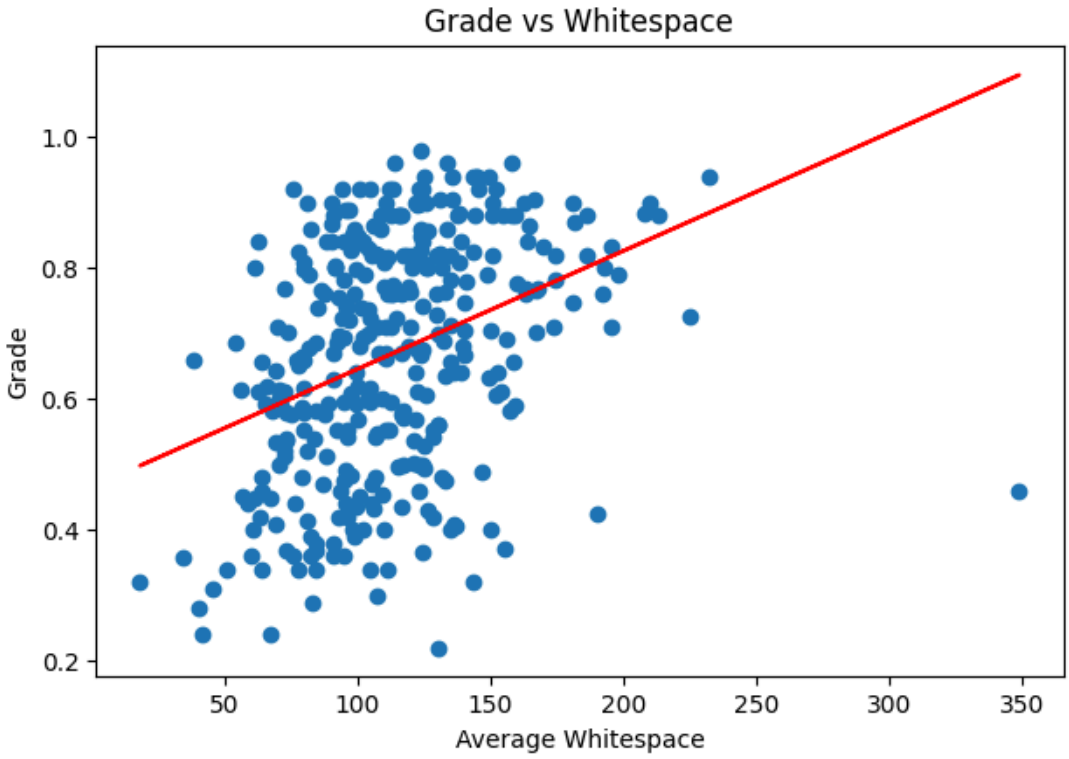
Educational Data Mining Report

Nathan Kuhn, Andrew Okerlund

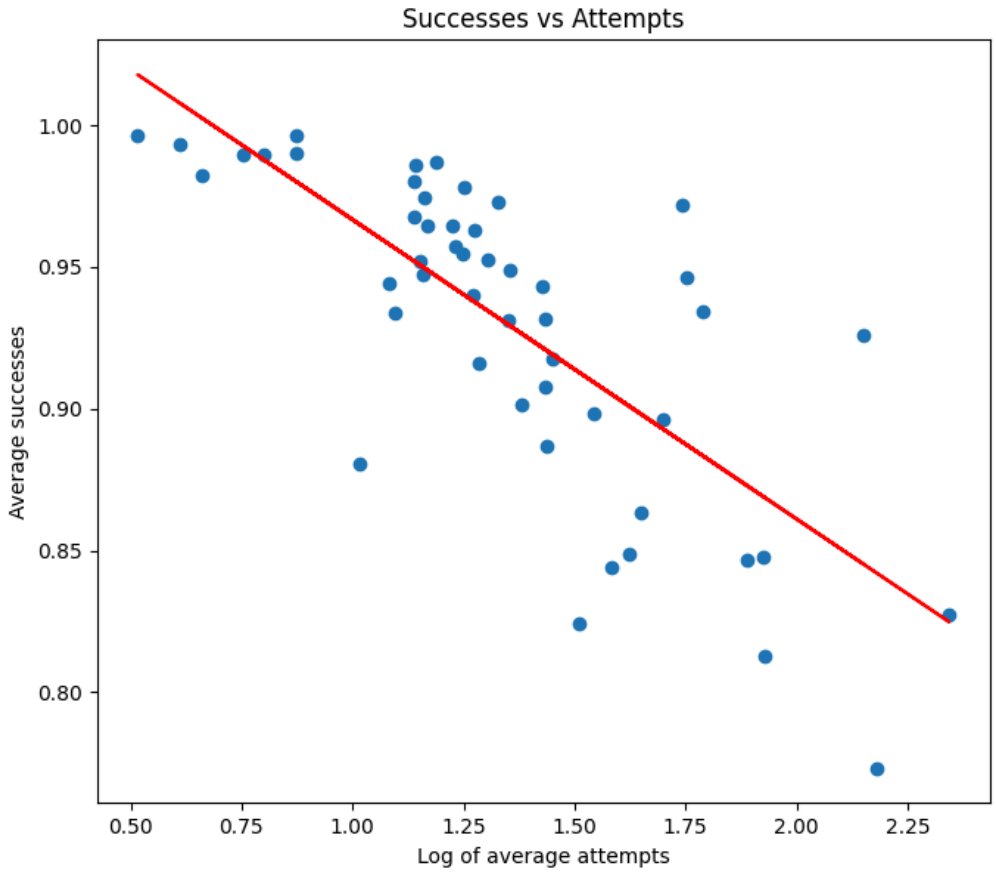
Whitespace Grade Correlation

One experiment we ran on the dataset was to see if there was any correlation between average whitespace in a student's code and grade. After processing the dataset and running a linear regression on the data, the resulting R2 value was 0.13 which indicates little to no correlation.



Problem Difficulty Correlation

To see if any correlation existed between problem success rate and number of attempts, a similar process was performed. The success rate of a problem may indicate difficulty and a linear regression model indicates a moderate correlation between success rate and log number of attempts. The model has an R2 value of 0.56, a slope of -0.11, and an intercept of 1.1. This model expresses how as number of attempts increases, success rate decreases:



Grade to Attempts Classifier

Another method used on the dataset was to try and find a relationship between early and late problem attempts and the grade of a student. A linear classifier was created to determine if a student would perform above the mean grade using four statistics:

- Early mean attempts per problem
- Early problem success rate
- Late mean attempts per problem
- Late problem success rate

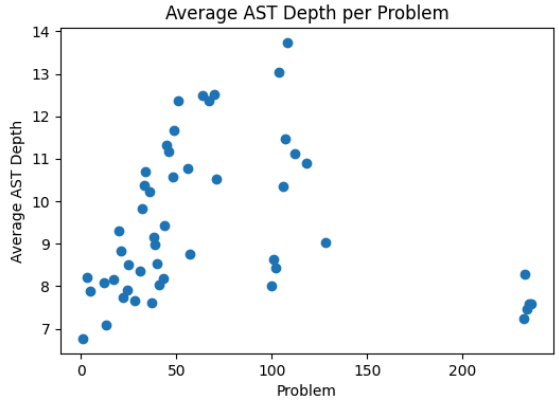
The resulting model had a test set accuracy of 78% and had the following coefficients:

Early attempts per problem	-0.02
Early success rate	1.18
Late attempts per problem	0.40
Late success rate	0.46

These coefficients show a high influence from early success rate and a medium influence from late average attempts and late success rate on if a student will receive an above-average grade.

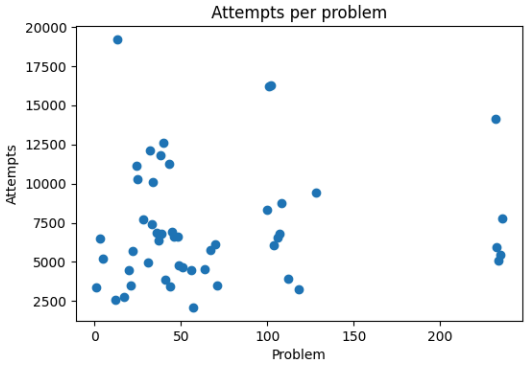
Abstract Syntax Tree Correlation

Looking at the Abstract syntax trees (ASTs) from student's code states had the allure of a deeper investigation. In this investigation the average AST depth was calculated in relationship to average number of attempts per problem, grade and students average AST depth and average AST depth per problem.

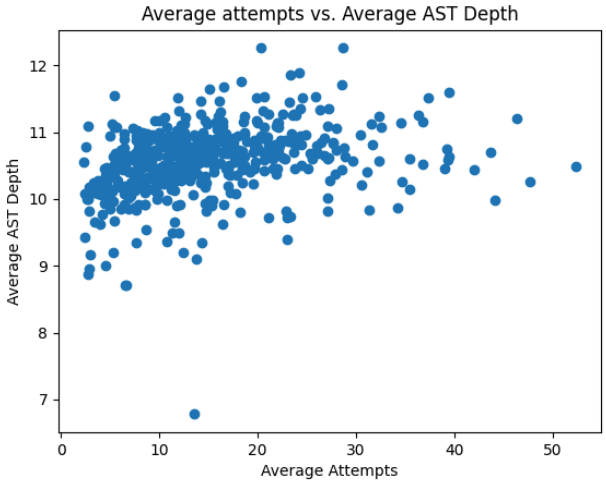



```
R^2: 0.0021382571874198897
Intercept: 9.549433711982228
Slope: [-0.0012882]
```

This plot is an output of Average AST depth per problem.

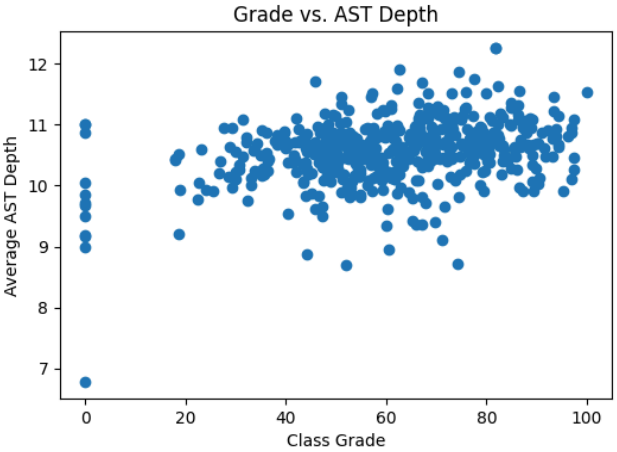


This plot shows the number of attempts per problem. Note that the scatter plot distribution looks similar to the average AST depth for problem plots. This shows correlation between the difficulty of the problem and a student's AST.




```
R^2: 0.10879623889923096
Intercept: 10.03925609588623
Slope: [0.0086454]
```

This plot shows average AST depth per problem over average number of attempts per problem. After processing the dataset with linear regression it is clear that there is no correlation between average number of attempts and average depth of AST. This is supported by the R^2 value of .108. The hypothesis was that a more complex AST could indicate more complex and therefore harder to read code, which could lead to more attempts on a given problem.




```
R^2: 0.09745872020721436
Intercept: 10.282143592834473
Slope: [0.01899639]
```

This plot shows average AST depth over class grade per student. After processing the data with linear regression an R^2 value of .09 was produced. This once again shows no

correlation between average AST depth and class grade. The hypothesis with this was that students who write code of a certain complexity would have a better grade in the course.

Based on the relationships investigated between average AST depth per problem and other attributes the only correlative finding is between AST depth and number of attempts. This indicates that more challenging problems, which required more attempts also often required a more complex AST to solve the problem.