```
> data<-read.csv("~/Downloads/base_metrics.csv")
> drop=c("Name")
> data = data[,!names(data) %in% drop]
> names(data)
 [1] "Code_Ownership_count"      "lines_added"               "lines_deleted"
 [4] "total_prerel_change"       "post_release_defects"      "pre_release_defects"
 [7] "AvgCyclomatic"             "AvgCyclomaticModified"     "AvgCyclomaticStrict"
[10] "AvgEssential"              "AvgLine"                   "AvgLineBlank"
[13] "AvgLineCode"               "AvgLineComment"            "CountDeclClass"
[16] "CountDeclClassMethod"      "CountDeclClassVariable"    "CountDeclExecutableUnit"
[19] "CountDeclFunction"         "CountDeclInstanceMethod"   "CountDeclInstanceVariable"
[22] "CountDeclMethod"           "CountDeclMethodDefault"    "CountDeclMethodPrivate"
[25] "CountDeclMethodProtected"  "CountDeclMethodPublic"     "CountLine"
[28] "CountLineBlank"            "CountLineCode"             "CountLineCodeDecl"
[31] "CountLineCodeExe"          "CountLineComment"          "CountStmt"
[34] "CountStmtDecl"             "CountStmtExe"              "MaxCyclomatic"
[37] "MaxCyclomaticModified"     "MaxCyclomaticStrict"       "MaxEssential"
[40] "MaxNesting"                "RatioCommentToCode"        "SumCyclomatic"
[43] "SumCyclomaticModified"     "SumCyclomaticStrict"       "SumEssential"
> drop=c("post_release_defects")
> independant=data[,!(names(data) %in% drop)]
> correlations <- cor(independant, method="spearman")
> highCorr <- findCorrelation(correlations, cutoff = .75)
> highCorr
 [1] 34 28 32 30 43 41 42 33 26 44 35 36 37 29 27 17 39 21 18 38 10 12  8  6  7 19  2
> low_cor_names=names(independant[, -highCorr])
> low_cor_data= independant[(names(independant) %in% low_cor_names)]
> dataforredun=low_cor_data
> redun_obj = redun (~. ,data = dataforredun ,nk =0)
> after_redun= dataforredun[,!(names(dataforredun) %in% redun_obj $Out)]
> form=as.formula(paste("post_release_defects>0~",paste(names(after_redun),collapse="+")))
> model=glm(formula=form, data=log10(data+1), family = binomial(link = "logit"))
> summary(model)

Call:
glm(formula = form, family = binomial(link = "logit"), data = log10(data +
    1))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0881  -0.1794  -0.1066  -0.0482   3.2526

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -5.377529   0.864767  -6.218 5.02e-10 ***
Code_Ownership_count       1.362463   0.352892   3.861 0.000113 ***
lines_deleted             -0.280057   0.203393  -1.377 0.168536
total_prerel_change        1.136215   0.127372   8.920  < 2e-16 ***
pre_release_defects        0.981500   1.085831   0.904 0.366040
AvgEssential              -1.691635   1.536865  -1.101 0.271025
AvgLineBlank               0.442462   0.689653   0.642 0.521151
AvgLineComment             0.114196   0.548014   0.208 0.834932
CountDeclClass            -0.129481   0.481706  -0.269 0.788086
CountDeclClassMethod      -0.612176   0.281185  -2.177 0.029471 *
CountDeclClassVariable     0.439551   0.278342   1.579 0.114296
CountDeclInstanceVariable  0.253984   0.284661   0.892 0.372267
CountDeclMethodDefault    -0.496733   0.274221  -1.811 0.070074 .
CountDeclMethodPrivate    -0.009205   0.298948  -0.031 0.975436
CountDeclMethodProtected  -0.074352   0.289478  -0.257 0.797295
CountDeclMethodPublic     -0.159106   0.311542  -0.511 0.609559
CountLineComment           1.240191   0.644537   1.924 0.054335 .
RatioCommentToCode        -7.755152   1.844414  -4.205 2.61e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 1295.94  on 5207  degrees of freedom
Residual deviance:  875.61  on 5190  degrees of freedom
AIC: 911.61

Number of Fisher Scoring iterations: 9

> newform= post_release_defects>0~Code_Ownership_count+ total_prerel_change + CountDeclClassMethod +
RatioCommentToCode
> newmodel=glm(formula=newform, data=log10(data+1), family = binomial(link = "logit"))
> summary(newmodel)

Call:
glm(formula = newform, family = binomial(link = "logit"), data = log10(data +
    1))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5398  -0.1798  -0.1157  -0.0658   3.3461

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -4.6864     0.3239 -14.469  < 2e-16 ***
Code_Ownership_count   2.0051     0.2786   7.198 6.11e-13 ***
total_prerel_change    1.1941     0.1012  11.796  < 2e-16 ***
CountDeclClassMethod  -0.4286     0.2360  -1.816   0.0694 .
RatioCommentToCode    -5.3795     1.0665  -5.044 4.56e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1295.94  on 5207  degrees of freedom
Residual deviance:  902.29  on 5203  degrees of freedom
AIC: 912.29

Number of Fisher Scoring iterations: 8

> newform= post_release_defects>0~Code_Ownership_count+ total_prerel_change+ RatioCommentToCode
> newmodel=glm(formula=newform, data=log10(data+1), family = binomial(link = "logit"))
> summary(newmodel)

Call:
glm(formula = newform, family = binomial(link = "logit"), data = log10(data +
    1))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5988  -0.1809  -0.1174  -0.0646   3.3557

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -4.5940     0.3191 -14.395  < 2e-16 ***
Code_Ownership_count   1.7938     0.2527   7.097 1.27e-12 ***
total_prerel_change    1.1719     0.1000  11.716  < 2e-16 ***
RatioCommentToCode    -5.5967     1.0842  -5.162 2.44e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1295.94  on 5207  degrees of freedom
Residual deviance:  905.72  on 5204  degrees of freedom
AIC: 913.72

Number of Fisher Scoring iterations: 8
```

```
> anova(newmodel)
Analysis of Deviance Table

Model: binomial, link: logit

Response: post_release_defects > 0

Terms added sequentially (first to last)


                    Df Deviance Resid. Df Resid. Dev
NULL                                 5207    1295.94
Code_Ownership_count  1  210.110      5206    1085.83
total_prerel_change   1  140.054      5205     945.78
RatioCommentToCode    1   40.056      5204     905.72
>
>
> testdata=data.frame(Code_Ownership_count =log10(mean(data$Code_Ownership_count)+1),
+ total_prerel_change=log10(mean(data$total_prerel_change)+1),
+ RatioCommentToCode =log10(mean(data$RatioCommentToCode)+1))
> predict(newmodel,testdata, type="response")
         1
0.01174439
> testdata=data.frame(Code_Ownership_count =log10(mean(data$Code_Ownership_count)*2+1),
+ total_prerel_change=log10(mean(data$total_prerel_change)+1),
+ RatioCommentToCode =log10(mean(data$RatioCommentToCode)+1))
> predict(newmodel,testdata, type="response")
         1
0.01832323
> testdata=data.frame(Code_Ownership_count =log10(mean(data$Code_Ownership_count)+1),
+ total_prerel_change=log10(mean(data$total_prerel_change)*2+1),
+ RatioCommentToCode =log10(mean(data$RatioCommentToCode)+1))
> predict(newmodel,testdata, type="response")
         1
0.01629254
> testdata=data.frame(Code_Ownership_count =log10(mean(data$Code_Ownership_count)+1),
+ total_prerel_change=log10(mean(data$total_prerel_change)+1),
+ RatioCommentToCode =log10(mean(data$RatioCommentToCode)*2+1))
> predict(newmodel,testdata, type="response")
          1
0.003692426
>
```