



University of Wales Trinity Saint David

School of Engineering

Engineering Applications & Study Skills



Prifysgol Cymru

Y Drindod Dewi Sant

University of Wales

Trinity Saint David

Shay Quinn

2319936

BEng Mechanical Engineering Degree

28/05/2024

Engineering Applications & Study Skills

ACMA4007 (PT)



Table of Contents

1	Introduction.....	3
2	Data Import & Summary Statistics.....	3
2.1	Correlation Analysis.....	4
2.2	Grouped Data for Exam 1, Exam 2 & O/All Results.....	6
3	Discussion & Conclusion.....	7
4	References.....	8
5	Appendix.....	8



1 Introduction

In this report I will be discussing the MATLAB script I wrote for the application of statistical analysis methods on a dataset containing student grades and information. Using this script it was then used for statistical analysis to gain an insight into different factors that affect students performances and how it reflects on their grades, such things as gender, attendance and age are taken into account, however, in this assignment only the gender was asked of, so a hypothesis test was done to assess if gender was a factor of affecting the students grades. Along with this the mean, median, mode, range (Including minimum & maximum) and the standard deviation for each variable, age, attendance, Exam 1, Exam 2 & Overall Exam Score. The objective of this report will be to perform various statistical analyses on the dataset of student grades provided to gain an insight into factors effecting their grades. A grouping of data will done to the dataset to ensure all data is accounted for and none is missed, a histogram, and ogive frequency graph will then be made to display the findings from the dataset.

2 Data Import & Summary Statistics

The process of importing the dataset started firstly, importing the Student Grades excel file to MATLAB's and then reading the file with the function 'readmatrix(Student_Grades.csv)' turning the dataset into a matrix allowing it to be used within the script and to allow MATLAB to read the data imported, from this I then created variables for row of data, firstly gathering the row of data then finding the mean, mode, median, range, minimum, maximum and finally standard deviation. I done this for the 4 rows then once completed creating a new row for the overall exam mark done adding the two exams together and dividing by 2 this when they give their average mark of the 2 exams completed using this I then continued to do the above again for this row of data. Once done, I created the variable names for the rows, defined by VN for the table names, then creating a results table as shown in figure 1 below. Below is what the different summary statistics mean and an advantage and disadvantage of each.

Mean – Mean is a simple mathematical average of set of two or more sets of numbers within a dataset. One advantage of this is that it takes account of all values to calculate the average, one disadvantage of this would be that a very small or very large value compared to the rest of the values in the dataset.

Median – This is the middle number in a sorted list of numbers and can be more descriptive of a dataset than the average. One advantage of this is that it is not affected by very large or very small numbers, one disadvantage is that if the number of values is not even then the middle is found by averaging the two middle numbers meaning the number found may not be in the original dataset.

Mode – The mode is the value within the dataset that appears the most frequently, this can typically be the same value as the mean and median. One advantage is that the mode is not affected by any extreme and unusual values, one disadvantage is that there can often be more than one mode, only one mode or no mode at all.

Range – This refers to the difference between the low and the high of the values from the dataset, one advantage is that the range is easy to calculate for statistics, the disadvantage is that it is very sensitive to miscellaneous datasets that don't follow the pattern.



Standard Deviation – This measures the amount of variation in relation to the mean. One advantage is that the standard deviation is based on all observations of a dataset and all values. One disadvantage is that it is very sensitive to random values not in pattern with the dataset .

Variable	Mean	Median	Mode	Range		Standard Deviation
"Age"	36.65	37.5	48	19	53	9.7643
"Attendance"	60.58	57	27	20	100	24.249
"1st Exam Score"	58.27	58	59	7	98	21.673
"2nd Exam Score"	55.08	53.5	51	4	98	22.428
"O/All Score"	56.675	54.5	25.5	9.5	95	21.452

Figure 1 – Results Table

From Figure 1, it shows the table created and the data gathered from this, with the mean, median, mode, range & standard deviation being displayed for each of the variable rows. A limitation of the mean median and mode used is that these summary statistics are not useful for extremely skewed data, that is within the dataset, the median and mode sometimes may not achieve the actual average value of the datasets, and another limitation may be outliers affecting the Mean results of the data. Other things need to be accounted for such as the median may it make it difficult to get an accurate understanding of relative magnitude, leading to difficulties with comparisons.

2.1 Correlation Analysis

What is Correlation Analysis? Correlation analysis is a statistical method that is used to identify relationships between variables and datasets, including the potential relationships that could be occurring. The higher the correlation the stronger the relationship between the variables, while a lower correlation would indicate a weaker relationship. Correlation analysis can be used to analyse quantitative dataset, which identify the relationship, patterns, connections and trends. A strong correlation between datasets or variables would be classed as a positive correlation, while a weak correlation between datasets or variables would be classed as a negative correlation.

Implications of Correlation Analysis

Outliers – Identifying unusual correlations can help understand unusual correlations which may highlight anomalies within the data results.

No Causation – Correlation may not indicate causation, due to variables potentially moving together for a strong relationship, due to unseen factors within the dataset or another variable not taken into account, this then can affect the results altogether causing them to be unreliable.

Spurious Correlations – Correlations may sometime happen by chance, although not all the time it can happen and this is then due to larger datasets. From this it can lead to wrong conclusions.



CorrelationResults =

1.0000	-0.0362	-0.0364	-0.0506	-0.0448
-0.0362	1.0000	0.8912	0.8297	0.8839
-0.0364	0.8912	1.0000	0.8929	0.9719
-0.0506	0.8297	0.8929	1.0000	0.9738
-0.0448	0.8839	0.9719	0.9738	1.0000

Figure 2 – Correlation Coefficient Results

Figure 2 is the correlation coefficient, these typically range from -1 to 1 and how quantify how closely variables move together. However, sometimes variables may be correlated without one causing the other, sometimes due to lurking variables or coincidences. As shown, there are some very strong relationships and some weaker relationships from the results, which can be seen by the Numbers within the results, 1's indicating a strong correlation and -1's indicating a weak correlation while 0's indicate a neutral correlation. Figure 2 then also corresponds with Figure 3, showing the relationship strengths.

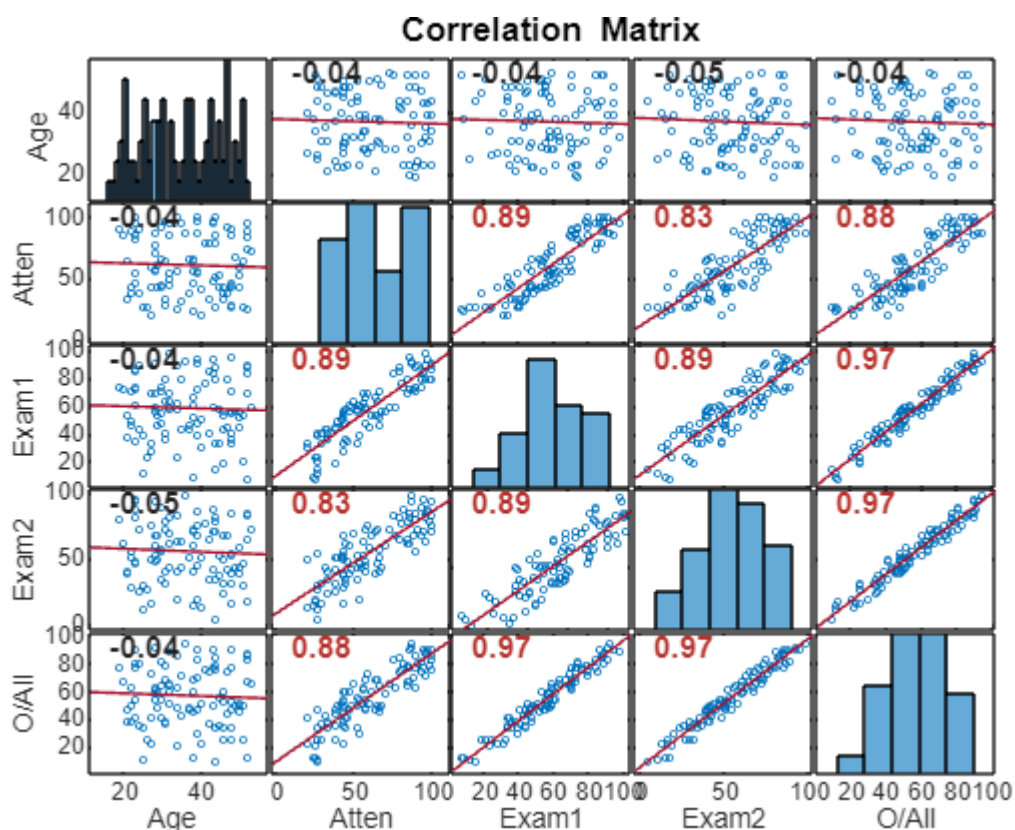


Figure 3 – Correlation Matrix Results

Figure 3 is a correlation matrix, this is a statistical technique to find the relationship between two or more variables in a dataset, the above table represents the relationships between the



age, attendance, Exam 1 Scores, Exam 2 Scores & Overall Exam Score. As seen above, there is a strong correlation and a large positive correlation from all the Exam Scores indicating a strong relationship between them which is expected. From the Correlation matrix it is seen that age is not an affecting factor into Exam Score with the results having no pattern throughout the age groups, Attendance can be seen with an increasing result as the attendance percentage goes up as well, this is showing how the attendance of the students is a factor on their grades.

2.2 Grouped Data for Exam 1, Exam 2 & O/All Results

What is Grouped Data? This is grouping data in intervals such as 0-10,10-20 etc. Ungrouped data are individual points or raw data that is unsorted when first gathered from a dataset. Grouped data can be shown in histograms and frequency tables. Grouping data allows to focus on important subgroups and overlooks the trivial groups, increasing the efficiency and accuracy of the estimation. Grouped data allows for faster testing and more informed decisions based on clear data.

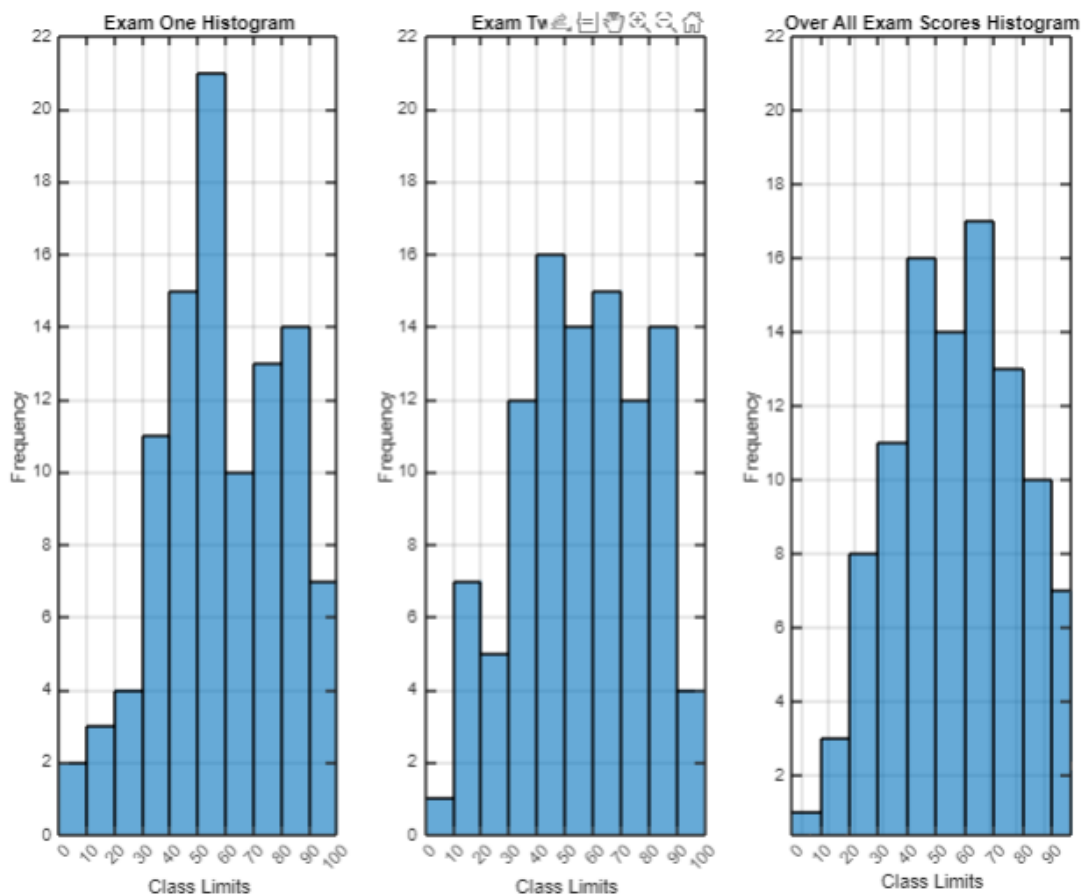


Figure 4 – Exam 1, Exam 2, & Overall Exam Scores Histograms

Figure 4 Shows the histograms created for the Exam 1, Exam 2 and Overall Exam Scores, the grouped data used, the intervals used and all the possibly groups shown in the histogram. This accurately shows all the results from the dataset, including the Overall Exam



Score, revealing that in Exam 1 there were more people who scored 50%, while in Exam 2 it shows that the exam scores are more evenly made in Exam 2.

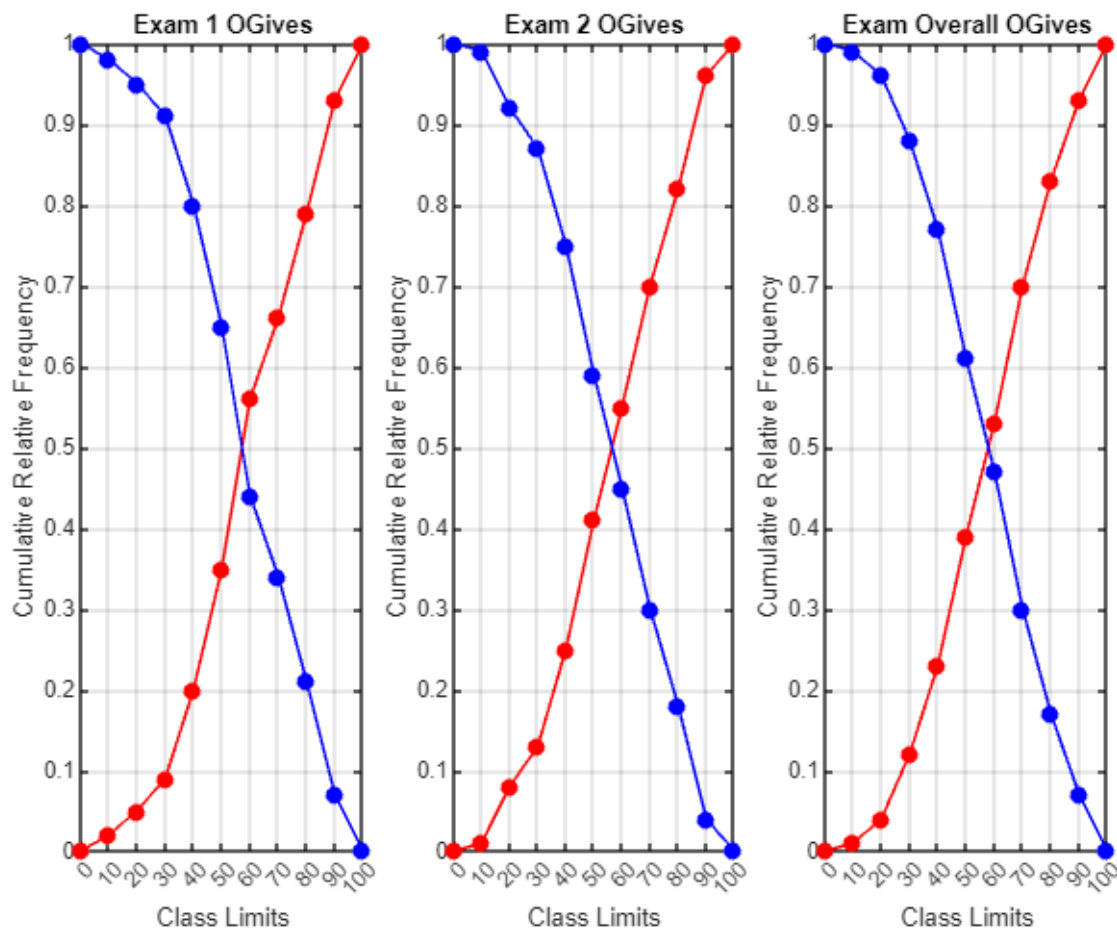


Figure 5 – Exam 1, Exam 2 & Overall Exam Score Ogives Graph

Figure 5 show the 3 Ogive graphs for Exam 1, Exam 2 & Overall Exam Scores, an Ogive graph is defined as a frequency distribution graph of a series, this graph explains the cumulative relative frequency distribution and reverse cumulative relative frequency distribution of Male and Female Exam Results for each of the exams and the overall score.

3.0 Discussion & Conclusion

In conclusion it was found that more females than males had a higher exam score, therefore showing there are factors that may affect the students grades, gender affecting the students grades and attendance affecting the students exam scores, it was found that the lower their attendance the lower their exam scores were. Throughout the statistical analysis of potential factors affecting student grades there were many limitations and implications, firstly, there were the statistical limitations of the mean, median, mode & range, these all individually but, all have their advantages. Such as some central tendencies are susceptible to skewness and outliers, with this being a major drawback with most statistical analysis methods which can affect the outcome of the result, however, these are extremely easy and quick to calculate and allow for easy statistical analysis and are very valuable methods for



comparison. Although there were some limitations within the statistical analysis methods, not many presented themselves within the dataset.

4 References

[Documentation - MATLAB & Simulink - MathWorks United Kingdom](#) – MathWorks Support

[Limitations Of Statistics | TutorOcean Questions & Answers](#) – Tutor Ocean

[Advantages and Disadvantages of Mean, Median, and Mode - Javatpoint](#)

[Grouping of Data - Definition, Advantages and Examples \(vedantu.com\)](#)

[Grouped Data / Ungrouped Data: Definition, Examples - Statistics How To](#)

[Ogive \(Cumulative Frequency Curve\) - Definition and its Types \(byjus.com\)](#)

5 Appendix

MATLAB SCRIPT

```
clc
clear
close

format short
%Question 1 a.)
D = readmatrix("student_grades.csv");

%Question 1 b.)
Age = D(1:100, 3); %Retrieves all the numbers from the 3rd row in the excel file.
AgeMean = mean(Age, "all"); %Mean of age
AgeMedian = median(Age); %Median of age
AgeMode = mode(Age); %Mode of age
AgeRange = range(Age); %Finds range of age
AgeMin = min(Age); %Finds the minimum Age from all the data
AgeMax = max(Age); %Finds the maximum Age from all the data
AgeStd = std(Age); %Creates a standard deviation of the age

Attendance = D(1:100, 4); %Retrieves all the data from the 4th row from the data
set
AttMean = mean(Attendance, "all"); %Finds the mean of the attendance
AttMedian = median(Attendance); %Finds the median of the attendance
AttMode = mode(Attendance); %Finds the mode of the attendance
AttRange = range(Attendance); %Finds the range of all the attendance data
AttMin = min(Attendance); %Finds the minimum value of all the attendance data
AttMax = max(Attendance); %Finds the maximum value of all the attendance data
AttStd = std(Attendance); %Finds the standard deviation of the minimum & maximum
value of the dataset
```



```
ExamOne = D(1:100, 5);%Retrieves all the Exam One data from the 5th row in the
dataset
ExOneMean = mean(ExamOne,"all");%Finds the mean of all the values of the Exam One
dataset
ExOneMedian = median(ExamOne);%Finds the median of all the values of the Exam One
dataset
ExOneMode = mode(ExamOne);%Finds the mode of Exam One dataset
ExOneRange = range(ExamOne);%Finds the range of the Exam One dataset
ExOneMin = min(ExamOne);%Finds the minimum value of the Exam One dataset
ExOneMax = max(ExamOne);%Finds the maximum value of the Exam One dataset
ExOneStd = std(ExamOne);%Finds the standard deviation of the Exam One dataset
CME1 = cumsum(ExamOne);%Finds the cumulative sum of the Exam One dataset
```

```
ExamTwo = D(1:100, 6);%Retrieves all the Exam Two data from the 6th row in the
dataset
ExTwoMean = mean(ExamTwo,"all");%Finds the mean of all the values of the Exam Two
dataset
ExTwoMedian = median(ExamTwo);%Finds the median of all the values of the Exam Two
dataset
ExTwoMode = mode(ExamTwo);%Finds the mode of the Exam Two dataset
ExTwoRange = range(ExamTwo);%Finds the range of the Exam Two dataset
ExTwoMin = min(ExamTwo);%Finds the minimum value of the Exam Two dataset
ExTwoMax = max(ExamTwo);%Finds the maximum value of the Exam Two dataset
ExTwoStd = std(ExamTwo);%Finds the standard deviation of the Exam Two dataset
CME2 = cumsum(ExamTwo);%Finds the cumulative sum of the Exam Two dataset
```

```
OvrAll = (ExamOne + ExamTwo)/2;%Takes the Exam One & Exam Two datasets, adds them
together and then divides them by the number of exams there are in this case 2
OvrAllMean = mean(OvrAll,"all");%Finds the mean of all the values of the Overall
Exam Scores
OvrAllMedian = median(OvrAll);%Finds the median of all the values of the Overall
Exam Scores
OvrAllMode = mode(OvrAll);%Finds the mode of the Overall Exam Scores
OvrAllRange = range(OvrAll);%Finds the range of the Overall Exam Scores
OvrAllMin = min(OvrAll);%Finds the minimum value of the Overall Exam Scores
dataset
OvrAllMax = max(OvrAll);%Finds the maximum value of the Overall Exam Scores
dataset
OvrAllStd = std(OvrAll);%Finds the standard deviation of the Overall Exam Scores
dataset
OvrAllT = table(OvrAll);%Creates a table descending of all the Overall Exam Scores
CMOA = cumsum(OvrAll);%Finds the cumulative sum of all the Overall Exam Scores
dataset
```

```
AllMeans = [AgeMean;AttMean;ExOneMean;ExTwoMean;OvrAllMean];%gets all the mean
results together
AllMedians = [AgeMedian;AttMedian;ExOneMedian;ExTwoMedian;OvrAllMedian];%Gets all
the median results together
AllModes = [AgeMode;AttMode;ExOneMode;ExTwoMode;OvrAllMode];%Gets all the modes
results together
AllRanges = [AgeRange;AttRange;ExOneRange;ExTwoRange;OvrAllRange];%Gets all the
range results together
AllMinimums = [AgeMin;AttMin;ExOneMin;ExTwoMin;OvrAllMin];%Gets all the minimum
results together
AllMaximums = [AgeMax;AttMax;ExOneMax;ExTwoMax;OvrAllMax];%Gets all the maximum
results together
```



```

AllStd = [AgeStd;AttStd;ExOneStd;ExTwoStd;OvrAllStd];%Gets all the Standard
Deviation results together
AllMinMax =
[AgeMin,AgeMax;AttMin,AttMax;ExOneMin,ExOneMax;ExTwoMin,ExTwoMax;OvrAllMin,OvrAllM
ax];%Displays the minimum and maximum of each variable and pairs them

Variable = ["Age";"Attendance";"1st Exam Score";"2nd Exam Score";"O/All Score"];

VN = {'Variable','Mean','Median','Mode','Range','Standard Deviation'}; %Creating
names for the table

%Question 1 C.)
ResultsTable =
table(Variable,AllMeans,AllMedians,AllModes,AllMinMax,AllStd,'VariableNames',VN);%
Creates the results table

ImportedDataTable = D(:,3:6);%Takes the dataset and creates a table
AllTable = [ImportedDataTable,OvrAll];%Uses the table above and the Overall Score
and creates a table with an extra row to include the Overall Exam Score

disp(ResultsTable) %Displays Results Table for Question 1 C.)
%Question 2 a.)
CorrelationResults = corr(AllTable) %Displays the correlation findings of dataset
"AllTable"

%Question 2 b.)
CMVN = {'Age','Attendance','Exam1','Exam2','O/All'};%Creates the variable names
for the rows
figure(1)
CM = corrplot(AllTable,"Rows","all",TestR="on",VarNames=CMVN);%Creates a
correlation matrix using the all table above

%Question 2 c.)

Gndrs1 = readtable("student_grades.csv"); %Creates a read table variable to allow
the code below to to read the cells as text an not displayed as NaN
Gndrs = [Gndrs1,OvrAllT];%Adds the Overall Exam Scores to the table above

M1 = Gndrs.gender == ("Male"); %Gets all the Male data and stores into the
variable M1
FM1 = Gndrs.gender == ("Female"); %Gets all the Female data and stores into the
variable FM1

SpM = Gndrs(M1,:); %Creates a table for all Male data from the file
SpFM = Gndrs(FM1,:); %Creates a table for all Female data from the file.

Ex1 = Gndrs(:,5);%Retrieves all the Exam One Results from the readtable above
Ex2 = Gndrs(:,6);%Retrieves all the Exam Two Results from the readtable above

%Hypothesis Test for the overall score on the exam between Male & Females.

[h0,p0] = ttest2(SpM{:,7},SpFM{:,7})%Overall Score Hypothesis Test between Male &
Female

%histcounts
edges = 0:10:100;%Edges for the Exam Results, as student can score between 0 & 100
with intervals of 10

```



```

H1 = histcounts(ExamOne, edges)
H2 = histcounts(ExamTwo, edges)
H0 = histcounts(OvrAll, edges)
tot = [H1 H2 H0];%Total of hiscounts
maxfreq1 = max(tot);%Finds the highest number in the dataset

%Histogram for Exam One
figure(2)

maxfreq = maxfreq1 + 1;%Uses the histcounts to find the highest number and adds 1
to increase the histogram frequency to 22
subplot(1,3,1)%Plots the Histogram on one Figure out of 3
HistExOne = histogram(ExamOne,'BinEdges',edges)%Histogram for Exam One
title('Exam One Histogram');%Title for Histogram above
ylim([0 maxfreq]);%Y Axis Limit
xlim([0 100]);%X Axis Limit
xlabel('Class Limits');%X Axis Label
ylabel('Frequency');%Y Axis Label
xticks(edges);%States minimum and maximum of the X Axis along with the interval
set
grid on;%Turns the grid on for the Histogram

%Histogram for Exam Two
subplot(1,3,2);%Plots the Histogram on one Figure out of 3
HistExTwo = histogram(ExamTwo, 'BinEdges',edges) %Histogram for Exam Two
title('Exam Two Histogram');%Title for Histogram above
ylim([0 maxfreq]);%Y Axis Limit
xlim([0 100]);%X Axis Limit
xlabel('Class Limits');%X Axis Label
ylabel('Frequency');%Y Axis Label
xticks(edges);%States minimum and maximum of the X Axis along with the interval
set
grid on;%Turns the grid on for the Histogram

%Histogram for Overall Score of exams
subplot(1,3,3);%Plots the Histogram on one Figure out of 3
HistOvrAll = histogram(OvrAll, 'BinEdges',edges) %Histogram for Over All Exam
Score
title('Over All Exam Scores Histogram');%Title for Histogram above
ylim([0 maxfreq]);%Y Axis Limit
xlim([0 100]);%X Axis Label
xlabel('Class Limits');%X Axis Label
ylabel('Frequency');%Y Axis Label
xticks(edges);%States minimum and maximum of the X Axis along with the interval
set
grid on;%Turns the grid on for the Histogram

%Ogives graph for Exam 1
figure(3)
subplot(1,3,1);%Plots the Ogives graphs on one figure
Ex1CmF = cumsum(H1);%Cumulative Sum of Exam One Histcounts for the Cumulative
Frequency
Ex1CmRF = Ex1CmF/100;%Cumulative Relative Frequency using the Cumulative Frequency
found above
plot(edges, [0 Ex1CmRF], "-
o",'Color','red','MarkerEdgeColor','red','MarkerFaceColor','red')%Plots Cumulative
Relative Frequency for the Exam One Ogive Graph
hold on;%Holds on to compare the two plots of Ogives together

```



```

Ex1RevCmF = cumsum(H1,"reverse");%Reverse the Cumulative Frequency
Ex1RevCmRF = Ex1RevCmF/100;%Finds the Reverse Cumulative Relative Frequency
plot(edges, [Ex1RevCmRF 0], "-
o", 'Color', 'Blue', 'MarkerEdgeColor', 'Blue', 'MarkerFaceColor', 'Blue')%Plots the
Reverse Cumulative Relative Frequency for Exam One Ogive Graph
hold off;%Stops comparison of Ogives graphs
xticks(edges);%States minimum and maximum of the X Axis along with the interval
set
grid on;%Turns grid on for the Ogive graph
title('Exam 1 OGives')%Title for Ogive graph for Exam One
xlabel('Class Limits')%X Axis Label
ylabel('Cumulative Relative Frequency')%Y Axis Label

%Ogives graph for Exam 2
subplot(1,3,2)%Plots the Ogives graphs on one figure
Ex2CmF = cumsum(H2);%Cumulative Sum of Exam Two Histcounts for the Cumulative
Frequency
Ex2CmRF = Ex2CmF/100;%Cumulative Relative Frequency using the Cumulative Frequency
found above
plot(edges,[0 Ex2CmRF], "-
o", 'Color', 'red', 'MarkerEdgeColor', 'red', 'MarkerFaceColor', 'red')%Plots Cumulative
Relative Frequency for the Exam Two Ogive Graph
hold on;%Holds on to compare the two plots of Ogives together
Ex2RevCmF = cumsum(H2, "reverse");%Reverse the Cumulative Frequency
Ex2RevCmRF = Ex2RevCmF/100;%Finds the Reverse Cumulative Relative Frequency
plot(edges,[Ex2RevCmRF 0], "-
o", 'Color', 'Blue', 'MarkerEdgeColor', 'Blue', 'MarkerFaceColor', 'Blue')%Plots the
Reverse Cumulative Relative Frequency for Exam Two Ogive Graph
hold off;%Stops comparison of Ogives graphs
grid on;%Turns grid on for the Ogive graph
xticks(edges);%States minimum and maximum of the X Axis along with the interval
set
xlabel('Class Limits');%X Axis Label
ylabel('Cumulative Relative Frequency');%Y Axis Label
title('Exam 2 OGives');%Title for Ogive graph for Exam Two

%Ogives for Overall Exam Score
subplot(1,3,3)%Plots the Ogives graphs on one figure
ExOACmF = cumsum(HO);%Cumulative Sum of Overall Exam Score Histcounts for the
Cumulative Frequency
ExOACmRF = ExOACmF/100;%Cumulative Relative Frequency using the Cumulative
Frequency found above
plot(edges, [0 ExOACmRF], "-
o", 'Color', 'red', 'MarkerEdgeColor', 'red', 'MarkerFaceColor', 'red')%Plots Cumulative
Relative Frequency for the Overall Exam Score Ogive Graph
hold on;%Holds on to compare the two plots of Ogives together
ExOARevCmF = cumsum(HO, "reverse");%Reverse the Cumulative Frequency
ExOARevCmRF = ExOARevCmF/100;%Finds the Reverse Cumulative Relative Frequency
plot(edges, [ExOARevCmRF 0], "-
o", 'Color', 'Blue', 'MarkerEdgeColor', 'Blue', 'MarkerFaceColor', 'Blue')%Plots the
Reverse Cumulative Relative Frequency for the Overall Exam Score Ogive Graph
hold off;%Stops comparison of Ogives graphs
grid on;%Turns grid on for the Ogive graph
xticks(edges);%States minimum and maximum of the X Axis along with the interval
set
xlabel('Class Limits');%X Axis Label
ylabel('Cumulative Relative Frequency');%Y Axis Label
title('Exam Overall OGives');%Title for Ogive graph for Overall Exam Score

```

