

Analysis of Student Grades using Statistical Methods in MATLAB

Engineering Practice and Study Skills ACMA4007

Tyler Powell 2318809

1600 words max

Contents:

Title page	[page 1]
Contents	[page 2]
Data Import and Summary Statistics	[page 3]
Correlation Analysis	[page 4 – 6]
Visualisation of grouped data for Exam 1, Exam 2, and O/ALL results	[page 7 – 9]
MATLAB Code/Appendix	[page 10 – 15]

Data Import and Summary Statistics:

Mean, median, mode, range and standard deviation values tabled:

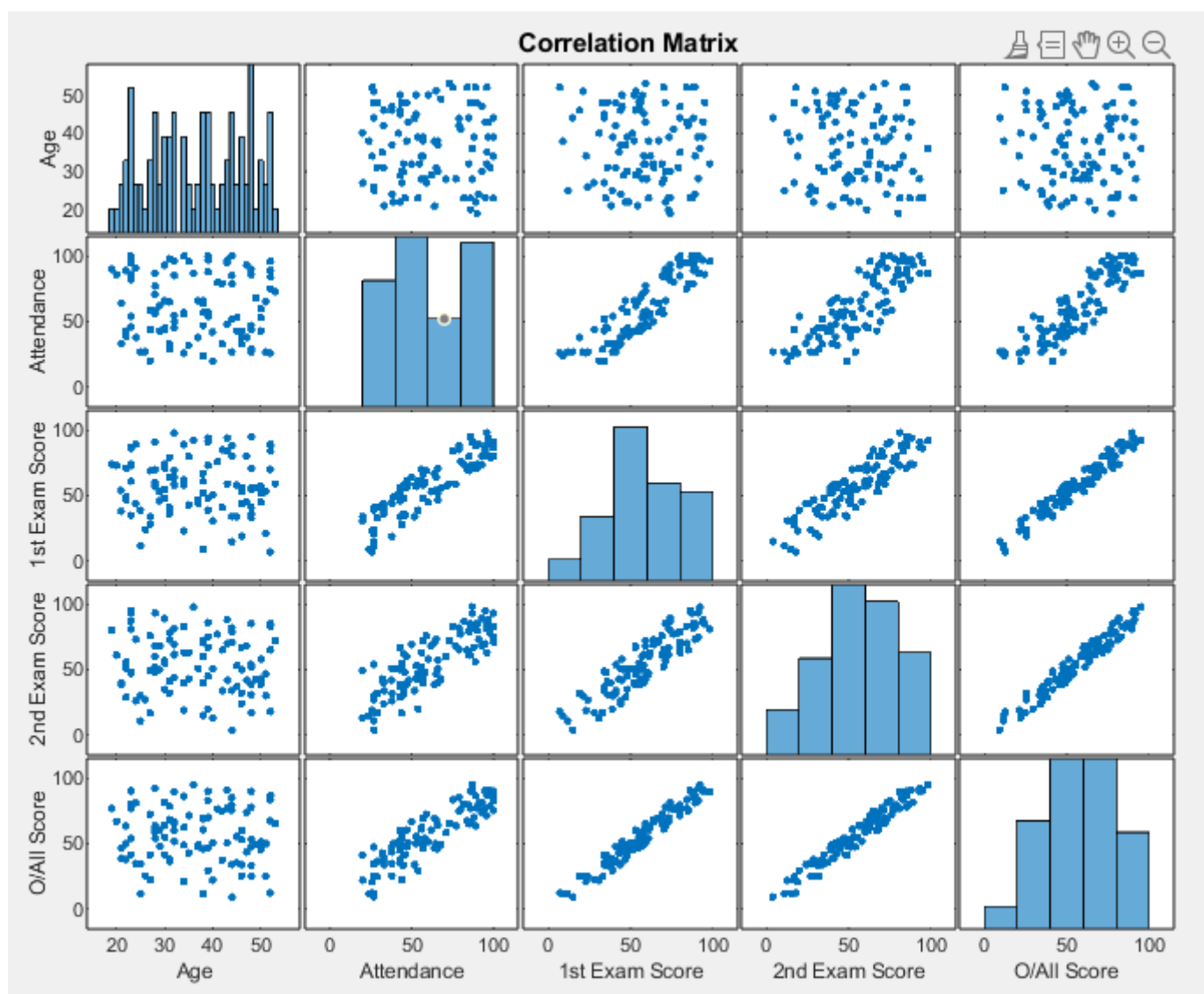
Variable	Mean	Median	Mode	Range	Std
"Age"	36.65	37.5	48	34	9.7643
"Attendance"	60.58	57	27	80	24.249
"1st Exam Score"	58.27	58	59	91	21.673
"2nd Exam Score"	55.08	53.5	51	94	22.428
"O/All Score"	56.675	54.5	25.5	85.5	21.452

Correlation Analysis:

The correlation coefficients:

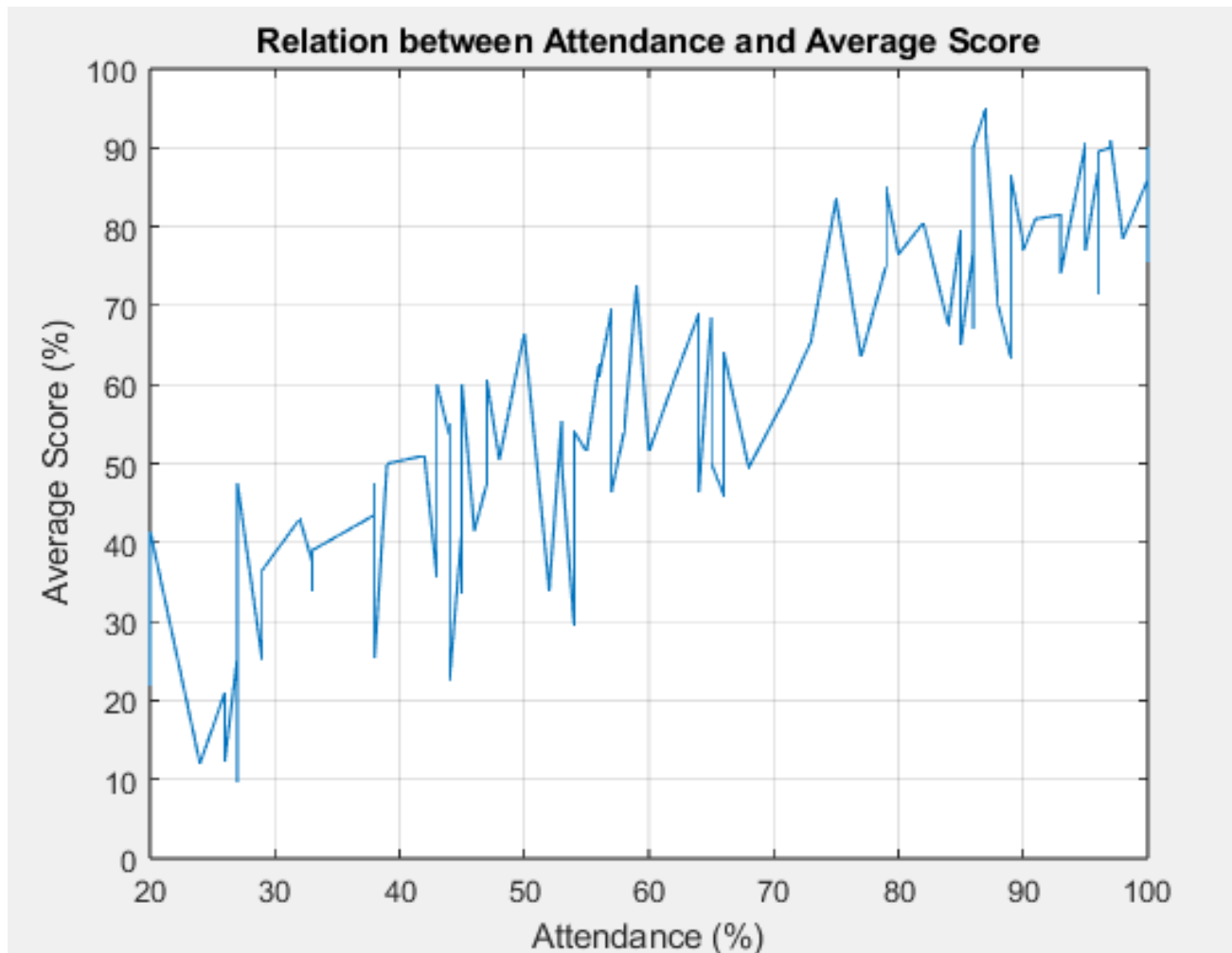
1.0000	-0.0362	-0.0364	-0.0506	-0.0448
-0.0362	1.0000	0.8912	0.8297	0.8839
-0.0364	0.8912	1.0000	0.8929	0.9719
-0.0506	0.8297	0.8929	1.0000	0.9738
-0.0448	0.8839	0.9719	0.9738	1.0000

Plotting the correlation coefficients:



The correlation matrix above shows the relationships between the different variables. From the matrix its clear to see age has no influence over the exam scores and that attendance directly correlates to a higher score in both exams. And obviously, exam 1 score directly correlates with exam score 2.

As mentioned above the key variable that strongly correlates with the exam score is Attendance. Here is a better a look at the relationship between attendance and score:



As you can see, attendance directly correlates with score. The more attendance, the higher the score.

Relationship between gender and overall module results:

```
The ttest2 function failed to reject the null hypothesis that the two data  
samples are from populations with equal means
```

```
The probability of this data reappearing is
```

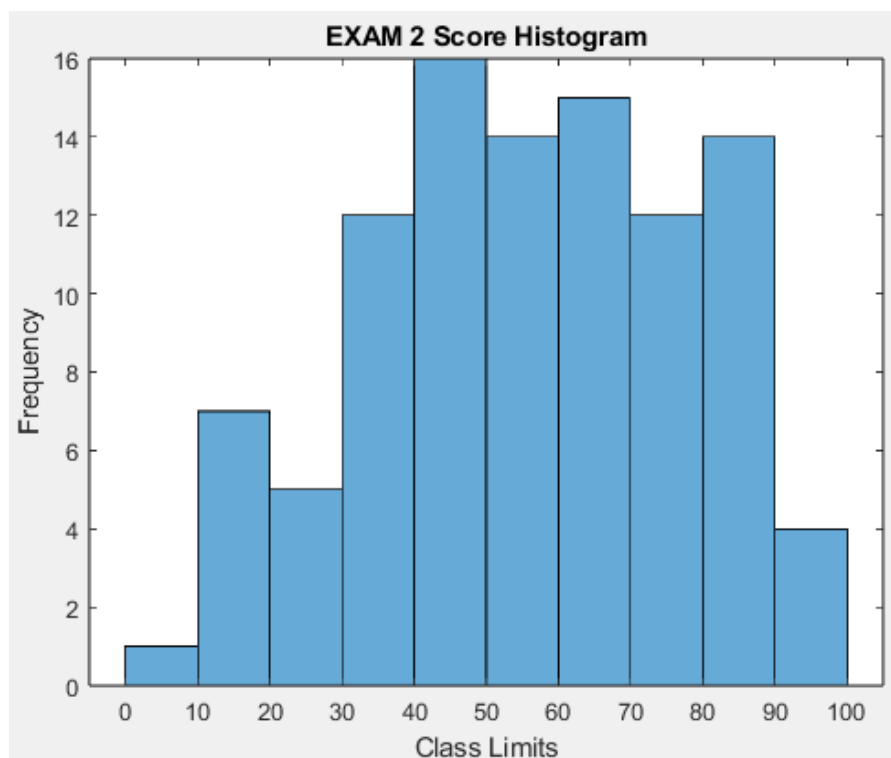
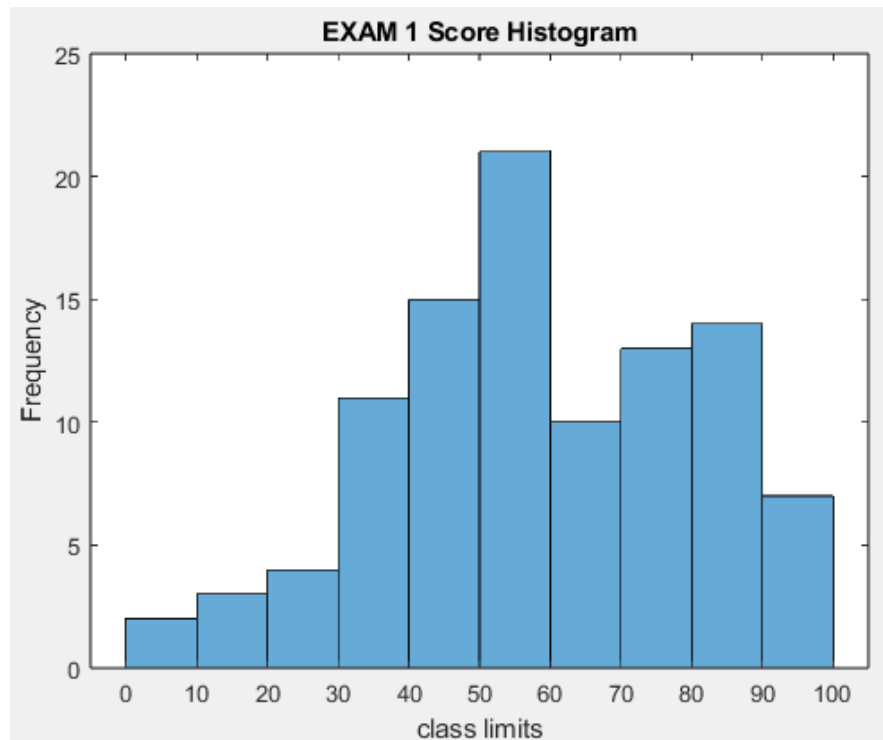
```
43.700617%
```

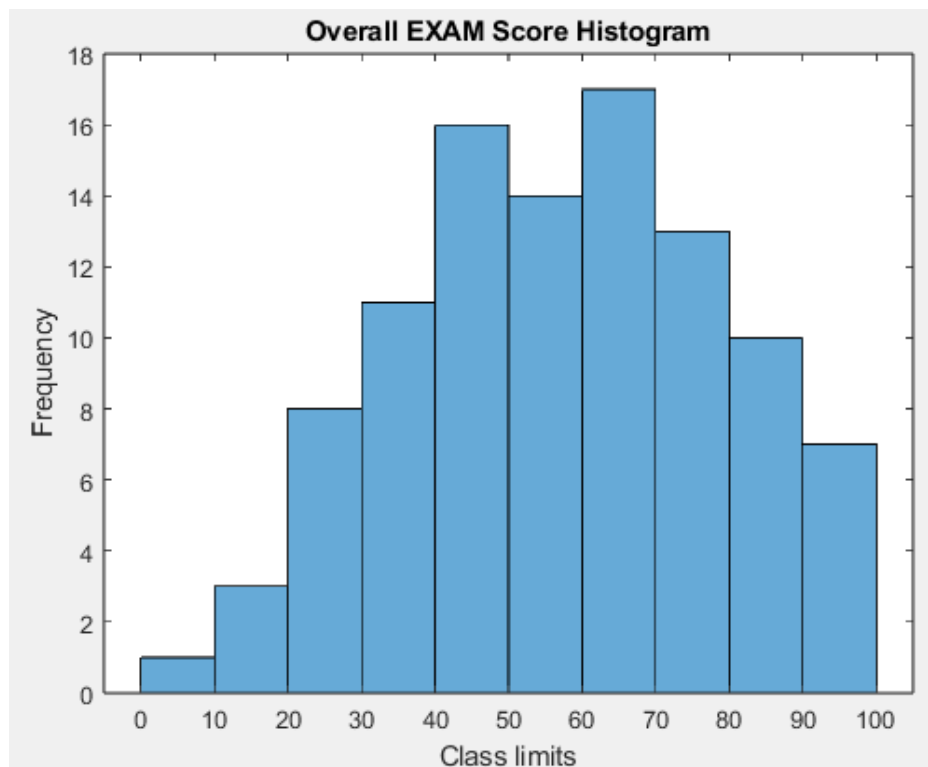
```
This means it is not a random occurrence but doesn't fully represent the  
categorical variables effect on the average score
```

Gender does have an effect on overall results; however, the results don't equal sample sizes, there are more females than males, this further suggests that if the sample sizes were equal in size, the gender would surely have a larger impact on overall score because currently there are less males in the sample pool. Also, since the probability is 43.7%, gender cannot fully predict overall exam score, only an unlikely prediction can be made from it.

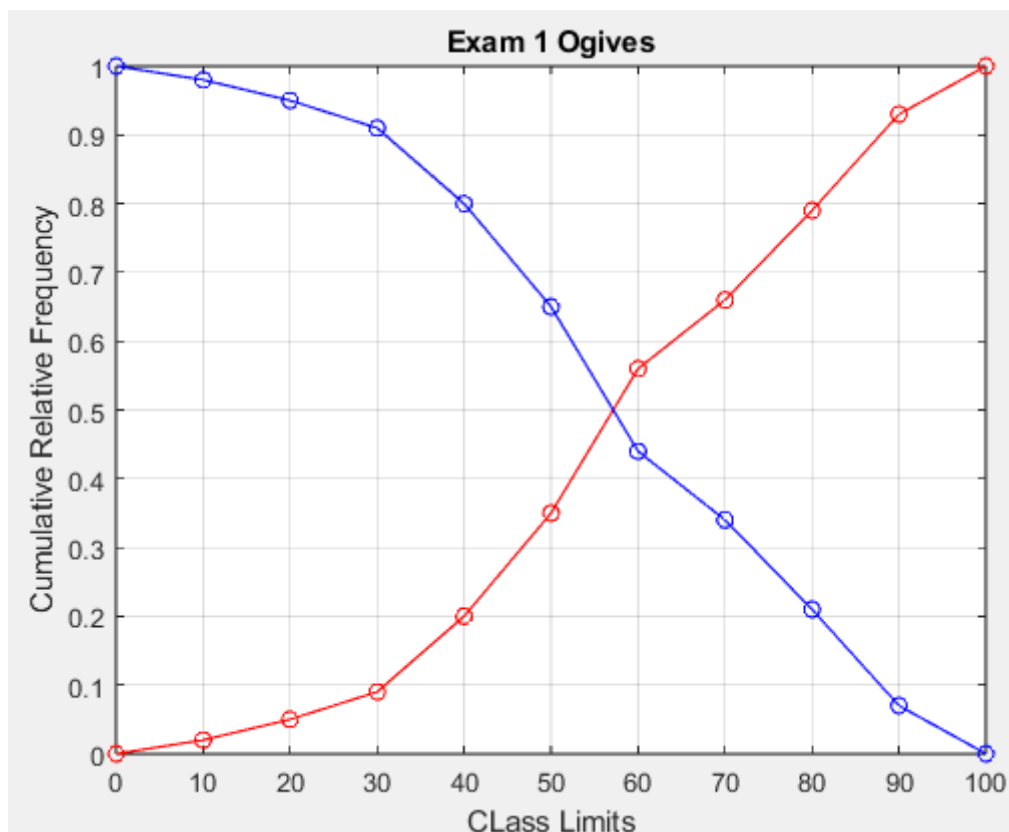
Visualisation of grouped data for Exam 1, Exam 2, and O/ALL results:

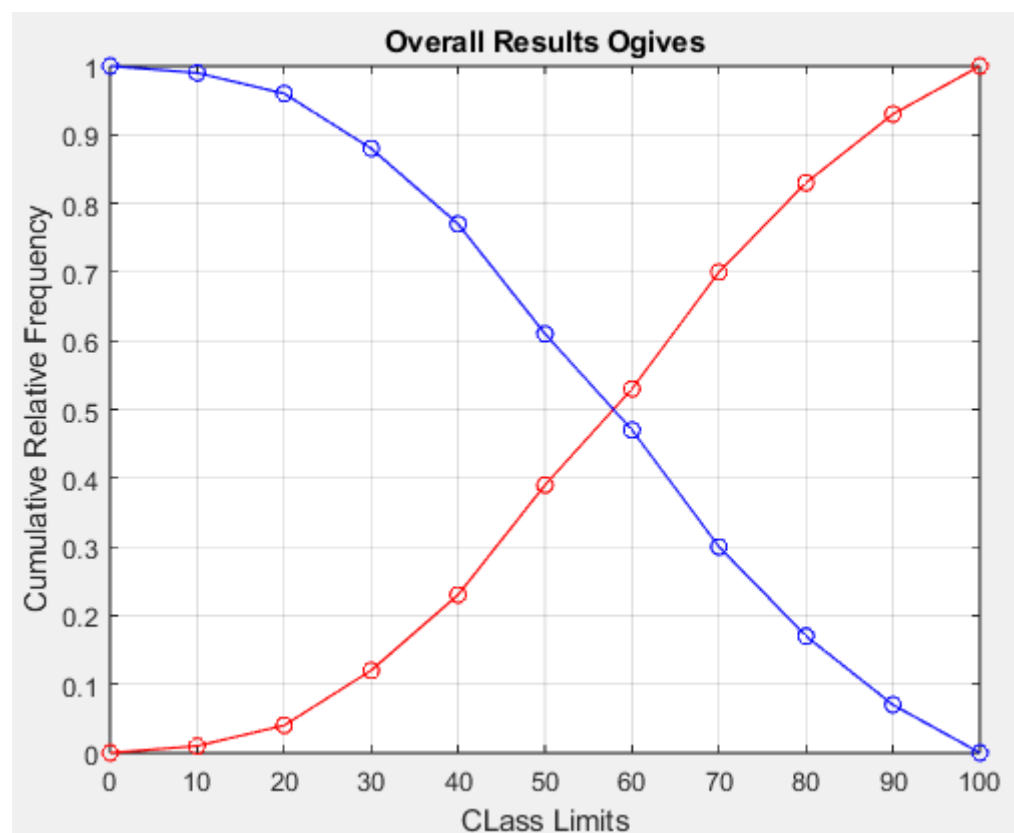
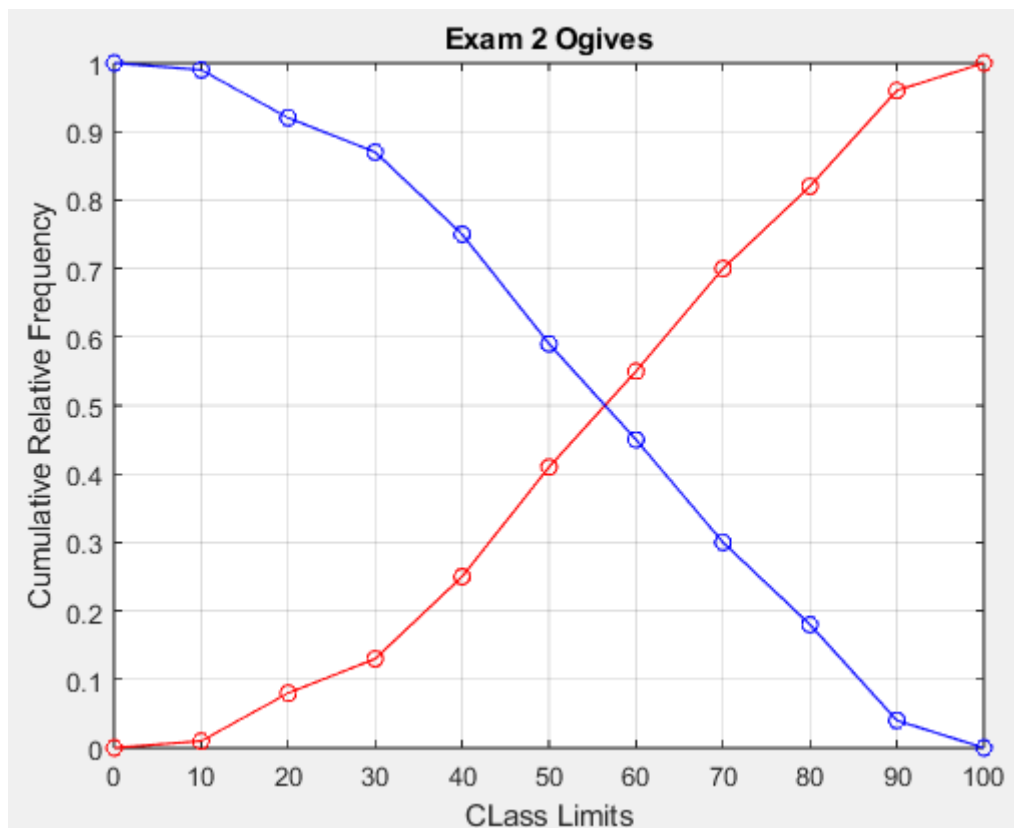
Histograms:





Less than/More than Ogives:





MATLAB Code:

```
%% ANALYSIS OF STUDENT GRADES USING STATISTICAL METHODS IN MATLAB

%% [1] DATA IMPORT AND SUMMARY STATISTICS
clc
clear
close all

% [a]
% Import Data name simplification
Student_DATA = readtable("student_grades.csv");

% Data Values
DATA_VALUES = table2array(Student_DATA(:,{'age', 'attendancePercentage', 'Exam_1', 'Exam_2'}));

% [b]
% Mean, median and mode table data
MEAN_DATA = mean(DATA_VALUES);
MEDIAN_DATA = median(DATA_VALUES);
MODE_DATA = mode(DATA_VALUES);

% Range table data
MINRANGE_DATA = min(DATA_VALUES);
MAXRANGE_DATA = max(DATA_VALUES);
RANGE_DATA = MAXRANGE_DATA - MINRANGE_DATA;

% Standard deviation table data
DEVIATION_DATA = std(DATA_VALUES);

% [c]
% Overall exam score data
% Exam score mean, median and mode
AVERAGESCORE_DATA = (DATA_VALUES(:,3) + DATA_VALUES(:,4))/2;
MEANEXAM_DATA = mean(AVERAGESCORE_DATA);
MEDIANEXAM_DATA = median(AVERAGESCORE_DATA);
MODEEXAM_DATA = mode(AVERAGESCORE_DATA);

% Exam score range
MINRANGEEXAM_DATA = min(AVERAGESCORE_DATA);
MAXRANGEEXAM_DATA = max(AVERAGESCORE_DATA);
RANGEEXAM_DATA = MAXRANGEEXAM_DATA - MINRANGEEXAM_DATA;

% Exam score standard deviation
DEVIATIONEXAM_DATA = std(AVERAGESCORE_DATA);

%Table data and variables
Variable = ["Age";"Attendance";"1st Exam Score";"2nd Exam Score";"O/All Score"];

MEAN_DATA = MEAN_DATA.';
Mean = vertcat(MEAN_DATA,MEANEXAM_DATA);

MEDIAN_DATA = MEDIAN_DATA.';
Median = vertcat(MEDIAN_DATA,MEDIANEXAM_DATA);
```

```
MODE_DATA = MODE_DATA.';
Mode = vertcat(MODE_DATA,MODEEXAM_DATA);

RANGE_DATA = RANGE_DATA.';
Range = vertcat(RANGE_DATA,RANGEEEXAM_DATA);

DEVIATION_DATA = DEVIATION_DATA.';
Std = vertcat(DEVIATION_DATA,DEVIATIONEXAM_DATA);

% Mean, median, mode, range and std value in a table
MMMRS_TABLEDATA = table(Variable,Mean,Median,Mode,Range,Std);

disp(MMMRS_TABLEDATA)

%% [2] CORRELATION ANALYSIS
clc
close all
% [a]
% Linear correlation coefficients
FULL_DATA_VALUES = DATA_VALUES;
FULL_DATA_VALUES(:,5) = AVERAGESCORE_DATA;
CORRELATION_COEFFICIENT = corrcoef(FULL_DATA_VALUES);
disp(CORRELATION_COEFFICIENT)

% [b]
% Plotting correlation graphs

% NOTE: corrplot function isnt available on my version of MATLAB so I have
% attempted a create a similar graphed layout using 'plotmatrix' function

Variable = Variable.';
[~,ax] = plotmatrix(FULL_DATA_VALUES);
title("Correlation Matrix")

iterations = size(ax,1);
for i = 1:iterations
    ax(i,1).YLabel.String = Variable(i);
    ax(iterations,i).XLabel.String = Variable(i);
end

% [c]
% Creating gender and results table
GENDER_DATA = table2array(Student_DATA(:,{'gender'}));
GENDER_DATA_TABLE = table(GENDER_DATA,AVERAGESCORE_DATA);

% Results sorted
GENDER_DATA_TABLE_SORT = sortrows(GENDER_DATA_TABLE,1);
% 53 Female : 47 Male

% Results separated by gender
MALE_RESULTS = GENDER_DATA_TABLE_SORT(54:100,:);
FEMALE_RESULTS = GENDER_DATA_TABLE_SORT(1:53,:);

MALE_RESULTS_DATA = table2array(MALE_RESULTS(:,2));
FEMALE_RESULTS_DATA = table2array(FEMALE_RESULTS(:,2));
```

```
MEAN_MALE_DATA = mean(MALE_RESULTS_DATA);
MEAN_FEMALE_DATA = mean(FEMALE_RESULTS_DATA);

[h,p] = ttest2(MALE_RESULTS_DATA,FEMALE_RESULTS_DATA);

if h == 0
    disp(['The ttest2 fucntion failed to reject the null hypothesis that the two data' newline
'samples are from populations with equal means'])
else
    disp(['The ttest2 fucntion successfully rejected the null hypothesis that the' newline 'two
data samples are from populations with equal means'])
end

ppercent = p*100;
fprintf('\nThe probability of this data reappearing is\n\n%4f%%\n\n',pppercent);
disp(['This means it is not a random occurrence but doesn't fully represent the' newline
'categorical variables effect on the average score']);

% [d]
% Variable effecting score
% Age?
AGE_DATA = table2array(Student_DATA(:,{'age'}));
AGE_DATA_TABLE = table(AGE_DATA,AVERAGESCORE_DATA);

AGE_DATA_TABLE_SORT = sortrows(AGE_DATA_TABLE,1);
% To break up the age ranges
% 19 - 25 (1:17)
% 26 - 30 (18:32)
% 31 - 40 (33:62)
% 41+ (63:100)

AGE1925_RESULTS = table2array(AGE_DATA_TABLE_SORT(1:17,2));
AGE2630_RESULTS = table2array(AGE_DATA_TABLE_SORT(18:32,2));
AGE3140_RESULTS = table2array(AGE_DATA_TABLE_SORT(33:62,2));
AGE41PLUS_RESULTS = table2array(AGE_DATA_TABLE_SORT(63:100,2));

MEAN1925_DATA = mean(AGE1925_RESULTS);
MEAN2630_DATA = mean(AGE2630_RESULTS);
MEAN3140_DATA = mean(AGE3140_RESULTS);
MEAN41PLUS_DATA = mean(AGE41PLUS_RESULTS);

% Age has little effect on exam score

% Attendance?
ATTEN_DATA = table2array(Student_DATA(:,{'attendancePercentage'}));
ATTEN_DATA_TABLE = table(ATTEN_DATA,AVERAGESCORE_DATA);

ATTEN_DATA_TABLE_SORT = sortrows(ATTEN_DATA_TABLE,1);

figure('name','Score in relation to Attendance');
ATTENDANCE_SORTED = table2array(ATTEN_DATA_TABLE_SORT(:,1));
AVERAGESCORE_ATTENSORT = table2array(ATTEN_DATA_TABLE_SORT(:,2));
plot(ATTENDANCE_SORTED,AVERAGESCORE_ATTENSORT)

xlabel("Attendance (%)")
ylabel("Average Score (%)")
grid on
title("Relation between Attendance and Average Score")
```

```
hold off
```

```
%% [3] Visualisation of grouped data for Exam 1, Exam 2, and O/ALL results:  
clc
```

```
% [a]  
% Using histcounts  
HISTCOUNT_EXAM1 = histcounts(FULL_DATA_VALUES(:,3),10);  
HISTCOUNT_EXAM2 = histcounts(FULL_DATA_VALUES(:,4),10);  
HISTCOUNT_OVERALL = histcounts(FULL_DATA_VALUES(:,5),0:10:100);
```

```
% [b]  
% Using histograms  
% Histogram 1  
figure('name','HISTOGRAM_EXAM1');  
HISTOGRAM_EXAM1 = histogram(FULL_DATA_VALUES(:,3),10);  
xlabel("class limits");  
ylabel("Frequency");  
title("EXAM 1 Score Histogram")
```

```
HANDLE_HISTOGRAM_EXAM1 = ancestor(HISTOGRAM_EXAM1,'figure');  
HANDLE_HISTOGRAM_EXAM1.Visible = 'on';
```

```
% Histogram 2  
figure('name','HISTOGRAM_EXAM2');  
HISTOGRAM_EXAM2 = histogram(FULL_DATA_VALUES(:,4),10);  
xlabel("Class Limits");  
ylabel("Frequency");  
title("EXAM 2 Score Histogram")
```

```
HANDLE_HISTOGRAM_EXAM2 = ancestor(HISTOGRAM_EXAM2,'figure');  
HANDLE_HISTOGRAM_EXAM2.Visible = 'on';
```

```
% Histogram 3  
figure('name','HISTOGRAM_OVERALL');  
HISTOGRAM_OVERALL = histogram(FULL_DATA_VALUES(:,5),0:10:100);  
xlabel("Class limits");  
ylabel("Frequency");  
title("Overall EXAM Score Histogram")
```

```
HANDLE_HISTOGRAM_OVERALL = ancestor(HISTOGRAM_OVERALL,'figure');  
HANDLE_HISTOGRAM_OVERALL.Visible = 'on';
```

```
% [c]  
% Plotting Ogives  
BOUNDARIES = [0 10 20 30 40 50 60 70 80 90 100];
```

```
% Exam 1 ogive  
figure('name','Exam 1 Ogives');
```

```
EXAM1_RFREQUENCY = HISTCOUNT_EXAM1/100;
```

```
for i = 2:1:10
```

```
    EXAM1_RFREQUENCY(i) = EXAM1_RFREQUENCY(i) + EXAM1_RFREQUENCY(i - 1);  
end
```

```
EXAM1_CUMFREQUENCY = zeros(11,1);
EXAM1_CUMFREQUENCY(2:11) = EXAM1_RFREQUENCY;

EXAM1_LESSOGIVE = EXAM1_CUMFREQUENCY;
EXAM1_MOREOGIVE = ones(11,1) - EXAM1_CUMFREQUENCY;

plot(BOUNDARIES,EXAM1_LESSOGIVE,'r-o')

grid on
title("Exam 1 Ogives")

hold on

plot(BOUNDARIES,EXAM1_MOREOGIVE,'b-o')

xlabel("Class Limits")
ylabel("Cumulative Relative Frequency")

hold off

% Exam 2 ogive
figure('name','Exam 2 Ogives');

EXAM2_RFREQUENCY = HISTCOUNT_EXAM2/100;

for i = 2:1:10

    EXAM2_RFREQUENCY(i) = EXAM2_RFREQUENCY(i) + EXAM2_RFREQUENCY(i - 1);
end

EXAM2_CUMFREQUENCY = zeros(11,1);
EXAM2_CUMFREQUENCY(2:11) = EXAM2_RFREQUENCY;

EXAM2_LESSOGIVE = EXAM2_CUMFREQUENCY;
EXAM2_MOREOGIVE = ones(11,1) - EXAM2_CUMFREQUENCY;

plot(BOUNDARIES,EXAM2_LESSOGIVE,'r-o')

grid on
title("Exam 2 Ogives")

hold on

plot(BOUNDARIES,EXAM2_MOREOGIVE,'b-o')

xlabel("Class Limits")
ylabel("Cumulative Relative Frequency")

hold off

% Exam 3 ogive
figure('name','Exam 3 Ogives');

EXAM3_RFREQUENCY = HISTCOUNT_OVERALL/100;

for i = 2:1:10
```

```
EXAM3_RFREQUENCY(i) = EXAM3_RFREQUENCY(i) + EXAM3_RFREQUENCY(i - 1);  
end  
  
EXAM3_CUMFREQUENCY = zeros(11,1);  
EXAM3_CUMFREQUENCY(2:11) = EXAM3_RFREQUENCY;  
  
EXAM3_LESSOGIVE = EXAM3_CUMFREQUENCY;  
EXAM3_MOREOGIVE = ones(11,1) - EXAM3_CUMFREQUENCY;  
  
plot(BOUNDARIES, EXAM3_LESSOGIVE, 'r-o')  
  
grid on  
title("Overall Results Ogives")  
  
hold on  
  
plot(BOUNDARIES, EXAM3_MOREOGIVE, 'b-o')  
  
xlabel("Class Limits")  
ylabel("Cumulative Relative Frequency")  
  
hold off
```