LungCancer\_Survival\_NeminChen

Nemin

11/28/2022

### Import Data

clinical<-read.csv(file="/Users/neminchen/OneDrive/Stats/codes/clinical.csv")  
str(clinical)

## 'data.frame': 190 obs. of 16 variables:  
## $ ID : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Outcome : Factor w/ 2 levels "Alive","Dead": 1 2 2 2 2 2 2 2 2 1 ...  
## $ Survival.Months : num 9 19 13 15 10 11 13 13 19 9 ...  
## $ Age : int 67 73 72 69 76 62 72 72 72 83 ...  
## $ Grade : int 4 2 3 9 9 9 2 2 9 3 ...  
## $ Num.Primaries : int 0 0 0 1 0 0 0 0 0 0 ...  
## $ T : Factor w/ 9 levels "1","1a","1b",..: 9 9 4 2 9 7 8 7 8 9 ...  
## $ N : Factor w/ 5 levels "0","1","2","3",..: 3 3 3 1 5 3 5 5 5 5 ...  
## $ M : Factor w/ 3 levels "0","1","NULL": 3 1 1 2 3 3 2 1 1 3 ...  
## $ Radiation : int 0 5 0 0 0 0 5 0 0 5 ...  
## $ Stage : Factor w/ 9 levels "1B","IA","IB",..: 8 8 6 2 6 9 8 4 4 8 ...  
## $ Primary.Site : Factor w/ 9 levels "Both Lung","Left Hilar",..: 3 9 9 9 2 2 6 9 9 4 ...  
## $ Histology : Factor w/ 3 levels "Adenocarcinoma",..: 3 1 1 1 2 1 1 1 3 3 ...  
## $ Tumor.Size : Factor w/ 20 levels "1","1.4","1.5",..: 2 20 3 20 20 20 16 8 8 20 ...  
## $ Num.Mutated.Genes: int 8 2 1 4 3 4 4 2 3 3 ...  
## $ Num.Mutations : int 8 2 1 4 3 5 4 2 3 3 ...

table(clinical$Outcome, useNA = "always")

##   
## Alive Dead <NA>   
## 40 150 0

table(clinical$Survival.Months, useNA = "always")

##   
## 9 9.5 10 11 13 15 16 18 19 22 23 24 26 29 32   
## 7 3 23 27 21 7 8 2 6 7 6 1 1 3 11   
## 33 34 35 36 37 38 39 40 41 42 46 50 71 <NA>   
## 8 4 6 18 1 9 2 1 1 3 1 1 2 0

table(clinical$Age, useNA = "always")

##   
## 56 59 60 62 63 67 68 69 70 71 72 73 74 76 77   
## 5 1 1 27 8 22 4 13 10 14 20 13 7 26 4   
## 78 80 82 83 84 <NA>   
## 1 2 7 4 1 0

table(clinical$Grade, useNA = "always") #needs cleaning--too many missing

##   
## 2 3 4 9 <NA>   
## 29 22 43 96 0

table(clinical$Num.Primaries, useNA = "always")

##   
## 0 1 <NA>   
## 147 43 0

table(clinical$T, useNA = "always") #needs cleaning--too many missing

##   
## 1 1a 1b 2 2a 2b 3 4 UNK <NA>   
## 1 26 2 12 16 10 38 23 62 0

table(clinical$N, useNA = "always") #needs cleaning--too many missing

##   
## 0 1 2 3 NULL <NA>   
## 52 9 58 6 65 0

table(clinical$M, useNA = "always") #needs cleaning--too many missing

##   
## 0 1 NULL <NA>   
## 86 8 96 0

table(clinical$Radiation, useNA = "always") #needs cleaning

##   
## 0 5 <NA>   
## 127 63 0

table(clinical$Stage, useNA = "always")

##   
## 1B IA IB IIA IIB IIIA IIIB IV IVB <NA>   
## 1 32 1 8 11 43 24 45 25 0

table(clinical$Primary.Site, useNA = "always")

##   
## Both Lung Left Hilar Left Lower Lobe Left Upper Lobe   
## 5 31 17 21   
## Righ Upper Lobe Right Hilar Right Lower Lobe Right Middle Lobe   
## 2 33 25 3   
## Right Upper Lobe <NA>   
## 53 0

table(clinical$Histology, useNA = "always")

##   
## Adenocarcinoma Large-cell carcinoma Squamous cell carcinoma   
## 86 27 77   
## <NA>   
## 0

table(clinical$Tumor.Size, useNA = "always") #needs cleaning--too many missing

##   
## 1 1.4 1.5 1.6 1.8 1.9 10 2 2.5 3.5 3.6 4 4.4 5.3 5.4   
## 1 2 13 1 2 2 8 20 2 2 9 7 1 2 2   
## 5.5 8 8.5 9 NULL <NA>   
## 6 2 6 10 92 0

table(clinical$Num.Mutated.Genes, useNA = "always")

##   
## 0 1 2 3 4 5 6 7 8 <NA>   
## 6 33 53 55 21 15 4 1 2 0

table(clinical$Num.Mutations, useNA = "always")

##   
## 0 1 2 3 4 5 6 7 8 <NA>   
## 6 26 45 50 21 25 11 3 3 0

genomics<-read.csv(file="/Users/neminchen/OneDrive/Stats/codes/genomics.csv")  
str(genomics)

## 'data.frame': 510 obs. of 2 variables:  
## $ ID : int 1 158 88 132 18 62 6 17 38 49 ...  
## $ Gene: Factor w/ 50 levels "AKT1","ALK\_Col1",..: 1 1 2 2 3 3 4 4 4 4 ...

table(genomics$Gene)

##   
## AKT1 ALK\_Col1 ALK\_Col2 APC ATM\_Col1 ATM\_Col2 BRAF   
## 2 2 2 19 4 1 1   
## CCND2 CDKN2A CTNNB1 DNMT3A EGFR ERBB3 ERBB4   
## 2 45 4 3 6 1 2   
## ESR1 FBXW7 FGFR1 FGFR3 FLT4 FOXL2 GNAS   
## 1 8 4 2 2 2 7   
## HNF1A KRAS\_Col1 KRAS\_Col2 MAP2K2 MET MLH\_Col2 MSH2   
## 1 55 1 1 9 1 30   
## MSH6 NF\_Col1 NF\_Col2 NF\_Col3 NF\_Col5 NOTCH1 NTRK1   
## 7 5 10 7 2 4 7   
## PDGFRB PIK3CA PIK3CB POLD\_Col2 PTCH1 PTEN RB1   
## 5 7 11 5 4 7 4   
## SMARCA4 SMARCB1 SMO STK11 TERT TP53\_Col1 TP53\_Col2   
## 1 9 8 23 10 117 8   
## TSC2   
## 31

### Clean the Dataset

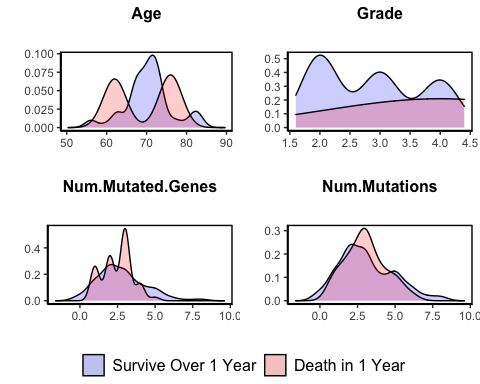
clinical[,c("Grade")][clinical[,c("Grade")] == 9] <- NA  
clinical[,c("T")][clinical[,c("T")] == "UNK"] <- NA  
clinical[,c("N")][clinical[,c("N")] == 'NULL'] <- NA  
clinical[,c("M")][clinical[,c("M")] == 'NULL'] <- NA  
clinical[,c("Radiation")][clinical[,c("Radiation")] == 5] <- 1  
clinical$Radiation<-factor(as.factor(clinical$Radiation), levels=c(0,1))  
clinical[,c("Tumor.Size")][clinical[,c("Tumor.Size")] == 'NULL'] <- NA  
  
clinical$deathin1yr<-NA  
for (i in 1:nrow(clinical)) {  
 if (clinical$Outcome[i]=="Dead" & clinical$Survival.Months[i]<=12) {  
 clinical$deathin1yr[i]=1  
 }  
 else {  
 clinical$deathin1yr[i]=0  
 }  
   
}  
clinical$deathin1yr <- factor(clinical$deathin1yr, levels=c(0,1), labels=c("No", "Yes"))

### Descriptive Analyses: Explore the Variables

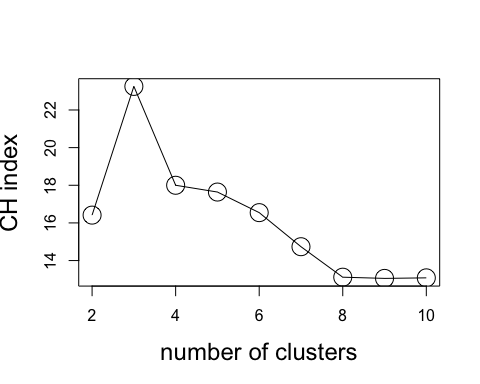
var <- c("Age", "Grade", "Num.Primaries", "T", "N", "M",  
 "Radiation", "Stage", "Primary.Site", "Histology",  
 "Tumor.Size", "Num.Mutated.Genes", "Num.Mutations")  
  
t <- CreateTableOne(vars = var, strata = "deathin1yr", data = clinical)  
print(t, nonnormal=c("Grade", "Num.Primaries", "N", "M",  
 "Tumor.Size", "Num.Mutated.Genes", "Num.Mutations"),  
 exact=c( "Num.Primaries", "T", "Radiation", "Stage", "Primary.Site", "Histology"), quote=T)

## "Stratified by deathin1yr"  
## "" "No"   
## "n" " 140"   
## "Age (mean (SD))" "70.49 (5.78)"   
## "Grade (median [IQR])" " 3.00 [2.00, 4.00]"  
## "Num.Primaries (median [IQR])" " 0.00 [0.00, 1.00]"  
## "T (%)" " "   
## " 1" " 1 ( 1.0) "   
## " 1a" " 26 (25.7) "   
## " 1b" " 2 ( 2.0) "   
## " 2" " 12 (11.9) "   
## " 2a" " 16 (15.8) "   
## " 2b" " 10 ( 9.9) "   
## " 3" " 11 (10.9) "   
## " 4" " 23 (22.8) "   
## " UNK" " 0 ( 0.0) "   
## "N (%)" " "   
## " 0" " 52 (52.5) "   
## " 1" " 9 ( 9.1) "   
## " 2" " 32 (32.3) "   
## " 3" " 6 ( 6.1) "   
## " NULL" " 0 ( 0.0) "   
## "M (%)" " "   
## " 0" " 86 (91.5) "   
## " 1" " 8 ( 8.5) "   
## " NULL" " 0 ( 0.0) "   
## "Radiation = 1 (%)" " 63 (45.0) "   
## "Stage (%)" " "   
## " 1B" " 1 ( 0.7) "   
## " IA" " 32 (22.9) "   
## " IB" " 1 ( 0.7) "   
## " IIA" " 8 ( 5.7) "   
## " IIB" " 11 ( 7.9) "   
## " IIIA" " 18 (12.9) "   
## " IIIB" " 24 (17.1) "   
## " IV" " 45 (32.1) "   
## " IVB" " 0 ( 0.0) "   
## "Primary.Site (%)" " "   
## " Both Lung" " 5 ( 3.6) "   
## " Left Hilar" " 0 ( 0.0) "   
## " Left Lower Lobe" " 17 (12.1) "   
## " Left Upper Lobe" " 21 (15.0) "   
## " Righ Upper Lobe" " 2 ( 1.4) "   
## " Right Hilar" " 14 (10.0) "   
## " Right Lower Lobe" " 25 (17.9) "   
## " Right Middle Lobe" " 3 ( 2.1) "   
## " Right Upper Lobe" " 53 (37.9) "   
## "Histology (%)" " "   
## " Adenocarcinoma" " 62 (44.3) "   
## " Large-cell carcinoma" " 1 ( 0.7) "   
## " Squamous cell carcinoma" " 77 (55.0) "   
## "Tumor.Size (%)" " "   
## " 1" " 1 ( 1.0) "   
## " 1.4" " 2 ( 2.0) "   
## " 1.5" " 13 (13.3) "   
## " 1.6" " 1 ( 1.0) "   
## " 1.8" " 2 ( 2.0) "   
## " 1.9" " 2 ( 2.0) "   
## " 10" " 8 ( 8.2) "   
## " 2" " 20 (20.4) "   
## " 2.5" " 2 ( 2.0) "   
## " 3.5" " 2 ( 2.0) "   
## " 3.6" " 9 ( 9.2) "   
## " 4" " 7 ( 7.1) "   
## " 4.4" " 1 ( 1.0) "   
## " 5.3" " 2 ( 2.0) "   
## " 5.4" " 2 ( 2.0) "   
## " 5.5" " 6 ( 6.1) "   
## " 8" " 2 ( 2.0) "   
## " 8.5" " 6 ( 6.1) "   
## " 9" " 10 (10.2) "   
## " NULL" " 0 ( 0.0) "   
## "Num.Mutated.Genes (median [IQR])" " 3.00 [2.00, 4.00]"  
## "Num.Mutations (median [IQR])" " 3.00 [2.00, 4.00]"  
## "Stratified by deathin1yr"  
## "" "Yes" "p"   
## "n" " 50" ""   
## "Age (mean (SD))" "69.28 (7.07)" " 0.232"  
## "Grade (median [IQR])" " 4.00 [4.00, 4.00]" "<0.001"  
## "Num.Primaries (median [IQR])" " 0.00 [0.00, 0.00]" "<0.001"  
## "T (%)" " " " NA"   
## " 1" " 0 ( 0.0) " ""   
## " 1a" " 0 ( 0.0) " ""   
## " 1b" " 0 ( 0.0) " ""   
## " 2" " 0 ( 0.0) " ""   
## " 2a" " 0 ( 0.0) " ""   
## " 2b" " 0 ( 0.0) " ""   
## " 3" " 27 (100.0) " ""   
## " 4" " 0 ( 0.0) " ""   
## " UNK" " 0 ( 0.0) " ""   
## "N (%)" " " " NaN"   
## " 0" " 0 ( 0.0) " ""   
## " 1" " 0 ( 0.0) " ""   
## " 2" " 26 (100.0) " ""   
## " 3" " 0 ( 0.0) " ""   
## " NULL" " 0 ( 0.0) " ""   
## "M (%)" " " " NaN"   
## " 0" " 0 ( NaN) " ""   
## " 1" " 0 ( NaN) " ""   
## " NULL" " 0 ( NaN) " ""   
## "Radiation = 1 (%)" " 0 ( 0.0) " "<0.001"  
## "Stage (%)" " " " NA"   
## " 1B" " 0 ( 0.0) " ""   
## " IA" " 0 ( 0.0) " ""   
## " IB" " 0 ( 0.0) " ""   
## " IIA" " 0 ( 0.0) " ""   
## " IIB" " 0 ( 0.0) " ""   
## " IIIA" " 25 ( 50.0) " ""   
## " IIIB" " 0 ( 0.0) " ""   
## " IV" " 0 ( 0.0) " ""   
## " IVB" " 25 ( 50.0) " ""   
## "Primary.Site (%)" " " " NA"   
## " Both Lung" " 0 ( 0.0) " ""   
## " Left Hilar" " 31 ( 62.0) " ""   
## " Left Lower Lobe" " 0 ( 0.0) " ""   
## " Left Upper Lobe" " 0 ( 0.0) " ""   
## " Righ Upper Lobe" " 0 ( 0.0) " ""   
## " Right Hilar" " 19 ( 38.0) " ""   
## " Right Lower Lobe" " 0 ( 0.0) " ""   
## " Right Middle Lobe" " 0 ( 0.0) " ""   
## " Right Upper Lobe" " 0 ( 0.0) " ""   
## "Histology (%)" " " "<0.001"  
## " Adenocarcinoma" " 24 ( 48.0) " ""   
## " Large-cell carcinoma" " 26 ( 52.0) " ""   
## " Squamous cell carcinoma" " 0 ( 0.0) " ""   
## "Tumor.Size (%)" " " " NaN"   
## " 1" " 0 ( NaN) " ""   
## " 1.4" " 0 ( NaN) " ""   
## " 1.5" " 0 ( NaN) " ""   
## " 1.6" " 0 ( NaN) " ""   
## " 1.8" " 0 ( NaN) " ""   
## " 1.9" " 0 ( NaN) " ""   
## " 10" " 0 ( NaN) " ""   
## " 2" " 0 ( NaN) " ""   
## " 2.5" " 0 ( NaN) " ""   
## " 3.5" " 0 ( NaN) " ""   
## " 3.6" " 0 ( NaN) " ""   
## " 4" " 0 ( NaN) " ""   
## " 4.4" " 0 ( NaN) " ""   
## " 5.3" " 0 ( NaN) " ""   
## " 5.4" " 0 ( NaN) " ""   
## " 5.5" " 0 ( NaN) " ""   
## " 8" " 0 ( NaN) " ""   
## " 8.5" " 0 ( NaN) " ""   
## " 9" " 0 ( NaN) " ""   
## " NULL" " 0 ( NaN) " ""   
## "Num.Mutated.Genes (median [IQR])" " 3.00 [2.00, 3.00]" " 0.536"  
## "Num.Mutations (median [IQR])" " 3.00 [2.00, 4.00]" " 0.887"  
## "Stratified by deathin1yr"  
## "" "test"   
## "n" ""   
## "Age (mean (SD))" ""   
## "Grade (median [IQR])" "nonnorm"  
## "Num.Primaries (median [IQR])" "nonnorm"  
## "T (%)" "exact"   
## " 1" ""   
## " 1a" ""   
## " 1b" ""   
## " 2" ""   
## " 2a" ""   
## " 2b" ""   
## " 3" ""   
## " 4" ""   
## " UNK" ""   
## "N (%)" ""   
## " 0" ""   
## " 1" ""   
## " 2" ""   
## " 3" ""   
## " NULL" ""   
## "M (%)" ""   
## " 0" ""   
## " 1" ""   
## " NULL" ""   
## "Radiation = 1 (%)" "exact"   
## "Stage (%)" "exact"   
## " 1B" ""   
## " IA" ""   
## " IB" ""   
## " IIA" ""   
## " IIB" ""   
## " IIIA" ""   
## " IIIB" ""   
## " IV" ""   
## " IVB" ""   
## "Primary.Site (%)" "exact"   
## " Both Lung" ""   
## " Left Hilar" ""   
## " Left Lower Lobe" ""   
## " Left Upper Lobe" ""   
## " Righ Upper Lobe" ""   
## " Right Hilar" ""   
## " Right Lower Lobe" ""   
## " Right Middle Lobe" ""   
## " Right Upper Lobe" ""   
## "Histology (%)" "exact"   
## " Adenocarcinoma" ""   
## " Large-cell carcinoma" ""   
## " Squamous cell carcinoma" ""   
## "Tumor.Size (%)" ""   
## " 1" ""   
## " 1.4" ""   
## " 1.5" ""   
## " 1.6" ""   
## " 1.8" ""   
## " 1.9" ""   
## " 10" ""   
## " 2" ""   
## " 2.5" ""   
## " 3.5" ""   
## " 3.6" ""   
## " 4" ""   
## " 4.4" ""   
## " 5.3" ""   
## " 5.4" ""   
## " 5.5" ""   
## " 8" ""   
## " 8.5" ""   
## " 9" ""   
## " NULL" ""   
## "Num.Mutated.Genes (median [IQR])" "nonnorm"  
## "Num.Mutations (median [IQR])" "nonnorm"

densityplot <- function(data, indep\_var) {  
 mydata <- data[!is.na(data[,indep\_var]), c(indep\_var, "deathin1yr")]   
   
 mydata$y <- mydata[,indep\_var]  
   
 myrange.u <- max(mydata[,indep\_var]) - min(mydata[,indep\_var])  
 mybuffer.u <- 0.20 \* myrange.u  
   
 myrange.t <- max(mydata$y) - min(mydata$y)  
 mybuffer.t <- 0.20 \* myrange.t  
 myn <- nrow(mydata)  
   
 p <- ggplot(mydata, aes(get(indep\_var), fill = factor(deathin1yr), color = factor(deathin1yr))) +  
 geom\_density(alpha = 0.2) +   
 scale\_color\_manual(values = c("black","black"), guide = FALSE) +  
 scale\_fill\_manual(name = "",  
 values = c("#0000FF66","#FF000066"),   
 labels = c("Survive Over 1 Year","Death in 1 Year")) +   
 xlab("") + ylab("") +   
 ggtitle(indep\_var, sub = " ") +   
 scale\_x\_continuous(limits = c(min(mydata[,indep\_var]) - mybuffer.u, max(mydata[,indep\_var]) + mybuffer.u )) +  
 theme(panel.grid.major = element\_blank(),   
 panel.grid.minor = element\_blank(),  
 panel.background = element\_blank(),   
 axis.line = element\_line(colour = "black"),  
 panel.border = element\_rect(colour = "black", fill=NA, size=1),  
 plot.title = element\_text(face = "bold", size = 12, hjust = 0.5),  
 plot.subtitle = element\_text(hjust = 0.5, size = 12),  
 legend.text = element\_text(size=12),  
 legend.position = "bottom",  
 legend.direction = "horizontal")  
 return(p)  
}  
  
p1 <- densityplot(clinical, colnames(clinical)[4])  
p2 <- densityplot(clinical, colnames(clinical)[5])  
p3 <- densityplot(clinical, colnames(clinical)[15])  
p4 <- densityplot(clinical, colnames(clinical)[16])  
  
  
grid\_arrange\_shared\_legend <- function(...,   
 ncol = length(list(...)),  
 nrow = 1,  
 position = c("bottom", "right")){  
 plots <- list(...)  
 position <- match.arg(position)  
 g <-  
 ggplotGrob(plots[[1]] + theme(legend.position = position))$grobs  
 legend <- g[[which(sapply(g, function(x)  
 x$name) == "guide-box")]]  
 lheight <- sum(legend$height)  
 lwidth <- sum(legend$width)  
 gl <- lapply(plots, function(x)  
 x + theme(legend.position = "none"))  
 gl <- c(gl, ncol = ncol, nrow = nrow)  
   
 combined <- switch(  
 position,  
 "bottom" = arrangeGrob(  
 do.call(arrangeGrob, gl),  
 legend,  
 ncol = 1,  
 heights = unit.c(unit(1, "npc") - lheight, lheight)  
 ),  
 "right" = arrangeGrob(  
 do.call(arrangeGrob, gl),  
 legend,  
 ncol = 2,  
 widths = unit.c(unit(1, "npc") - lwidth, lwidth)  
 )  
 )  
   
 grid.newpage()  
 grid.draw(combined)  
   
 # return gtable invisibly  
 invisible(combined)  
}   
  
grid\_arrange\_shared\_legend(p1, p2, p3, p4, ncol=2, nrow=2)



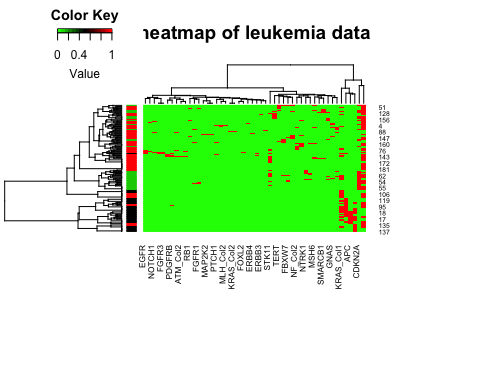
#Any clustering of gene mutation?  
genomics\_wide<-genomics%>%mutate(mutation=1)%>%pivot\_wider(names\_from = Gene, values\_from = mutation)  
genomics\_wide[,2:51][is.na(genomics\_wide[,2:51])] <- 0  
dat<-genomics\_wide%>%dplyr::select(-c("ID"))  
CH<-NbClust(data=dat,min.nc = 2, max.nc = 10,method="kmeans",index="ch")  
plot(2:10,CH$All.index,type="o",xlab="number of clusters",ylab="CH index","cex.lab"=1.5,cex=2.5)



a<-kmeans(dat,nstart=10,centers = 3)  
cluster<-a$cluster  
table(cluster)

## cluster  
## 1 2 3   
## 48 91 45

row\_distance = dist(dat, method = "euclidean")  
row\_cluster = hclust(row\_distance, method = "ward.D")  
col\_distance = dist(t(dat), method = "euclidean")  
col\_cluster = hclust(col\_distance, method = "ward.D")  
#par(cex.main=0.8)  
#colnames(per\_MiRNA)<-cancerType1  
heatmap.2(as.matrix(dat),  
 #RowSideColors = c( # grouping row-variables into different  
 # rep("red", num\_sample1), # categories, Measurement 1-3: green # Measurement 4-6: blue  
 # rep("blue", num\_sample1)), # Measurement 7-10: red  
 main = paste("heatmap of leukemia data",sep=""),  
 #xlab=cancerType1,# heat map title  
 notecol="black", # change font color of cell labels to black  
 density.info="none", # turns off density plot inside color legend  
 trace="none",  
 # turns off trace lines inside the heat map  
 margins =c(12,9),  
 RowSideColors = as.character(cluster),  
 col=greenred, # use on color palette defined earlier  
 #dendrogram="both", # only draw a row dendrogram  
 Rowv=as.dendrogram(row\_cluster),  
 Colv =as.dendrogram(col\_cluster))



cluster\_result<-cbind(genomics\_wide%>%dplyr::select("ID"), cluster)  
cluster\_result$cluster<-as.factor(cluster\_result$cluster)  
  
final\_data<-merge(clinical, cluster\_result, by.x="ID", by.y="ID")%>%dplyr::select(-c("ID", "Outcome", "Survival.Months",   
 "Grade", "T", "N", "M", "Tumor.Size", #missing values  
 "Num.Primaries", "Radiation")) ##only one category in outcome=1  
final\_data<-final\_data[complete.cases(final\_data),]  
  
table(final\_data$deathin1yr, final\_data$cluster)

##   
## 1 2 3  
## No 5 84 45  
## Yes 43 7 0

### 1. Please walk us through how you cleaned up this dataset. Do you see any interesting trends?

* Replaced all the values indicating a missing with NA
* Created a binary variable of whether or not dead within 1 year based on the survival outcome and survival month variable.
* Clustered the patients into three classes based on the mutations genes using k means clustering. Merged the class info with main dataset.
* Deleted multiple variables with large proportion of missingness before modeling.
* Deleted multiple variables with only one category among non survivals.
* Include rows with complete data.
* According to table 1 and density plot, survival and non survivals differ the most significantly in age, grade, stage, primary site, histology, and the cluster based on types of gene mutation. Patients who are around 62 or 76, have a higher tumor grade, have a IIIA or IVB tumor stage at diagnosis, have a primary tumor location at Left Hilar or Right Hilar, have a Adenocarcinoma or Large-cell carcinoma tumor, and who are in class 1, are more likely to die within one year of diagnosis.

### 2. Tell us how you decided which features to use for your model. Are there any new features it might be productive to engineer from the current features?

* I decide to use age, stage, primary site, histology, number of mutated genes, number of mutations, and class based on type of gene mutations.
* The reasons for choosing the features are:

1. They have none or limited missing data.
2. There are more than 1 values for each outcome group, so they are eligible to be included in the modeling.
3. They are related to the survival outcome from the descriptive analyses.

* I used k means clustering analysis to define the class of patients based on where the genes were found to have mutation. It looks like that not the number of mutation but the site of mutation that has an impact on the survival outcome.

### Prediction Modeling

plot.new()  
set.seed(1)  
n<-nrow(final\_data)  
index.train<-sample(1:n,size = round(0.6\*n))  
index.test<-(1:n)[-index.train]  
data\_train<-final\_data[index.train,]  
data\_test<-final\_data[index.test,]  
#Linear Discriminant Analysis  
lda.fit=lda(deathin1yr~.,data=data\_train)  
lda.fit

## Call:  
## lda(deathin1yr ~ ., data = data\_train)  
##   
## Prior probabilities of groups:  
## No Yes   
## 0.7363636 0.2636364   
##   
## Group means:  
## Age StageIA StageIB StageIIA StageIIB StageIIIA  
## No 71.16049 0.2345679 0.01234568 0.03703704 0.08641975 0.1234568  
## Yes 68.27586 0.0000000 0.00000000 0.00000000 0.00000000 0.4482759  
## StageIIIB StageIV StageIVB Primary.SiteLeft Hilar  
## No 0.1975309 0.2962963 0.0000000 0.0000000  
## Yes 0.0000000 0.0000000 0.5517241 0.5172414  
## Primary.SiteLeft Lower Lobe Primary.SiteLeft Upper Lobe  
## No 0.1481481 0.1728395  
## Yes 0.0000000 0.0000000  
## Primary.SiteRigh Upper Lobe Primary.SiteRight Hilar  
## No 0.02469136 0.1234568  
## Yes 0.00000000 0.4827586  
## Primary.SiteRight Lower Lobe Primary.SiteRight Middle Lobe  
## No 0.1234568 0.01234568  
## Yes 0.0000000 0.00000000  
## Primary.SiteRight Upper Lobe HistologyLarge-cell carcinoma  
## No 0.345679 0.0000000  
## Yes 0.000000 0.5172414  
## HistologySquamous cell carcinoma Num.Mutated.Genes Num.Mutations  
## No 0.5308642 3.024691 3.395062  
## Yes 0.0000000 2.275862 2.896552  
## cluster2 cluster3  
## No 0.6419753 0.3209877  
## Yes 0.2068966 0.0000000  
##   
## Coefficients of linear discriminants:  
## LD1  
## Age 0.02483981  
## StageIA -0.25019250  
## StageIB -0.76808351  
## StageIIA -0.06548663  
## StageIIB -0.92395319  
## StageIIIA 1.07988817  
## StageIIIB -0.03494186  
## StageIV -0.42944532  
## StageIVB 3.60154900  
## Primary.SiteLeft Hilar 3.57522071  
## Primary.SiteLeft Lower Lobe -0.62415422  
## Primary.SiteLeft Upper Lobe -0.30212066  
## Primary.SiteRigh Upper Lobe 0.85221248  
## Primary.SiteRight Hilar 1.16900763  
## Primary.SiteRight Lower Lobe -0.16624834  
## Primary.SiteRight Middle Lobe -0.40258249  
## Primary.SiteRight Upper Lobe -0.48697145  
## HistologyLarge-cell carcinoma 1.74973532  
## HistologySquamous cell carcinoma -0.27152153  
## Num.Mutated.Genes 0.03395892  
## Num.Mutations -0.12224871  
## cluster2 -2.74882608  
## cluster3 -2.63156859

#plot(lda.fit)  
lda.pred=predict(lda.fit, data\_test)  
table(data\_test$deathin1yr, lda.pred$class)

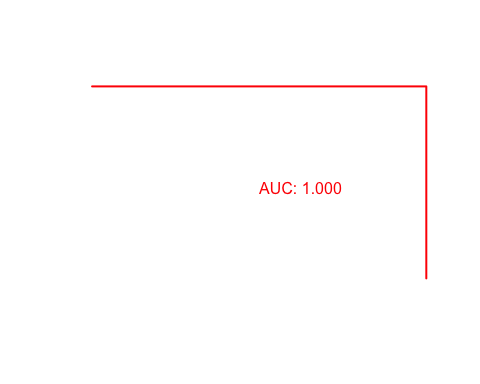
##   
## No Yes  
## No 53 0  
## Yes 0 21

roc(as.factor(data\_test$deathin1yr), lda.pred$x, plot=TRUE,print.auc=TRUE,col=2,legacy.axe=TRUE,add=TRUE,print.auc.y=0.5)

## Setting levels: control = No, case = Yes

## Warning in roc.default(as.factor(data\_test$deathin1yr), lda.pred$x, plot  
## = TRUE, : Deprecated use a matrix as predictor. Unexpected results may be  
## produced, please pass a numeric vector.

## Setting direction: controls < cases



##   
## Call:  
## roc.default(response = as.factor(data\_test$deathin1yr), predictor = lda.pred$x, plot = TRUE, print.auc = TRUE, col = 2, legacy.axe = TRUE, add = TRUE, print.auc.y = 0.5)  
##   
## Data: lda.pred$x in 53 controls (as.factor(data\_test$deathin1yr) No) < 21 cases (as.factor(data\_test$deathin1yr) Yes).  
## Area under the curve: 1

### 3. Which algorithm did you use for the prediction and why?

* Linear discriminant analysis. Reason is that we have a relatively large number of sample size compared with number of features. LDA has a good performance under this condition. The results of LDA are easier to interpret and more transparent compared to other machine learning methods designed for large features. LDA also has equal or better prediction performance compared with logistic regression.

### 4. How did you assess the predictive model’s quality? Summarize your findings.

* I used cross-validation to separate the original datasets into two parts, one for training the prediction and one for testing. Tihs is to avoid overfitting. I assessed the error rate, which is 100%. I also generated the ROC curve, the AUC is 1. The testing results both suggest perfect performance of the prediction model.

### 5. Next steps? What might you do with more time or access to additional data or expertise?

* Since the performance of the model is already perfect, I would gather the existing features on greater number of patients, and also in a setting with patients more diversed in demographics (gender, race, social economic status, etc). I will replicate the prediction model to see if the model could be generalized to a larger population before deploying it in clinical setting.