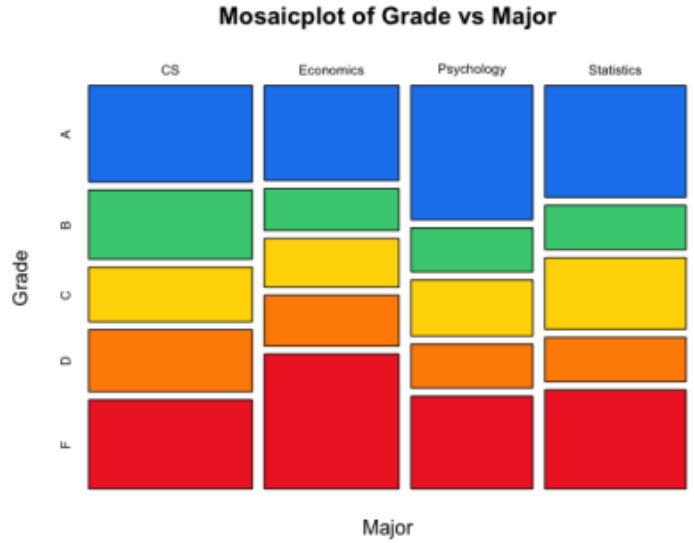
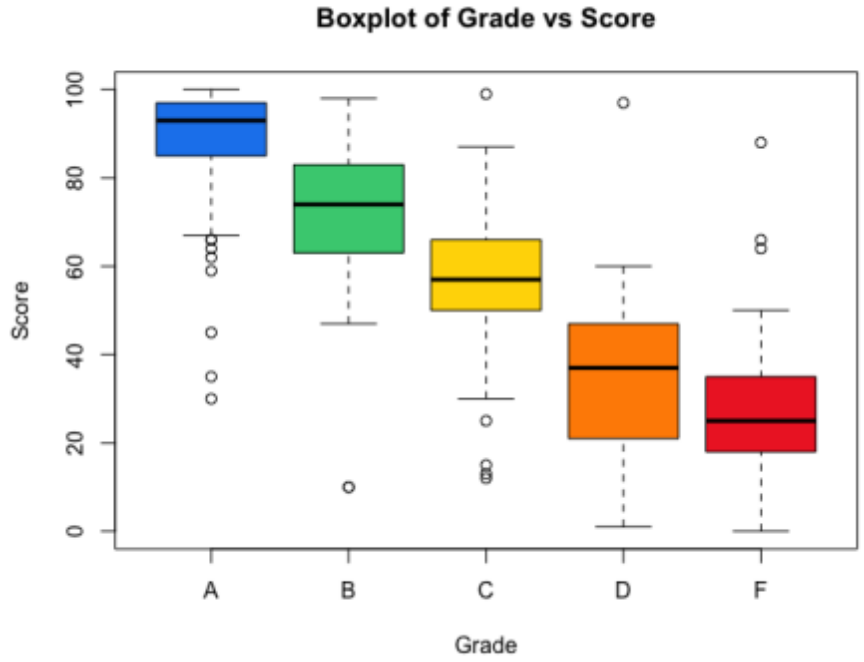
Prediction Challenge 1

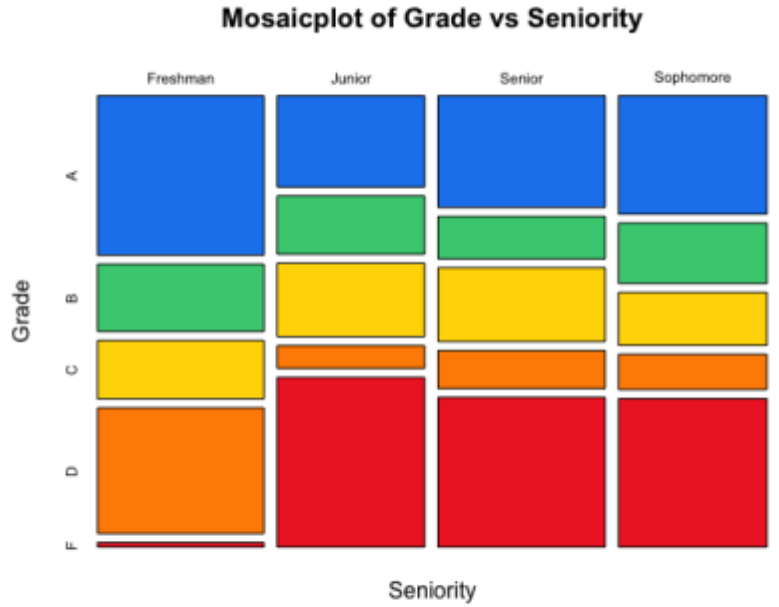
Jeevanandan Ramasamy

Initial Analysis of Data

We cannot make any conclusions based on these plots







Thought Process

- Before making assumptions about the distributions of scores for each category, I opened the training data in a spreadsheet format
- After sorting the data by Score, Grade, Major, and Seniority, I made a list of ranges of values, including potential outliers
- I noticed that each major has a different distribution of grades and freshmen have different distributions from the rest of the seniority groups

Cleaning Data

- Then, I removed outliers that seemed too far away from the range of values for each category
- I decided to keep some outliers that do not fit in the ranges of other categories
- Afterward, I combined the data for all other seniorities other than freshman into one category (‘Other’) due to their similarities

Prediction Model

- I made a few assumptions to make my prediction model due to lack of data for certain aspects
  - The ranges I made had gaps, so I had to extend them in order to cover all score values for each category
  - There were only 2 Freshmen with F's in the entire dataset and there were no patterns, so I assumed that Freshmen generally do not receive F's in Moody's class
- After applying my prediction model to the training dataset, I got an error of 0.04285714, which is about 4.3%

Final Results

- After my initial submission to Kaggle, I received an error of 3.3% based on 10% the of test data
- After all the test data was released, my error became 8.15%
- I got first place, which I am really excited for!