Perfomance Task

Karl Mbouombouo

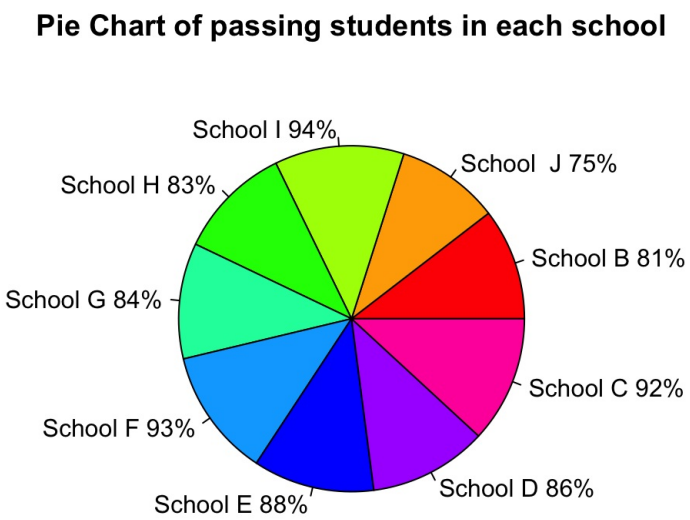
2023-03-19

Packages

Import, and “clean” the data

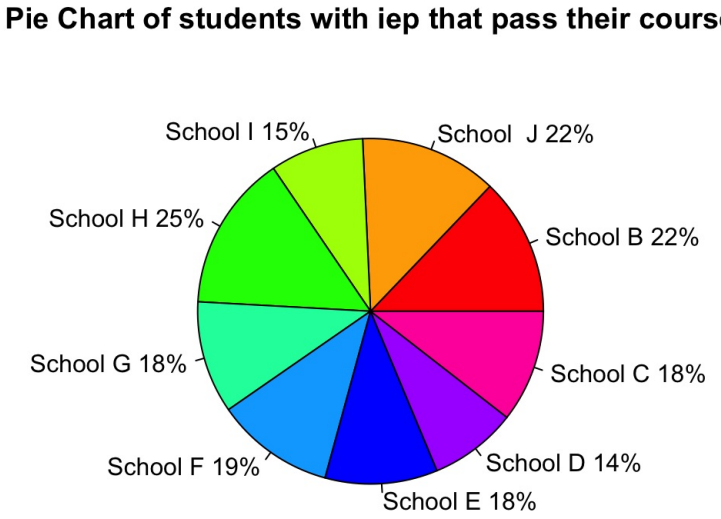
ETL

Top 3 schools with the most students passing their course

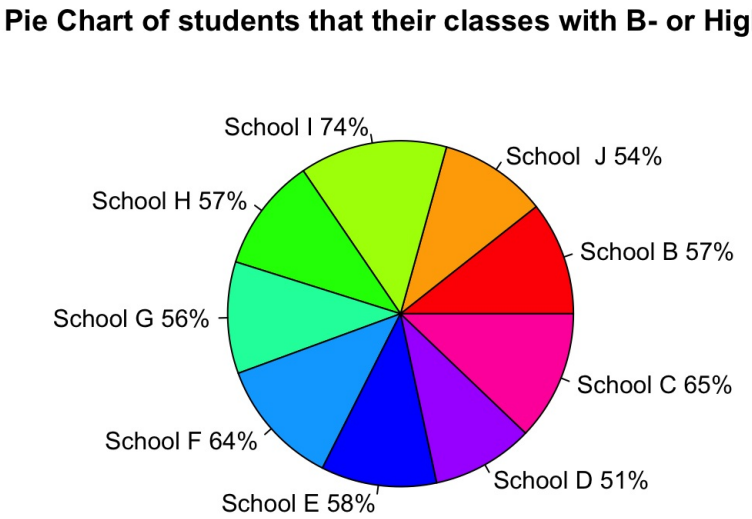


The Top 3 schools with the most students that passe their course are schools I with 94%, school F with 93% and school C with 92% Overall ranking: 1- School I: 94% 2- School F: 93% 3- School C: 92% 4- School E: 88% 5- School D: 86% 6- School G: 84% 7- School H: 83% 8- School B: 81% 9- School J: 75%

```
slices3 <- c(22, 22, 15, 25, 18, 19, 18, 14, 18)
labels <- c("School B", "School  J", "School I", "School H", "School G",
          "School F", "School E", "School D", "School C")
labels <- paste(labels, slices3)
labels <- paste(labels,"%",sep="")
pie(slices3,labels = labels, col=rainbow(length(labels)),
   main="Pie Chart of students with iep that pass their course")
```



Top 3 schools with the most students passing their course with B- or higher



The Top 3 schools with the most students that passe their course with B- or higher are schools I with 74%, school C with 65% and school F with 64% Overall ranking: 1- School I 74% 2- School C 65% 3- School F 64% 4- School E 58% 5- School B 57% 6- School H 57% 7- School G 56% 8- School J 54% 9- School D 51%

```
survey_by_school_ieptg <- group_by(df3, school_name, iep == '1', ranking == "1")
print(summarize(survey_by_school_ieptg, counts = n()), n=36)
```

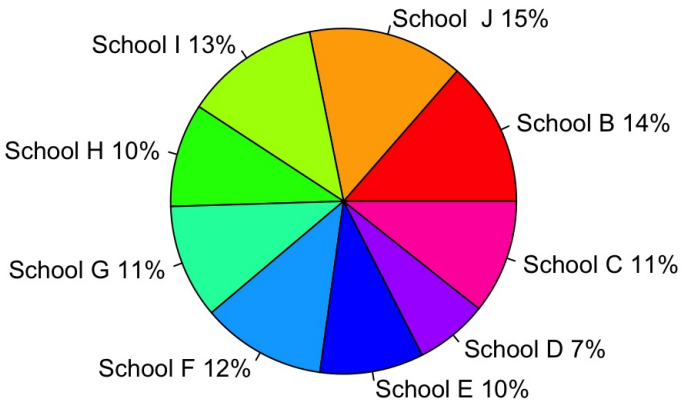
```
## `summarise()` has grouped output by 'school_name', 'iep == "1"'. You can
## override using the `.groups` argument.
```

```
## # A tibble: 36 × 4
## # Groups:   school_name, iep == "1" [18]
##    school_name `iep == "1"` `ranking == "1"` counts
##    <chr>       <lgl>        <lgl>              <int>
##  1 SCHOOL B    FALSE        FALSE                247
##  2 SCHOOL B    FALSE        TRUE                 204
##  3 SCHOOL B    TRUE         FALSE                 61
##  4 SCHOOL B    TRUE         TRUE                  66
##  5 SCHOOL C    FALSE        FALSE                240
##  6 SCHOOL C    FALSE        TRUE                 375
##  7 SCHOOL C    TRUE         FALSE                 61
##  8 SCHOOL C    TRUE         TRUE                  77
##  9 SCHOOL D    FALSE        FALSE                272
## 10 SCHOOL D    FALSE        TRUE                 214
## 11 SCHOOL D    TRUE         FALSE                 44
## 12 SCHOOL D    TRUE         TRUE                  32
## 13 SCHOOL E    FALSE        FALSE                241
## 14 SCHOOL E    FALSE        TRUE                 246
## 15 SCHOOL E    TRUE         FALSE                 47
## 16 SCHOOL E    TRUE         TRUE                  53
## 17 SCHOOL F    FALSE        FALSE                285
## 18 SCHOOL F    FALSE        TRUE                 414
## 19 SCHOOL F    TRUE         FALSE                 69
## 20 SCHOOL F    TRUE         TRUE                  99
## 21 SCHOOL G    FALSE        FALSE                285
## 22 SCHOOL G    FALSE        TRUE                 235
## 23 SCHOOL G    TRUE         FALSE                 54
## 24 SCHOOL G    TRUE         TRUE                  61
## 25 SCHOOL H    FALSE        FALSE                347
## 26 SCHOOL H    FALSE        TRUE                 299
## 27 SCHOOL H    TRUE         FALSE                 84
## 28 SCHOOL H    TRUE         TRUE                  84
## 29 SCHOOL I    FALSE        FALSE                241
## 30 SCHOOL I    FALSE        TRUE                 506
## 31 SCHOOL I    TRUE         FALSE                 27
## 32 SCHOOL I    TRUE         TRUE                 106
## 33 SCHOOL J    FALSE        FALSE                325
## 34 SCHOOL J    FALSE        TRUE                 195
## 35 SCHOOL J    TRUE         FALSE                 69
## 36 SCHOOL J    TRUE         TRUE                  75
```

```
#Percent of student with iep that passed their course with a B- or higher per school
ieptg_B = 66/471
ieptg_C = 77/691
ieptg_D = 32/481
ieptg_E = 53/519
ieptg_F = 99/805
ieptg_G = 61/532
ieptg_H = 84/814
ieptg_I = 106/829
ieptg_J = 75/496
```

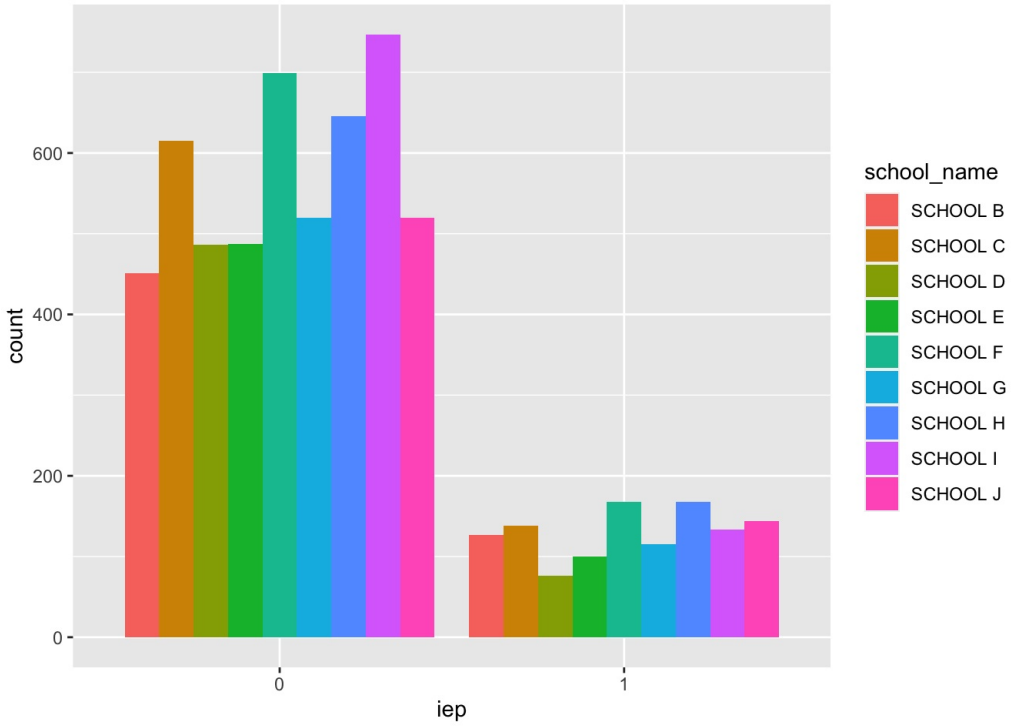
```
slices3 <- c(14, 15, 13, 10, 11, 12, 10, 7, 11)
labels <- c("School B", "School  J", "School I", "School H", "School G",
          "School F", "School E", "School D", "School C")
labels <- paste(labels, slices3)
labels <- paste(labels,"%",sep="")
pie(slices3,labels = labels, col=rainbow(length(labels)),
   main="Pie Chart of students with iep that their classes with B- or Higher")
```

Pie Chart of students with iep that their classes with B- or Higher

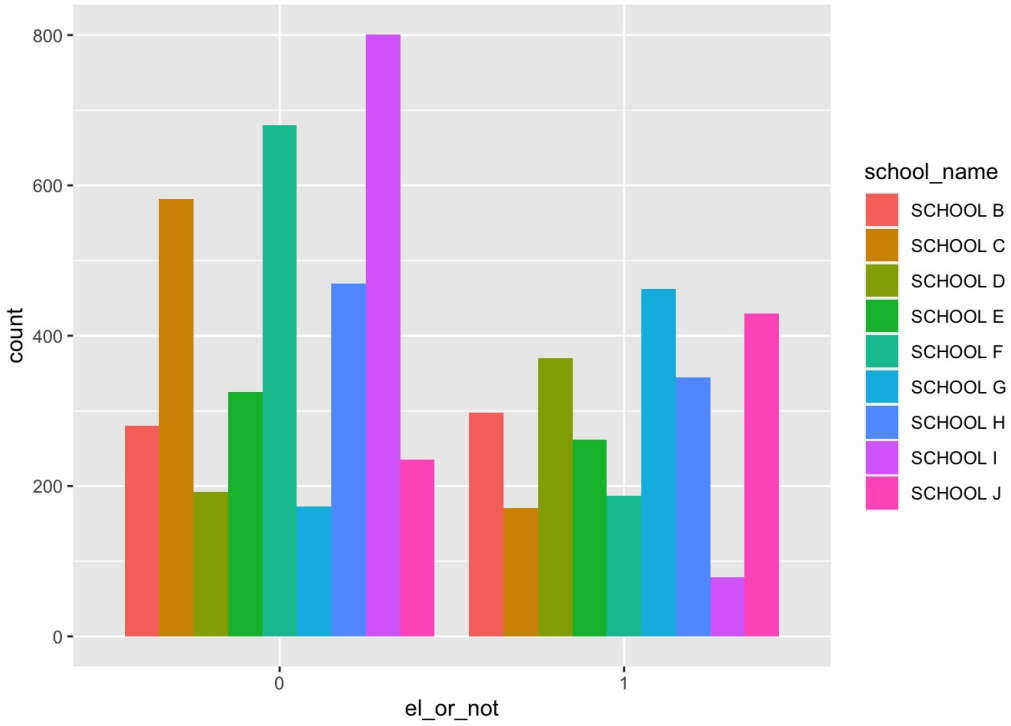


Analyse iep, el_or_not and ecodis

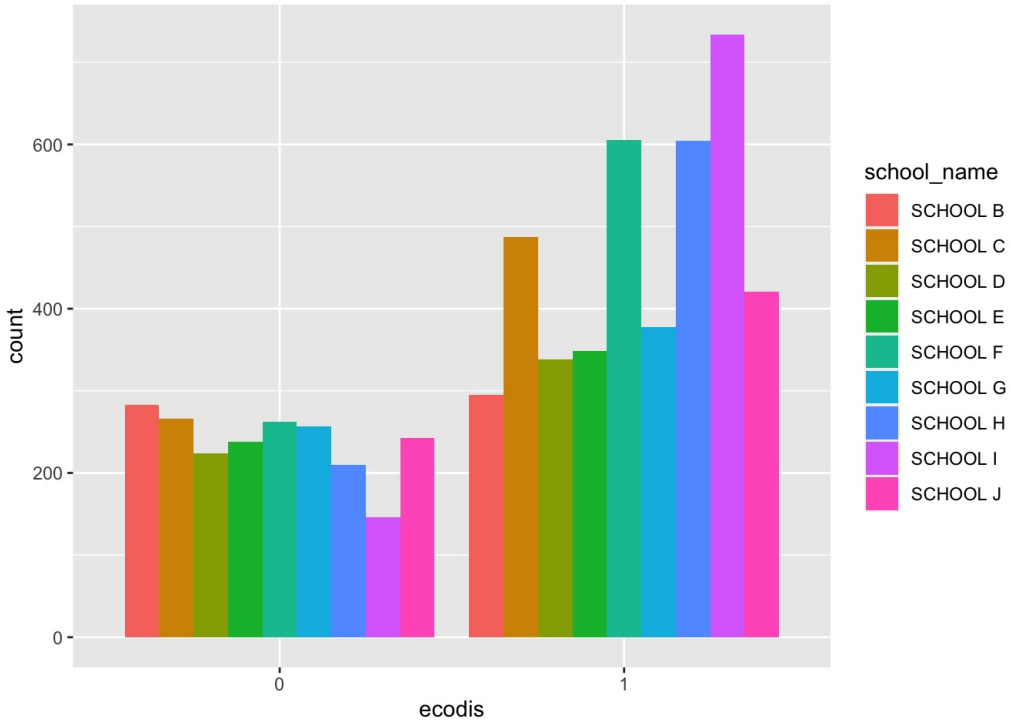
```
ggplot(data = df4, aes(x = iep, fill = school_name)) +
  geom_bar(position = position_dodge())
```




```
ggplot(data = df4, aes(x = el_or_not, fill = school_name)) +
  geom_bar(position = position_dodge())
```




```
ggplot(data = df4, aes(x = ecodis, fill = school_name)) +
  geom_bar(position = position_dodge())
```

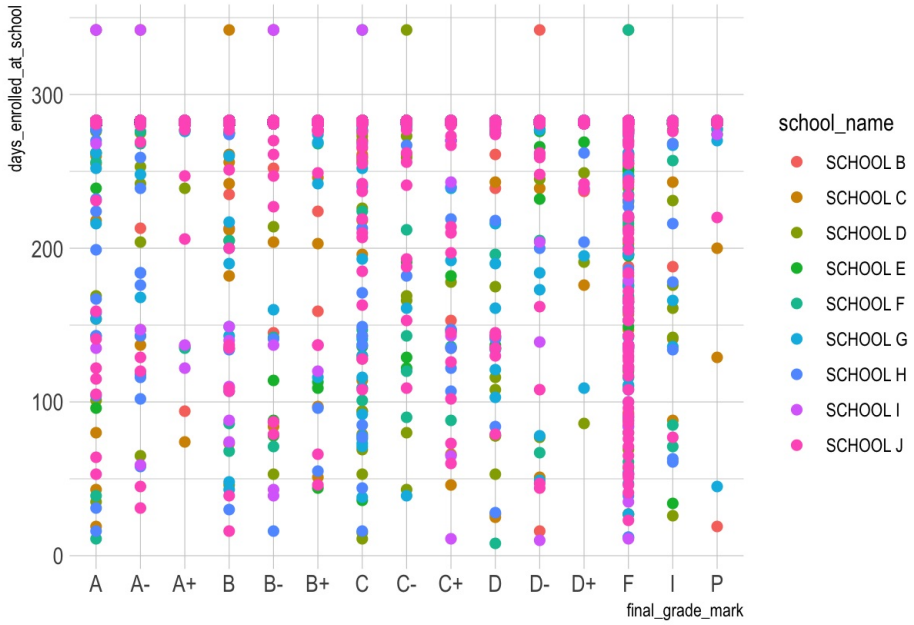


1- The schools that have the most individualized education plan installed are school H, F, J, C. School F and C are the schools with the highest percentage with students passing their classes with B- or higher. So, it might not be in relationship with the final grade because they do not have much students with disabilities than others. So we might considerate the effect that have individualized education plan on the final grades

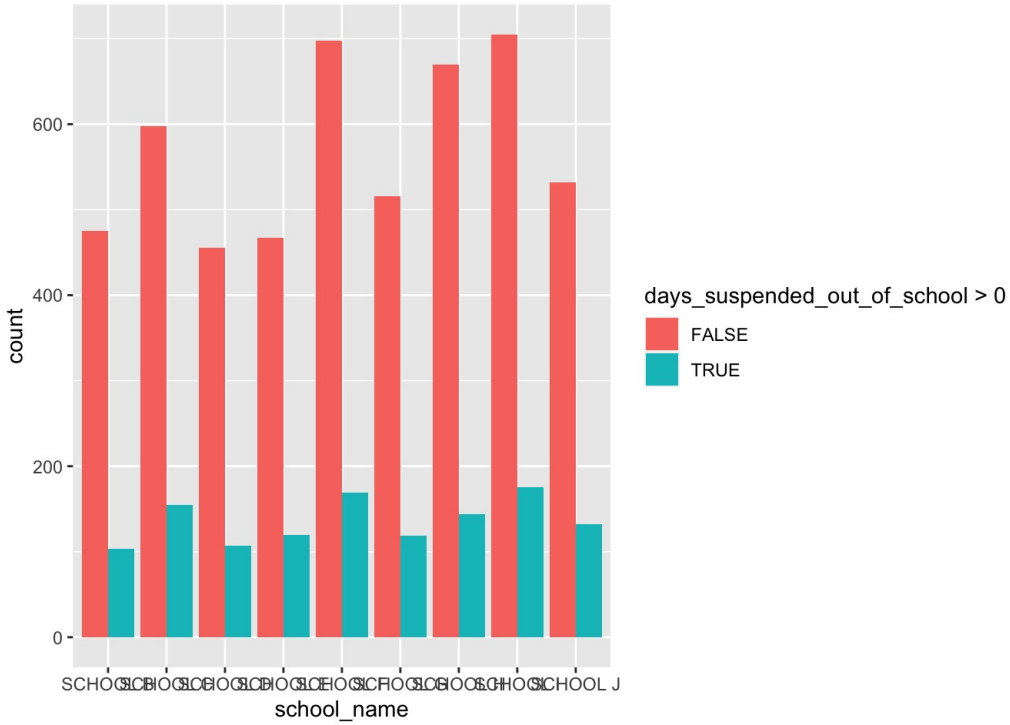
2- The schools with the most student who have never been considered an English Learner are school I, F, C. Those are the school that have the most percentage with student that finish their course with a final great with a B- or higher. So we might exist a relationship between the students that never been considerate as English learning and their final grades.

3- The schools with the most student classified as economically disadvantaged are schools I, F, H. Once again school I and F are the second that have the best results regarding their final grades results. So, a relationship might exist between student budget and their final grades.

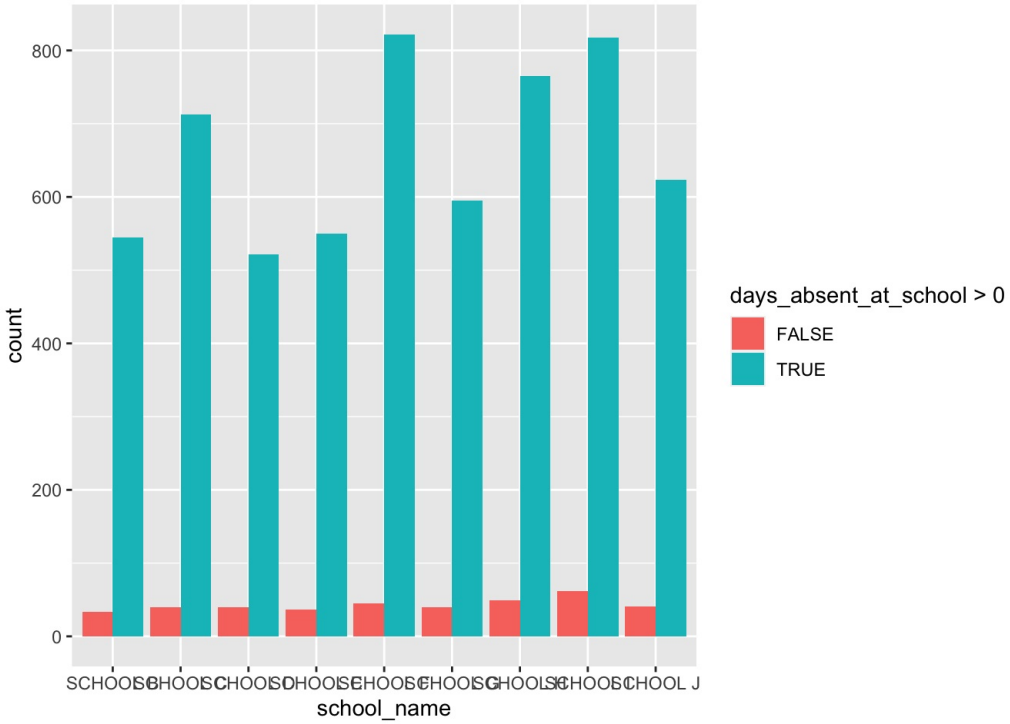
```
ggplot(df3, aes(x=final_grade_mark, y=days_enrolled_at_school, color=school_name)) +
    geom_point(size=2) +
    theme_ipsum()
```




```
ggplot(data = df3, aes(x = school_name, fill = days_suspended_out_of_school > 0)) +
  geom_bar(position = position_dodge())
```




```
ggplot(data = df3, aes(x = school_name, fill = days_absent_at_school > 0)) +
  geom_bar(position = position_dodge())
```



The top 3 schools that have the highest percentage with students that finish their course with a B- or higher are school I, C and F. But also those school are the school that have the most students passing their classes. But those school got a bigger sample of students, so they have more students to study and might affect the results of students with final grade with a B- or higher.

Machine Learning - Logistics regression

```
#Put variable as factor
df4$success <- as.factor(df4$success) # Passing class = 1, Failing class =2
df4$ranking <- as.factor(df4$ranking)
df4$iep <- as.factor(df4$iep) # Don't Have a EP = 1, Have a EP = 2
df4$el_status <- as.factor(df4$el_or_not) # Current EL = 1, Former EL = 2, Never EL = 3
df4$ecodis <- as.factor(df4$ecodis) # not economically disadvantaged = 1, economically disadvantaged = 2
str(df4)
```

```
## tibble [6,340 × 11] (S3: tbl_df/tbl/data.frame)
##  $ school_name                : chr [1:6340] "SCHOOL B" "SCHOOL B" "SCHOOL B" "SCHOOL B" ...
##  $ el_status                  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ iep                        : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 2 1 2 1 ...
##  $ ecodis                     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ days_enrolled_at_school    : num [1:6340] 283 283 279 283 283 283 283 283 283 206 ...
##  $ days_absent_at_school      : num [1:6340] 0 4 19 6 3 3 6 2 15 5 ...
##  $ days_suspended_out_of_school: num [1:6340] 0 0 0 0 0 0 0 0 0 0 ...
##  $ final_grade_mark           : chr [1:6340] "F" "D-" "F" "D" ...
##  $ ranking                    : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 1 1 1 ...
##  $ success                    : Factor w/ 2 levels "No","Yes": 1 2 1 2 2 2 2 2 2 1 ...
##  $ el_or_not                  : chr [1:6340] "1" "1" "1" "1" ...
```

```
#Set the seed
set.seed(1234567)

#Use 70% of dataset as training set and remaining 30% as testing set
sample <- sample(c(TRUE, FALSE), nrow(df4), replace=TRUE, prob=c(0.7,0.3))
train <- df4[sample, ]
test <- df4[!sample, ]

#Built the predictive model
model <- glm(success~iep+el_or_not+ecodis+days_enrolled_at_school+days_absent_at_school+
               days_suspended_out_of_school,
               family="binomial", data=train)
summary(model)
```

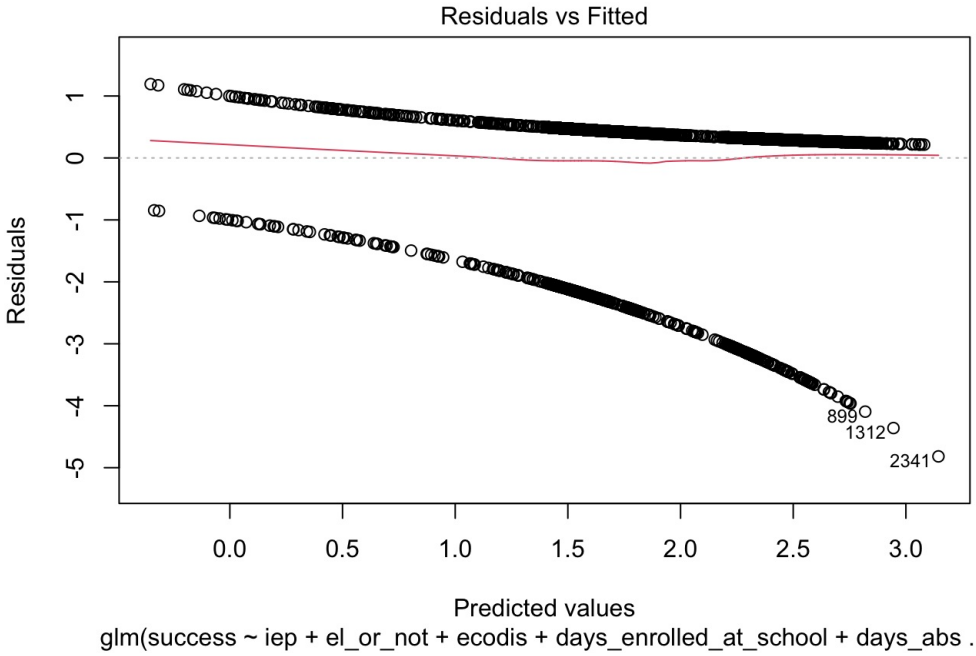
```
##
## Call:
## glm(formula = success ~ iep + el_or_not + ecodis + days_enrolled_at_school +
##     days_absent_at_school + days_suspended_out_of_school, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5245   0.3911   0.4358   0.5866   1.3299
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   0.2716679  0.2336758   1.163   0.2450
## iep1                          0.2464135  0.1247023   1.976   0.0482 *
## el_or_not1                   -0.7828579  0.0935284  -8.370   <2e-16 ***
## ecodis1                       0.0842742  0.0954950   0.882   0.3775
## days_enrolled_at_school       0.0068566  0.0007992   8.579   <2e-16 ***
## days_absent_at_school         0.0059197  0.0027158   2.180   0.0293 *
## days_suspended_out_of_school -0.0085189  0.0121099  -0.703   0.4818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3452.5  on 4464  degrees of freedom
## Residual deviance: 3285.3  on 4458  degrees of freedom
## AIC: 3299.3
##
## Number of Fisher Scoring iterations: 5
```

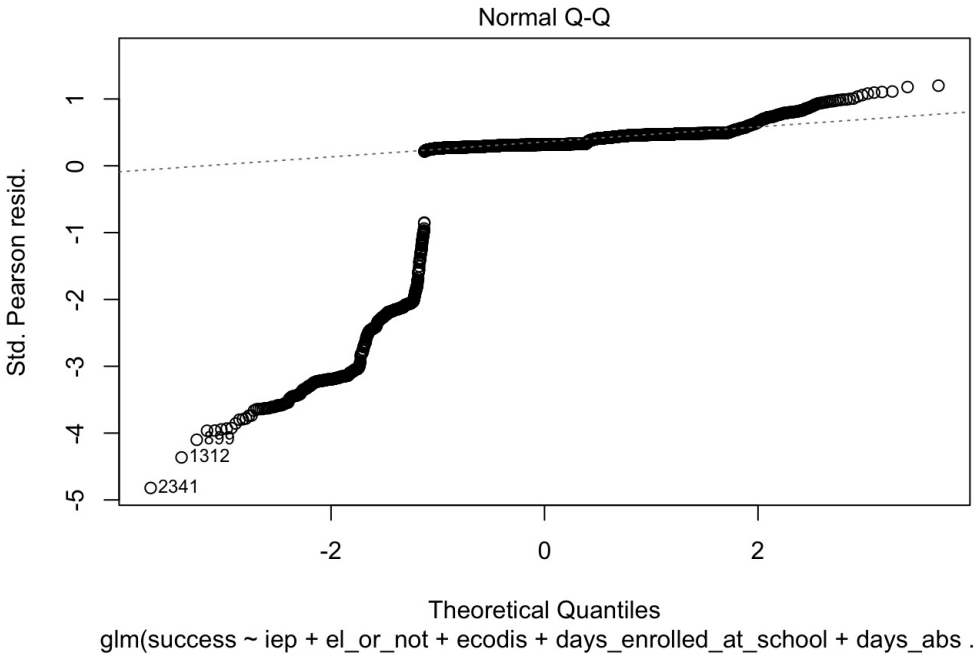
```
pscl::pR2(model)["McFadden"]
```

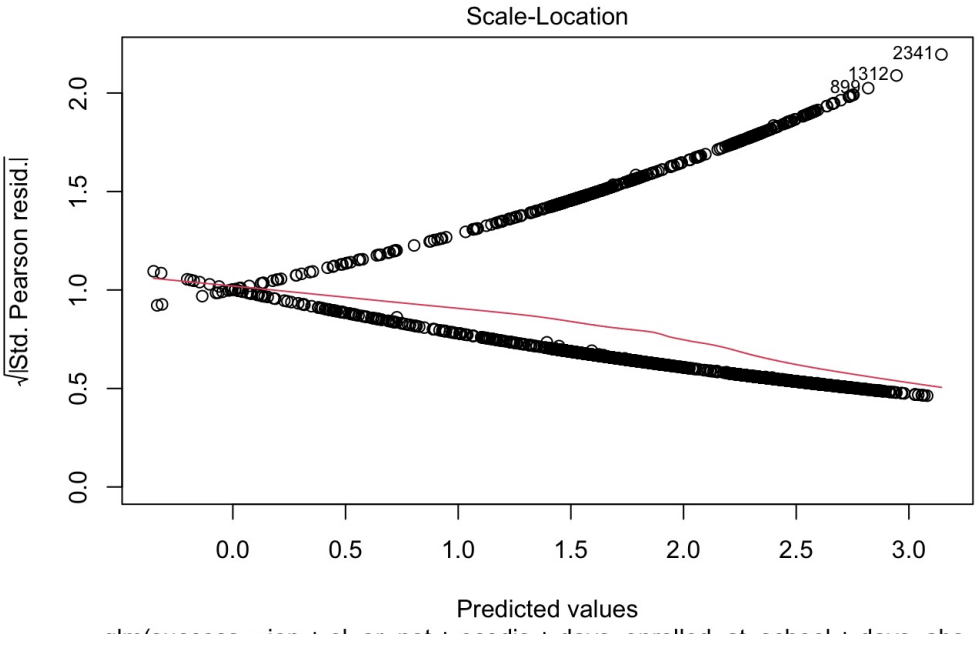
```
## fitting null model for pseudo-r2
```

```
##   McFadden
## 0.04843518
```

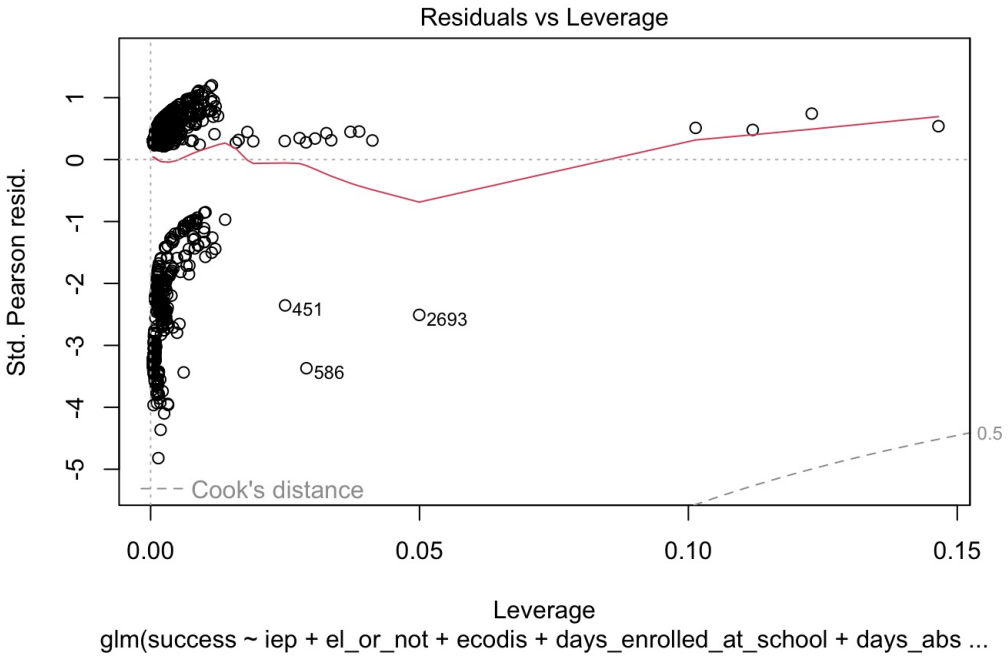
```
plot(model)
```


Residuals vs Fitted
Residuals
Predicted values
glm(success ~ iep + el_or_not + ecodis + days_enrolled_at_school + days_abs ...
899
1312
2341


Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
glm(success ~ iep + el_or_not + ecodis + days_enrolled_at_school + days_abs ...
899
1312
2341


Scale-Location
√|Std. Pearson resid.|
Predicted values
2341
1312
899

glm(success ~ iep + el_or_not + ecodis + days_enrolled_at_school + days_abs ...



glm(success ~ iep + el_or_not + ecodis + days_enrolled_at_school + days_abs ...

```
caret::varImp(model)
```

```
##                               Overall
## iep1                        1.9760148
## el_or_not1                  8.3702681
## ecodis1                     0.8824987
## days_enrolled_at_school     8.5794025
## days_absent_at_school       2.1797480
## days_suspended_out_of_school 0.7034669
```

```
# calculate probability of default for each individual in test dataset
predicted <- predict(model, test, type="response")

#convert success from "passed" and "failed" to 1's and 0's
test$success <- ifelse(test$success=="Yes", 1, 0)

#find optimal cutoff probability to use to maximize accuracy
optimal <- optimalCutoff(test$success, predicted)[1]
optimal
```

```
## [1] 0.4332874
```

```
confusionMatrix(test$success, predicted)
```

```
##     0    1
## 0   1    6
## 1 264 1604
```

```
#calculate sensitivity
sensitivity(test$success, predicted)
```

```
## [1] 0.9962733
```

```
#calculate specificity
specificity(test$success, predicted)
```

```
## [1] 0.003773585
```

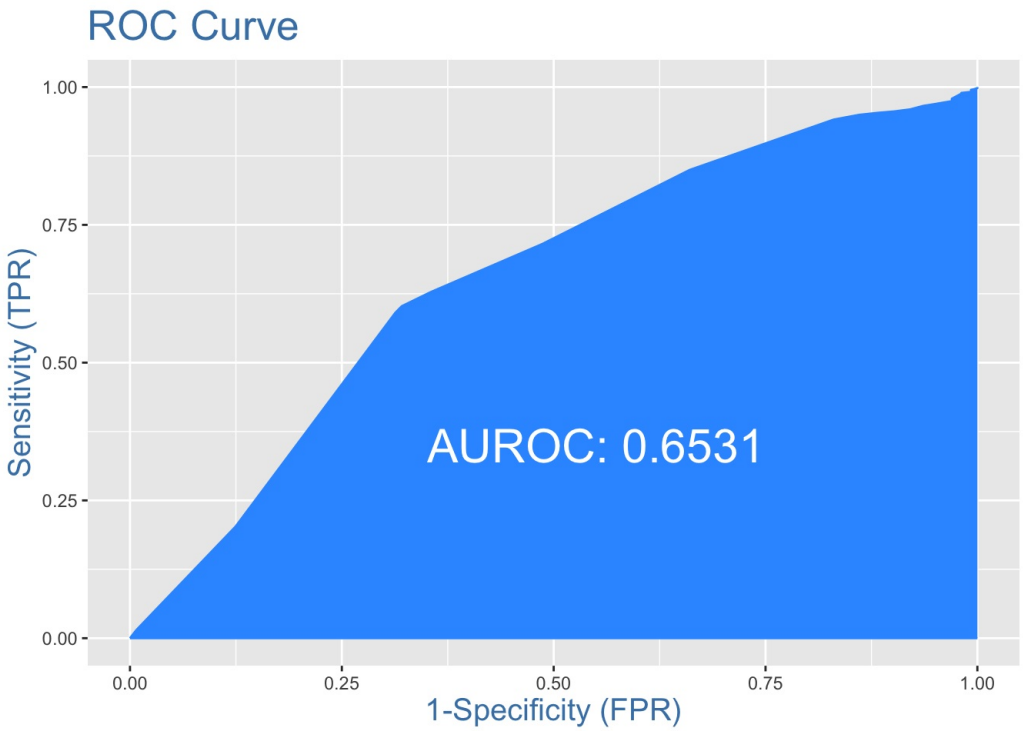
#Misclassification

```
#calculate total misclassification error rate
misClassError(test$success, predicted, threshold=optimal)
```

```
## [1] 0.1419
```

#Roc Curve

```
#plot the ROC curve
plotROC(test$success, predicted)
```



We can see that the AUC is 0.8437, which is quite high. This indicates that our model does a good job of predicting whether or not an individual will pass the course.