Breast_cancer_analysis

Barry

2/3/2022

Intro

We are going to use scatterplots to analyse 3 subtypes of breast cancer:

- Luminal A
- Luminal B
- Triple negative

> The goal here is to see which genes are good at separating each subtype of breast cancer, and classify 2 new patients into the correct subtype category.

We have a dataset with 150 patients, 50 for each subtype. Load the dataset below:

```
library(ggpubr)
```

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

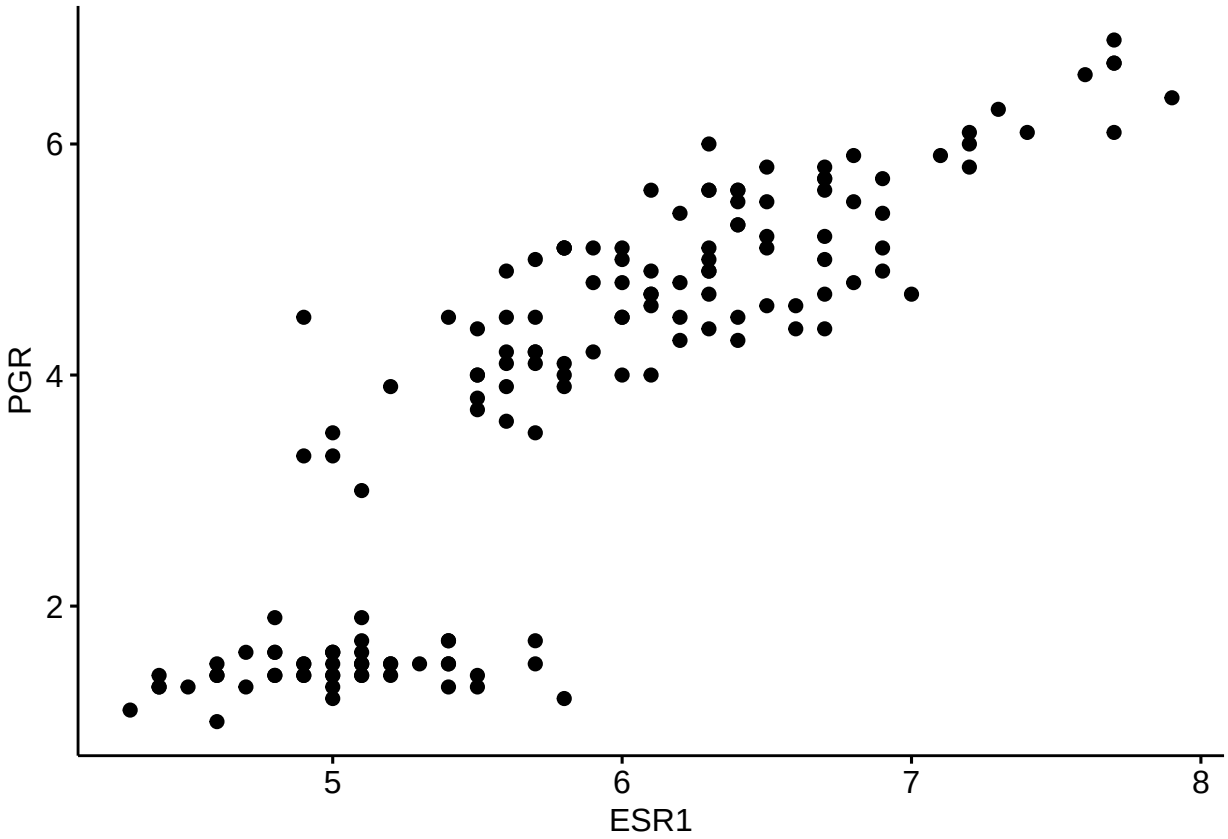
```
## Warning in register(): Can't find generic 'scale_type' in package ggplot2 to
## register S3 method.
```

```
dataset <- read.delim("https://raw.githubusercontent.com/BarryDigby/TY_workshop/master/docs/source/works
```

Walkthrough

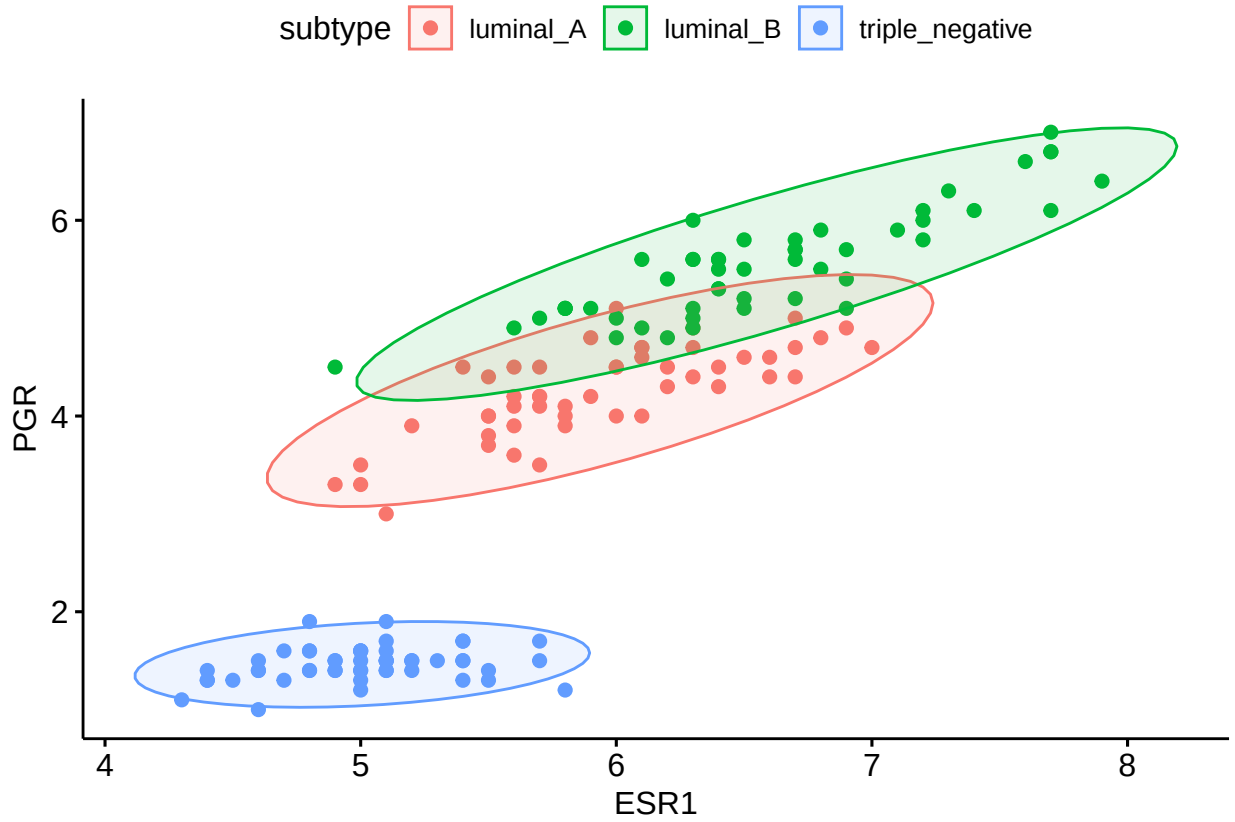
Make a basic scatter plot of `ESR1` and `PGR`:

```
ggscatter(dataset, x="ESR1", y="PGR")
```



The plot is hard to interpret. We will make it easier to identify the subtypes by adding colors and clusters to the plot:

```
ggscatter(dataset, x="ESR1", y="PGR", color="subtype", ellipse = TRUE)
```

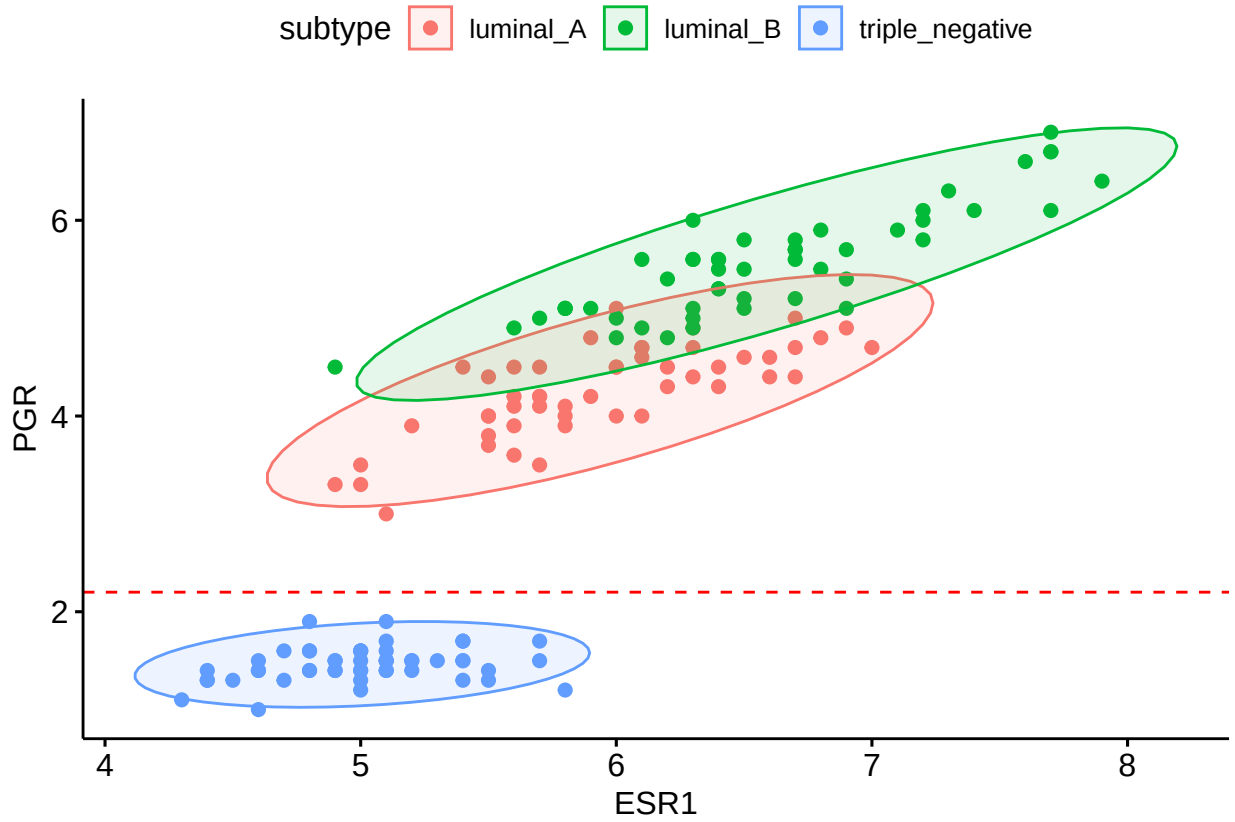


Triple negative breast cancer subtypes have low Estrogen receptor (ESR1) and low Progesterone receptor (PGR) expression levels. Our plot makes sense, the blue cluster represents triple negative breast cancer (TNBC) patients which are at the bottom left of the plot - lower expression that luminal A and luminal B subtypes.

PGR looks to be an excellent gene for classifying TNBC patients. If the patient has a PGR expression level below 2, the patient is probably TNBC.

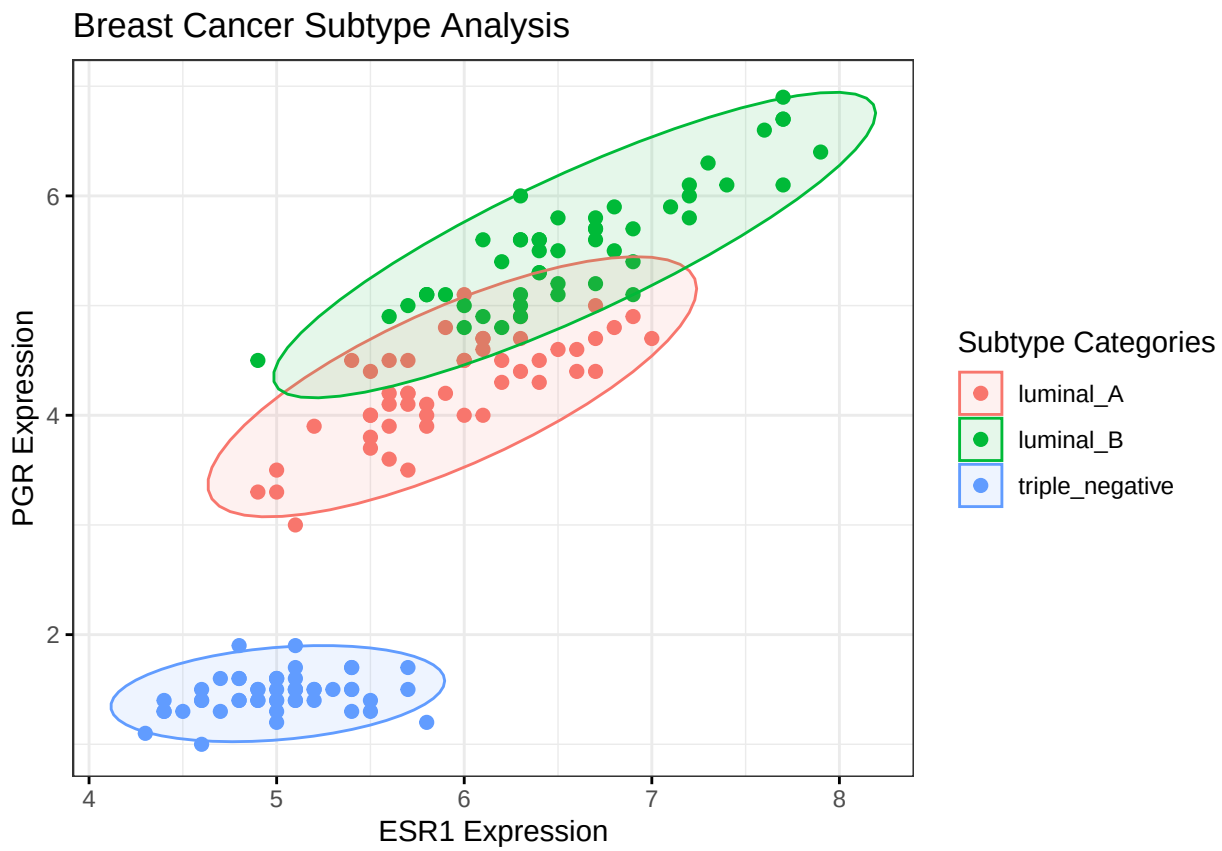
Run the code block below to visualise this threshold:

```
ggscatter(dataset, x="ESR1", y="PGR", color="subtype", ellipse = T) + geom_hline(yintercept=2.2, color=
```



Finally, customaise your plot with appropriate labels

```
ggscatter(dataset, x="ESR1", y="PGR", color="subtype", ellipse = T, title = "Breast Cancer Subtype Analy
```



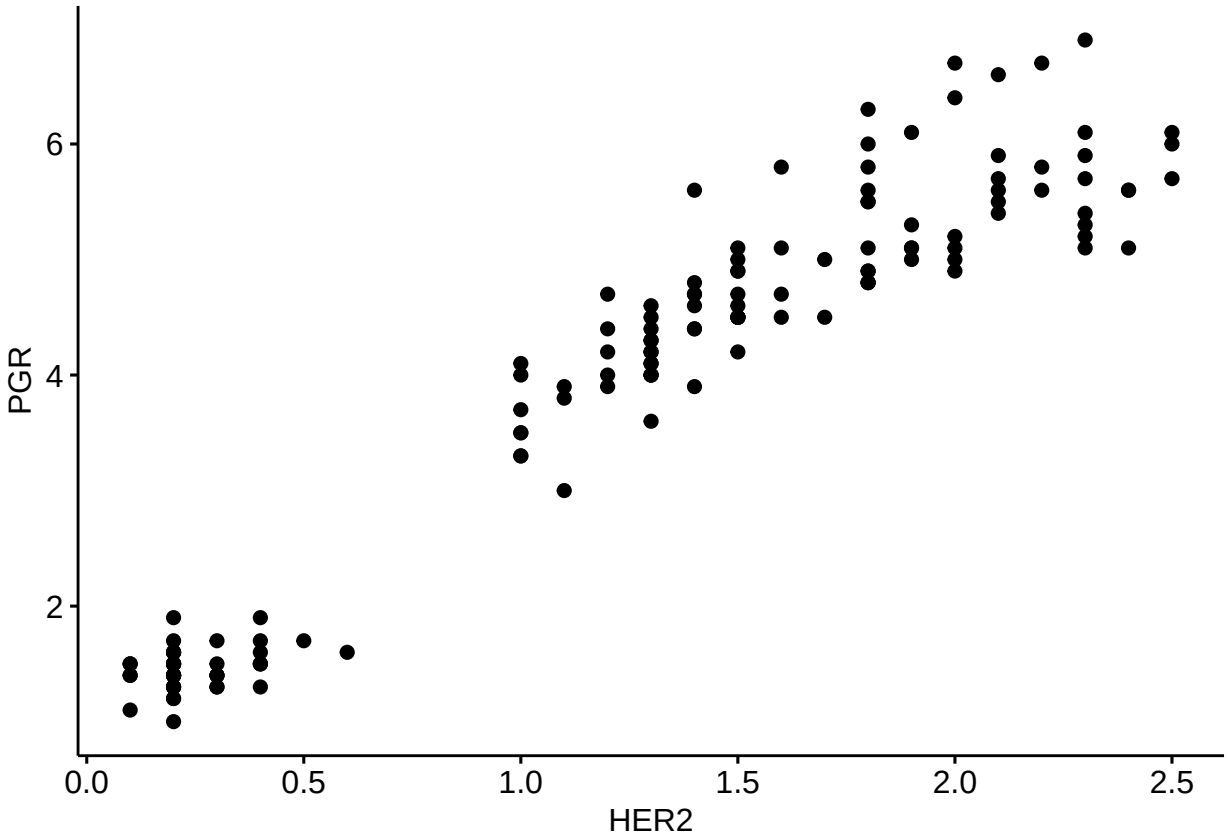
Exercise 1

Research states that luminal A breast cancers are HER2 negative and luminal B cancers are HER2 positive. This means we expect HER2 expression to be higher in luminal B when compared to luminal A patients.

Your job is to make scatterplots to show this trend. Use the same types of plots we made above, substituting `x="ESR1"` for `x="HER2"`. Below is a basic plot to get started, add colors and clusters to the plot to enhance it.

Can we see luminal A patients separate from luminal B patients according to `HER2` expression?

```
ggscatter(dataset, x="HER2", y="PGR")
```



Exercise 2

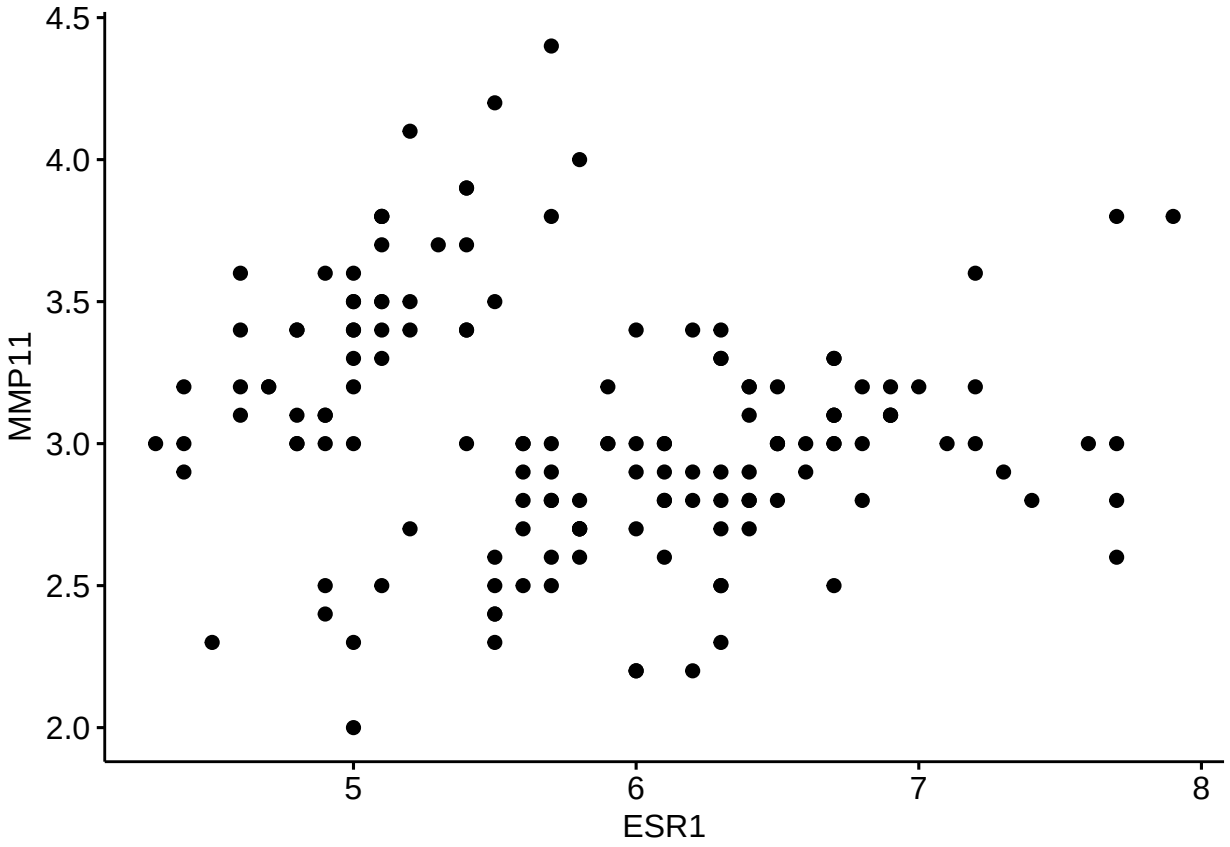
`PGR` and `HER2` are good genes because they help us identify luminal_A, luminal_B and triple negative breast cancers. These are known as 'informative genes'.

Your next job is to identify which gene is 'non-informative' in the dataset (it will be `MMP11` or `ESR1`). If it is not good a separating subtypes, we can stop using this gene in our gene panel and save money.

Make scatterplots of `MMP11` vs. `ESR1` and make a decision on which gene to remove. In the code block below, add colors and clusters to help aid your decision.

Which gene would you remove?

```
ggscatter(dataset, x="ESR1", y="MMP11")
```



Exercise 3

2 new patients have come to the clinic and we have measured their gene expression for `ESR1`, `MMP11`, `PGR` and `HER2`. Run the code block below to add the patients to your dataset. (Do not change anything in the block)

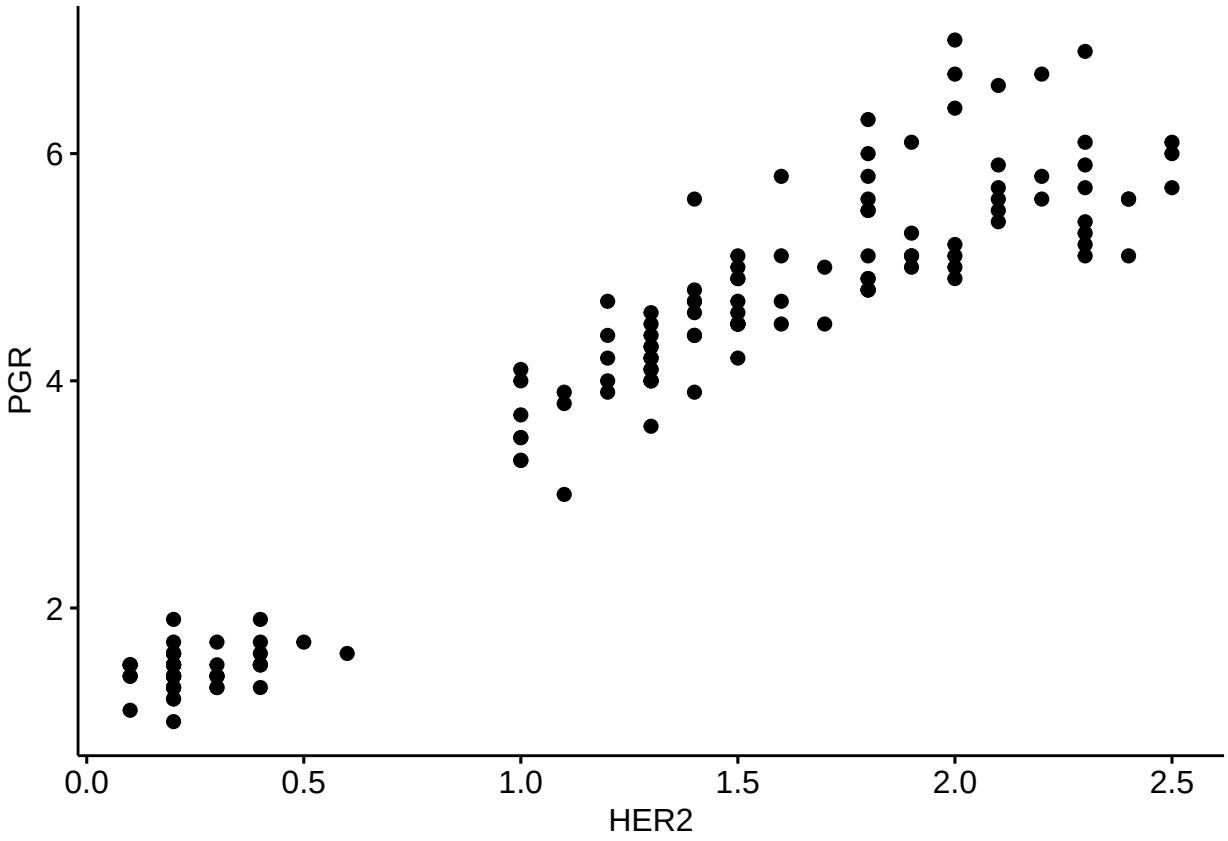
```
new_patient_1 <- c(5.0, 3.0, 1.5, 0.1, "new_patient_1")
new_patient_2 <- c(7.8, 3.9, 7, 2, "new_patient_2")
dataset <- rbind(dataset, new_patient_1, new_patient_2)
dataset$subtype <- as.factor(dataset$subtype)
dataset[,1:4] <- as.numeric(unlist(dataset[,1:4]))
```

Your job is to figure out which subtype the new patients belong to.

Make scatterplots using `HER2` and `PGR` to see which group they fall into. Add colors and clusters to the plot.

Can you figure out which breast cancer subtype the two new patients belong to?

```
ggscatter(dataset, x="HER2", y="PGR")
```


PGR
6
4
2
0.0
0.5
1.0
1.5
2.0
2.5
HER2