Outlier Detection

Kelvin Li

2022/3/19

```r
# DATA PREP ---------------------
Sys.setenv(LANG = "en")
library(tidyverse)
library(ggplot2)
library(GGally)
library(olsrr)
library(gridExtra)
library(cowplot)
library(factoextra)
library(caretEnsemble)
library(caret)
library(mlbench)
library(Metrics)
library(gtsummary)

dt <- read_csv('data_stat450.csv')
dt <- rename(dt, MEVA = 'Mean Evaporation (m_per_year)',
             MPEVA = 'Mean Potential Evaporation (m_per_year)',
             MSDEN = 'Mean Snow Density (kg3_per_m3)',
             MSDEP = "Mean Snow Depth (m)",
             MSDEPWEQ = "Mean Snow Depth, Snow Water Equiv (m_of_swe)",
             MSFAL = "Mean Snowfall (m_per_year)",
             MSMELT = "Mean Snowmelt (m_per_year)",
             MTEMP = "Mean Temperature (deg_C)",
             MPRECIP = "Mean Total Precip (m_per_year)",
             Q = "q_mean") |>
  mutate(gridcode = as.factor(gridcode),
         year = as.integer(year),
         MPEVA = MPEVA / 1000)

grid <- read_csv('grid_codes.csv') |>
  select(gridcode, Longitude, Latitude) |>
  mutate(gridcode = as.factor(gridcode))

dt <- left_join(grid,dt)
dt <- rename(dt, LONG = 'Longitude', LAT = 'Latitude')
dt.small <- dt |> select(year, gridcode, MEVA,MPEVA,MSDEN,MSDEP, MSDEPWEQ, MSFAL,MSMELT,MTEMP,MPRECIP,Q
# ---------------------
```

Outlier Detection

```
# SCALED EXPLANATORY VARS ---------------------
dt.valid <- na.omit(dt) |> mutate(gridcode = as.integer(gridcode))
outlier.count <- c()
outlier <- c()
for (i in 5:13){
qt <- quantile(dt.valid[,i],na.rm = TRUE)
iqr <- qt[4]-qt[2]
upper.bd <- qt[4]+iqr*1.5
lower.bd <- qt[2]-iqr*1.5
current.col.name = colnames(dt.valid[,i])
rename.year = as.symbol(paste(current.col.name, "year",sep="."))
rename.gridcode = as.symbol(paste(current.col.name, "gridcode",sep="."))
abnormal <- filter(dt.valid, !!as.symbol(current.col.name) > upper.bd | !!as.symbol(current.col.name) <
  select(!!rename.year:=year,
         !!rename.gridcode:=gridcode)
outlier.count <- append(outlier.count, dim(abnormal)[1])
outlier <- append(outlier, abnormal)
}
outlier.count <- as.data.frame(t(outlier.count))
colnames(outlier.count) <- colnames(dt.valid[,5:13])
outlier.count
```

```
##   MEVA MPEVA MSDEN MSDEP MSDEPWEQ MSFAL MSMELT MTEMP MPRECIP
## 1   16    66     0    12       13    10      7     5       7
```

```
outlier
```

```
## $MEVA.year
##  [1] 1983 1987 1988 1991 1995 1998 1999 2002 2005 2006 2007 2010 2011 2012 2016
## [16] 2017
##
## $MEVA.gridcode
##  [1] 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10
##
## $MPEVA.year
##  [1] 1985 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000
## [16] 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
## [31] 2016 2017 2018 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992
## [46] 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007
## [61] 2008 2009 2010 2011 2012 2013
##
## $MPEVA.gridcode
##  [1] 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17
## [26] 17 17 17 17 17 17 17 17  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
## [51]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##
## $MSDEN.year
## integer(0)
##
## $MSDEN.gridcode
## integer(0)
##
## $MSDEP.year
##  [1] 1992 1999 2000 2012 1991 1991 2007 1991 1996 1999 2007 1999
##
## $MSDEP.gridcode
##  [1] 17 17 17 17 15 19 19 20 20 20 20  5
##
## $MSDEPWEQ.year
##  [1] 1992 2000 2012 1991 1991 1996 2007 1991 1996 1999 2007 2010 1999
##
## $MSDEPWEQ.gridcode
##  [1] 17 17 17 15 19 19 19 20 20 20 20 20  5
##
## $MSFAL.year
##  [1] 1999 1990 1990 1999 2007 1990 1999 2003 1990 1999
##
## $MSFAL.gridcode
##  [1] 17 15 19 19 19 20 20 20  5  5
##
## $MSMELT.year
## [1] 2000 1991 1991 2007 1991 1999 1999
##
## $MSMELT.gridcode
## [1] 17 15 19 19 20 20  5
##
## $MTEMP.year
## [1] 1996 1982 1996 1998 2012
##
## $MTEMP.gridcode
## [1] 17  1  1 10 10
##
## $MPRECIP.year
## [1] 1998 1990 1990 1990 1999 2017 2019
##
## $MPRECIP.gridcode
## [1] 16 15 19  5  5 10 10
```