Correlation PDF

Vinod Kumar Ahuja

November 21, 2017

Correlation between Closed issues, number of commits, number of watchers and time it takes to close an issue over a period of time

```
# Loading Libraries
library(RODBC) # for database connection
library(sqldf) # for sql query
library(tidyverse) # for joins
library(ggpubr) # for correlation plotting
library(PerformanceAnalytics) # for correlation plotting
library(corrplot) # for corrplot
library(GGally) # for ggcorr


# Connecting to GHData MySQL database
connect <- odbcConnect("ghtorrent")

#Select Project
Project_Name <- "rust"
Project_Owner <- "rust-lang"

#Finding Project ID from GHTorrent data
project_id <- sqlQuery(connect, paste("SELECT projects.id FROM projects INNER JOIN users ON projects.ow



##Commits ##
# Selecting number of commits
query <- sqlQuery(connect, paste("select id as ID, created_at as 'Date' from commits where project_id =

# Sorting by Date
number_of_commits <- query[order(query$Date),]

# Correcting date format
number_of_commits$Date <- as.Date(number_of_commits$Date, "%Y-%m-%d")

# Counting number of commits
number_of_commits$Commit_Count <- seq.int(nrow(number_of_commits))

# selecting maximum values and one date from the repeated dates
selected_commits <- sqldf(' select Date, Commit_Count from number_of_commits Group by Date')




## Closed Issues ##
# Selecting number of closed issues
```

```
query <- sqlQuery(connect, paste("SELECT issue_events.event_id as 'ID', issue_events.action as 'Issue_s

# Sorting by Date
Closed_Issues <- query[order(query$Date),]

# Correcting date format
Closed_Issues$Date <- as.Date(Closed_Issues$Date, "%Y-%m-%d")

# Counting number of commits
Closed_Issues$Issues_Count <- seq.int(nrow(Closed_Issues))

# selecting maximum values and one date from the repeated dates
selected_closed_issues <- sqldf(' select Date, Issues_Count from Closed_Issues Group by Date')




# Watchers
# Selecting number of watchers
query <- sqlQuery(connect, paste("select watchers.repo_id , watchers.user_id, watchers.created_at as Da

# Sorting by Date
number_of_watchers <- query[order(query$Date),]

# Correcting date format
number_of_watchers$Date <- as.Date(number_of_watchers$Date, "%Y-%m-%d")

# Counting number of watchers
number_of_watchers$watch_Count <- seq.int(nrow(number_of_watchers))

# selecting maximum values and one date from the repeated dates
selected_watchers <- sqldf(' select Date, watch_Count from number_of_watchers Group by Date')



## Time takes to close an Issue ##
# Selecting issues and closure time
query <- sqlQuery(connect, paste("SELECT issues.created_at as 'Date', DATEDIFF(closed.created_at, issues

# Sorting by Date
time_to_close_issue <- query[order(query$Date),]

# Correcting date format
time_to_close_issue$Date <- as.Date(time_to_close_issue$Date, "%Y-%m-%d")

# Converting days_to_close to int
time_to_close_issue$days_to_close <- as.integer(time_to_close_issue$days_to_close)


# selecting maximum values and one date from the repeated dates
selected_time_to_close_issues <- sqldf(' select * from time_to_close_issue Group by Date')
```

```r
# Merging #
# Merge all the variables
# join for multiple tables
join <- full_join(selected_closed_issues,selected_commits, by='Date')%>%
  full_join(.,selected_watchers, by='Date')%>%
  full_join(.,selected_time_to_close_issues, by='Date')

# Sort join by date
join <- join[order(join$Date),]


# Drop Missing values
join <- na.omit(join)

# Converting to time series
join<- xts(join[, -1], order.by=as.POSIXct(join$Date))



# Correlation Matrix
cor(join, method = "pearson", use = "complete.obs")
```

```
##               Issues_Count Commit_Count watch_Count days_to_close
## Issues_Count     1.0000000    0.9976144   0.9962787    -0.2183982
## Commit_Count     0.9976144    1.0000000   0.9920878    -0.2181532
## watch_Count      0.9962787    0.9920878   1.0000000    -0.2237032
## days_to_close   -0.2183982   -0.2181532  -0.2237032     1.0000000
```

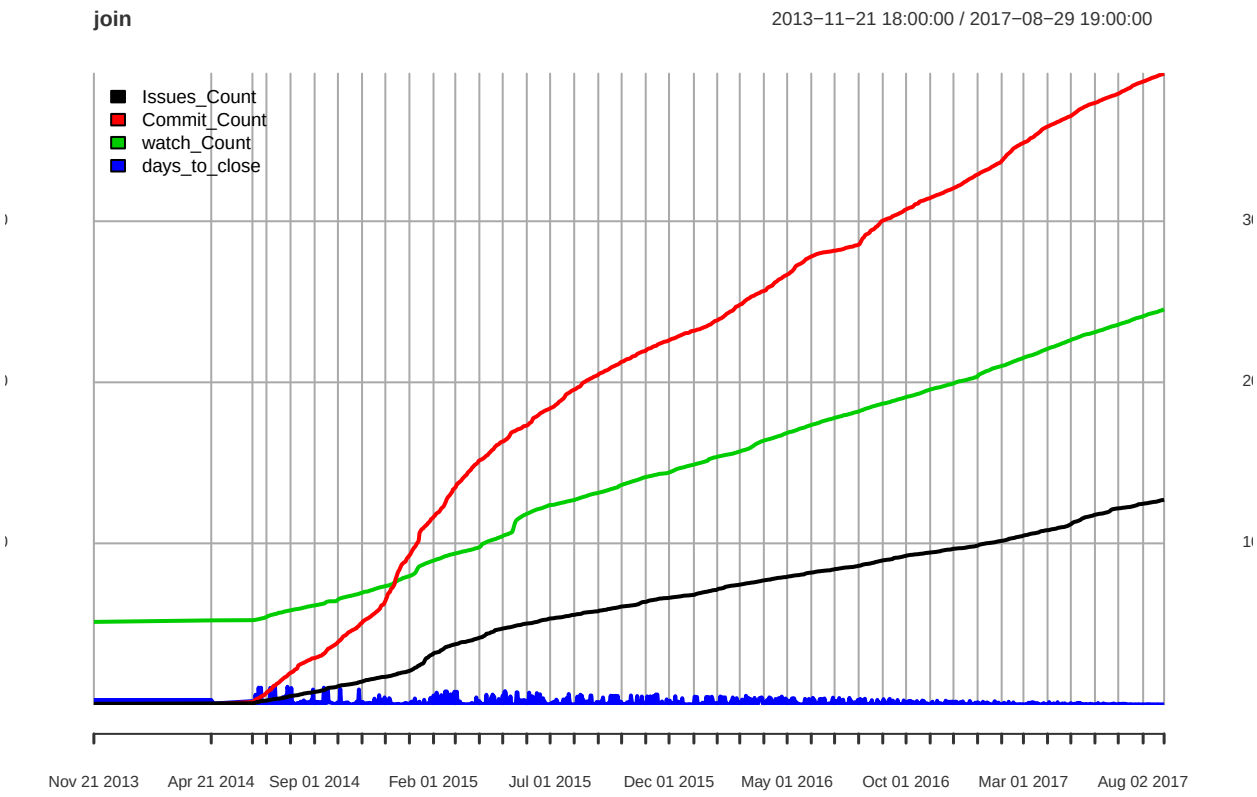
```r
cor(join, method = "kendall", use = "complete.obs")
```

```
##               Issues_Count Commit_Count watch_Count days_to_close
## Issues_Count     1.0000000    1.0000000   1.0000000    -0.1233691
## Commit_Count     1.0000000    1.0000000   1.0000000    -0.1233691
## watch_Count      1.0000000    1.0000000   1.0000000    -0.1233691
## days_to_close   -0.1233691   -0.1233691  -0.1233691     1.0000000
```

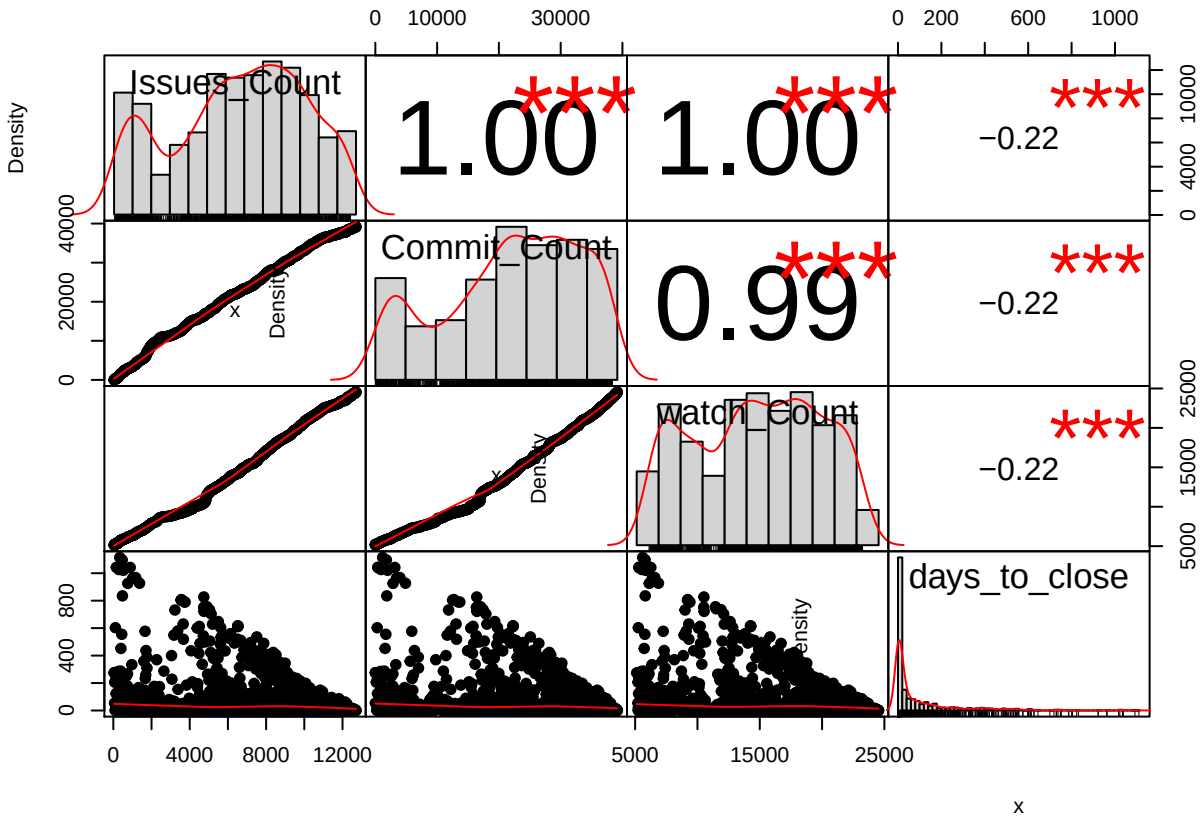
```r
cor(join, method = "spearman", use = "complete.obs")
```

```
##               Issues_Count Commit_Count watch_Count days_to_close
## Issues_Count     1.0000000    1.0000000   1.0000000    -0.1735745
## Commit_Count     1.0000000    1.0000000   1.0000000    -0.1735745
## watch_Count      1.0000000    1.0000000   1.0000000    -0.1735745
## days_to_close   -0.1735745   -0.1735745  -0.1735745     1.0000000
```

```r
# Initial plot
plot(join, legend.loc = T)
```

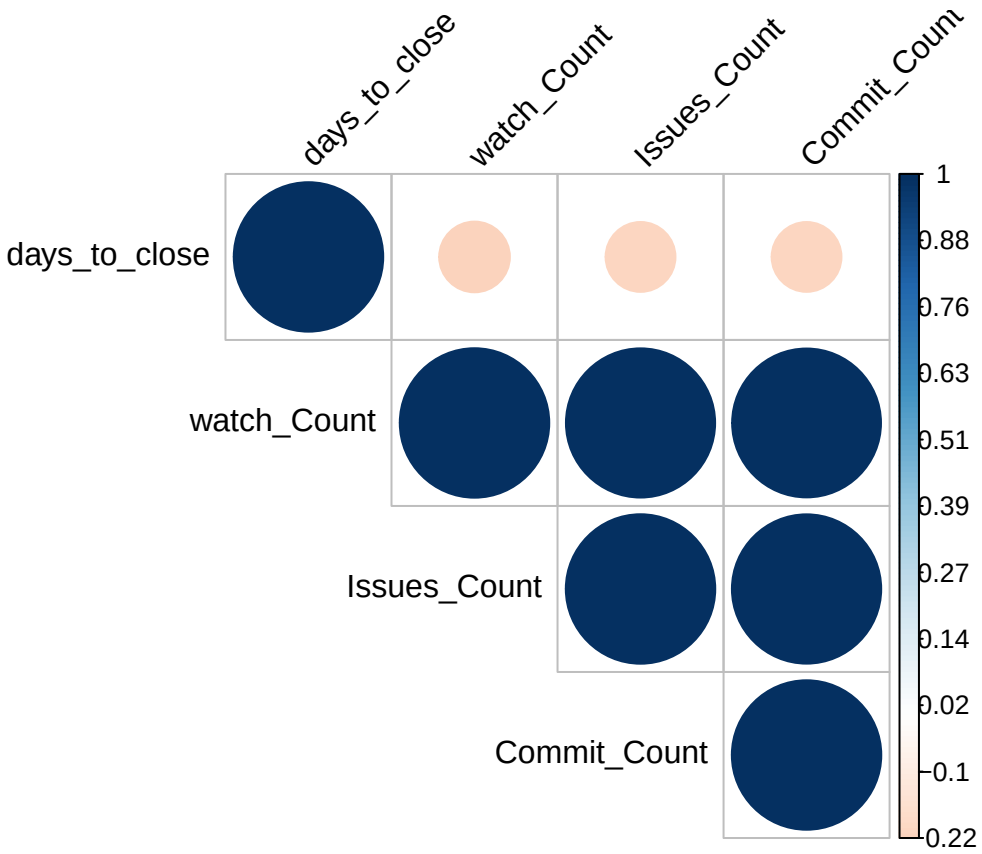



```
# Plot using performance analysis library
chart.Correlation(join, histogram=TRUE, pch=19)
```



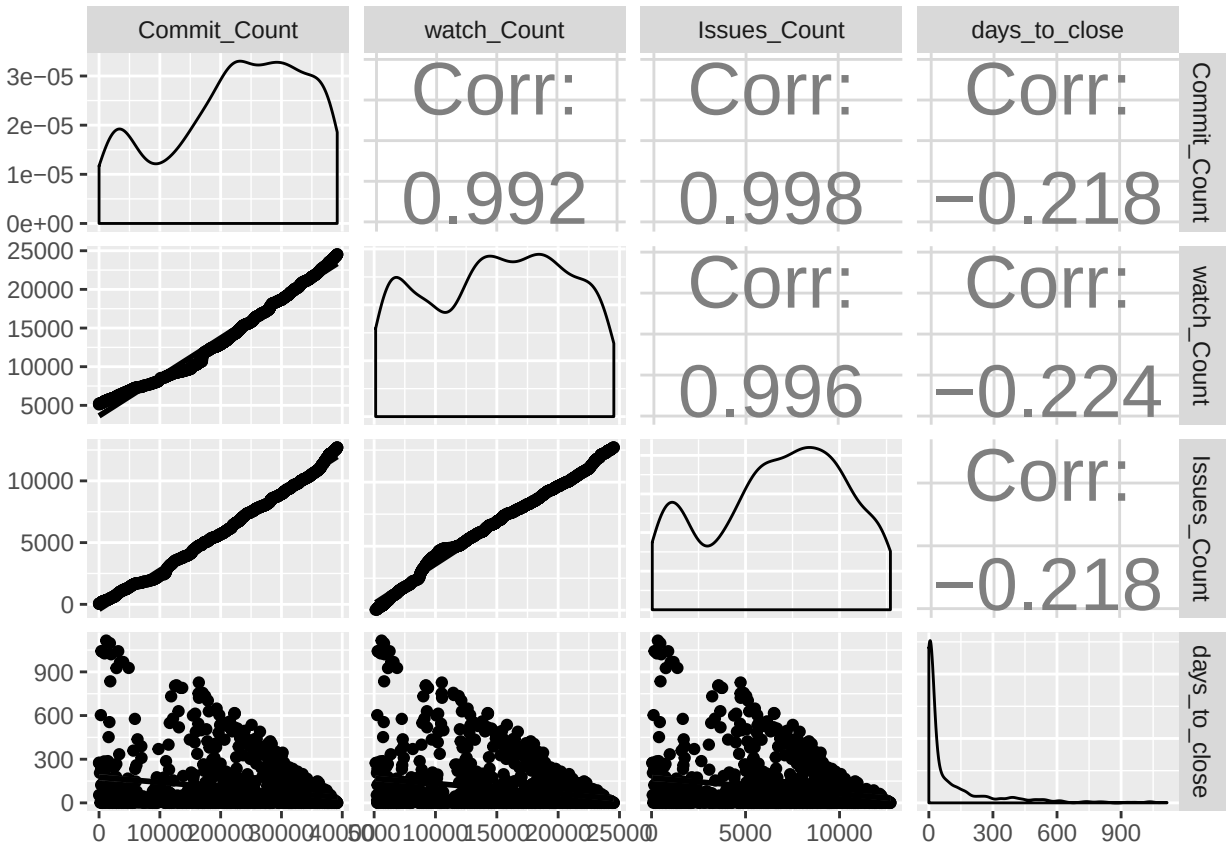

```
# Various Plots

corrplot(cor(join), type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45,  use="complete.obs", is.corr=FALSE)
```




```
ggcorr(join,  label = T)
```


```
ggpairs(join,
        columns = c("Commit_Count", "watch_Count", "Issues_Count","days_to_close"),
        upper = list(continuous = wrap("cor",
                                       size = 10)),
        lower = list(continuous = "smooth"))
```


Commit_Count
watch_Count
Issues_Count
days_to_close
Corr:
0.992
Corr:
0.998
Corr:
−0.218
Corr:
0.996
Corr:
−0.224
Corr:
−0.218
3e−05
2e−05
1e−05
0e+00
25000
20000
15000
10000
5000
0
900
600
300
0
0
10000
20000
30000
40000
5000
10000
15000
20000
25000
0
5000
10000
0
300
600
900