Mahalanobis Outliers

## -- Attaching packages ------------------------------------------------------------------------ tidyverse 1.3.0 --

## v ggplot2 3.2.1 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts --------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

## Removeing Outliers using the Mahalanobis distance

Find Outliers using Mahalanobis distance of “distance” ,“stocking\_rate”, “forest\_reserves”, “labourperhectare”, “milkperhectare”, “tfp”. Mahalanobis distance calculates the standard Euclidean distance of each point to the mean. It is unitless, scale-invariant and includes correlations (Wikipedia!! Scite correctly)

df <- c(14,32,12,36,37,40) #"distance" ,"stocking\_rate", "forest\_reserves", "labourperhectare", "milkperhectare", "tfp"  
  
summary(si.df[df])

## distance stocking\_rate forest\_reserves   
## Min. : 161.8 Min. : 0.009804 Min. :0.00000   
## 1st Qu.:10151.3 1st Qu.: 0.414235 1st Qu.:0.02603   
## Median :14224.9 Median : 0.663836 Median :0.05882   
## Mean :13793.5 Mean : 1.213961 Mean :0.08960   
## 3rd Qu.:19575.2 3rd Qu.: 1.052614 3rd Qu.:0.12500   
## Max. :25536.8 Max. :16.723810 Max. :0.42857   
## labourperhectare milkperhectare tfp   
## Min. : 0.02222 Min. : 44.02 Min. : 16.7   
## 1st Qu.: 0.15417 1st Qu.: 306.34 1st Qu.: 405.3   
## Median : 0.42123 Median : 505.93 Median : 1519.1   
## Mean : 1.95097 Mean : 657.83 Mean : 12987.5   
## 3rd Qu.: 1.28205 3rd Qu.: 804.78 3rd Qu.: 4862.3   
## Max. :40.84500 Max. :3332.61 Max. :620500.0

si.mhlnbs\_outl <- mahalanobis(si.df[,df],colMeans(si.df[,df]),cov(si.df[,df]))  
si.mhlnbs\_outl # Calculated Mahalanobis distances

## [1] 0.6436295 0.3074113 1.0541423 3.4515976 3.3566030 1.4645417  
## [7] 1.2018209 0.5400544 0.9946445 0.7935193 0.4985301 2.5592997  
## [13] 1.6052084 0.7785060 1.5412891 2.3799707 9.8614702 1.5906536  
## [19] 9.1586440 1.4920944 1.9514362 13.9021279 30.8747245 1.9160046  
## [25] 4.1092570 3.6552885 3.9827511 3.0249763 1.7713324 13.2845811  
## [31] 3.3484059 2.2235614 2.3426741 3.5048117 2.0433856 3.1991237  
## [37] 3.5395783 15.7153428 3.0168986 4.5405425 3.8719739 3.4915823  
## [43] 2.5922116 3.6138579 4.7855495 3.9067734 6.1824944 1.8604930  
## [49] 6.0565970 1.2569814 1.9115669 5.9271918 1.1260630 0.6669026  
## [55] 0.6131164 4.1863792 5.9698281 12.7691294 6.6989573 3.4532844  
## [61] 61.3768055 5.3532786 1.6436838 10.1301537 8.3304519 6.7266365  
## [67] 2.5883246 1.7081596 2.0646059 2.6550061 0.5079262 2.8421102  
## [73] 1.6062298 1.0906901 2.1529358 1.5376230 88.3946158 5.4220172  
## [79] 1.1017296 2.0040140 1.9240082 1.1016511 2.7867120 4.0111205  
## [85] 1.9186700 1.9435575 73.1577810 0.4612373 0.5443305 1.4034975  
## [91] 1.2588815 4.0238542 2.1293725 3.9369599

Add Mahalanobis distance to dataframe si.df

si.df$mhlnbs <- round(si.mhlnbs\_outl,3)  
summary(si.df$mhlnbs)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.307 1.539 2.574 5.936 4.088 88.395

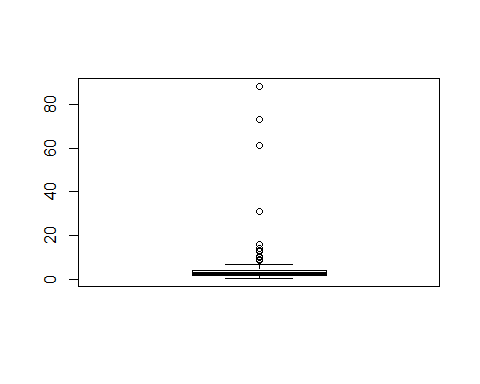
#save(si.df, file = "Dataframes/si.df.rda")

Boxplot Statistics of Mahalanobis distance

boxplot.stats(si.df$mhlnbs)

## $stats  
## [1] 0.3070 1.5380 2.5735 4.1090 6.7270  
##   
## $n  
## [1] 94  
##   
## $conf  
## [1] 2.154518 2.992482  
##   
## $out  
## [1] 9.861 9.159 13.902 30.875 13.285 15.715 12.769 61.377 10.130 8.330  
## [11] 88.395 73.158

boxplot(si.df$mhlnbs)



The boxplot identifies a Mahalanobis distance of bigger than 8.3 as outlier.

Create binomial variable of Mahalanobis Outliers, level 8.3 is chosen

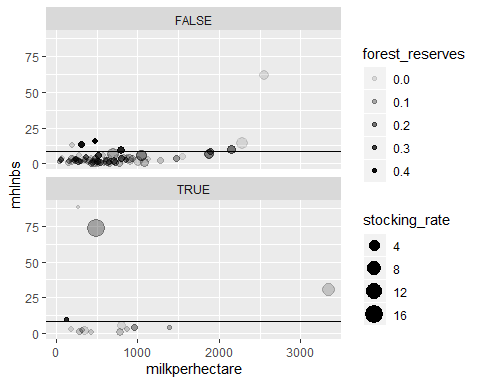
outlier <- 8.3  
si.df$outlier\_mhlnbs <- ifelse(si.df$mhlnbs > outlier, TRUE, FALSE)  
summary(si.df$outlier\_mhlnbs)

## Mode FALSE TRUE   
## logical 82 12

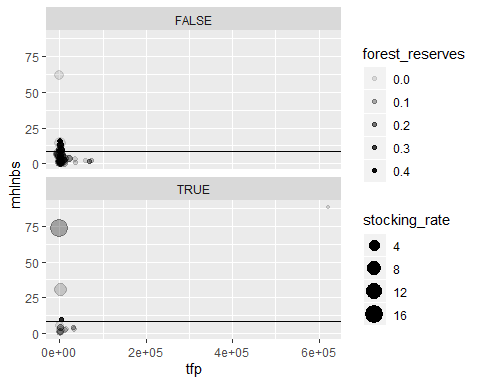
12 outliers are identified

Visualize Outliers - Point clouds, above are SPS=FALSE, below SPS=TRUE. y axis displays Mahalanobis distance, the horizontal line shows the 14 boundary.

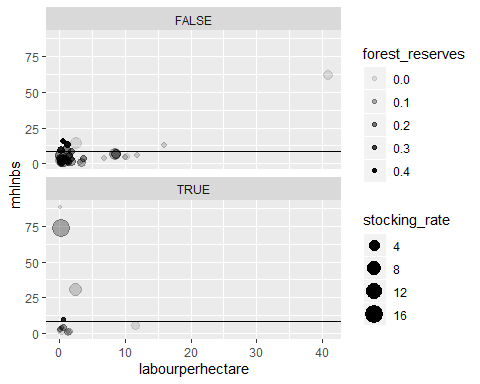
ggplot(data = si.df)+  
 geom\_point(mapping = aes(x=mhlnbs, y= milkperhectare, size=stocking\_rate, alpha=forest\_reserves))+  
 geom\_vline(xintercept = outlier)+  
 facet\_wrap(~SPS, dir = "v")+  
 coord\_flip()



ggplot(data = si.df)+  
 geom\_point(mapping = aes(x=mhlnbs, y= tfp, size=stocking\_rate, alpha=forest\_reserves))+  
 geom\_vline(xintercept = outlier)+  
 facet\_wrap(~SPS, dir = "v")+  
 coord\_flip()



ggplot(data = si.df,mapping = aes(x=mhlnbs, y= labourperhectare, size=stocking\_rate, alpha=forest\_reserves ))+  
 geom\_point()+  
 geom\_vline(xintercept = outlier)+  
 facet\_wrap(~SPS, dir = "v")+  
 coord\_flip()



Build subset without Mahalanobis Outliers

si.mhlnbsoutl <- filter(si.df, outlier\_mhlnbs==FALSE)  
save(si.mhlnbsoutl, file="Dataframes/si.mhlnbsoutl.rda" )

Boxplot Statistics of Mahalanobis distance

boxplot.stats(si.mhlnbsoutl$mhlnbs)

## $stats  
## [1] 0.307 1.403 2.097 3.540 6.727  
##   
## $n  
## [1] 82  
##   
## $conf  
## [1] 1.724132 2.469868  
##   
## $out  
## numeric(0)

boxplot(si.mhlnbsoutl$mhlnbs)

