**Kriangkrai Saetan 6205019**

**Hand-on day 1**

* Load the data “gbsg.csv” (read data information at “gbsg data.doc”)

gbsg\_data <- read.csv("../data/gbsg.csv", header = TRUE, as.is = TRUE)

* Check for the 10 first lines

pid age meno size grade nodes pgr er hormon rfstime status

1 132 49 0 18 Moderate 2 0 0 0 1838 Alive

2 1575 55 1 20 Severe 16 0 0 0 403 Dead

3 1140 56 1 40 Severe 3 0 0 0 1603 Alive

4 769 45 0 25 Severe 1 0 4 0 177 Alive

5 130 65 1 30 Moderate 5 0 36 1 1855 Alive

6 1642 48 0 52 Moderate 11 0 0 0 842 Dead

7 475 48 0 21 Severe 8 0 0 0 293 Dead

8 973 37 0 20 Moderate 9 0 0 1 42 Alive

9 569 67 1 20 Moderate 1 0 0 1 564 Dead

10 1180 45 0 30 Moderate 1 0 0 0 1093 Dead

* Get a summary of each variable

pid age meno size grade nodes

Min. : 1.0 Min. :21.00 Min. :0.0000 Min. : 3.00 Mild : 81 Min. : 1.00

1st Qu.: 580.8 1st Qu.:46.00 1st Qu.:0.0000 1st Qu.: 20.00 Moderate:444 1st Qu.: 1.00

Median :1015.5 Median :53.00 Median :1.0000 Median : 25.00 Severe :161 Median : 3.00

Mean : 966.1 Mean :53.05 Mean :0.5773 Mean : 29.33 Mean : 5.01

3rd Qu.:1340.5 3rd Qu.:61.00 3rd Qu.:1.0000 3rd Qu.: 35.00 3rd Qu.: 7.00

Max. :1819.0 Max. :80.00 Max. :1.0000 Max. :120.00 Max. :51.00

pgr er hormon rfstime status

Min. : 0.0 Min. : 0.00 Min. :0.0000 Min. : 8.0 Alive:387

1st Qu.: 7.0 1st Qu.: 8.00 1st Qu.:0.0000 1st Qu.: 567.8 Dead :299

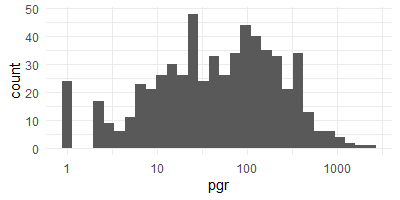
Median : 32.5 Median : 36.00 Median :0.0000 Median :1084.0

Mean : 110.0 Mean : 96.25 Mean :0.3586 Mean :1124.5

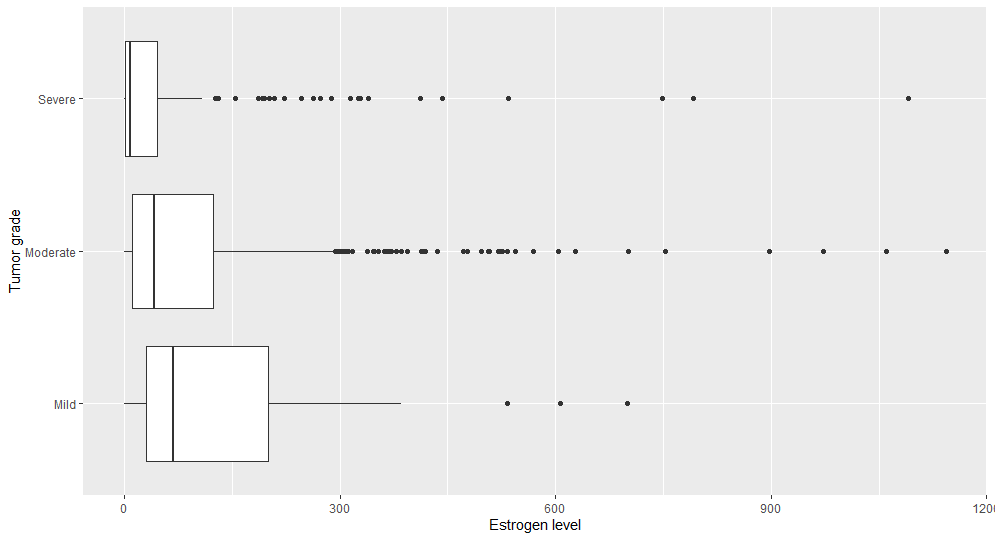
3rd Qu.: 131.8 3rd Qu.: 114.00 3rd Qu.:1.0000 3rd Qu.:1684.8

Max. :2380.0 Max. :1144.00 Max. :1.0000 Max. :2659.0

* Plot a histogram of progesterone level and scaling the x axis as log scale. Do you have any comments?



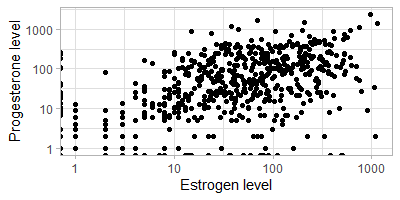
* Plot a boxplot the distribution of estrogen level (x-axis) by tumor grade (y-axis). Do you have any comment?



* Change the class of variable “status” to factor which labels 0 = “Alive” and 1 = “Dead”.

gbsg\_data$status <- factor(gbsg\_data$status, labels=c("Alive","Dead"))

* Plot a scatter plot between estrogen level (x-axis) and progesterone level (y-axis) and faceting by the status. Also, using a logarithmic scale on both x- and y-axis.



* Change the class of variable “grade” to factor which labels 1 = “Mild”, 2 = “Moderate” and 3 = “Severe”.

gbsg\_data$grade <- factor(gbsg\_data$grade, labels=c("Mild","Moderate","Severe"))

* Plot a scatter plot between estrogen level (x-axis) and progesterone level (y-axis) and faceting by the grade (x-panel) and status (y-panel). Apply color of each point by their observed age (variable “age”), and scaling the color from “yellow” for low age to “darkgreen” for high age. Using a logarithmic scale on both x- and y-axis. Add an average trend line using method = “loess”. Save the plot as “hand\_on\_plot.pdf”.

Chart

Description automatically generated

* Export the data of patients who had Tumor grade == “Severe” and estrogen level <= 5 to the new dataframe in a \*.txt file format using tab as column separator called *High\_risk\_patients.txt*

high\_risk <- gbsg\_data[gbsg\_data$grade == "Severe" & gbsg\_data$er <=5,]

write.table(file = "High\_risk\_patients.txt",

high\_risk, row.names = FALSE, sep = "\t")