**Data manipulation and visualisation in R**

**Mauzum shamil a m 20bsc132**

# using an external csv file

Manipulation

### **Getting started with tidyverse packages**

#

install.packages("tidyverse")

library(tidyverse)

We will use the starwars data from dplyr - you may be noticing a theme here…

#

starwars <- dplyr::starwars

### **Selecting columns**

# with names

starwars[, c('name', 'homeworld')]

# with indices

starwars[, c(1, 9)]

With dplyr we can do the following:

starwars %>% select(name, homeworld)

select is more powerful than just this. See some examples below:

# choose all columns BUT name

starwars %>% select(-name)

# choose only columns containing an underscore

starwars %>% select(contains("\_"))

# choose only columns beginning with "s"

starwars %>% select(starts\_with("s"))

### **Filtering data**

#

starwars %>% filter(species == "Human")

#

starwars %>% filter(species == "Human", homeworld == "Tatooine")

You can see how dplyr makes filtering on multiple variables much more straightforward and cleaner.

### **Filtering AND selecting data**

starwars %>% filter(species == "Human") %>% select(name, height, birth\_year)

### **Summarising data**

starwars %>% group\_by(species) %>% tally()

We can do a lot more than just count occurrences with this functionality from dplyr.

#

starwars %>% group\_by(species) %>% summarise(mean\_height = mean(height, na.rm = T),

mean\_mass = mean(mass, na.rm = T))

#

starwars %>% group\_by(species) %>% summarise\_at(vars(height, mass), mean, na.rm = T)

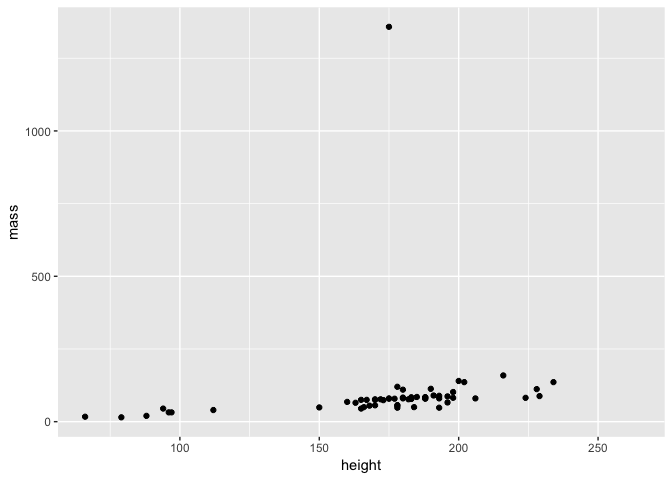
Data visualization

## Plotting with ggplot2

starwars <- dplyr::starwars

Let’s do a simple scatter plot of the height and mass of the 87 characters in this dataset.

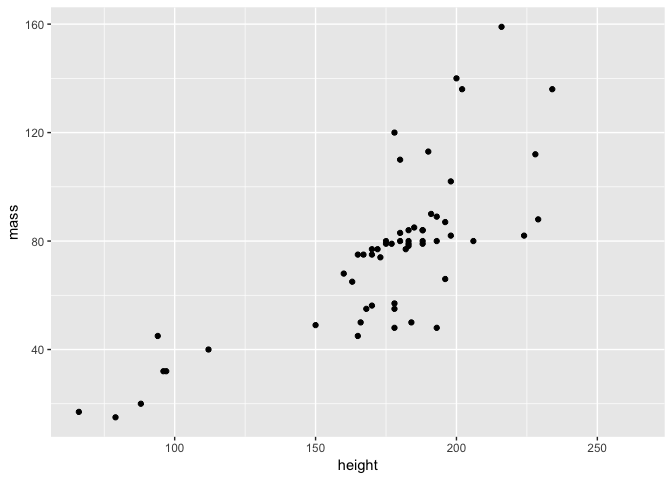
ggplot(starwars, aes(height, mass)) + geom\_point()



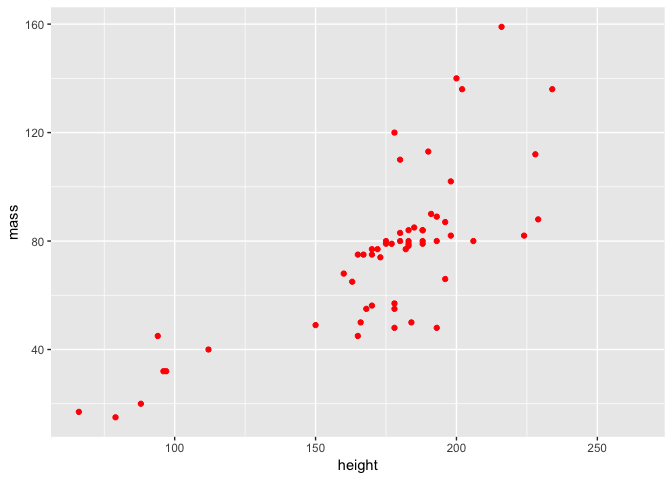
starwars2 <- filter(starwars, name != "Jabba Desilijic Tiure")

Now we can replot the data and see what the relationship between height and mass is.

ggplot(starwars2, aes(height, mass)) + geom\_point()

**Customising a plot**

ggplot(starwars2, aes(height, mass)) + geom\_point(colour = "red")



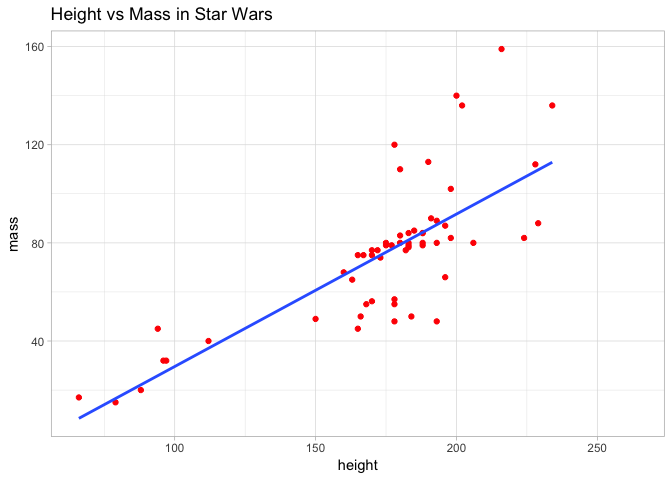
a <- ggplot(starwars2, aes(height, mass))

a <- a + geom\_point(colour = "red")

a <- a + geom\_smooth(method = lm, se = FALSE)

a <- a + theme\_light() + ggtitle("Height vs Mass in Star Wars")

a

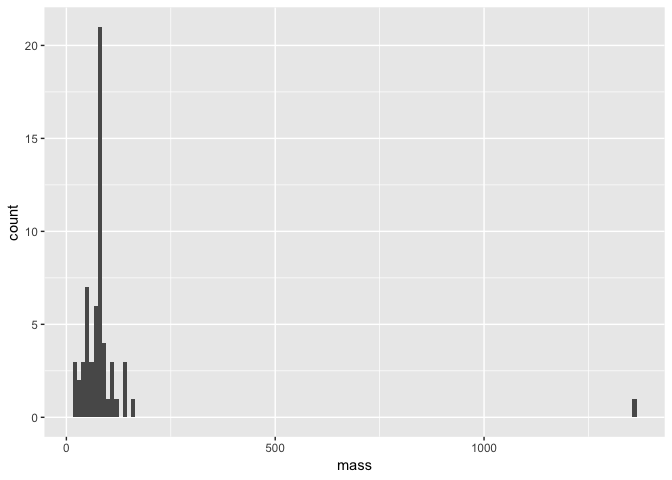


### **Different types of plot with ggplot2**

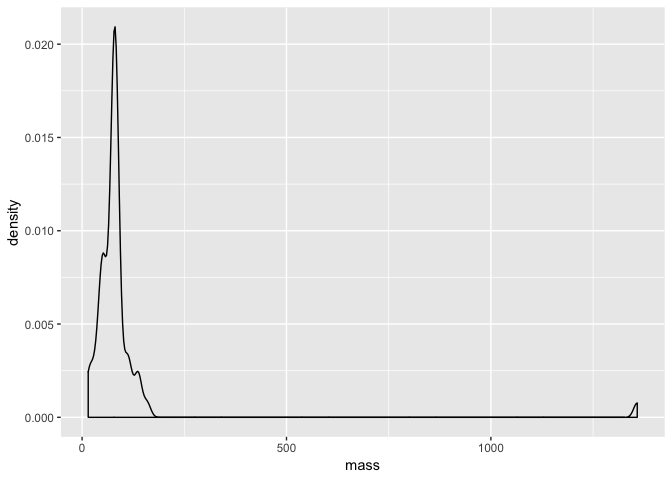
#### 

#### **Visualising data distributions**

a <- ggplot(starwars, aes(mass))

a + geom\_histogram(binwidth = 10)

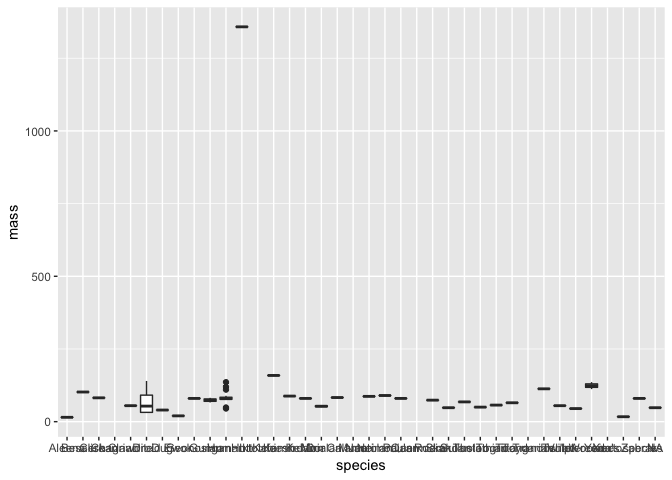
a + geom\_density()



#### **Comparing data distributions**

a <- ggplot(starwars, aes(species, mass))

a + geom\_boxplot()



# subset data

hde <- starwars %>% subset(species == "Human" | species == "Droid" | species == "Ewok")

# plot the distributions

a <- ggplot(hde, aes(species, mass))

a + geom\_boxplot()

