Part 1 - Data visualization and descriptive statistics

Ricardo Martins Ferreira

2024-03-11

Introduction

This workflow presents basic tools for data visualization and descriptive statistics based on the ggplot2 package.

```
setwd("C:/Users/ricar/Documents/Workshop_2024/")

getwd()
```

```
## [1] "C:/Users/ricar/Documents/Workshop_2024"
```

Install Bioconductor and packages

Bioconductor is an R project that provides software for the analysis of genomic data.

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install(version = "3.18")
```

Now, install packages using Bioconductor. - ggplot2 - readxl...

```
#BiocManager::install(c("ggplot2", "readxl"))

# load one by one
library(ggplot2)

#load multiple
libs <- c("ggplot2","readxl")
suppressMessages(
  suppressWarnings(sapply(libs, require, character.only = TRUE))
)
```

```
## ggplot2  readxl
##    TRUE    TRUE
```

Load files from location

```r
df <- read_excel("data/coldata_part1.xlsx",sheet = "coldata")
df <- read.csv("data/coldata_part1.csv",sep=";")
df <- read.table("data/coldata_part1.txt", header = TRUE)
```

Subset data tables

```r
#columns
df_sub <- df[,c(1,2,3)]

#rows
df_sub <- df[c(1,2,3),]

#based on categorical variables
df_sub <- df[df$Group=="A",]
df_sub <- df[df$Group %in% c("A","B"),]
df_sub <- df[!df$Group == "C",]

#based on numeric variables
df_sub <- df[df$Age>50,]
df_sub <- df[df$Age<50,]
df_sub <- df[df$Age>=50,]
df_sub <- df[df$Age<=50,]

#based on two combined variables
df_sub <- df[df$Group=="A" & df$Age>50,]

#subset a variable based on other
df_sub <- df$Sample[df$Sex=="M"]

#exclude based on more than one variable
df_sub <- df[!(df$Group=="A" & df$Sex == "M"),]

#subset based on a vector of samples
rownames(df)<- df$Sample
samples <- c("C1","A4","B5")
df_sub <- df[samples,]
```

Exercise 1

1.1. Create a data table from df with people over 40 years old, excluding females from Group A and B.

1.2. Identify the female individuals that are not in group C.

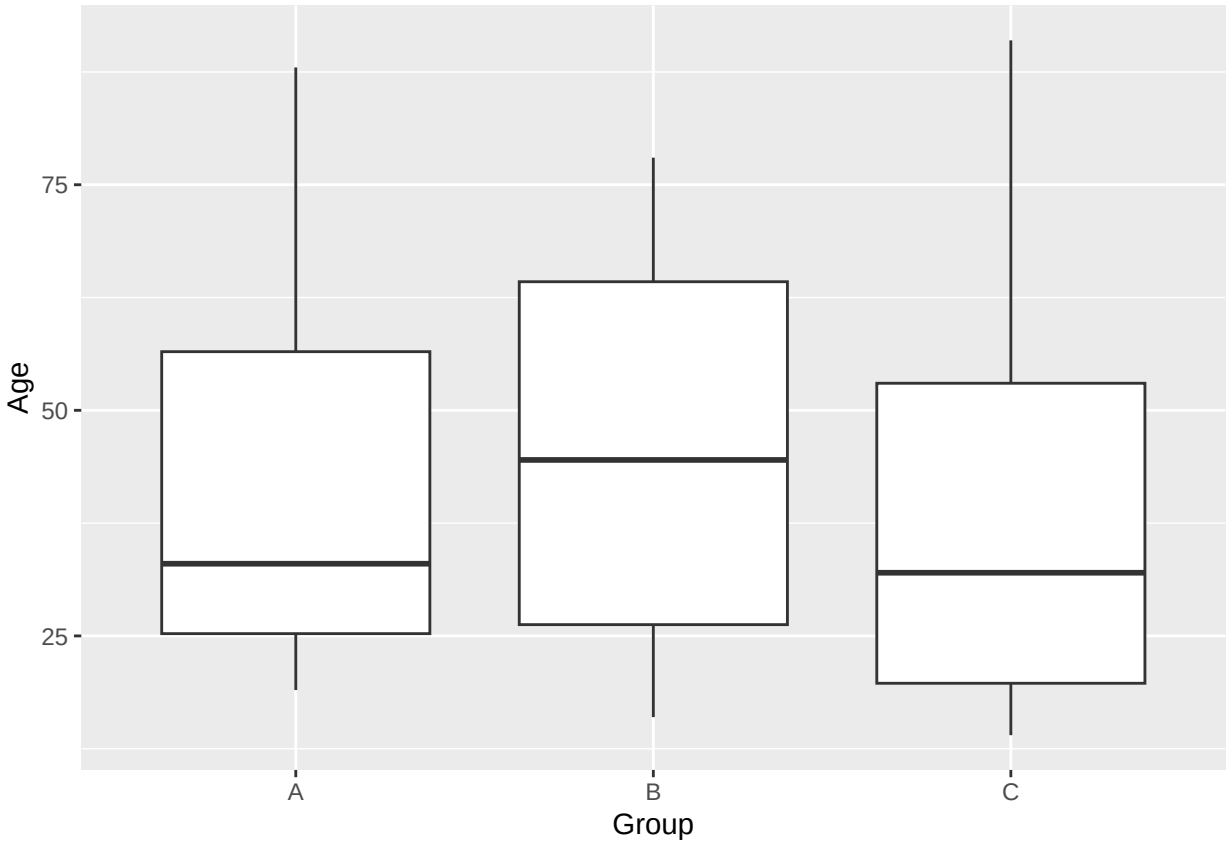
Visual representation with ggplot2

For detailed exploration of the ggplot2 tool, check https://ggplot2.tidyverse.org/.

Representation of numeric variable (Age)

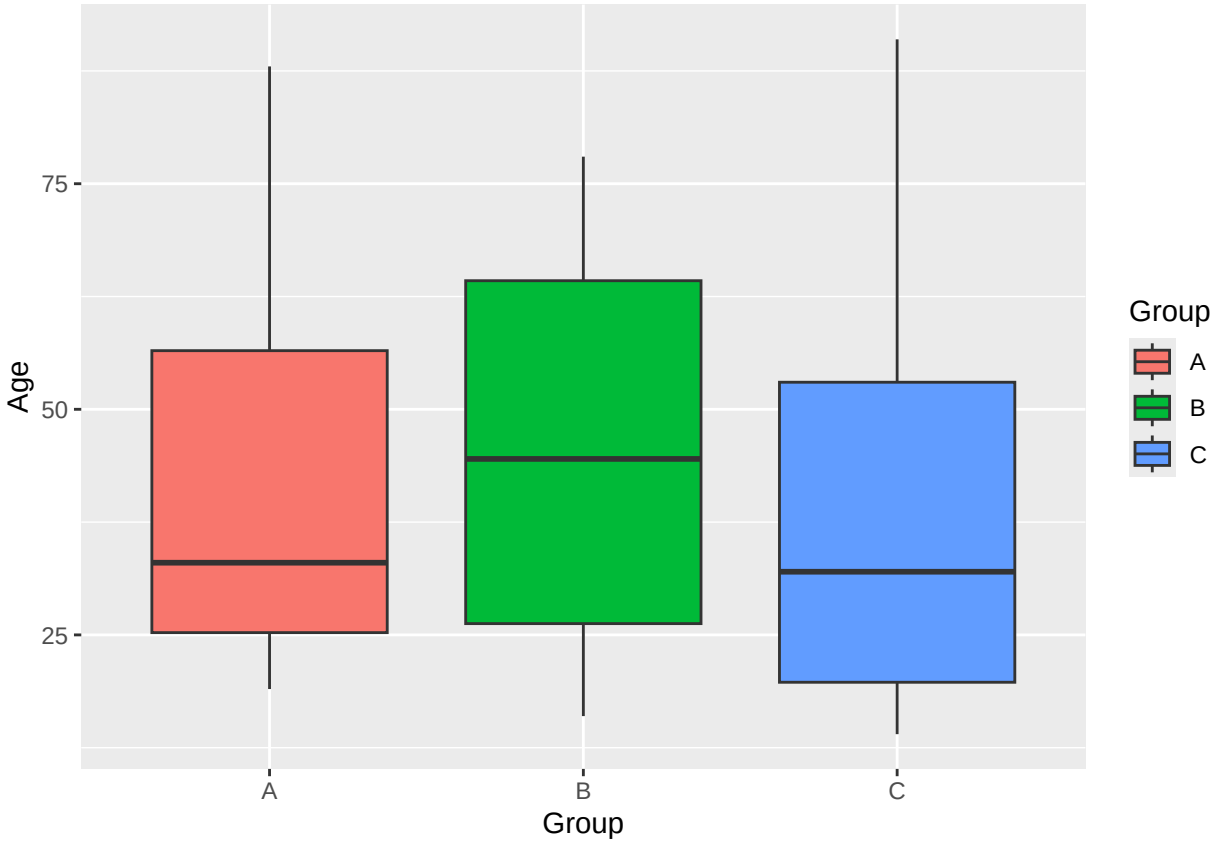
Box plot (box and whiskers). Description: The boxplot compactly displays the distribution of a continuous variable. It visualises five summary statistics (the median, two hinges and two whiskers), and all "outlying" points individually.The middle line represents the median, the hinges the first and third quartiles (25th and 75th percentiles), and the whiskers the max and minimun values.

```
df$Age <- as.numeric(df$Age)

ggplot(df, aes(x = Group,y =Age))+
  geom_boxplot()
```



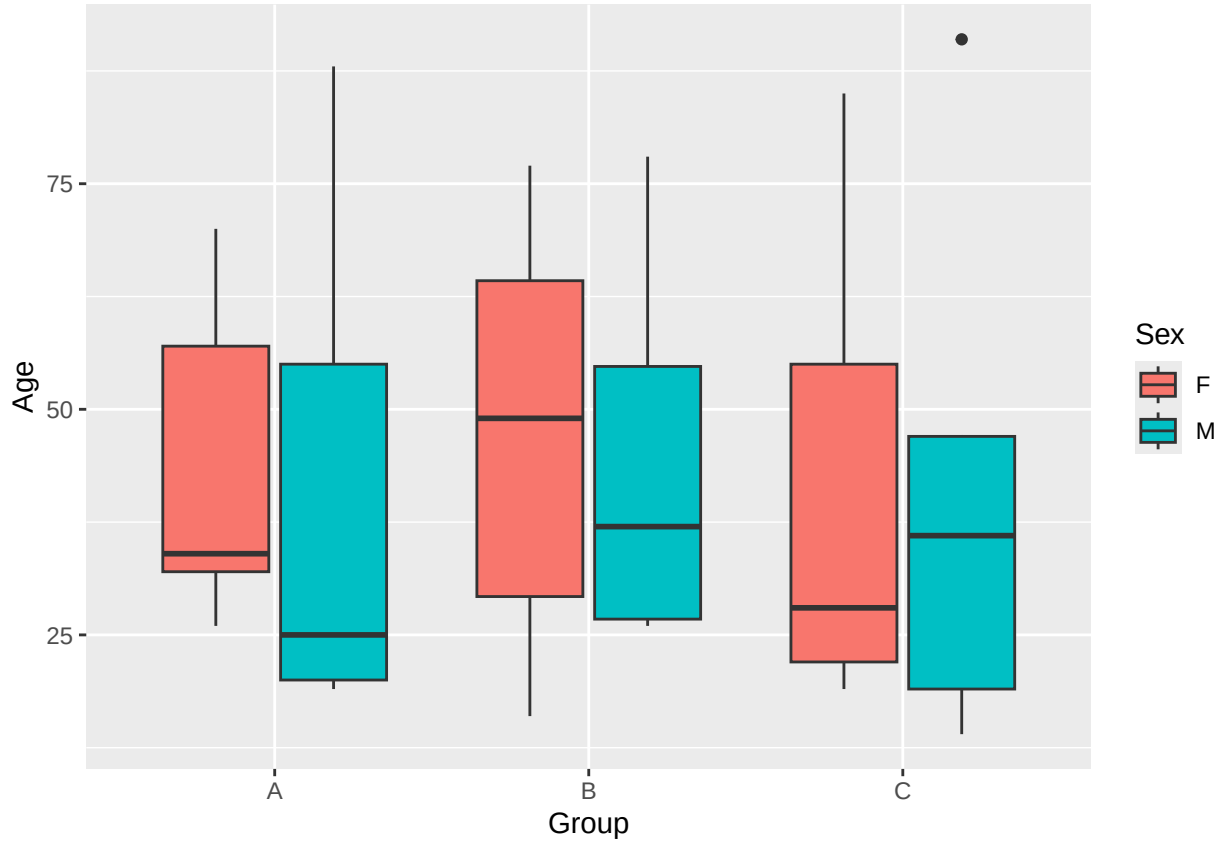
Aesthetics Color the boxes by Group.

```
ggplot(df, aes(x = Group,y =Age, fill=Group))+
  geom_boxplot()
```



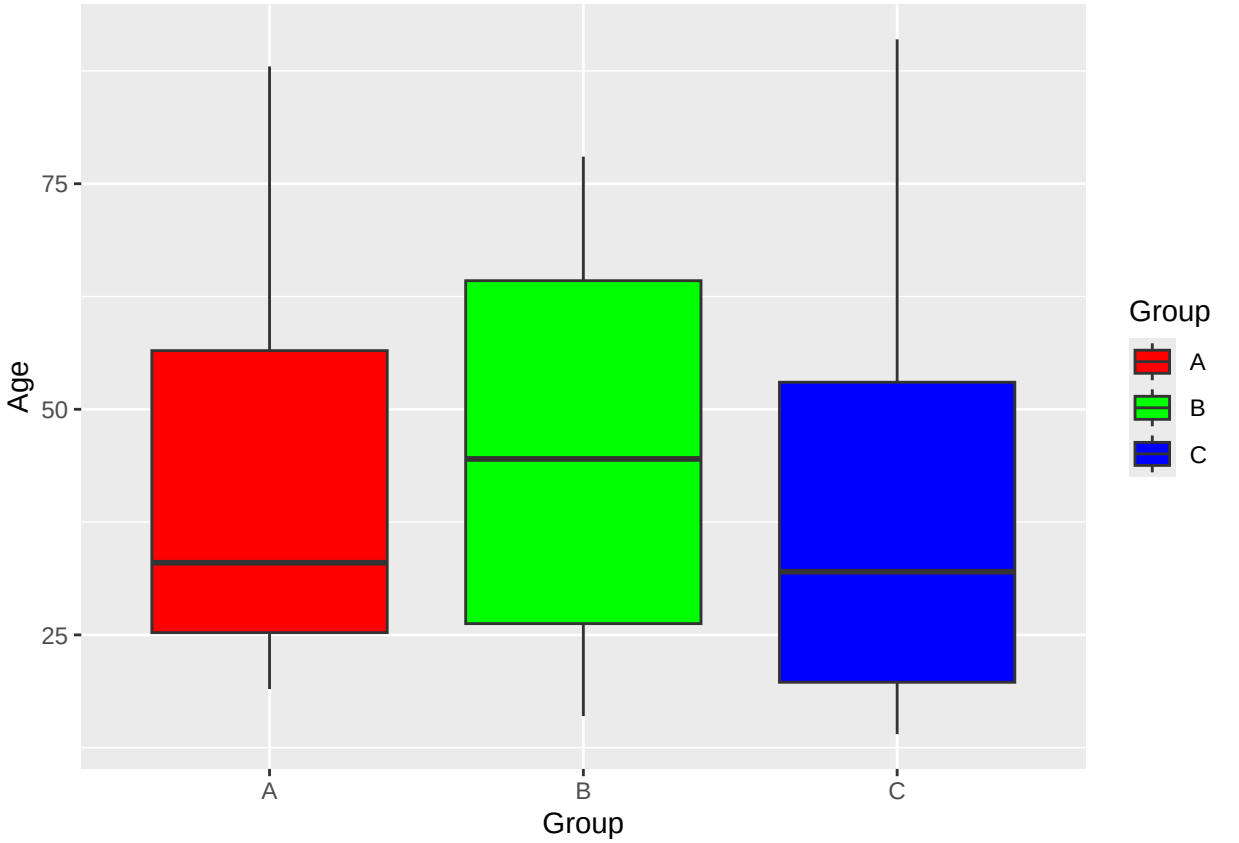
Group the boxes by Sex.

```
ggplot(df, aes(x = Group,y =Age,fill=Sex, Group=Sex))+
  geom_boxplot()
```



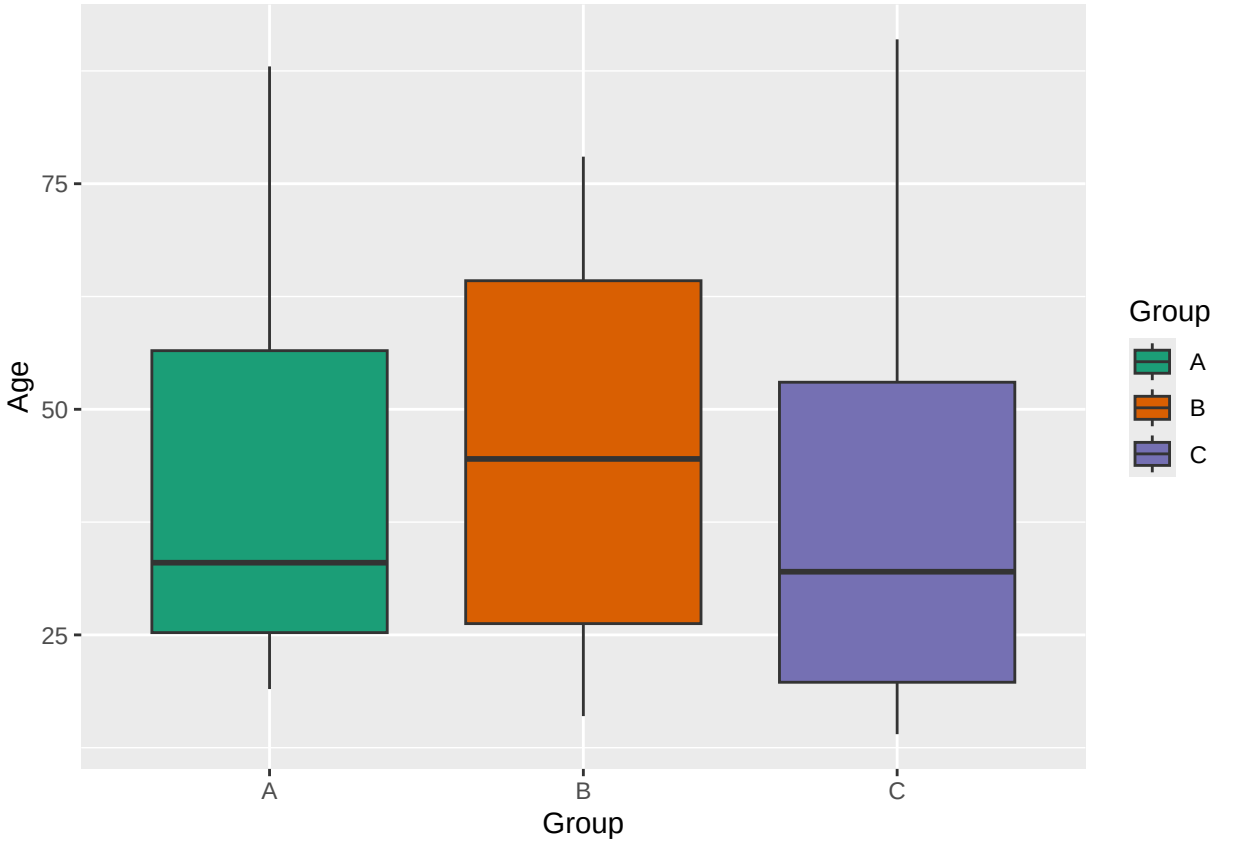
Change the colors manually

```
ggplot(df, aes(x = Group,y =Age, fill=Group))+
  geom_boxplot()+
  scale_fill_manual(values = c("red","green","blue"))
```



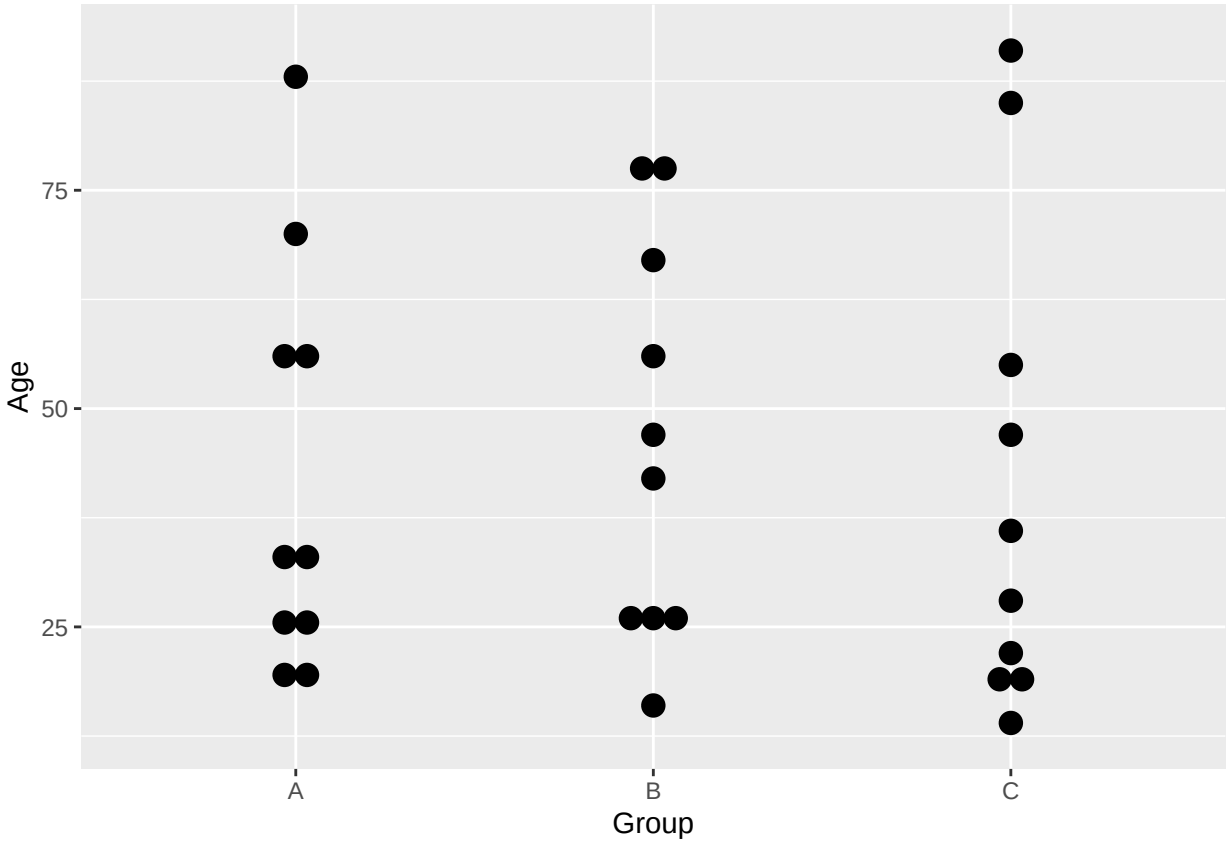
Change the colors based on existing palettes (https://r-graph-gallery.com/38-rcolorbrewers-palettes.html).

```
ggplot(df, aes(x = Group,y =Age, fill=Group))+
  geom_boxplot()+
  scale_fill_brewer(palette = "Dark2")
```



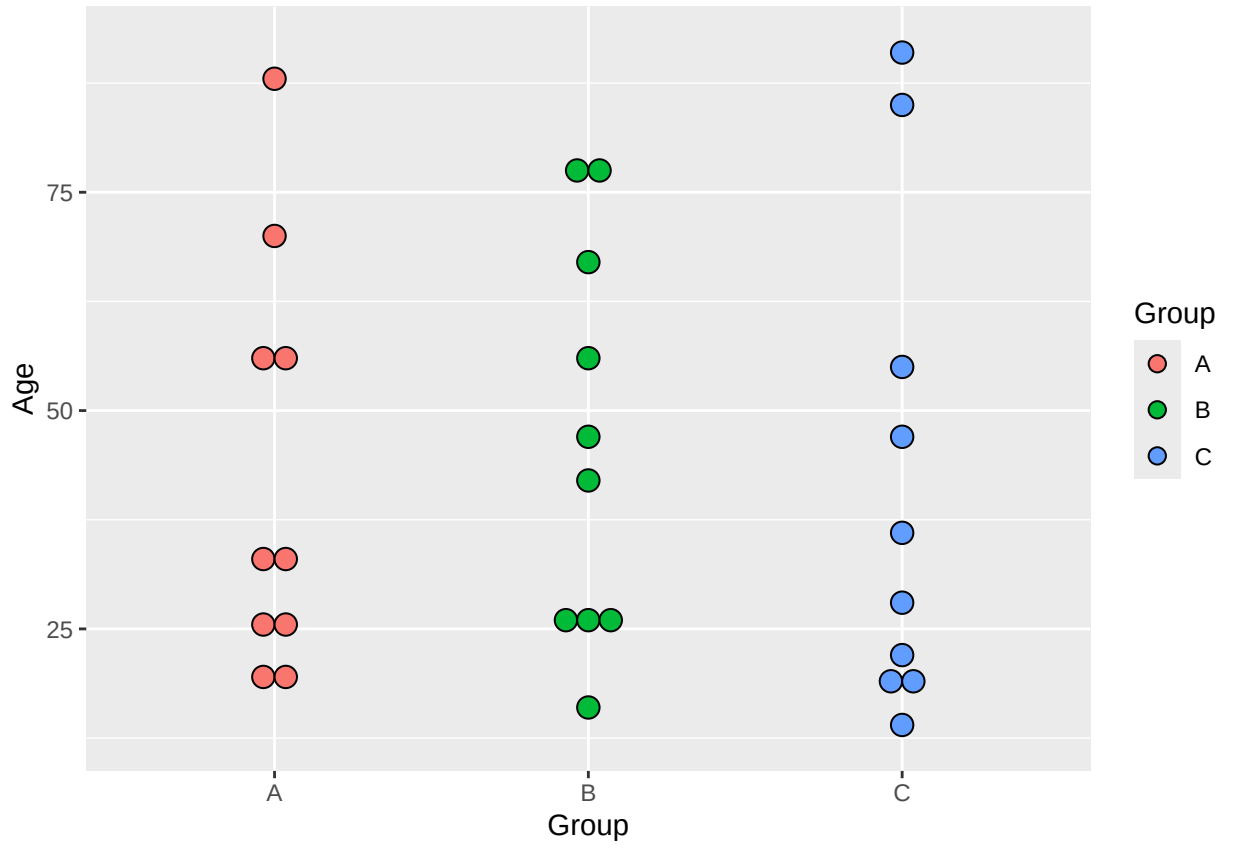
Dot plot. Description: In a dot plot, the width of a dot corresponds to the bin width (or maximum width, depending on the binning algorithm), and dots are stacked, with each dot representing one observation.

```
ggplot(df, aes(x = Group, y = Age)) +
  geom_dotplot(binaxis = "y", stackdir = "center")
```



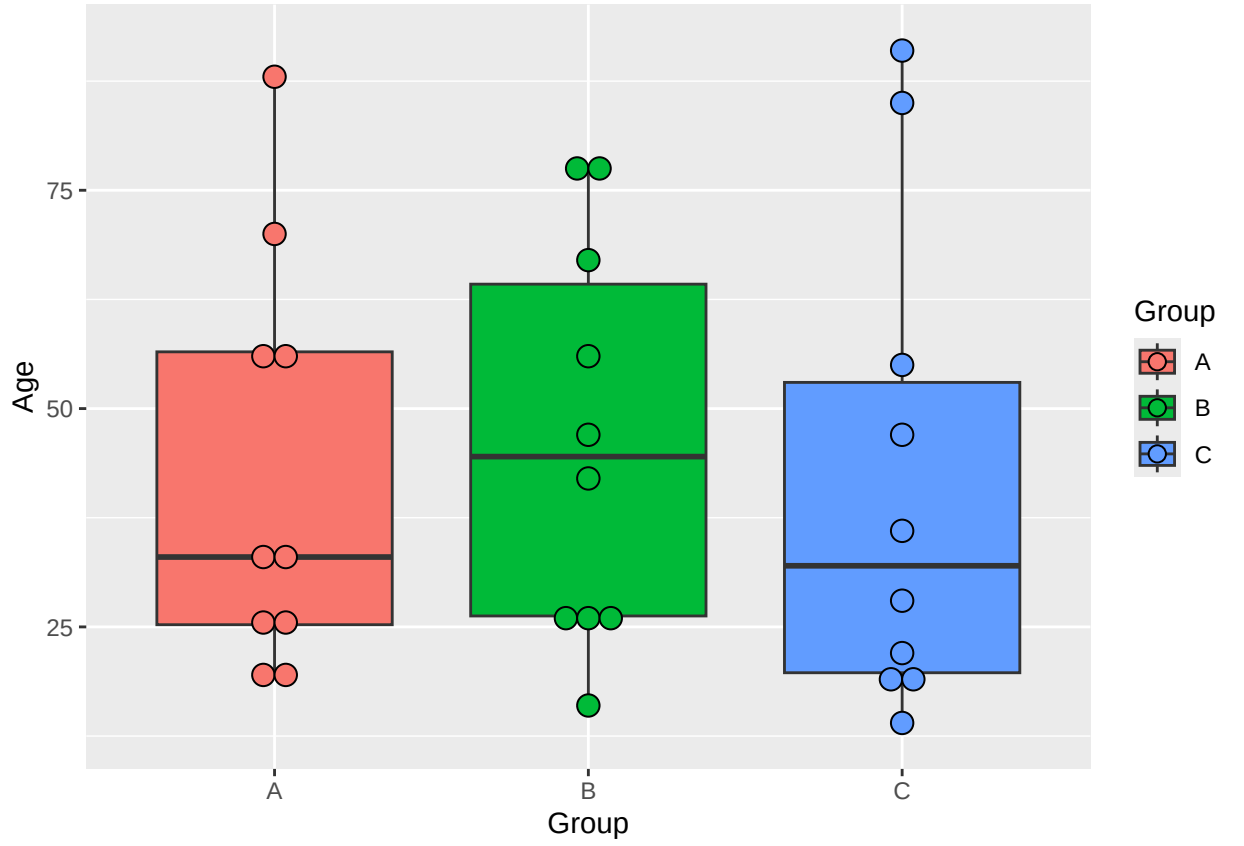
Aesthetics Color dot plots.

```
ggplot(df, aes(x = Group, y = Age, fill=Group)) +
  geom_dotplot(binaxis = "y", stackdir = "center")
```



####Combine box plot and dot plot

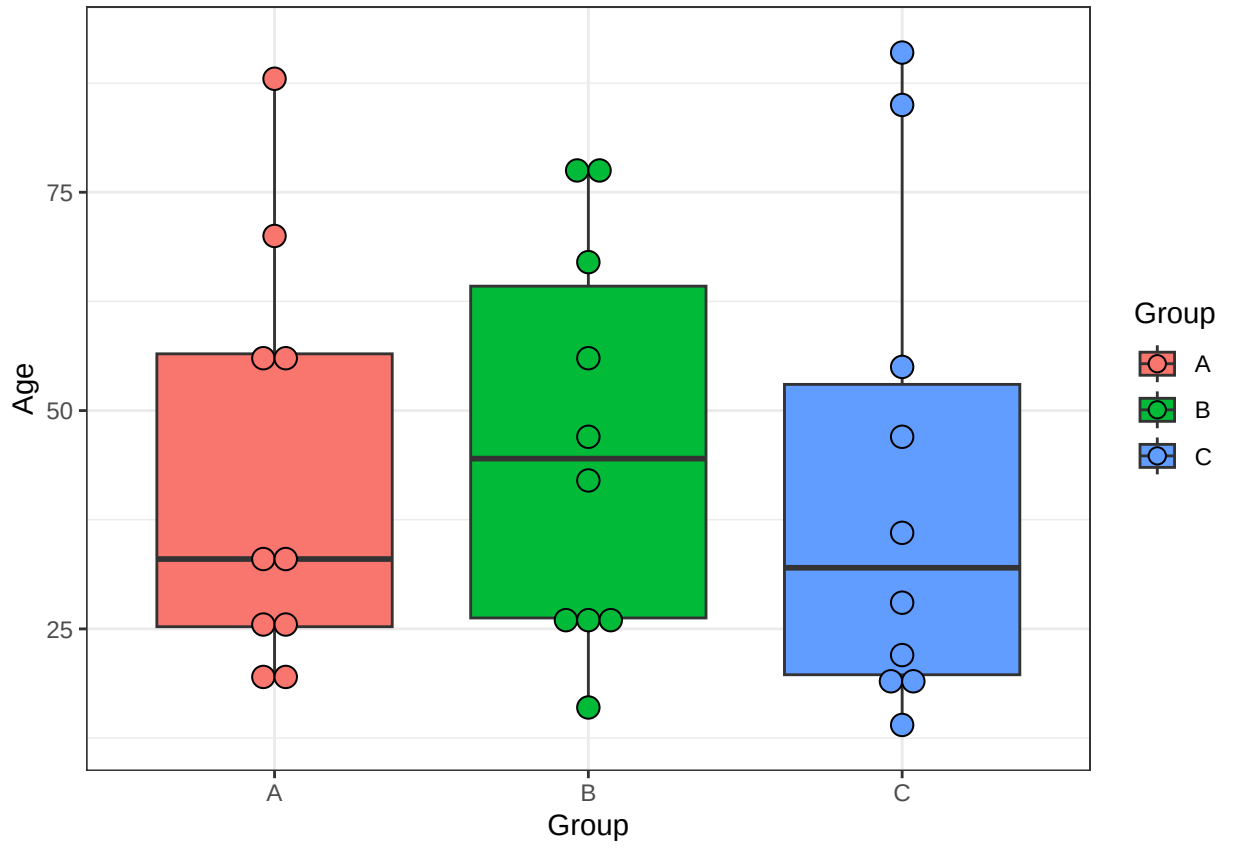
```
ggplot(df, aes(x = Group, y = Age, fill=Group)) +
  geom_boxplot()+
  geom_dotplot(binaxis = "y", stackdir = "center")
```



#####Additional aesthetics

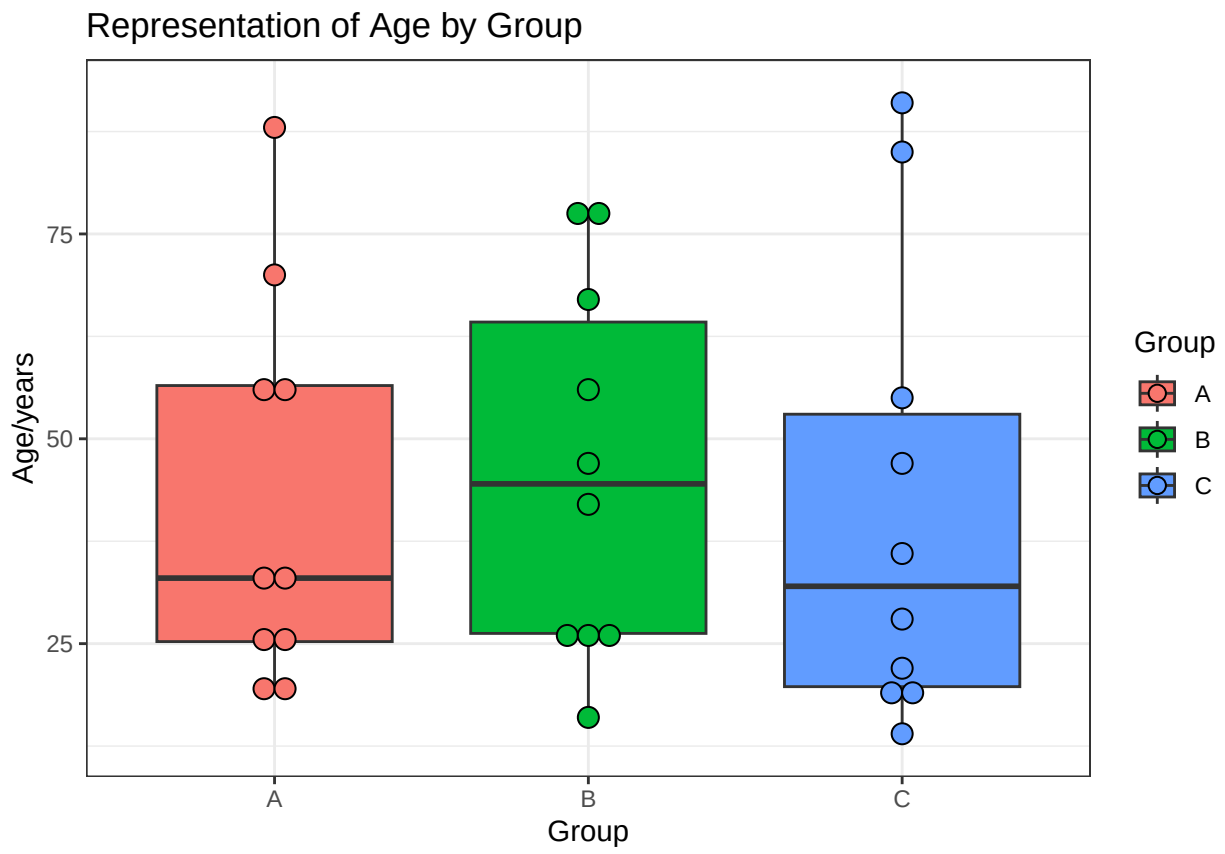
Edit non-data display.

```
ggplot(df, aes(x = Group, y = Age, fill=Group)) +
  geom_boxplot()+
  geom_dotplot(binaxis = "y", stackdir = "center")+
  theme_bw()
```




```
  #+theme_minimal()
  #+theme_classic()
  #+theme_linedraw()
  #+theme_dark()
```

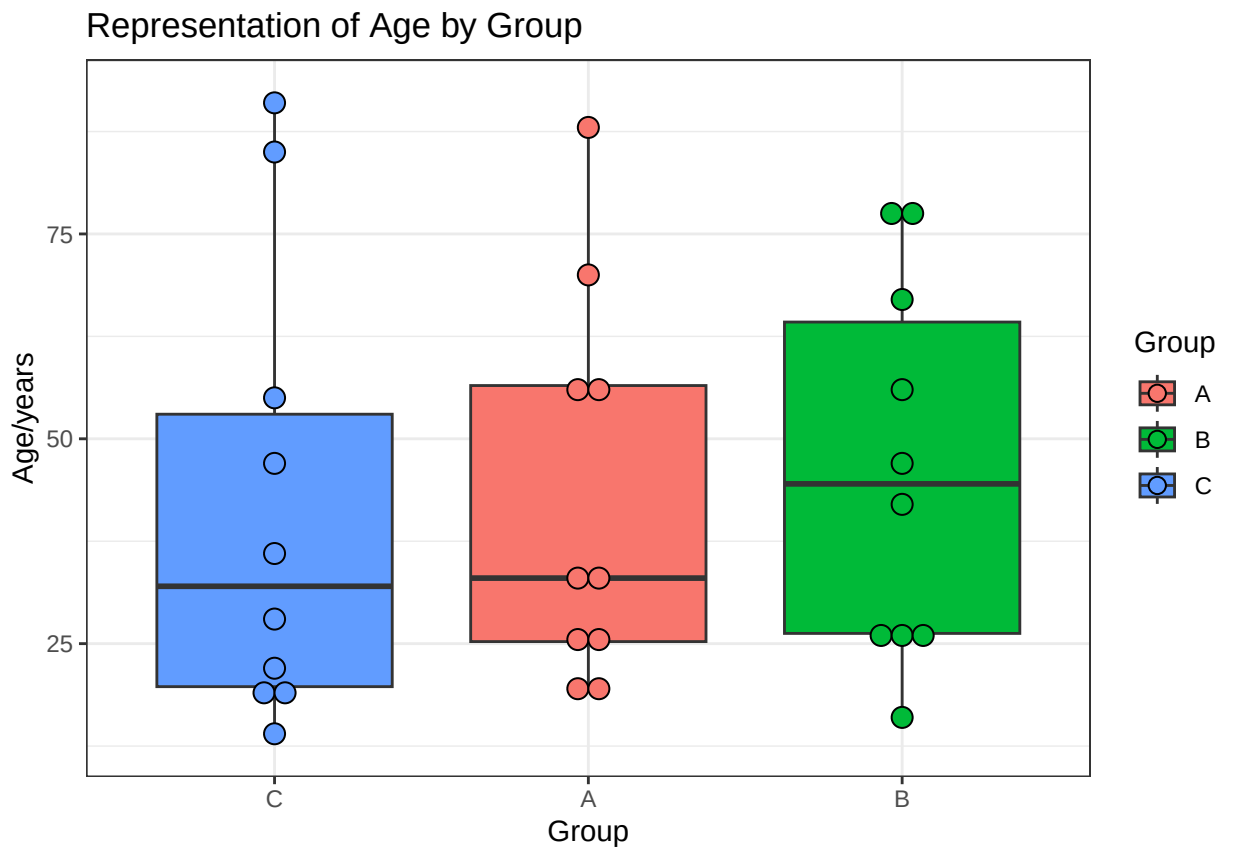
```
#add labels
ggplot(df, aes(x = Group, y = Age, fill=Group)) +
  geom_boxplot()+
  geom_dotplot(binaxis = "y", stackdir = "center")+
  labs(title="Representation of Age by Group", y = "Age/years")+
  theme_bw()
```




```
  #+theme_minimal()
  #+theme_classic()
  #+theme_linedraw()
  #+theme_dark()
```

Change x order.

```
#add labels
ggplot(df, aes(x = Group, y = Age, fill=Group)) +
  geom_boxplot()+
  scale_x_discrete(limits=c("C","A","B"))+
  geom_dotplot(binaxis = "y", stackdir = "center")+
  labs(title="Representation of Age by Group", y = "Age/years")+
  theme_bw()
```




```
  #+theme_minimal()
  #+theme_classic()
  #+theme_linedraw()
  #+theme_dark()
```

Statistical tests To calculate the statistical significance of the differences of continuous variables, the following tests can be used:

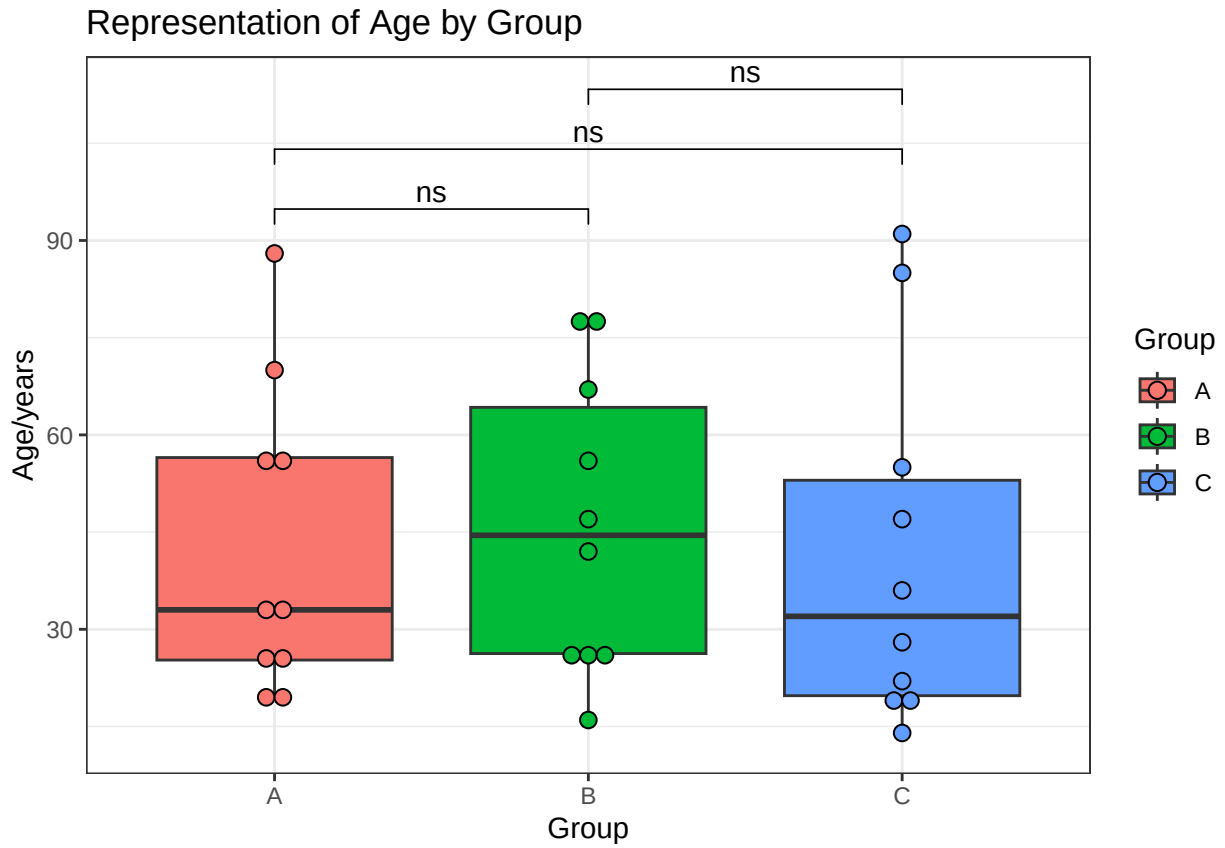
Between two groups:

t-test (independent or paired); Non-parametric tests (Mann-Whitney U test, independent; Wilcoxon Signed-RanK test, paired)

Between two or more groups: One-Way or Two-Way ANOVA (independent): Kruskal-Wallis test (paired)

```
#BiocManager::install("ggpubr")

library(ggpubr)

#establish comparisons
my_comparisons <- list(c("A","B"),c("A","C"),c("C","B"))

ggplot(df, aes(x = Group, y = Age, fill=Group)) +
  geom_boxplot()+
  geom_dotplot(binaxis = "y", stackdir = "center")+
  labs(title="Representation of Age by Group", y = "Age/years")+
```

```
  stat_compare_means(data = df,method = "t.test", paired=FALSE,comparisons = my_comparisons,label="p.sig
  theme_bw()
```

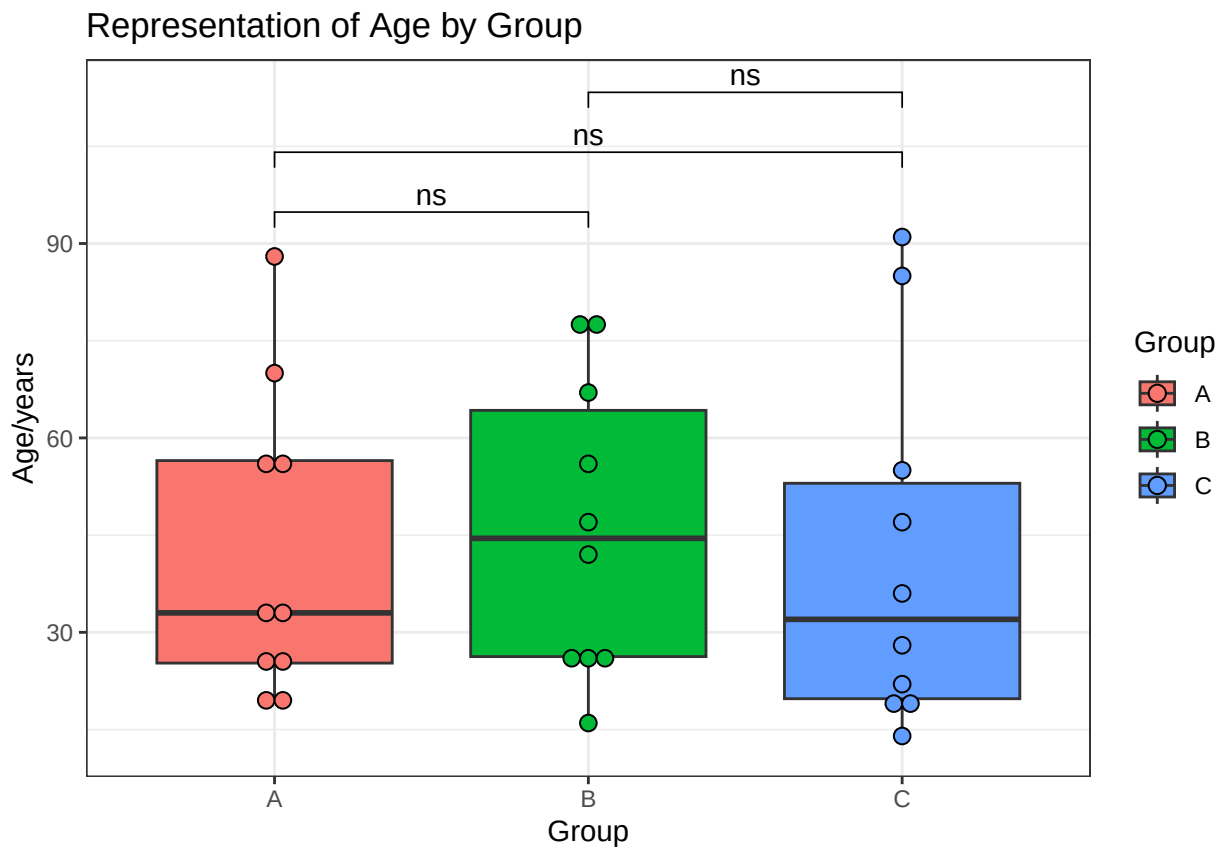



```
ggplot(df, aes(x = Group, y = Age, fill=Group)) +
  geom_boxplot()+
  geom_dotplot(binaxis = "y", stackdir = "center")+
  labs(title="Representation of Age by Group", y = "Age/years")+
  stat_compare_means(data = df,method = "wilcox.test", paired=FALSE,comparisons = my_comparisons,label=
  theme_bw()
```

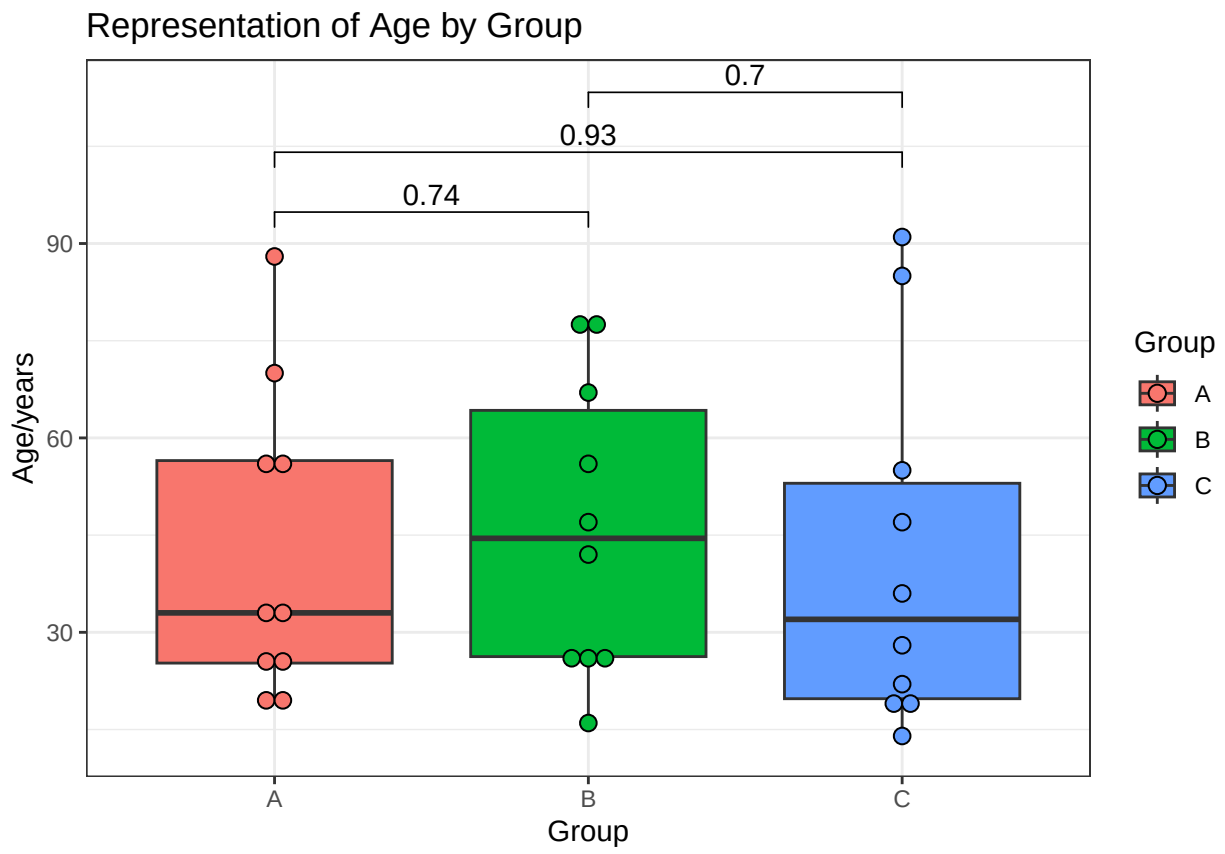
```
## Warning in wilcox.test.default(c(20, 55, 34, 25, 57, 88, 26, 32, 19, 70), :
## cannot compute exact p-value with ties
```

```
## Warning in wilcox.test.default(c(20, 55, 34, 25, 57, 88, 26, 32, 19, 70), :
## cannot compute exact p-value with ties
```

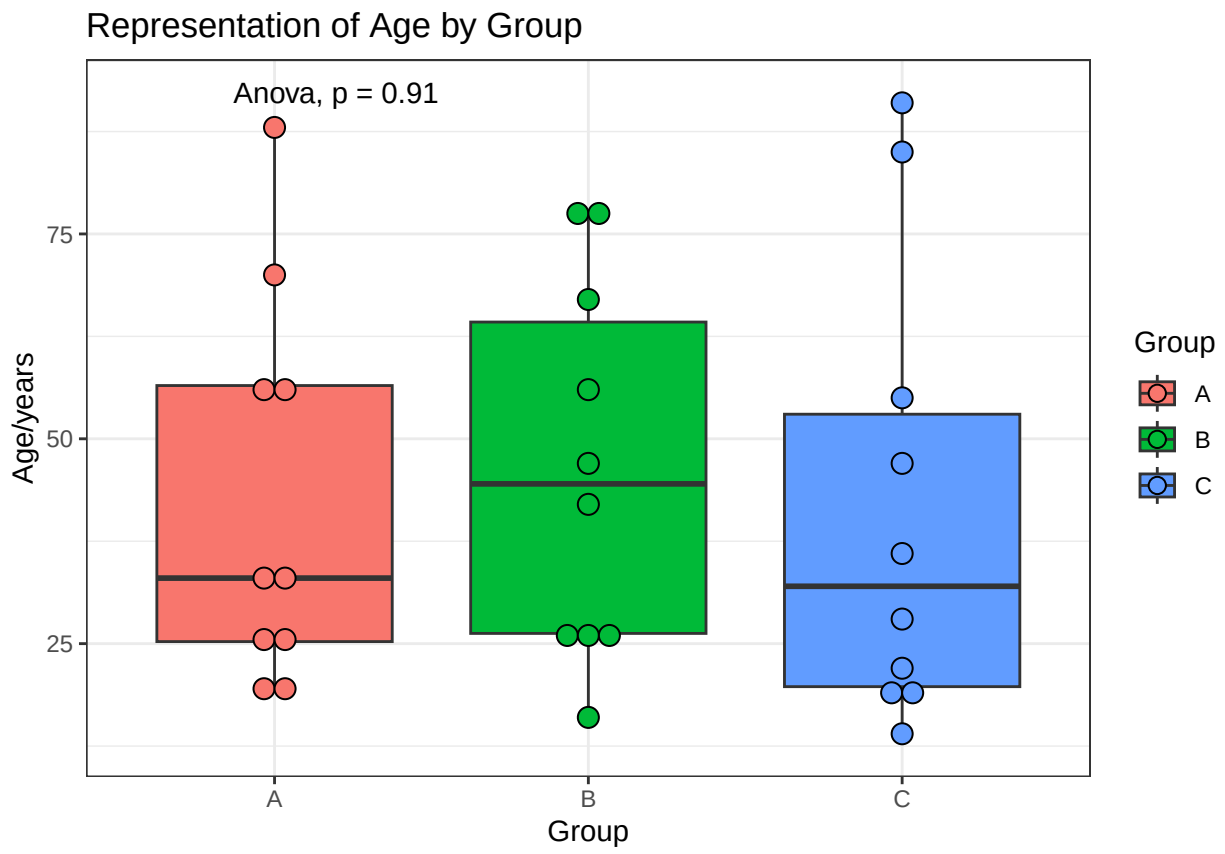
```
## Warning in wilcox.test.default(c(47, 22, 36, 19, 14, 55, 91, 85, 19, 28), :
## cannot compute exact p-value with ties
```



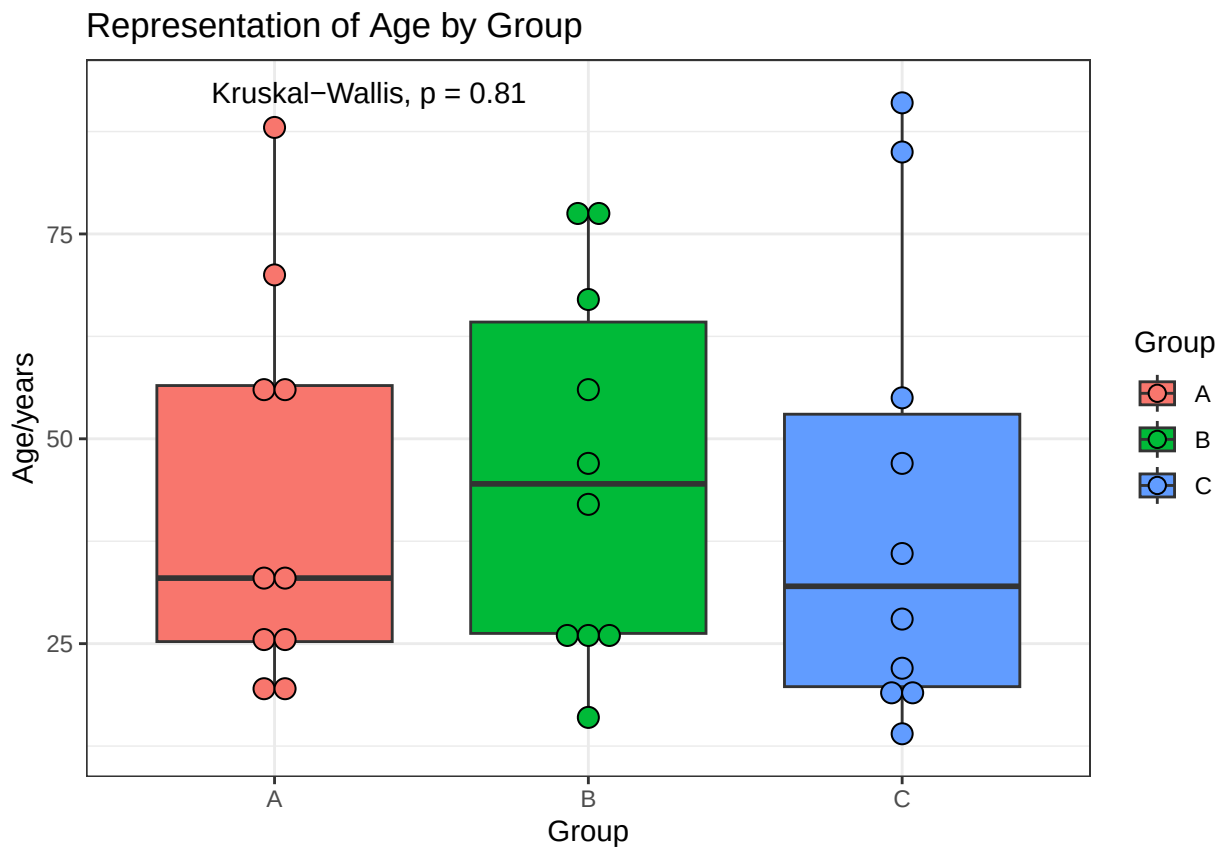

```
ggplot(df, aes(x = Group, y = Age, fill=Group)) +
  geom_boxplot()+
  geom_dotplot(binaxis = "y", stackdir = "center")+
  labs(title="Representation of Age by Group", y = "Age/years")+
  stat_compare_means(data = df,method = "t.test", paired=FALSE,comparisons = my_comparisons,label="p.fo
  theme_bw()
```




```
ggplot(df, aes(x = Group, y = Age, fill=Group)) +
  geom_boxplot()+
  geom_dotplot(binaxis = "y", stackdir = "center")+
  labs(title="Representation of Age by Group", y = "Age/years")+
  stat_compare_means(data = df,method = "anova")+
  theme_bw()
```




```
ggplot(df, aes(x = Group, y = Age, fill=Group)) +
  geom_boxplot()+
  geom_dotplot(binaxis = "y", stackdir = "center")+
  labs(title="Representation of Age by Group", y = "Age/years")+
  stat_compare_means(data = df,method = "kruskal.test")+
  theme_bw()
```

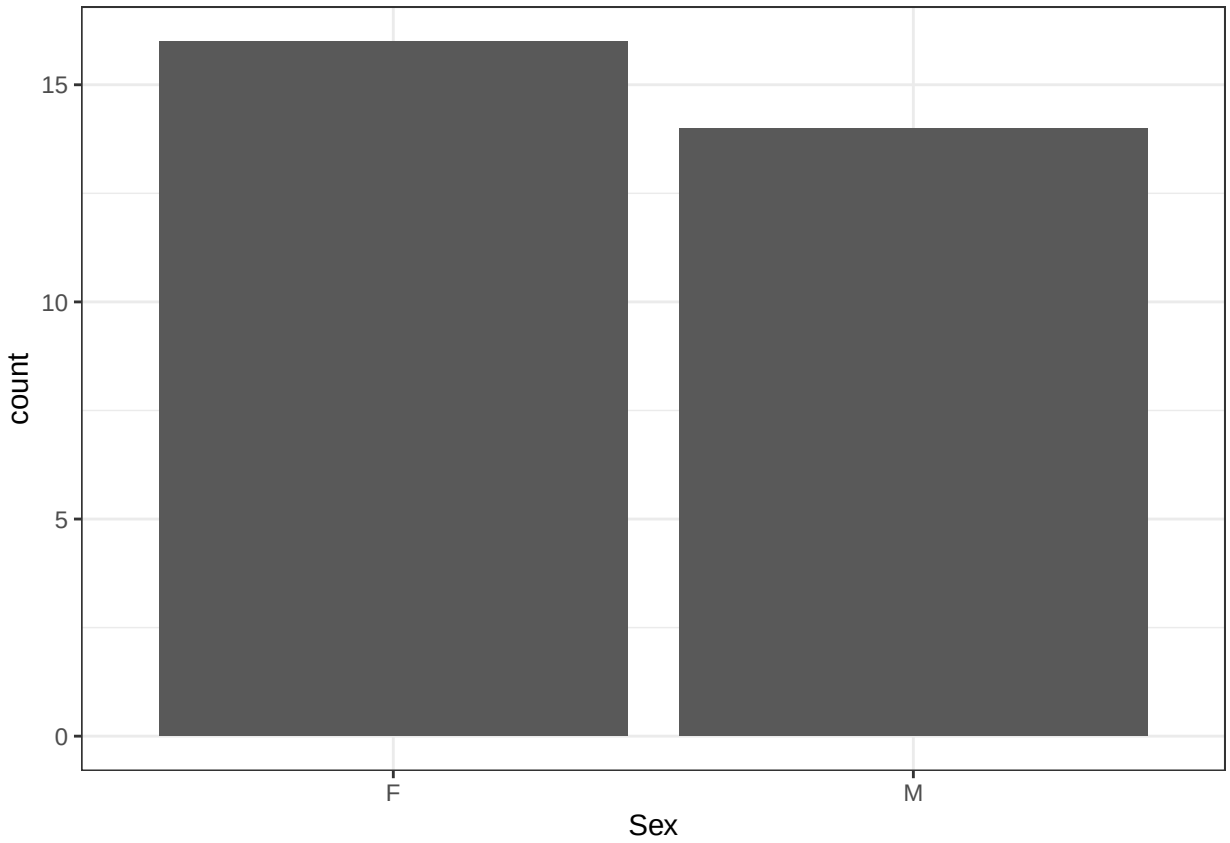


Representation of categorical variable (Sex)

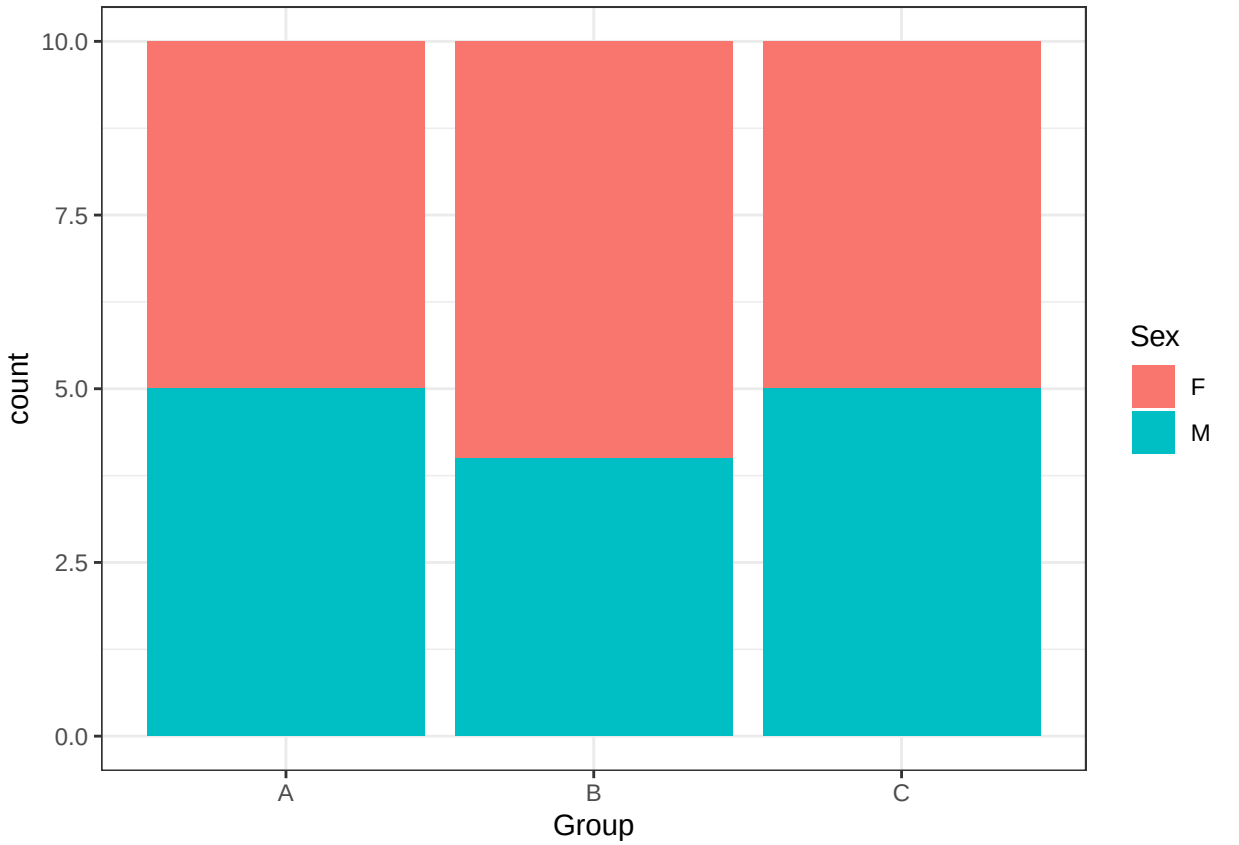
Bar plot Description: There are two types of bar charts: geom_bar() and geom_col(). geom_bar() makes the height of the bar proportional to the number of cases in each group (or if the weight aesthetic is supplied, the sum of the weights). If you want the heights of the bars to represent values in the data, use geom_col() instead. geom_bar() uses stat_count() by default: it counts the number of cases at each x position. geom_col() uses stat_identity(): it leaves the data as is.

With total number (counts).

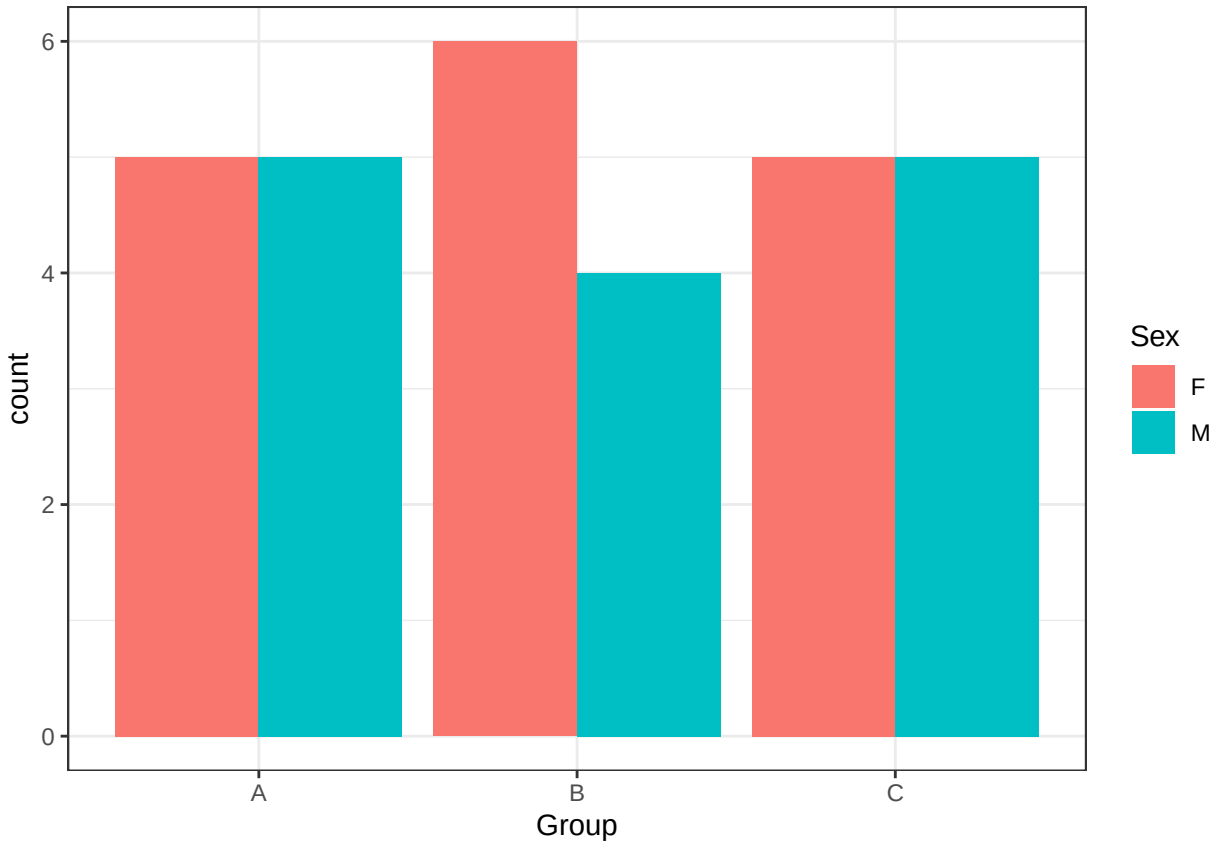
```
#basic
ggplot(df, aes(x = Sex)) +
  geom_bar()+
  theme_bw()
```




```
#by group stacked
ggplot(df, aes(x = Group, fill=Sex)) +
  geom_bar()+
  theme_bw()
```

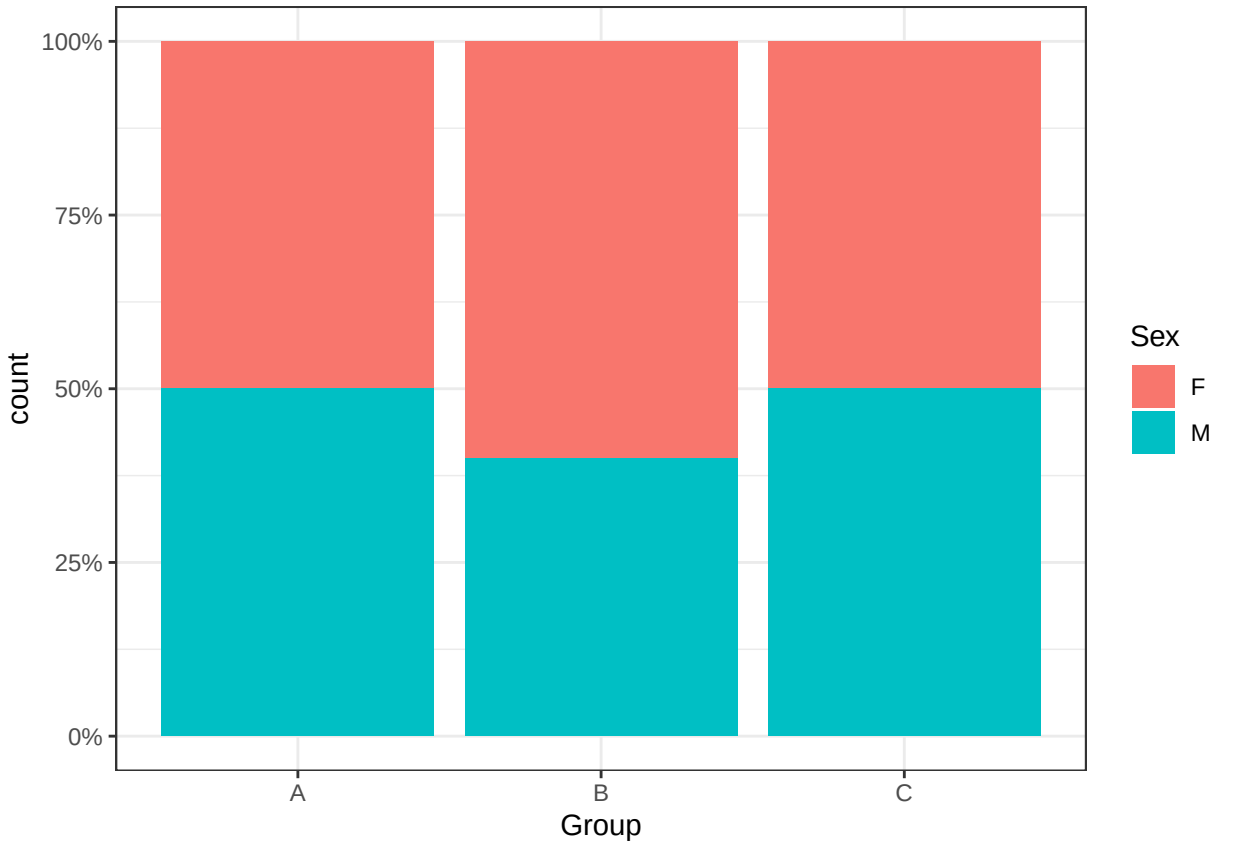



```
#by group split
ggplot(df, aes(x = Group, fill=Sex)) +
  geom_bar(position = "dodge", stat = "count")+
  theme_bw()
```

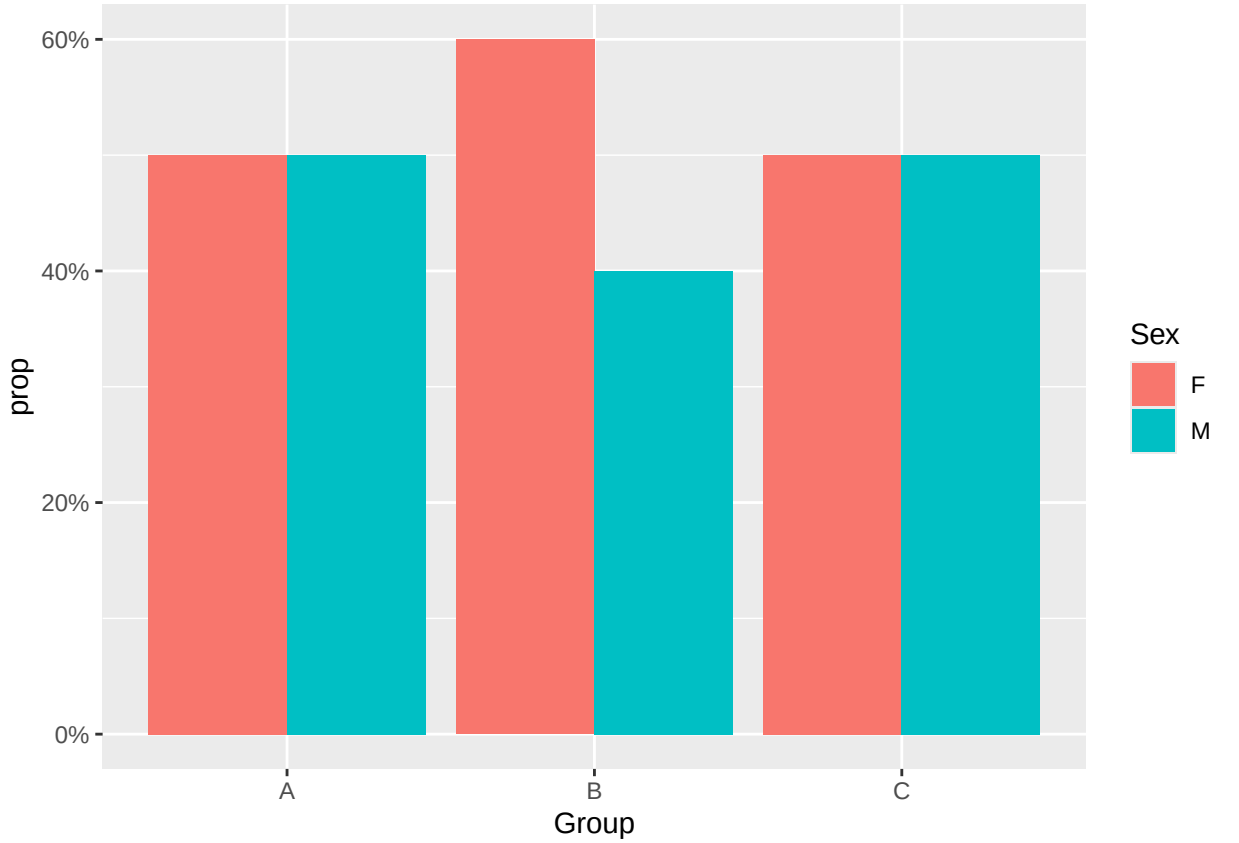


With percentages.

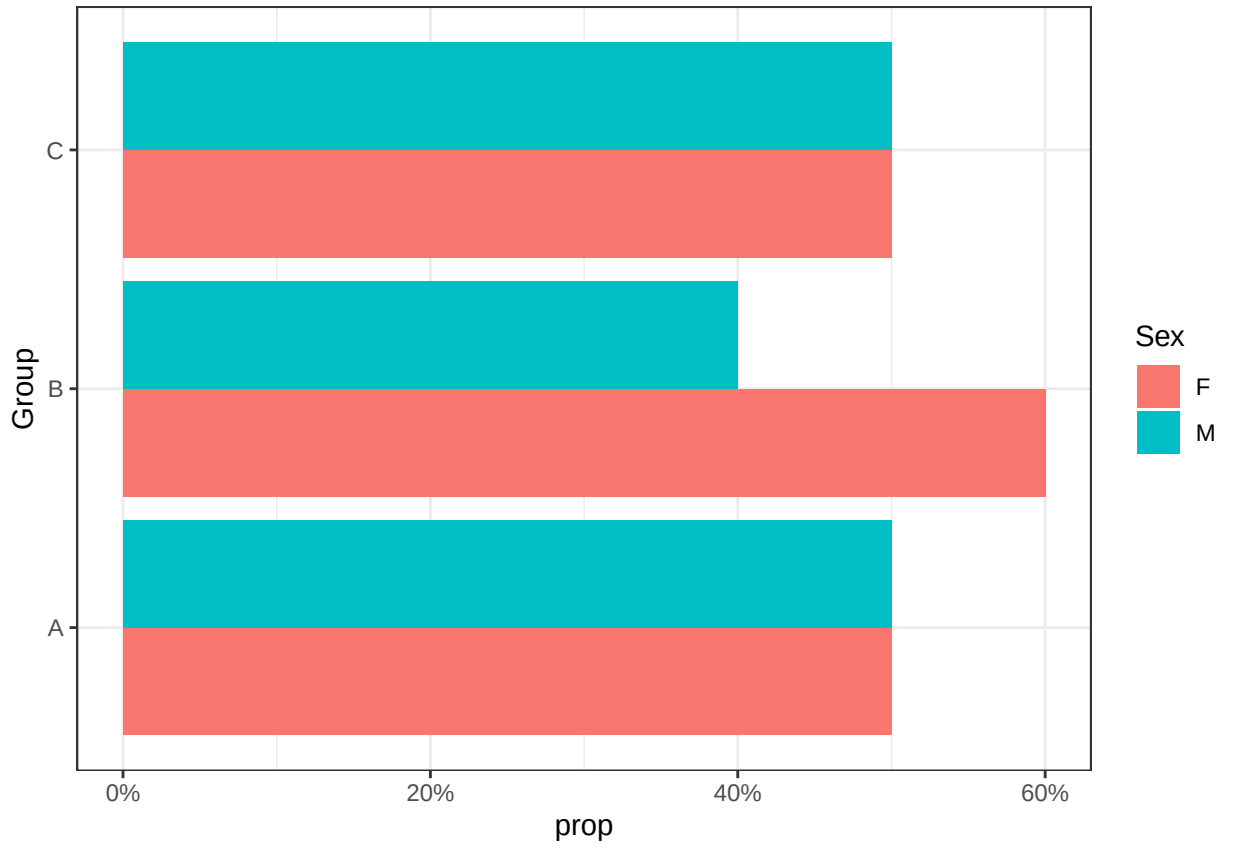
```
#by group stacked
ggplot(df, aes(x = Group, fill=Sex)) +
  geom_bar(position = "fill",stat="count")+
  scale_y_continuous(labels = scales::percent_format())+
  theme_bw()
```




```
#by group split

#BiocManager::install("dplyr")

library(dplyr)

df_summary <- df %>%
  group_by(Group, Sex) %>%
  summarise(count = n()) %>%
  ungroup() %>%
  group_by(Group) %>%
  mutate(prop = count / sum(count))

ggplot(df_summary, aes(x = Group, y = prop, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_y_continuous(labels = scales::percent_format())
```




```
#coord_flip
ggplot(df_summary, aes(x = Group, y = prop, fill = Sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_y_continuous(labels = scales::percent_format())+
  coord_flip()+
  theme_bw()
```



Calculate significant of the differential distribution and if there is an association between two categorical variable.

Fisher's exact test.

```
fisher.test(df$Group, df$Sex)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  df$Group and df$Sex
## p-value = 1
## alternative hypothesis: two.sided
```

Exercise 2

2.1. Plot the distribution of Age by Group in females and verify if the differences are statistically significant.

2.2. Plot the distribution of Group by Sex in percentage in individuals under 40 years of age and verify if the differences are statistically significant.