Cancer Data Using R

Carbajal,Daniel

Sept 12, 2021

library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

## Set the working directory to the root of your DSC 520 directory  
setwd("C:/Users/dan35/Documents/DSC520\_GitRepository/dsc520/data/BreastCancer/")  
  
breastCancer\_df <- read.csv("breast\_cancer.csv")  
head(breastCancer\_df)

## id diagnosis radius\_mean texture\_mean perimeter\_mean area\_mean  
## 1 842302 M 17.99 10.38 122.80 1001.0  
## 2 842517 M 20.57 17.77 132.90 1326.0  
## 3 84300903 M 19.69 21.25 130.00 1203.0  
## 4 84348301 M 11.42 20.38 77.58 386.1  
## 5 84358402 M 20.29 14.34 135.10 1297.0  
## 6 843786 M 12.45 15.70 82.57 477.1  
## smoothness\_mean compactness\_mean concavity\_mean concave.points\_mean  
## 1 0.11840 0.27760 0.3001 0.14710  
## 2 0.08474 0.07864 0.0869 0.07017  
## 3 0.10960 0.15990 0.1974 0.12790  
## 4 0.14250 0.28390 0.2414 0.10520  
## 5 0.10030 0.13280 0.1980 0.10430  
## 6 0.12780 0.17000 0.1578 0.08089  
## symmetry\_mean fractal\_dimension\_mean radius\_se texture\_se perimeter\_se  
## 1 0.2419 0.07871 1.0950 0.9053 8.589  
## 2 0.1812 0.05667 0.5435 0.7339 3.398  
## 3 0.2069 0.05999 0.7456 0.7869 4.585  
## 4 0.2597 0.09744 0.4956 1.1560 3.445  
## 5 0.1809 0.05883 0.7572 0.7813 5.438  
## 6 0.2087 0.07613 0.3345 0.8902 2.217  
## area\_se smoothness\_se compactness\_se concavity\_se concave.points\_se  
## 1 153.40 0.006399 0.04904 0.05373 0.01587  
## 2 74.08 0.005225 0.01308 0.01860 0.01340  
## 3 94.03 0.006150 0.04006 0.03832 0.02058  
## 4 27.23 0.009110 0.07458 0.05661 0.01867  
## 5 94.44 0.011490 0.02461 0.05688 0.01885  
## 6 27.19 0.007510 0.03345 0.03672 0.01137  
## symmetry\_se fractal\_dimension\_se radius\_worst texture\_worst perimeter\_worst  
## 1 0.03003 0.006193 25.38 17.33 184.60  
## 2 0.01389 0.003532 24.99 23.41 158.80  
## 3 0.02250 0.004571 23.57 25.53 152.50  
## 4 0.05963 0.009208 14.91 26.50 98.87  
## 5 0.01756 0.005115 22.54 16.67 152.20  
## 6 0.02165 0.005082 15.47 23.75 103.40  
## area\_worst smoothness\_worst compactness\_worst concavity\_worst  
## 1 2019.0 0.1622 0.6656 0.7119  
## 2 1956.0 0.1238 0.1866 0.2416  
## 3 1709.0 0.1444 0.4245 0.4504  
## 4 567.7 0.2098 0.8663 0.6869  
## 5 1575.0 0.1374 0.2050 0.4000  
## 6 741.6 0.1791 0.5249 0.5355  
## concave.points\_worst symmetry\_worst fractal\_dimension\_worst  
## 1 0.2654 0.4601 0.11890  
## 2 0.1860 0.2750 0.08902  
## 3 0.2430 0.3613 0.08758  
## 4 0.2575 0.6638 0.17300  
## 5 0.1625 0.2364 0.07678  
## 6 0.1741 0.3985 0.12440

# Loop to graph histograms  
names<-names(breastCancer\_df)  
names2 = names(breastCancer\_df)  
classes<-sapply(breastCancer\_df,class)  
  
library(plyr)

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##   
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

breastCancer\_df$diagnosis <- revalue(breastCancer\_df$diagnosis, c("M"=1))  
breastCancer\_df$diagnosis <- revalue(breastCancer\_df$diagnosis, c("B"=0))  
breastCancer\_df$diagnosis<- as.numeric(breastCancer\_df$diagnosis)   
head(breastCancer\_df$diagnosis)

## [1] 1 1 1 1 1 1

n = 1  
for(name in names[classes == 'numeric'])  
{  
   
 cor\_data <- (cor(breastCancer\_df$diagnosis,breastCancer\_df[,name], use="pairwise.complete.obs"))  
 print(names[n])  
 print(cor\_data)  
 n = n + 1  
   
}

## [1] "id"  
## [1] 0.7300285  
## [1] "diagnosis"  
## [1] 0.4151853  
## [1] "radius\_mean"  
## [1] 0.7426355  
## [1] "texture\_mean"  
## [1] 0.7089838  
## [1] "perimeter\_mean"  
## [1] 0.35856  
## [1] "area\_mean"  
## [1] 0.5965337  
## [1] "smoothness\_mean"  
## [1] 0.6963597  
## [1] "compactness\_mean"  
## [1] 0.7766138  
## [1] "concavity\_mean"  
## [1] 0.3304986  
## [1] "concave.points\_mean"  
## [1] -0.0128376  
## [1] "symmetry\_mean"  
## [1] 0.5671338  
## [1] "fractal\_dimension\_mean"  
## [1] -0.008303333  
## [1] "radius\_se"  
## [1] 0.5561407  
## [1] "texture\_se"  
## [1] 0.5482359  
## [1] "perimeter\_se"  
## [1] -0.06701601  
## [1] "area\_se"  
## [1] 0.2929992  
## [1] "smoothness\_se"  
## [1] 0.2537298  
## [1] "compactness\_se"  
## [1] 0.4080423  
## [1] "concavity\_se"  
## [1] -0.006521756  
## [1] "concave.points\_se"  
## [1] 0.07797242  
## [1] "symmetry\_se"  
## [1] 0.7764538  
## [1] "fractal\_dimension\_se"  
## [1] 0.4569028  
## [1] "radius\_worst"  
## [1] 0.7829141  
## [1] "texture\_worst"  
## [1] 0.733825  
## [1] "perimeter\_worst"  
## [1] 0.4214649  
## [1] "area\_worst"  
## [1] 0.5909982  
## [1] "smoothness\_worst"  
## [1] 0.6596102  
## [1] "compactness\_worst"  
## [1] 0.793566  
## [1] "concavity\_worst"  
## [1] 0.4162943  
## [1] "concave.points\_worst"  
## [1] 0.3238722

library(reshape2)  
melt.df1<-melt(breastCancer\_df[3:12])

## No id variables; using all as measure variables

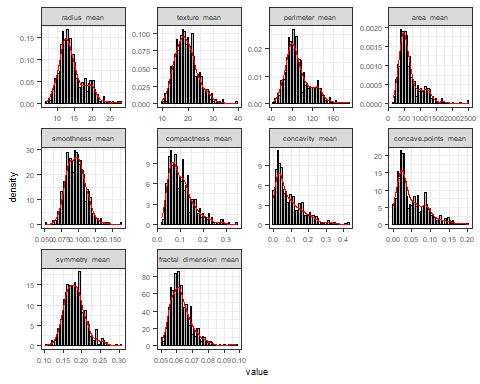
melt.df2<-melt(breastCancer\_df[13:22])

## No id variables; using all as measure variables

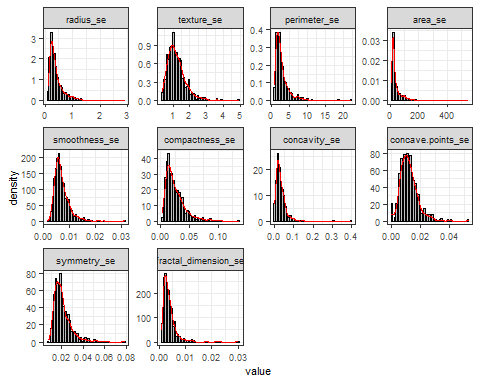
melt.df3<-melt(breastCancer\_df[23:32])

## No id variables; using all as measure variables

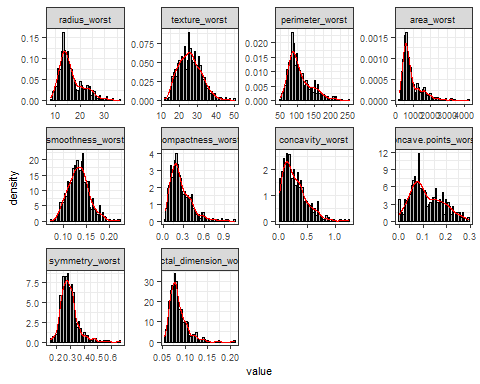
library(ggplot2)  
ggplot(melt.df1, aes(x = value)) +  
 geom\_histogram(aes(y = ..density..),col = "black", fill = "grey",bins=50,na.rm=TRUE) + geom\_density(colour = "red") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 7))



ggplot(melt.df2, aes(x = value)) +  
 geom\_histogram(aes(y = ..density..),col = "black", fill = "grey",bins=50,na.rm=TRUE) + geom\_density(colour = "red") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 8))

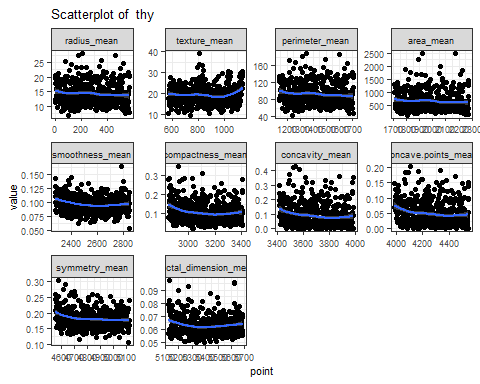


ggplot(melt.df3, aes(x = value)) +  
 geom\_histogram(aes(y = ..density..),col = "black", fill = "grey",bins=50,na.rm=TRUE) + geom\_density(colour = "red") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 8))



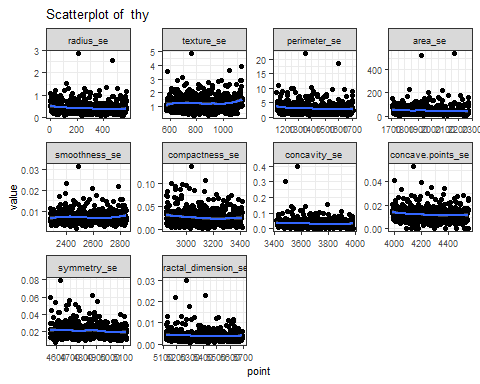
ggplot(melt.df1,aes(x = 1:5690, y = value)) + geom\_point() + geom\_smooth() + labs(title = paste("Scatterplot of ","thy"), x = "point", y = "value") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 8))

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



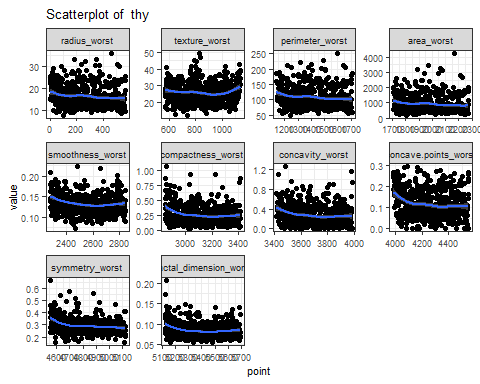
ggplot(melt.df2,aes(x = 1:5690, y = value)) + geom\_point() + geom\_smooth() + labs(title = paste("Scatterplot of ","thy"), x = "point", y = "value") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 8))

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



ggplot(melt.df3,aes(x = 1:5690, y = value)) + geom\_point() + geom\_smooth() + labs(title = paste("Scatterplot of ","thy"), x = "point", y = "value") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 8))

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



breastCancer\_Malignant\_df <- subset(breastCancer\_df, breastCancer\_df$diagnosis == 1)  
melt.Mdf1<-melt(breastCancer\_Malignant\_df[3:12])

## No id variables; using all as measure variables

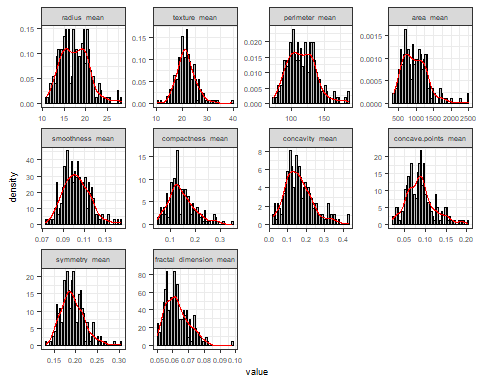
melt.Mdf2<-melt(breastCancer\_Malignant\_df[13:22])

## No id variables; using all as measure variables

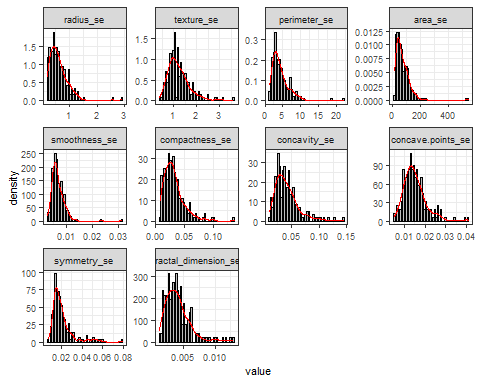
melt.Mdf3<-melt(breastCancer\_Malignant\_df[23:32])

## No id variables; using all as measure variables

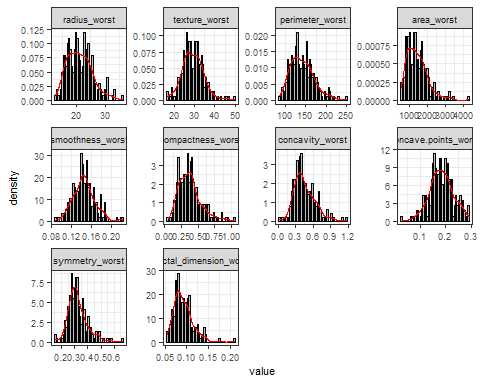
ggplot(melt.Mdf1, aes(x = value)) +  
 geom\_histogram(aes(y = ..density..),col = "black", fill = "grey",bins=50,na.rm=TRUE) + geom\_density(colour = "red") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 7))



ggplot(melt.Mdf2, aes(x = value)) +  
 geom\_histogram(aes(y = ..density..),col = "black", fill = "grey",bins=50,na.rm=TRUE) + geom\_density(colour = "red") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 8))



ggplot(melt.Mdf3, aes(x = value)) +  
 geom\_histogram(aes(y = ..density..),col = "black", fill = "grey",bins=50,na.rm=TRUE) + geom\_density(colour = "red") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 8))



breastCancer\_Benign\_df <- subset(breastCancer\_df, breastCancer\_df$diagnosis == 0)  
melt.Bdf1<-melt(breastCancer\_Benign\_df[3:12])

## No id variables; using all as measure variables

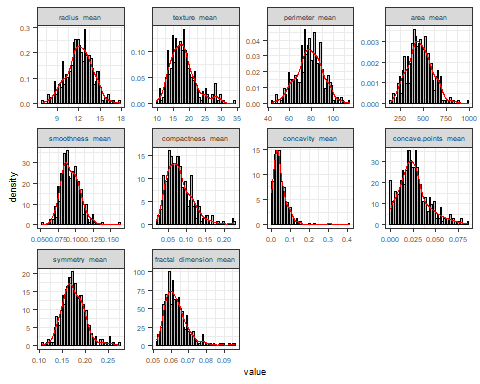
melt.Bdf2<-melt(breastCancer\_Benign\_df[13:22])

## No id variables; using all as measure variables

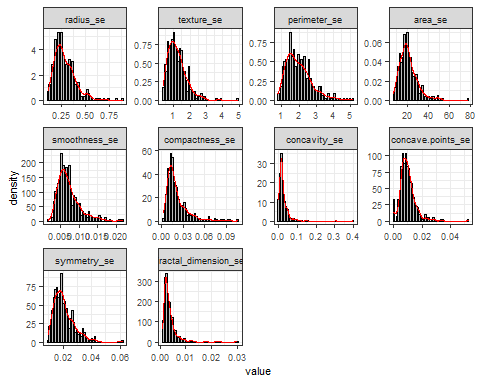
melt.Bdf3<-melt(breastCancer\_Benign\_df[23:32])

## No id variables; using all as measure variables

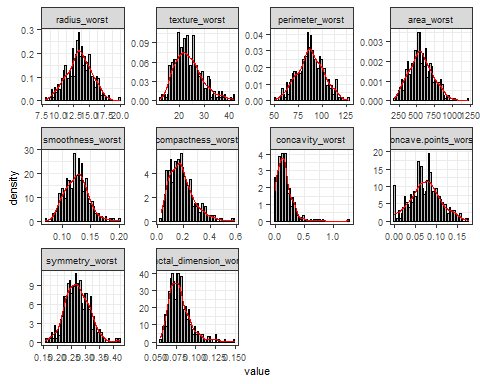
ggplot(melt.Bdf1, aes(x = value)) +  
 geom\_histogram(aes(y = ..density..),col = "black", fill = "grey",bins=50,na.rm=TRUE) + geom\_density(colour = "red") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 7))



ggplot(melt.Bdf2, aes(x = value)) +  
 geom\_histogram(aes(y = ..density..),col = "black", fill = "grey",bins=50,na.rm=TRUE) + geom\_density(colour = "red") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 8))

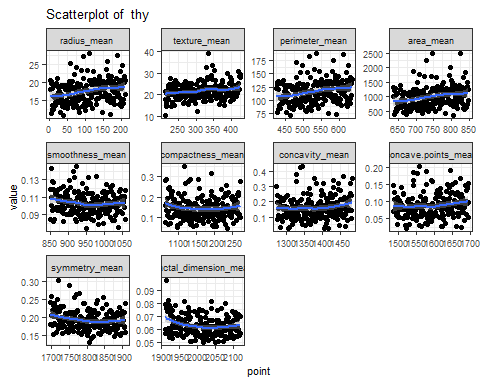


ggplot(melt.Bdf3, aes(x = value)) +  
 geom\_histogram(aes(y = ..density..),col = "black", fill = "grey",bins=50,na.rm=TRUE) + geom\_density(colour = "red") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 8))



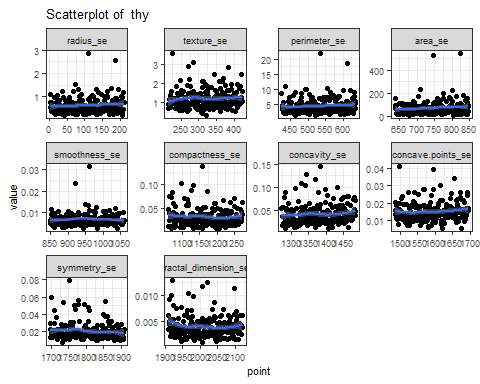
ggplot(melt.Mdf1,aes(x = 1:2120, y = value)) + geom\_point() + geom\_smooth() + labs(title = paste("Scatterplot of ","thy"), x = "point", y = "value") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 8))

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



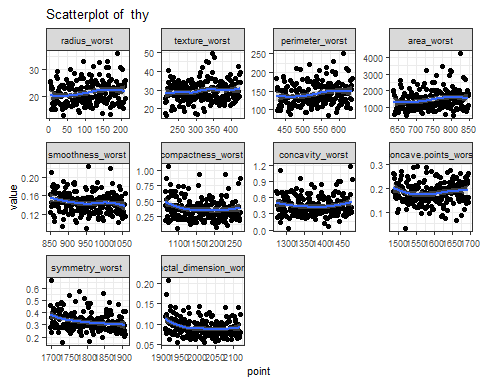
ggplot(melt.Mdf2,aes(x = 1:2120, y = value)) + geom\_point() + geom\_smooth() + labs(title = paste("Scatterplot of ","thy"), x = "point", y = "value") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 8))

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



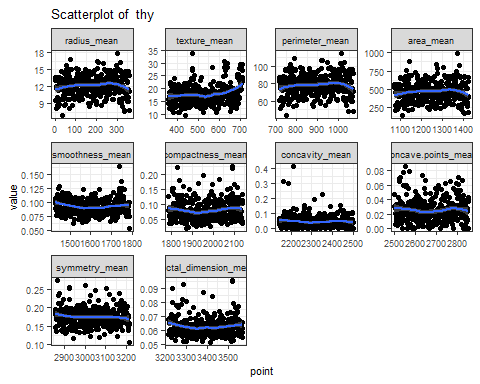
ggplot(melt.Mdf3,aes(x = 1:2120, y = value)) + geom\_point() + geom\_smooth() + labs(title = paste("Scatterplot of ","thy"), x = "point", y = "value") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 8))

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



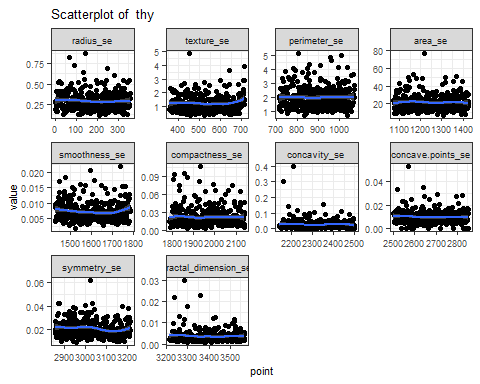
ggplot(melt.Bdf1,aes(x = 1:3570, y = value)) + geom\_point() + geom\_smooth() + labs(title = paste("Scatterplot of ","thy"), x = "point", y = "value") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 8))

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



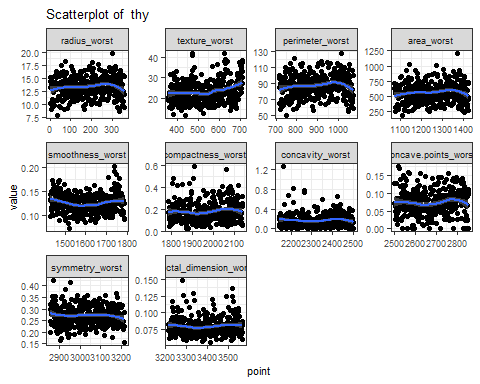
ggplot(melt.Bdf2,aes(x = 1:3570, y = value)) + geom\_point() + geom\_smooth() + labs(title = paste("Scatterplot of ","thy"), x = "point", y = "value") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 8))

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



ggplot(melt.Bdf3,aes(x = 1:3570, y = value)) + geom\_point() + geom\_smooth() + labs(title = paste("Scatterplot of ","thy"), x = "point", y = "value") + facet\_wrap(~variable, scales="free") + theme\_bw() + theme(text = element\_text(size = 8))

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



## Will now use logistic regression to will all variables to see which variables are significant  
  
## Set the working directory to the root of your DSC 520 directory  
setwd("C:/Users/dan35/Documents/DSC520\_GitRepository/dsc520/data/BreastCancer/")  
  
breastCancer\_df <- read.csv("breast\_cancer.csv")  
  
#breastCancer\_df$ndiagnosis[breastCancer\_df$diagnosis == "M"] <- 1  
#breastCancer\_df$ndiagnosis[breastCancer\_df$diagnosis == "B"] <- 0  
#breastCancer\_df$ndiagnosis <- factor(breastCancer\_df$ndiagnosis)  
  
library(plyr)  
breastCancer\_df$diagnosis <- revalue(breastCancer\_df$diagnosis, c("M"=1))  
breastCancer\_df$diagnosis <- revalue(breastCancer\_df$diagnosis, c("B"=0))  
breastCancer\_df$diagnosis<- as.numeric(breastCancer\_df$diagnosis)   
#breastCancer\_df$diagnosis<- factor(breastCancer\_df$diagnosis)   
  
glm\_bc0 <- glm(diagnosis ~ radius\_mean + texture\_mean + perimeter\_mean + area\_mean + smoothness\_mean + compactness\_mean + concavity\_mean +   
 symmetry\_mean + concavity\_se + fractal\_dimension\_mean + radius\_se + texture\_se + perimeter\_se + area\_se + smoothness\_se +   
 compactness\_se + symmetry\_se + fractal\_dimension\_se + radius\_worst + texture\_worst + perimeter\_worst + area\_worst +   
 smoothness\_worst + compactness\_worst + concavity\_worst + symmetry\_worst + fractal\_dimension\_worst, data=breastCancer\_df, family=binomial())

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(glm\_bc0)

##   
## Call:  
## glm(formula = diagnosis ~ radius\_mean + texture\_mean + perimeter\_mean +   
## area\_mean + smoothness\_mean + compactness\_mean + concavity\_mean +   
## symmetry\_mean + concavity\_se + fractal\_dimension\_mean + radius\_se +   
## texture\_se + perimeter\_se + area\_se + smoothness\_se + compactness\_se +   
## symmetry\_se + fractal\_dimension\_se + radius\_worst + texture\_worst +   
## perimeter\_worst + area\_worst + smoothness\_worst + compactness\_worst +   
## concavity\_worst + symmetry\_worst + fractal\_dimension\_worst,   
## family = binomial(), data = breastCancer\_df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5428 -0.0101 -0.0003 0.0000 4.0573   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.890e+01 4.190e+01 -1.645 0.1001   
## radius\_mean -1.082e+01 1.802e+01 -0.601 0.5481   
## texture\_mean -2.563e-01 3.278e-01 -0.782 0.4342   
## perimeter\_mean 1.174e+00 2.461e+00 0.477 0.6332   
## area\_mean 3.106e-02 6.924e-02 0.449 0.6538   
## smoothness\_mean 2.665e+02 1.708e+02 1.560 0.1188   
## compactness\_mean -9.127e+01 9.313e+01 -0.980 0.3271   
## concavity\_mean 6.310e+01 6.796e+01 0.928 0.3532   
## symmetry\_mean -2.546e+00 3.996e+01 -0.064 0.9492   
## concavity\_se -6.930e+01 8.199e+01 -0.845 0.3980   
## fractal\_dimension\_mean -1.142e+02 2.940e+02 -0.388 0.6977   
## radius\_se -1.553e+00 2.942e+01 -0.053 0.9579   
## texture\_se -4.062e+00 2.264e+00 -1.794 0.0727 .  
## perimeter\_se 1.367e+00 3.149e+00 0.434 0.6641   
## area\_se 2.080e-01 2.883e-01 0.721 0.4708   
## smoothness\_se 9.736e+02 4.682e+02 2.080 0.0376 \*  
## compactness\_se 1.170e+02 2.232e+02 0.524 0.6003   
## symmetry\_se -6.376e+01 2.084e+02 -0.306 0.7596   
## fractal\_dimension\_se -3.157e+03 1.910e+03 -1.653 0.0983 .  
## radius\_worst 4.942e+00 5.478e+00 0.902 0.3670   
## texture\_worst 8.249e-01 3.568e-01 2.312 0.0208 \*  
## perimeter\_worst -2.264e-01 5.465e-01 -0.414 0.6787   
## area\_worst -1.860e-02 5.316e-02 -0.350 0.7265   
## smoothness\_worst -9.353e+01 9.137e+01 -1.024 0.3060   
## compactness\_worst -2.014e+01 4.137e+01 -0.487 0.6264   
## concavity\_worst 1.775e+01 2.813e+01 0.631 0.5279   
## symmetry\_worst 2.199e+01 2.594e+01 0.848 0.3967   
## fractal\_dimension\_worst 3.553e+02 2.199e+02 1.616 0.1062   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 751.440 on 568 degrees of freedom  
## Residual deviance: 46.871 on 541 degrees of freedom  
## AIC: 102.87  
##   
## Number of Fisher Scoring iterations: 11

glm\_bc1<- glm(diagnosis ~ texture\_mean + perimeter\_mean + smoothness\_mean + compactness\_mean + symmetry\_se + radius\_worst + texture\_worst + area\_worst +   
 smoothness\_worst + compactness\_worst,data=breastCancer\_df,family=binomial())

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(glm\_bc1)

##   
## Call:  
## glm(formula = diagnosis ~ texture\_mean + perimeter\_mean + smoothness\_mean +   
## compactness\_mean + symmetry\_se + radius\_worst + texture\_worst +   
## area\_worst + smoothness\_worst + compactness\_worst, family = binomial(),   
## data = breastCancer\_df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0436 -0.0933 -0.0146 0.0005 3.7470   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -36.70833 13.72541 -2.674 0.00748 \*\*  
## texture\_mean 0.18000 0.16874 1.067 0.28608   
## perimeter\_mean -0.08195 0.10385 -0.789 0.43004   
## smoothness\_mean 91.21286 70.10038 1.301 0.19320   
## compactness\_mean -19.57726 27.15174 -0.721 0.47089   
## symmetry\_se 93.07391 48.39892 1.923 0.05447 .   
## radius\_worst 0.08847 1.33747 0.066 0.94726   
## texture\_worst 0.17045 0.12026 1.417 0.15639   
## area\_worst 0.02127 0.01354 1.571 0.11616   
## smoothness\_worst 48.84496 31.73906 1.539 0.12382   
## compactness\_worst 8.40333 6.37085 1.319 0.18716   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 751.440 on 568 degrees of freedom  
## Residual deviance: 90.478 on 558 degrees of freedom  
## AIC: 112.48  
##   
## Number of Fisher Scoring iterations: 9

anova(glm\_bc0, glm\_bc1, test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: diagnosis ~ radius\_mean + texture\_mean + perimeter\_mean + area\_mean +   
## smoothness\_mean + compactness\_mean + concavity\_mean + symmetry\_mean +   
## concavity\_se + fractal\_dimension\_mean + radius\_se + texture\_se +   
## perimeter\_se + area\_se + smoothness\_se + compactness\_se +   
## symmetry\_se + fractal\_dimension\_se + radius\_worst + texture\_worst +   
## perimeter\_worst + area\_worst + smoothness\_worst + compactness\_worst +   
## concavity\_worst + symmetry\_worst + fractal\_dimension\_worst  
## Model 2: diagnosis ~ texture\_mean + perimeter\_mean + smoothness\_mean +   
## compactness\_mean + symmetry\_se + radius\_worst + texture\_worst +   
## area\_worst + smoothness\_worst + compactness\_worst  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 541 46.871   
## 2 558 90.478 -17 -43.607 0.0003912 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# predict the test dataset  
pred1 <- predict(glm\_bc0, breastCancer\_df, type="response")   
  
predicted1 <- round(pred1) # round of the value; >0.5 will convert to 1 else 0  
# Creating a contigency table  
tab <- table(Predicted = predicted1, Reference = breastCancer\_df$diagnosis)  
tab

## Reference  
## Predicted 0 1  
## 0 354 5  
## 1 3 207

##Predicted FALSE TRUE  
##0 390 67  
##1 10 3  
  
  
accuracy1 <- (354+207) / (354+207+3+5)  
accuracy1\*100

## [1] 98.59402

# predict the test dataset  
pred2 <- predict(glm\_bc1, breastCancer\_df, type="response")   
  
predicted2 <- round(pred2) # round of the value; >0.5 will convert to 1 else 0  
# Creating a contigency table  
tab2<- table(Predicted = predicted2, Reference = breastCancer\_df$diagnosis)  
tab2

## Reference  
## Predicted 0 1  
## 0 353 10  
## 1 4 202

##Predicted FALSE TRUE  
##0 390 67  
##1 10 3  
##Reference  
  
accuracy2 <- (353+202) / (353+202+3+10)  
accuracy2\*100

## [1] 97.71127

## Run a clustering algorithm on the data   
  
library(gmodels)  
library(ggplot2)  
library(cluster)  
library(ClusterR)

## Warning: package 'ClusterR' was built under R version 4.0.4

## Loading required package: gtools

library(ggpubr)

## Warning: package 'ggpubr' was built under R version 4.0.4

##   
## Attaching package: 'ggpubr'

## The following object is masked from 'package:plyr':  
##   
## mutate

library(factoextra)

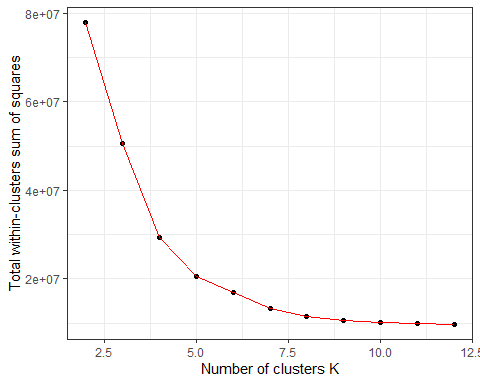
## Warning: package 'factoextra' was built under R version 4.0.4

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

##Fit the dataset using the k-means algorithm from k=2 to k=12. Create a scatter plot of the resultant clusters for each value of k.\*\*  
   
set.seed(20)  
k\_values <- c(2:12)  
total\_withinss <- NULL  
avg\_distance <- NULL  
for (i in 1:length(k\_values))  
{  
 temp\_df <- breastCancer\_df[2:32]  
 clusters <- kmeans(temp\_df, k\_values[i])  
 temp\_df$cluster <- as.factor(clusters$cluster)  
 x\_dist <- clusters$centers[temp\_df$cluster] - temp\_df$x  
 y\_dist <- clusters$centers[temp\_df$cluster] - temp\_df$y  
 tot\_dist <- sqrt((x\_dist \*\* 2) + (y\_dist \*\* 2))  
 avg\_distance <- c(avg\_distance, mean(tot\_dist))  
 total\_withinss <- c(total\_withinss, clusters$tot.withinss)  
}  
  
withinss\_elbow\_df <- data.frame(k\_values, total\_withinss)  
withinss\_elbow\_df

## k\_values total\_withinss  
## 1 2 77943168  
## 2 3 50516882  
## 3 4 29226594  
## 4 5 20535217  
## 5 6 16974353  
## 6 7 13339815  
## 7 8 11582727  
## 8 9 10648653  
## 9 10 10162710  
## 10 11 9919083  
## 11 12 9706689

ggplot(data = withinss\_elbow\_df, aes(x=k\_values, y=total\_withinss)) + xlab("Number of clusters K") + ylab("Total within-clusters sum of squares") + theme\_bw() + geom\_point() + geom\_line(color = "red")



kmodel <- kmeans(temp\_df, 5)  
kmodel$cluster = as.numeric(kmodel$cluster)  
temp\_df$cluster <- as.numeric(clusters$cluster)  
library(factoextra)  
fviz\_cluster(kmodel, data = temp\_df)

