Professor Moody

•••

By: Rohit Manjunath

Breaking the Data

- To analyze the dataset, and make predictions based on the analysis we first need to break the data into 2 parts: test data and train data.

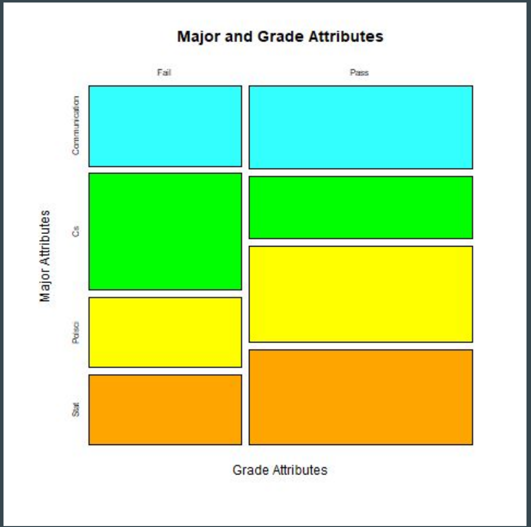
- Code:

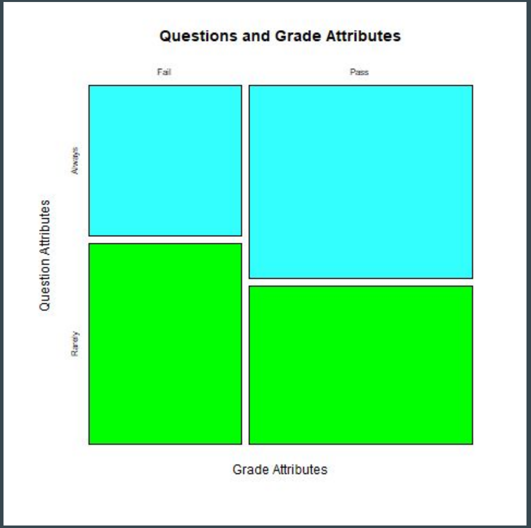
```
dt = sort(sample(nrow(moody), nrow(moody)*.8))

train <- moody[dt,]#7571 rows

test <- moody[-dt,]#1893 rows
```

- As you can see in the above code, 80% of the data is split into train data and the rest is test data. Now we can analyze the train data by each major and scores.

Analyzing the Train Data | Part 1

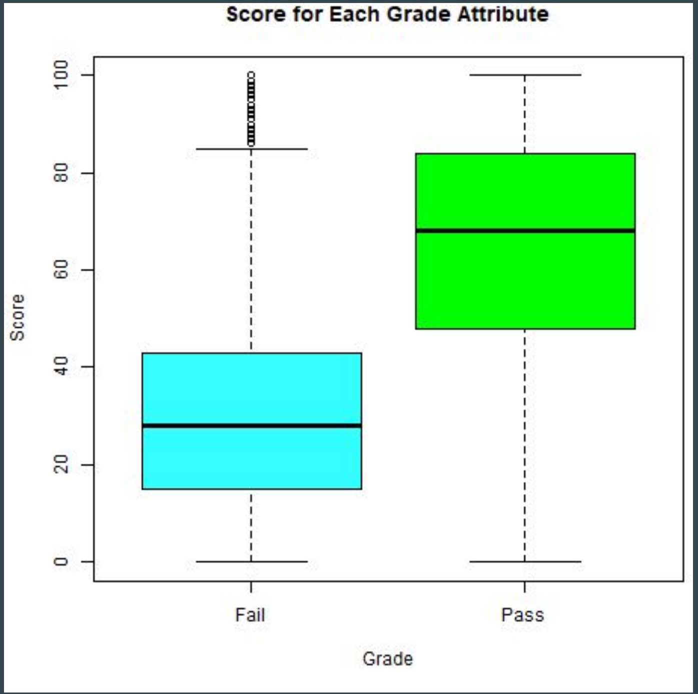
- Before we further split the data based on student's majors, we can make plots on the train data to understand the overall data better.
- More Cs students fail than pass. So it's safe to assume that the parameters involved in deciding if the student failed or passed is much more 'stricter' than other majors.
- More Polsci and stat students pass than fail. So it's safe to assume that these classes are generally less 'strict' while evaluating the students attributes.
- Generally students that always ask question have a higher chance of passing. This may vary for different majors.





Analyzing the Train Data | Part 2

- We can see a boxplot of the scores for students that passed and failed.
- There are two critical numbers we can take from this. The lower quartile and median.
- Students with a score below 50(the lower quartile), are very likely to fail. This varies with each major.
- Students with a score above 70(the median), are very likely to pass. This too varies with each major.
- Students with a score between 50 and 70 are influenced by other parameters the most.

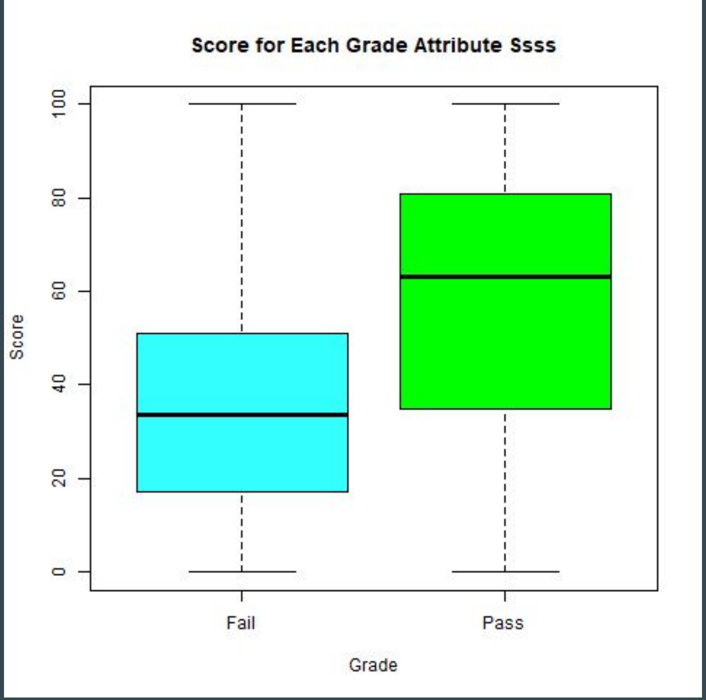


Analyzing the Segments - Statistics

- You can see in the boxplot on the right that most students that score above 40 pass.

- Not much can be said about the other graphs that contain grades of only students from statistics.

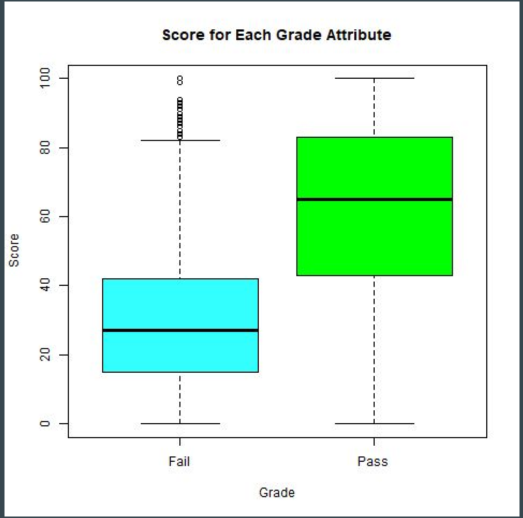
- On further evaluation you can see there is nothing special about this major in terms of how professor moody decides to grade them. It looks like the statistics major follows the general trend.



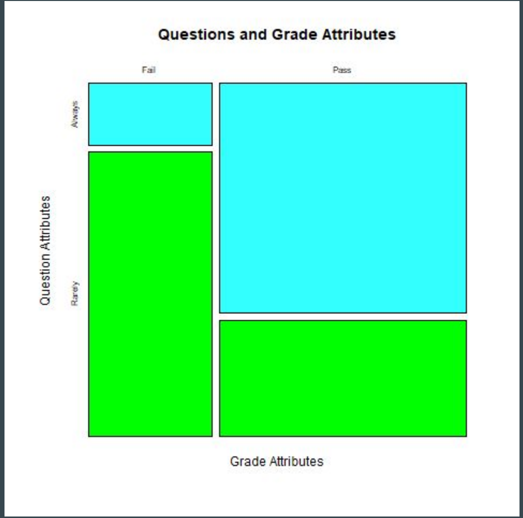
Analyzing the Segments - Political Science

- The boxplot on the right tells us more students pass when their score is more than 42.

- As you can see, the mosaic graph on the right tell us that students that ask questions always are more likely to pass than fail. This major agrees with the general trend(Overall data analysis).



Score for Each Grade Attribute

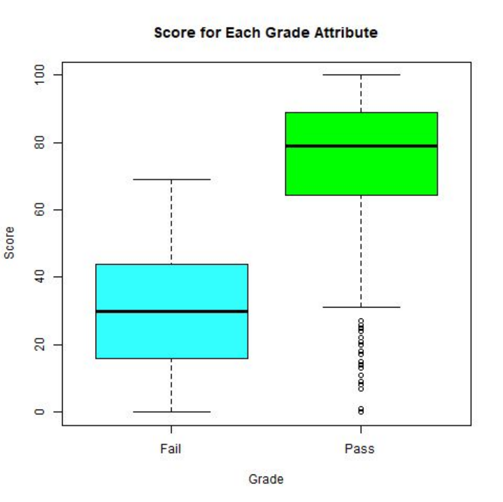


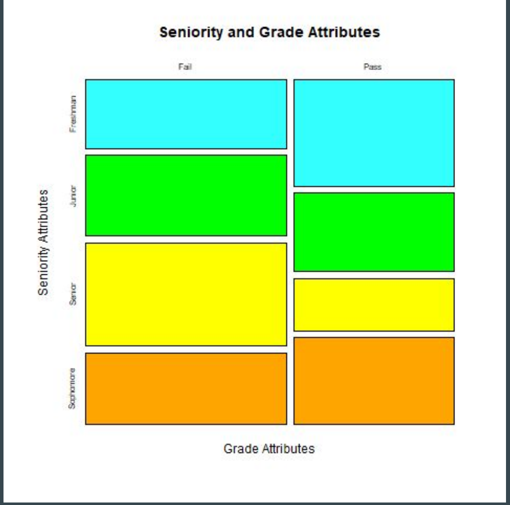
Questions and Grade Attributes

Analyzing the Segments - Computer Science

- As you can see in the boxplot in the right, students that score more than 70 generally pass. This tells us that you need a higher score than other fields to have a higher chance to pass. This also tells us any student that is in computer science and got a score below 50(lower quartile) will most likely fail.

- Freshman and sophomores have a higher chance to pass than juniors and seniors.

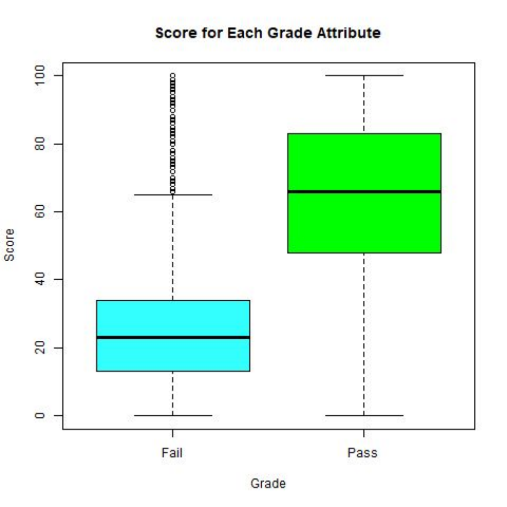


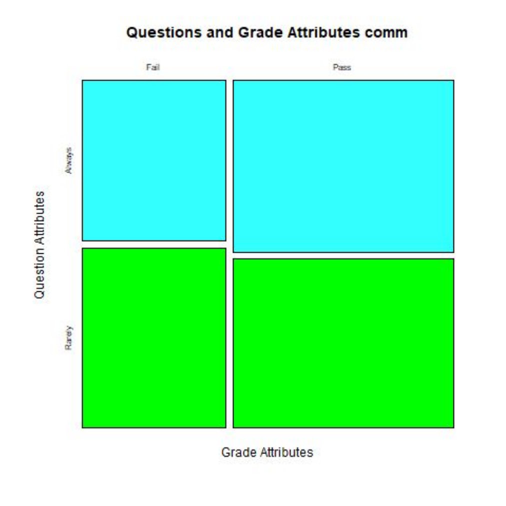


Analyzing the Segments - Communication

- A student in communication with a score greater than 42 has a good chance of passing.

- The mosaic graph on the right tells us that students that always ask questions have a greater chance of passing than when students that rarely text. This major kind of agrees with the general trend(Overall data analysis) too.





Summary of the Analysis

- If the score is below 50, and the student does major in computer science then the student most likely failed.
- Students that got below 50 and majored in courses that have higher passing rates such as polisci are more likely to pass when they always ask questions.
- Students that got below 50 and major in communications are more likely to pass if their score is greater than 42 and always ask questions.
- Students with a score greater than 70, most likely pass among all majors.
- Students with a score greater than 50 but less than 70 is also most likely passed by all majors except computer science.
- If the student is majoring in computer science and scores between 50 and 70, then we know freshmen and sophomores are more likely to pass.

Building the Model

- Using all the above data from the various plots we can finally create a model based on the provided parameters and analysis:

```
decision[test$Score >= 70] = "Pass"

decision[test$Score < 50 & test$Major != 'Cs' & test$Questions == 'Always' & test$Major != 'Communication'] = "Pass"

decision[test$Score < 50 & test$Questions == 'Always' & test$Major == 'Communication' & test$Score > 42] = "Pass"

decision[test$Score >= 50 & test$Score < 70 & test$Major != "Cs"] = "Pass"

decision[test$Score >= 50 & test$Score < 70 & test$Major == "Cs" & test$Seniority != "Junior" & test$Seniority != "Senior"] = "Pass"
```

- The variable decision is a vector that has the same length as the ‘test’ dataset. It has default values of “Fail’. The above model changes it to ‘Pass’ if the conditions are met.

Prediction Accuracy

- Code:

  ```
  error <- mean(Predict$Grade != Predict$GradePredict)
  error
  ```

- “error” rate gives me an average of .15 or 15%.

Thank you!