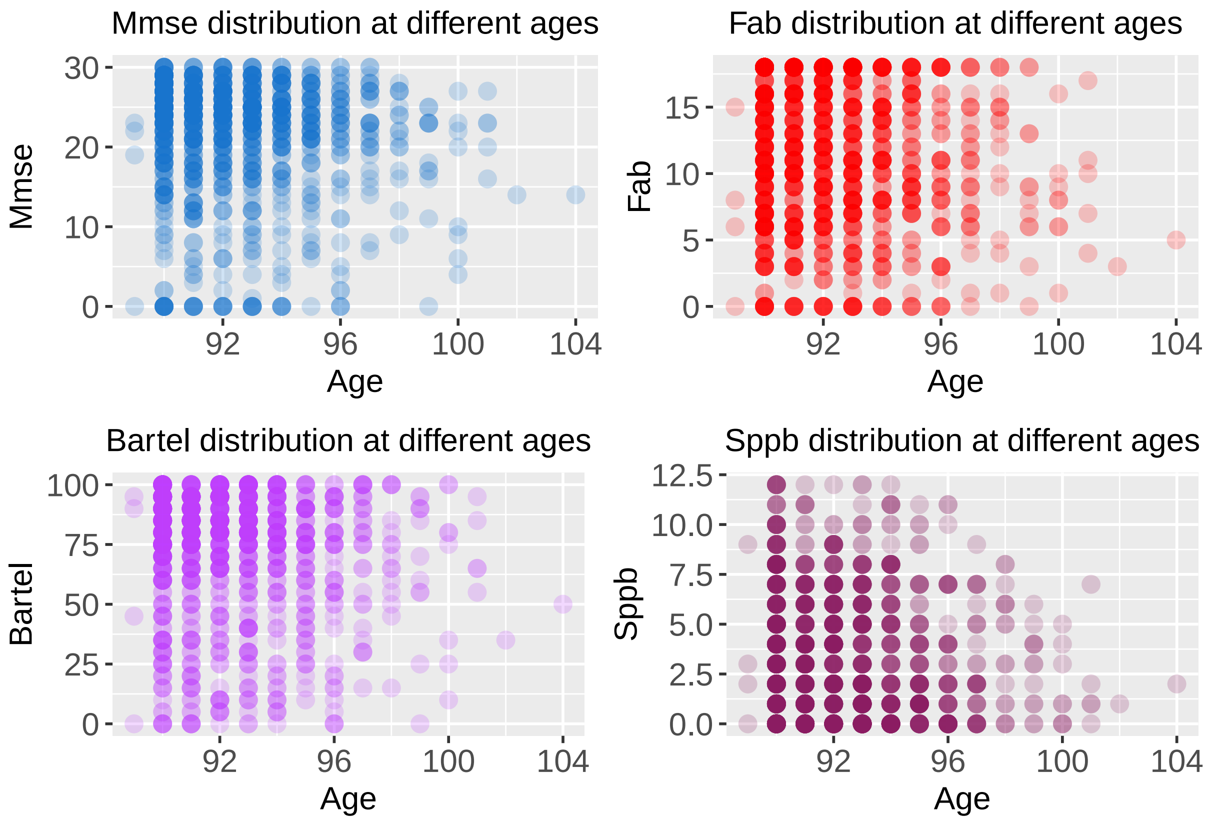



**The distribution of the scales differs between genders, especially in the case of the sppb scale.**

### Distribution of scales by age:

```
p5 <- ggplot(scales_lipids, aes(x = age, y = mmse)) + geom_point(shape = 16,
    size = 3, color = adjustcolor("#1874CD"), alpha = 0.2) +
```

```r
    labs(title = "Mmse distribution at different ages", x = "Age",
        y = "Mmse") + theme(plot.title = element_text(hjust = 0.5,
    size = 12), axis.title.x = element_text(size = 12), axis.title.y = element_text(size = 12),
    legend.text = element_text(size = 12), legend.title = element_text(size = 12)) +
    theme(axis.text.x = element_text(size = 12), axis.text.y = element_text(size = 12))

p6 <- ggplot(scales_lipids, aes(x = age)) + geom_point(aes(y = fab),
    shape = 16, size = 3, color = adjustcolor("Red"), alpha = 0.2) +
    labs(title = "Fab distribution at different ages", x = "Age",
        y = "Fab") + theme(plot.title = element_text(hjust = 0.5,
    size = 12), axis.title.x = element_text(size = 12), axis.title.y = element_text(size = 12),
    legend.text = element_text(size = 12), legend.title = element_text(size = 12)) +
    theme(axis.text.x = element_text(size = 12), axis.text.y = element_text(size = 12))

p7 <- ggplot(scales_lipids, aes(x = age)) + geom_point(aes(y = bartel),
    shape = 16, size = 3, color = adjustcolor("#BF3EFF"), alpha = 0.2) +
    labs(title = "Bartel distribution at different ages", x = "Age",
        y = "Bartel") + theme(plot.title = element_text(hjust = 0.5,
    size = 12), axis.title.x = element_text(size = 12), axis.title.y = element_text(size = 12),
    legend.text = element_text(size = 12), legend.title = element_text(size = 12)) +
    theme(axis.text.x = element_text(size = 12), axis.text.y = element_text(size = 12))

p8 <- ggplot(scales_lipids, aes(x = age)) + geom_point(aes(y = sppb),
    shape = 16, size = 3, color = adjustcolor("#8B1C62"), alpha = 0.2) +
    labs(title = "Sppb distribution at different ages", x = "Age",
        y = "Sppb") + theme(plot.title = element_text(hjust = 0.5,
    size = 12), axis.title.x = element_text(size = 12), axis.title.y = element_text(size = 12),
    legend.text = element_text(size = 12), legend.title = element_text(size = 12)) +
    theme(axis.text.x = element_text(size = 12), axis.text.y = element_text(size = 12))

(p5 + p6)/(p7 + p8)
```



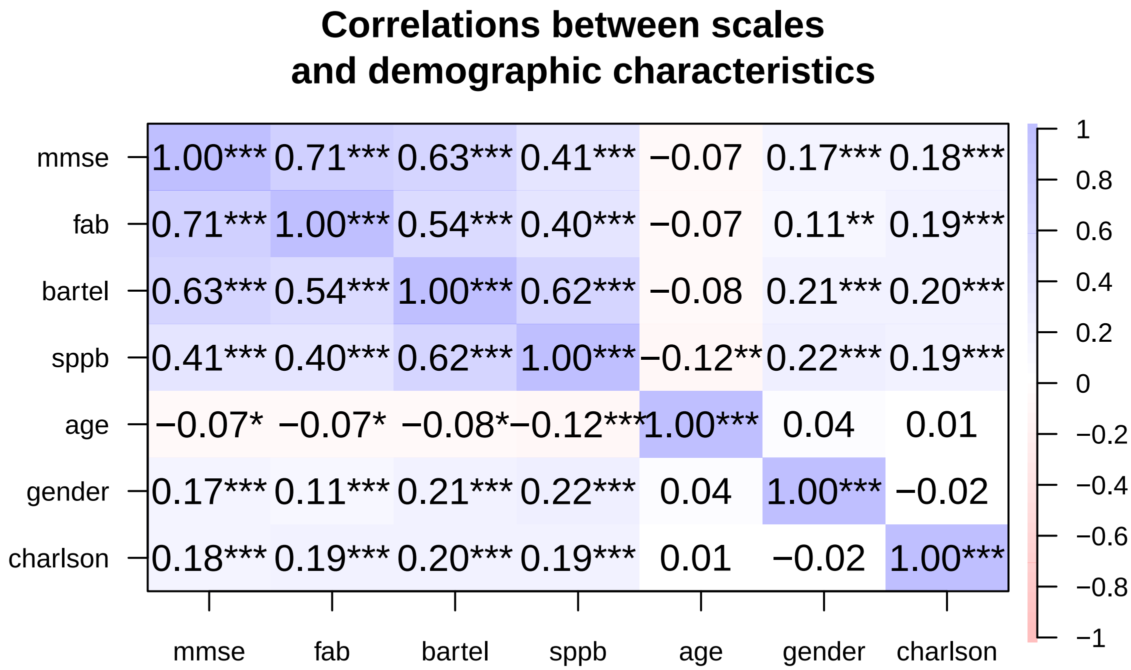

## Let's evaluate the correlation between scales and demographic factors:

#### Let's convert gender data to a numeric type:

```
names(scales_lipids)[names(scales_lipids) == "sex"] <- "gender"
names(scales_lipids)[names(scales_lipids) == "(-1)Charlson"] <- "charlson"
scales_lipids$gender <- ifelse(scales_lipids$gender == "f", 0,
    1)
```

### Let's draw a correlation plot:

```
corPlot(scales_lipids[, c("mmse", "fab", "bartel", "sppb", "age",
    "gender", "charlson")], cex = 1.2, stars = TRUE, alpha = 0.25,
    cex.axis = 0.8, main = "Correlations between scales \n and demographic characteristics",
    cex.main = 1)
```




**The scales are highly correlated with each other, but not with demographic factors.**