Leukemia Analysis

Emily Kophs

12/5/2021

```
require(data.table)
```

```
## Loading required package: data.table
```

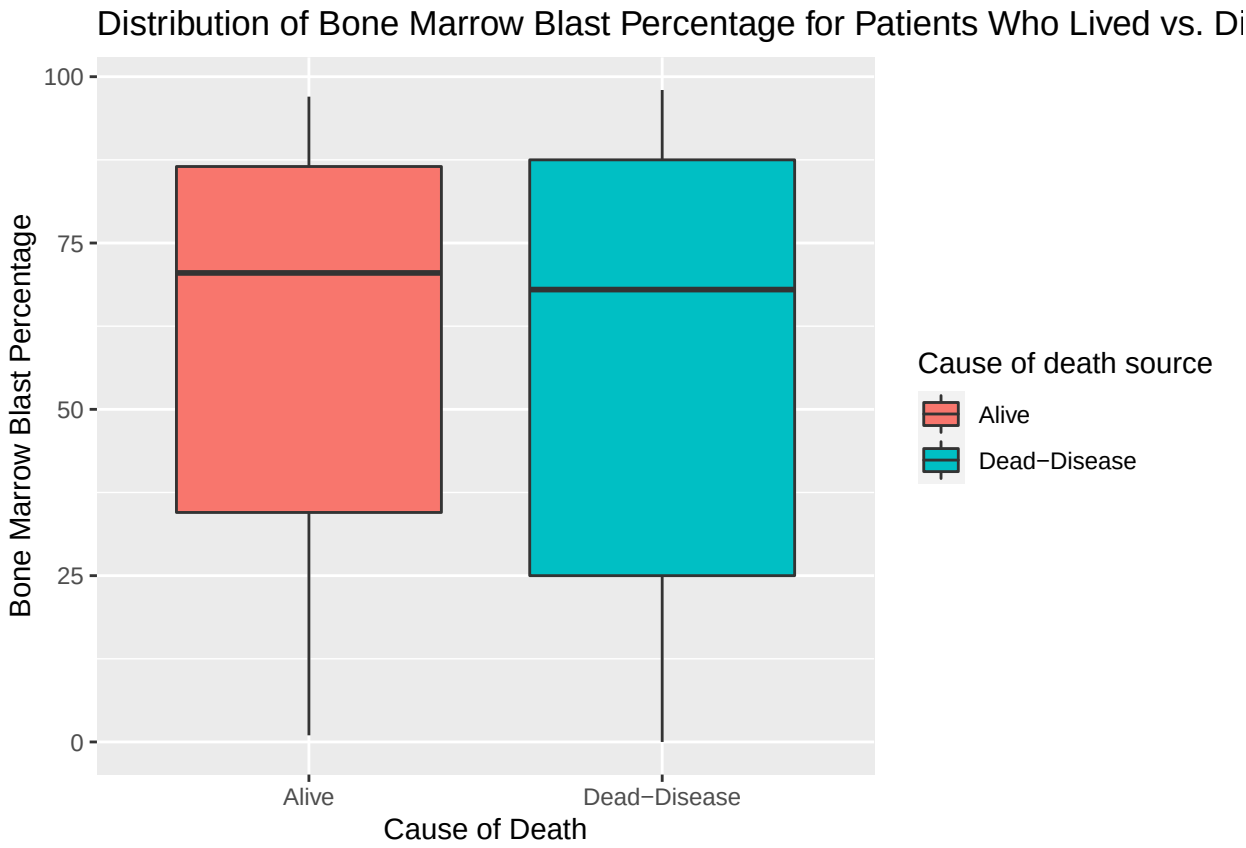
```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

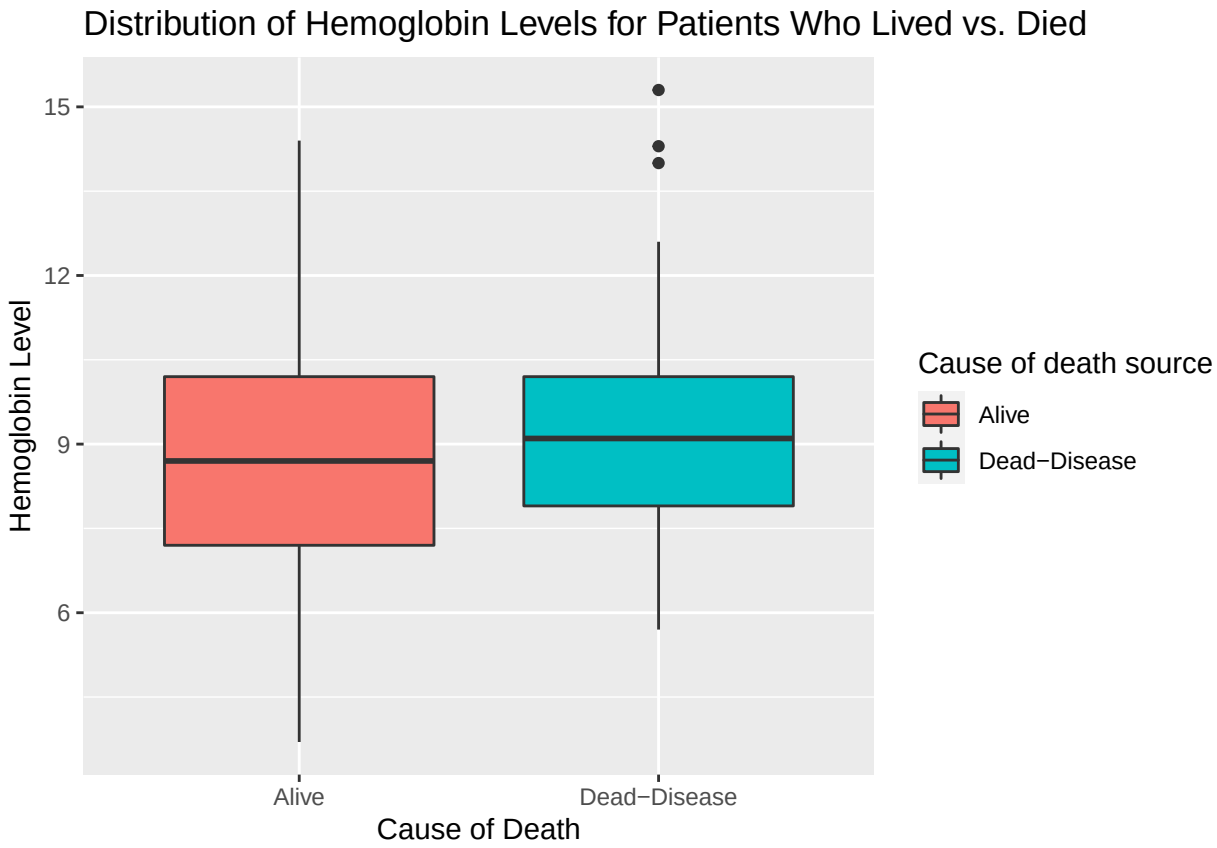
```
data<-as.data.frame(fread("~/Downloads/aml_ohsu_2018_clinical_data.tsv"))

leukemia_data<-data[c(4,6,8,10,14,21,41)]
leukemia_data<-leukemia_data[leukemia_data$`Cancer Type` %in% "Leukemia",]
leukemia_data<-na.omit(leukemia_data)
leukemia_data$`Cause of death source`[leukemia_data$`Cause of death source`=="Dead-Other"]<-"Alive"
leukemia_data<-leukemia_data[!(leukemia_data$`Cause of death source` %in% "Dead-Unknown"),]
leukemia_data<-leukemia_data[!(leukemia_data$`Cause of death source` %in% "Dead-Treatment"),]
leukemia_data$`Cause of death source`<-as.factor(leukemia_data$`Cause of death source`)
leukemia_data$Chemotherapy<-as.factor(leukemia_data$Chemotherapy)


ggplot(leukemia_data,aes(x=`Cause of death source`, y=`Bone Marrow Blast Percentage`))+geom_boxplot(aes
```




```
ggplot(leukemia_data,aes(x=`Cause of death source`, y=`Hemoglobin level`))+geom_boxplot(aes(fill =`Caus
```




```
logREG<-glm(formula=as.factor(`Cause of death source`)~`Bone Marrow Blast Percentage` + `Hemoglobin leve
summary(logREG)
```

```
##
## Call:
## glm(formula = as.factor(`Cause of death source`) ~ `Bone Marrow Blast Percentage` +
##     `Hemoglobin level`, family = binomial, data = leukemia_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4161  -1.0938  -0.9535   1.2405   1.4207
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.124330   0.708106  -1.588   0.1123
## `Bone Marrow Blast Percentage` -0.001624   0.004298  -0.378   0.7055
## `Hemoglobin level`             0.116083   0.068572   1.693   0.0905 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 323.90  on 234  degrees of freedom
## Residual deviance: 320.58  on 232  degrees of freedom
## AIC: 326.58
##
## Number of Fisher Scoring iterations: 4
```

```
leukemia_data$fitted_values<-logREG$fitted.values
leukemia_data$prediction<-ifelse(leukemia_data$fitted_values <.5,"Dead-Disease","Alive")
mean(leukemia_data$prediction == leukemia_data$`Cause of death source`)
```

```
## [1] 0.4808511
```

```
ggplot(leukemia_data, aes(x = `Cause of death source`, y = fitted_values)) + geom_boxplot(aes(fill = `Ca
```