AP Grades Report

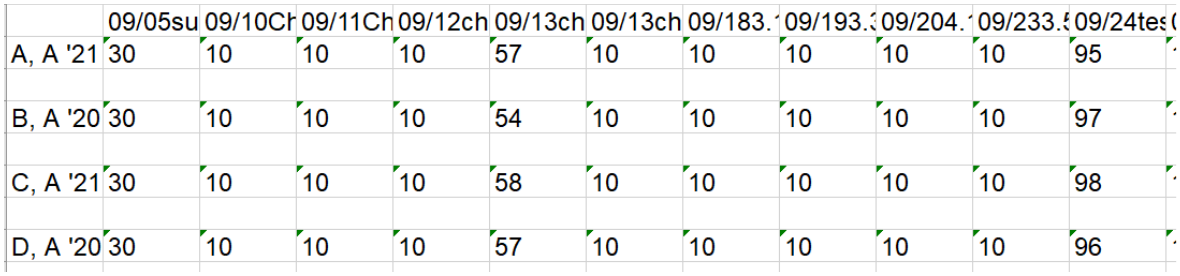
Introduction

The goal of this project was to use grades data from my AP classes to predict scores on the AP Exam. AP Exams are given to high school students who wish to possibly get college credit in a course. The exams are graded on a 1 - 5 scale, 1 being the worst, 5 being the best. "Passing" an AP exam and possibly getting college credit has traditionally been seen as getting at least a 3 on the exam. Having been a mathematics teachers for the past 5 years I have accumulated a decent amount of data from year to year in my courses. I have data within AP Calculus BC and AP Statistics. For the purposes of this project I will only be using AP Calculus BC data and will be cleaning, visualizing, and doing predictive modeling. These processes should be the exam same for the AP Statistic data as well.

Data Collection:

Grades data was collected from a learning management system, all names were randomized and anonomized. Ordered pairs of initials are used to differentiate from student to student. The order pairs have no relationship to students actual initials unless by chance.

Initial data looked in this form:



	09/05su	09/10Ch	09/11Ch	09/12ch	09/13ch	09/13ch	09/183.1	09/193.3	09/204.1	09/233.5	09/24tes
A, A '21	30	10	10	10	57	10	10	10	10	10	95
B, A '20	30	10	10	10	54	10	10	10	10	10	97
C, A '21	30	10	10	10	58	10	10	10	10	10	98
D, A '20	30	10	10	10	57	10	10	10	10	10	96

This snippet of data is for one quarter of the school year for one AP Calculus BC class. There are 3 main categories for grades:

- Tests - Worth 65% of cummulative grade
- Quizzes - Worth 25% of cummulative grade
- Homework - Worth 10% of cummulative grade

All data file labels have the following layout:

YEAR_class#_Q#

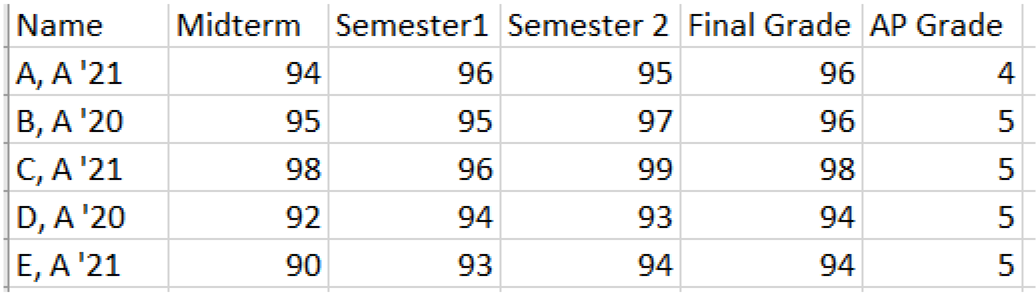
- YEAR - the year in which the class began
- class# - The class label and section number - ex: "BC1"
- Q# - The quarter in which the data is for: Ranging from Q1 to Q4

Data Cleaning

Cleaning of the data would go as follows:

- Remove homework grades entirely - Homework would generally average to a 100%, making it a poor indicator for any predictions
- Calculate average quiz and test grades from a given quarter
- Using the name column extract students graduation year - Always the last 2 digits of their name
- Using graduation year along with current school year (found in filename) find the class standing of a student (i.e Freshman, Sophomore, Junior, or Senior)

In addition to raw grades data I compiled another data set for each class consisting of some grades not found in the learning management system. Below is what this data looks like:



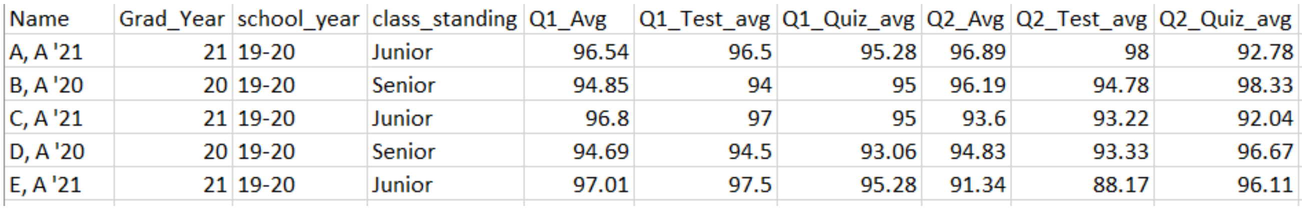
Name	Midterm	Semester1	Semester 2	Final Grade	AP Grade
A, A '21	94	96	95	96	4
B, A '20	95	95	97	96	5
C, A '21	98	96	99	98	5
D, A '20	92	94	93	94	5
E, A '21	90	93	94	94	5

"AP Grade" is the variable we are trying to predict. The Semester grades are calculated based on a students quarter averages

Compile all transformations accross all years, sections, and quarters to get our cleaned data.

The data is now cleaned and all compiled into one file, "ALL_BC_cleaned.csv"

Here is a snippit of the data for quarters 1 and 2:



Name	Grad_Year	school_year	class_standing	Q1_Avg	Q1_Test_avg	Q1_Quiz_avg	Q2_Avg	Q2_Test_avg	Q2_Quiz_avg
A, A '21	21	19-20	Junior	96.54	96.5	95.28	96.89	98	92.78
B, A '20	20	19-20	Senior	94.85	94	95	96.19	94.78	98.33
C, A '21	21	19-20	Junior	96.8	97	95	93.6	93.22	92.04
D, A '20	20	19-20	Senior	94.69	94.5	93.06	94.83	93.33	96.67
E, A '21	21	19-20	Junior	97.01	97.5	95.28	91.34	88.17	96.11

Note: our cleaned data has 79 responses and 21 variables in total

Data Exploration

For data visualizations and prediction of AP grades I decided to only use the Quarter averages (Q1_Avg, Q2_Avg, Q3_Avg, Q4_Avg), semester grades (Semester1, Semester 2), midterm grade, and final grade. These grades are made up of the test and quiz averages found in the cleaning step. I removed any student who does not have an AP Grade listed. This leaves use with 70 responses and 9 variables for prediction. "class_standing" is also kept for visualization purposes.

Below is a snippet of the data selected along with the summary statistics for the quantitative variables

	class_standing	Q1_Avg	Q2_Avg	Midterm	Q3_Avg	Q4_Avg	Semester1	Semester 2	Final Grade	AP Grade
0	Junior	96.54	96.89	94	89.67	98.56	96.0	95.0	96.0	4.0
1	Senior	94.85	96.19	95	93.08	97.92	95.0	97.0	96.0	5.0
2	Junior	96.80	93.60	98	89.65	88.54	96.0	99.0	98.0	5.0
3	Senior	94.69	94.83	92	89.56	95.50	94.0	93.0	94.0	5.0
4	Junior	97.01	91.34	90	89.93	98.47	93.0	94.0	94.0	5.0

	Q1_Avg	Q2_Avg	Midterm	Q3_Avg	Q4_Avg	Semester1	Semester 2	Final Grade	AP Grade
count	70.00	70.00	70.00	70.00	70.00	70.00	70.00	70.00	70.00
mean	91.67	91.93	88.11	86.36	94.56	91.13	90.81	91.18	3.66
std	6.34	6.20	8.95	10.35	5.83	6.29	7.39	6.57	1.31
min	72.26	72.04	60.00	48.54	70.62	73.80	70.00	72.00	1.00
25%	89.81	89.36	83.00	83.13	91.72	88.65	88.88	90.06	3.00
50%	93.65	93.64	90.00	89.66	96.46	92.80	93.75	93.20	4.00
75%	96.02	96.19	94.75	93.05	98.54	95.85	96.00	96.00	5.00
max	99.57	101.22	100.00	99.50	100.62	99.20	100.00	100.00	5.00

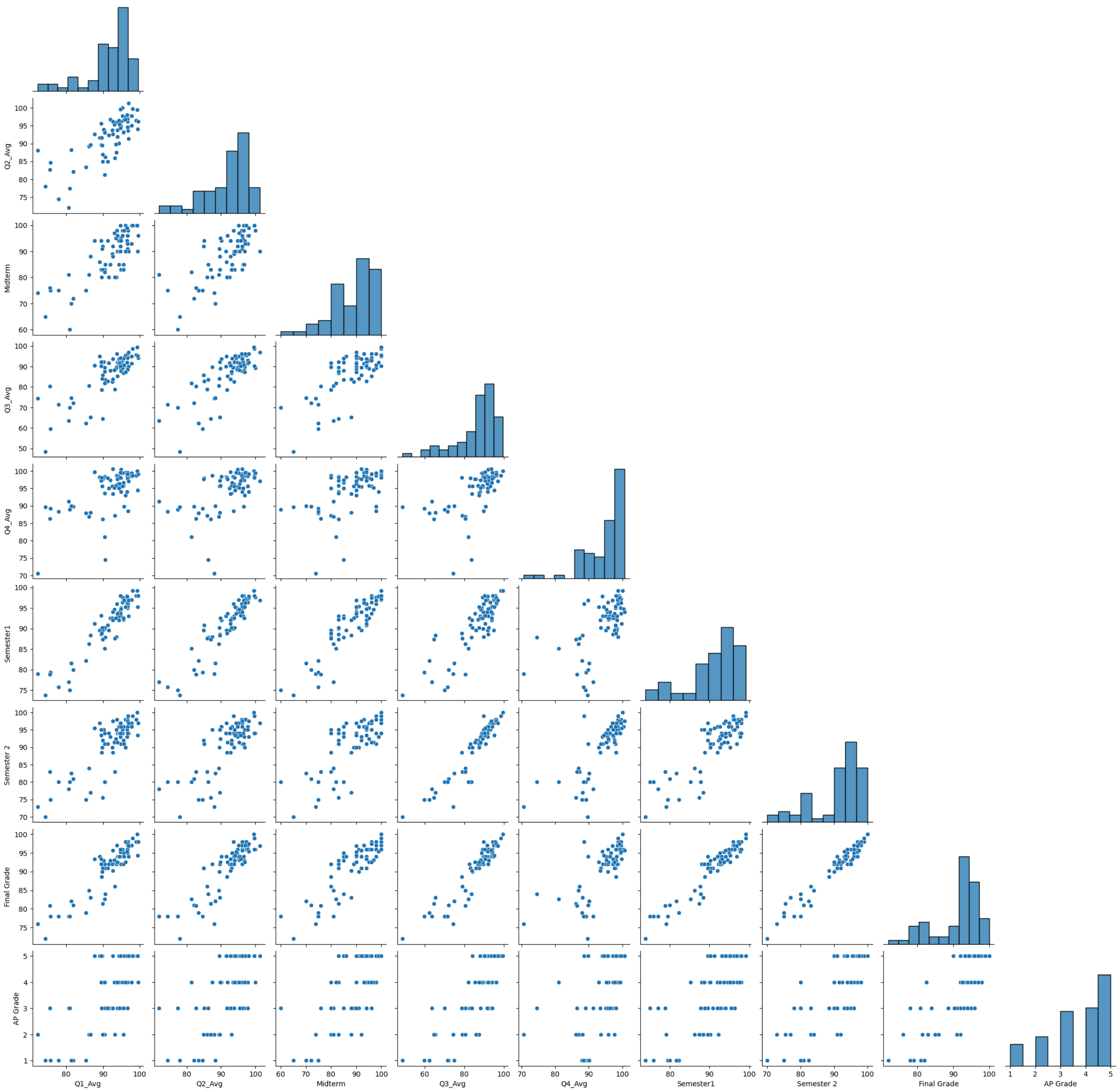
Next, a quick view at the relationships between the variables having scatterplots for each pair of quantitative variables and histograms along the main diagonal. Below these graphs I've also included the correlation coefficients for each pairing.

We see that when looking at grades that I give (everything but the AP Grade) there is generally a strong positive linear relationship with most correlations being above 0.7. Students who do well in one area of the year tend to do well in all the others.

The relationships between AP Grades (i.e. the last row below) do show some positive linear relationships that are only slightly lower than others.

The variable I find most interesting here would be Q4_Avg. We see correlations dip as low as 0.53. In the world of AP classes quarter 4 is generally the easiest quarter. Half of the quarter is devoted to review for the AP Exam and the other half takes place after the AP Exam is completed. Generally this leads to higher grades (94.6% from the table above) even for those students who struggled the entire year.

```
<seaborn.axisgrid.PairGrid at 0x2ac151cd6d0>
```

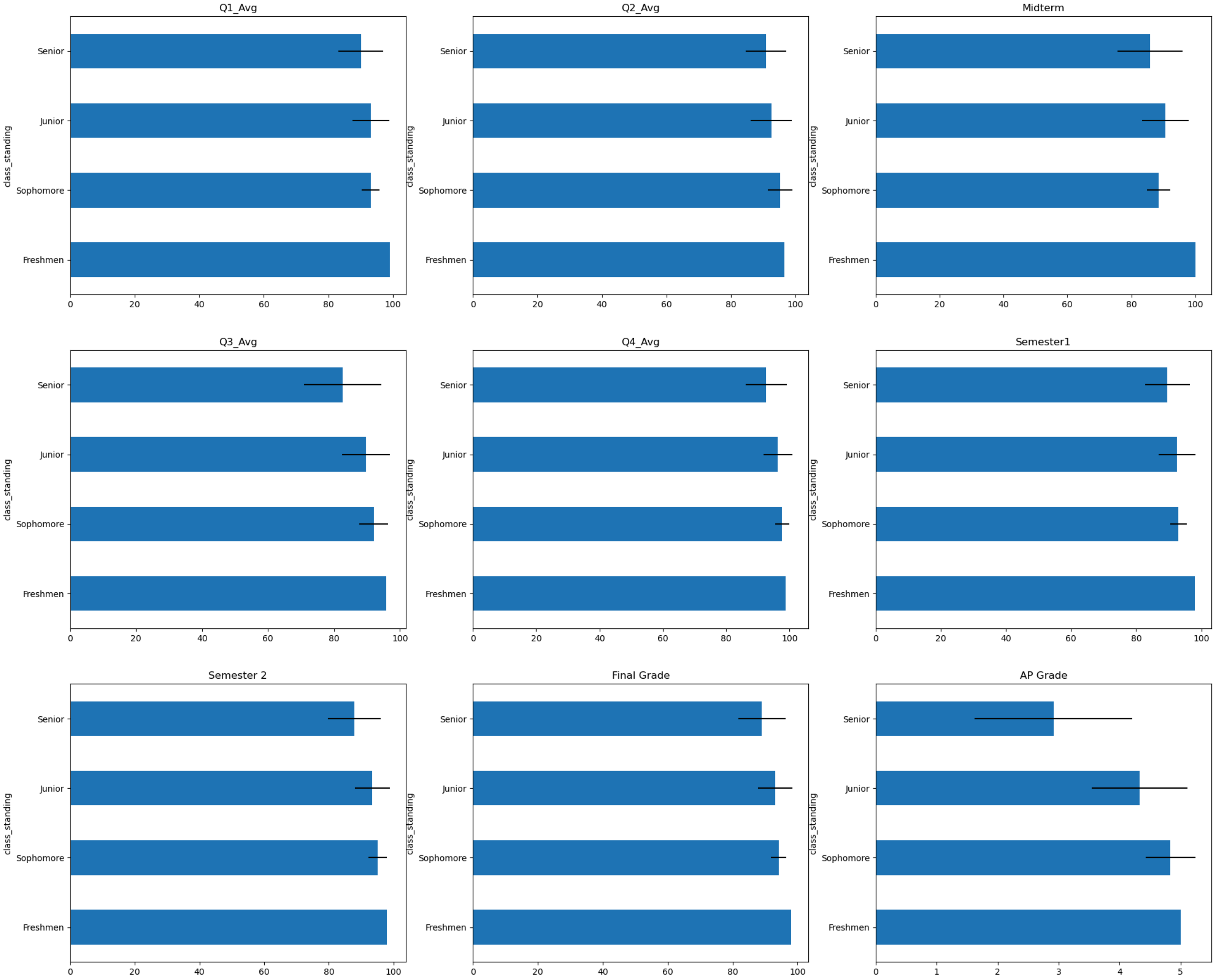

Q2_Avg
100
95
90
85
80
75
Midterm
100
90
80
70
60
Q3_Avg
100
90
80
70
60
50
Q4_Avg
100
95
90
85
80
75
70
Semester1
100
95
90
85
80
75
Semester 2
100
95
90
85
80
75
70
Final Grade
100
95
90
85
80
75
AP Grade
5
4
3
2
1
80
90
100
Q1_Avg
80
90
100
Q2_Avg
60
70
80
90
100
Midterm
60
80
100
Q3_Avg
70
80
90
100
Q4_Avg
80
90
100
Semester1
70
80
90
100
Semester 2
80
90
100
Final Grade
1
2
3
4
5
AP Grade

	Q1_Avg	Q2_Avg	Midterm	Q3_Avg	Q4_Avg	Semester1	Semester 2	Final Grade	AP Grade
Q1_Avg									
Q2_Avg	0.78								
Midterm	0.78	0.73							
Q3_Avg	0.80	0.79	0.73						
Q4_Avg	0.62	0.56	0.56	0.61					
Semester1	0.93	0.92	0.90	0.84	0.63				
Semester 2	0.82	0.78	0.75	0.94	0.81	0.85			
Final Grade	0.90	0.87	0.85	0.93	0.76	0.95	0.97		
AP Grade	0.68	0.64	0.71	0.76	0.53	0.72	0.76	0.77	

Class Standing Visualizations

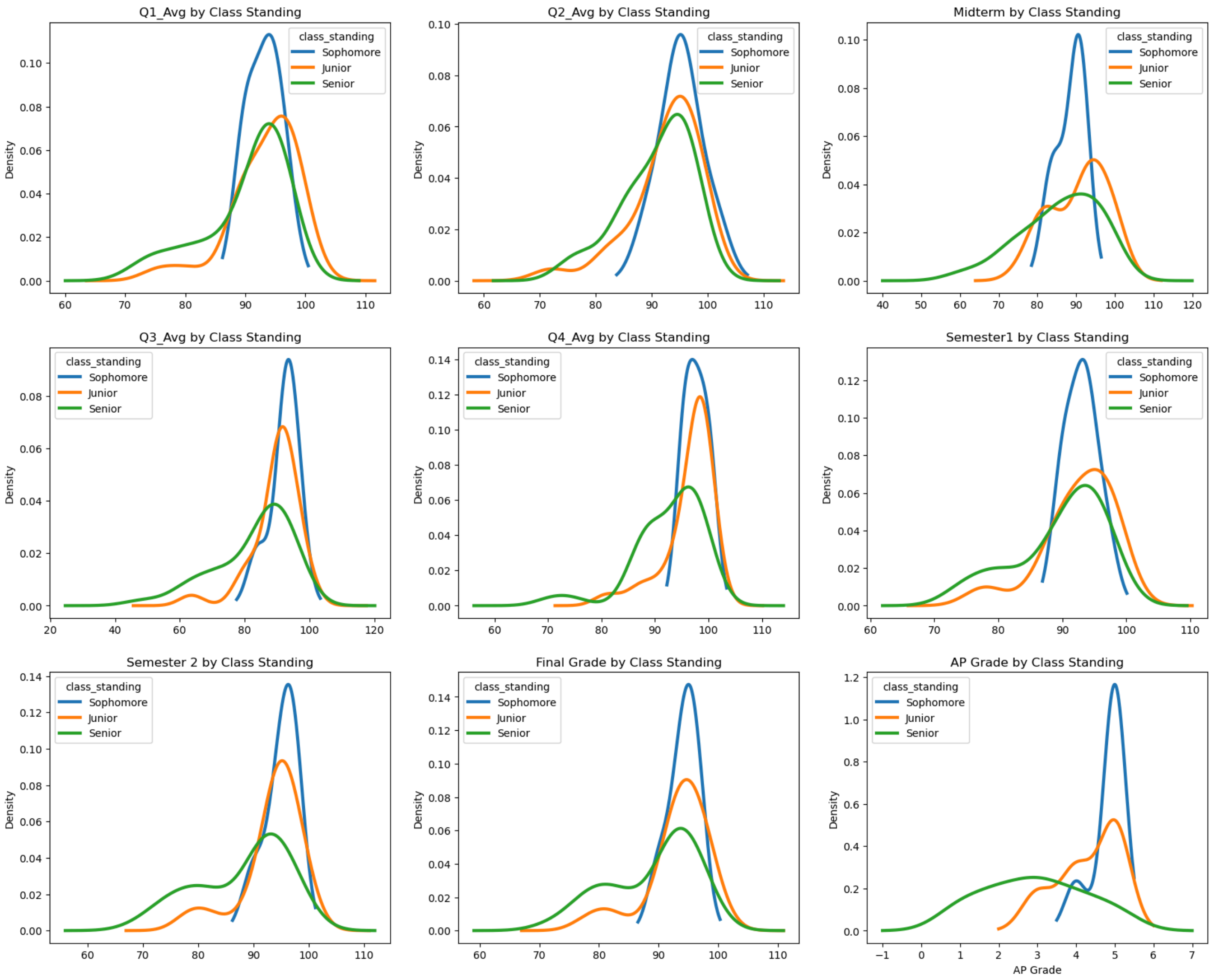
From here I wanted to look at grade distributions by class standing (freshman, sophomore, junior, senior). First via bar charts on the average grades for each class standing. Note that there was only 1 freshman in my data set so no error bar is shown.

We see that seniors generally are the lowest performers with the highest variance.



To go further into the relationship between class standing and grades I made density plots for all variables. The 1 freshman data point could not be included in this data because a density plot can't be constructed with only one data point.

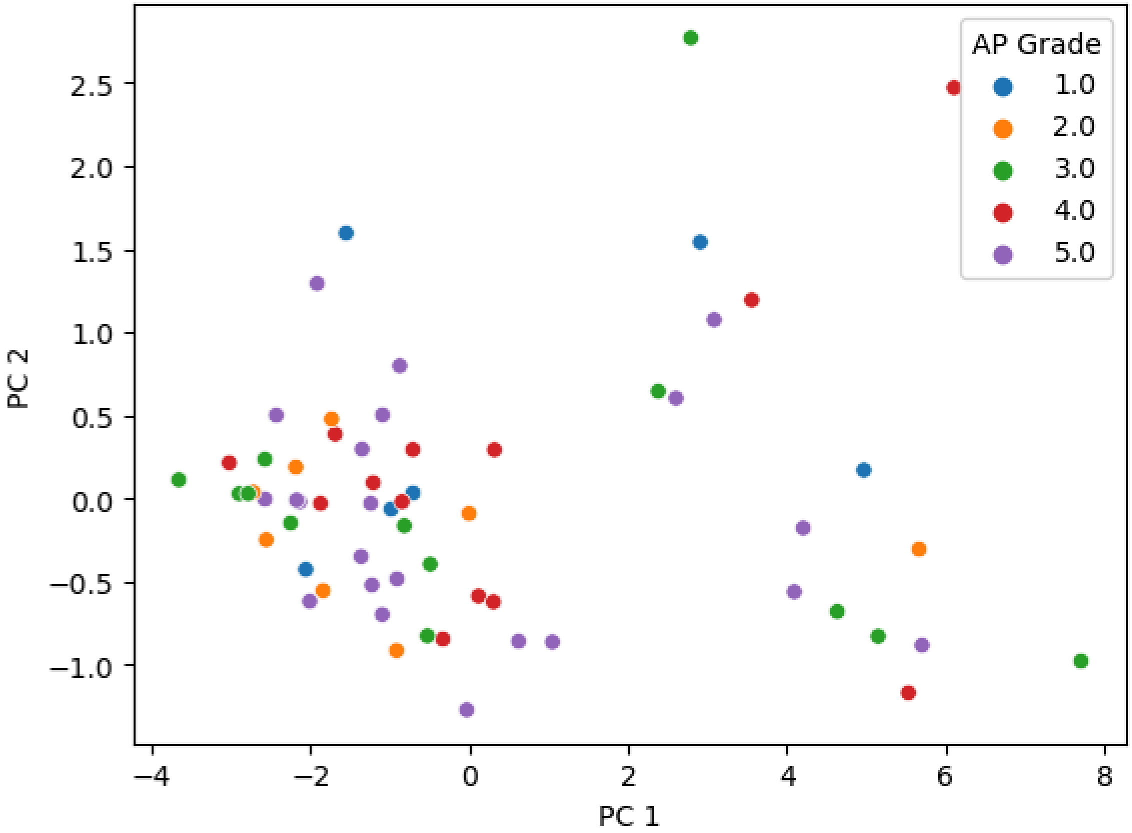
From these plots we can see that skewness tends to increase as class standing increases (i.e. students get older). With a student being able to make it to AP Calculus BC by sophomore or junior year is an impressive feat so these students are generally high achieving. Seniors on the other hand often fall victum to "senioritis", AKA they get lazy their final year or they are overworked from the college application process.



Finally, I wanted to analyze dimensional reduction (principal components) to see if any trend could be seen on AP Grade. I first standardized the data before running principal components. Below I graphed the first 2 principal components colored by AP Grade. The variance explained for the first 3 principal components is shown as well above the graph, we see that the first component takes into account nearly 81% of the variances. These grades data are all intertwined quite strongly!

From this visualization I can't make out any obvious trends as to a students AP grade and their component scores.

```
     Variance Explained
PC1               0.806
PC2               0.065
PC3               0.047
```



Prediction Modeling

These are the 6 models I will be using for predicting AP Grades

Multi-Class Classifcations:

- k-Nearest Neighbors
- Decision Trees
- Gaussian Naive Bayes
- Multi Naive Bayes
- Random Forest
- Gradient Boosting

For each method bootstrapping was used across 100 random train/test samples at an 80-20 split to get accuracy scores for each.

From the output below we see that Gaussian Naive Bayes was the most accurate model at 47.6%. Considering that there are 5 possible outcomes this is not too bad. For a comparison there are 25 students who recieved a 5 on the exam, this is about 36% of responses (25/70) so the models are all doing slightly better than if it were to just predict 5s for every student.

	Accuracy
KNN	0.401
Decision Tree	0.379
Gaussian Naive Bayes	0.476
Multi Naive Bayes	0.395
Random Forest	0.444
Gradient Boosting	0.445

Pass/Fail Predictions

A passing grade is considered to be at least a 3 on the AP Exam. Because of this we can make our model into a binary prediction instead of multiclass prediction. The variable AP Grade was replaced with the variable "PassFail" where 1 is pass and 0 is fail. Again bootstrapping is done 100 times at an 80-20 split for train/test. The 5 classification techniques I used are as follows:

Binary Classifications:

- Logistic Regression
- k-Nearest Neighbors
- Decision Trees
- Support Vector Machine
- Naive Bayes (Gaussian)

From the output below we see that Gaussian Naive Bayes again performed the best with 88.6% accuracy, quite strong. Looking at the ratio of passing students to failing there are 14 students who failed (20%) and 56 who passed (80%). This means apart from the decision tree model all processes have a slightly better accuracy than just pure guessing based off prior knowledge.

	Accuracy
KNN	0.840
Decision Tree	0.797
Gaussian Naive Bayes	0.886
Logistic Regression	0.832
Support Vector Machine	0.835