Capstone project ~ Starting Data Visualization

Elena Tortosa

2/1/2019

Libraries and work directory

I loaded all required libraries and set the work directory.

```
library(tidyr)
library(dplyr)
library(ggplot2)
library(gridExtra)
setwd("/Users/Tortosae/Desktop/Data science course/Capstone_project")
```

Data tables

All data was provided in 14 different excel sheets. I saved all as .csv and loaded them in R. They were named with a number (df#) followed by the name of the perturbation + treatment each table contains.

```
df1_cd3cd28 <- read.table(file="1.cd3cd28.csv",sep=",", header=TRUE)
df2_cd3cd28icam2 <- read.table(file="2. cd3cd28icam2.csv",sep=",", header=TRUE)
df3_cd3cd28aktinhib <- read.table(file="3. cd3cd28aktinhib.csv",sep=",", header=TRUE)
df4_cd3cd28g0076 <- read.table(file="4. cd3cd28g0076.csv",sep=",", header=TRUE)
df5_cd3cd28psitect <- read.table(file="5. cd3cd28psitect.csv",sep=",", header=TRUE)
df6_cd3cd28u0126 <- read.table(file="6. cd3cd28u0126.csv",sep=",", header=TRUE)
df7_cd3cd28ly <- read.table(file="7. cd3cd28ly.csv",sep=",", header=TRUE)
df8_pma <- read.table(file="8. pma.csv",sep=",", header=TRUE)
df9_b2camp <- read.table(file="9. b2camp.csv",sep=",", header=TRUE)
df10_cd3cd28icam2aktinhib <- read.table(file="10. cd3cd28icam2aktinhib.csv",sep=",", header=TRUE)
df11_cd3cd28icam2g0076 <- read.table(file="11. cd3cd28icam2g0076.csv",sep=",", header=TRUE)
df12_cd3cd28icam2psit <- read.table(file="12. cd3cd28icam2psit.csv",sep=",", header=TRUE)
df13_cd3cd28icam2u0126 <- read.table(file="13. cd3cd28icam2u0126.csv",sep=",", header=TRUE)
df14_cd3cd28icam2ly <- read.table(file="14. cd3cd28icam2ly.csv",sep=",", header=TRUE)
```

Column names

Only df8 contained two columns with different names (in lower case). I unfied column names.

```
df8_pma <- df8_pma %>% rename (PIP2 = pip2, PIP3 = pip3)
```

New column for perturbations

Measurements are obtained from two different perturbations ("general perturbation" : GP1 and GP2). At the same time, these perturbations are combined with different treatments (or conditions) (see below: condition columns and dummy variables). I added a new column called GP to each table to classify the data depending on the general perturbation is applied : GP = 1 for GP1 and GP = 2 for GP2.

```
df1_cd3cd28 <- df1_cd3cd28 %>%
               mutate(treatment = "cd3cd28") %>% mutate (GP = 1)
```

```
df2_cd3cd28icam2 <- df2_cd3cd28icam2 %>%
                    mutate(treatment = "cd3cd28icam2") %>% mutate (GP = 2)
df3_cd3cd28aktinhib <- df3_cd3cd28aktinhib %>%
                      mutate(treatment = "cd3cd28aktinhib") %>% mutate (GP = 1)
df4_cd3cd28g0076 <- df4_cd3cd28g0076 %>%
                    mutate(treatment = "cd3cd28g0076") %>% mutate (GP = 1)
df5_cd3cd28psitect <- df5_cd3cd28psitect %>%
                      mutate(treatment = "cd3cd28psitect") %>% mutate (GP = 1)
df6_cd3cd28u0126 <- df6_cd3cd28u0126 %>%
                    mutate(treatment = "cd3cd28u0126") %>% mutate (GP = 1)
df7_cd3cd28ly <- df7_cd3cd28ly %>%
                  mutate(treatment = "cd3cd28ly") %>% mutate (GP = 1)
df8_pma <- df8_pma %>%
            mutate(treatment = "pma") %>% mutate (GP = 1)
df9_b2camp <- df9_b2camp %>%
              mutate(treatment = "b2camp") %>% mutate (GP = 1)
df10_cd3cd28icam2aktinhib <-df10_cd3cd28icam2aktinhib %>%
                           mutate(treatment = "cd3cd28icam2aktinhib") %>% mutate (GP = 2)
df11_cd3cd28icam2g0076 <-df11_cd3cd28icam2g0076 %>%
                         mutate(treatment = "cd3cd28icam2g0076") %>% mutate (GP = 2)
df12_cd3cd28icam2psit <- df12_cd3cd28icam2psit %>%
                        mutate(treatment = "cd3cd28icam2psit") %>% mutate (GP = 2)
df13_cd3cd28icam2u0126 <- df13_cd3cd28icam2u0126 %>%
                         mutate(treatment = "cd3cd28icam2u0126") %>% mutate (GP = 2)
df14_cd3cd28icam2ly <- df14_cd3cd28icam2ly %>%
                        mutate(treatment = "cd3cd28icam2ly") %>% mutate (GP = 2)
```

New column for conditions

As mentioned before, measurments are obtained from two different perturbations (GP1 and GP2). At the same time, these perturbations are combined with different treatments. I added a new column called "condition" to each table to classify the data depending on the treatment applied : 0 <- no treatment 4 <- MEK_inh 1 <- Akt_inh1 5 <- Akt_inh2 2 <- PKC_inh 6 <- PKC_act 3 <- PIP2_inh 7 <- PKA_act

```
#Add new columns with conditions names
df1_cd3cd28 <- df1_cd3cd28 %>%
  mutate(treatment = "cd3cd28") %>% mutate (condition = 0)
df2_cd3cd28icam2 <- df2_cd3cd28icam2 %>%
  mutate(treatment = "cd3cd28icam2") %>% mutate (condition = 0)
df3_cd3cd28aktinhib <- df3_cd3cd28aktinhib %>%
  mutate(treatment = "cd3cd28aktinhib") %>% mutate (condition = 1)
df4_cd3cd28g0076 <- df4_cd3cd28g0076 %>%
  mutate(treatment = "cd3cd28g0076") %>% mutate (condition = 2)
df5_cd3cd28psitect <- df5_cd3cd28psitect %>%
  mutate(treatment = "cd3cd28psitect") %>% mutate (condition = 3)
df6_cd3cd28u0126 <- df6_cd3cd28u0126 %>%
  mutate(treatment = "cd3cd28u0126") %>% mutate (condition = 4)
df7_cd3cd28ly <- df7_cd3cd28ly %>%
  mutate(treatment = "cd3cd28ly") %>% mutate (condition = 5)
df8_pma <- df8_pma %>%
  mutate(treatment = "pma") %>% mutate (condition = 6)
df9_b2camp <- df9_b2camp %>%
  mutate(treatment = "b2camp") %>% mutate (condition = 7)
```

```
df10_cd3cd28icam2aktinhib <-df10_cd3cd28icam2aktinhib %>%
  mutate(treatment = "cd3cd28icam2aktinhib") %>% mutate (condition = 1)
df11_cd3cd28icam2g0076 <-df11_cd3cd28icam2g0076 %>%
  mutate(treatment = "cd3cd28icam2g0076") %>% mutate (condition = 2)
df12_cd3cd28icam2psit <- df12_cd3cd28icam2psit %>%
  mutate(treatment = "cd3cd28icam2psit") %>% mutate (condition = 3)
df13_cd3cd28icam2u0126 <- df13_cd3cd28icam2u0126 %>%
  mutate(treatment = "cd3cd28icam2u0126") %>% mutate (condition = 4)
df14_cd3cd28icam2ly <- df14_cd3cd28icam2ly %>%
  mutate(treatment = "cd3cd28icam2ly") %>% mutate (condition = 5)
```

Unique table

I created a unique table for all the perturbations/treatments.

```
alldf <- bind_rows(df1_cd3cd28, df2_cd3cd28icam2, df3_cd3cd28aktinhib, df4_cd3cd28g0076,     df5_cd3cd
```

Dummy variables for the different treatments

I created dummy variables for each of the treatments (independently of the perturbation)

```
alldf <- alldf %>%
          mutate (Akt_inh1 = ifelse (treatment == c("cd3cd28aktinhib", "cd3cd28icam2aktinhib"), 1, 0)) %
          mutate (PKC_inh = ifelse (treatment == c("cd3cd28g0076", "cd3cd28icam2g0076"), 1, 0)) %>%
          mutate (PIP2_inh = ifelse (treatment == c("cd3cd28psitect", "cd3cd28icam2psit"), 1, 0)) %>%
          mutate (MEK_inh = ifelse (treatment == c("cd3cd28u0126", "cd3cd28icam2u0126"), 1, 0)) %>%
          mutate (Akt_inh2 = ifelse (treatment == c("cd3cd28ly", "cd3cd28icam2ly"), 1, 0)) %>%
          mutate (PKC_act = ifelse (treatment == "pma", 1, 0)) %>%
          mutate (PKA_act = ifelse (treatment == "b2camp", 1, 0))
```

Reorder columns

I reorder the columns to have the treatment names and dummy variables first, and after that all the measurments done.

```
alldf <- alldf %>% select(treatment, GP, condition, Akt_inh1, PKC_inh, PIP2_inh, MEK_inh, Akt_inh2, PKC
```

Table visualization

```
head(alldf)
```

```
##   treatment GP condition Akt_inh1 PKC_inh PIP2_inh MEK_inh Akt_inh2
## 1   cd3cd28  1         0        0       0        0       0        0
## 2   cd3cd28  1         0        0       0        0       0        0
## 3   cd3cd28  1         0        0       0        0       0        0
## 4   cd3cd28  1         0        0       0        0       0        0
## 5   cd3cd28  1         0        0       0        0       0        0
## 6   cd3cd28  1         0        0       0        0       0        0
##   PKC_act PKA_act praf  pmek  plcg  PIP2  PIP3 p44.42 pakts473 PKA   PKC
## 1       0       0 26.4 13.20  8.82 18.30 58.80   6.61     17.0 414 17.00
```

```
## 2       0       0 35.9 16.50 12.30 16.80  8.13  18.60     32.5 352  3.37
## 3       0       0 59.4 44.10 14.60 10.20 13.00  14.90     32.5 403 11.40
## 4       0       0 73.0 82.80 23.10 13.50  1.29   5.83     11.8 528 13.70
## 5       0       0 33.7 19.80  5.19  9.73 24.80  21.10     46.1 305  4.66
## 6       0       0 18.8  3.75 17.60 22.10 10.90  11.90     25.7 610 13.70
##    P38 pjnk
## 1 44.9 40.0
## 2 16.5 61.5
## 3 31.9 19.5
## 4 28.6 23.1
## 5 25.7 81.3
## 6 49.1 57.8
```

Data subseting

I grouped the data to help for the data visualization

```
GP1 <- subset(alldf, GP == "1", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP2", "PI
GP2 <- subset(alldf, GP == "2", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP2", "PI
GP1_0<- subset(GP1, condition == "0", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP
GP1_1<- subset(GP1, condition == "1", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP
GP1_2<- subset(GP1, condition == "2", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP
GP1_3<- subset(GP1, condition == "3", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP
GP1_4<- subset(GP1, condition == "4", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP
GP1_5<- subset(GP1, condition == "5", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP
GP1_6<- subset(GP1, condition == "6", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP
GP1_7<- subset(GP1, condition == "7", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP
GP2_0<- subset(GP2, condition == "0", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP
GP2_1<- subset(GP2, condition == "1", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP
GP2_2<- subset(GP2, condition == "2", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP
GP2_3<- subset(GP2, condition == "3", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP
GP2_4<- subset(GP2, condition == "4", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP
GP2_5<- subset(GP2, condition == "5", select = c("treatment", "condition", "praf", "pmek", "plcg", "PIP
```

Data visualization

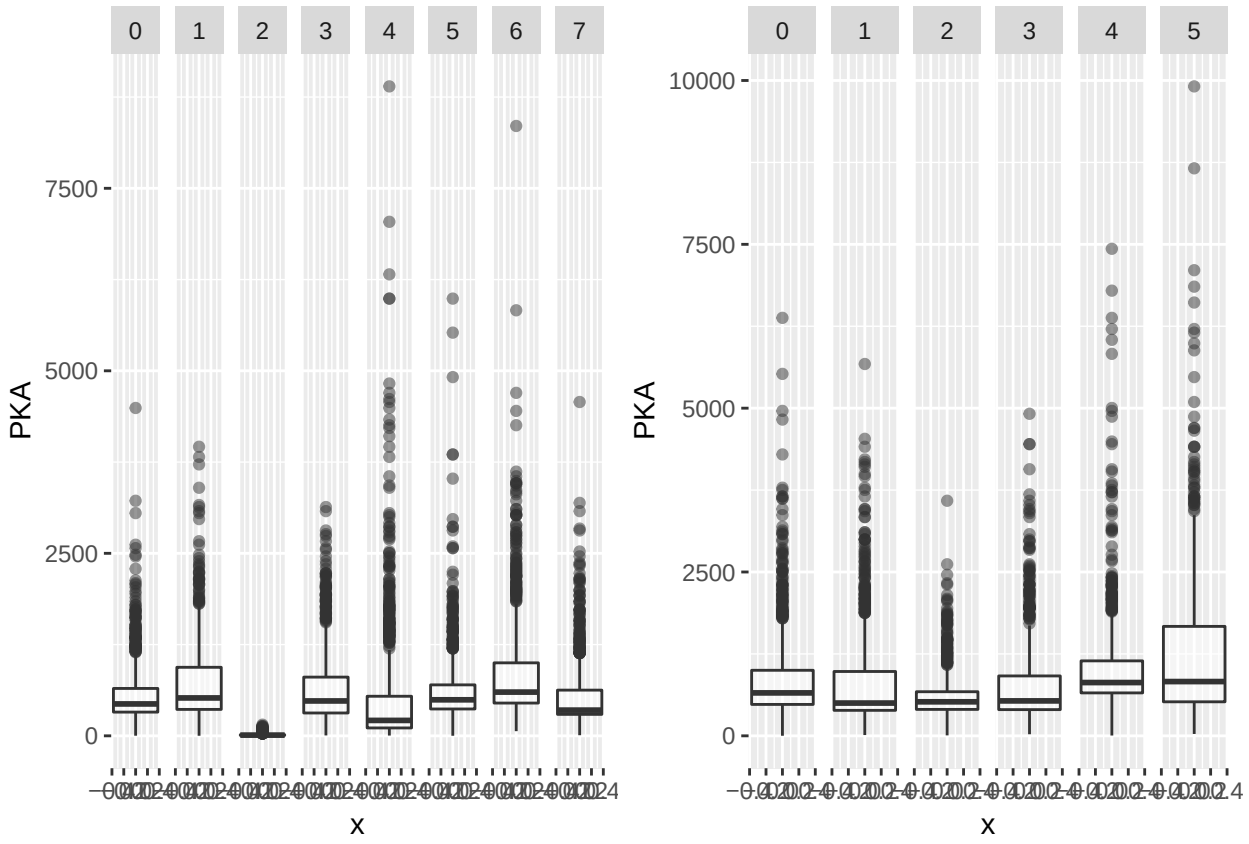
Representation of a certein output for the different perturbations and conditions

The idea would be to do the same graph for the different outputs. In this case I just show the graphs for PKA (I didn't play with X and Y axes labels, legends,... yet)

Option 1

I represent a graph for each perturbation (GP1 and GP2) and for each single treatment.

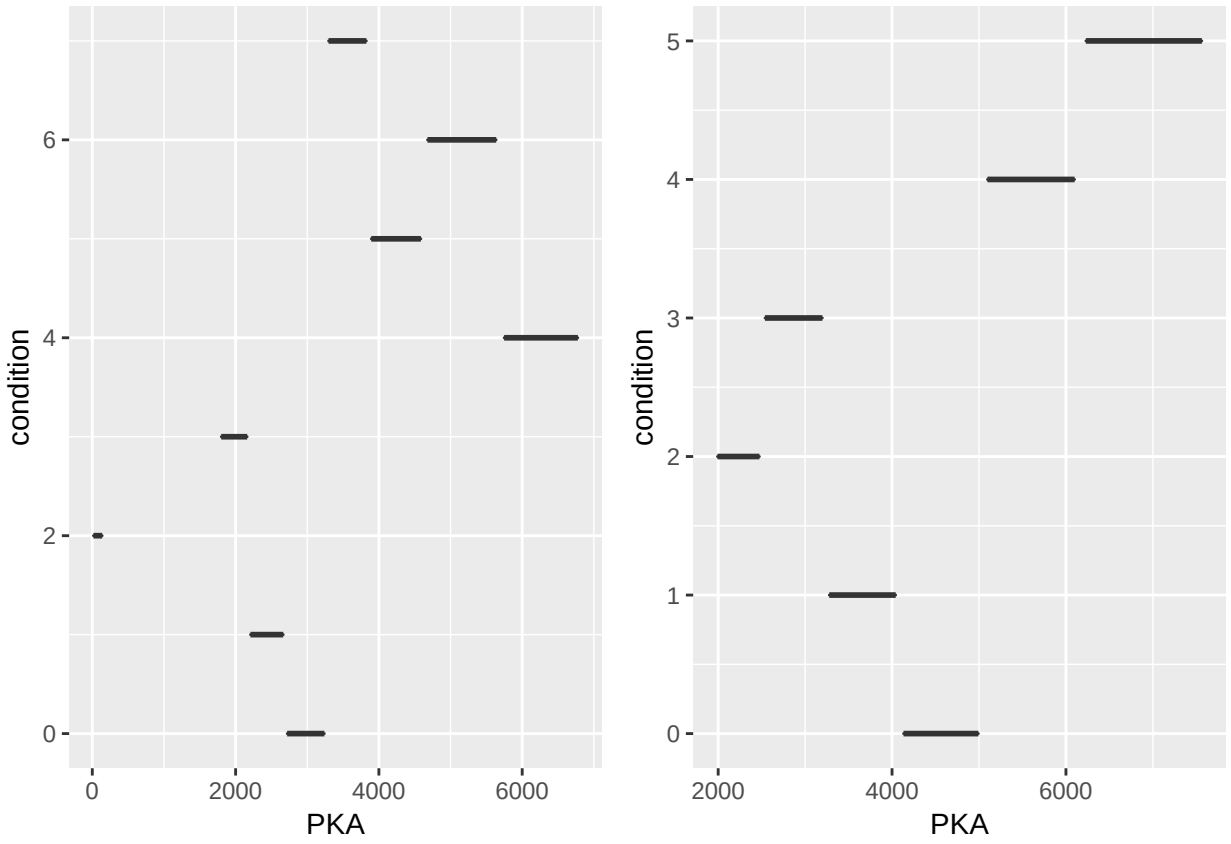
```
PKA1.1 <- ggplot(GP1, aes(x = 0, y = PKA)) +
  geom_boxplot(alpha = 0.5) +
  facet_grid(. ~ condition)

PKA2.1 <- ggplot(GP2, aes(x = 0, y = PKA)) +
  geom_boxplot(alpha = 0.5) +
  facet_grid(. ~ condition)
grid.arrange(PKA1.1, PKA2.1, nrow = 1)
```



Option 2

It's the same idea but with a different plot type.

```
PKA1.2 <- ggplot(GP1, aes(x = PKA, y = condition, group = condition)) +
  geom_boxplot(alpha = 0.5)
PKA2.2 <- ggplot(GP2, aes(x = PKA, y = condition, group = condition)) +
  geom_boxplot(alpha = 0.5)
grid.arrange(PKA1.2,PKA2.2, nrow = 1)
```

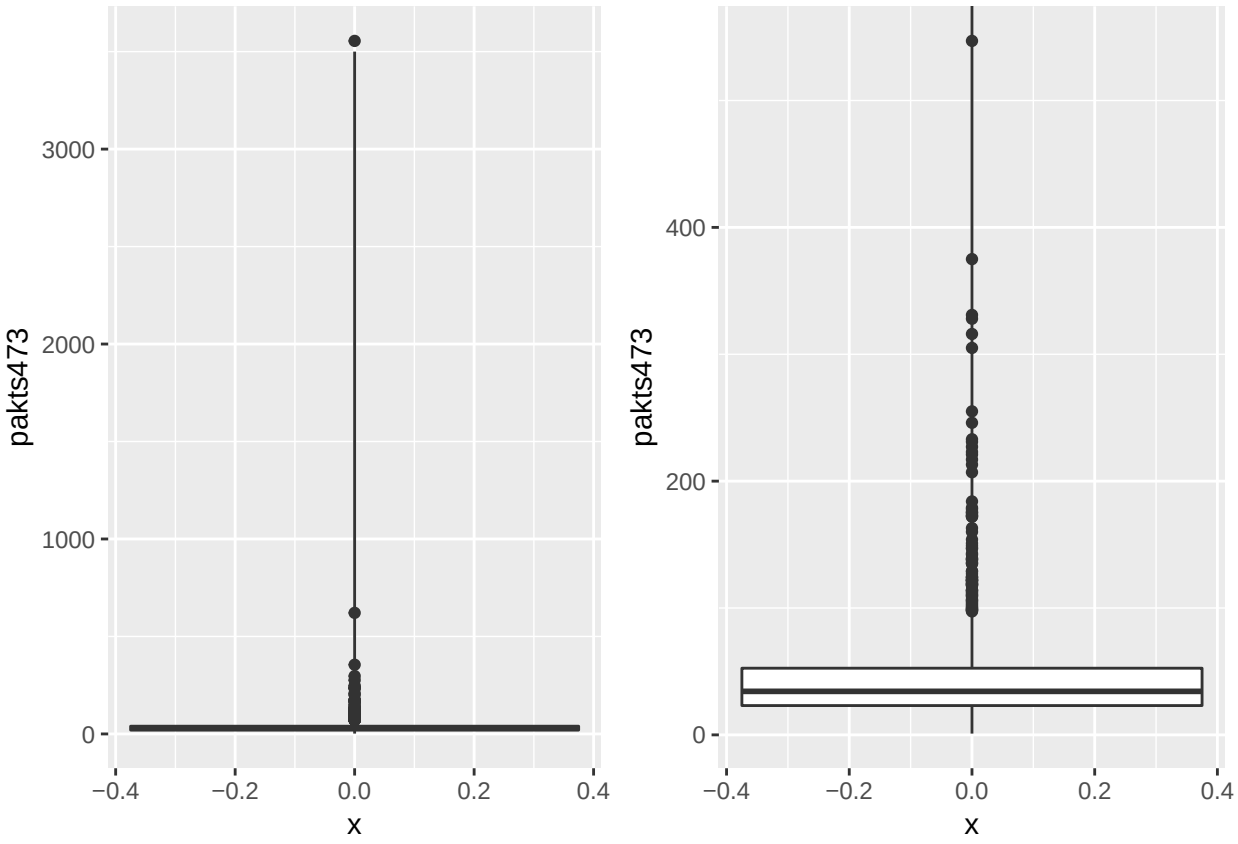


Some controls...

Comparing two treatments: control vs inhibition.

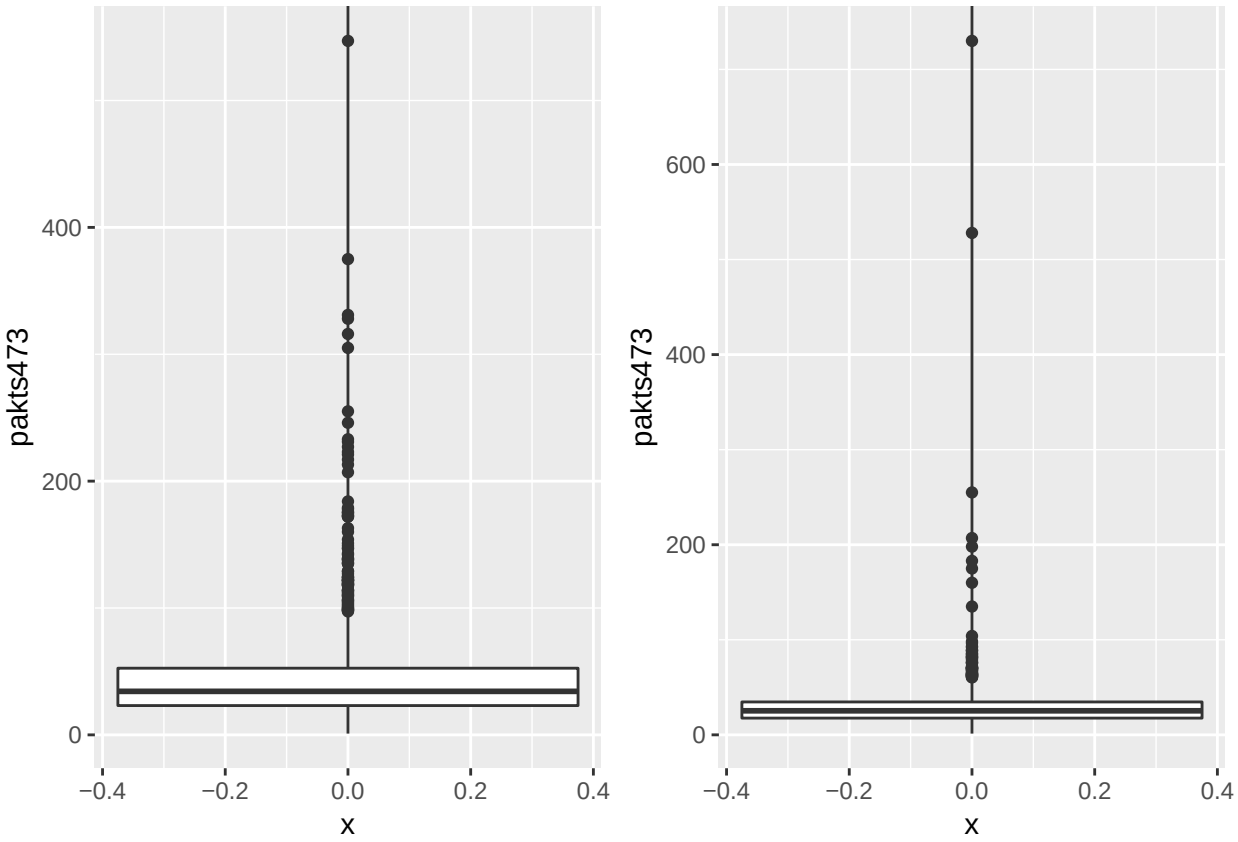
How do I combine both in one graph? How do you fix the Y axes values?

```
p1 <- ggplot(GP1_0, aes (x = 0, y = pakts473)) +
  geom_boxplot(ymax = 3500)
p2 <- ggplot(GP1_1, aes (x = 0, y = pakts473)) +
  geom_boxplot(ymax = 3500)
grid.arrange(p1,p2, nrow = 1)
```



Comparing two treatments that are supose to do the same

```
p1 <- ggplot(GP1_1, aes (x = 0, y = pakts473)) +
  geom_boxplot(ymax = 3500)
p5 <- ggplot(GP1_5, aes (x = 0, y = pakts473)) +
  geom_boxplot(ymax = 3500)
grid.arrange(p1,p5, nrow = 1)
```



It would be nice to have a correlation between the two treatments

Save table

I saved the tabble as .csv document (capstone_project.csv)

```
write.table(alldf, file = "capstone_project.csv", sep = ",", col.names = NA)
```