water fluoridation ~ caries 12 years 1990 2003 2012 2019

Namhuynh

Preapre data from sav file

```
# 12 year-old
data12 <- read.spss ("~/Library/Mobile Documents/com~apple~CloudDocs/Nam/New project/20 years of water

## re-encoding from latin1

data12a <- data12[-c(1,15,20,22,23,24 )]
colnames(data12a)[c(3,13,14,15,16,17,18)] <- c("fluoride condition","caries","fluorosis teeth","DEANinc
colnames(data12a)[c(5:12)] <- c("DT","MT","FT","DMFT","DS","MS","FS","DMFS")

data12a$code <- as.character(data12a$code)
data12a$year <- as.factor(data12a$year)
firstup <- function(x) {
  x <- tolower(x)
  substr(x, 1, 1) <- toupper(substr(x, 1, 1))
  x
}
data12a$DEANindex <- firstup(data12a$DEANindex)
#data12a$DEANindex[is.na(data12a$DEANindex)] <- c("Normal")
data12a$DEANindex <- recode factor(data12a$DEANindex, "Normal" = "Normal", "Questinable" = "Questional
                                   "Moderate" = "Moderate", "Serve" = "Severe")

#data12a$fluoride condition[is.na(data12a$fluoride condition)] <- c("khong fluor hoa")
data12a$fluoride condition <- recode factor(data12a$fluoride condition, `khong fluor hoa` = "none fluo
                                            `Fluor hoa khong on dinh` = "unstable fluoridation",
                                            `Fluor hoa on dinh` = "stable fluoridation")

data12a$gender <- recode factor(data12a$gender, `NAM` = "male", `NU` = "female")
data12a$caries <- recode factor(data12a$caries,
                      `CO`="TRUE",`KHONG`="FALSE")%>%as.logical()
#data12a$fluorosis[is.na(data12a$fluorosis)] <- c("no")
data12a$fluorosis <- recode factor(data12a$fluorosis,
                      `yes`="TRUE",`no`="FALSE")%>%as.logical()
data12a$age <- c("12")
data12a$age <- as.factor(data12a$age)

# There were F+ in 1990
data12a$fluoride concentration <-  as.factor(ifelse(data12a$year == 2012 & data12a$area == "F+", "0.5
                                                    ifelse(data12a$year == 2003 & data12a$area == "F
```

```
# 2019
data_2019 <- read.spss ("~/Library/Mobile Documents/com~apple~CloudDocs/Nam/New project/20 years of wat
```

```
## Warning in read.spss("~/Library/Mobile Documents/com~apple~CloudDocs/Nam/New
## project/20 years of water fluoridation/Data/DATA2019/DATATPHCM1215tuoi2019-
## Phantich-WFProject.sav", : ~/Library/Mobile Documents/com~apple~CloudDocs/Nam/
## New project/20 years of water fluoridation/Data/DATA2019/DATATPHCM1215tuoi2019-
## Phantich-WFProject.sav: Long string value labels record found (record type 7,
## subtype 21), but ignored
```

```
## re-encoding from latin1
```

```
data_2019 <- data_2019[-c(1,3,6)]
colnames(data_2019)[c(4,5,9,10)] <- c("DEANindex","fluorosis_teeth","DMFT","caries")
data_2019$DEANindex <- firstup(data_2019$DEANindex)
data_2019$DEANindex <- recode_factor(data_2019$DEANindex, "Normal" = "Normal", "Questionnable" = "Questi
                                     "Moderate" = "Moderate", "Serve" = "Severe")
data_2019$district <- NA
data_2019$fluoride_condition <- NA
data_2019$year <- 2019
data_2019$year <- as.factor(data_2019$year)
data_2019$fluorosis <- as.factor(ifelse(data_2019$DEANindex == "Normal","FALSE","TRUE"))%>%as.logical()
data_2019$fluoride_concentration <-  as.factor(ifelse(data_2019$area == "F+", "0.5_ppm","0_ppm"))
data_2019 <- data_2019[c(1,16,17,2,6:9,11:14,10,5,4,18,15,19,3,20)]
data_2019$age <- recode_factor(data_2019$age, "12 year-old" = 12, "15 year-old" = 15)
data_2019$gender <- recode_factor(data_2019$gender, "NAM       " = "male","Nam       "="male",
                                  "NU        " = "female","2         "="female")
data_2019$caries <- recode_factor(data_2019$caries,
                      `Yes`="TRUE",`No`="FALSE")%>%as.logical()
data_2019$fluorosis_teeth <- ifelse(data_2019$DEANindex == "Normal",0,data_2019$fluorosis_teeth)

# 12 year old, 2019
data_12_2019 <- subset(data_2019, data_2019$age == "12")

# merge data
data <- rbind(data_12a, data_12_2019)

write.table(data,"~/Library/Mobile Documents/com~apple~CloudDocs/Nam/New project/20 years of water fluo
```

bar_missing

```
bar_missing <- function(x){
  require(reshape2)
  x %>%
    is.na %>%
    melt %>%
    ggplot(data = .,
           aes(x = Var2)) +
    geom_bar(aes(y=(..count..),fill=value),alpha=0.7)+
```

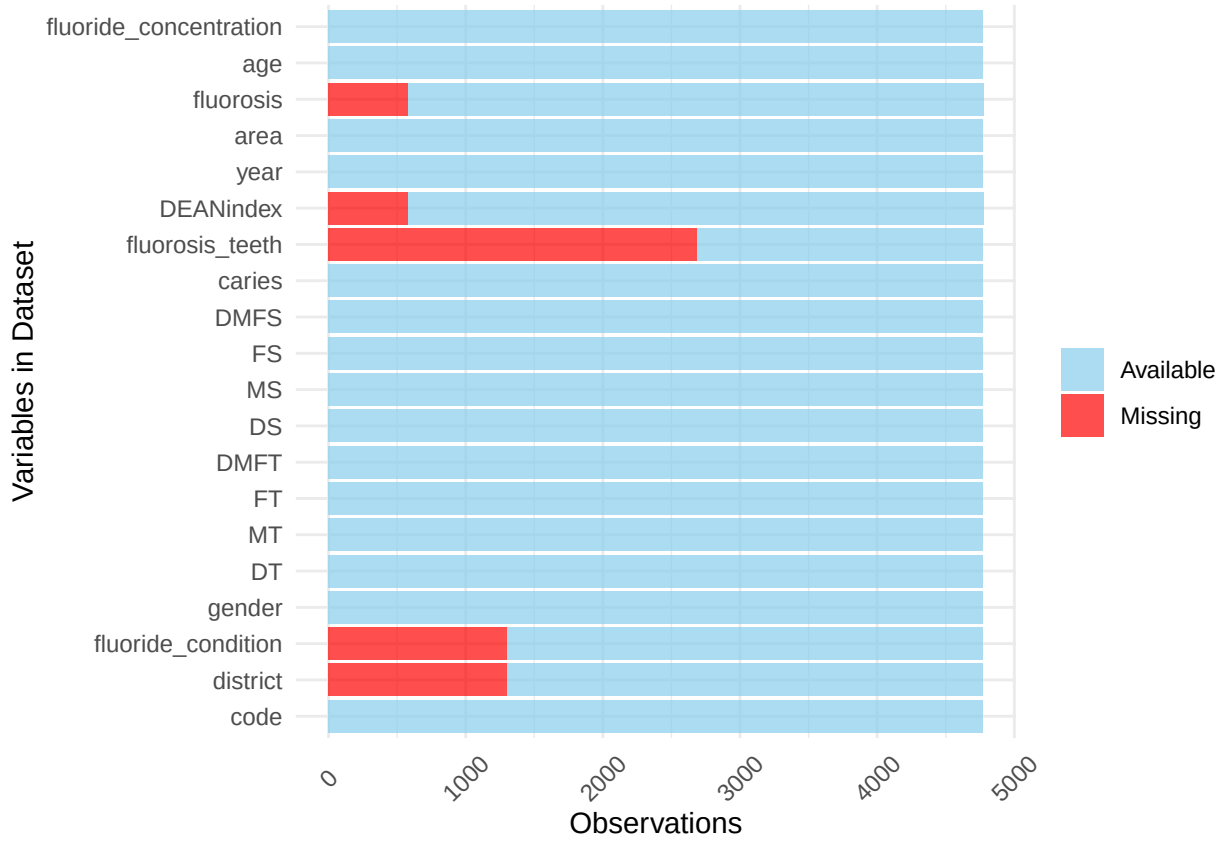
```
    scale_fill_manual(values=c("skyblue","red"),
                      name = "",
                      labels = c("Available","Missing"))+
    theme_minimal()+
    theme(axis.text.x = element_text(angle=45, vjust=0.5)) +
    labs(x = "Variables in Dataset",
         y = "Observations")+coord_flip()
}

bar_missing(data)
```

```
## Loading required package: reshape2

##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```



Prepare data w/o NA or unnecessary variables

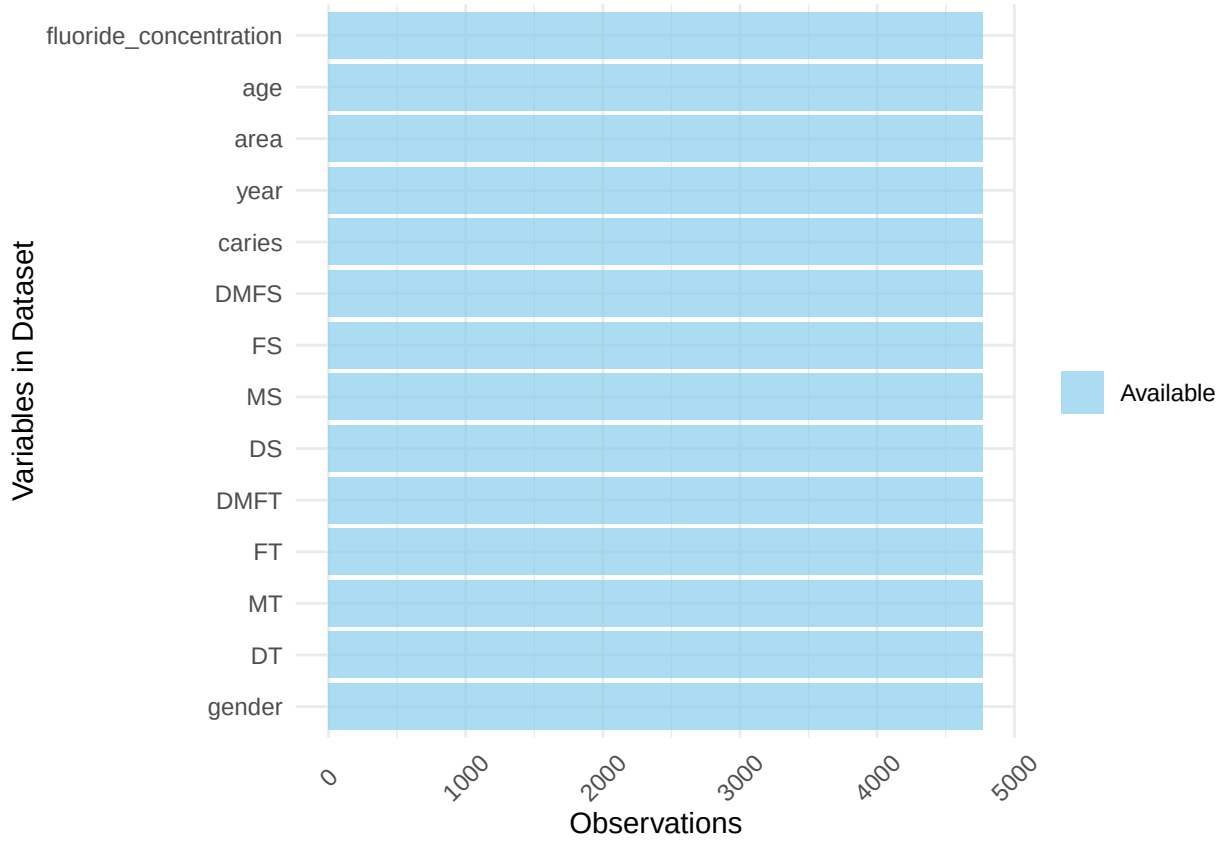
percent <- list() for (i in 1:ncol(data_b)) { percent[[i]] <- tabyl(data_b[[i]], sort = F) print(i) print(percent[[i]]) }

```
library(janitor)
```

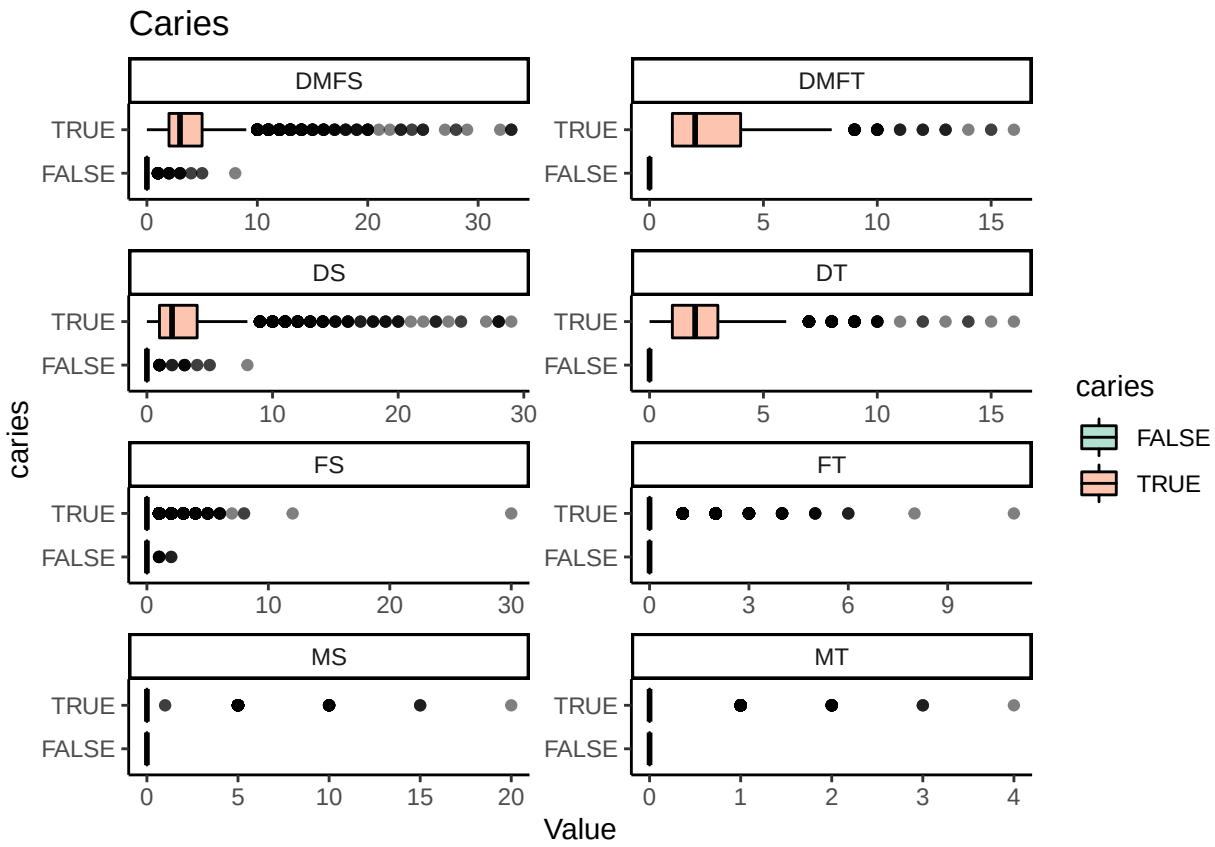
```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

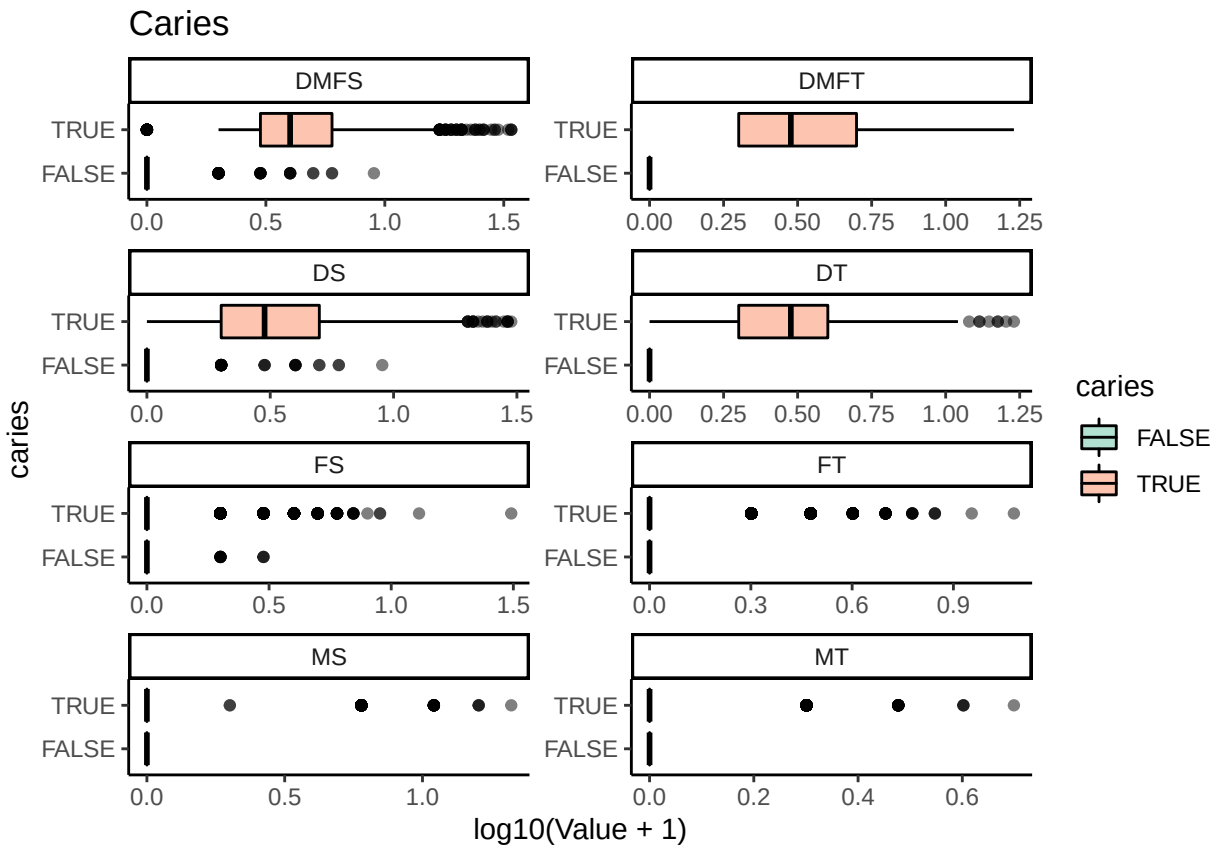
```
data_b <- data[-c(1:3,14:15,18)]
bar_missing(data_b)
```



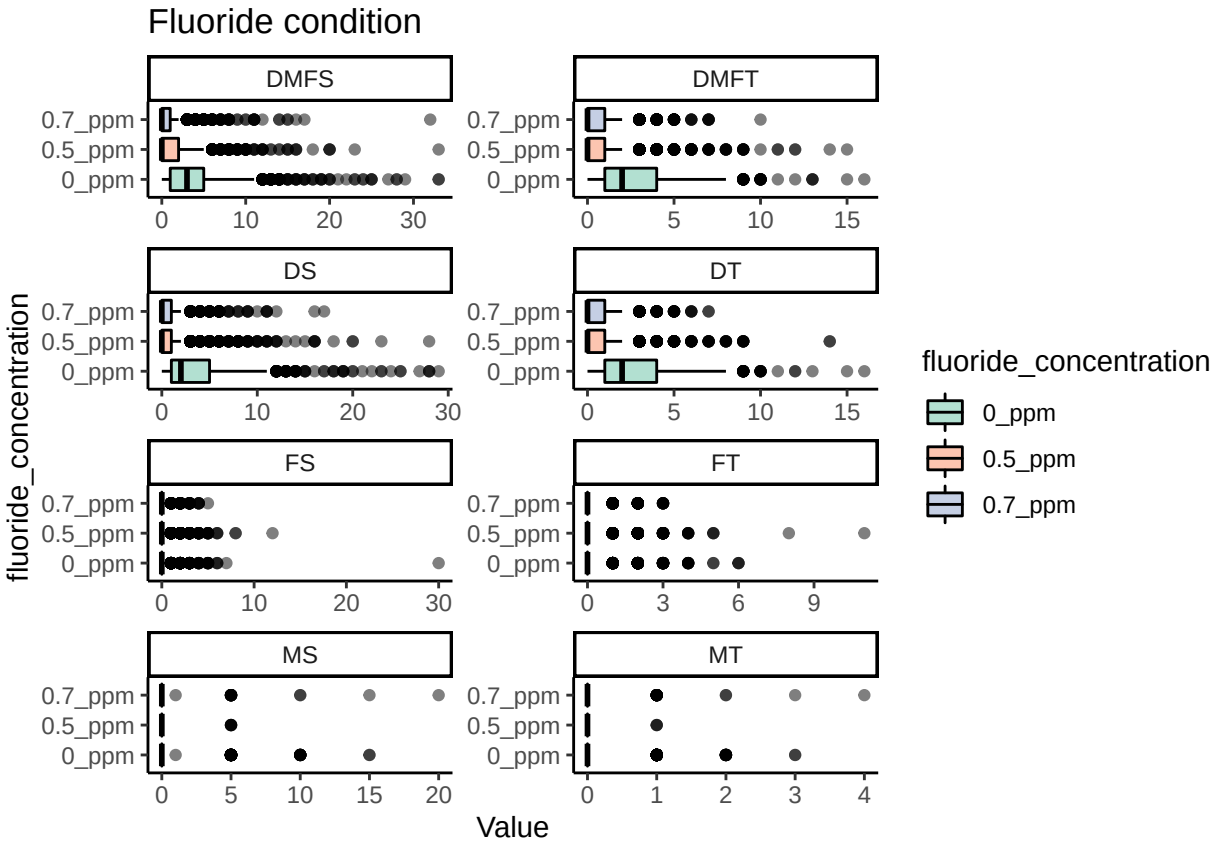

```
data_b %>%gather(c(DT:DMFS),
                 key = "Parameter",
                 value="Value")%>%
  ggplot(aes(x=caries, y=Value,fill=caries))+
  geom_boxplot(alpha=0.5,col="black")+
  ggtitle("Caries")+
  facet_wrap(~Parameter,ncol=2,scales = "free")+
  coord_flip()+
  scale_fill_manual(values = brewer.pal(n = 4, name = "Set2"))+ theme_classic()
```



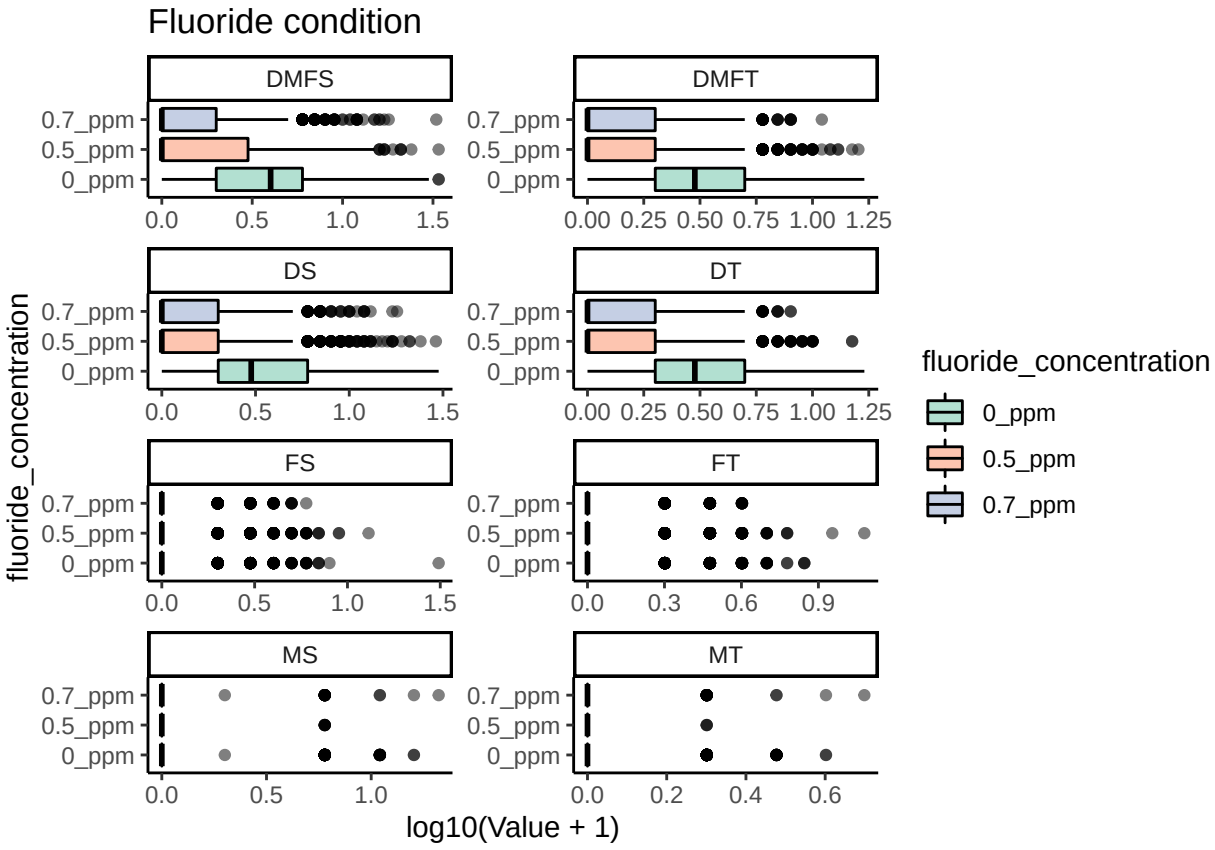

```
#transformation
data_b %>%gather(c(DT:DMFS),
                 key = "Parameter",
                 value="Value")%>%
  ggplot(aes(x=caries, y=log10(Value+1),fill=caries))+
  geom_boxplot(alpha=0.5,col="black")+
  ggtitle("Caries")+
  facet_wrap(~Parameter,ncol=2,scales = "free")+
  coord_flip()+
  scale_fill_manual(values = brewer.pal(n = 4, name = "Set2"))+ theme_classic()
```




```
data_b %>%gather(c(DT:DMFS),
                    key = "Parameter",
                    value="Value")%>%
  ggplot(aes(x=fluoride_concentration, y=Value,fill=fluoride_concentration))+
  geom_boxplot(alpha=0.5,col="black"
               )+
  ggtitle("Fluoride condition")+
  facet_wrap(~Parameter,ncol=2,scales = "free")+
  coord_flip()+
  scale_fill_manual(values = brewer.pal(n = 4, name = "Set2"))+ theme_classic()
```




```
#transformation
data_b %>%gather(c(DT:DMFS),
                 key = "Parameter",
                 value="Value")%>%
  ggplot(aes(x=fluoride_concentration, y=log10(Value+1),fill=fluoride_concentration))+
  geom_boxplot(alpha=0.5,col="black"
               )+
  ggtitle("Fluoride condition")+
  facet_wrap(~Parameter,ncol=2,scales = "free")+
  coord_flip()+
  scale_fill_manual(values = brewer.pal(n = 4, name = "Set2"))+ theme_classic()
```



k

```
#FS, FT, MS, MT are too low -> remove
#area is the same with fluoride_concentration -> remove
#remove age
data_b1 <- data_b[-c(3,4,7,8,12,13)]

#transformation
data_b2 <- data_b1
data_b2[,c(2:5)] <- log10(data_b1[,c(2:5)]+1)

scaled_data_b <- data_b2 %>%as.data.frame()
scaled_data_b[,c(2:5)] <- scale(scaled_data_b[,c(2:5)])

#remove the diagnosis variables
X_mat <- scaled_data_b %>% select(-c("caries"))

Es <- numeric(10)
for(i in 1:10){
  kpres <- kproto(X_mat,
                  k = i, nstart = 5,
                  lambda = lambdaest(X_mat),
                  verbose = FALSE)
  Es[i] <- kpres$tot.withinss}
```

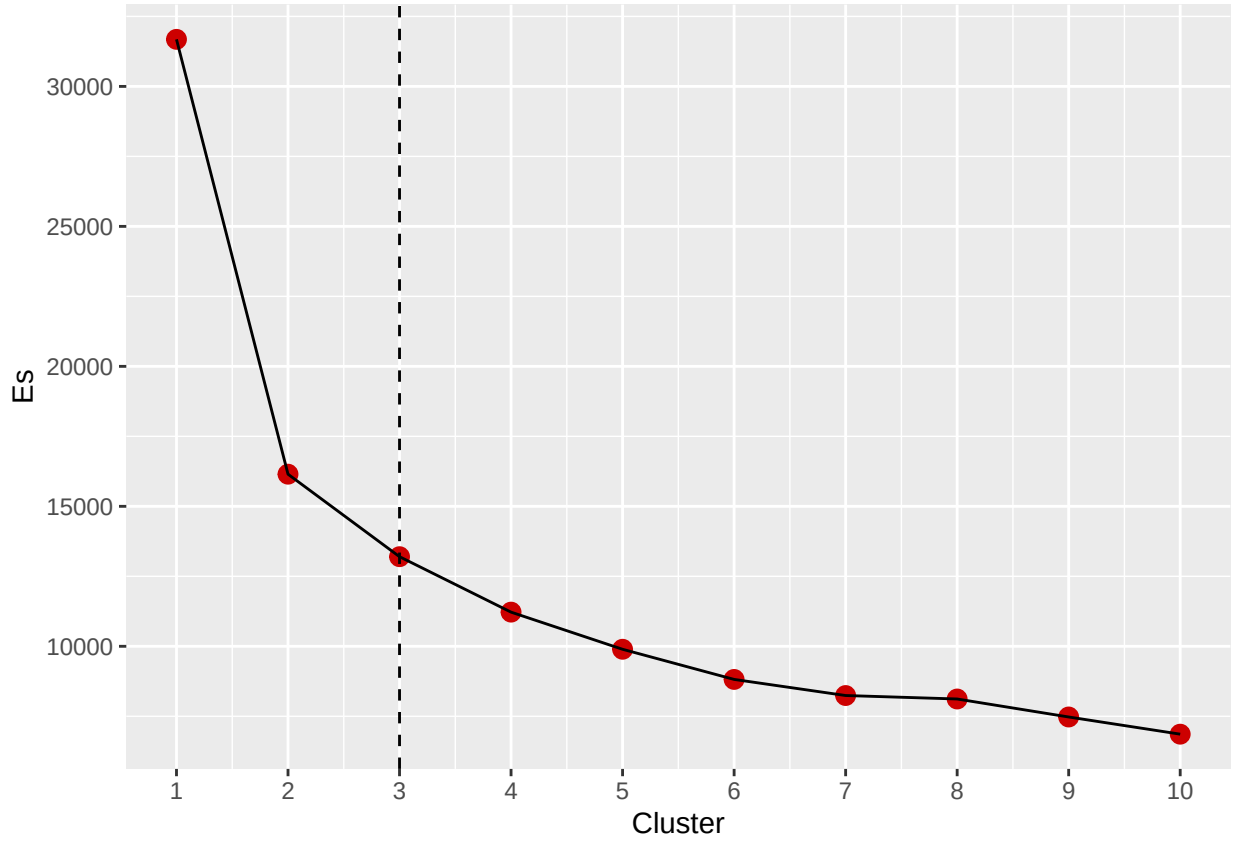
```
## Numeric variances:
##   DT DMFT   DS DMFS
```

```
##    1    1    1    1
## Average numeric variance: 1
##
## Heuristic for categorical variables: (method = 1)
##                 gender                   year fluoride_concentration
##              0.4987253              0.6857270              0.6418736
## Average categorical variation: 0.6087753
##
## Estimated lambda: 1.642642
##
## Numeric variances:
##   DT DMFT   DS DMFS
##    1    1    1    1
## Average numeric variance: 1
##
## Heuristic for categorical variables: (method = 1)
##                 gender                   year fluoride_concentration
##              0.4987253              0.6857270              0.6418736
## Average categorical variation: 0.6087753
##
## Estimated lambda: 1.642642
##
## Numeric variances:
##   DT DMFT   DS DMFS
##    1    1    1    1
## Average numeric variance: 1
##
## Heuristic for categorical variables: (method = 1)
##                 gender                   year fluoride_concentration
##              0.4987253              0.6857270              0.6418736
## Average categorical variation: 0.6087753
##
## Estimated lambda: 1.642642
##
## Numeric variances:
##   DT DMFT   DS DMFS
##    1    1    1    1
## Average numeric variance: 1
##
## Heuristic for categorical variables: (method = 1)
##                 gender                   year fluoride_concentration
##              0.4987253              0.6857270              0.6418736
## Average categorical variation: 0.6087753
##
## Estimated lambda: 1.642642
##
## Numeric variances:
##   DT DMFT   DS DMFS
##    1    1    1    1
## Average numeric variance: 1
##
## Heuristic for categorical variables: (method = 1)
##                 gender                   year fluoride_concentration
##              0.4987253              0.6857270              0.6418736
```

```
## Average categorical variation: 0.6087753
##
## Estimated lambda: 1.642642
##
## Numeric variances:
##   DT DMFT   DS DMFS
##    1    1    1    1
## Average numeric variance: 1
##
## Heuristic for categorical variables: (method = 1)
##                gender                  year fluoride_concentration
##             0.4987253             0.6857270              0.6418736
## Average categorical variation: 0.6087753
##
## Estimated lambda: 1.642642
##
## Numeric variances:
##   DT DMFT   DS DMFS
##    1    1    1    1
## Average numeric variance: 1
##
## Heuristic for categorical variables: (method = 1)
##                gender                  year fluoride_concentration
##             0.4987253             0.6857270              0.6418736
## Average categorical variation: 0.6087753
##
## Estimated lambda: 1.642642
##
## Numeric variances:
##   DT DMFT   DS DMFS
##    1    1    1    1
## Average numeric variance: 1
##
## Heuristic for categorical variables: (method = 1)
##                gender                  year fluoride_concentration
##             0.4987253             0.6857270              0.6418736
## Average categorical variation: 0.6087753
##
## Estimated lambda: 1.642642
##
## Numeric variances:
##   DT DMFT   DS DMFS
##    1    1    1    1
## Average numeric variance: 1
##
## Heuristic for categorical variables: (method = 1)
##                gender                  year fluoride_concentration
##             0.4987253             0.6857270              0.6418736
## Average categorical variation: 0.6087753
##
## Estimated lambda: 1.642642
##
## Numeric variances:
##   DT DMFT   DS DMFS
```

```
##    1    1    1    1
## Average numeric variance: 1
##
## Heuristic for categorical variables: (method = 1)
##                  gender                    year fluoride_concentration
##               0.4987253               0.6857270               0.6418736
## Average categorical variation: 0.6087753
##
## Estimated lambda: 1.642642
```

```
tibble(Cluster = c(1:10), Es = Es) %>%
  ggplot(aes(x = Cluster, y = Es)) +
  geom_point(size = 3,
             col ="red3") +
  geom_path() +
  geom_vline(xintercept = 3,
             linetype = 2)+
  scale_x_continuous(breaks = c(1:10))
```



Clustering

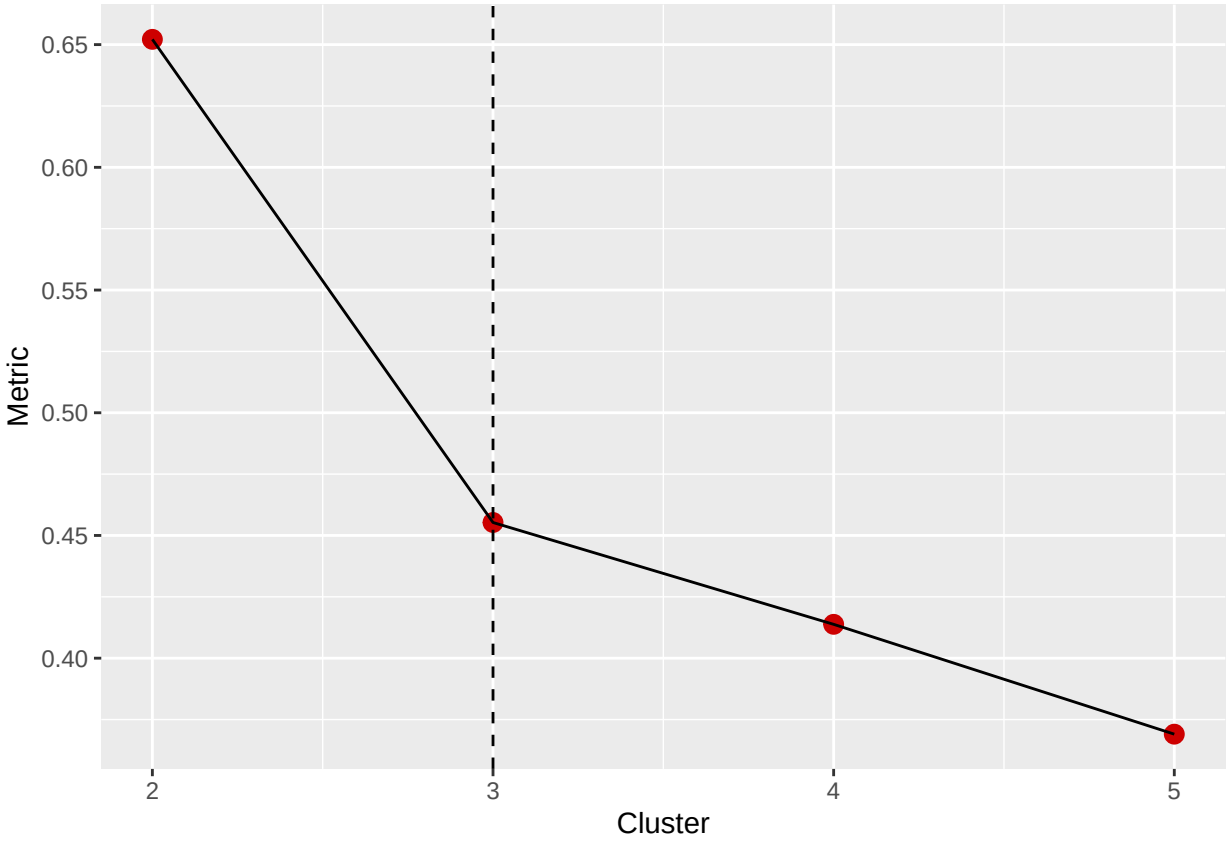
```
k_opt1 <- validation_kproto(method="silhouette", data=X_mat, lambda=lambdaest(X_mat), k=2:5,
                            kp_obj="optimal", nstart = 5, verbose = FALSE)
```

```
## Numeric variances:
##   DT DMFT   DS DMFS
##    1    1    1    1
## Average numeric variance: 1
##
## Heuristic for categorical variables: (method = 1)
##                 gender                   year fluoride_concentration
##              0.4987253              0.6857270              0.6418736
## Average categorical variation: 0.6087753
##
## Estimated lambda: 1.642642
```

```
saveRDS(k_opt1,"~/Library/Mobile Documents/com~apple~CloudDocs/Nam/New project/20 years of water fluori

k_opt1 <- readRDS("~/Library/Mobile Documents/com~apple~CloudDocs/Nam/New project/20 years of water flu

#
tibble(Cluster = c(2:5),
       Metric = as.vector(k_opt1$indices)) %>%
  ggplot(aes(x = Cluster,
             y = Metric)) +
  geom_point(size = 3,
             col ="red3") +
  geom_path() +
  geom_vline(xintercept = 3,
             linetype = 2)+
  scale_x_continuous(breaks = c(2:5))
```




```
#k=3
kpres <- kproto(x = X_mat,
                k = 3,
                lambda = lambdaest(X_mat), nstart = 5)
```

```
## Numeric variances:
##   DT DMFT   DS DMFS
##    1    1    1    1
## Average numeric variance: 1
##
## Heuristic for categorical variables: (method = 1)
##                 gender                   year fluoride_concentration
##              0.4987253              0.6857270              0.6418736
## Average categorical variation: 0.6087753
##
## Estimated lambda: 1.642642
##
## # NAs in variables:
##                 gender                     DT                   DMFT
##                      0                      0                      0
##                     DS                   DMFS                   year
##                      0                      0                      0
## fluoride_concentration
##                      0
## 0 observation(s) with NAs.
##
```

```
## # NAs in variables:
##                  gender                      DT                    DMFT
##                       0                       0                       0
##                      DS                    DMFS                    year
##                       0                       0                       0
## fluoride_concentration
##                       0
## 0 observation(s) with NAs.
##
## # NAs in variables:
##                  gender                      DT                    DMFT
##                       0                       0                       0
##                      DS                    DMFS                    year
##                       0                       0                       0
## fluoride_concentration
##                       0
## 0 observation(s) with NAs.
##
## # NAs in variables:
##                  gender                      DT                    DMFT
##                       0                       0                       0
##                      DS                    DMFS                    year
##                       0                       0                       0
## fluoride_concentration
##                       0
## 0 observation(s) with NAs.
##
## # NAs in variables:
##                  gender                      DT                    DMFT
##                       0                       0                       0
##                      DS                    DMFS                    year
##                       0                       0                       0
## fluoride_concentration
##                       0
## 0 observation(s) with NAs.
```

```
kpres
```

```
## Numeric predictors: 4
## Categorical predictors: 3
## Lambda: 1.642642
##
## Number of Clusters: 3
## Cluster sizes: 1764 1244 1765
## Within cluster error: 3594.863 2781.844 6718.473
##
## Cluster prototypes:
##   gender         DT       DMFT         DS      DMFS year fluoride_concentration
## 1   male -0.6657260 -0.6294098 -0.6577884 -0.622106 2012                0.5_ppm
## 2 female -0.6748476 -0.6918046 -0.6713487 -0.687082 2003                0.7_ppm
## 3 female  1.1409921  1.1166481  1.1305929  1.106020 2012                  0_ppm
```

```
summary(kpres)
```

```
## gender
##
## cluster  male female
##       1 0.620  0.380
##       2 0.356  0.644
##       3 0.414  0.586
##
## ------------------------------------------------------------------
## DT
##         Min.    1st Qu.     Median       Mean    3rd Qu.      Max.
## 1 -0.8470149 -0.8470149 -0.8470149 -0.6657260 -0.8470149 1.1774777
## 2 -0.8470149 -0.8470149 -0.8470149 -0.6748476 -0.8470149 0.7573575
## 3 -0.8470149  0.7573575  1.1774777  1.1409921  1.5033482 3.2905042
##
## ------------------------------------------------------------------
## DMFT
##         Min.    1st Qu.     Median       Mean     3rd Qu.     Max.
## 1 -0.9298146 -0.9298146 -0.9298146 -0.6294098  0.04668298 1.594395
## 2 -0.9298146 -0.9298146 -0.9298146 -0.6918046 -0.92981460 1.337543
## 3 -0.9298146  0.6178974  1.0231806  1.1166481  1.33754256 3.061583
##
## ------------------------------------------------------------------
## DS
##        Min.    1st Qu.     Median       Mean   3rd Qu.      Max.
## 1 -0.826115 -0.8261150 -0.8261150 -0.6577884 -0.826115 1.4291039
## 2 -0.826115 -0.8261150 -0.8261150 -0.6713487 -0.826115 0.9187606
## 3 -0.826115  0.5566662  0.9187606  1.1305929  1.623127 3.4548417
##
## ------------------------------------------------------------------
## DMFS
##          Min.    1st Qu.     Median      Mean     3rd Qu.     Max.
## 1 -0.90239862 -0.9023986 -0.9023986 -0.622106 -0.07058007 1.432811
## 2 -0.90239862 -0.9023986 -0.9023986 -0.687082 -0.90239862 2.424876
## 3 -0.07058007  0.7612385  1.0290242  1.106020  1.43281129 3.329447
##
## ------------------------------------------------------------------
## year
##
## cluster  1990  2003  2012  2019
##       1 0.049 0.011 0.782 0.158
##       2 0.056 0.760 0.047 0.137
##       3 0.239 0.224 0.380 0.156
##
## ------------------------------------------------------------------
## fluoride_concentration
##
## cluster 0_ppm 0.5_ppm 0.7_ppm
##       1 0.154   0.846   0.000
##       2 0.264   0.092   0.645
##       3 0.642   0.244   0.114
##
```

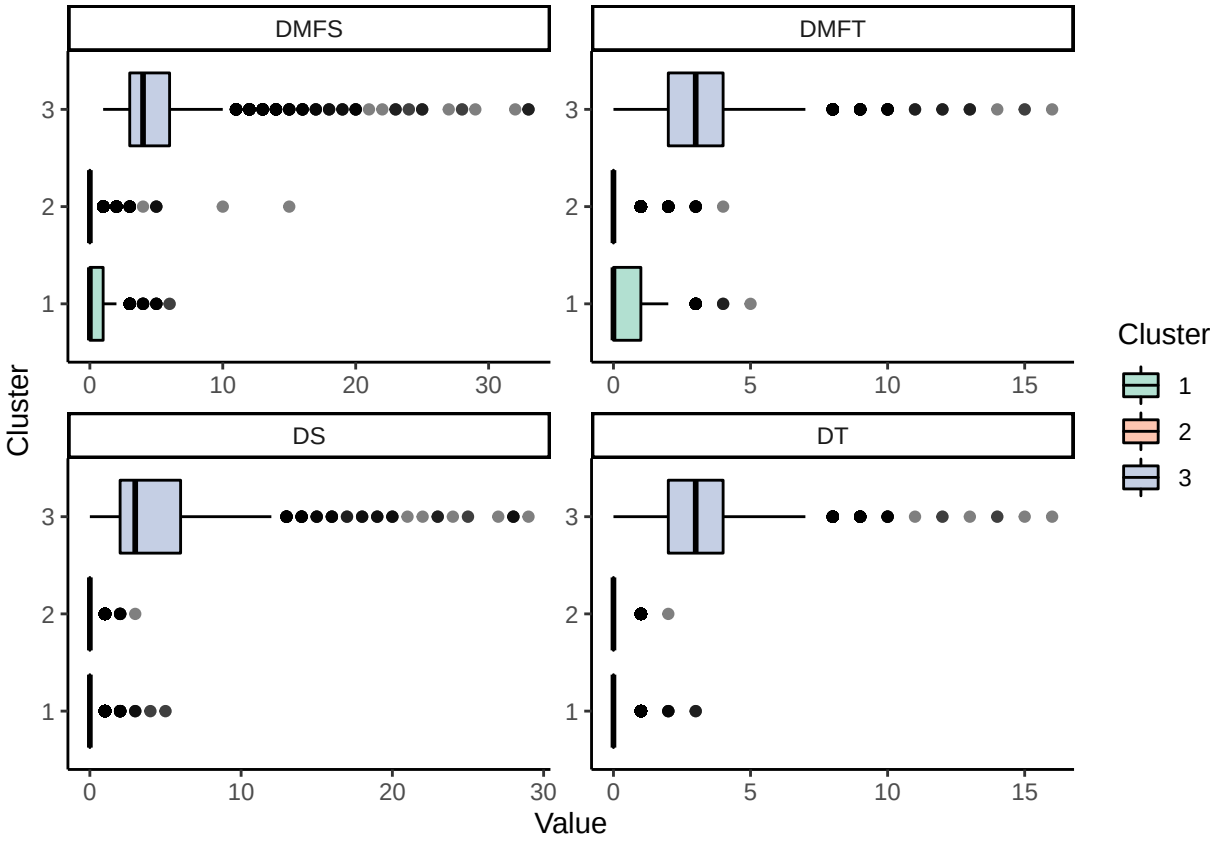
```
## ------------------------------------------------------------------
```

```
saveRDS(kpres,"~/Library/Mobile Documents/com~apple~CloudDocs/Nam/New project/20 years of water fluorida
kpres <- readRDS("~/Library/Mobile Documents/com~apple~CloudDocs/Nam/New project/20 years of water fluo
```

Plot

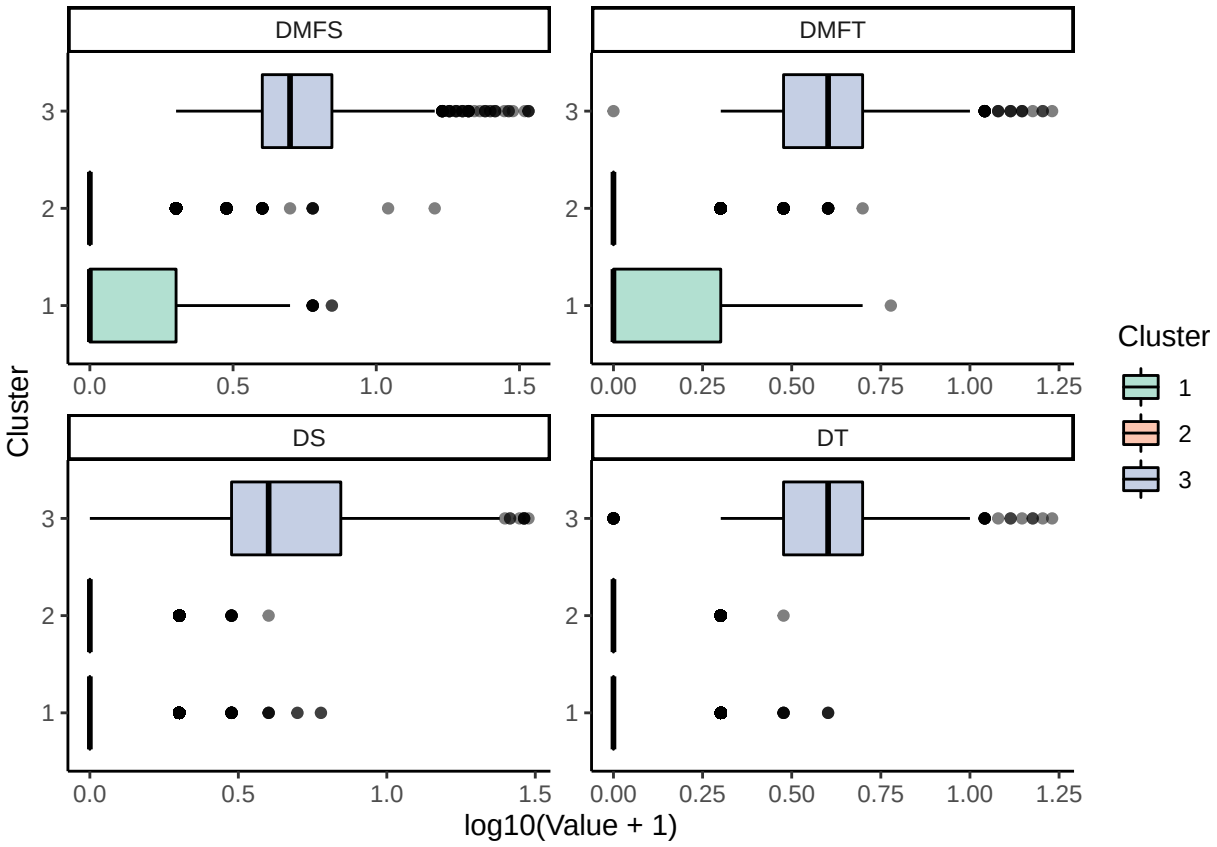
```
valid_df <- data_b1 %>% mutate(Cluster = as.factor(kpres$cluster))

valid_df %>%gather(c(DT:DMFS),
                   key = "Parameter",
                   value="Value")%>%
  ggplot(aes(x = Cluster, y=Value, fill = Cluster))+
  geom_boxplot(alpha=0.5,col="black")+
  facet_wrap(~Parameter,ncol=2,scales = "free")+
  coord_flip()+
  scale_fill_manual(values = brewer.pal(n = 4, name = "Set2"))+ theme_classic()
```




```
#transformation
valid_df %>% gather(c(DT:DMFS),
                   key = "Parameter",
                   value="Value")%>%
```

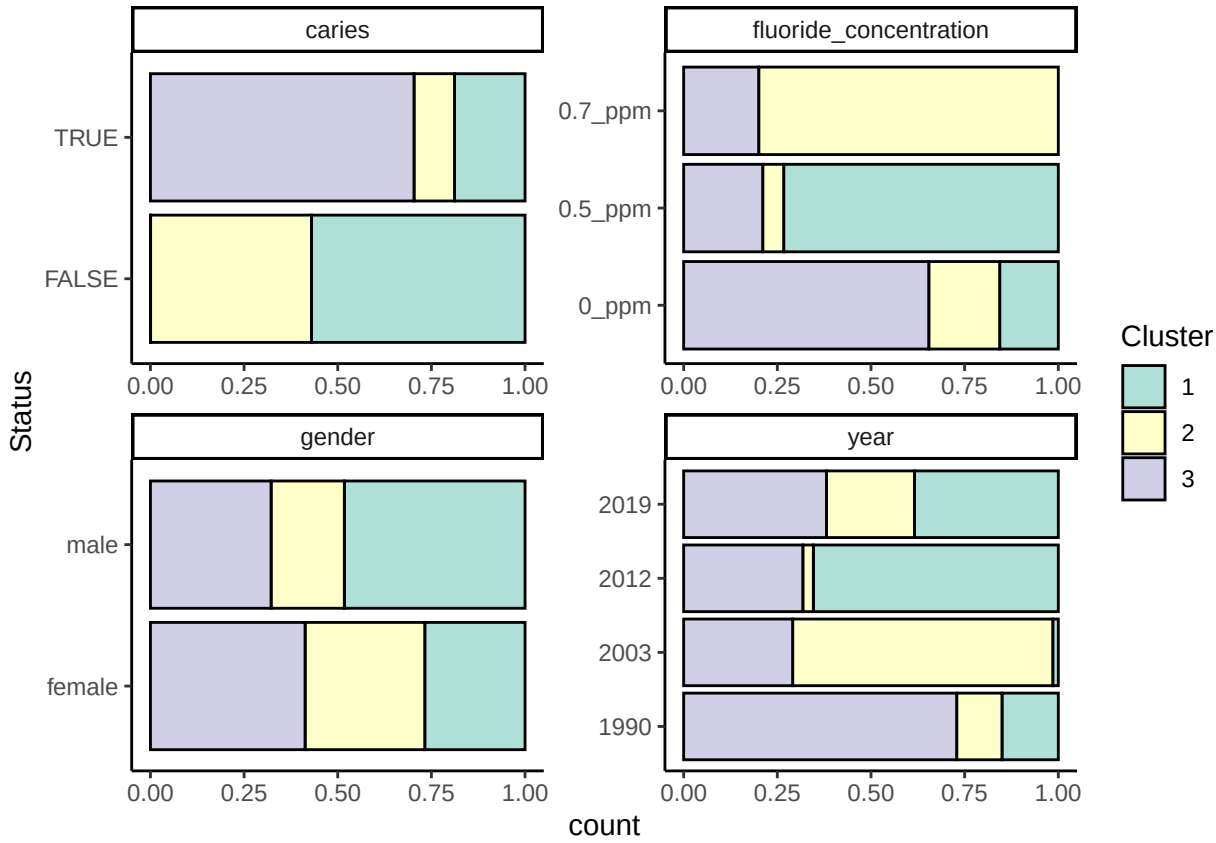
```
  ggplot(aes(x = Cluster, y=log10(Value+1), fill = Cluster))+
  geom_boxplot(alpha=0.5,col="black")+
  facet_wrap(~Parameter,ncol=2,scales = "free")+
  coord_flip()+
  scale_fill_manual(values = brewer.pal(n = 4, name = "Set2"))+ theme_classic()
```



Plot

```
valid_df %>% gather(gender, fluoride_concentration, caries, year,
                    key="Factor",value="Status")%>%
  ggplot(aes(x=Status,fill=Cluster))+
  geom_bar(stat = "count",position="fill",show.legend = T,alpha=0.7,col="black")+
  coord_flip()+
  facet_wrap(~Factor,ncol=2,scales = "free")+
  scale_fill_manual(values = brewer.pal(n = 4, name = "Set3"))+ theme_classic()
```

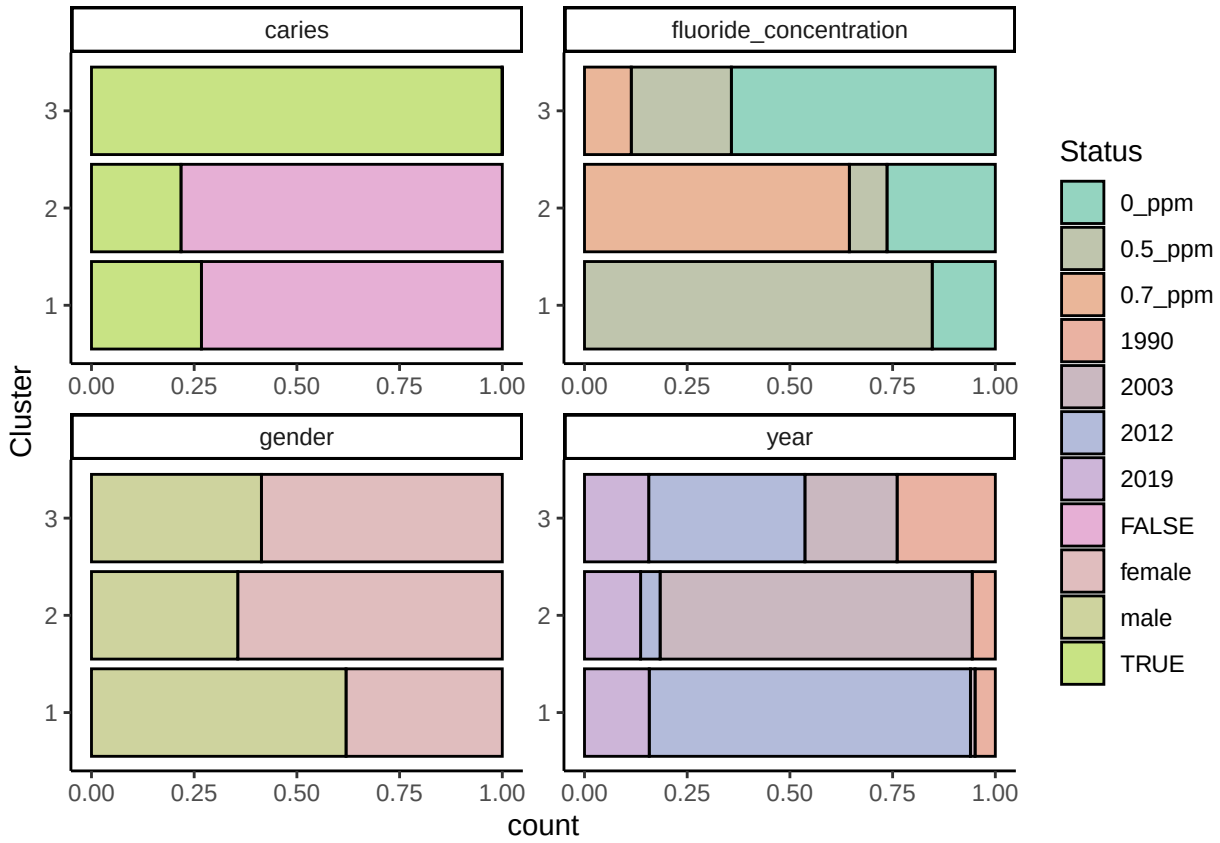
```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```




```
valid_df %>% gather(gender, fluoride_concentration, caries, year,
                    key="Factor",value="Status")%>%
  ggplot(aes(x=Cluster,fill= Status))+
  geom_bar(stat = "count",position="fill",show.legend = T,alpha=0.7,col="black")+
  coord_flip()+
  facet_wrap(~Factor,ncol=2,scales = "free")+
  scale_fill_manual(values = colorRampPalette(brewer.pal(n = 12, name = "Set2"))(18)) + theme_classic()
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
## Warning in brewer.pal(n = 12, name = "Set2"): n too large, allowed maximum for palette Set2 is 8
## Returning the palette you asked for with that many colors
```



#grid.col = c(1 =“#f7286d”, 2 =“#1faae0”, TRUE= “#2968c2”, FALSE= “#97c425”,)

```
library(circlize)
```

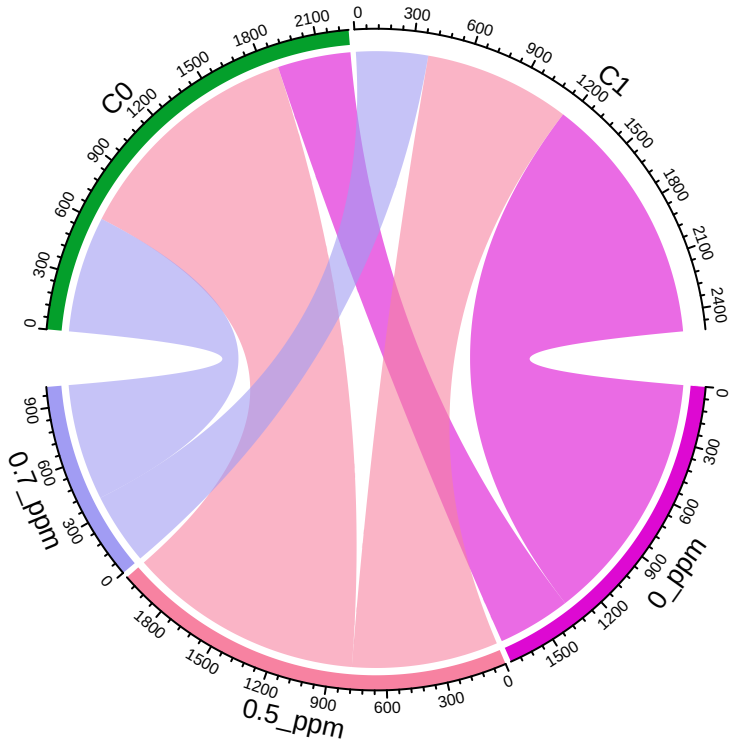
```
## ========================================
## circlize version 0.4.13
## CRAN page: https://cran.r-project.org/package=circlize
## Github page: https://github.com/jokergoo/circlize
## Documentation: https://jokergoo.github.io/circlize_book/book/
##
## If you use it in published research, please cite:
## Gu, Z. circlize implements and enhances circular visualization
##   in R. Bioinformatics 2014.
##
## This message can be suppressed by:
##   suppressPackageStartupMessages(library(circlize))
## ========================================
```

```
#caries, fluoride_concentration
xtb_y <- valid_df %>%
    mutate(Id=rownames(valid_df))%>%
    gather(caries,
           key="Pathology",
           value="Diagnosis")%>%
  group_by(fluoride_concentration,Diagnosis,Pathology)%>%
  summarise(frequency = n())
```

```
## `summarise()` has grouped output by 'fluoride_concentration', 'Diagnosis'. You can override using the
```

```
xtb_y$Diagnosis <- ifelse(xtb_y$Pathology == "caries" & xtb_y$Diagnosis == "FALSE", "C0",
                          ifelse(xtb_y$Pathology == "caries" & xtb_y$Diagnosis == "TRUE", "C1",
                                 ifelse(xtb_y$Pathology == "fluorosis" & xtb_y$Diagnosis == "FALSE", "F
xtb_y <- xtb_y%>%.[,c(1,2,4)]

chordDiagram(as.data.frame(xtb_y),
             transparency = 0.4,
#             grid.col = grid.col,
             column.col ="black")
```

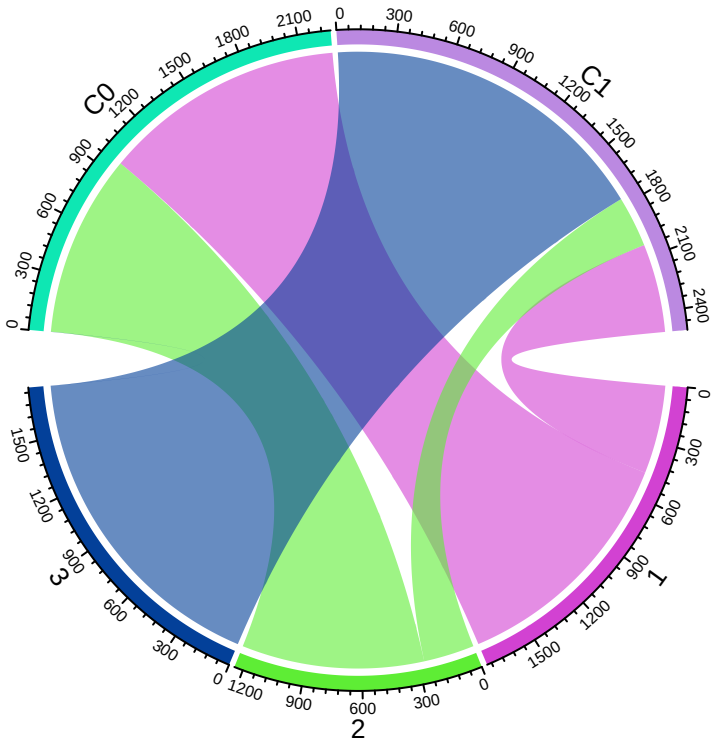



```
#caries, Cluster
xtb_y <- valid_df %>%
    mutate(Id=rownames(valid_df))%>%
    gather(caries,
           key="Pathology",
           value="Diagnosis")%>%
  group_by(Cluster,Diagnosis,Pathology)%>%
  summarise(frequency = n())
```

```
## `summarise()` has grouped output by 'Cluster', 'Diagnosis'. You can override using the `.groups` argu
```

```
xtb_y$Diagnosis <- ifelse(xtb_y$Pathology == "caries" & xtb_y$Diagnosis == "FALSE", "C0",
                          ifelse(xtb_y$Pathology == "caries" & xtb_y$Diagnosis == "TRUE", "C1",
                                 ifelse(xtb_y$Pathology == "fluorosis" & xtb_y$Diagnosis == "FALSE", "F
xtb_y <- xtb_y%>%.[,c(1,2,4)]

chordDiagram(as.data.frame(xtb_y),
             transparency = 0.4,
 #            grid.col = grid.col,
             column.col ="black")
```



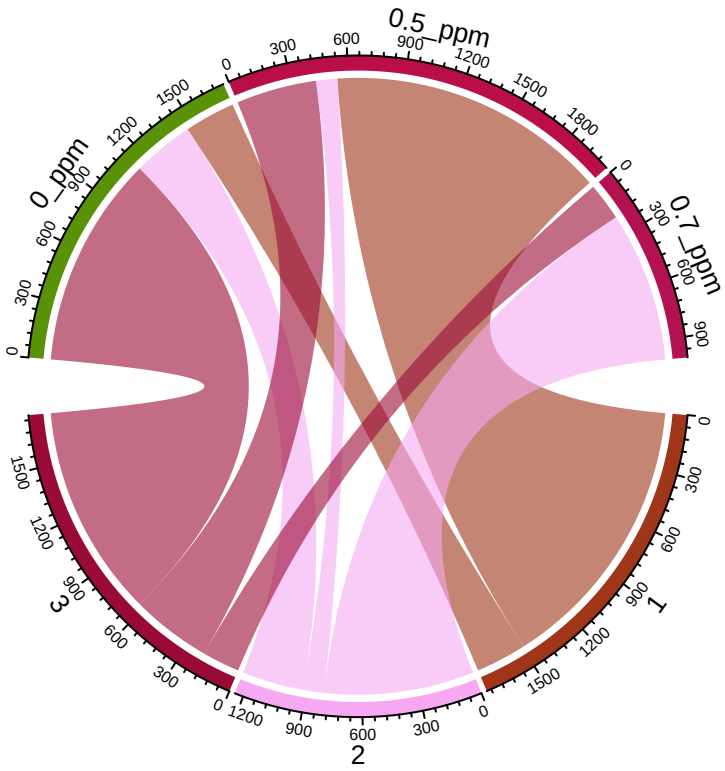

```
#fluoride_concentration, Cluster
xtb_y <- valid_df %>%
    mutate(Id=rownames(valid_df))%>%
    gather(fluoride_concentration,
           key="Treatment",
           value="Dose")%>%
  group_by(Cluster,Dose,Treatment)%>%
  summarise(frequency = n())
```

```
## `summarise()` has grouped output by 'Cluster', 'Dose'. You can override using the `.groups` argument
```

```
xtb_y <- xtb_y%>%.[,c(1,2,4)]

chordDiagram(as.data.frame(xtb_y),
```

```
            transparency = 0.4,
#            grid.col = grid.col,
            column.col ="black")
```



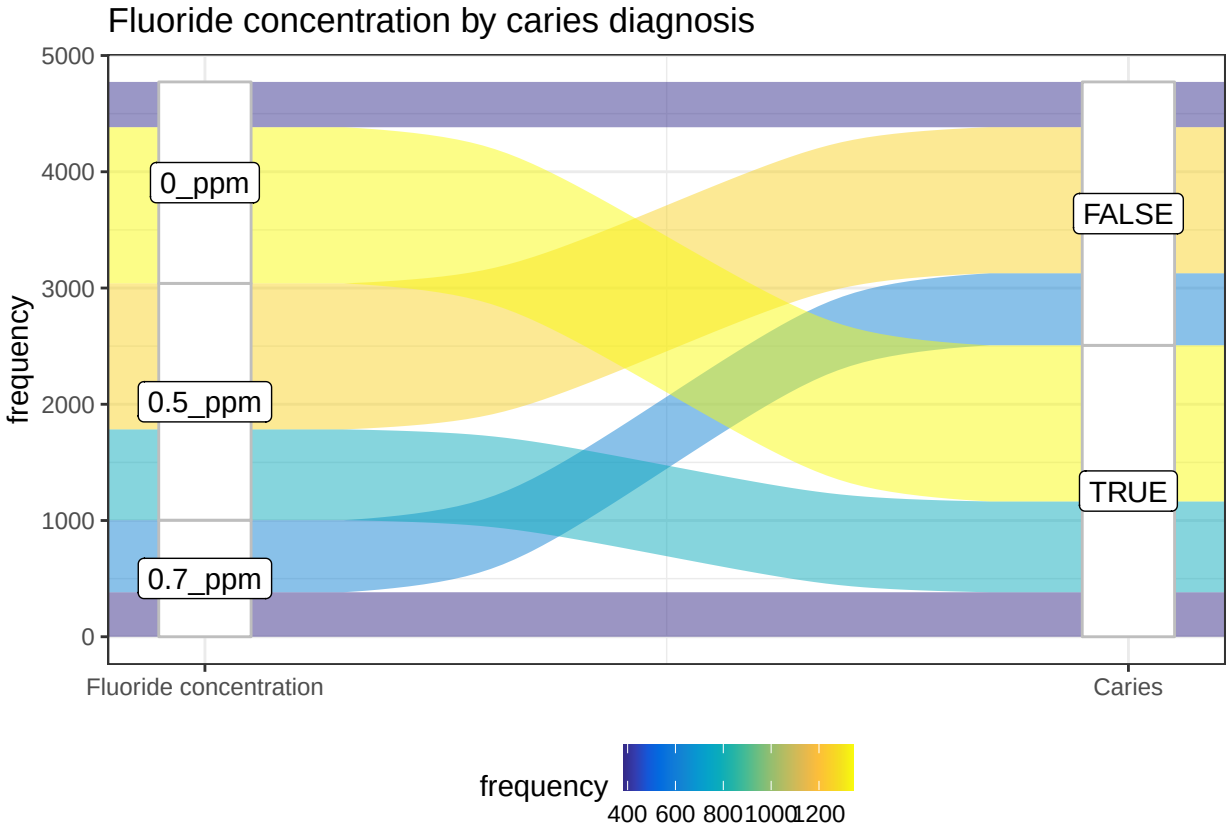

```
library(ggalluvial)
library(pals)
#
xtb_y2 <- valid_df %>%
  dplyr::group_by(caries, fluoride_concentration)%>%
  summarise(frequency = n())
```

```
## `summarise()` has grouped output by 'caries'. You can override using the `.groups` argument.
```

```
xtb_y2 %>% ggplot(aes(y = frequency,
                  axis1= fluoride_concentration,
                  axis2 = caries,
                  )) +
  geom_alluvium(aes(fill = frequency), width = 0.3) +
  geom_stratum(width = 1/10,
               fill = "white",
               color = "grey") +
  geom_label(stat = "stratum",
             infer.label = TRUE) +
  scale_x_continuous(breaks = 1:2,
```

```
                    labels = c("Fluoride concentration", "Caries")) +
  scale_fill_gradientn(colours = pals::parula(n=500))+
  theme_bw()+theme(legend.position="bottom") +
  ggtitle("Fluoride concentration by caries diagnosis")
```

```
## Warning: The parameter 'infer.label' is deprecated.
## Use 'aes(label = after_stat(stratum))'.
```

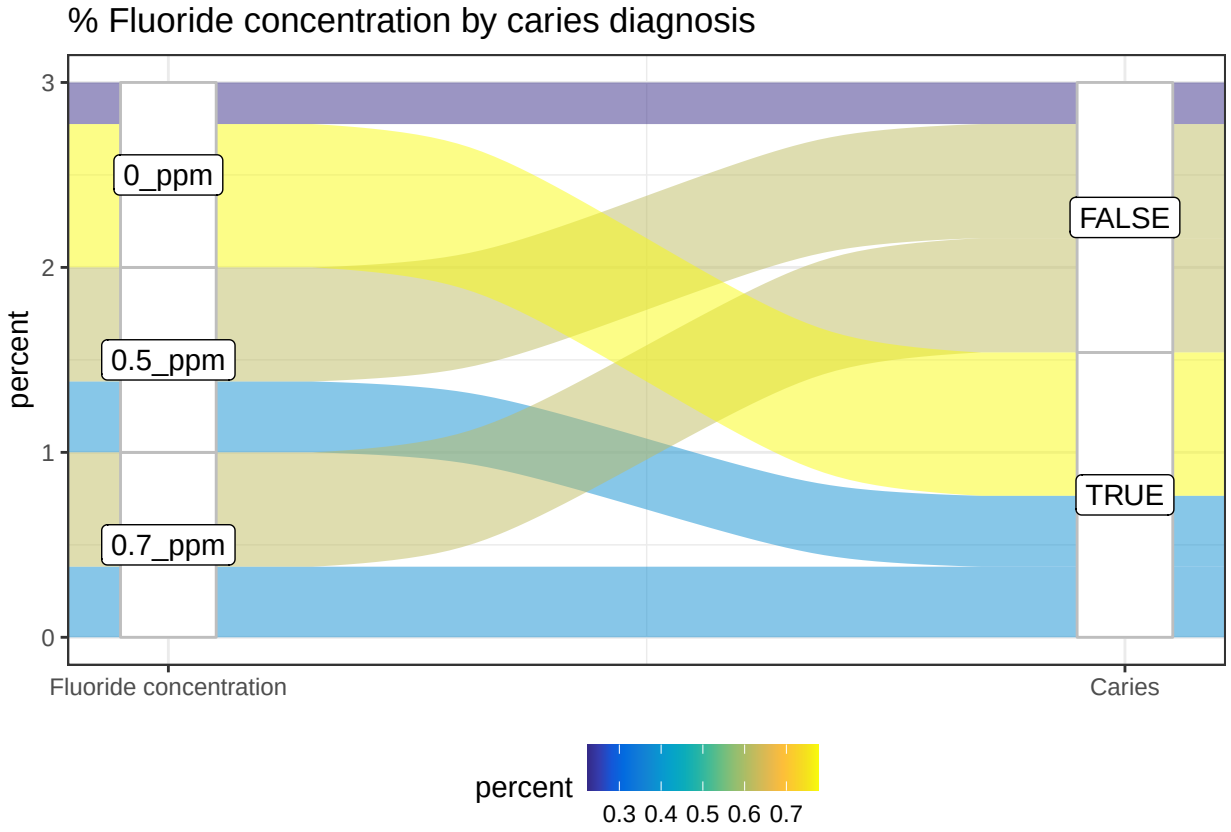



```
#
xtb_y2 <- xtb_y2 %>%
  dplyr::group_by(fluoride_concentration)%>% mutate(percent = frequency/sum(frequency))
xtb_y2$sum <- xtb_y2$frequency/xtb_y2$percent

xtb_y2 %>% ggplot(aes(y = percent,
                      axis1= fluoride_concentration,
                      axis2 = caries,
                      )) +
  geom_alluvium(aes(fill = percent), width = 0.3) +
  geom_stratum(width = 1/10,
               fill = "white",
               color = "grey") +
  geom_label(stat = "stratum",
             infer.label = TRUE) +
  scale_x_continuous(breaks = 1:2,
```

```
                      labels = c("Fluoride concentration", "Caries")) +
  scale_fill_gradientn(colours = pals::parula(n=500))+
  theme_bw()+theme(legend.position="bottom") +
  ggtitle("% Fluoride concentration by caries diagnosis")
```

```
## Warning: The parameter 'infer.label' is deprecated.
## Use 'aes(label = after_stat(stratum))'.
```



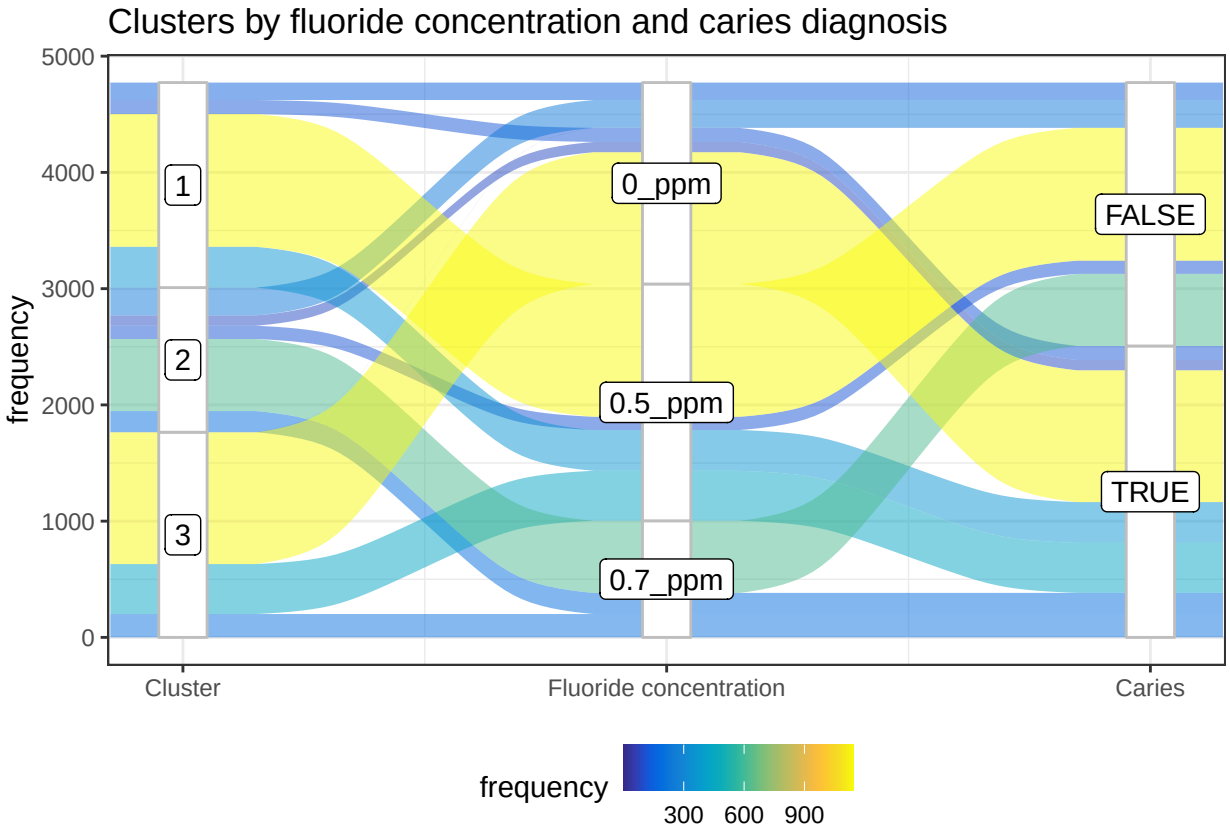

```
#
xtb_y3 <- valid_df %>%
  dplyr::group_by(caries, fluoride_concentration, Cluster)%>%
  summarise(frequency = n())
```

```
## 'summarise()' has grouped output by 'caries', 'fluoride_concentration'. You can override using the '
```

```
xtb_y3 %>% ggplot(aes(y = frequency,
                      axis1 = Cluster,
                      axis2= fluoride_concentration,
                      axis3 = caries,
                      )) +
  geom_alluvium(aes(fill = frequency), width = 0.3) +
  geom_stratum(width = 1/10,
               fill = "white",
               color = "grey") +
```

```
  geom_label(stat = "stratum",
             infer.label = TRUE) +
  scale_x_continuous(breaks = 1:3,
                     labels = c("Cluster", "Fluoride concentration", "Caries")) +
  scale_fill_gradientn(colours = pals::parula(n=500))+
  theme_bw()+theme(legend.position="bottom") +
  ggtitle("Clusters by fluoride concentration and caries diagnosis")
```

```
## Warning: The parameter 'infer.label' is deprecated.
## Use 'aes(label = after_stat(stratum))'.
```




```
#
xtb_y3 <- xtb_y3 %>%
  dplyr::group_by(Cluster)%>% mutate(percent = frequency/sum(frequency))
xtb_y3$sum <- xtb_y3$frequency/xtb_y3$percent

xtb_y3 %>% ggplot(aes(y = percent,
                  axis1 = Cluster,
                  axis2= fluoride_concentration,
                  axis3 = caries,
                  )) +
  geom_alluvium(aes(fill = percent), width = 0.3) +
  geom_stratum(width = 1/10,
               fill = "white",
```

```
              color = "grey") +
  geom_label(stat = "stratum",
             infer.label = TRUE) +
  scale_x_continuous(breaks = 1:3,
                     labels = c("Cluster", "Fluoride concentration", "Caries")) +
  scale_fill_gradientn(colours = pals::parula(n=500))+
  theme_bw()+theme(legend.position="bottom") +
  ggtitle("% Clusters by fluoride concentration and caries diagnosis")
```

```
## Warning: The parameter ‘infer.label‘ is deprecated.
## Use ‘aes(label = after_stat(stratum))‘.
```