STAT 3010: Computer Application of Statistics SAS Project

Roman Alvarez
5/5/2023

Dataset: Exams Data

1. The analysis of the Exams dataset is important in better understanding the factors that potentially influence a student's test scores. This data is significant in finding correlations that may exist and providing solutions to improve the scores of students.

This data set was pulled from Kaggle and contains test score data from three different high schools in the United States. The variables included are gender, whether the student completed a test preparation course, whether the student receives free or reduced-price lunch, standardized math test score, and standardized reading test score.

2.

Table 1: Variable Description

Variable Label	Variable Description	General Type	Specific Type	Measurement Units
Gender	Gender of student	Categorical	Nominal	N/A
Test Preparation Course	Whether student has completed a test preparation course	Categorical	Nominal	N/A
Lunch	Whether student has free or reduced lunch	Categorical	Nominal	N/A
Math Score	Student Math score	Quantitative	Discrete	Points
Reading Score	Student Reading score	Quantitative	Discrete	Points

3.

a)

Table 2: Descriptive Statistics for Quantitative Variables

The MEANS Procedure

Variable	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Quartile Range
math_score	67.81	15.25	15.00	58.00	68.00	79.50	100.00	21.50
reading_score	70.38	14.11	25.00	61.00	70.50	80.00	100.00	19.00

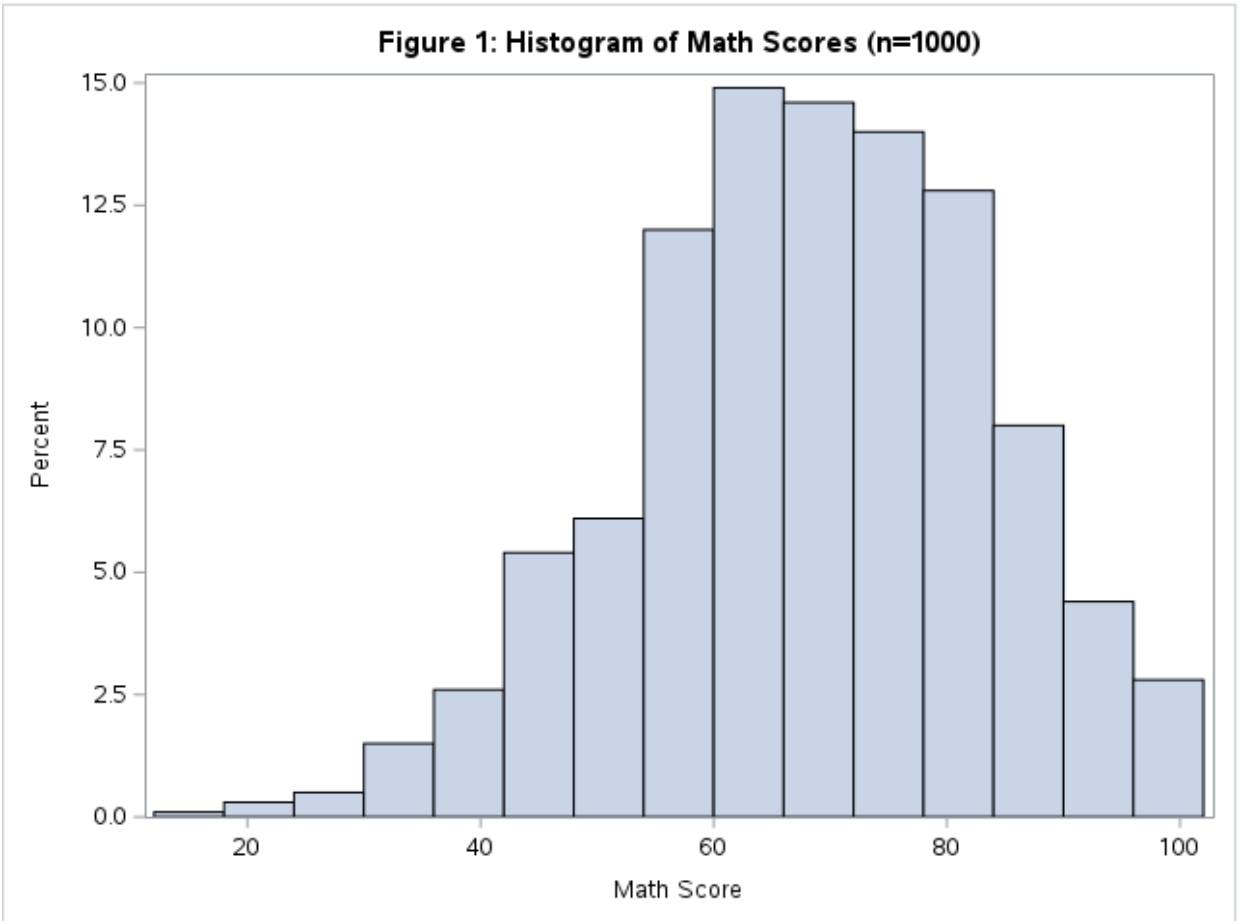


Figure 2: Histogram of Reading Scores (n=1000)

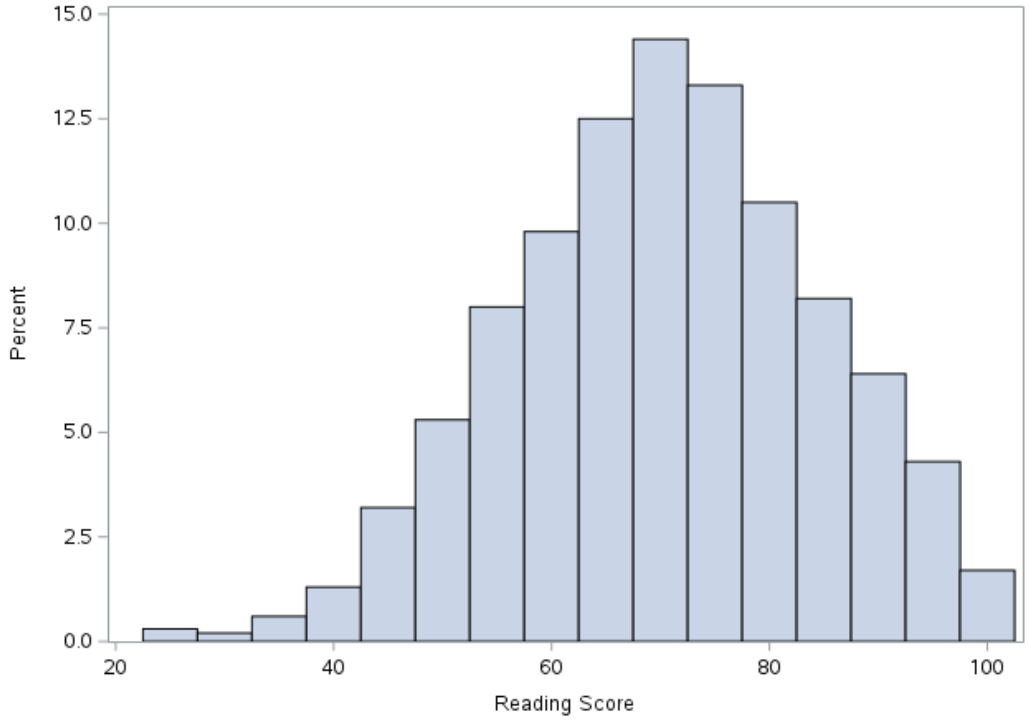
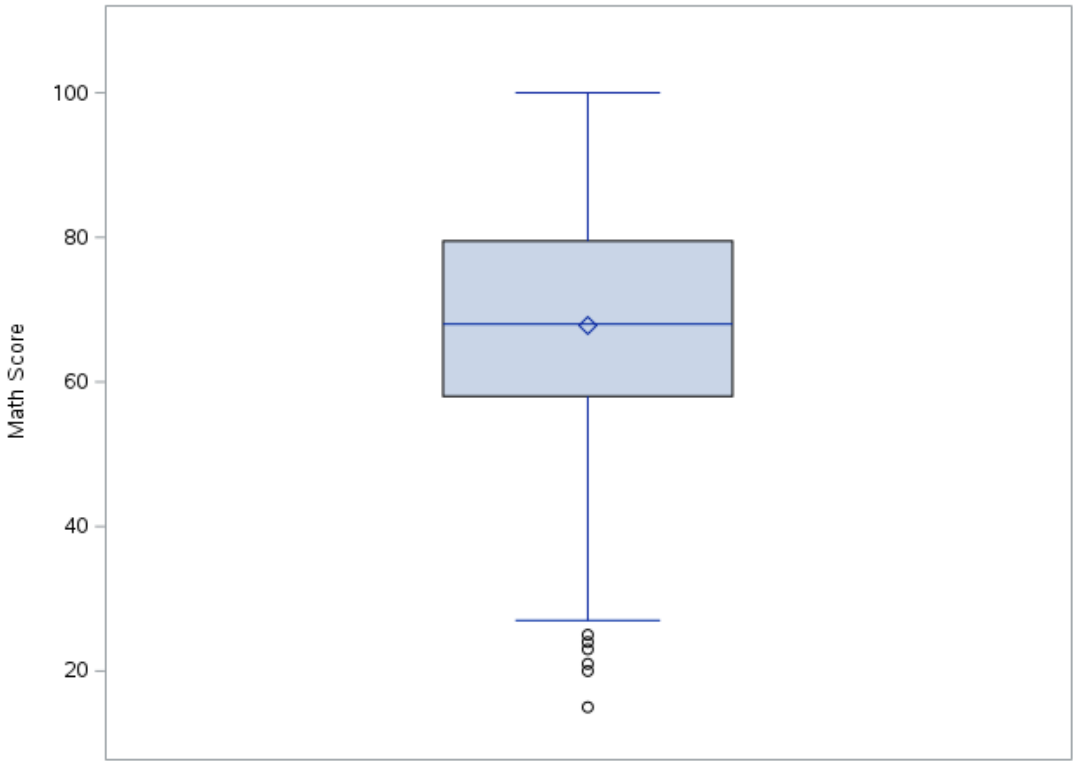
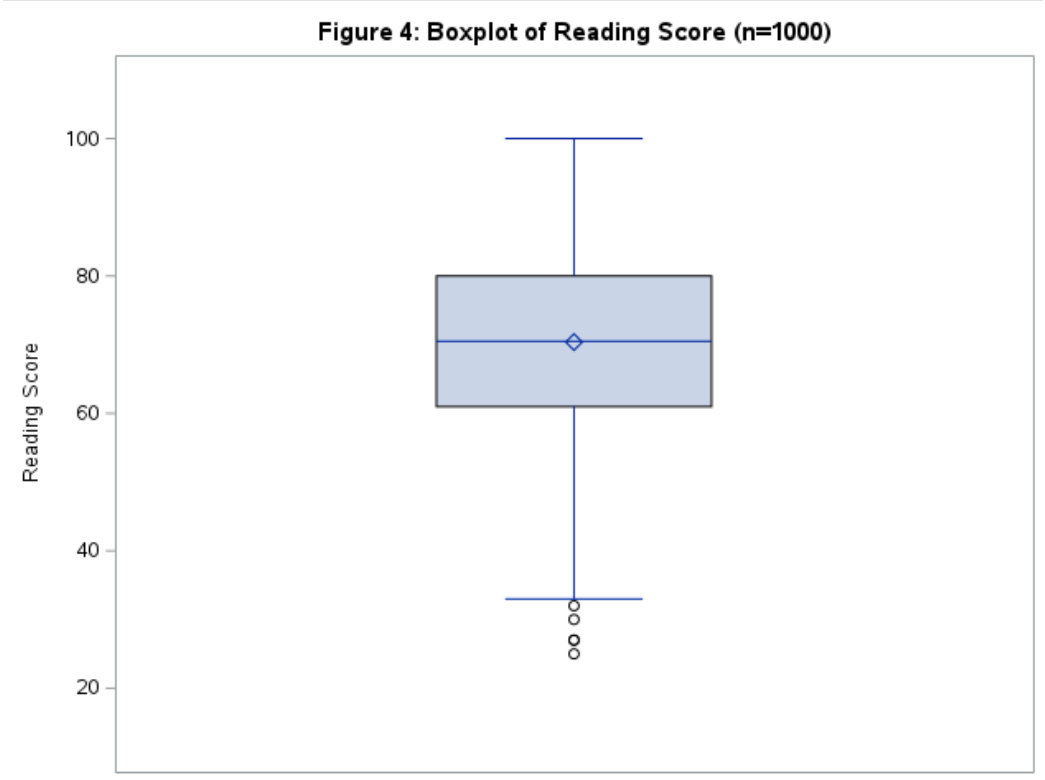


Figure 3: Boxplot of Math Score (n=1000)





From looking at both reading and math test score histograms, we can see that they have a normal distribution with the math score histogram being slightly more uniform. The math test has a lower average score of 67.81% in comparison to the average reading test score of 70.38%. Additionally, the math test has a higher quartile range as a result of the scores being more spread. The spread for math scores is represented by the larger number of outliers it has. The minimum score for the math test is a 15%, and 25% for the reading test. Overall, it looks like the students struggled more with the math test in comparison to the reading test.

b)

Table 3: Frequency Table of Student Gender

The FREQ Procedure

gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
female	492	49.20	492	49.20
male	508	50.80	1000	100.00

Table 4: Frequency Table of Student Lunch

The FREQ Procedure

lunch	Frequency	Percent	Cumulative Frequency	Cumulative Percent
free/reduced	340	34.00	340	34.00
standard	660	66.00	1000	100.00

Table 5: Frequency Table of Test Preparation Course

The FREQ Procedure

test preparation course	Frequency	Percent	Cumulative Frequency	Cumulative Percent
completed	344	34.40	344	34.40
none	656	65.60	1000	100.00

Figure 5: Pie Chart of Student Gender (n=1000)

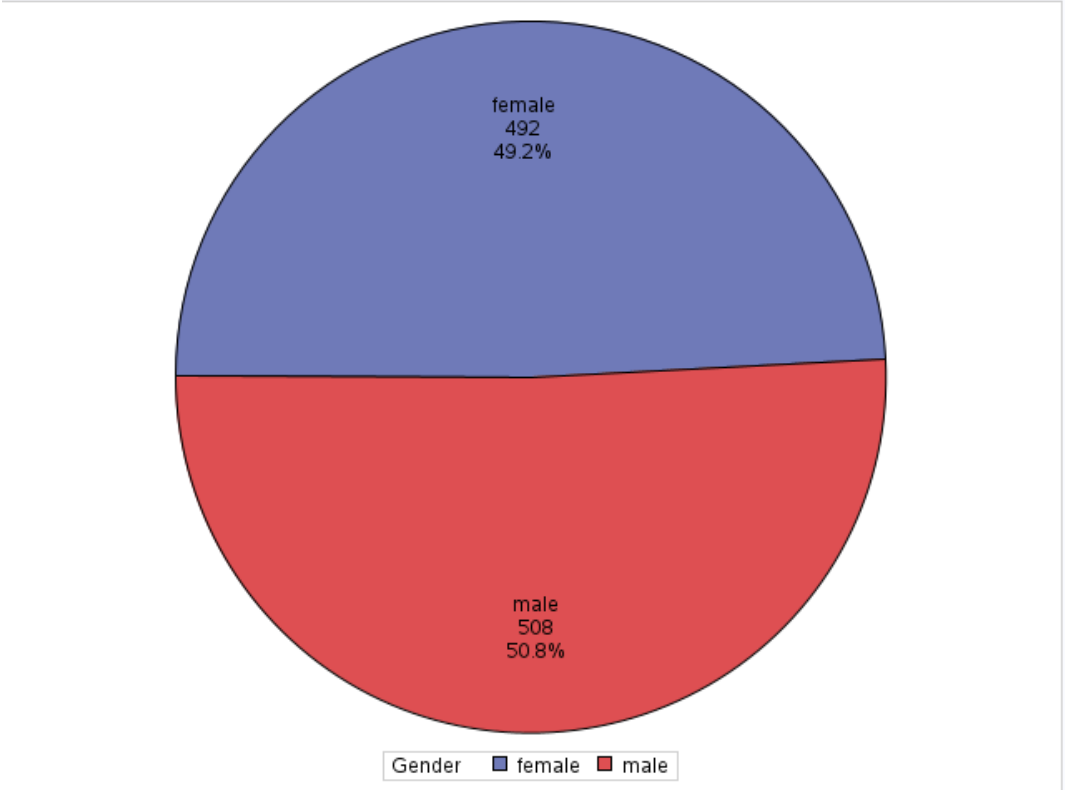


Figure 6: Pie Chart of Student Lunch (n=1000)

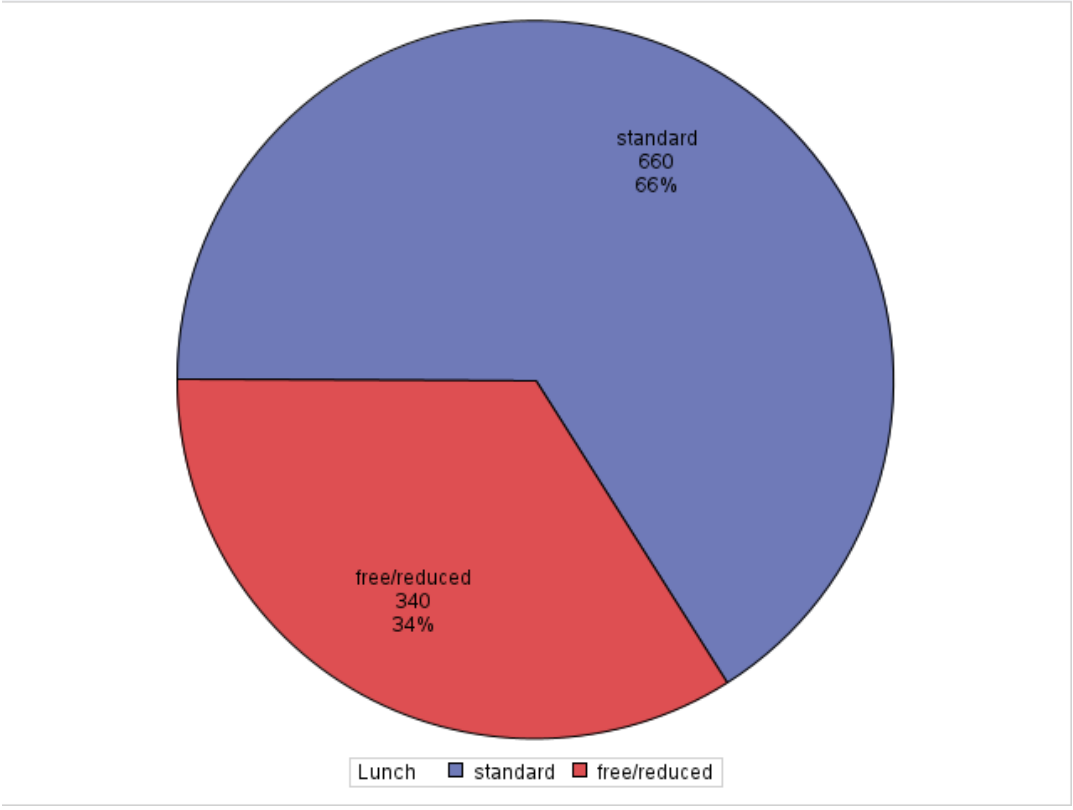


Figure 7: Pie Chart of Test Preparation Course Completion (n=1000)

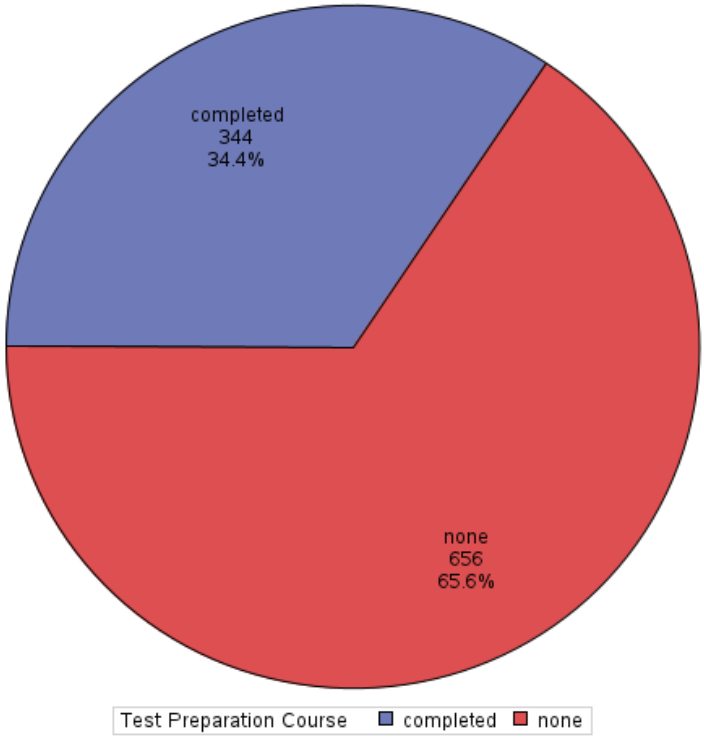


Figure 8: Bar Chart of Student Gender (n=1000)

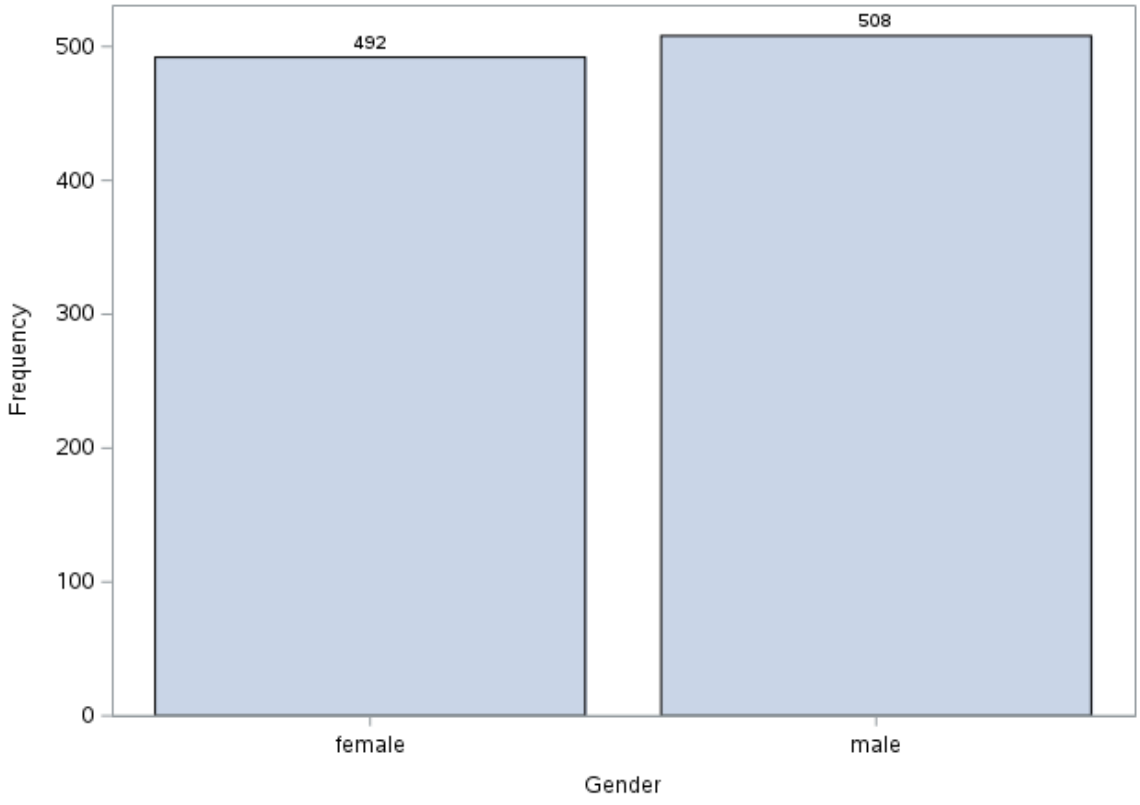


Figure 9: Bar Chart of Student Lunch (n=1000)

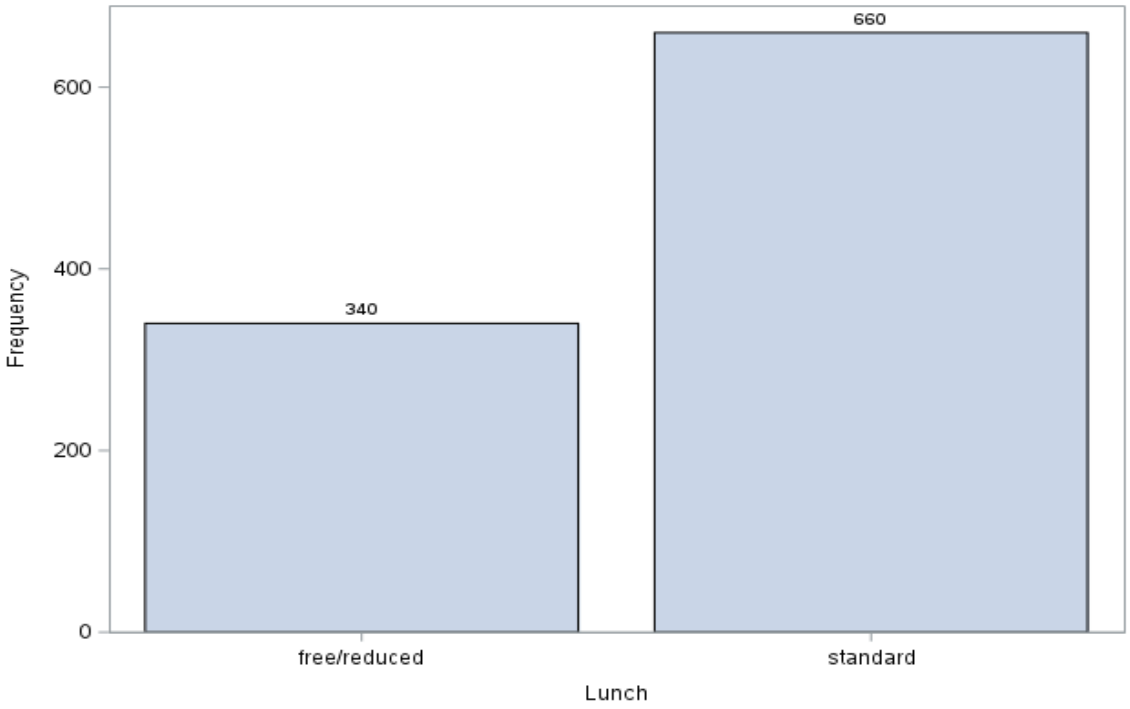
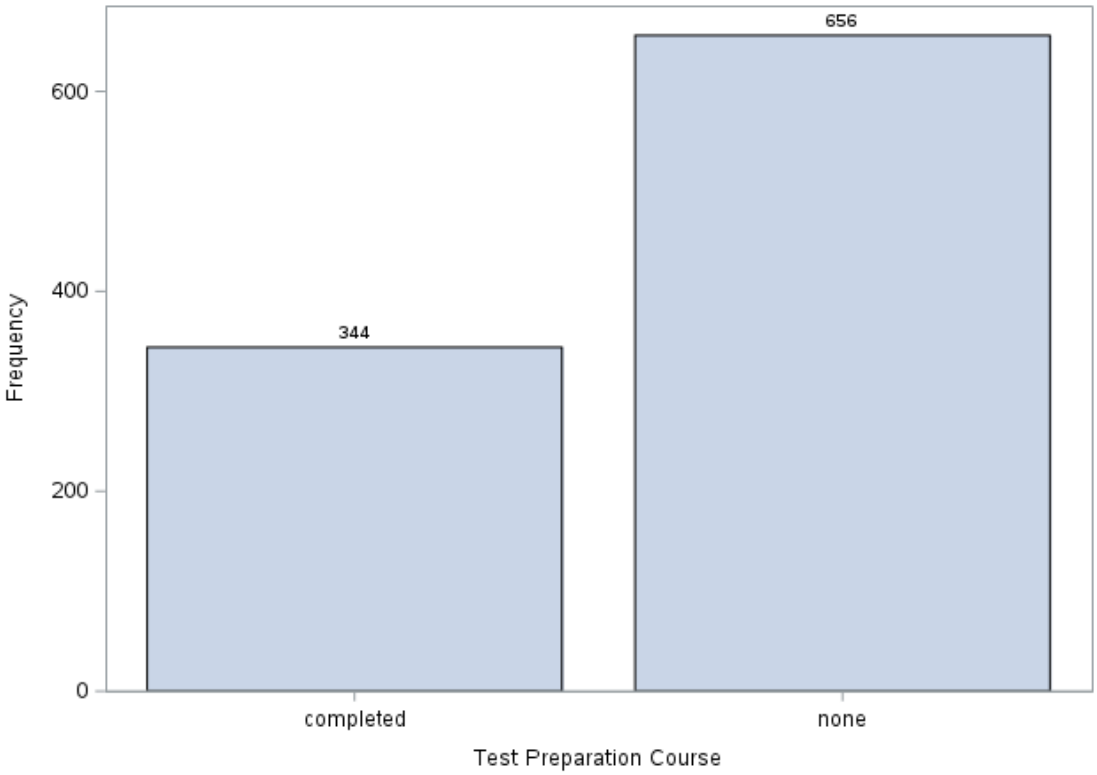


Figure 10: Bar Chart of Test Preparation Course Completion (n=1000)



The frequency spread of student gender is almost equivalent with a total of 492 females or 49.2% and 508 males or 50.8%.

Roughly a third of the students receive free or reduced-price lunch totaling 340 students or 34%; 660 or 66% of the students do not.

The test preparation course was completed by 344 students or 34.4%, and not completed by the remaining 656 students, or 65.6%.

There are no missing values in these categorical data sets.

4.

a)

Table 6: Frequency Table of Reading Performance

The FREQ Procedure

performance	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Unsatisfactory	223	22.30	223	22.30
Less Than Satisfactory	251	25.10	474	47.40
Satisfactory	265	26.50	739	73.90
Good	164	16.40	903	90.30
Excellent	97	9.70	1000	100.00

b)

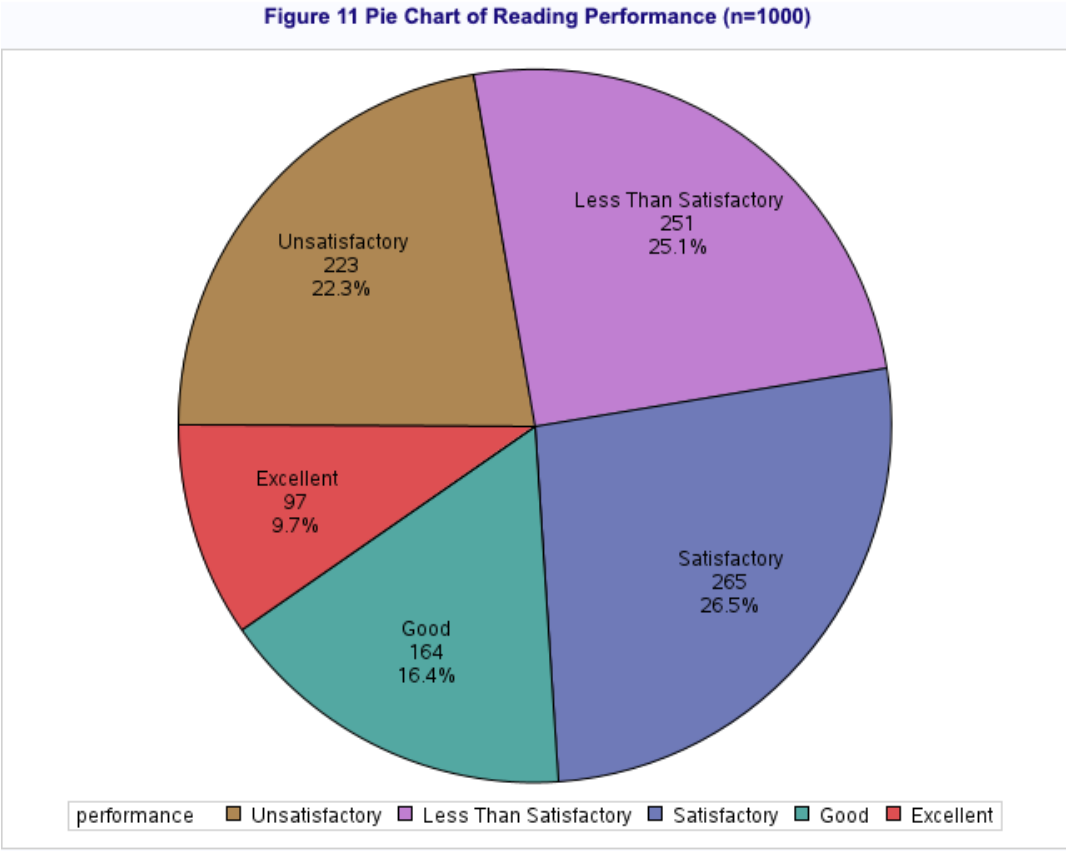


Figure 11 Pie Chart of Reading Performance (n=1000)
Less Than Satisfactory
251
25.1%
Unsatisfactory
223
22.3%
Excellent
97
9.7%
Satisfactory
265
26.5%
Good
164
16.4%
performance
Unsatisfactory
Less Than Satisfactory
Satisfactory
Good
Excellent

c)

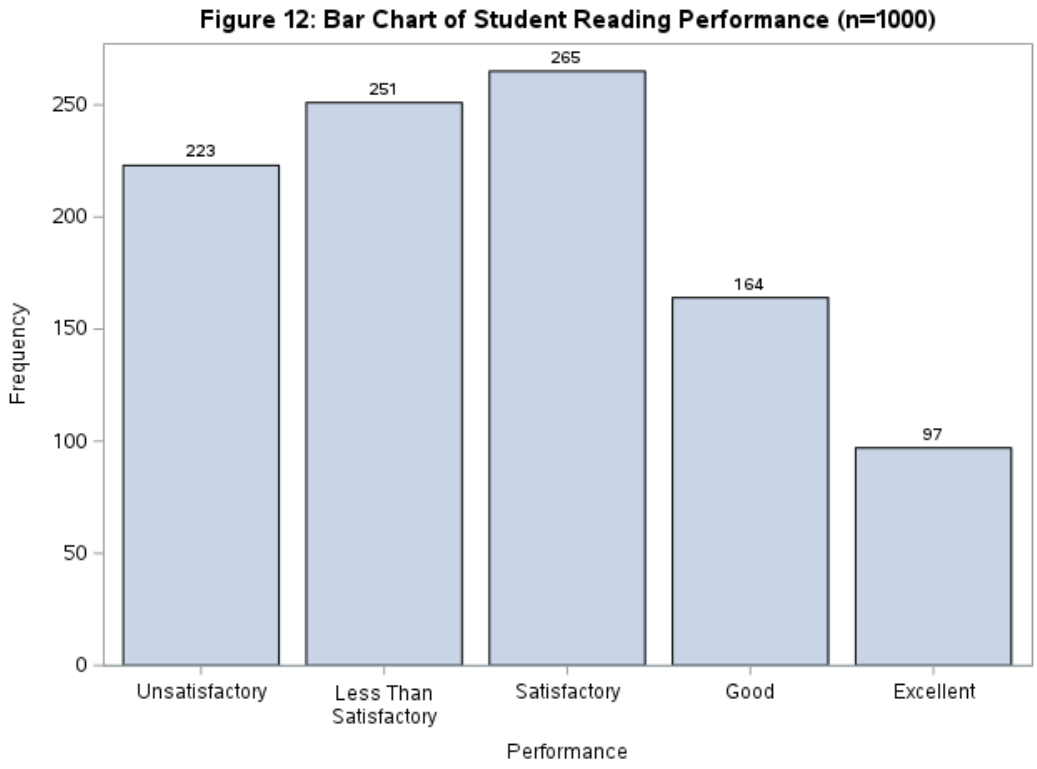


Figure 12: Bar Chart of Student Reading Performance (n=1000)
265
251
223
164
97
250
200
150
100
50
0
Frequency
Unsatisfactory
Less Than Satisfactory
Satisfactory
Good
Excellent
Performance

d) The new categorical variable that I introduced was a performance indicator of the reading test. I categorized each grade into new classifications. These classifications are distributed from the grade the students received. From 0-59 is “Unsatisfactory”, 60-69 is “Less Than Satisfactory”, 70-79 is “Satisfactory”, 80-89 is “Good”, and 90-100 is “Excellent”.

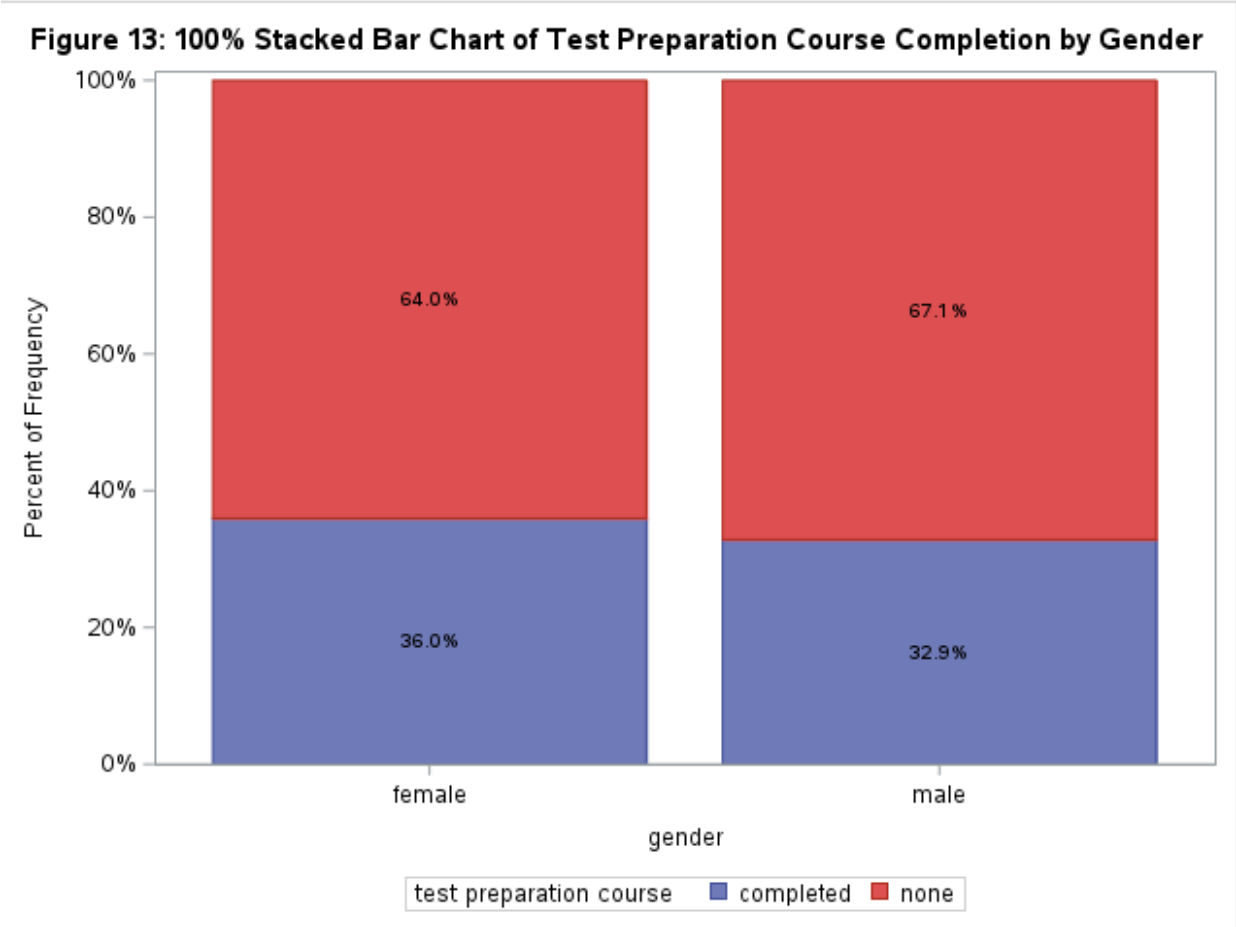
5.

2 Categorical

The two categorical variables I compared were Gender and Test Preparation Course completion. My hypothesis was that the number of students who took the preparation course was going to be evenly distributed between females and males. This hypothesis is proven to be accurate as 17.7% of females and 16.7% of males completed the course. This is represented in Figure 13, as both bar charts look almost equivalent. The explanatory variable is the student's gender, and the response variable is test preparation course completion.

Table 7: Gender by Test Preparation Course Completion

	test preparation course		
gender	completed	none	total
female	17.7%	31.5%	49.2%
male	16.7%	34.1%	50.8%
total	34.4%	65.6%	100.0%



2 Quantitative

The two quantitative variables that were compared were the math and reading test scores of the students. My hypothesis was that students who received a high grade on one test would also score high on the other. This hypothesis was proven to be true, as the correlation coefficient for the test scores was 0.8117. This coefficient represents a strong correlation between the scores of both tests. Figure 14 portrays this, showing a positive linear slope. The explanatory variable is the students' math score, and the response variable is the reading score.

Table 8: Correlation Coefficent of Math Score and Reading Score

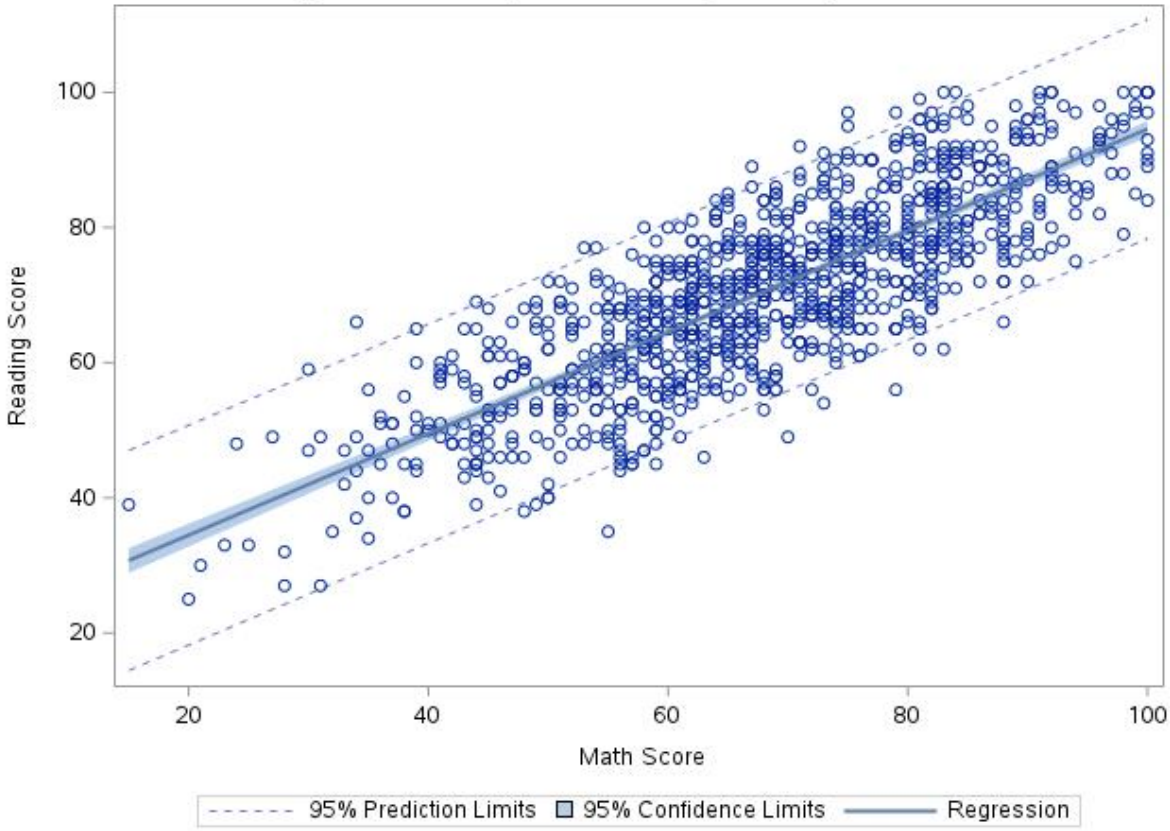
The CORR Procedure

2 Variables:	reading_score math_score

Pearson Correlation Coefficients, N = 1000
Prob > |r| under H0: Rho=0

	reading_score	math_score
reading_score	1.00000	0.81177 <.0001
math_score	0.81177 <.0001	1.00000

Figure 14: Scatterplot of Reading Score by Math Score



1 Categorical 1 Quantitative

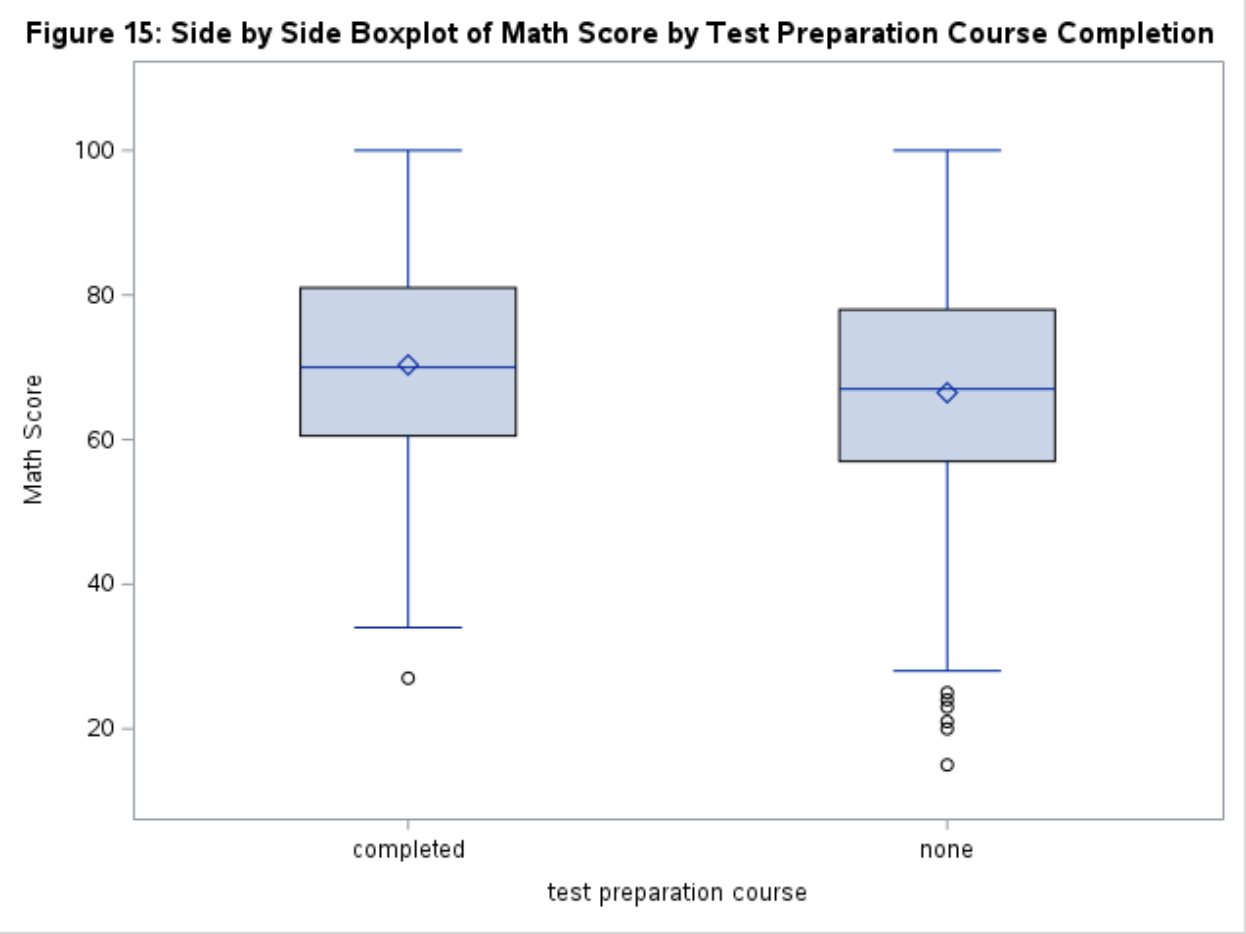
The categorical variable I used was the Test Preparation Course and the quantitative variable is the students' math scores. I hypothesized that the students who completed the preparation course were going to score higher on the test. This hypothesis was proven to be true, with the average score of a student who completed the preparation course being 70.33%, and 66.49% for the students who did not. The lowest score for students who completed the preparation course was 27%, and 15% for those who did not. Completing the preparation course influenced the students' scores. Figure 15 represents this conclusion, showing a higher median value and fewer low-scoring outliers for those who completed the course. The explanatory variable was the completion of the Test Preparation Course, and the response variable is the math test scores.

Table 9: Descriptive Statistics of Math Score by Test Preparation Course Completion

The MEANS Procedure

Analysis Variable : math_score									
test preparation course	N Obs	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Quartile Range
completed	344	70.33	14.69	27.00	60.50	70.00	81.00	100.00	20.50
none	656	66.49	15.38	15.00	57.00	67.00	78.00	100.00	21.00

Figure 15: Side by Side Boxplot of Math Score by Test Preparation Course Completion



Conclusion

From the bivariate analysis, there are some connections to be made. Students who score high on one test, score high on their other tests. This is a strong correlation that could be studied further. How are they scoring higher? One attribute that could affect this is preparation. The students who took the test preparation course outperformed those who did not. However, why was the percentage of students who completed the course so low? Looking at Figure 12, 47.4% of students had a classification of “Less Than Satisfactory” or lower. Pushing students to attend the preparation course can have a large effect on these numbers and improve their scores on the tests. My suggestion to further the findings of this research is to implement a strategy to get all the students to attend the

preparation course. In doing so, we can figure out more conclusions as to getting the test scores higher. Is their test-taking ability worse than other students? Do they struggle with comprehension? These are questions that can be answered if we can have every student prepare for the test in the same manner.

Appendix

```
/*Printing Data Set*/
proc print data= stat3010.widgethree;
run;
/*Making Permanent Library for data*/
data stat3010.widgethree;
set stat3010.widgethree;
run;
/*Quant Analysis*/
proc means data=stat3010.widgethree mean std min q1 median q3 max
qrange maxdec=2;
        var math_score reading_score;
        title 'Table 2: Descriptive Statistics for Quantitative Variables';
run;
/*Histo for Quant Variables*/
proc sgplot data=stat3010.widgethree;
        title 'Figure 1: Histogram of Math Scores (n=1000)';
        histogram math_score;
        xaxis label = 'Math Score';
run;
proc sgplot data=stat3010.widgethree;
        title 'Figure 2: Histogram of Reading Score (n=1000)';
        histogram reading_score;
        xaxis label = 'Reading Scores';
run;
/*Boxplots for Quant*/
proc sgplot data=stat3010.widgethree;
        title 'Figure 3: Boxplot of Math Score (n=1000)';
        vbox math_score;
        yaxis label = 'Math Score' valueshint min=10 max=110;
run;
proc sgplot data=stat3010.widgethree;
        title 'Figure 4: Boxplot of Reading Score (n=1000)';
        vbox reading_score;
        yaxis label = 'Reading Score' valueshint min=10 max=110;
run;
/*Uni Cat Frequency Tables*/
proc freq data=stat3010.widgethree;
        tables gender;
```

```
title 'Table 3: Frequency Table of Student Gender';
run;
proc freq data=stat3010.widgethree;
	tables lunch;
title 'Table 4: Frequency Table of Student Lunch';
run;
proc freq data=stat3010.widgethree;
	tables 'test preparation course'n;
title 'Table 5: Frequency Table of Test Preparation Course';
run;
/*Uni Cat Pie Charts*/
title 'Figure 5: Pie Chart of Student Gender (n=1000)';
proc template;
	define statgraph pie;
		begingraph;
			layout Region;
				Piechart category = gender /
				datalabellocation = inside
				datalabelcontent = all
				categorydirection= clockwise
				start = 180 name = 'pie';
				discretelegend 'pie'/
				title = 'Gender';
			endlayout;
		endgraph;
	end;
run;
proc sgrender data=stat3010.widgethree
			template=pie;
run;
title 'Figure 6: Pie Chart of Student Lunch (n=1000)';
proc template;
	define statgraph pie;
		begingraph;
			layout Region;
				Piechart category = lunch /
				datalabellocation = inside
				datalabelcontent = all
				categorydirection= clockwise
				start = 180 name = 'pie';
```

```
                        discretelegend 'pie'/
                        title = 'Lunch';
                endlayout;
        endgraph;
    end;
run;
proc sgrender data=stat3010.widgethree
                template=pie;
run;
title 'Figure 7: Pie Chart of Test Preparation Course Completion (n=1000)';
proc template;
    define statgraph pie;
        begingraph;
            layout Region;
                Piechart category =  'test preparation course'n /
                datalabellocation = inside
                datalabelcontent = all
                categorydirection= clockwise
                start = 180 name = 'pie';
                discretelegend 'pie'/
                title = 'Test Preparation Course';
            endlayout;
        endgraph;
    end;
run;
proc sgrender data=stat3010.widgethree
                template=pie;
run;
/*Bar Charts for Cat Variables*/
proc sgplot data=stat3010.widgethree;
    title 'Figure 8: Bar Chart of Student Gender (n=1000)';
    vbar gender / datalabel;
    xaxis label= 'Gender';
run;
proc sgplot data=stat3010.widgethree;
    title 'Figure 9: Bar Chart of Student Lunch (n=1000)';
    vbar lunch / datalabel;
    xaxis label= 'Lunch';
run;
proc sgplot data=stat3010.widgethree;
```

```
	title 'Figure 10: Bar Chart of Test Preparation Course Completion
(n=1000)';
	vbar 'test preparation course'n / datalabel;
	xaxis label= 'Test Preparation Course';
run;
/*Variable Creation*/
data stat3010.widgefour;
	set stat3010.widgethree;
	if reading_score <60 then performance = 1;
	Else if reading_score=> 60 and reading_score<70 then performance
= 2;
	else if reading_score => 70 and reading_score<80 then performance
= 3;
	else if reading_score=> 80 and reading_score<90 then performance
= 4;
	else if reading_score => 90 then performance = 5;
run;
/*renaming categories*/
proc format;
	value per    1='Unsatisfactory'
			2='Less Than Satisfactory'
			3='Satisfactory'
			4='Good'
			5='Excellent';
run; quit;
data stat3010.widgefour;
	set stat3010.widgefour;
	format performance per.;
run;
/*Frequency Table*/
proc freq data=stat3010.widgefour;
	tables performance;
	title 'Table 6: Frequency Table of Reading Performance';
run;
title;
/*Creating Pie Chart*/
title "Figure 11 Pie Chart of Reading Performance (n=1000)";
proc template;
	define statgraph pie;
		begingraph;
```

```
                    layout Region;
                        Piechart category = performance /
                        datalabellocation = inside
                        datalabelcontent = all
                        categorydirection= clockwise
                        start = 180 name = 'pie';
                        discretelegend 'pie'/
                        title = 'performance';
                    endlayout;
                endgraph;
        end;
run;
proc sgrender data=stat3010.widgefour
                    template=pie;
run;
/*Bar Chart*/
proc sgplot data=stat3010.widgefour;
        title 'Figure 12: Bar Chart of Student Reading Performance (n=1000)';
        vbar performance / datalabel;
        xaxis label = 'Performance';
run;
/*2 Cat*/
/*Row Percent Contingency Table*/
/*Gender and Test Prep Course Completion*/
proc sort data=stat3010.widgefour;
        by Gender;
run;
proc sort data=stat3010.widgefour;
        by 'test preparation course'n;
run;
/*Edit Table in Excel*/
proc freq data=stat3010.widgefour;
        tables gender*'test preparation course'n;
run;
title;
/*100% Stacked Bar Chart*/
proc sgplot data=stat3010.widgefour pctlevel=group;
        vbar gender / group = 'test preparation course'n stat = pct seglabel;
        title 'Figure 13: 100% Stacked Bar Chart of Test Preparation Course
Completion by Gender';
```

```
run;
/*2 Quant Analysis*/
/*Correlation Coeff. of math and reading score*/
proc corr data=stat3010.widgefour;
	var reading_score math_score;
	title 'Table 8: Correlation Coefficent of Math Score and Reading
Score';
run;
/*Scatterplot*/
proc sgplot data=stat3010.widgefour;
	title 'Figure 14: Scatterplot of Reading Score by Math Score';
	reg x = math_score y = reading_score / clm cli;
	xaxis label = 'Math Score';
	yaxis label= 'Reading Score';
run;
/*1 Cat 1 Quant*/
/*Stratified Table of Descriptive Stats for Test Preparation Course and Math
Score*/
proc means data=stat3010.widgefour mean std min q1 median q3 max
qrange maxdec=2;
	var math_score;
	class 'test preparation course'n;
	title 'Table 9: Descriptive Statistics of Math Score by Test Preparation
Course Completion';
run;
/*Side by Side Boxplot*/
proc sgplot data=stat3010.widgefour;
	title 'Figure 15: Side by Side Boxplot of Math Score by Test
Preparation Course Completion';
	vbox math_score / category = 'test preparation course'n;
	yaxis label = 'Math Score' valueshint min = 10 max = 110;
run;
```