

Plots for Superconductor Dataset

Mariya Kim

10/8/2021

Plots for all **Superconductor Dataset** variables were additionally used for data analysis to select plots for the report and discover relationships and correlations.

- (1) Distributions of the variables. Histograms of all variables show, that there no normally distributed variables in this data set and some have low variance.
- (2) *critical_temp* vs variables, colored by *number_of_elements*. There are variables that can be stratified by *number_of_elements* and can be found using this set of graphs.
- (3) Correlations in the groups of predictors.
- (4) Cluster heatmaps of feature groups also show that patterns can be found across all plots; some features are paired:

maen and *gmean*;
std and *wtd_std*;
entropy and *wtd_entropy*;
wtd_mean and *wtd_gmean*.

indicating similar predictors that were obtained using similar calculations.

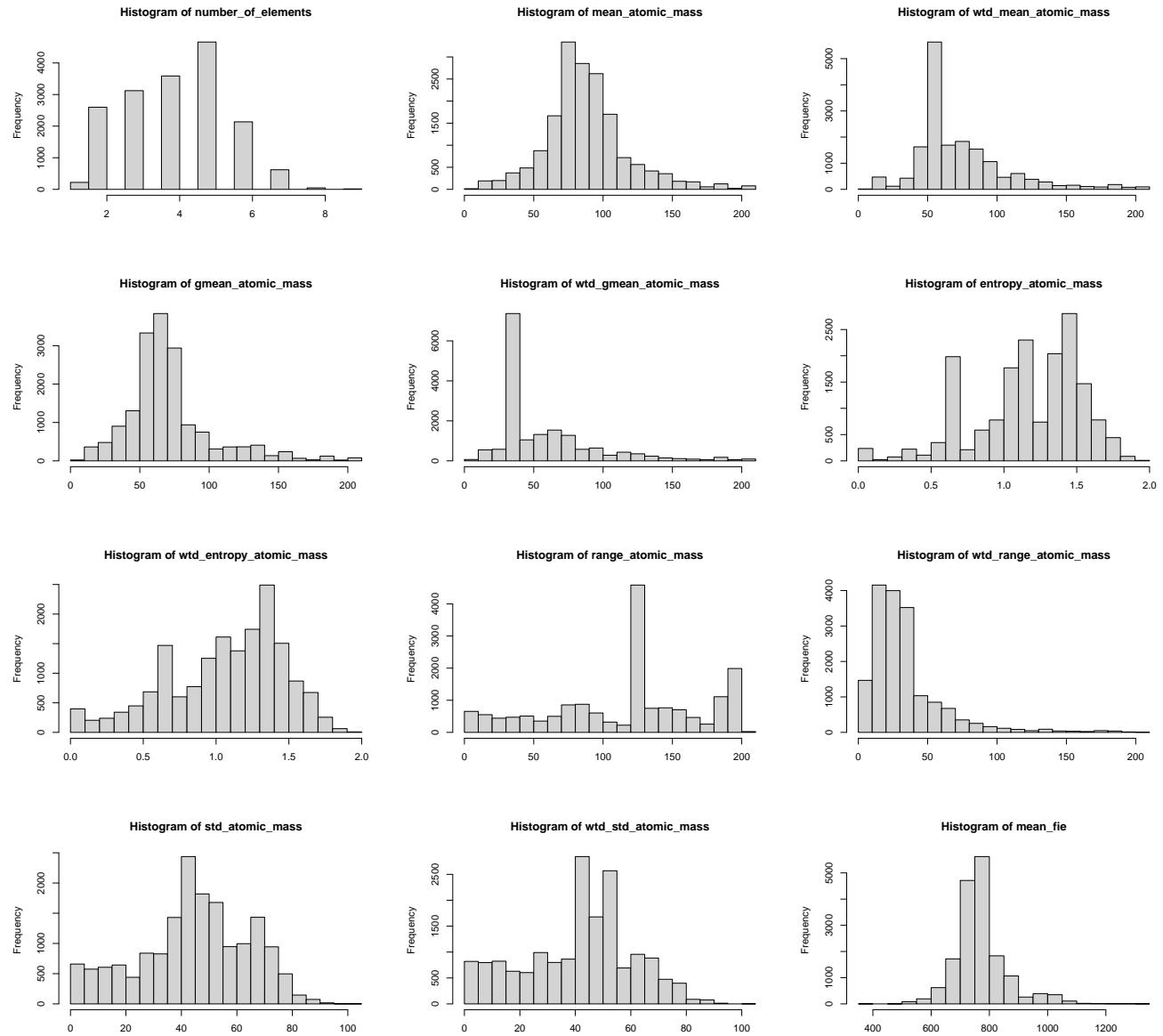
Also *_Density* variables stand out in many cluster heatmaps, as well as “range_” features.

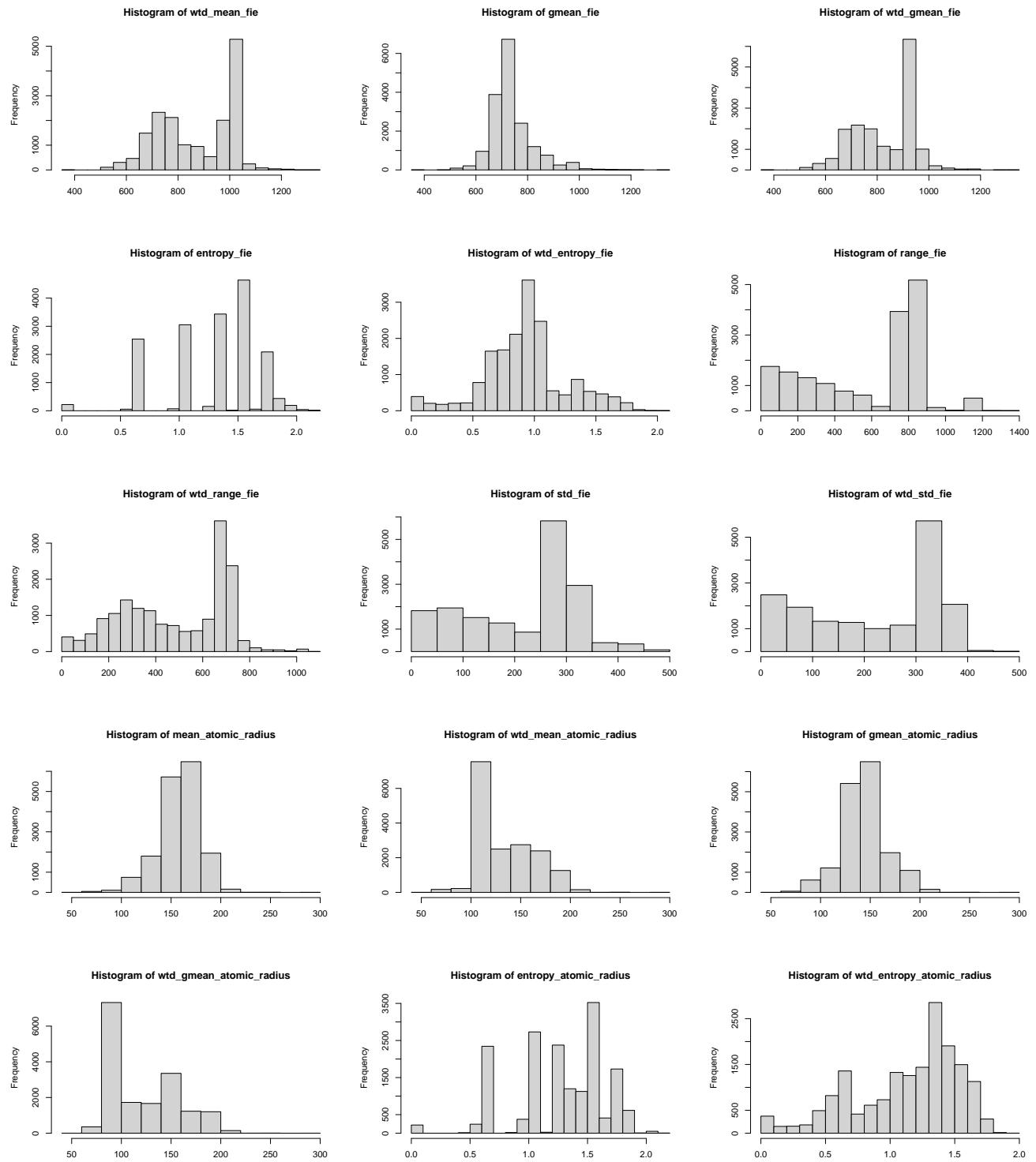
```
if(!require(tidyverse)) install.packages("tidyverse")
if(!require(readxl)) install.packages("readxl")
if(!require(ggplot2)) install.packages("ggplot2")
if(!require(caret)) install.packages("caret")
if(!require(tidyr)) install.packages("tidyr")
if(!require(reshape2)) install.packages("reshape2")
if(!require(tinytex)) install.packages("tinytex")
if(!require(dplyr)) install.packages("dplyr")
if(!require(stringr)) install.packages("stringr")
if(!require(RColorBrewer)) install.packages("RColorBrewer")
if(!require(matrixStats)) install.packages("matrixStats")
if(!require(tinytex)) install.packages("tinytex")
if(!require(knitr)) install.packages("knitr")

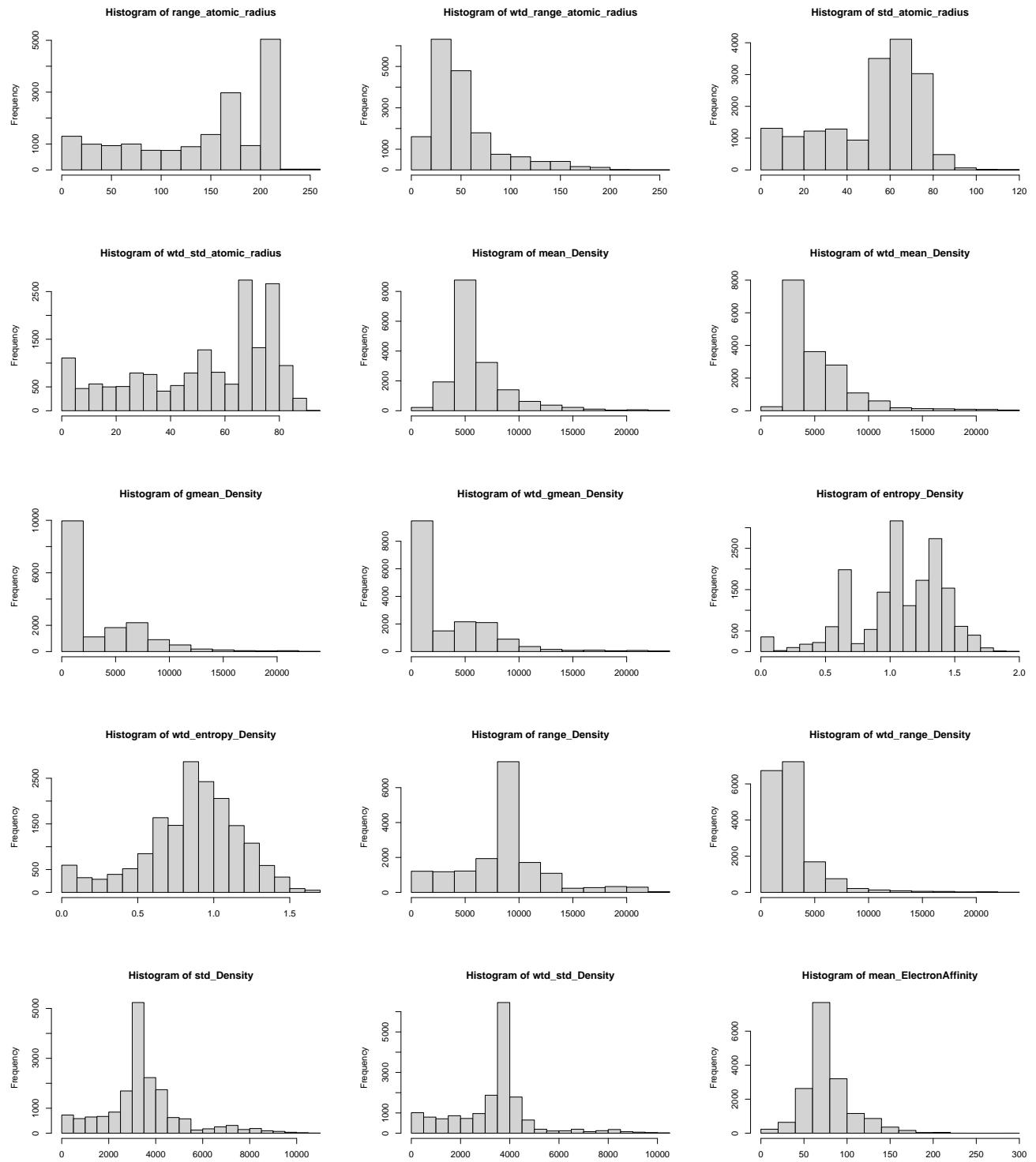
fileName <- "4_Superconductors.xlsx"
superconductors <- read_excel(fileName)
set.seed(1, sample.kind="Rounding") # if using R 3.5 or earlier, use `set.seed(1)`
index_val <- createDataPartition(y = superconductors$critical_temp,
                                 times = 1, p = 0.2, list = FALSE)
# sc set, that does not have validation set rows.
sc <- superconductors[-index_val, ]
# 25% of sc set is used for Critical_temp-vs-variables plots.
sc_small <- slice_sample(sc, prop = 0.25)
```

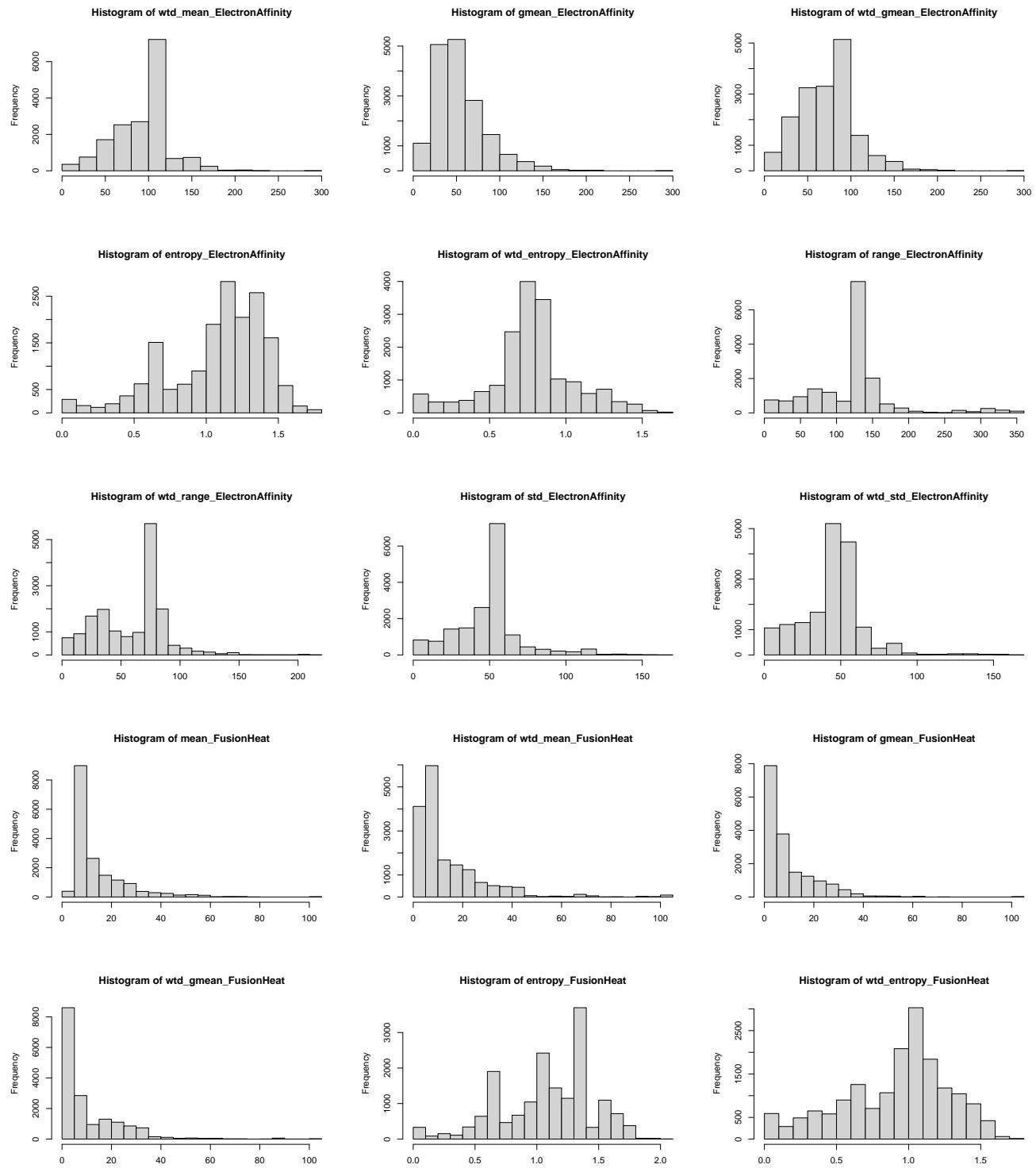
(1) Distributions of all variables in Superconductivity data set.

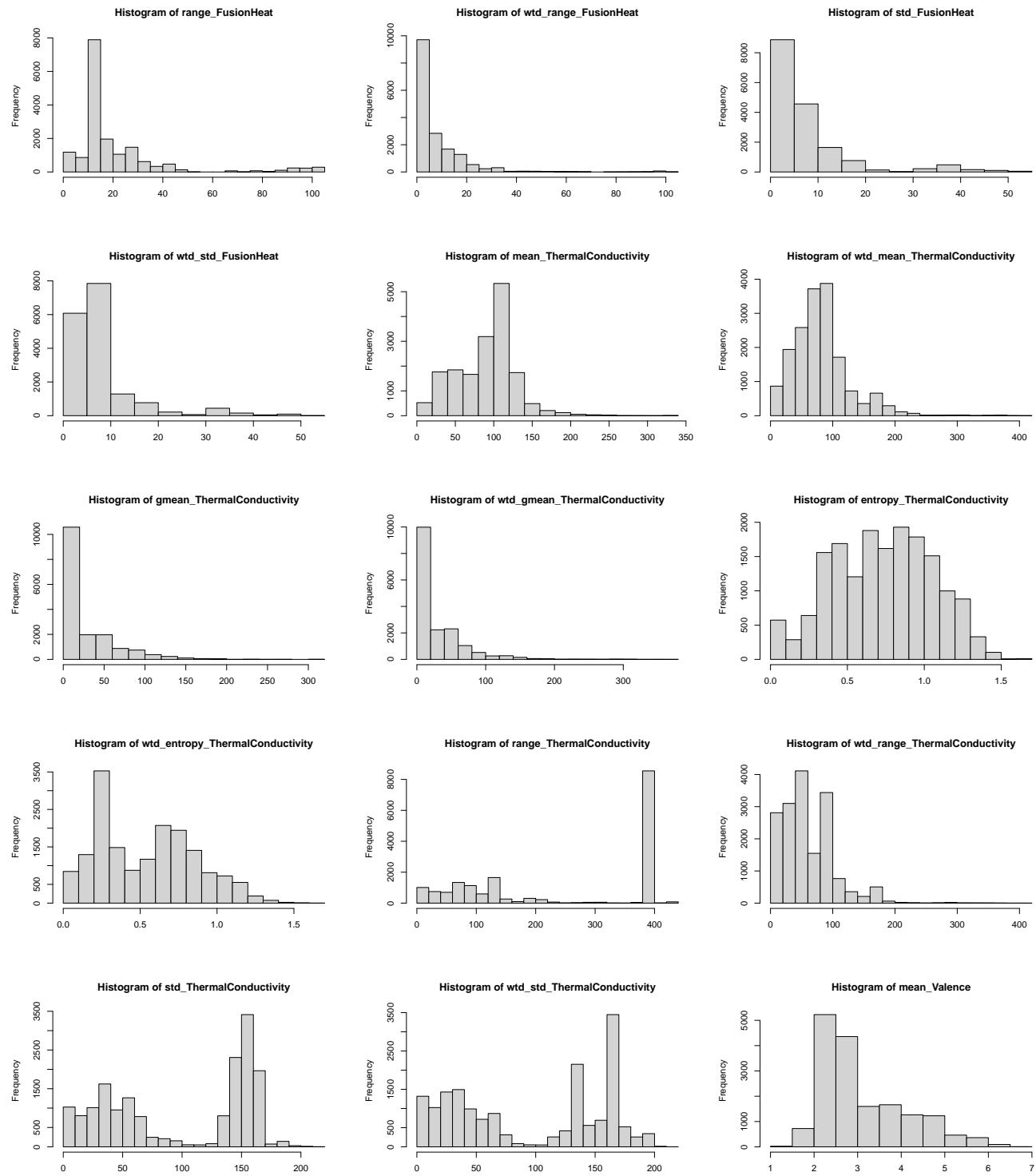
```
cols <- 1:82
plot_histogram_function <- sapply(cols, function(cols){
  df <- sc[,cols] %>% as.matrix()
  plot <- hist(df,
               xlab="",
               main = paste("Histogram of", colnames(sc)[cols]))
  plot
})
```

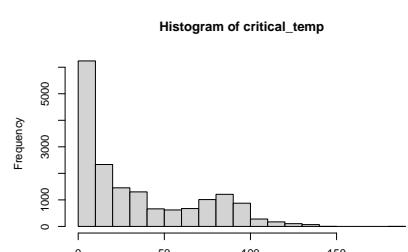
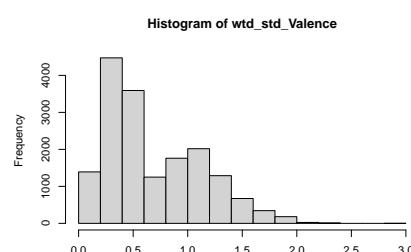
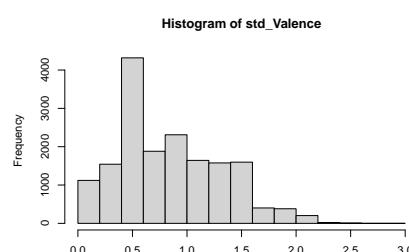
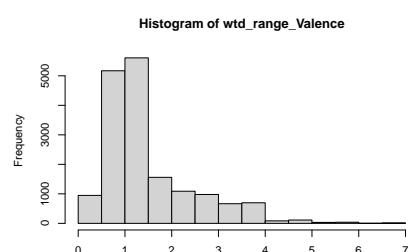
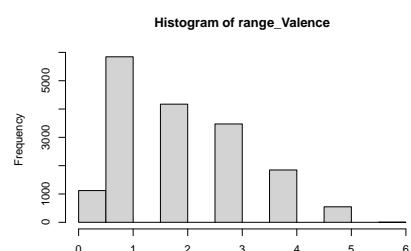
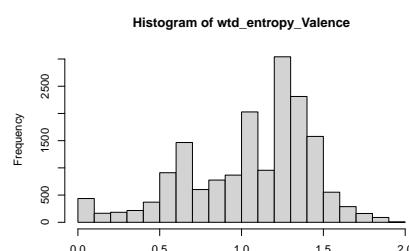
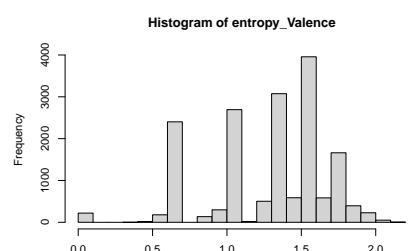
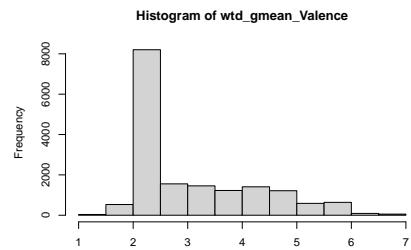
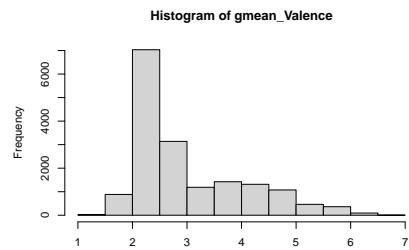
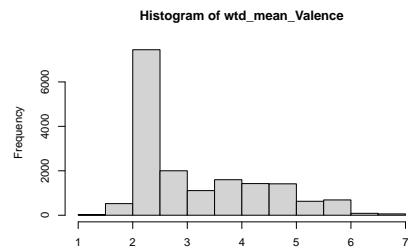






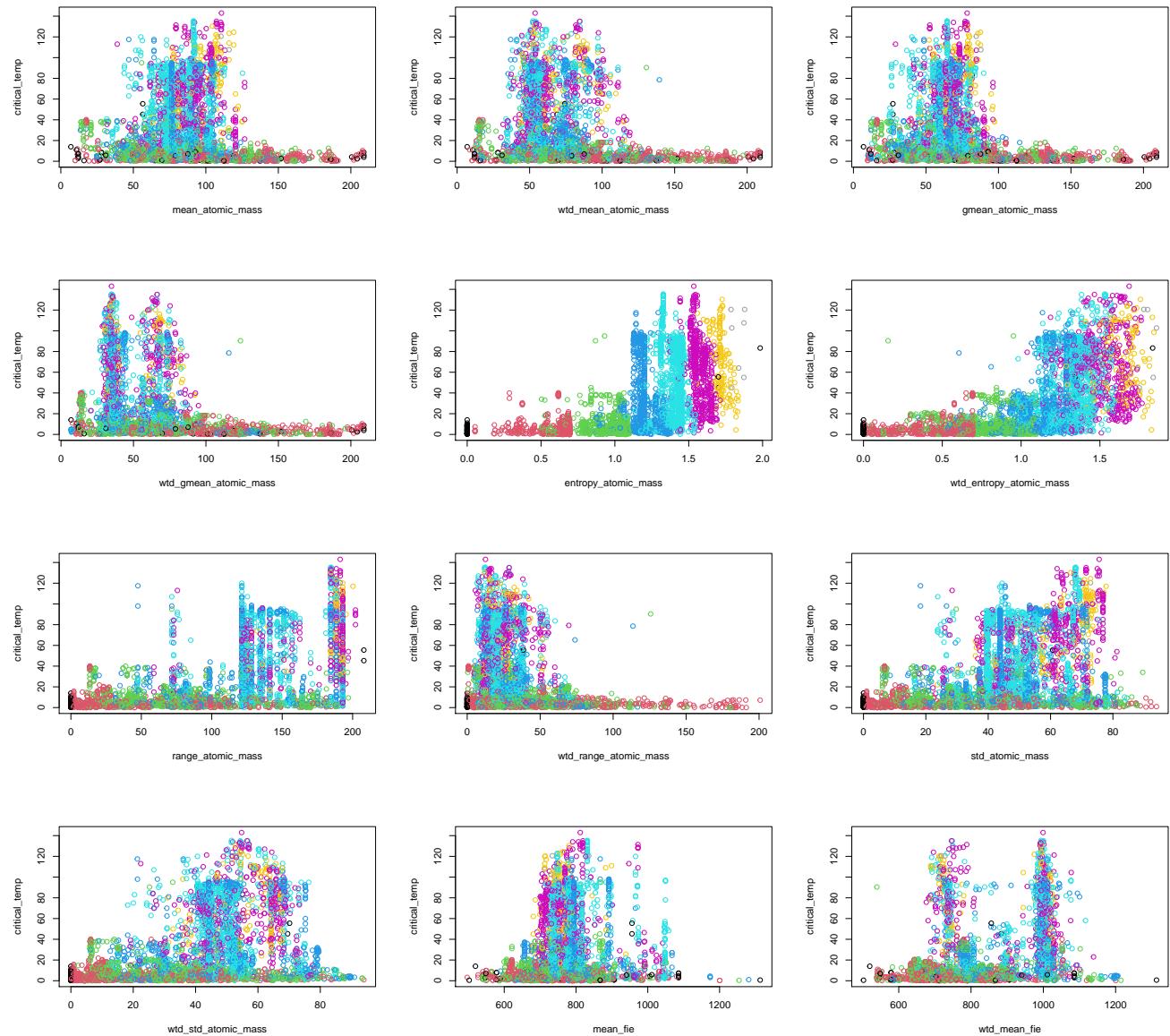


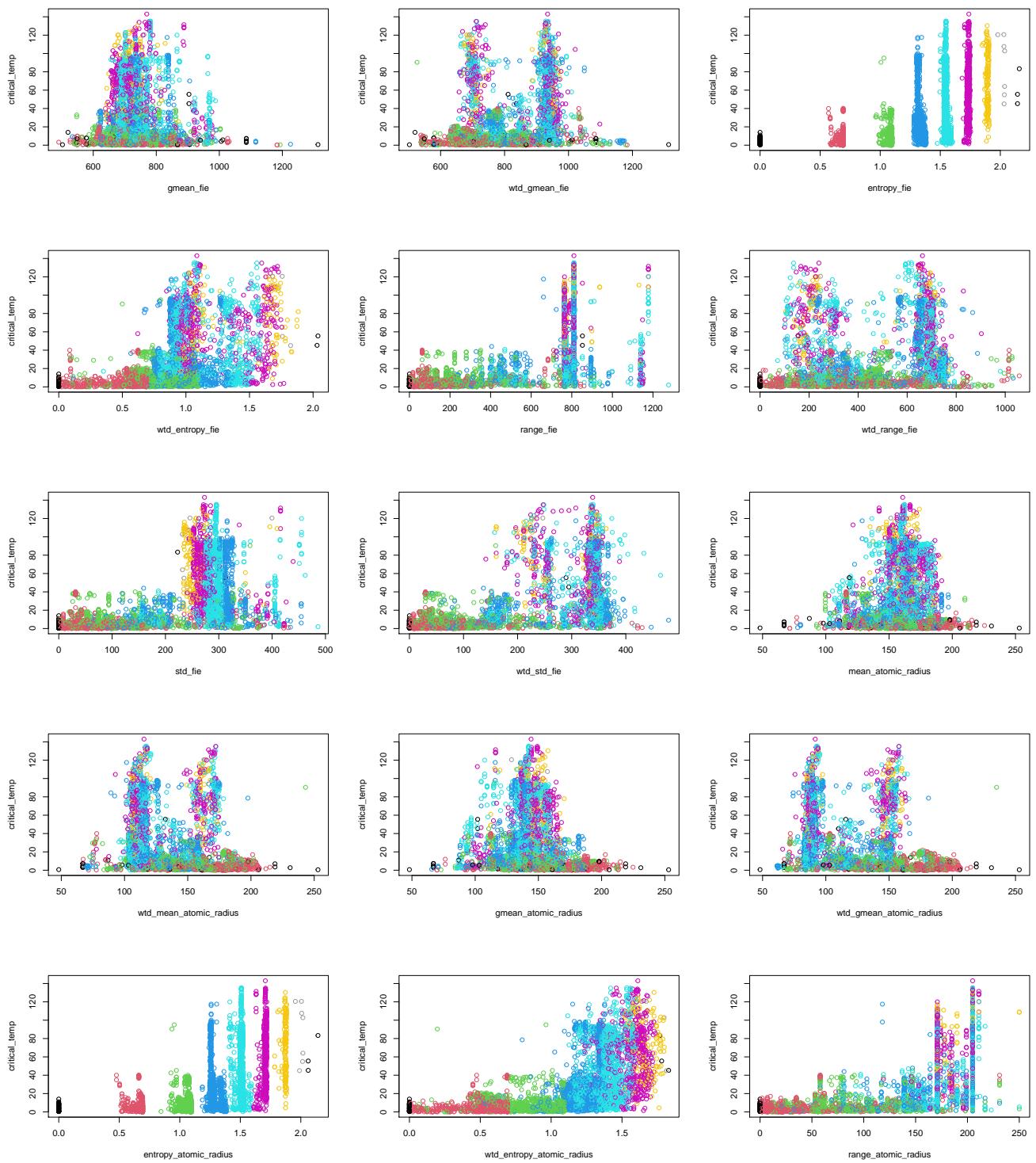


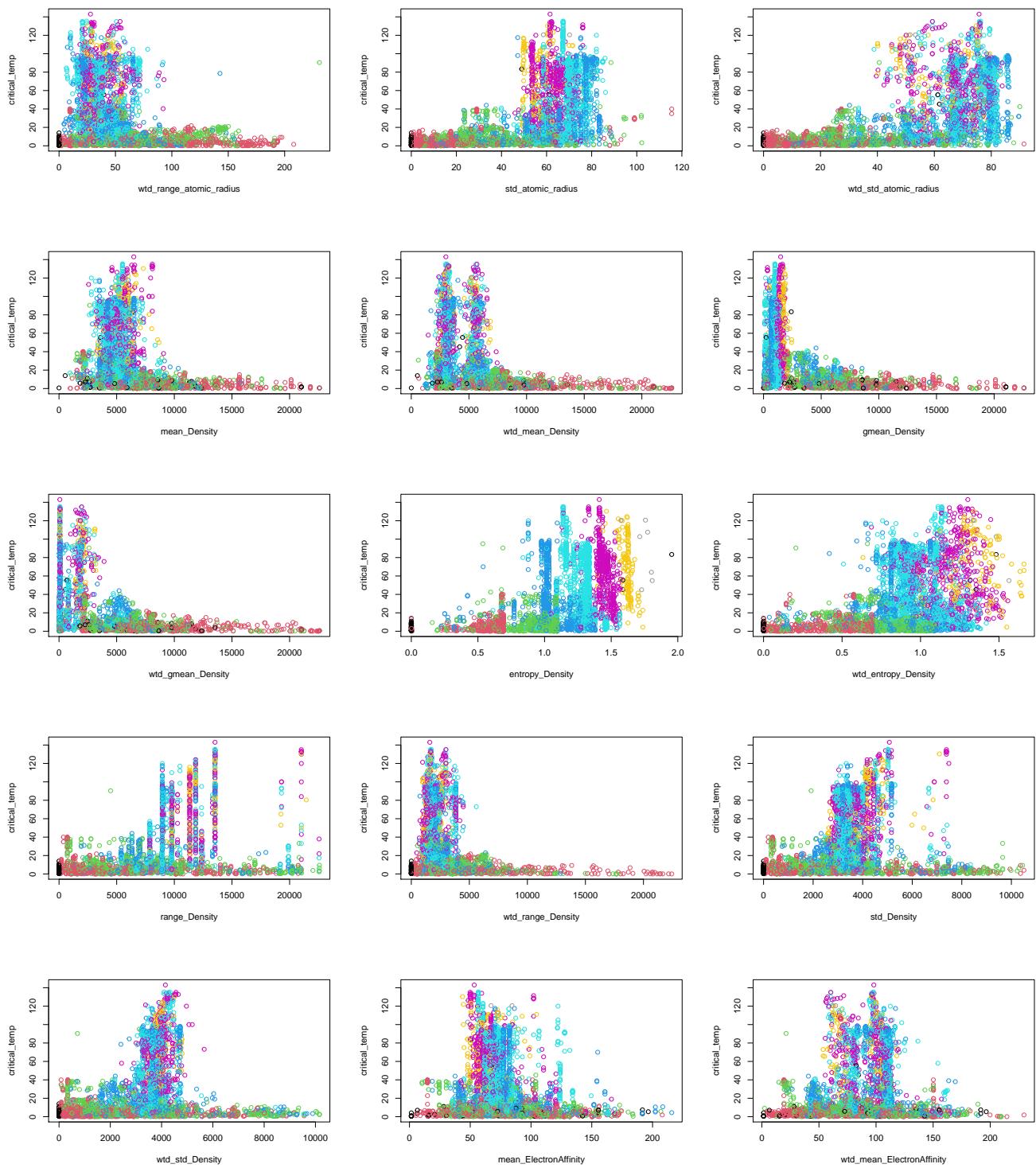


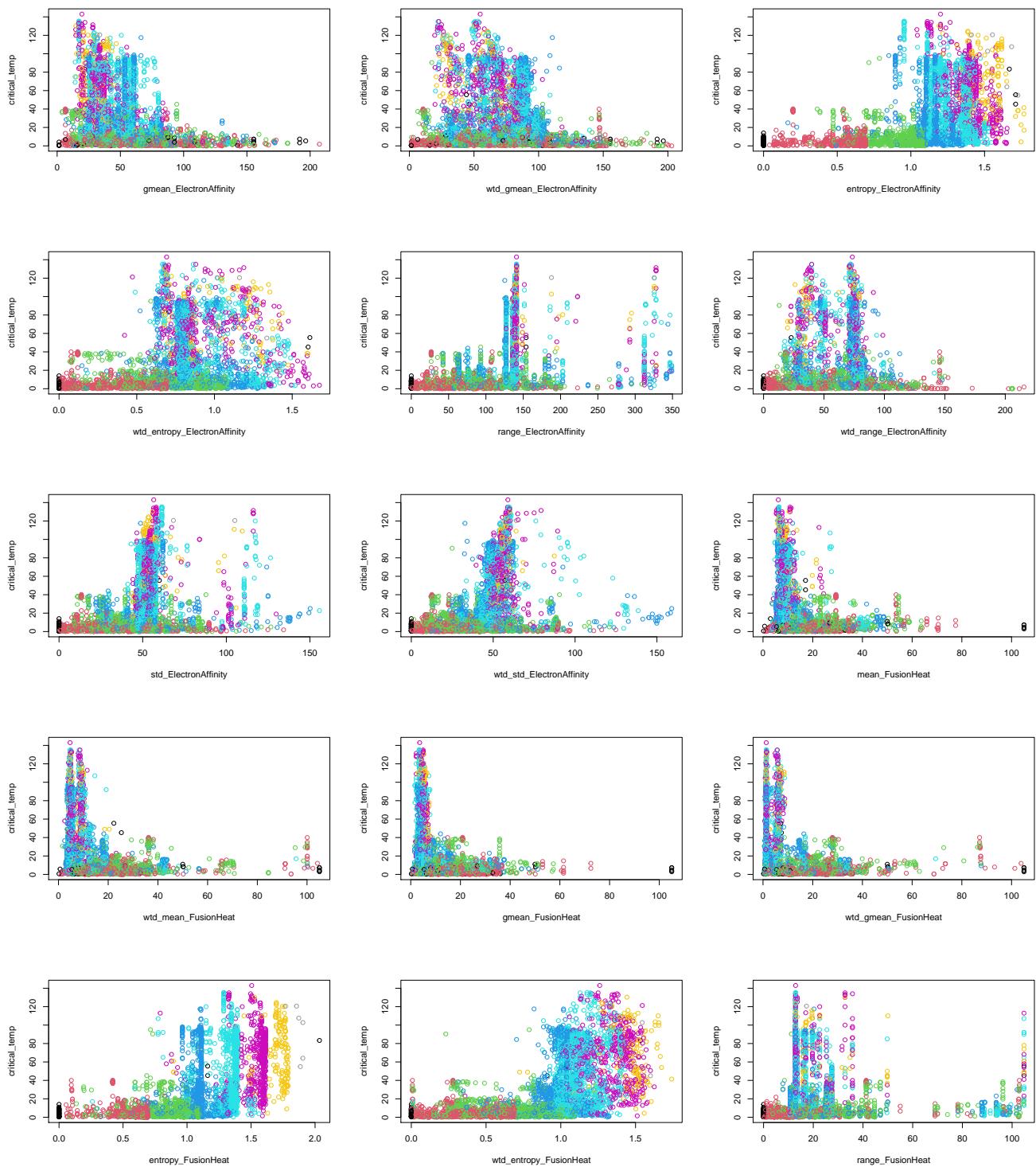
(2) Critical temperature vs variables, colored by *number_of_elements*.

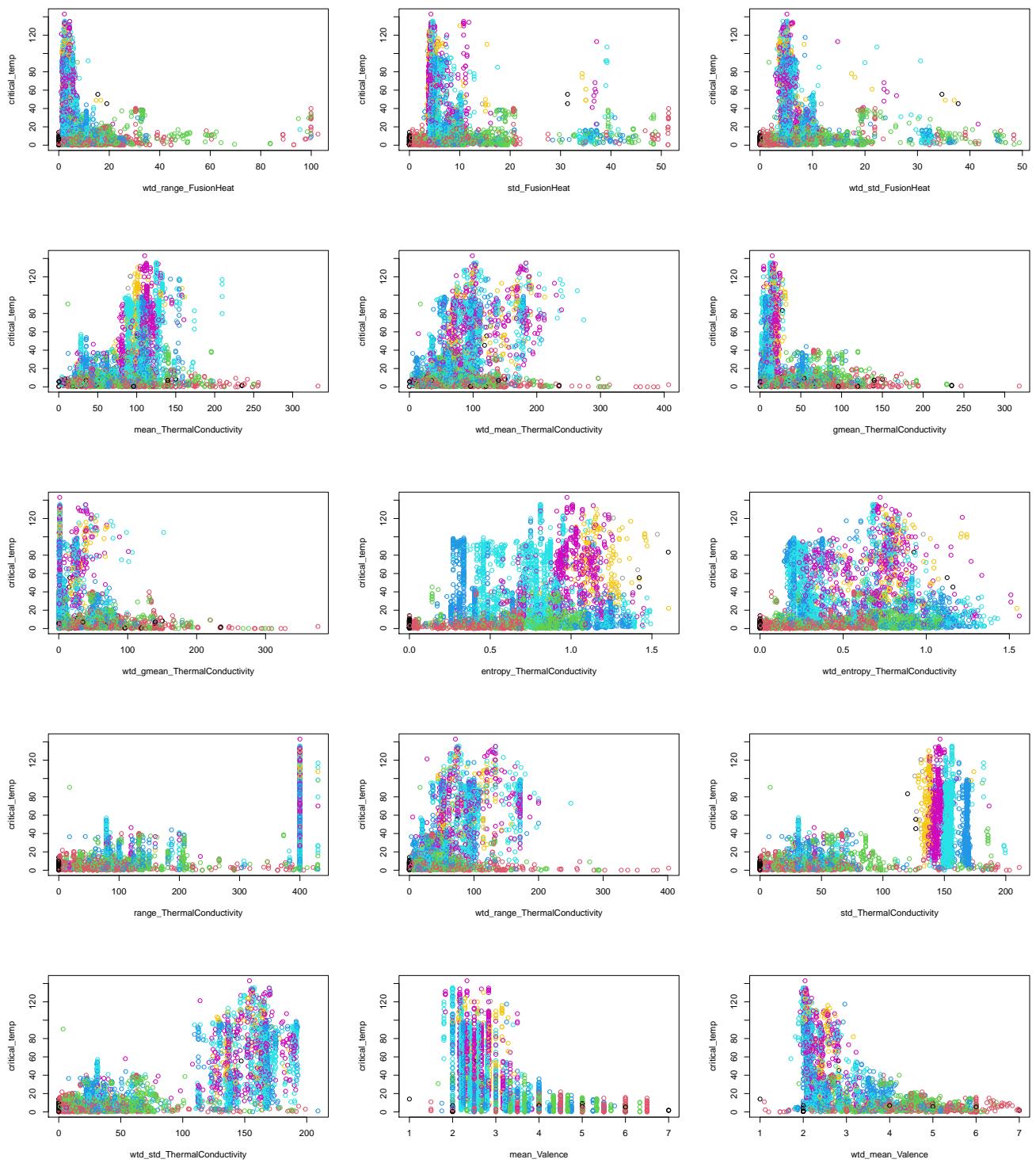
```
cols <- 2:81
plot_function <- sapply(cols, function(cols){
  df <- cbind(sc_small[,cols], sc_small[82])
  plot(df, col=sc_small$number_of_elements)
})
```

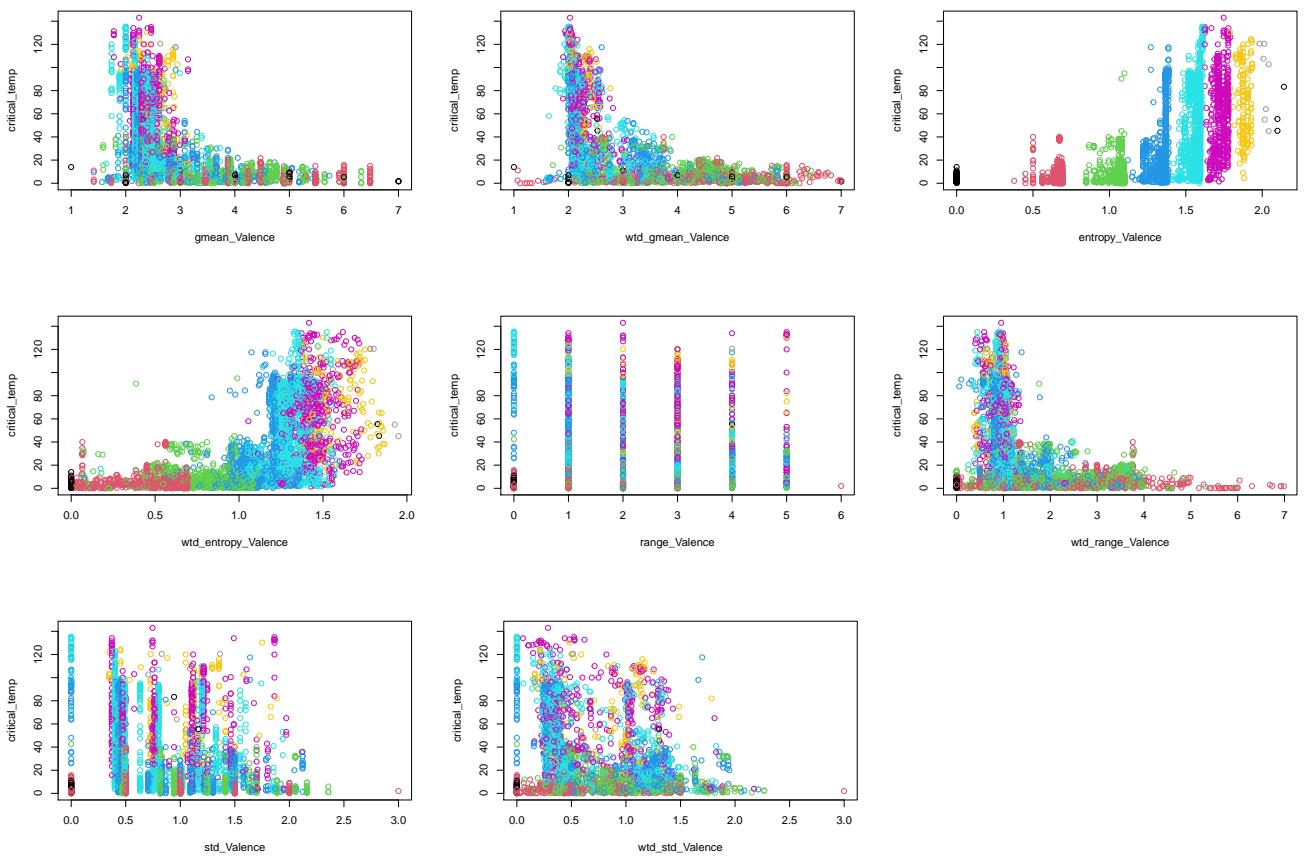












(3) Correlations in the groups of predictors.

```

col_name <- str_subset(colnames(sc), "atomic_mass") # length(col_name)
patterns1 <- str_replace(col_name, "_atomic_mass", "")
patt_unite <- data.frame(a = rep("^", 10), b = patterns1)
patterns1 <- unite(patt_unite, a, b, col = "c", sep = "") %>% pull()
string <- toString(patterns1)
string <- str_replace_all(string, ", ", "|")
patterns2 <- str_replace(colnames(sc), string, "") %>% unique()
patterns2 <- patterns2[2:9]
p <- c(patterns2, patterns1)

```

Correlation heatmaps of features that are grouped using patterns:

```

p

## [1] "_atomic_mass"      "_file"          "_atomic_radius"
## [4] "_Density"         "_ElectronAffinity"  "_FusionHeat"
## [7] "_ThermalConductivity"  "_Valence"        "^mean"
## [10] "^wtd_mean"       "^gmean"         "^wtd_gmean"
## [13] "^entropy"        "^wtd_entropy"    "^range"
## [16] "^wtd_range"      "^std"           "^wtd_std"

corr_heatmaps <- map(p, function(p){
  corr_group <- sc[str_subset(colnames(sc), p)] %>%
    cor() %>% data.frame()

  corr_group <- rownames_to_column(corr_group, "col_name")
  corr_group <- melt(corr_group) %>%
    set_names(c("col_name", "variable", "value"))

  plot <- corr_group %>%
    ggplot(aes(col_name, variable, fill = value)) +
    geom_tile() +
    scale_fill_distiller(palette = "Reds", direction = 1) +
    labs(caption = paste("Correlation heatmap of the variables in", p, "group.")) +
    theme_minimal() +
    theme(plot.caption.position = "panel",
          plot.caption = element_text(size = 14, face = "bold.italic", hjust = 0.65),
          axis.title.x = element_blank(),
          axis.title.y = element_blank(),
          axis.text.x = element_text(angle = 90, hjust = 1))
  plot
})

corr_heatmaps

## [[1]]
##
## [[2]]
##
## [[3]]
##
## [[4]]

```

```

## 
## [[5]]

## 
## [[6]]

## 
## [[7]]

## 
## [[8]]

## 
## [[9]]

## 
## [[10]]

## 
## [[11]]

## 
## [[12]]

## 
## [[13]]

## 
## [[14]]

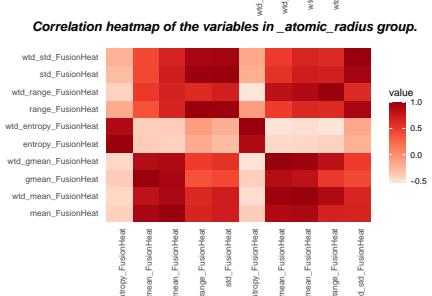
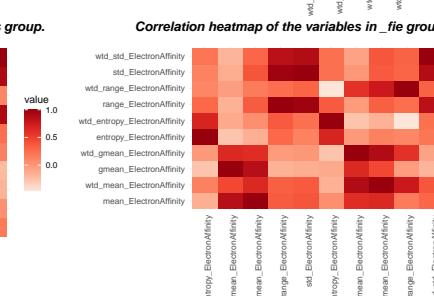
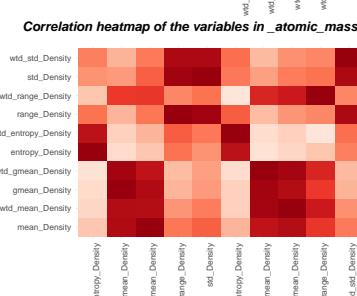
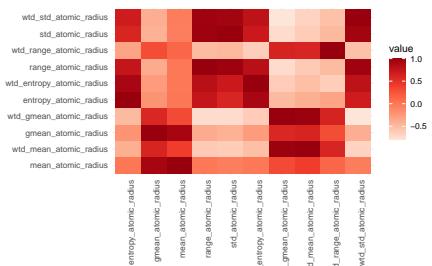
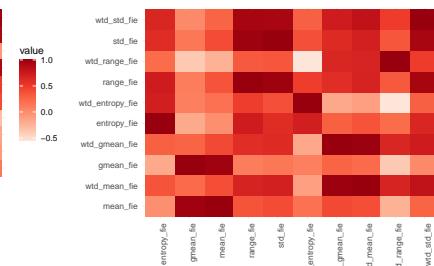
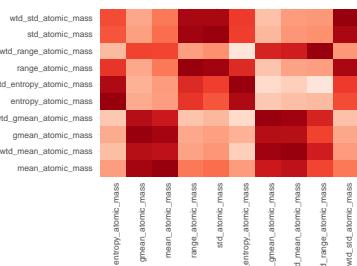
## 
## [[15]]

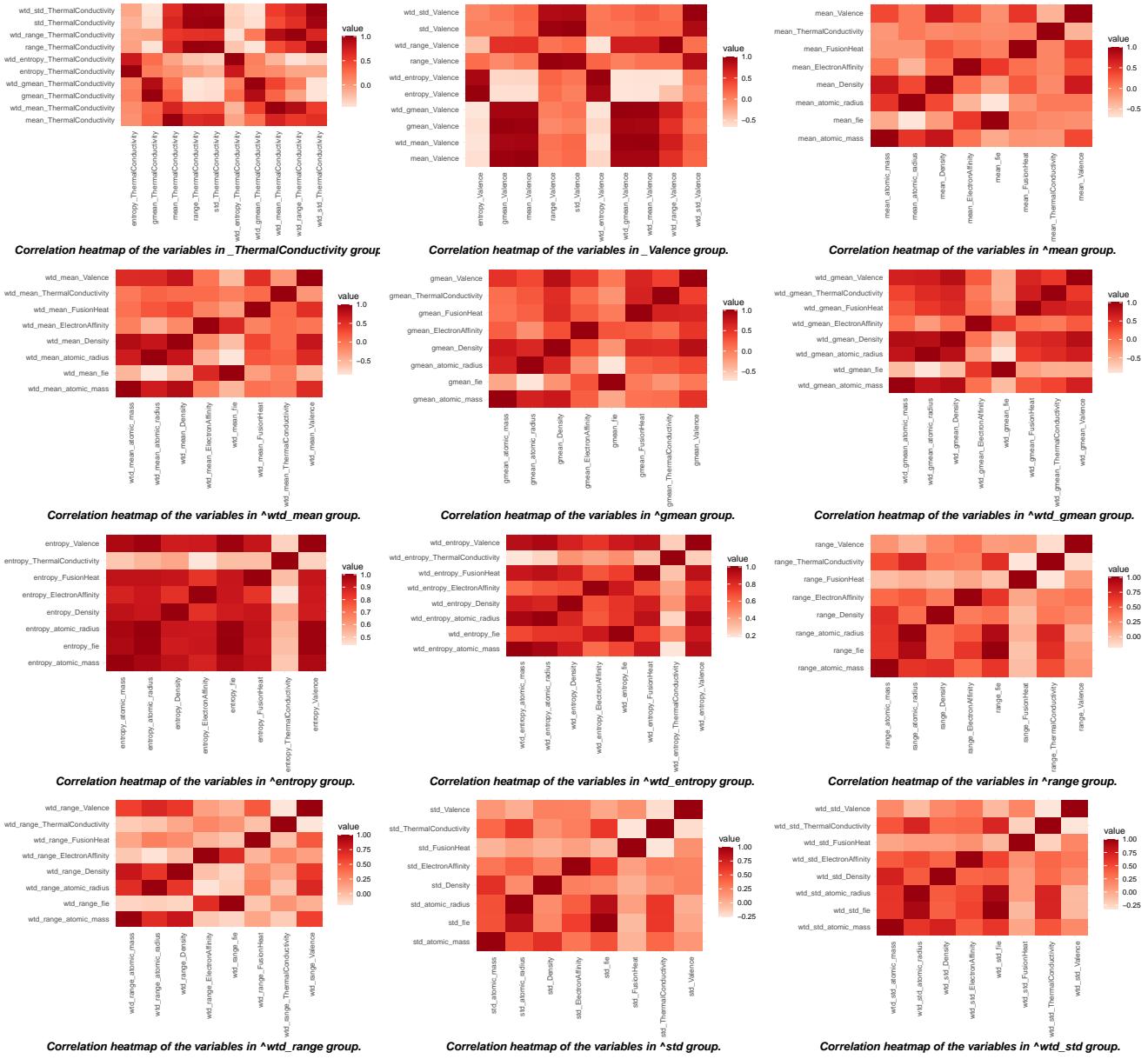
## 
## [[16]]

## 
## [[17]]

## 
## [[18]]

```





(4) discovering clusters in the groups of features using heatmap() function.

Features are grouped using patterns:

p

```
## [1] "_atomic_mass"      "_fie"          "_atomic_radius"
## [4] "_Density"         "_ElectronAffinity"  "_FusionHeat"
## [7] "_ThermalConductivity"  "_Valence"       "``mean"
## [10] ``wtd_mean"        ``gmean"         ``wtd_gmean"
## [13] ``entropy"          ``wtd_entropy"    ``range"
## [16] ``wtd_range"        ``std"           ``wtd_std"

x <- sc[1:81] %>% as.matrix()
y <- sc[82] %>% round() %>% pull()
rownames(x) <- y # critical_temp column is turned into row names
x <- sweep(x, 2, colMeans(x)) %>% as.matrix() # removing the center

heatm_clust_col <- sapply(p, function(p){
  ind <- which(colnames(sc) %in% colnames(sc[str_subset(colnames(sc), p)])))
  heatmap(t(x[,ind]), col = brewer.pal(11, "Spectral"),
          scale = "column")
})

heatm_clust_col
```

```
##      _atomic_mass _fie      _atomic_radius _Density
## rowInd Integer,10 Integer,10 Integer,10 Integer,10
## colInd Integer,17009 Integer,17009 Integer,17009 Integer,17009
## Rowv   NULL      NULL      NULL      NULL
## Colv   NULL      NULL      NULL      NULL
##      _ElectronAffinity _FusionHeat  _ThermalConductivity _Valence
## rowInd Integer,10     Integer,10     Integer,10     Integer,10
## colInd Integer,17009     Integer,17009     Integer,17009     Integer,17009
## Rowv   NULL      NULL      NULL      NULL
## Colv   NULL      NULL      NULL      NULL
##      ``mean"      ``wtd_mean" ``gmean"      ``wtd_gmean" ``entropy"
## rowInd Integer,8     Integer,8     Integer,8     Integer,8     Integer,8
## colInd Integer,17009     Integer,17009     Integer,17009     Integer,17009     Integer,17009
## Rowv   NULL      NULL      NULL      NULL      NULL
## Colv   NULL      NULL      NULL      NULL      NULL
##      ``wtd_entropy" ``range"      ``wtd_range" ``std"      ``wtd_std"
## rowInd Integer,8     Integer,8     Integer,8     Integer,8     Integer,8
## colInd Integer,17009     Integer,17009     Integer,17009     Integer,17009     Integer,17009
## Rowv   NULL      NULL      NULL      NULL      NULL
## Colv   NULL      NULL      NULL      NULL      NULL
```

