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)
library(corrplot)
library(GGally)
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_cd3cd2
```

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, everything())
```

Table visualization

```
head(alldf)
```

```
##   treatment GP condition praf  pmek  plcg  PIP2  PIP3 p44.42 pakts473 PKA
## 1   cd3cd28  1         0 26.4 13.20  8.82 18.30 58.80   6.61     17.0 414
## 2   cd3cd28  1         0 35.9 16.50 12.30 16.80  8.13  18.60     32.5 352
## 3   cd3cd28  1         0 59.4 44.10 14.60 10.20 13.00  14.90     32.5 403
## 4   cd3cd28  1         0 73.0 82.80 23.10 13.50  1.29   5.83     11.8 528
## 5   cd3cd28  1         0 33.7 19.80  5.19  9.73 24.80  21.10     46.1 305
## 6   cd3cd28  1         0 18.8  3.75 17.60 22.10 10.90  11.90     25.7 610
##     PKC  P38 pjnk
## 1 17.00 44.9 40.0
## 2  3.37 16.5 61.5
## 3 11.40 31.9 19.5
## 4 13.70 28.6 23.1
## 5  4.66 25.7 81.3
## 6 13.70 49.1 57.8
```

Data subseting

I grouped the data to help in 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
```

Summarise the data to see overall trends

stats_GP1 <- GP1 %>% group_by(condition) %>% summarise_at(vars(praf:pjnk), list(mean = mean, sd = sd), na.rm = TRUE) head(stats) stats_GP2 <- GP2 %>% group_by(condition) %>% summarise_at(vars(praf:pjnk), mean, na.rm = TRUE) head(stats)

Data visualization

Here, I put just few examples. The idea would be to do similar graphs for the different variables.

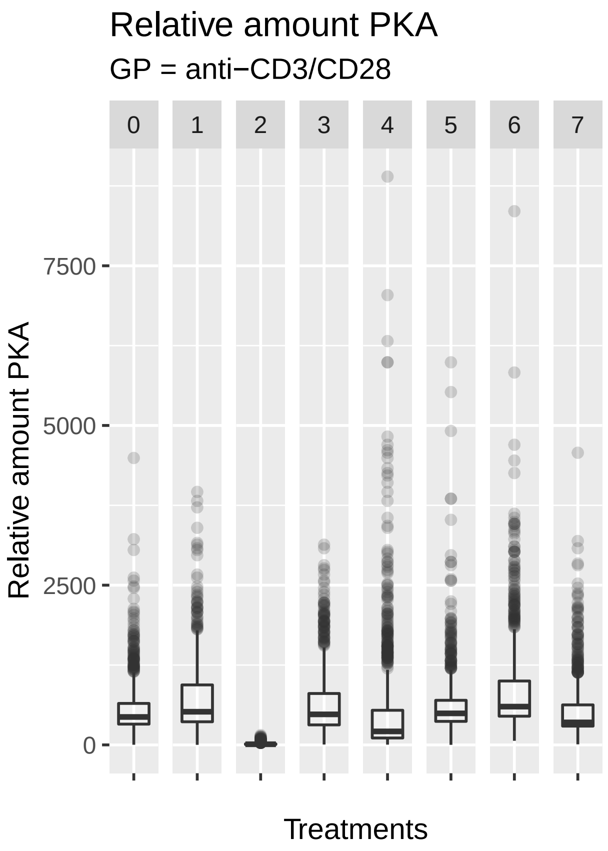
Box-plots per protein (PKA, PKC, p38, JNK,. . .)

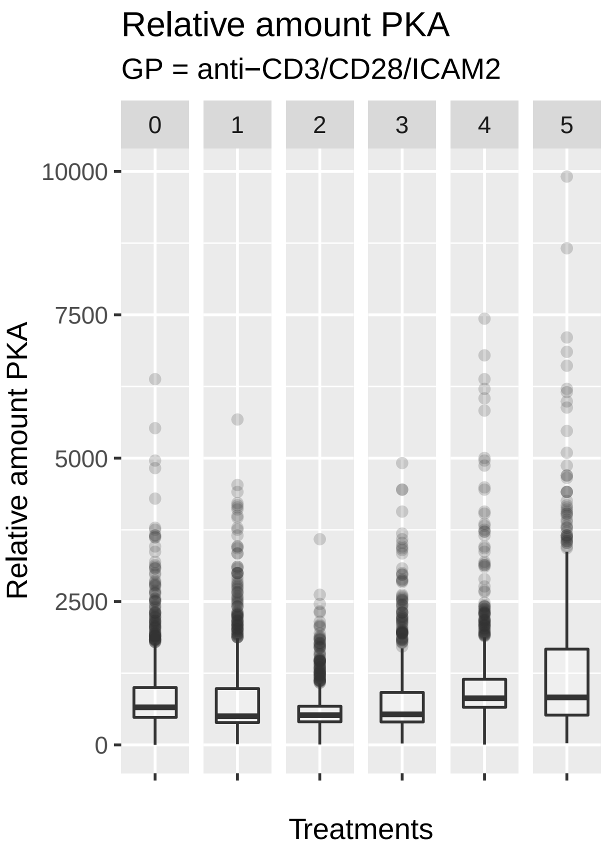
I represented a graph for each perturbation (GP1 and GP2) and for each single treatment (in this case only for PKA).

```
PKA_GP1 <-ggplot(GP1, aes(x = "", y = PKA)) +
    geom_boxplot(aes(), alpha=0.2) +
    facet_grid(.~condition) +
    labs(title="Relative amount PKA",
         subtitle="GP = anti-CD3/CD28",
         x="Treatments",
         y = "Relative amount PKA")

PKA_GP2 <-ggplot(GP2, aes(x = "", y = PKA)) +
  geom_boxplot(aes(), alpha=0.2) +
  facet_grid(.~condition) +
  labs(title="Relative amount PKA",
       subtitle="GP = anti-CD3/CD28/ICAM2",
       x="Treatments",
       y = "Relative amount PKA")

grid.arrange(PKA_GP1, PKA_GP2, nrow = 1)
```



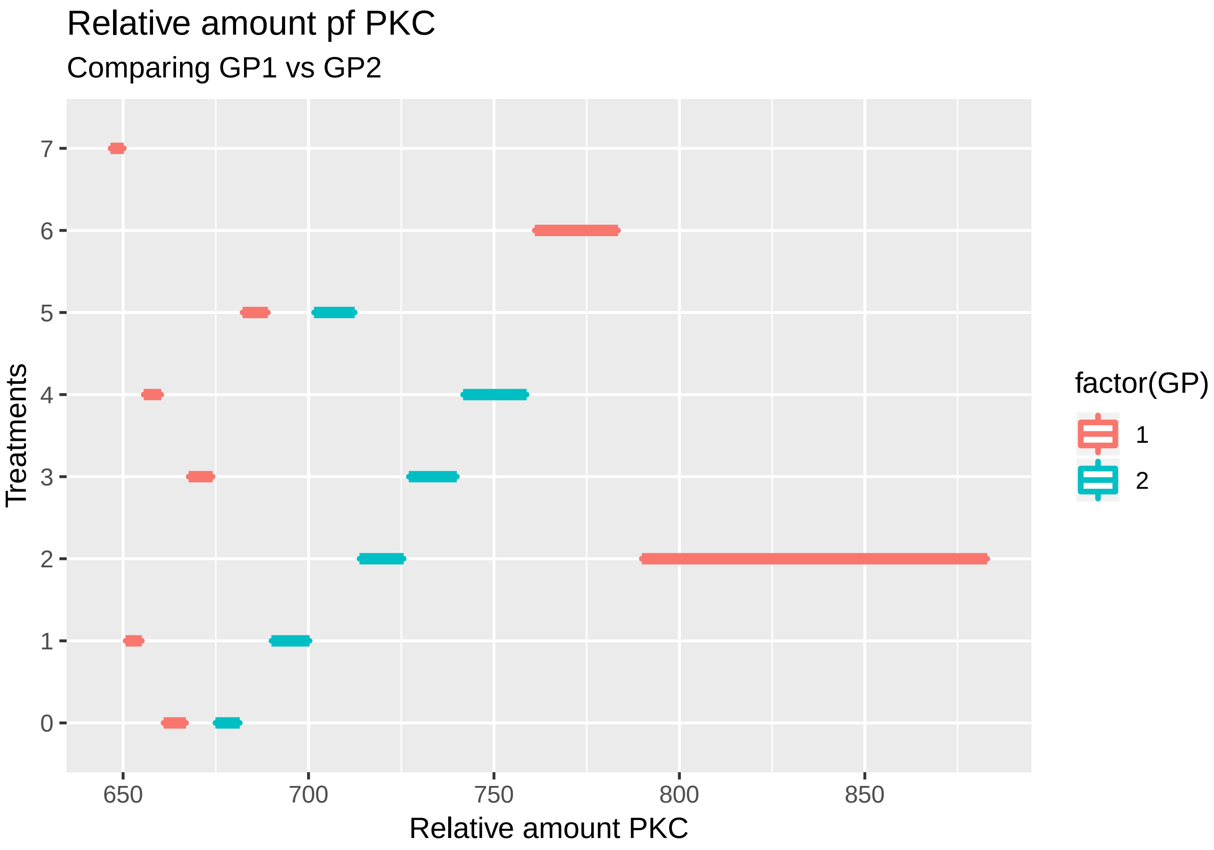


Comparing perturbations/conditions for PKC

Why are values in X different for Option1, 2 and 3?

Option1

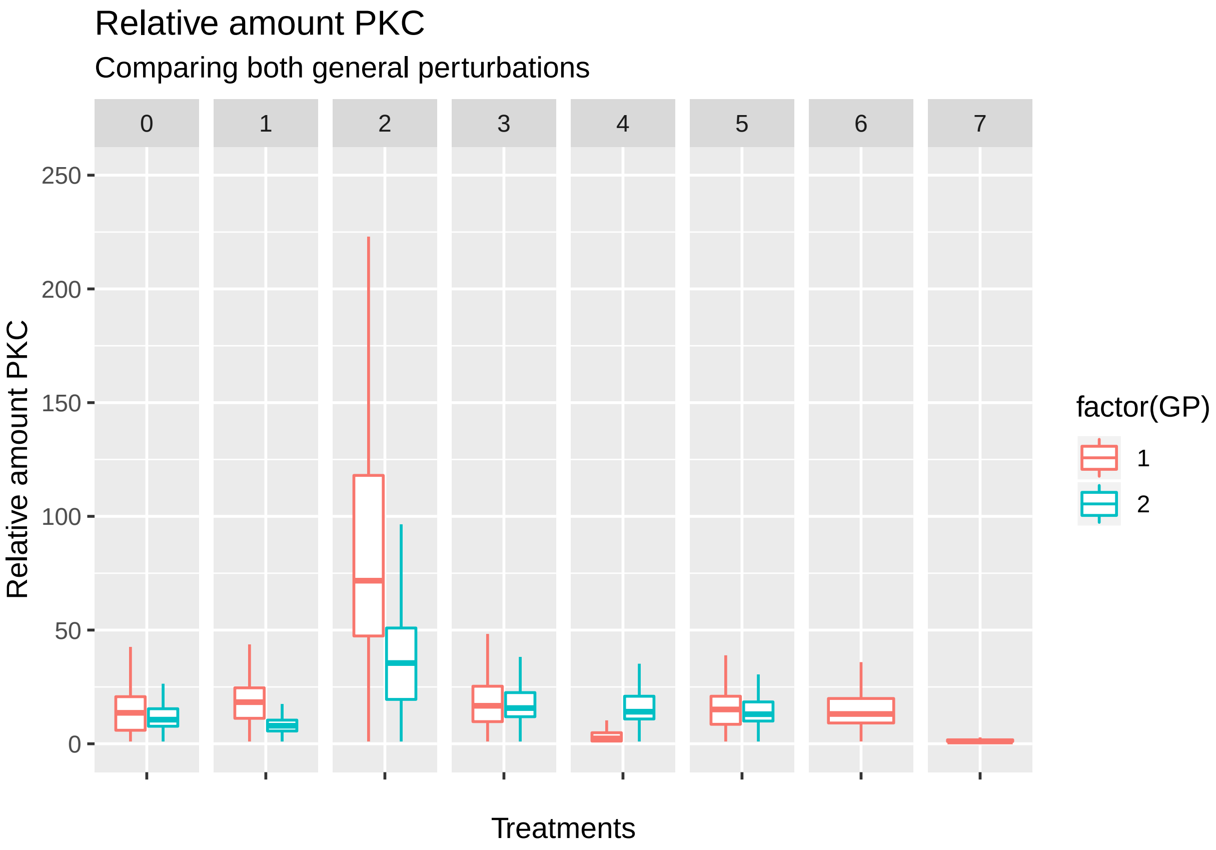
```
ggplot(alldf, aes(x = PKC, y = factor(condition), color = factor(GP))) +
  geom_boxplot(size = 1) +
  labs(title="Relative amount pf PKC",
       subtitle="Comparing GP1 vs GP2",
       x="Relative amount PKC",
       y = "Treatments")
```



Option 2

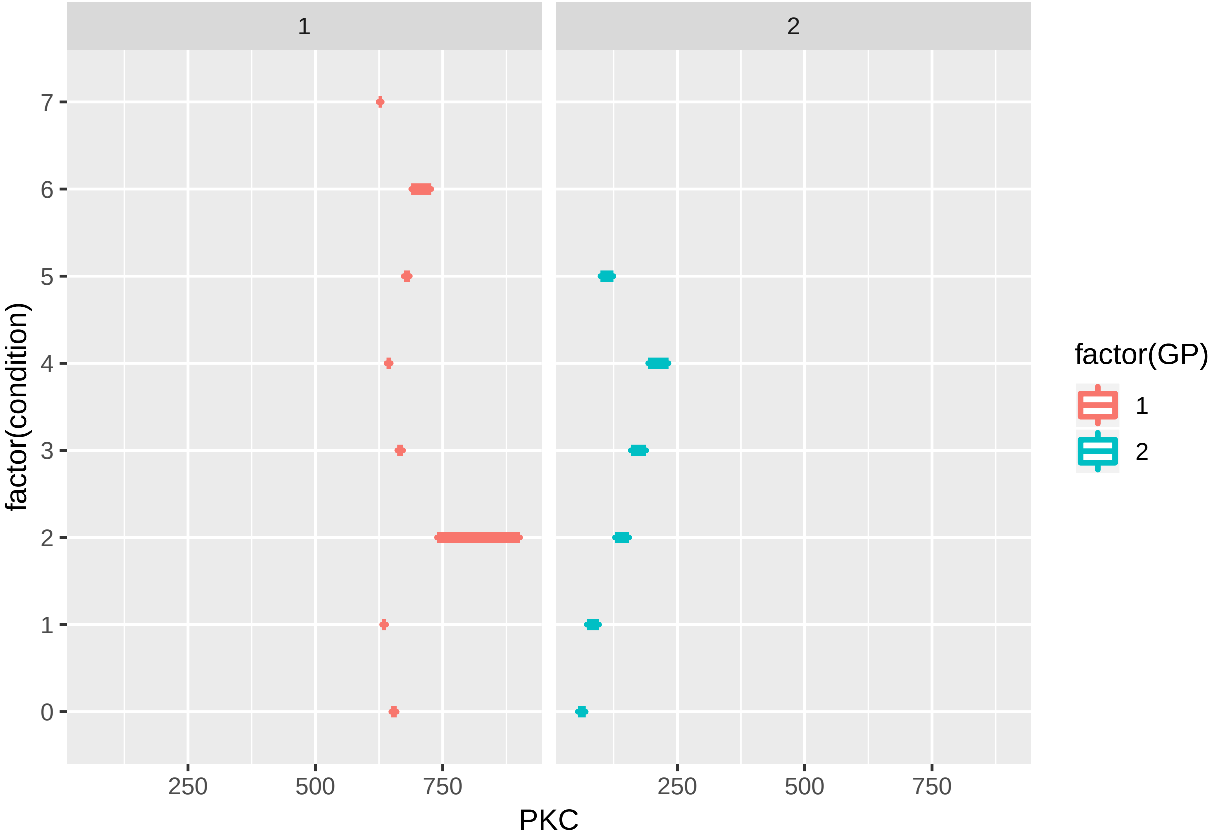
```
ggplot(alldf, aes(x = "", y = PKC, color = factor(GP))) +
  geom_boxplot(outlier.shape=NA) +
  ylim(0, 250) +
  facet_grid(. ~ condition) +
  labs(title="Relative amount PKC",
       subtitle="Comparing both general perturbations",
       x="Treatments",
       y = "Relative amount PKC")
```

```
## Warning: Removed 126 rows containing non-finite values (stat_boxplot).
```



Option 3

```
ggplot(alldf, aes(x = PKC, y = factor(condition), color = factor(GP))) +
  geom_boxplot(size = 1) +
  facet_grid(. ~ GP)
```

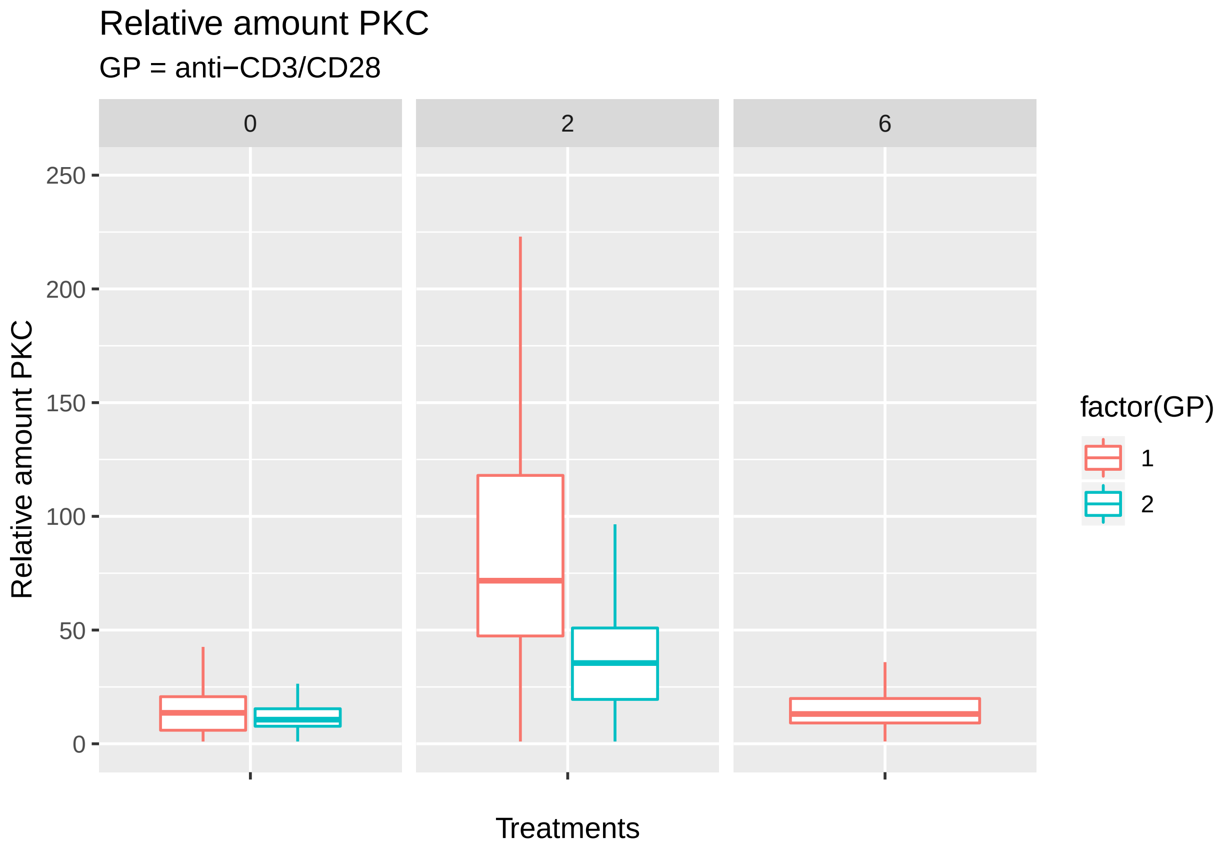


Checking activation vs inhibition

PKC control (cond = 0), activated (cond = 6) and inhibited (cond = 2)

```
PKC_actvsinh <- alldf %>% filter(condition %in% c("0", "2", "6"))
ggplot(PKC_actvsinh, aes(x = "", y = PKC, color = factor(GP))) +
  geom_boxplot(outlier.shape=NA ) +
  ylim(0, 250) +
  facet_grid(.~condition) +
  labs(title="Relative amount PKC",
       subtitle="GP = anti-CD3/CD28",
       x="Treatments",
       y = "Relative amount PKC")
```

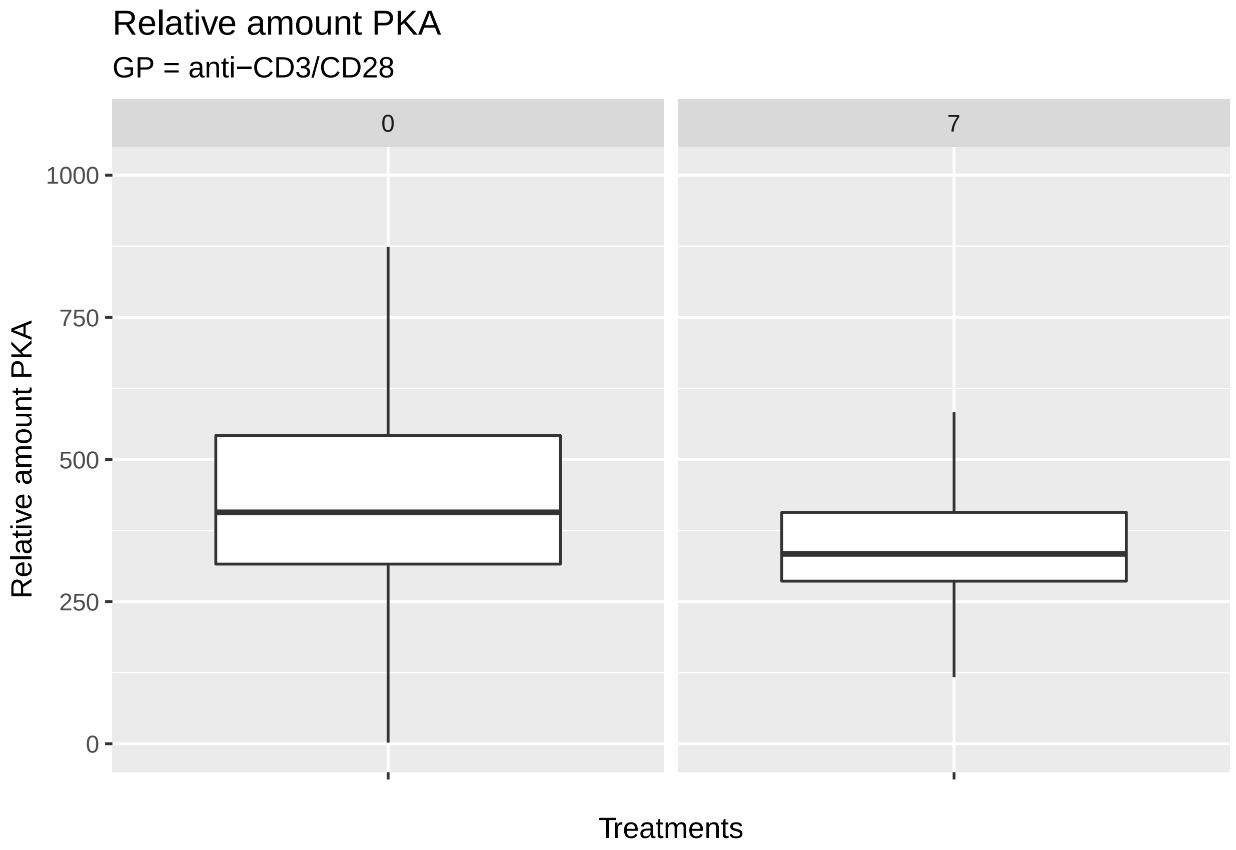
```
## Warning: Removed 124 rows containing non-finite values (stat_boxplot).
```



PKA control (cond = 0) vs activated (cond = 7)

```
PKA_act <- GP1 %>% filter(condition %in% c("0", "7"))
ggplot(PKA_act, aes(x = "", y = PKA)) +
  geom_boxplot(outlier.shape=NA ) +
  ylim(0, 1000) +
  facet_grid(.~condition) +
  labs(title="Relative amount PKA",
       subtitle="GP = anti-CD3/CD28",
       x="Treatments",
       y = "Relative amount PKA")
```

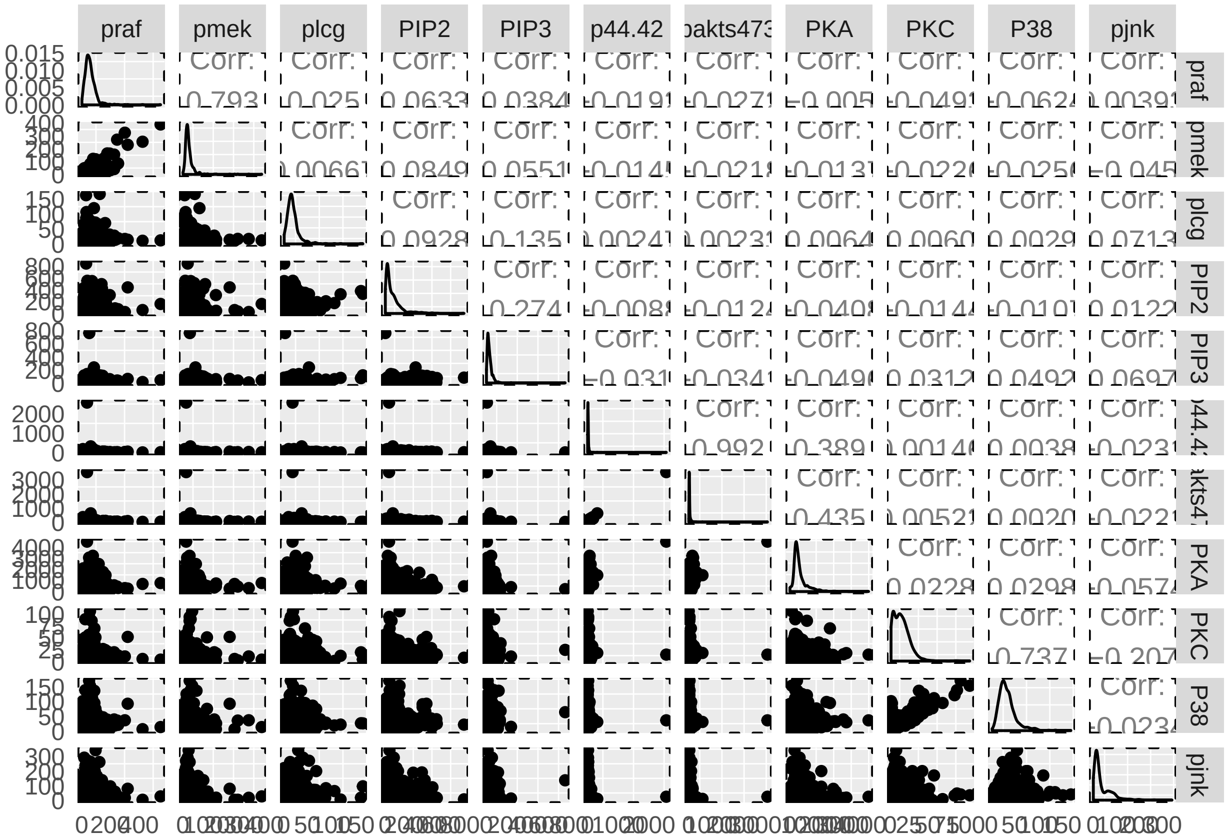
```
## Warning: Removed 210 rows containing non-finite values (stat_boxplot).
```



Looking at some correlations

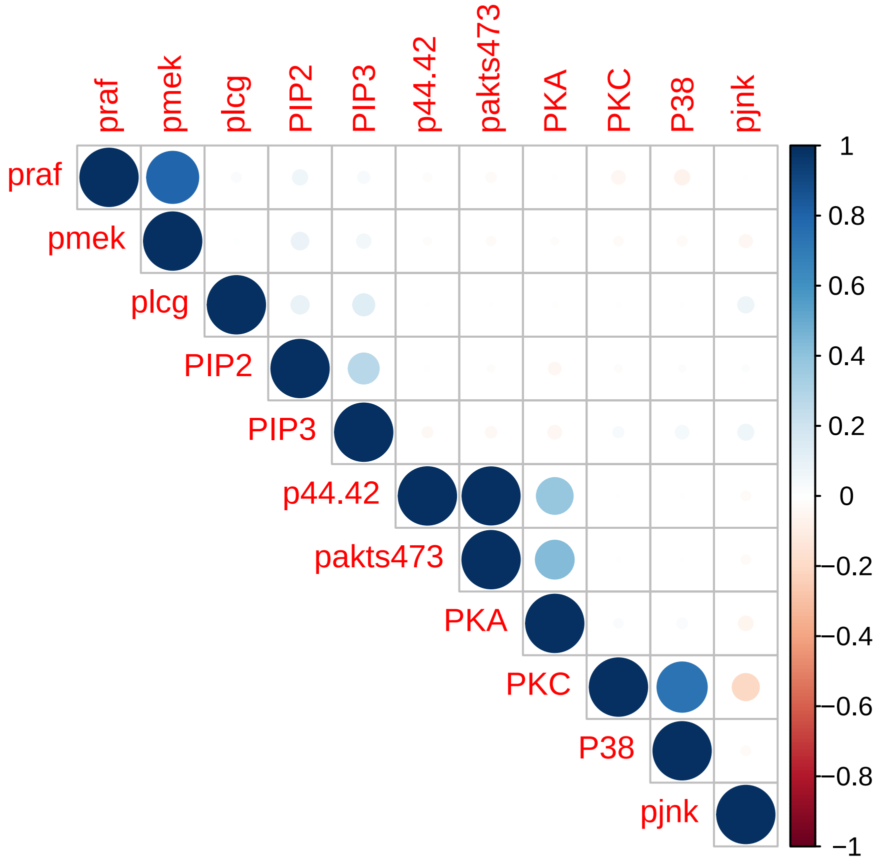
Option1

```
CorrPlot <- GP1 %>% filter(condition == "0")
ggpairs(CorrPlot, columns = 3:ncol(CorrPlot))+
  theme(legend.position = "none",
        panel.grid.major = element_blank(),
        axis.ticks = element_blank(),
        panel.border = element_rect(linetype = "dashed", colour = "black", fill = NA))
```

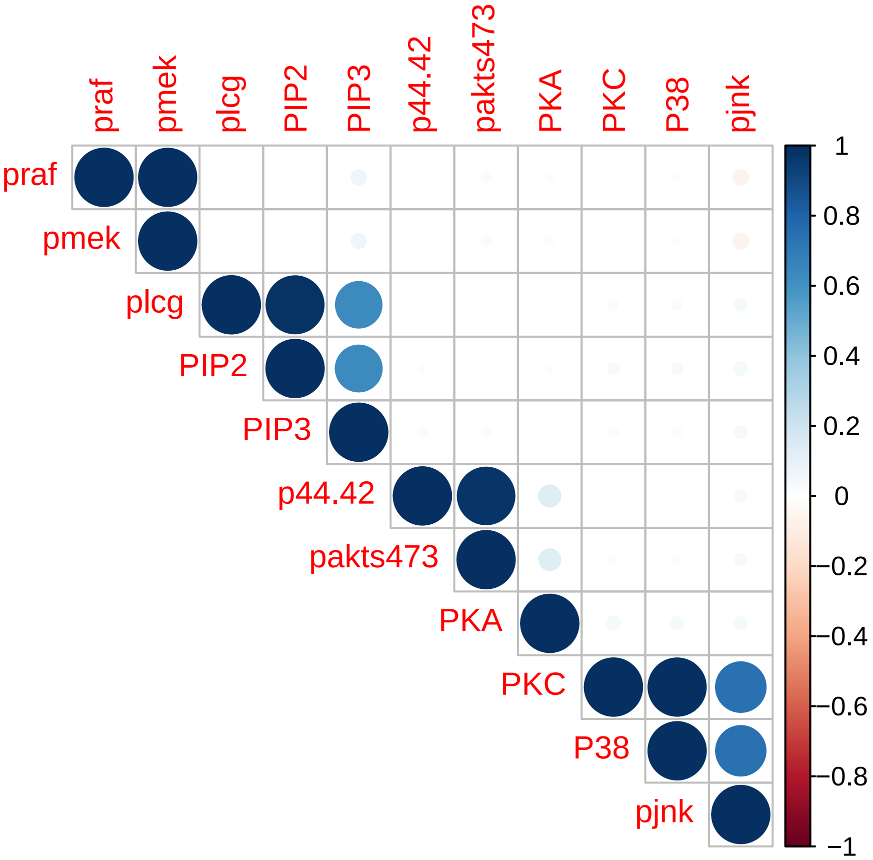


Option2

```
CM_0 <- GP1 %>% filter(condition == "0") %>% select(-treatment, -condition) %>% cor()
corrplot(CM_0, type="upper")
```

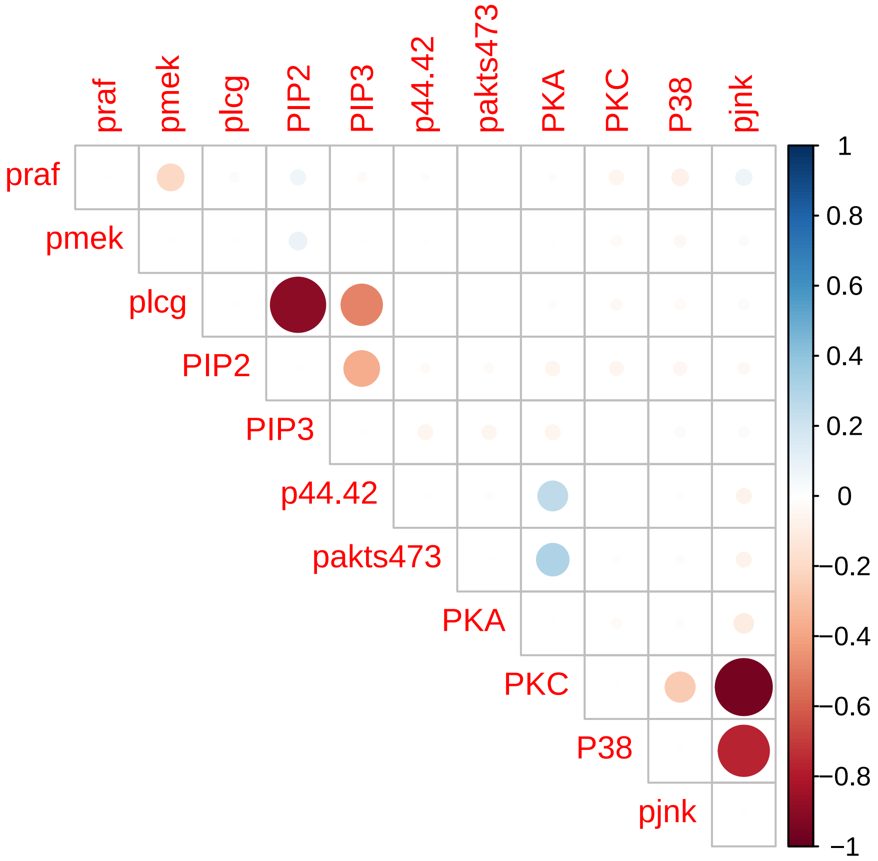


```
CM_2 <- GP1 %>% filter(condition == "2") %>% select(-treatment, -condition) %>% cor()
corrplot(CM_2, type="upper")
```



How do the correlations change between different conditions?

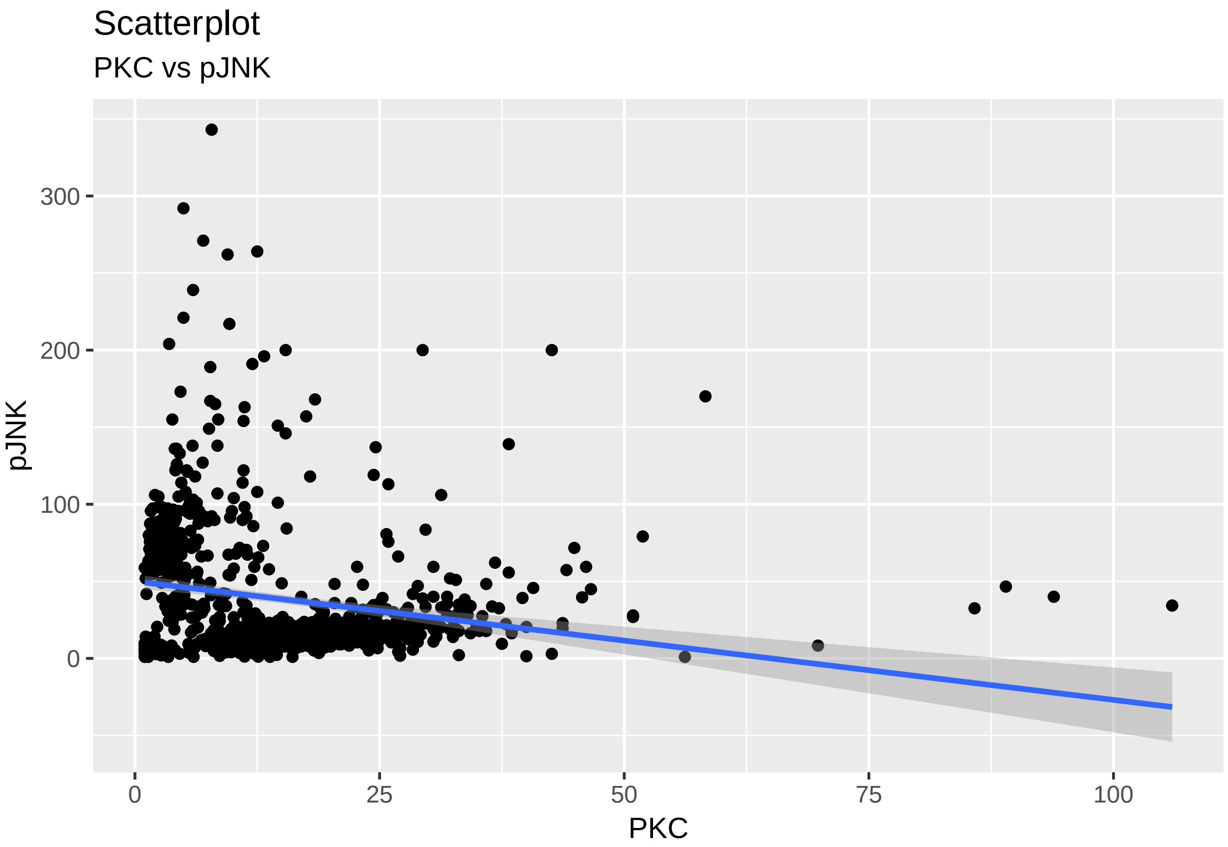
```
CM2m0 <- CM_0 - CM_2
corrplot(CM2m0, type="upper")
```



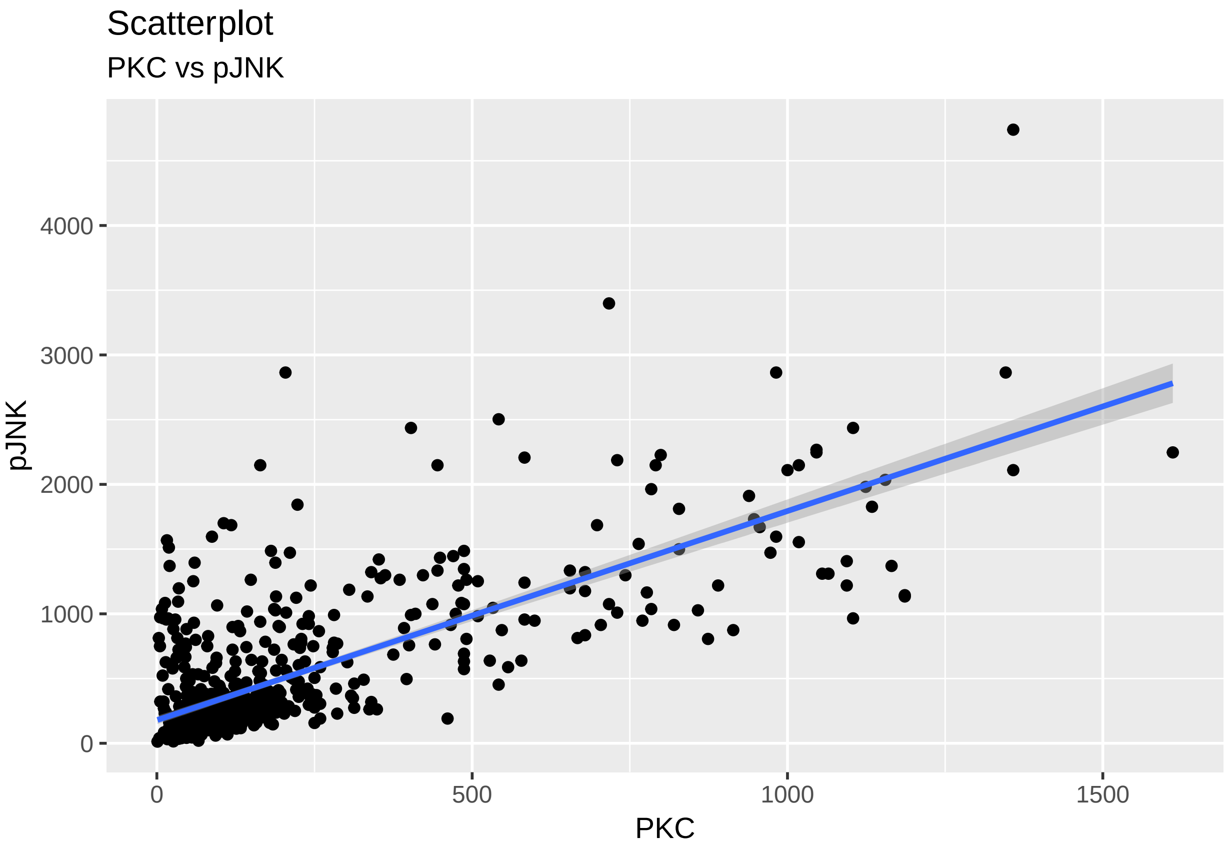
Some examples of good and bad correlation

PKC vs pJNK in cond = 0 and cond =2

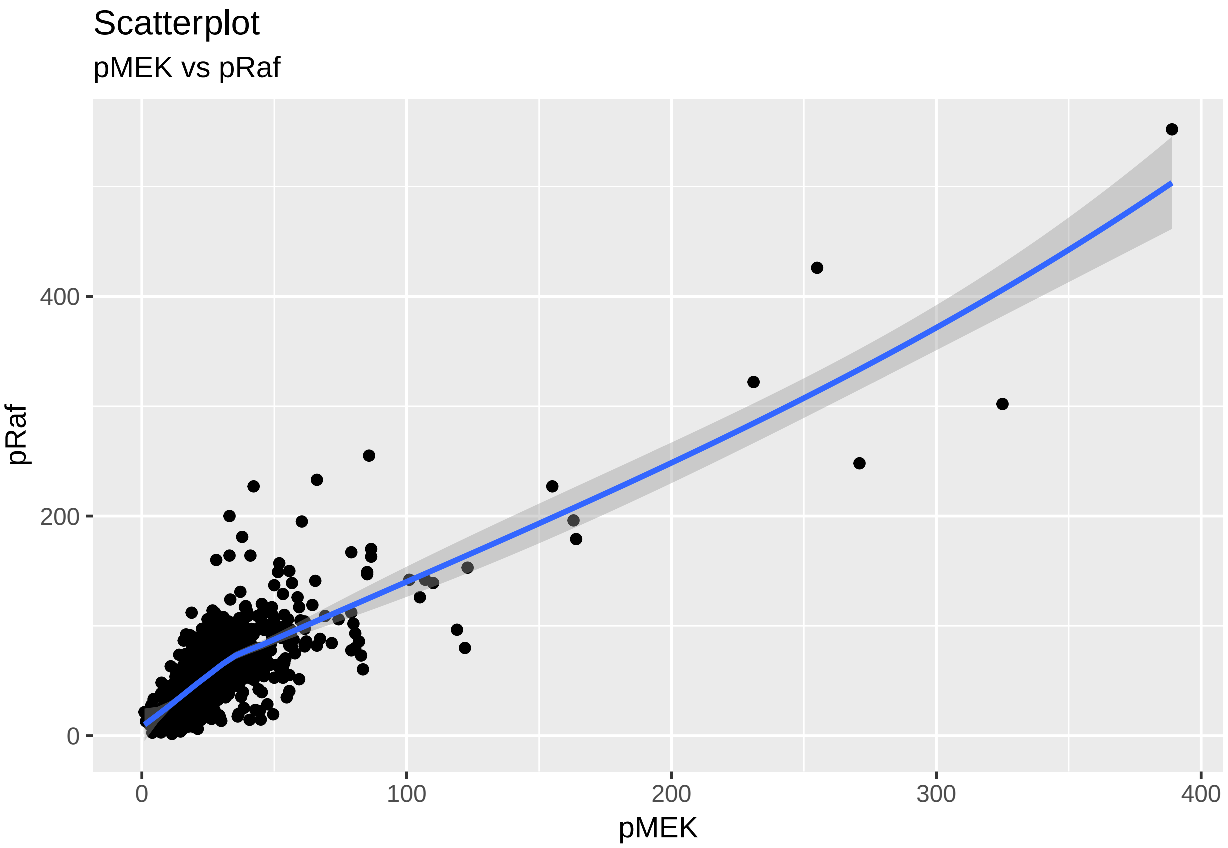
```
PKCvspJNK_0 <- GP1 %>% filter(condition == "0")
ggplot(PKCvspJNK_0, aes(x=PKC, y=pjnk)) +
  geom_point() +
  geom_smooth(method="lm") +
  labs(subtitle="PKC vs pJNK",
       y="pJNK",
       x="PKC",
       title="Scatterplot")
```



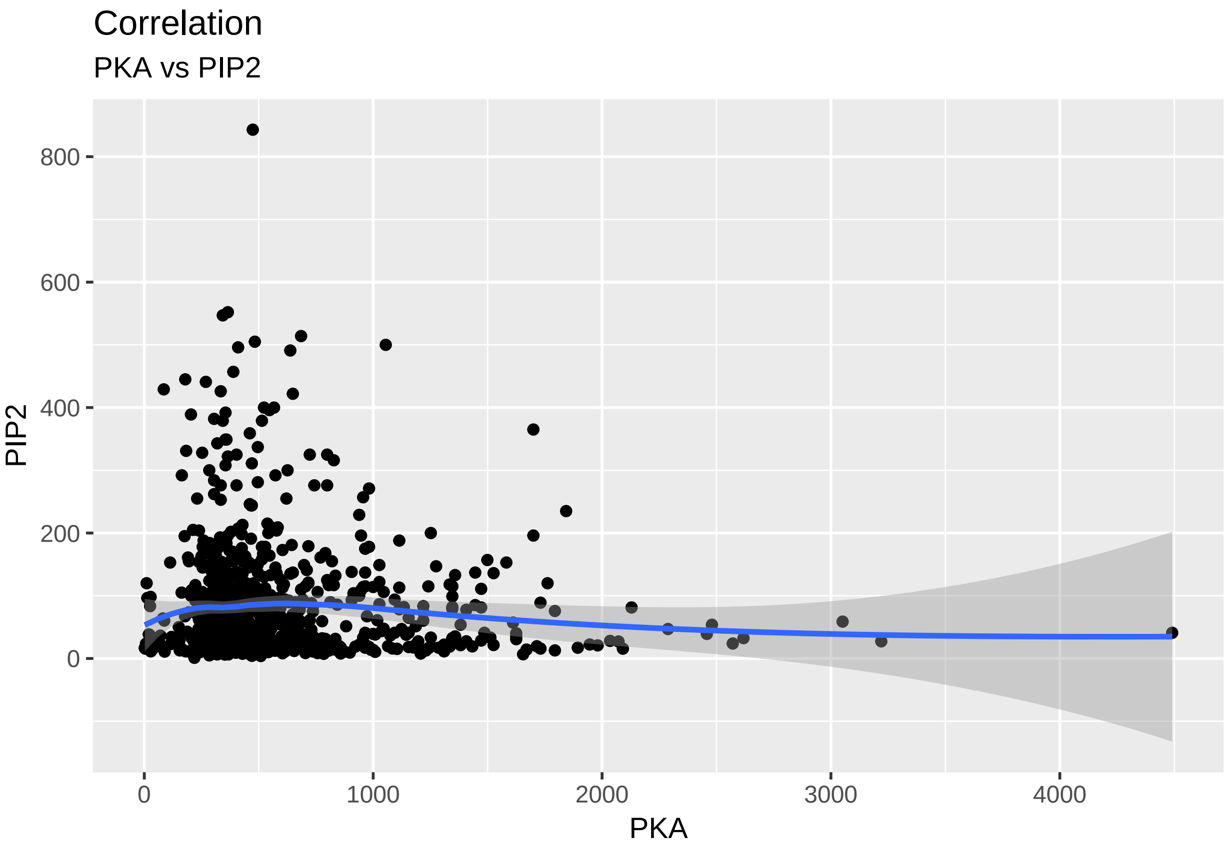
```
PKCvspJNK_2 <- GP1 %>% filter(condition == "2")
ggplot(PKCvspJNK_2, aes(x=PKC, y=pjnk)) +
  geom_point() +
  geom_smooth(method="lm") +
  labs(subtitle="PKC vs pJNK",
       y="pJNK",
       x="PKC",
       title="Scatterplot")
```



```
pmekvspraf <- GP1 %>% filter(condition == "0")
ggplot(pmekvspraf, aes(x=pmek, y=praf)) +
  geom_point() +
  geom_smooth(method="loess") +
  labs(subtitle="pMEK vs pRaf",
       y="pRaf",
       x="pMEK",
       title="Scatterplot")
```



```r
PKAvsPIP2 <- GP1 %>% filter(condition == "0")
ggplot(PKAvsPIP2, aes(x=PKA, y=PIP2)) +
  geom_point() +
  geom_smooth(method="loess") +
  labs(subtitle="PKA vs PIP2",
       y="PIP2",
       x="PKA",
       title="Correlation")
```



Save table

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

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