Title: Predicting Student Exam Performance

Author: Tina M. Kirk

Purpose: Predict whether a student will pass (1) or fail (0) an exam based on the number of study hours and their scores in the previous exam.

Dataset features [1]:

- **Study hours (numeric)-**number of hours a student spent studying for the upcoming exam.
- **Previous exam score (numeric)-**indicates the student's score on the previous exam.
- **Pass/Fail (binary)-**The target variable, where 1 represents a pass and 0 represents a fail on the current exam.
- **Dataset size:** consists of data for 500 students, with a diverse range of study patterns and previous exam performances.

Data Cleaning and Preparation

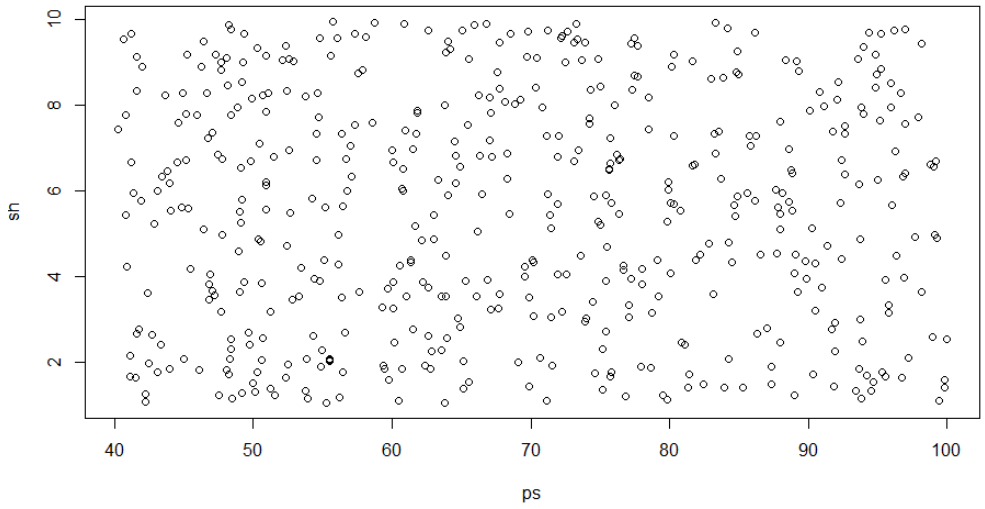
- Upon loading the CSV file into Excel, the numbers in the first two columns included the single quote character at the beginning, so it read each value as text instead of number. That character was not present in the raw data, but a special character in Excel to tell it to treat the rest of the string as text. I took care of this by converting the cells to number type.
- Imported the data into Rstudio as comma separated text file.
- To treat the Pass/Fail column as binary, this column was converted to factors in R. A factor of 1 is a Fail and a factor of 2 is Pass. For some applications, the numeric versions of Pass/Fail will be used.

Initial Analysis

- A summary of the data is below. The pass/fail column was read from the raw data, so I ignored that one and summarize it separately. We do see that there is 1 NA value in each column. We need to see where they are and remove them if they are problematic. After further analysis, there are no NA values in the data. They each occurred *after* the last row in each column. To cause R not to find NA values, I used the statement
  data = na.omit(data).

```
> summary(data)
  Study.Hours     Previous.Exam.Score   Pass.Fail
 Min.   :1.046   Min.   :40.28       Min.   :0.000
 1st Qu.:3.172   1st Qu.:53.75       1st Qu.:0.000
 Median :5.618   Median :68.31       Median :0.000
 Mean   :5.487   Mean   :68.92       Mean   :0.368
 3rd Qu.:7.805   3rd Qu.:83.58       3rd Qu.:1.000
 Max.   :9.937   Max.   :99.98       Max.   :1.000
 NA's   :1       NA's   :1           NA's   :1

> summary(pf)
   0    1 NA's
 316  184    1
```

- Plotting study hours vs. previous exam score produced a scatterplot that showed no relation between the two variables. They are, in fact, metrics from two different exams, so this may explain the lack of relationship. For example, if the two exams are on different topics, the amount needed to study for one may be completely different than the amount needed for the other.



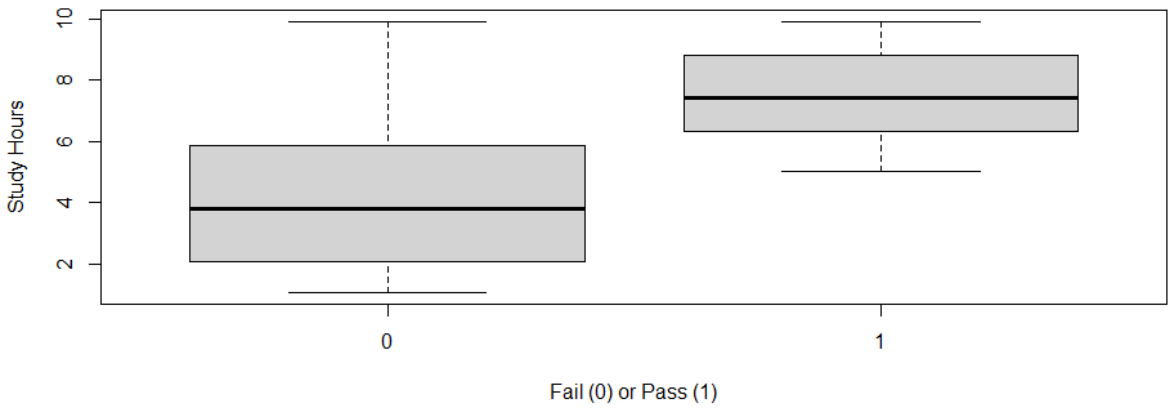
- A scatterplot of Study Hours for New Exam vs. Previous Exam Score, with students passing new exam in red. There is really no correlation between the two numeric variables, but we can see from the plot that students with both a high previous exam score AND around 6 or more hours of study, passed the new exam.

- The summary of study hours vs. previous exam score quantifies the extreme lack of correlation between the two predictors. Previous exam score is not important in predicting study hours for the next exam, in this case, because the p-value is very high and the r-squared value is very low.

```
Call:
lm(formula = Study.Hours ~ Previous.Exam.Score, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4537 -2.3287  0.1423  2.3306  4.4711

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.375144   0.499323  10.765   <2e-16 ***
Previous.Exam.Score 0.001623   0.007032   0.231    0.818
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.691 on 498 degrees of freedom
Multiple R-squared:  0.0001069, Adjusted R-squared:  -0.001901
F-statistic: 0.05326 on 1 and 498 DF,  p-value: 0.8176
```

Adding in the Pass/Fail categorical variable:

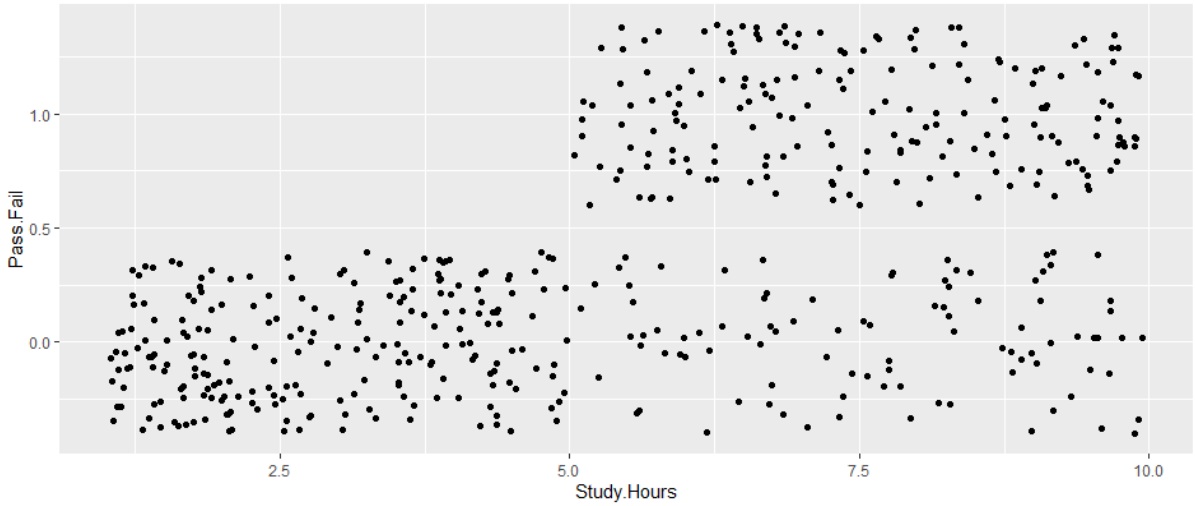
- Plotting study hours vs. pass/fail for the *same* exam, we see that the middle 50% of students who passed all studied between 6 and 9 hours, while the middle 50% of students who failed studied under 6 hours. Of the students that passed the current exam, the least number of hours they studied is around 5.5. Of the students that failed the current exam, 75% of them studied less than 6 hours.



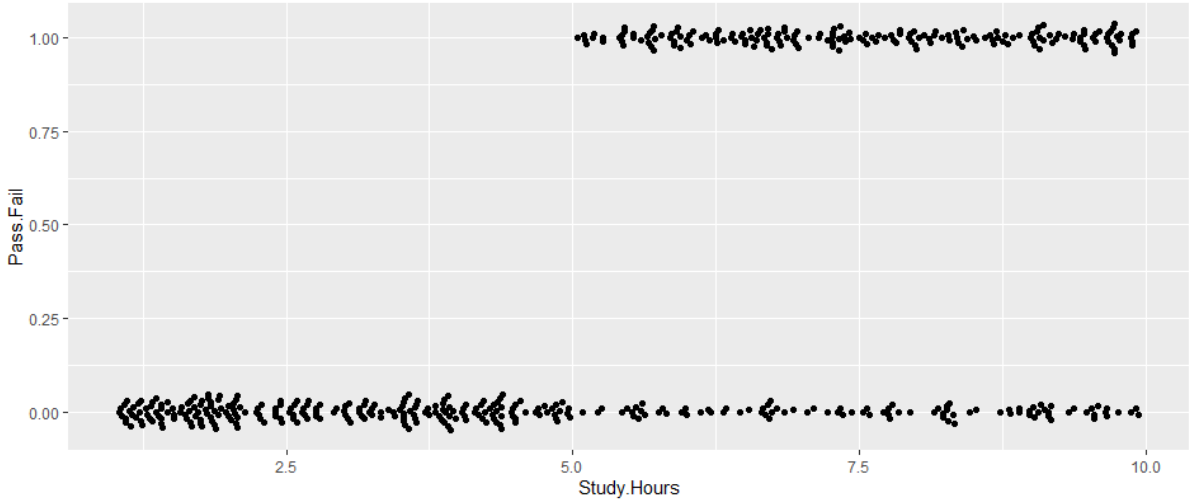
- Plotting previous exam score vs. pass/fail for the new exam, we see that of the students who passed the 2nd exam, they all had scores of 60 or higher on the previous exam. Only about half of the students who failed the 2nd exam had scores above 60 on the previous exam. There was more variation in the previous exam scores for students who failed the 2nd exam, and all of the students who passed the 2nd exam scored better on the previous exam than the lower 50% of students who failed the 2nd exam.

Some other interesting visualizations using Pass/Fail as numeric:

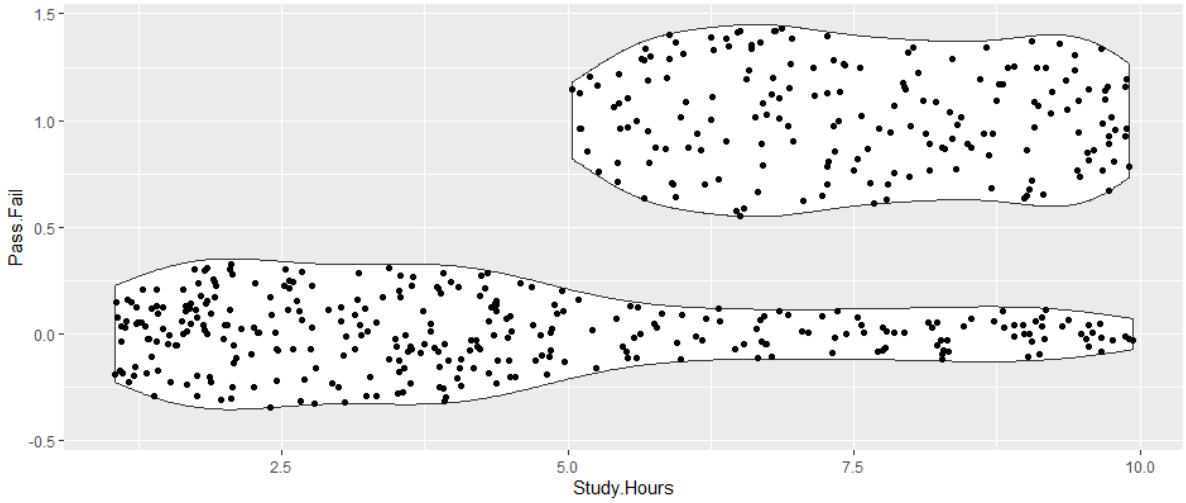
A clunky jitter version, interesting but could be improved.



A lined up beeswarm version (*flaring can be more or less with changes to cex argument*)



A version using the ggplot library methods of geom_sina() and geom_violin(), where geom_violin() provides the outline around the points:



Logistics Regression (Pass/Fail as factors)

```
Call:
glm(formula = Pass.Fail ~ Study.Hours + Previous.Exam.Score,
    family = binomial, data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.10292  -0.35071  -0.05671   0.26107   2.47196

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -17.75847    1.72325 -10.305   <2e-16 ***
Study.Hours         1.14573    0.11387  10.062   <2e-16 ***
Previous.Exam.Score   0.14312    0.01521   9.412   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 657.88  on 499  degrees of freedom
Residual deviance: 269.13  on 497  degrees of freedom
AIC: 275.13

Number of Fisher Scoring iterations: 7
```

Both study hours and previous exam score have very low p-values, so both are important. Take note of the AIC value of 275.13. We need that to be the lowest it can be. Let’s see what happens when one or the other predictor is removed.

Using only study hours, the AIC value is 465.6, and using only previous exam score, it is 557.04. Using the 2 predictors together is best.

The predict function can be used to predict the probability that the student will pass or fail, given the values of the predictors.

The probabilities for the first 10 records are:

```
           1           2           3           4           5           6           7           8
0.2628913984 0.9712127313 0.3357572343 0.9066115254 0.0323074757 0.0003768773 0.0781932723 0.9939794381
           9          10
0.9697076282 0.9336664990
```

These correspond nicely to pass/fail realities:

```
   Study.Hours Previous.Exam.Score Pass.Fail
1     4.370861             81.88970         0
2     9.556429             72.16578         1
3     7.587945             58.57166         0
4     6.387926             88.82770         1
5     2.404168             81.08387         0
6     2.403951             49.75702         0
7     1.522753             94.65563         0
8     8.795585             89.35223         1
9     6.410000             96.98799         1
10    7.372653             83.54000         1
```

Now to convert these predicted probabilities into class labels and see how accurate our model was, with a confusion matrix.

```
         Pass.Fail
glm.pred   0   1
    Fail 284  34
    Pass  32 150
```

The logistics model accurately predicted passing grades 150/184 times, or 81.5% of the time, and failing grades 284/316 times, or 89.9% of the time. Overall it was (284+150)/500, or 86.8% accurate. The training error rate is 13.2%. This most likely underestimates the training error rate, since the model was trained and tested on all the data. Now we'll better assess the model by holding out some testing data from the training set.

There are 500 observations, not sorted in any way, so it should be fine to use the first 400 for training and the last 100 for testing. In doing so, we get the following results.

```
         Pass.Fail.test
glm.pred  0  1
    Fail 52  7
    Pass 11 30
```

The model predicted the testing data with (52+30)/100 = 82% accuracy. This means the training error rate was 18%. It predicted the failings at 82.5% and the passings at 81.1%. The model is within about 1% accuracy in predicting passing and failing grades.

Final Model

Predicted probability = -17.75847 + 1.14573(study hours) + 0.14312(previous exam score), where

"Pass" if predicted probability > 0.5,

"Fail" if predicted probability <= 0.5.

References

1. MrSimple07. (2024). Student Exam Performance Prediction [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/7400136