STUDENTS PERFORMANCE PREDICTION

Nava Carlo
Passaro Jacopo

Agenda

- EDA
- CORRELATION
- CLUSTERING
- PCA
- LINEAR REGRESSION
- SVM
- MODEL COMPARISON

Dataset

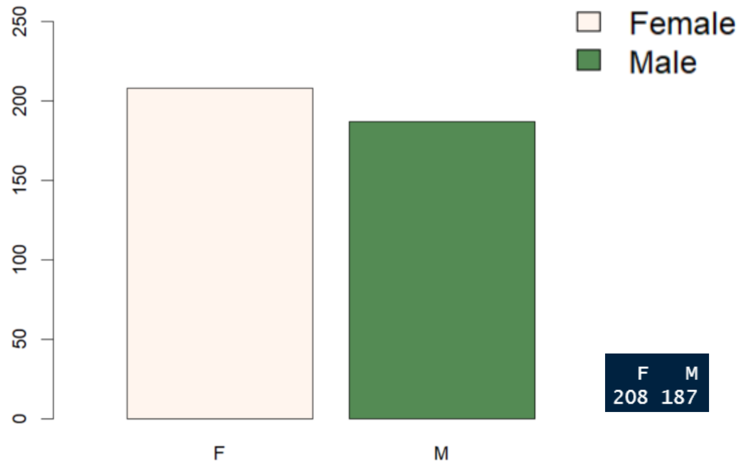
	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid
1	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no
2	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no
3	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	yes	no	yes
4	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	yes
5	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	yes
6	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	0	no	yes	yes
7	GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no	no	no
8	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	yes	yes	no
9	GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	no	yes	yes
10	GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	0	no	yes	yes
11	GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	1	2	0	no	yes	yes

activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
no	yes	yes	no	no	4	3	4	1	1	3	6	5	6	6
no	no	yes	yes	no	5	3	3	1	1	3	4	5	5	6
no	yes	yes	yes	no	4	3	2	2	3	3	10	7	8	10
yes	yes	yes	yes	yes	3	2	2	1	1	5	2	15	14	15
no	yes	yes	no	no	4	3	2	1	2	5	4	6	10	10
yes	yes	yes	yes	no	5	4	2	1	2	5	10	15	15	15
no	yes	yes	yes	no	4	4	4	1	1	3	0	12	12	11
no	yes	yes	no	no	4	1	4	1	1	1	6	6	5	6
no	yes	yes	yes	no	4	2	2	1	1	1	0	16	18	19

Variables: 33
Observations: 395
Target: G3

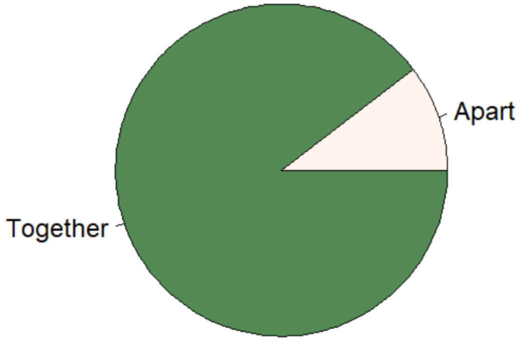
Exploratory Data Anlaysis

Gender Distribution



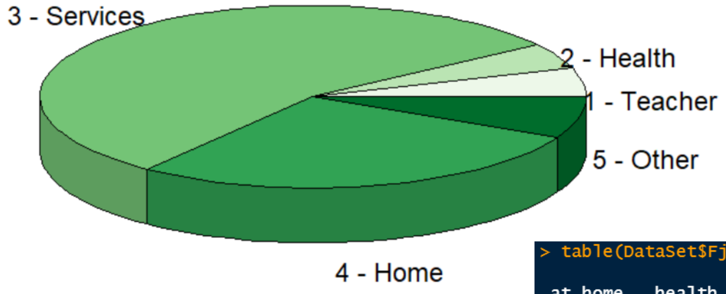
F	M
208	187

Parent Status



A	T
41	354

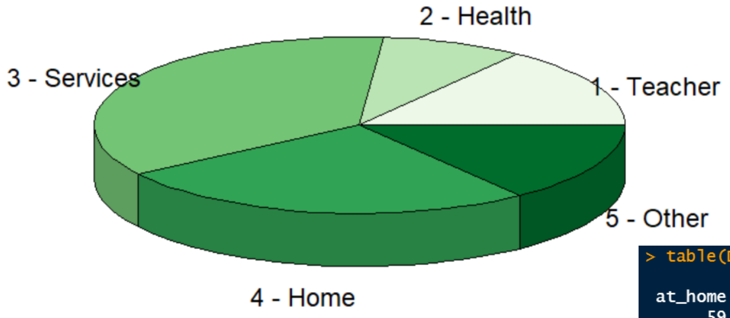
Father Occupation




```
> table(DataSet$Fjob)
```

at_home	health	other	services	teacher
20	18	217	111	29

Mother Occupation

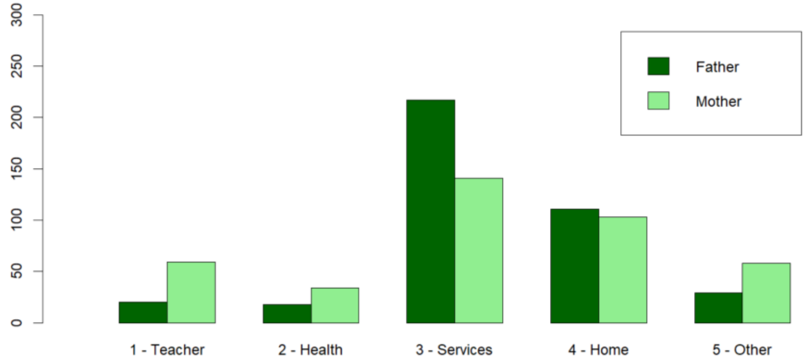



```
> table(DataSet$Mjob)
```

at_home	health	other	services	teacher
59	34	141	103	58

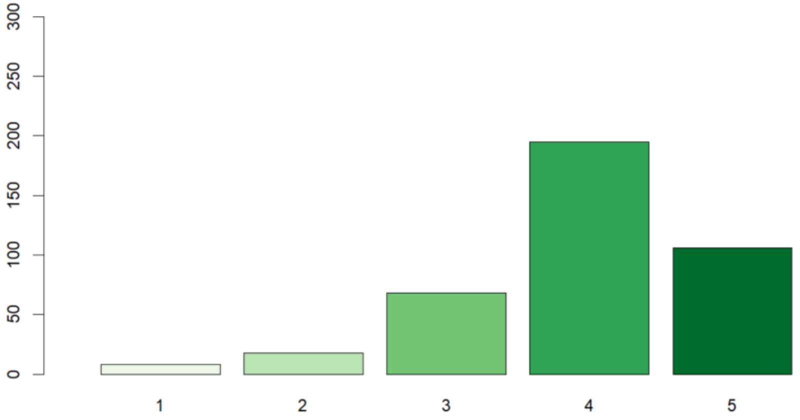
Exploratory Data Anlaysis

Father vs Mother Occupation



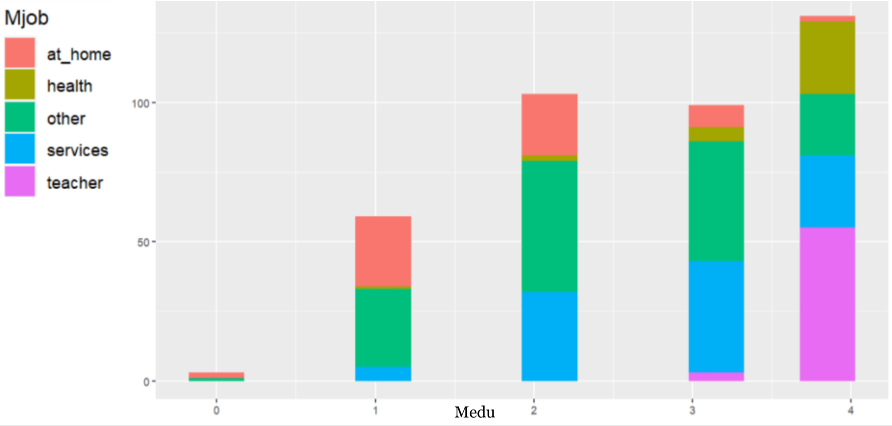
	1 - Teacher	2 - Health	3 - Services	4 - Home	5 - Other
Father	20	18	217	111	29
Mother	59	34	141	103	58

Family Relations

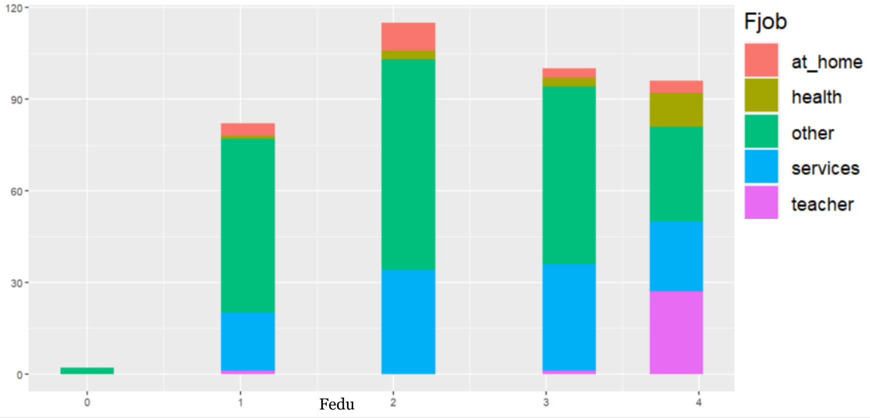



```
> table(DataSet$famrel)

  1   2   3   4   5
  8  18  68 195 106
```

Mother Job Distrib. Per Education

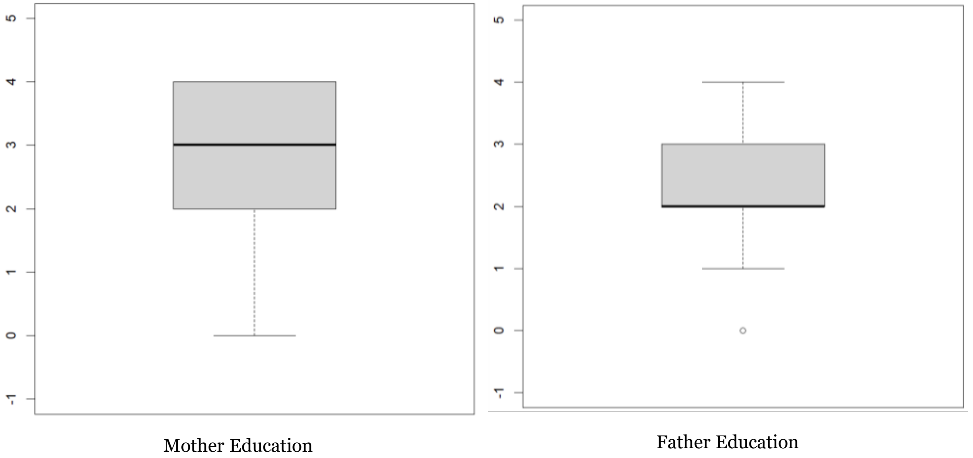


Father Job Distrib. Per Education

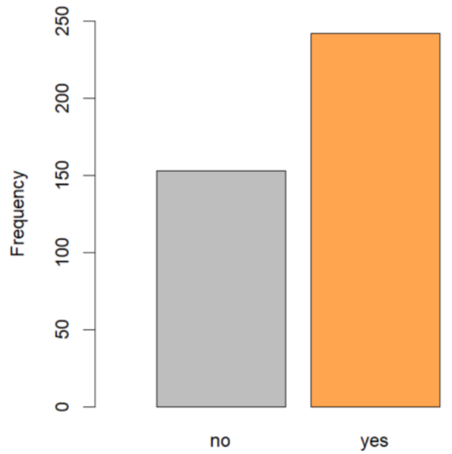


Exploratory Data Anlaysis

Father vs Mother Education



Family Support

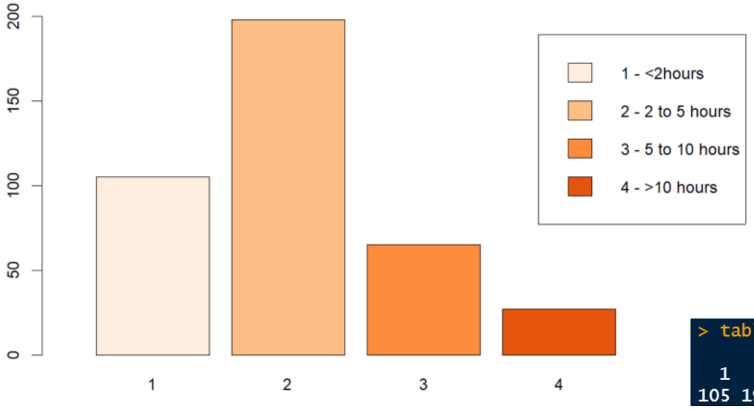



```
> table(DataSet$famsup)

 no yes
153 242
```

Failures – Family Support Matrix

```
> table(DataSet[,c(17,15)])
      failures
famsup   0   1   2   3
   no  115  25   5   8
   yes 197  25  12   8
```

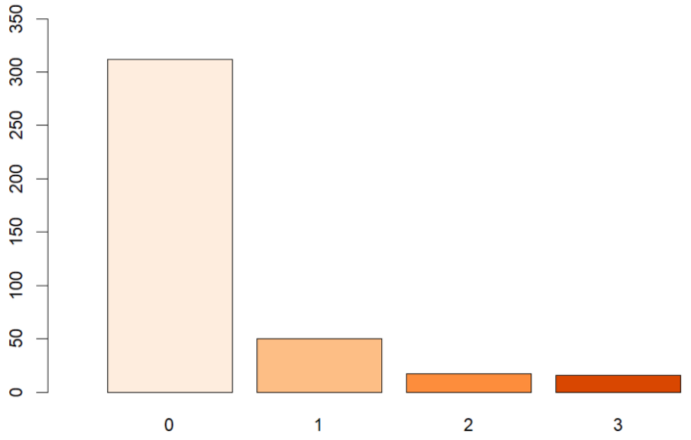
Weekly Studytime




```
> table(DataSet$studytime)

  1   2   3   4
105 198  65  27
```

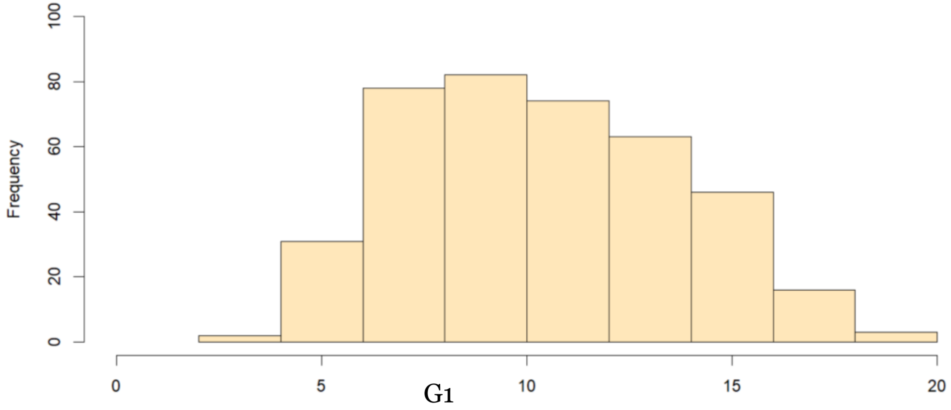
Exploratory Data Anlaysis

Count of Past Failures



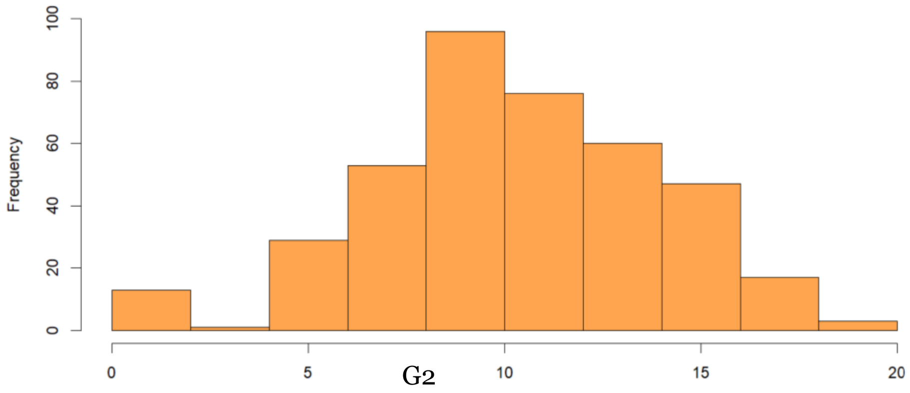

```
> table(DataSet$failures)

  0   1   2   3
312  50  17  16
```

Grade 1 Distribution



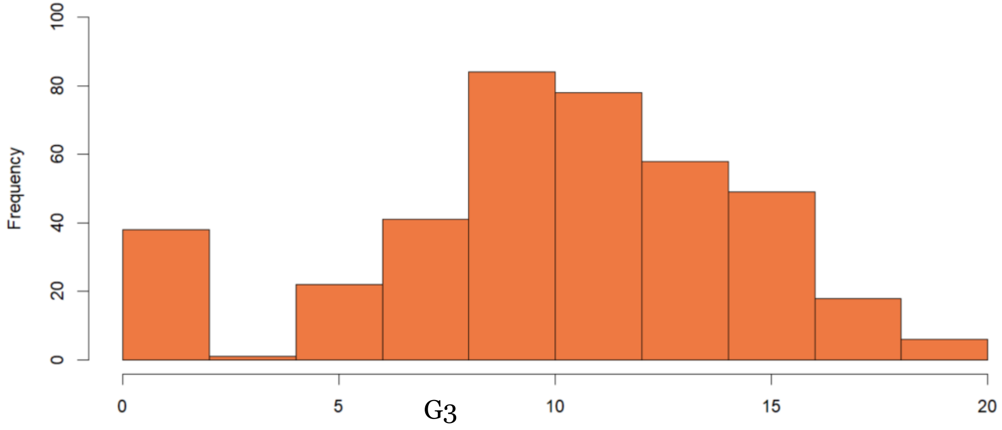

```
summary(DataSet$G1)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3.00    8.00   11.00   10.91   13.00   19.00
```

Grade 2 Distribution




```
summary(DataSet$G2)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00    9.00   11.00   10.71   13.00   19.00
```

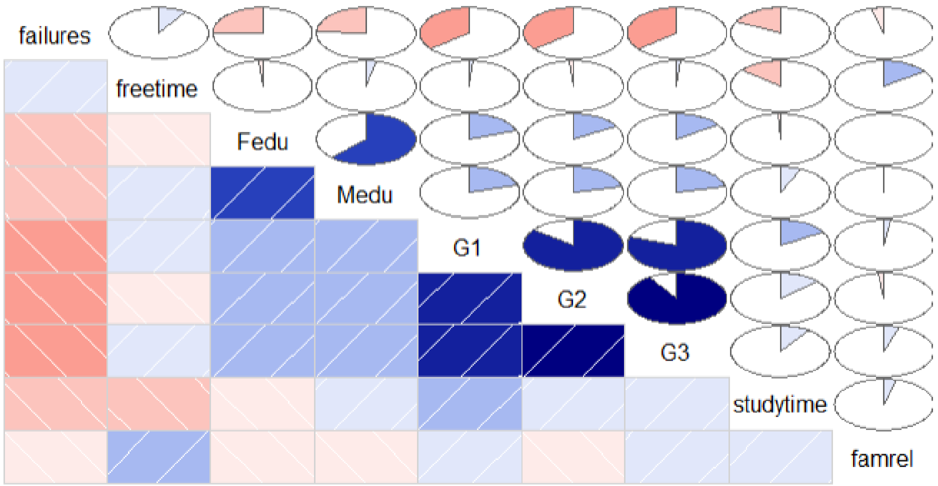
Grade 3 Distribution



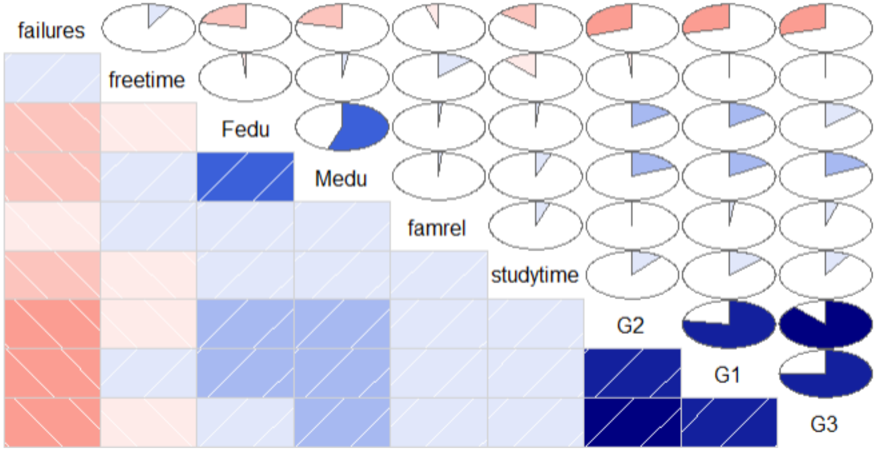

```
summary(DataSet$G3)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00    8.00   11.00   10.42   14.00   20.00
```

Correlation

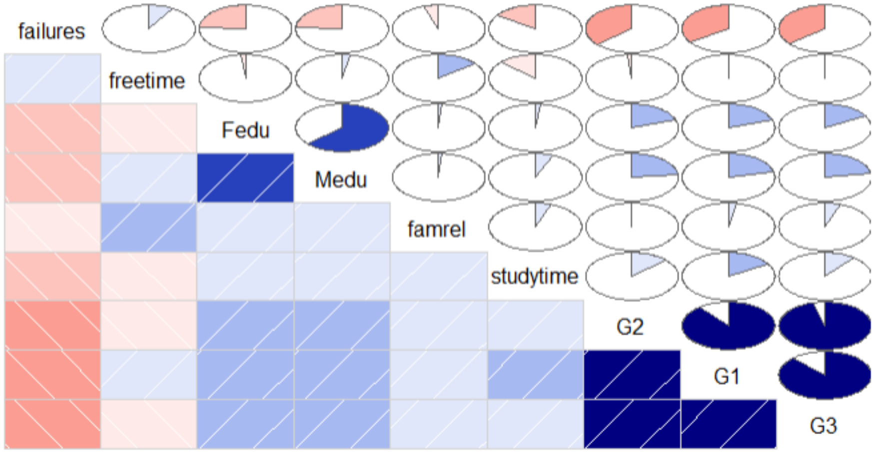
Correlation Matrix - Pearson



Correlation Matrix - Kendall

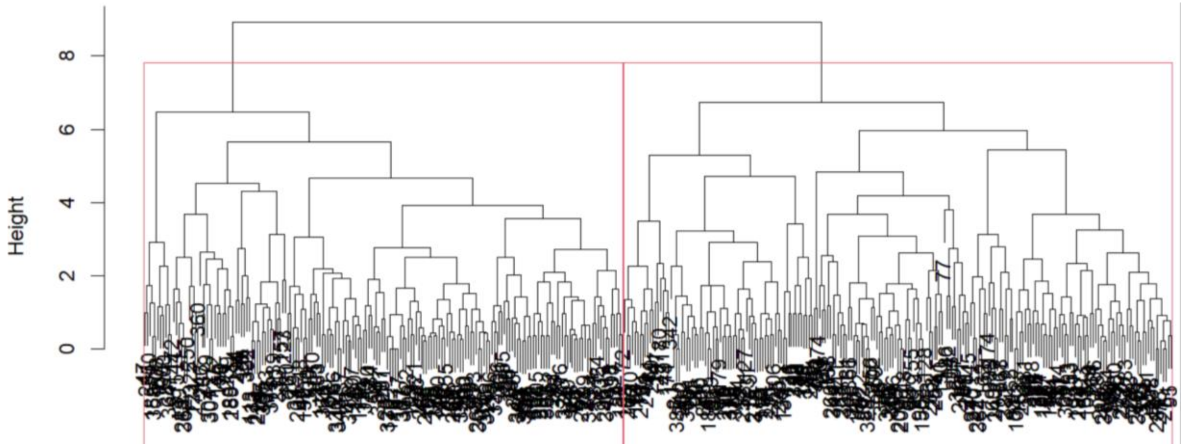


Correlation Matrix - Spearman



Clustering

Cluster Dendrogram



Variable Selected for the Cluster analysis

```
subset.clus <- subset(DataSet,select = c(Medu, Fedu, famrel, G1, G2, G3))
subset.sc <- scale(subset.clus)
subset.dist <- dist(subset.sc)
```

```
> subset.hc

Call:
hclust(d = subset.dist, method = "complete")

Cluster method   : complete
Distance         : euclidean
Number of objects: 395
```

Distribution of G1, G2, G3 in the Two Clusters

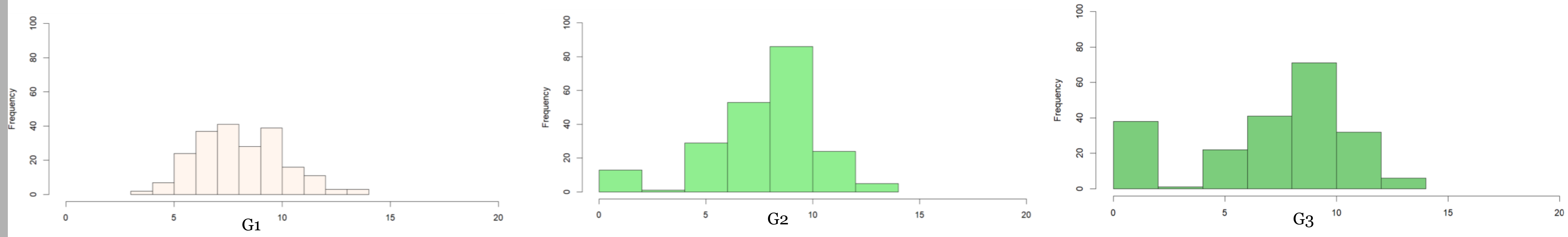
```
> table(subset.clus$G1, subset.hc.2)
    subset.hc.2
      1  2
  3   1  0
  4   1  0
  5   7  0
  6  24  0
  7  37  0
  8  41  0
  9  28  3
  10 39 12
  11 16 23
  12 11 24
  13  3 30
  14  3 27
  15  0 24
  16  0 22
  17  0  8
  18  0  8
  19  0  3
```

```
> table(subset.clus$G2, subset.hc.2)
    subset.hc.2
      1  2
  0  13  0
  4   1  0
  5  15  0
  6  14  0
  7  21  0
  8  32  0
  9  48  2
  10 38  8
  11 12 23
  12 12 29
  13  5 32
  14  0 23
  15  0 34
  16  0 13
  17  0  5
  18  0 12
  19  0  3
```

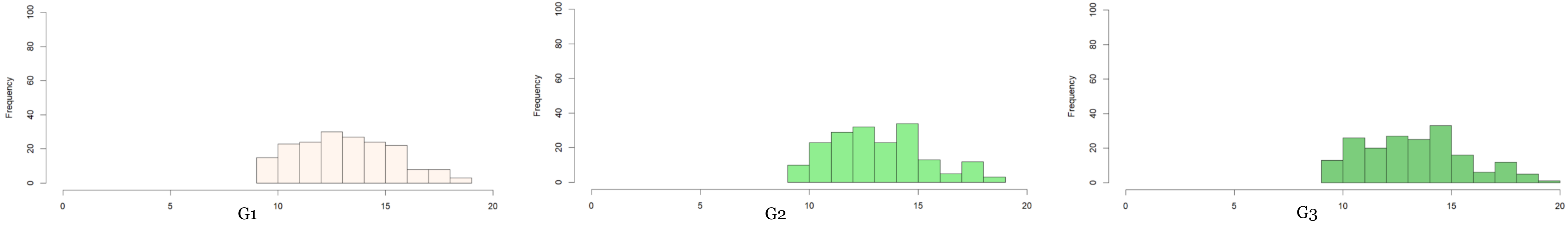
```
> table(subset.clus$G3, subset.hc.2)
    subset.hc.2
      1  2
  0  38  0
  4   1  0
  5   7  0
  6  15  0
  7   9  0
  8  32  0
  9  26  2
  10 45 11
  11 21 26
  12 11 20
  13  4 27
  14  2 25
  15  0 33
  16  0 16
  17  0  6
  18  0 12
  19  0  5
  20  0  1
```

Clustering

Distribution of G1, G2, G3 in the 1st Cluster



Distribution of G1, G2, G3 in the 2nd Cluster



Principal Component Analysis

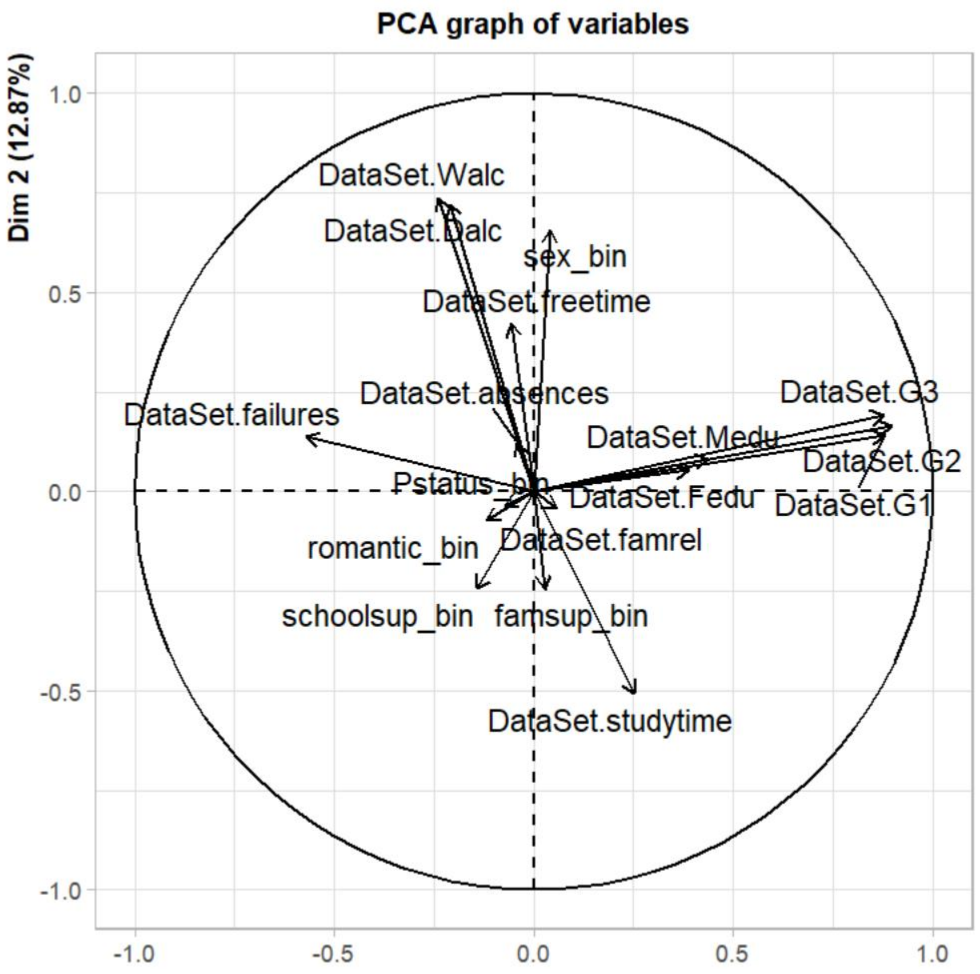
Convert some Categorical variables into Numerical

```
sex_bin <- ifelse(DataSet$sex == "M", 1, 0)
Pstatus_bin <- ifelse(DataSet$Pstatus == "T", 1, 0)
famsup_bin <- ifelse(DataSet$famsup == "yes", 1, 0)
schoolsup_bin <- ifelse(DataSet$schoolsup == "yes", 1, 0)
romantic_bin <- ifelse(DataSet$romantic == "yes", 1, 0)
```

```
## Replacing zeros in G2 and G3 with NA
Subset$G2 <- ifelse(Subset$G2==0, NA, Subset$G2)
Subset$G3 <- ifelse(Subset$G3==0, NA, Subset$G3)
## Omitting NAs
Subset <- na.omit(Subset)
```

Apply the PCA algorithm to the numerical Subset

```
> pca = prcomp(Subset, scale = TRUE)
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6    PC7    PC8
Standard deviation     1.7988 1.4793 1.30091 1.15165 1.05276 1.02990 1.0099 0.9229
Proportion of Variance 0.1903 0.1287 0.09955 0.07802 0.06519 0.06239 0.0600 0.0501
Cumulative Proportion  0.1903 0.3191 0.41862 0.49663 0.56183 0.62422 0.6842 0.7343
                           PC9    PC10    PC11    PC12    PC13    PC14    PC15
Standard deviation     0.91189 0.89207 0.81883 0.81330 0.77870 0.60715 0.57297
Proportion of Variance 0.04891 0.04681 0.03944 0.03891 0.03567 0.02168 0.01931
Cumulative Proportion  0.78323 0.83004 0.86948 0.90839 0.94406 0.96575 0.98506
                          PC16   PC17
Standard deviation     0.41514 0.2857
Proportion of Variance 0.01014 0.0048
Cumulative Proportion  0.99520 1.0000
```



Linear Regression

Linear Regression on all the features

```
lm0 <- lm(DataSet.G3 ~ ., data=Subset)
summary(lm0)
```

```
Call:
lm(formula = DataSet.G3 ~ ., data = Subset)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.1042 -0.5100  0.3037  0.9716  4.2735 
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -3.12840    0.81082  -3.858 0.000134 ***
sex_bin             0.08746    0.21712   0.403 0.687308    
Pstatus_bin        -0.26402    0.31688  -0.833 0.405271    
DataSet.Medu        0.11919    0.11583   1.029 0.304137    
DataSet.Fedu       -0.16640    0.11505  -1.446 0.148922    
famsup_bin          0.22552    0.20575   1.096 0.273744    
DataSet.famrel      0.33154    0.10892   3.044 0.002498 ** 
DataSet.studytime  -0.16523    0.12555  -1.316 0.188979    
DataSet.failures   -0.26186    0.14360  -1.824 0.069006 .  
schoolsup_bin       0.54869    0.29901   1.835 0.067289 .  
romantic_bin       -0.32589    0.20887  -1.560 0.119527    
DataSet.freetime    0.05040    0.10141   0.497 0.619490    
DataSet.Dalc       -0.15501    0.14302  -1.084 0.279134    
DataSet.Walc        0.17907    0.10083   1.776 0.076535 .  
DataSet.absences    0.03680    0.01243   2.960 0.003265 ** 
DataSet.G1          0.18883    0.05790   3.262 0.001209 ** 
DataSet.G2          0.95510    0.04985  19.161  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.88 on 378 degrees of freedom
Multiple R-squared:  0.8385,	Adjusted R-squared:  0.8317 
F-statistic: 122.7 on 16 and 378 DF,  p-value: < 2.2e-16
```

Linear Regression on grades G1 and G2

```
lm1 <- lm(DataSet.G3 ~ DataSet.G1 + DataSet.G2, data=Subset)
summary(lm1)
```

```
Call:
lm(formula = DataSet.G3 ~ DataSet.G1 + DataSet.G2, data = Subset)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5713 -0.3888  0.2885  0.9725  3.7089 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.83001    0.33531  -5.458 8.57e-08 ***
DataSet.G1   0.15327    0.05618   2.728  0.00665 ** 
DataSet.G2   0.98687    0.04957  19.909  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.937 on 392 degrees of freedom
Multiple R-squared:  0.8222,	Adjusted R-squared:  0.8213 
F-statistic: 906.1 on 2 and 392 DF,  p-value: < 2.2e-16
```

Linear Regression

Linear Regression on family variables

```
lm2 <- lm(DataSet.G3 ~ Pstatus_bin + DataSet.Fedu + DataSet.Medu + famsup_bin + DataSet.famrel, data=Subset)
summary(lm2)
```

```
Call:
lm(formula = DataSet.G3 ~ Pstatus_bin + DataSet.Fedu + DataSet.Medu +
    famsup_bin + DataSet.famrel, data = Subset)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2716  -2.0376   0.6069   2.9067   9.6411

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.4851     1.3934   5.372 1.34e-07 ***
Pstatus_bin     -0.4354     0.7451  -0.584  0.55932
DataSet.Fedu     0.1551     0.2662   0.583  0.56056
DataSet.Medu     0.8617     0.2656   3.245  0.00128 **
famsup_bin      -0.7723     0.4728  -1.633  0.10323
DataSet.famrel   0.2620     0.2516   1.041  0.29842
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.475 on 389 degrees of freedom
Multiple R-squared:  0.05786,   Adjusted R-squared:  0.04575
F-statistic: 4.778 on 5 and 389 DF,  p-value: 0.000299
```

Linear Regression on studytime and failures

```
lm3 <- lm(DataSet.G3 ~ DataSet.studytime + DataSet.failures, data=Subset)
summary(lm3)
```

```
Call:
lm(formula = DataSet.G3 ~ DataSet.studytime + DataSet.failures,
    data = Subset)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5342  -1.9556   0.0613   3.0359   9.2429

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        10.7402     0.5970  17.991  < 2e-16 ***
DataSet.studytime   0.1985     0.2610   0.761    0.447
DataSet.failures   -2.1815     0.2945  -7.407 7.97e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.281 on 392 degrees of freedom
Multiple R-squared:  0.1312,    Adjusted R-squared:  0.1267
F-statistic: 29.59 on 2 and 392 DF,  p-value: 1.072e-12
```

By comparing the Adjusted R-squared of the models we can asses that the best one is the linear regression on all the features.

Support Vector Machine

Set Train and Test Partition

```
sample.svm <- sample(nrow(Subset), nrow(Subset)*0.8)

train.svm <- Subset[sample.svm,]
train.svm <- data.frame(train.svm)

test.svm <- Subset[-sample.svm,]
test.svm <- data.frame(test.svm)
```

SVM on the whole Train Data

```
> subset.lsvm <- svm(DataSet.G3 ~., data = train.svm, type = "C-classification", kernel =
 "linear")
```

Make the prediction of G3 on Test Data

```
> pred <- predict(subset.lsvm, test.svm)
> table(pred, DataSet.G3)
    DataSet.G3
pred 0 5 6 7 8 9 10 11 12 13 14 15 16 17 18
  0  4 0 0 1 0 1  0  0  0  0  0  0  0  0  0
  4  0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
  5  0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
  6  0 0 2 0 2 0  0  0  0  0  0  0  0  0  0
  7  0 1 0 0 0 0  2  0  0  0  0  0  0  0  0
  8  1 1 2 1 1 0  0  1  0  0  0  0  0  0  0
  9  2 0 0 1 2 2  2  0  0  0  0  0  0  0  0
  10 1 0 0 0 1 2  3  1  0  0  0  0  0  0  0
  11 1 0 0 0 0 1  0  1  3  1  1  0  0  0  0
  12 0 0 0 0 0 0  1  2  0  0  0  0  0  0  0
  13 0 0 0 0 0 0  0  2  3  0  6  0  0  0  0
  14 0 0 0 0 0 0  0  0  2  0  2  1  0  0  0
  15 0 0 0 0 0 0  0  0  0  0  3  3  4  0  1
  16 0 0 0 0 0 0  0  0  0  0  0  1  1  1  1
  17 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
  18 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
  19 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
  20 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
```

SVM on Train Data – Related only to family features

```
> fam.lsvm <- svm(DataSet.G3 ~ Pstatus_bin + DataSet.Medu + DataSet.Fedu + DataSet.famrel
 + famsup_bin + schoolsup_bin, data = train.svm, type = "C-classification", kernel = "lin
ear")
```

Make the prediction of G3 on Test Data

```
> table(fam.pred, DataSet.G3)
        DataSet.G3
fam.pred 0 5 6 7 8 9 10 11 12 13 14 15 16 17 18
      0  1 0 0 0 0 0  0  1  0  0  0  1  0  0  0
      4  0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      5  0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      6  0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      7  0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      8  0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      9  0 0 1 0 0 0  0  0  0  0  0  0  0  0  0
      10 8 2 2 2 5 5  6  5  7  0  9  4  3  1  0
      11 0 0 1 0 1 0  0  1  1  1  2  0  1  0  0
      12 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      13 0 0 0 1 0 0  2  0  0  0  0  0  1  0  0
      14 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      15 0 0 0 0 0 1  0  0  0  0  1  0  0  0  2
      16 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      17 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      18 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      19 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      20 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
```

Support Vector Machine

SVM on Train Data – Related only to study and alcohol features

```
> stu.lsvm <- svm(DataSet.G3 ~ DataSet.studytime + DataSet.failures + DataSet.freetime +
 DataSet.Dalc + DataSet.Walc + DataSet.absences, data = train.svm, type = "C-classificati
on", kernel = "linear")
```

```
> stu.pred <- predict(stu.lsvm, test.svm)
> table(stu.pred, DataSet.G3)
        DataSet.G3
stu.pred 0 5 6 7 8 9 10 11 12 13 14 15 16 17 18
      0  7 0 0 0 0 1  2  0  0  0  1  1  0  0  0
      4  0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      5  0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      6  0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      7  0 0 0 0 0 0  1  0  0  0  0  0  0  0  0
      8  0 0 0 1 1 1  1  2  0  0  0  0  0  0  0
      9  0 0 0 1 0 0  0  0  0  0  0  0  0  0  0
      10 2 1 2 0 2 1  2  3  5  1  5  3  2  0  2
      11 0 0 2 1 2 2  0  1  2  0  4  1  1  0  0
      12 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      13 0 1 0 0 0 1  1  0  0  0  0  0  0  0  0
      14 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      15 0 0 0 0 1 0  1  1  1  0  2  0  2  1  0
      16 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      17 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      18 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      19 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
      20 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
```

Make the prediction of G3 on Test Data

SVM on Train Data – Related only to failures, G1, G2

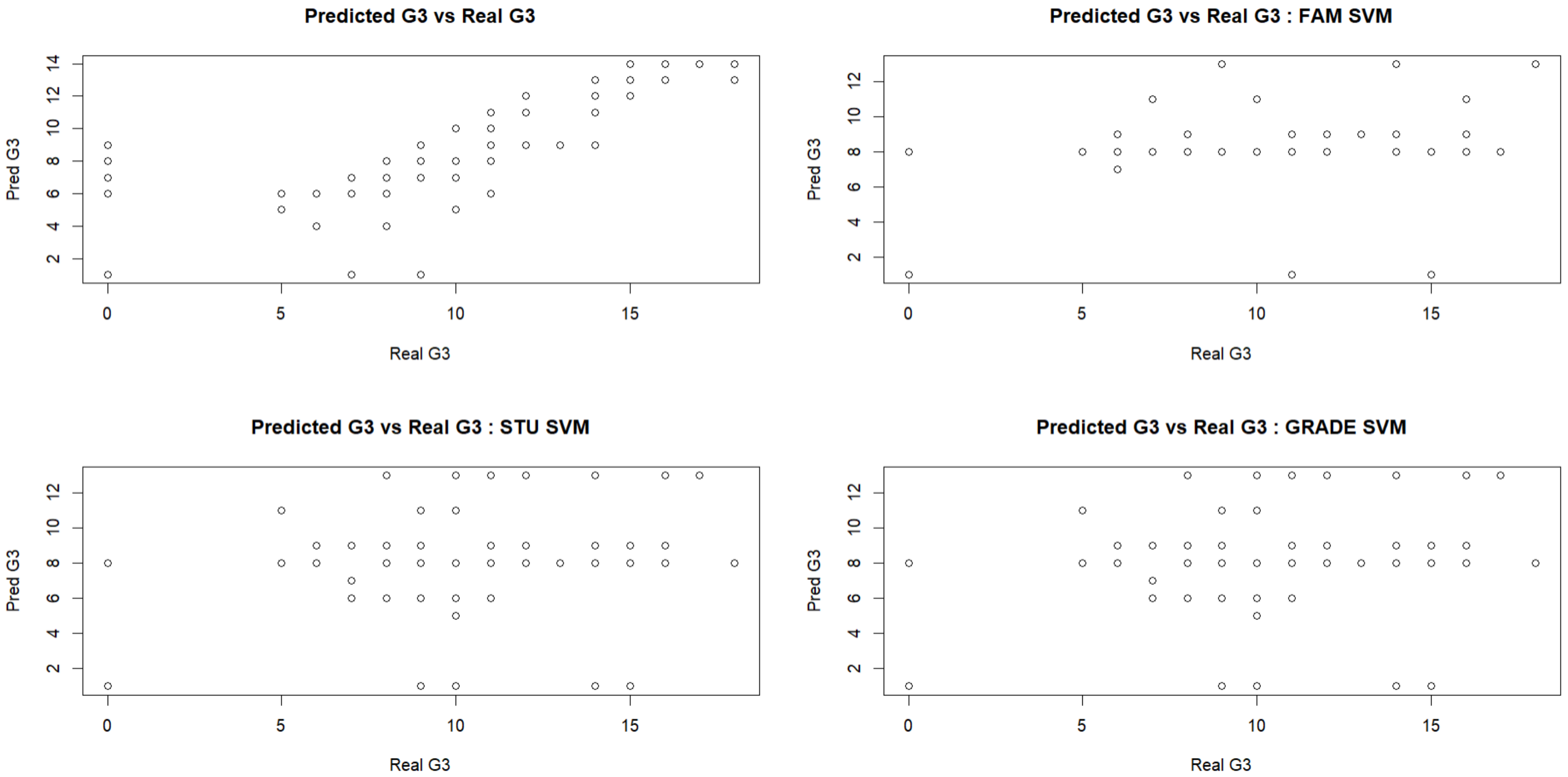
```
> grade.lsvm <- svm(DataSet.G3 ~ DataSet.failures + DataSet.G1 + DataSet.G2, data = trai
n.svm, type = "C-classification", kernel = "linear")
```

Make the prediction of G3 on Test Data

```
> grade.pred <- predict(stu.lsvm, test.svm)
> table(grade.pred, DataSet.G3)
          DataSet.G3
grade.pred 0 5 6 7 8 9 10 11 12 13 14 15 16 17 18
        0  7 0 0 0 0 1  2  0  0  0  1  1  0  0  0
        4  0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
        5  0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
        6  0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
        7  0 0 0 0 0 0  1  0  0  0  0  0  0  0  0
        8  0 0 0 1 1 1  1  2  0  0  0  0  0  0  0
        9  0 0 0 1 0 0  0  0  0  0  0  0  0  0  0
        10 2 1 2 0 2 1  2  3  5  1  5  3  2  0  2
        11 0 0 2 1 2 2  0  1  2  0  4  1  1  0  0
        12 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
        13 0 1 0 0 0 1  1  0  0  0  0  0  0  0  0
        14 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
        15 0 0 0 0 1 0  1  1  1  0  2  0  2  1  0
        16 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
        17 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
        18 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
        19 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
        20 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0
```

Support Vector Machine

Plot of Predicted Values against Real Values for all the previous models



Model Comparison

Linear Regression on G1, G2 vs SVM on failures, G1, G2

```
pred.lm1 <- predict(lm1, newdata = test.svm)
pred.grade.lsvm <- predict(grade.lsvm, newdata = test.svm)
```

```
rmse_lm1 <- rmse(DataSet.G3, pred.lm1)
rmse_grade.lsvm <- rmse(DataSet.G3, as.numeric(pred.grade.lsvm))
```

```
> rmse_lm1
[1] 2.377473
> rmse_grade.lsvm
[1] 3.05436
```

Linear Regression wins

Linear Regression on Family variables vs SVM on Family features

```
pred.lm2 <- predict(lm2, newdata = test.svm)
pred.fam.lsvm <- predict(fam.lsvm, newdata = test.svm)
```

```
rmse_lm2 <- rmse(DataSet.G3, pred.lm2)
rmse_fam.lsvm <- rmse(DataSet.G3, as.numeric(pred.fam.lsvm))
```

```
> rmse_lm2
[1] 4.620417
> rmse_fam.lsvm
[1] 5.112655
```

Linear Regression wins

Linear Regression on the whole Dataset vs SVM on the whole Dataset

```
pred.lm0 <- predict(lm0, newdata = test.svm)
pred.subset.lsvm <- predict(subset.lsvm, newdata = test.svm)
```

```
rmse_lm0 <- rmse(DataSet.G3, pred.lm1)
rmse_subset.lsvm <- rmse(DataSet.G3, as.numeric(pred.grade.lsvm))
```

```
> rmse_lm0
[1] 2.377473
> rmse_subset.lsvm
[1] 3.05436
```

Linear Regression wins

Thanks for your Attention

Nava Carlo
Passaro Jacopo