Exercise 3 - Data Manipulation

Keaan Amin

06 May 2021

Grab the data.csv file from the Github Directory (https://github.com/keaan95/virtual-elective/tree/master/Week1) and put it into your Working Directory.

EXERCISE 1

1.1. Add the data file to your ENVIRONMENT.

TIP - Remember to Set your Working Directory to the location of your data file.

```
# You can fetch your WORKING directory as follows
getwd()
```

```
## [1] "C:/Users/keaan/OneDrive - Newcastle University/HLA - Virtual Elective/Week1"
```

```
# Change to your local directory - in my case it is as follows:
setwd("C:/Users/keaan/OneDrive - Newcastle University/HLA - Virtual Elective/Week1/")
df <- read.csv("data.csv")
```

EXERCISE 2

2.1. Visualise the first five columns and rows.

```
df[1:5,1:5]
```

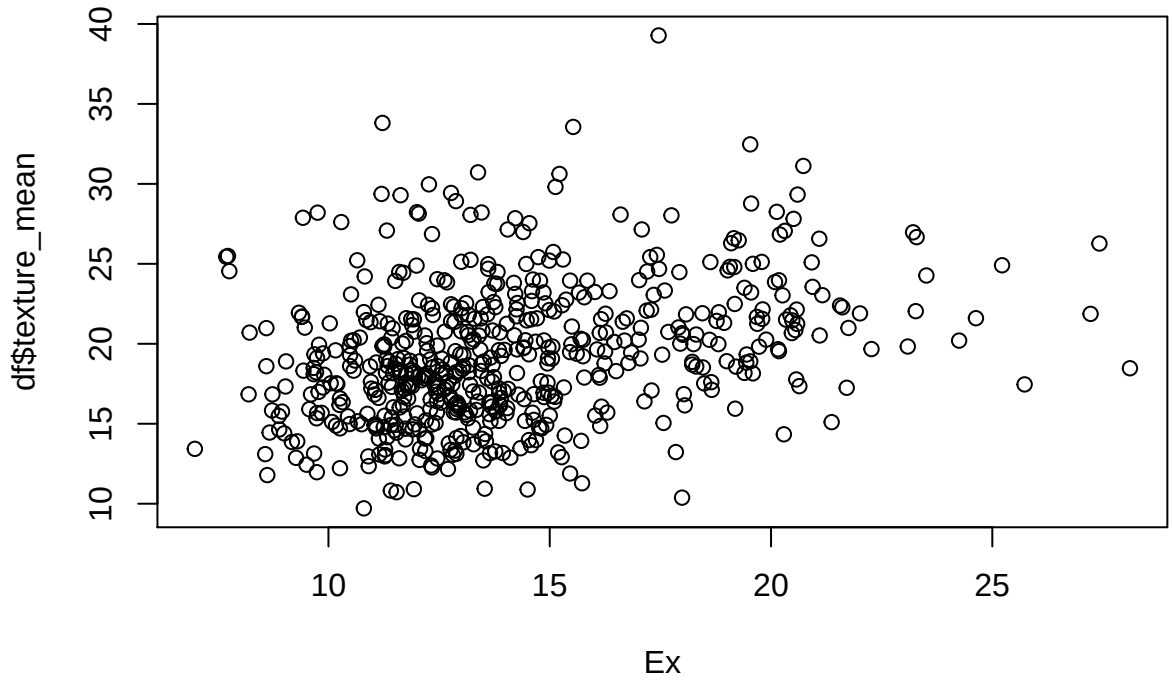
```
##         id diagnosis radius_mean texture_mean perimeter_mean
## 1   842302         M       17.99        10.38         122.80
## 2   842517         M       20.57        17.77         132.90
## 3 84300903         M       19.69        21.25         130.00
## 4 84348301         M       11.42        20.38          77.58
## 5 84358402         M       20.29        14.34         135.10
```

2.2. How many patients are in the data?

```
nrow (df)
```

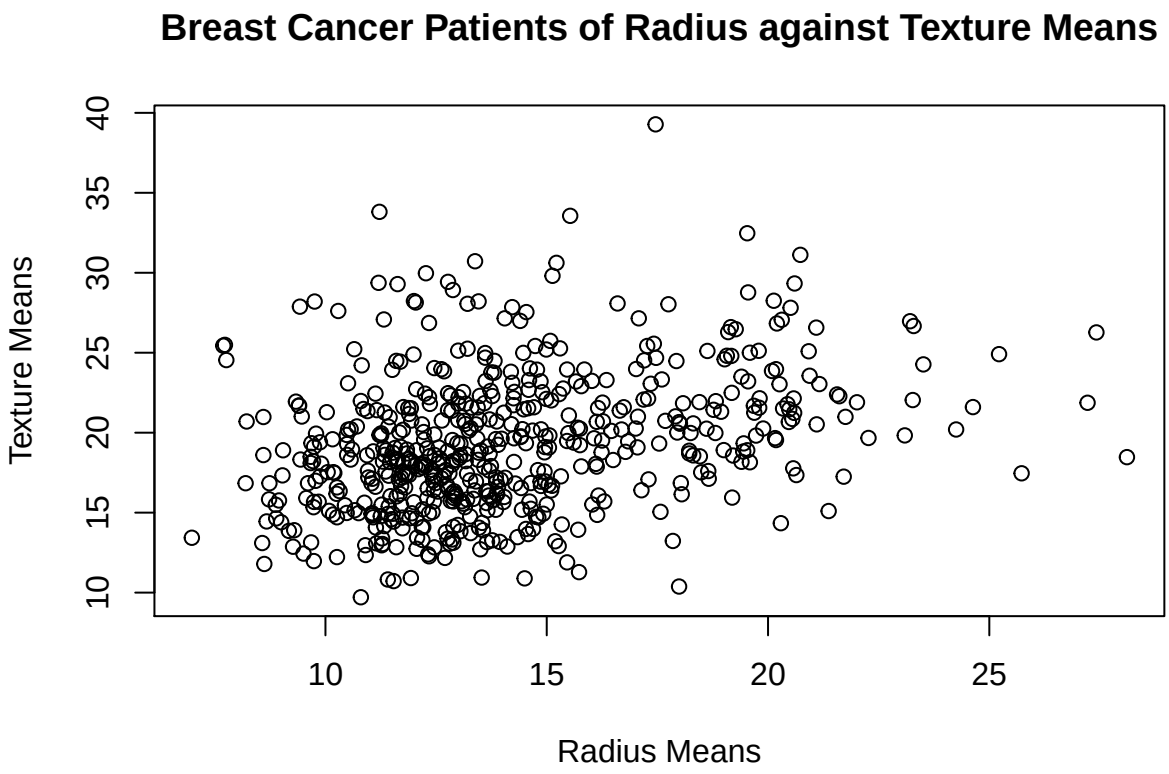
2.3. Let's Plot Radius against Perimeter Means of Our Patients

```
plot(df$radius_mean,df$texture_mean,
     xlab="Ex")
```



2.3. Let's Add some Legends and a Title

```
plot(df$radius_mean,df$texture_mean,
     xlab="Radius Means",
     ylab="Texture Means",
     main="Breast Cancer Patients of Radius against Texture Means")
```



2.4. Let's Colour By Malignant vs. Benign Patients

In order to do this we need to change our column 'data type'

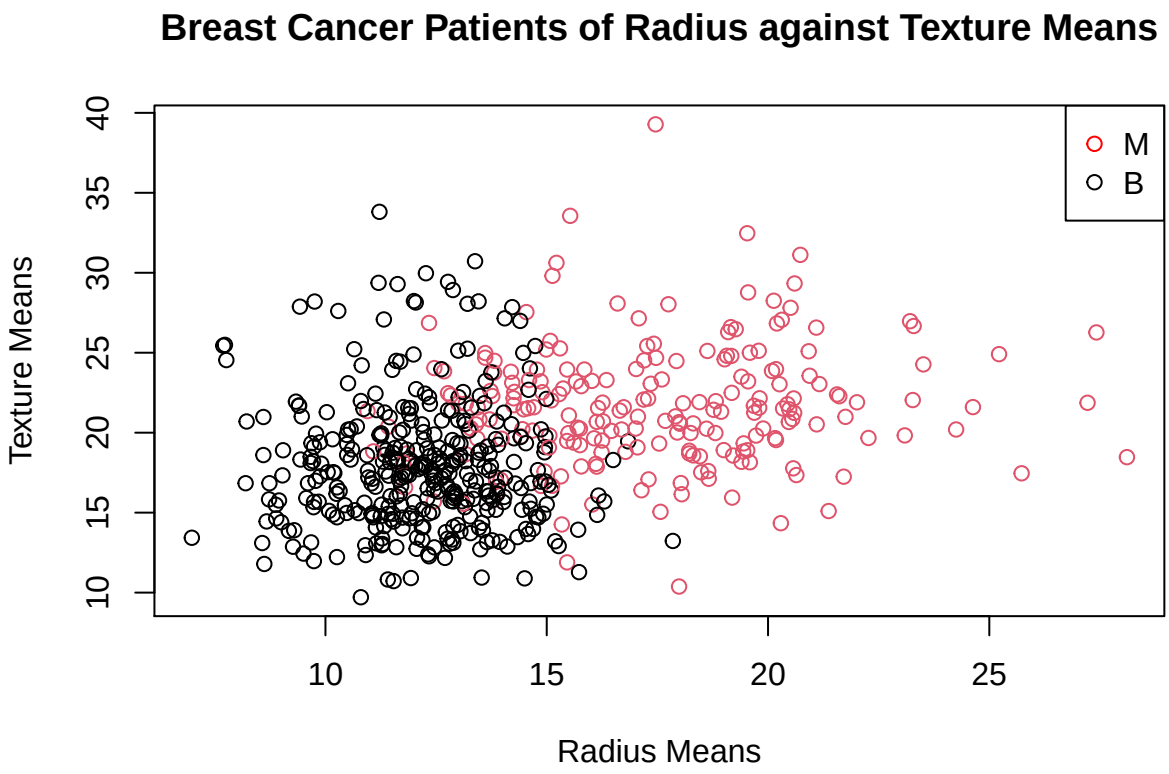
```
# View the data type of structures of our dataframe 'df'
str(df)
```

Our 'df$diagnosis' are a bunch of 'M's or 'B's - known as 'characters.' We need this data type to be a 'factor' so R recognises it belongs to a particular group. We can do this by:

```
as.factor(df$diagnosis)
```

Using the col variable, we can specify this:

```
plot(df$radius_mean,df$texture_mean,
     col=as.factor(df$diagnosis),
     xlab="Radius Means",
     ylab="Texture Means",
     main="Breast Cancer Patients of Radius against Texture Means")
legend("topright",legend = c("M", "B"), pch=c(1,1), col=c("red","black"))
```



EXERCISE 3

3.1. How many patients have Malignant Tumours?

```
sum(df$diagnosis=="M")
```

3.2. How many patients have a Radius Mean greater than or equal to 13?

```
sum(df$radius_mean >= 13)
```

3.3. How many Malignant Tumours have a Radius Mean greater than or equal to 20?

There are many solutions to this problem. A simple yet easy way is to create a Dataframe of only Malignant Patients.

```
mdf <- df[df$diagnosis=="M",]
```

Next we can apply our filters:

```r
# Using our newly created dataframe:
sum(mdf$radius_mean >= 20)

# Alternatively, we can count the rows.
nrow(mdf[mdf$radius_mean >= 20,])
```

That wraps-up all the preparatory work required ahead of our LIVE session!

If you have gotten this far, well done! The initial learning curve is really hard with coding. Do not worry, we will go through these answers in our live session!