foobar

Barry

2/2/2022

We will use the stroke dataset again to make plots that can quickly answer similar types of questions we were asking in the previous worksheet.

With plots, the most important thing to consider is:

- Are you plotting one or two variables?
- What goes on the x-axis? What goes on the y-axis?
- Think about the columns you want to plot and the information you want to portray!

Run the codeblock below to load the plotting library and the stroke dataset.

```
install.packages("ggpubr")
```

```
## Installing package into '/home/barry/R/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```
## Warning in register(): Can't find generic `scale_type` in package ggplot2 to
## register S3 method.
```

```
library(ggpubr)
stroke <- data.frame(read.delim("https://raw.githubusercontent.com/BarryDigby/Youth-Academy/master/data/
stroke <- stroke[which(stroke$gender != "Other"),]
```

Question 1

Let's investigate Body Mass Index (BMI) in the dataset.

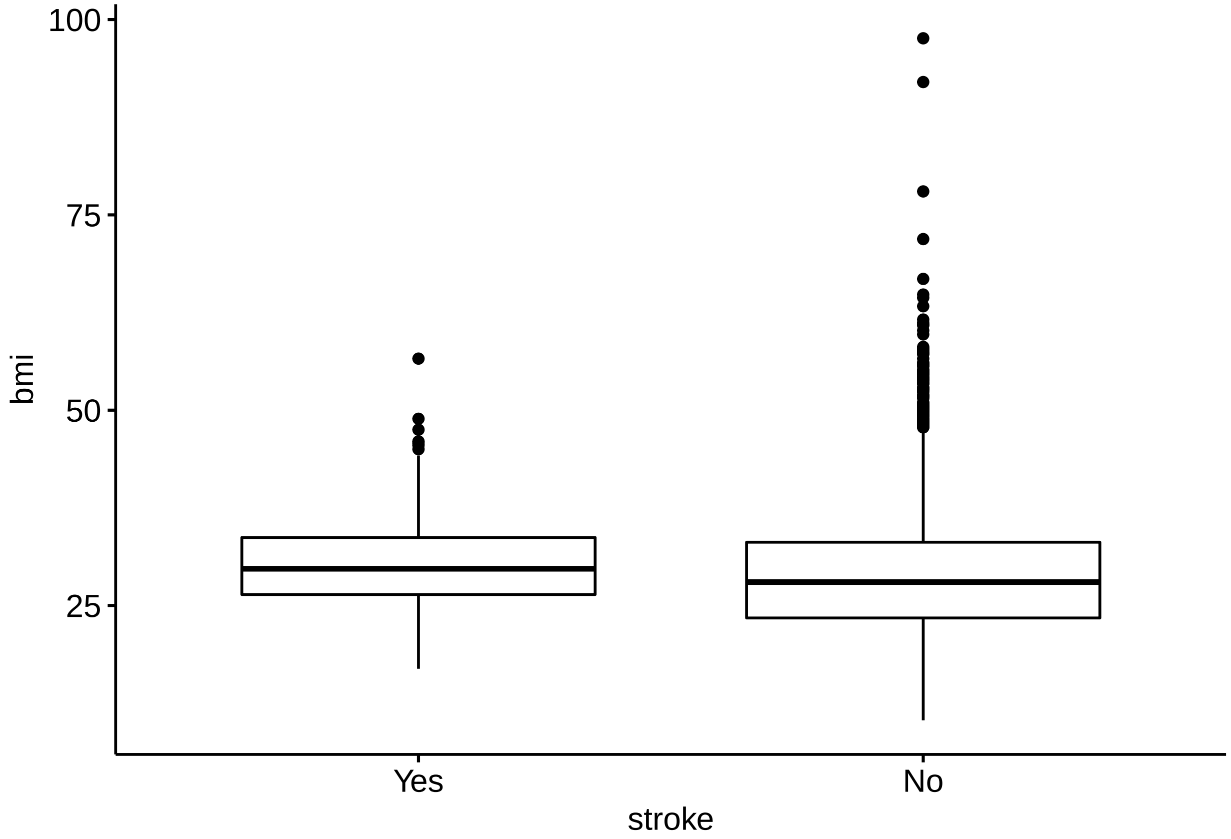
Using ggboxplot(), make a boxplot of bmi on the y-axis and stroke on the x-axis. This will allow us to see if people with higher BMI values are more susceptible to strokes.

```
install.packages("ggpubr")
```

```
## Installing package into '/home/barry/R/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library(ggpubr)
```

```
ggboxplot(stroke, x="stroke", y="bmi")
```

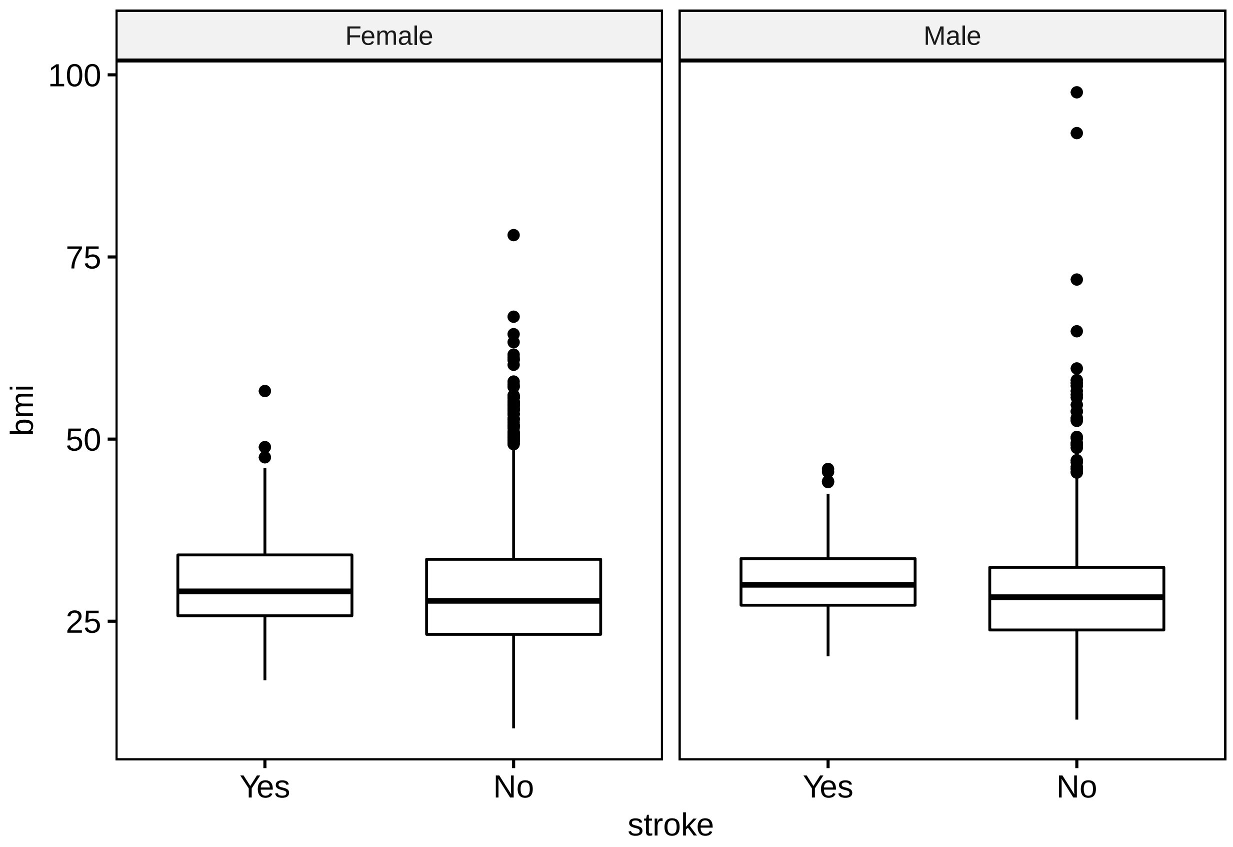


yes, higher bmi values could lead to a stroke

Question 2

Produce the same plot, but this time include 'facet.by' in the code to produce a plot for both females and males using the column 'gender'. Do the results look the same? (i.e are the plots similar for both females and males)

```
ggboxplot(stroke, x="stroke", y="bmi", facet.by = "gender")
```

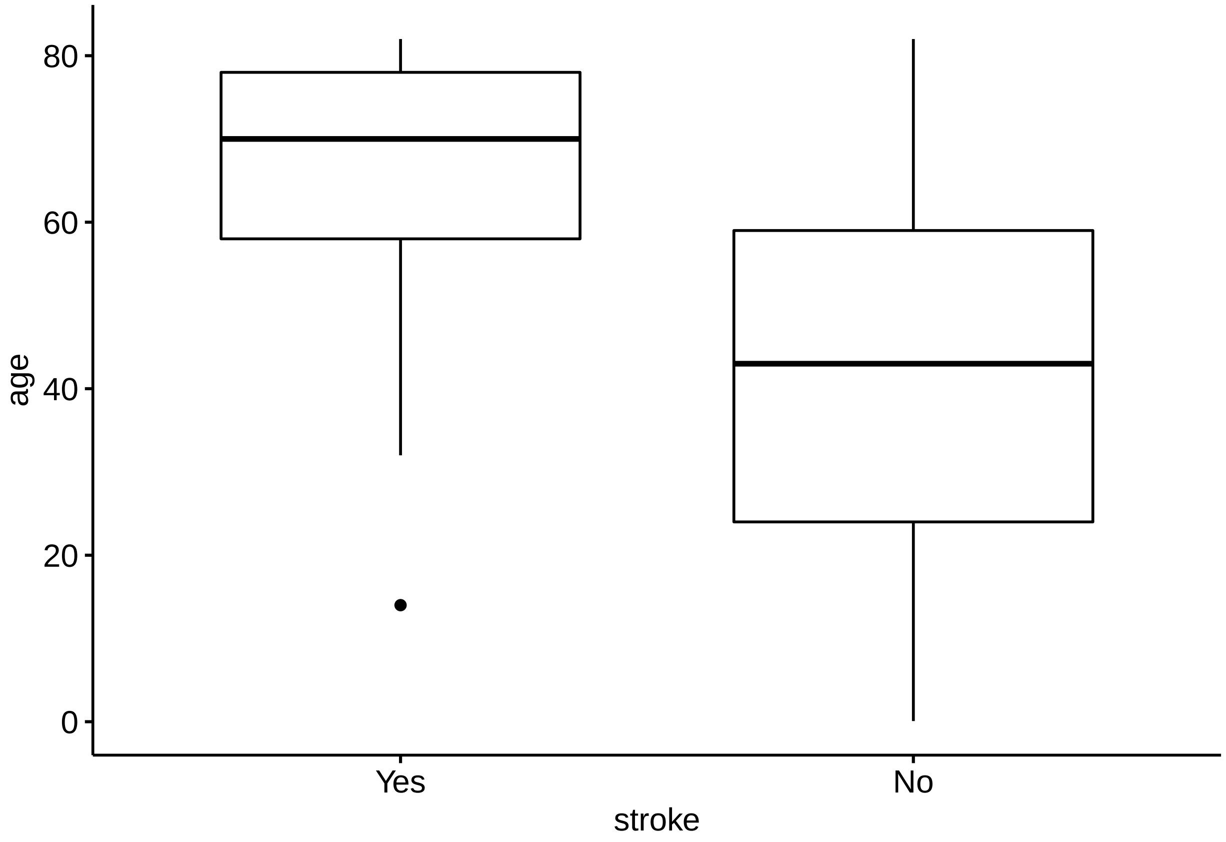


Question 3

Let's investigate Age in the dataset.

Using ggboxplot(), make a boxplot of age on the y-axis and stroke on the x-axis. This will allow us to see if people with higher ages are more susceptible to strokes.

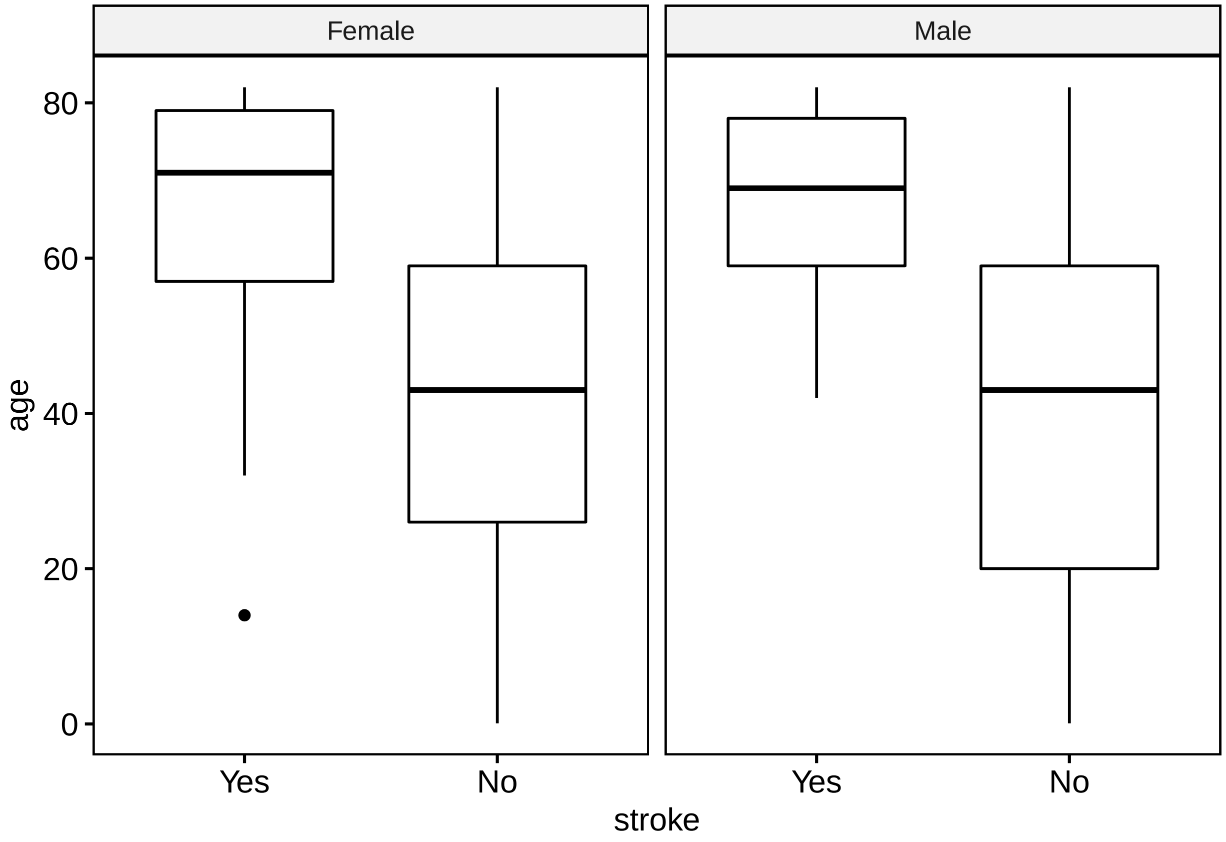
```
ggboxplot(stroke, x="stroke", y="age")
```



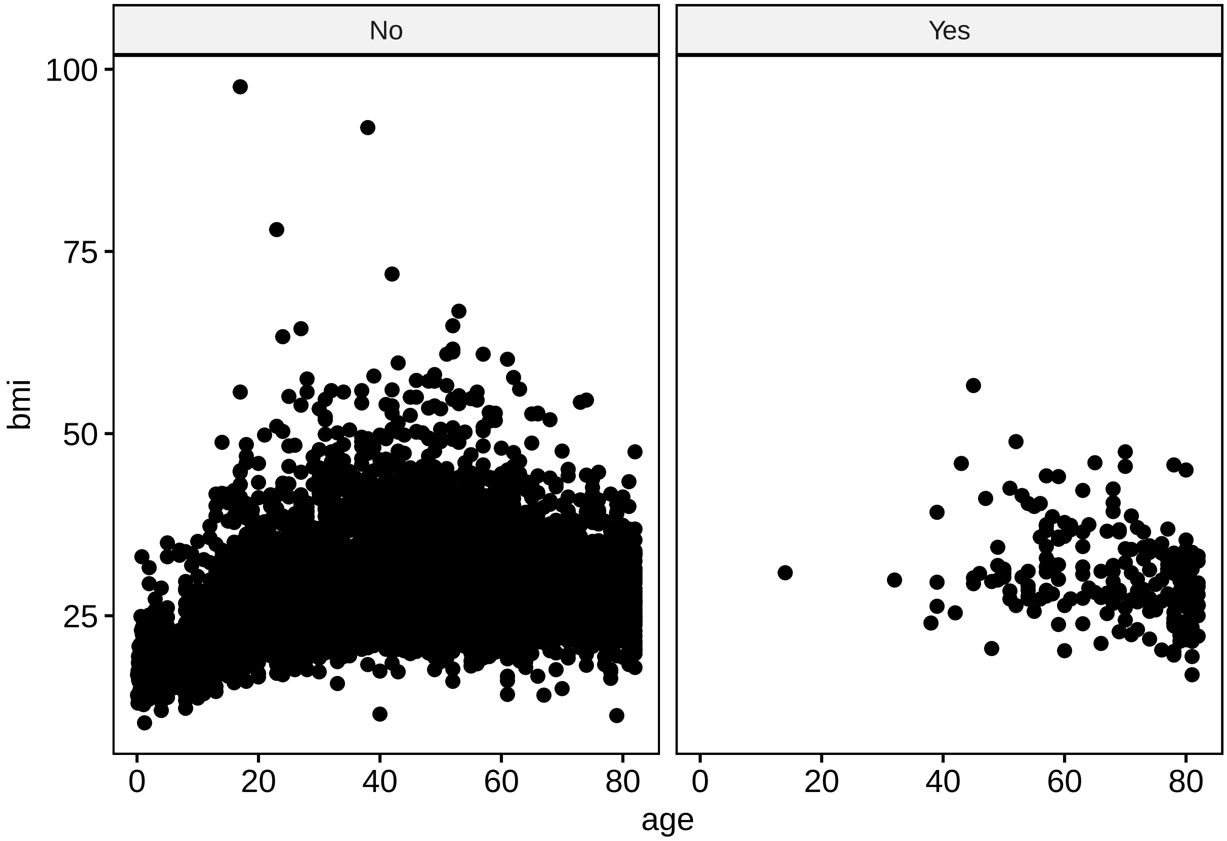
Question 4

Produce the same plot, but this time include 'facet.by' in the code to produce a plot for both females and males. Do the results look the same? (i.e are the plots similar for both females and males)

```
ggboxplot(stroke, x="stroke", y="age", facet.by="gender")
```



```
ggscatter(stroke, x="age", y="bmi", facet.by = "stroke")
```



Task 1

Create a new column in the dataset called 'diabetic' using an 'ifelse' statement to figure out if the patients are diabetic or not. The ifelse statement reads as follows:

> If patients avg_glucose_level is greater than 200: 'Yes', else: 'No'.

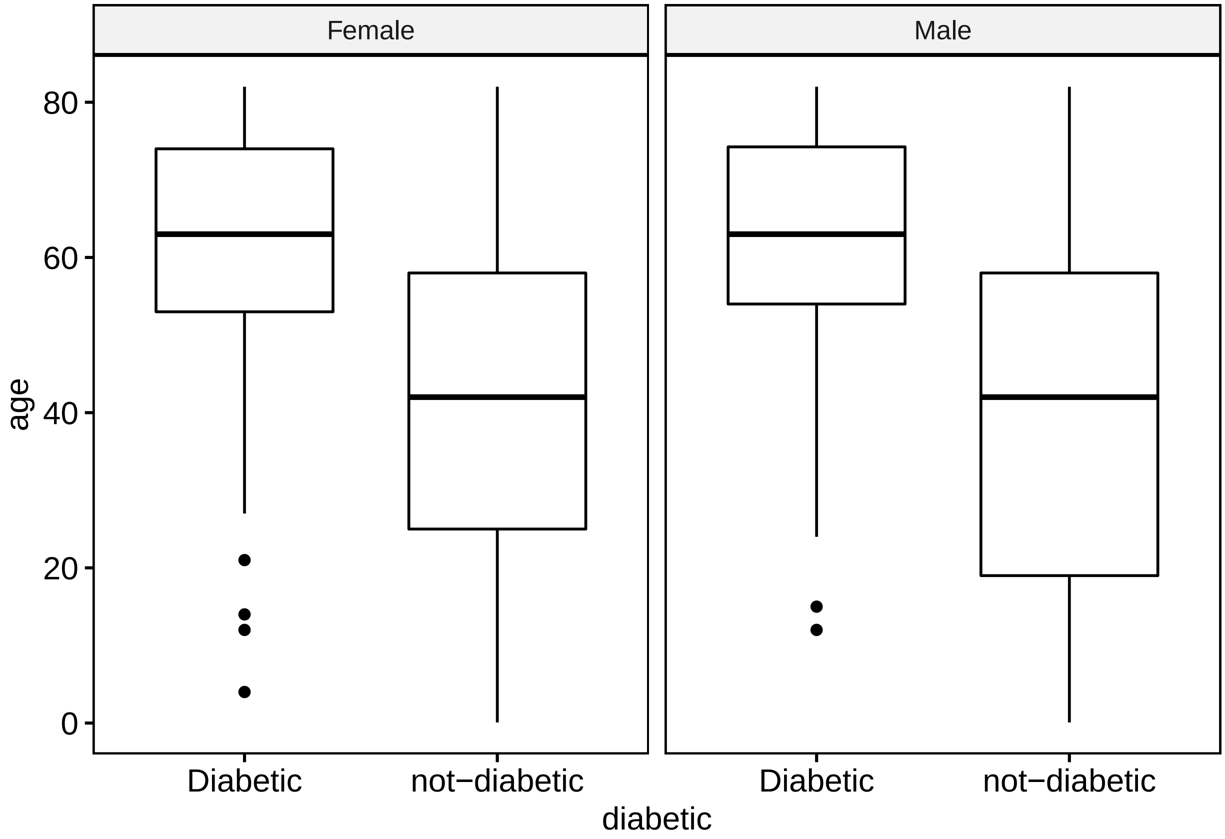
```
stroke$diabetic <- ifelse(stroke$avg_glucose_level > 200, "Diabetic", "not-diabetic")
```

Question 5

Are older patients more likely to be diabetic?

Make a boxplot with your new column 'diabetic' on the x-axis and 'age' on the y-axis. Facet the plot using 'gender' to check if it is true for both females and males.

```
ggboxplot(stroke, x="diabetic", y="age", facet.by = "gender")
```



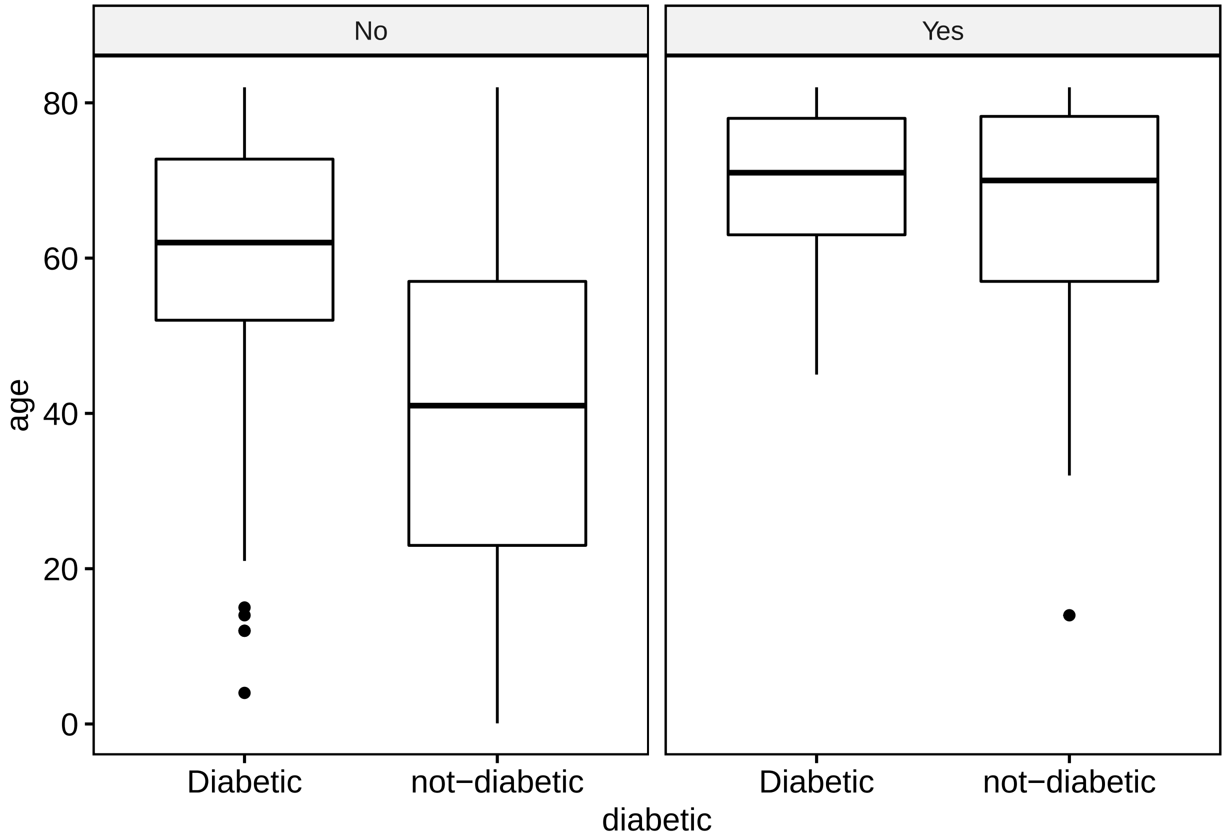
Question 6

We can see that older patients (in both females and males) tend to be diabetic. We also saw that older people are more likely to have strokes.

Does this mean that older people with diabetes are more likely to have a stroke?

Make a boxplot with your new column 'diabetic' on the x-axis and 'age' on the y-axis. Facet the plot using 'stroke'. Focus on the right panel showing patients who had a stroke. Are diabetic patients more likely to have a stroke? (boxplot is much higher/lower).

```r
ggboxplot(stroke, x="diabetic", y="age", facet.by="stroke")
```



This is a good example of teasing apart which variables are associated with the outcome (stroke). This takes practice, but once you are comfortable making plots, you can start to ask these types of questions yourself!

Gene expression example.

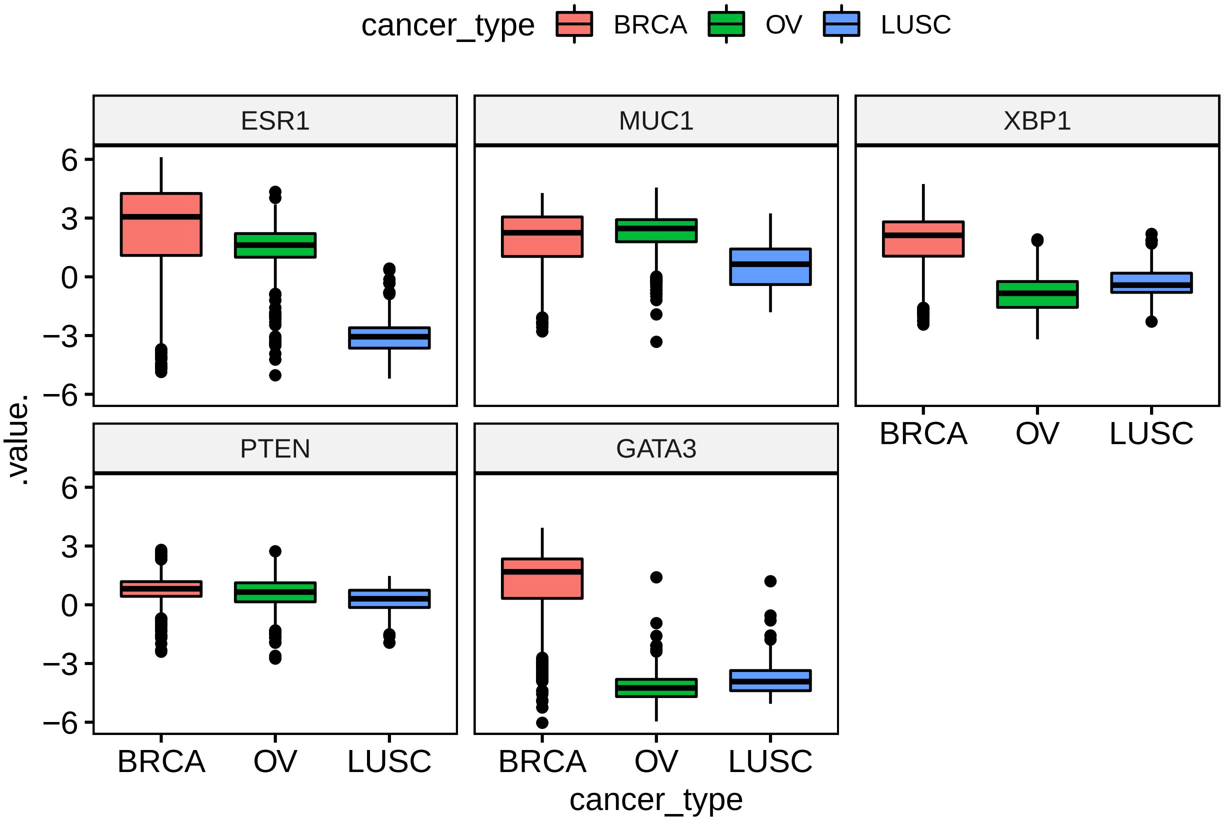
So far we have worked with clinical data, now we will load in a gene expression dataset and answer questions about it. This is section is about interpreting the plots and linking it to the underlying genomics. The three cancer types are:

- BRCA: Breast Cancer
- OV: Ovarian Cancer
- LUSC: Lung Cancer

Run the code block below to make the plots:

```
gene_data <- read.delim("https://raw.githubusercontent.com/BarryDigby/Youth-Academy/master/data/expr.tx


ggboxplot(gene_data, x="cancer_type", y=c("ESR1", "MUC1", "XBP1", "PTEN", "GATA3"), combine=T, color="b
```



Your challenge is to identify which genes are up-regulated in a cancer when compared to others. For example, we can say that GATA3 is up-regulated in BRCA (breast cancer). Google the gene name and cancer type and write a very short paragraph on your findings.

An example would be:

> Gene X is up-regulated in cancer A. This gene has a function in cell signalling, and has been found to be a clinical marker for cancer A.

Do not worry if you cannot find great information, it is not really the data scientists job to do this :)