DSL810-Assignment 4- Statistical DATA

5 FACULTIES vs 2 Category of Students and 10 Gradings

1) For sets of marks Provided to the post graduate students (of non-engg. and engg. orientation) for a project review by different faculty members of a design department of an institute.

From your analysis of the dataset, do you find any statistically significant difference between the marks given by different faculty to the students? What would you comment on the marks received by the students of engg. vs. non-engg. orientation? Is there any difference between the marks given by faculty to engg. vs. non-engg. students?

Can you compare two faculty members at a time to find out for which faculty members you find statistically significant differences in grading the students? Please also compute the Power (probability of avoiding a Type II error) of any of these comparisons and find out the sample size for which the Power >= 90%.

Please share any other insights that you have for this analysis. What could be other unknown factors, if any, if considered could have made the analysis more meaningful?

Please visualize this data using box plots or other means. Please show the calculations/ upload the code and/or the ANOVA tables involved.

2) Can you think of other real world applications of the Statistical Methods learnt in the class? Please outline any five of them.

With the Provided Data Sets, We started Analysing using MATLAB.

Although most of the assignment part are preformed in MATLAB, the data sets have been simultaneous evaluated in Excel as well for validation purpose.

```
global abspath datapath scriptpath datasetspath;
%path of the folder
abspath='X:\downloads\DSL819Assignment-4';
%path for the raw data
datapath=strcat(abspath,'\data\');

%path where the scripts are kept
scriptpath=strcat(abspath,'\scripts\');

%path where the results are kept
datasetspath=strcat(abspath,'\datasets\');
```

At first we converted data sets table to Arrays

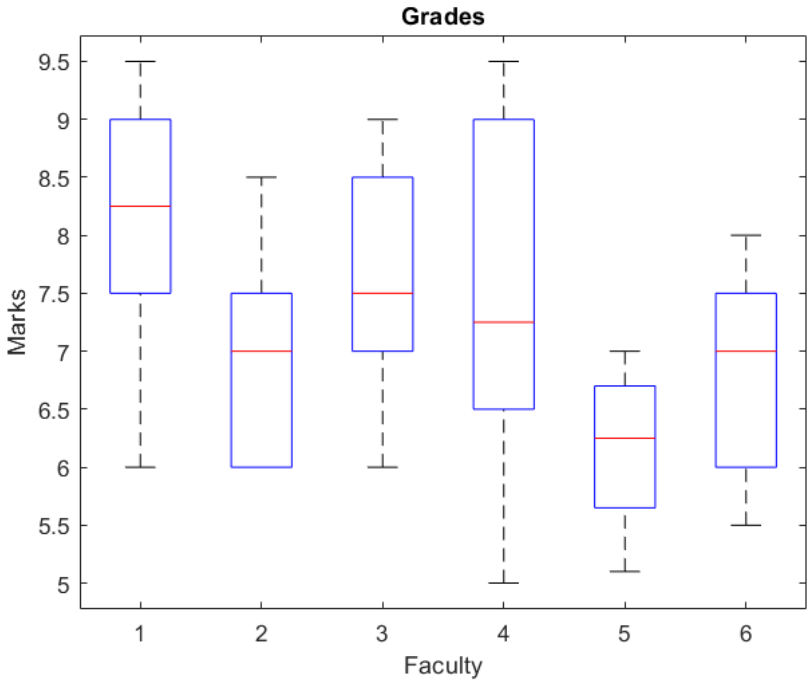
Convert Table values into Arrays

```
grades= DSLassignment4{:,2:7};
import =grades
```

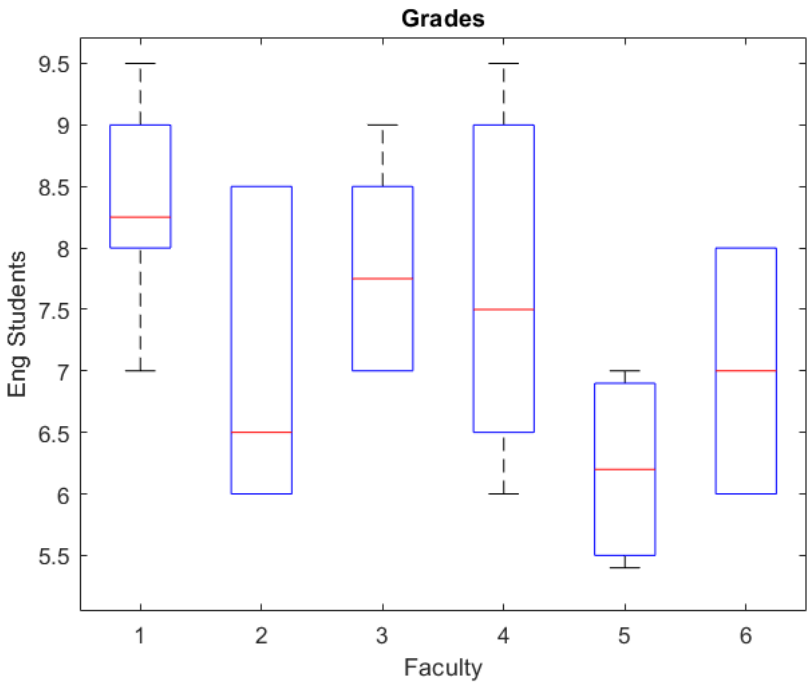
```
import = 20×6
    8.5000    7.0000    7.0000    7.0000    6.5000    8.0000
    8.5000    8.0000    7.0000    9.0000    6.6000    6.5000
    9.0000    7.0000    9.0000    9.5000    6.9000    7.0000
    8.5000    7.5000    8.0000    7.5000    5.8000    7.5000
    7.0000    7.5000    7.5000    6.5000    5.8000    6.5000
    6.0000    7.5000    7.5000    5.0000    5.1000    5.5000
    9.0000    7.5000    8.0000    8.0000    6.5000    7.0000
    7.5000    7.5000    7.0000    6.5000    6.0000    6.5000
    7.0000    6.8000    6.0000    9.0000    6.5000    6.0000
    8.0000    6.0000    8.5000    7.0000    5.4000    6.0000
      ⋮
```

Ploting BOX plots to Understand the relative Data overall behaviour,

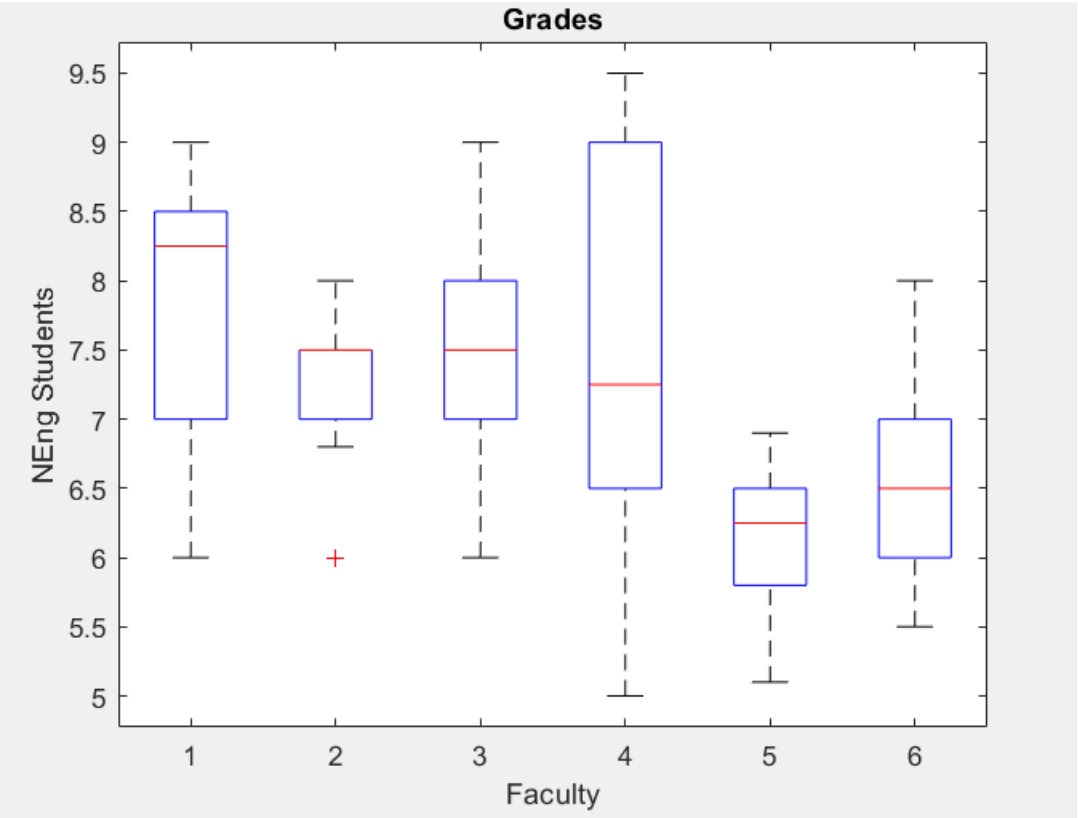
```
boxplot(grades);title('Grades');ylabel('Marks');xlabel('Faculty')
```




```
boxplot(EngMarks);title('Grades');ylabel('Eng Students');xlabel('Faculty');
```




```
boxplot(NEngMarks);title('Grades');ylabel('NEng Students');xlabel('Faculty');
```



From above plots it is very much clear that Faculty 2 has huge variation in evaluating students compared to other faculties.

Engg. Students scored better than Non-Engg. Students. Exception Faculty 2.

Faculty 4 is mostly graded high and Faculty 5 is least graded.

Performing Annova using Matlab

Matlab does whole calculations; Similar OneWay Anova is Verified using Excel as well.

```
[~,~,~] = anova2(grades,10)
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	48.634	5	9.72688	11.01	0
Rows	0.954	1	0.95408	1.08	0.3011
Interaction	1.8	5	0.36008	0.41	0.8427
Error	95.449	108	0.88379		
Total	146.838	119			

By Setting Significance levels as 0.05

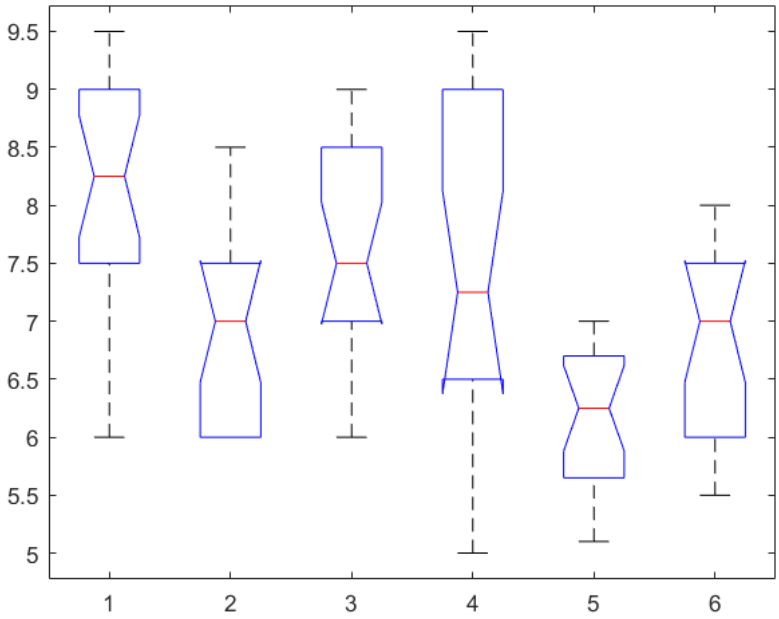
- The p-value - Interaction Between Faculty & Students is **0.8427**. This is not statistically significant at alpha level 0.05. This implies not any significant difference in marks given by faculty to Engineering & Non Engineering Students
- The p-value - For Eng Stud. vs Non-eng. is **0.3011**. This is not statistically significant at alpha level 0.05. This implies not any significant differences in marks provided to Eng. and Non-eng. students.
- The p-value for marks given by faculty is **0.** This is statistically significant at alpha level 0.05. This implies there is difference in the marks provided by faculties.

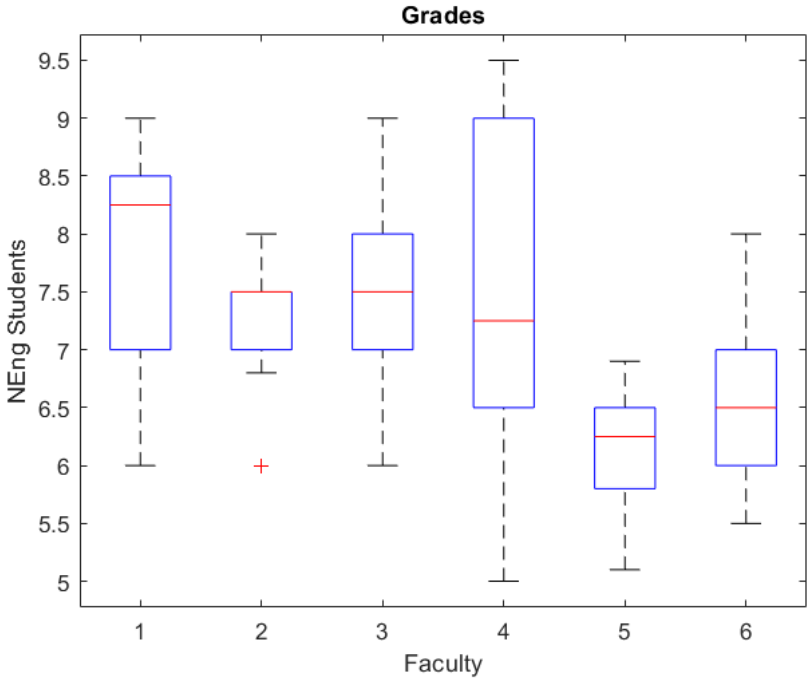
Oneway Annova was tried over the data to visualize the effects.

```
[~,tbl,stats]=anova1(grades)
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	48.634	5	9.72688	11.29	7.38558e-09
Error	98.204	114	0.86143		
Total	146.838	119			





tbl = 4×6 cell

	1	2	3	4	5	6
1	'Source'	'SS'	'df'	'MS'	'F'	'Prob>F'
2	'Columns'	48.6344	5	9.7269	11.2915	7.3856e-09
3	'Error'	98.2035	114	0.8614	[ ]	[ ]
4	'Total'	146.8379	119	[ ]	[ ]	[ ]

stats = *struct with fields:*
    gnames: [6×1 char]

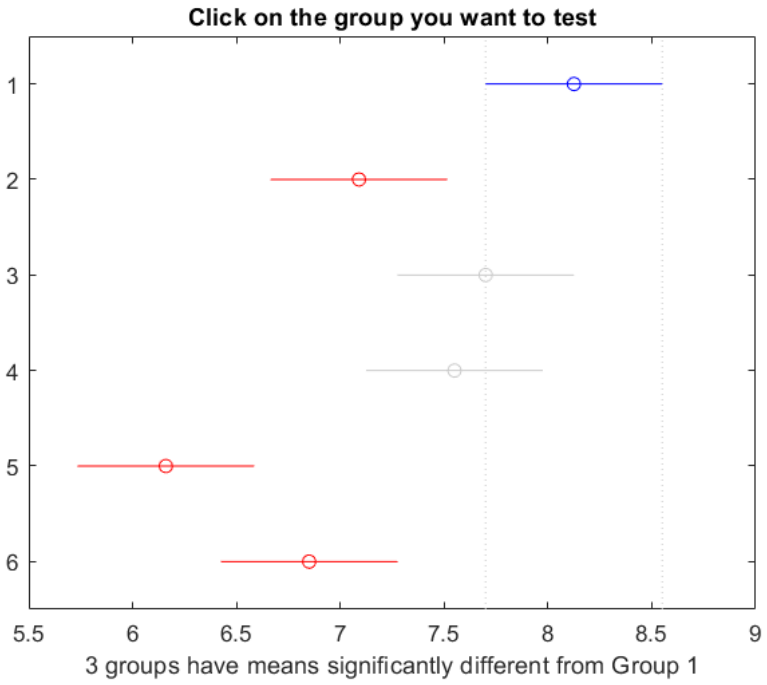
```
      n: [20 20 20 20 20 20]
 source: 'anova1'
  means: [8.1250 7.0900 7.7000 7.5500 6.1600 6.8500]
     df: 114
      s: 0.9281
```

Thats the stats above for the individual Groups..

The Complete DATA is compared withe Eachother

Now, we can compare point to point differences using Multcompare

```
[c,m,h,nms]=multcompare(stats)
```




```
c = 15×6
    1.0000    2.0000    0.1842    1.0350    1.8858    0.0078
    1.0000    3.0000   -0.4258    0.4250    1.2758    0.6976
    1.0000    4.0000   -0.2758    0.5750    1.4258    0.3723
    1.0000    5.0000    1.1142    1.9650    2.8158    0.0000
    1.0000    6.0000    0.4242    1.2750    2.1258    0.0004
    2.0000    3.0000   -1.4608   -0.6100    0.2408    0.3060
    2.0000    4.0000   -1.3108   -0.4600    0.3908    0.6216
    2.0000    5.0000    0.0792    0.9300    1.7808    0.0235
    2.0000    6.0000   -0.6108    0.2400    1.0908    0.9637
    3.0000    4.0000   -0.7008    0.1500    1.0008    0.9957
      ⋮
m = 6×2
    8.1250    0.2075
    7.0900    0.2075
    7.7000    0.2075
    7.5500    0.2075
```

```
    6.1600    0.2075
    6.8500    0.2075
h =
  Figure (boxplot) with properties:

      Number: 4
        Name: 'Multiple comparison of means'
       Color: [1 1 1]
    Position: [489 143 560 420]
       Units: 'pixels'

  Show all properties
nms = 6×1 char array
    '1'
    '2'
    '3'
    '4'
    '5'
    '6'
```

From the above Comparison, it is very Clear,

Faculty F1 & F2, F1 & F5 , F1 & F6, F2 &F5, F3 & F5, F4 & F5 have Differences in Grades provided.

For comparison of two faculties, F2 vs F6 and F1 vs F5 is considered for Power Calculation;

Power (probability of avoiding a Type II error)

Ideally t2 approach has to be implemented in getting the power & Sample Sizes,

However, t and z power & sample size calculations are also performed to visualize the effects.

```
pwrt = sampsizepwr ("t",[0 5.324],0.24, [],20)
```

```
pwrt = 0.0542
```

```
pwrt2=sampsizepwr("t2",[0 5.324],0.24, [],20)
```

```
pwrt2 = 0.0522
```

```
pwrz = sampsizepwr("z",[0 5.324],0.24, [],20)
```

```
pwrz = 0.0547
```

Sample size for which the Power >= 90%

```
sampt = sampsizepwr("t",[0 5.324], 0.24, 0.900, [])
```

```
sampt = 5173
```

```
sampt2 = sampsizepwr("t2",[0 5.324], 0.24, 0.900, [])
```

```
sampt2 = 10343
```

```
sampz = sampsizepwr("Z",[0 5.324], 0.24, 0.900, [])
```

```
sampz = 5171
```

If suppose, the normality functions are Not valid

using kruskal-walis method

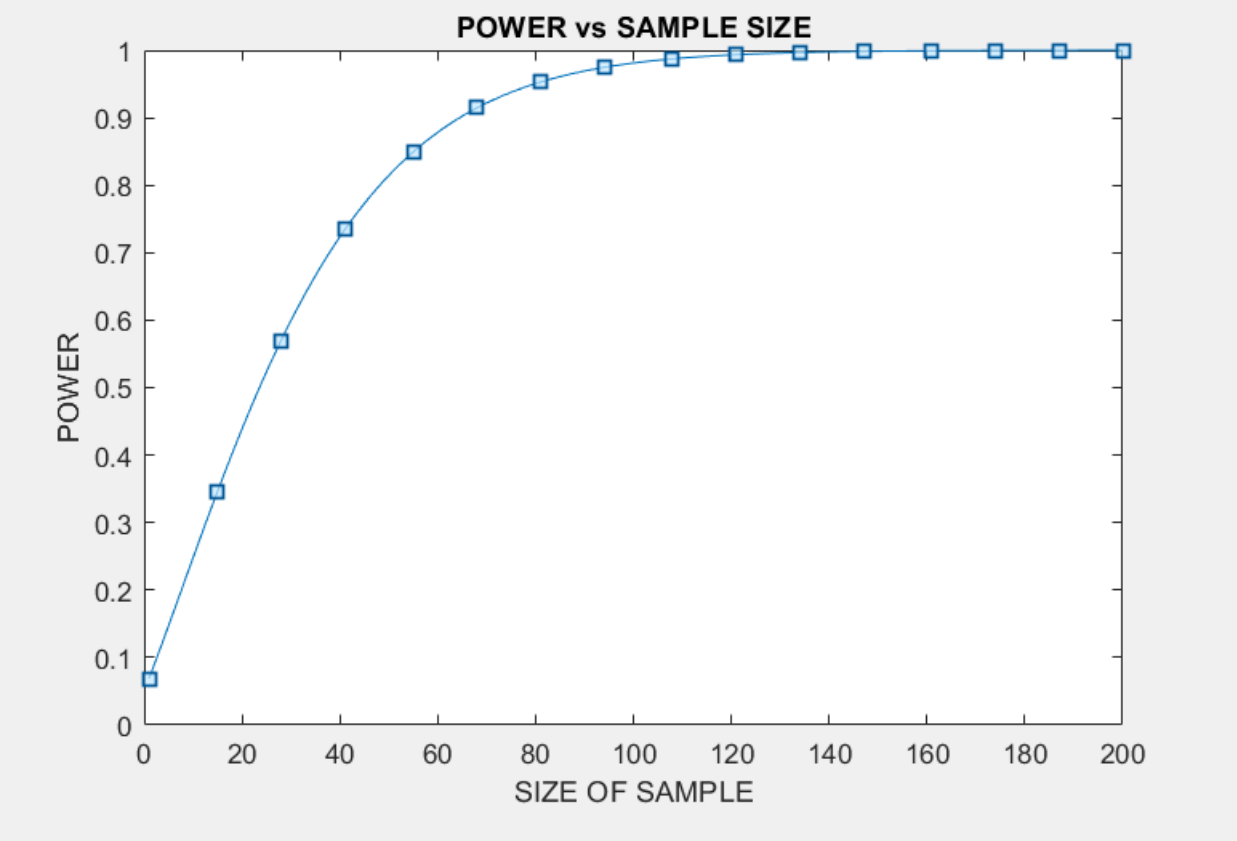
```
pwrt = sampsizepwr ("t",[0 4.88],1.97, [],20)
```

```
pwrt = 0.4030
```

```
pwrt2=sampsizepwr("t2",[0 4.88],1.97, [],20)
```

```
pwrt2 = 0.2378
```

```
pwrz = sampsizepwr("z",[0 4.88],1.97, [],20);
figure;
plot(n,pwrz);
title('POWER vs SAMPLE SIZE');
xlabel('SIZE OF SAMPLE');
ylabel("POWER");
```



Sample size for which the Power >= 90%

```
sampt = sampsizepwr("t",[0 4.88],1.97,0.9,[])
```

```
sampt = 67
```

```
sampt2 = sampsizepwr("t2",[0 4.88],1.97,0.9,[])
```

```
sampt2 = 130
```

```
sampz = sampsizepwr("Z",[0 4.88],1.97, 0.9,[])
```

```
sampz = 65
```

We Can Infact Plot the Power over number of Sample Plots.

For F2 & F6 10000+ sample size is required, however for F1& F5 65 sample size is enough as per Z method and 130 sample size for t2 approach for 90% of Power.

```

```

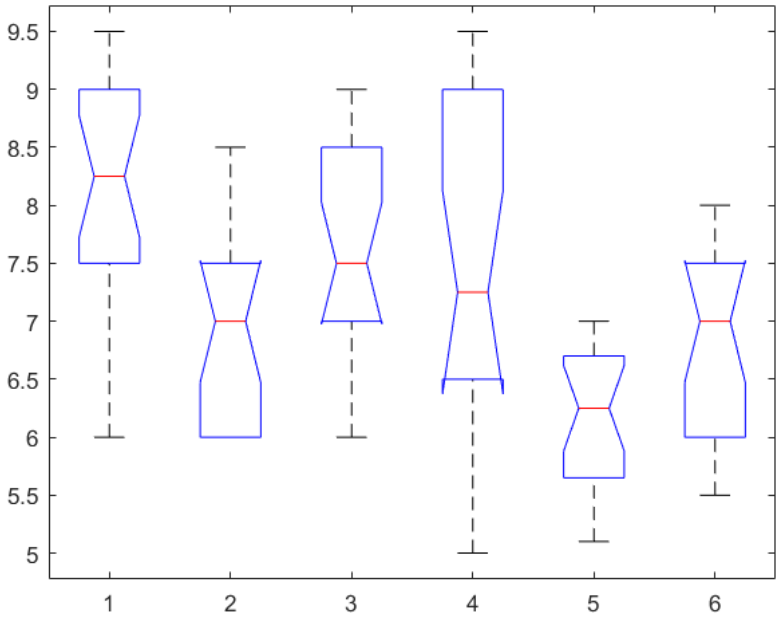
Suppose, if the normality functions are Not valid

using **kruskal-walis** method is used for understanding purpose.

```
p=kruskalwallis(grades)
```

Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	51561.8	5	10312.4	43.25	3.28439e-08
Error	90299.7	114	792.1		
Total	141861.5	119			




```
p = 3.2844e-08
```

The Chi-Sq is 43.35 whereas the Prob>Chi-sq is very negligible.

This assignment gives the complete exposure to Statistical Knacks & tricks which will be handy in handling other real time problems.

REAL WORLD APPLICATIONS OF LESSONS LEARNT

1) Google Ads:

   Google uses the statistics to take survey of number of peoples clicking on the ad links prompted in their webpage based on which they are providing relevant data links to the searches or clicks you click. Many online webpages works on this basis of statistics. Survey are conducted. (Source)

2) Exit Polls:

   Most of the elections exit polls are conducted with the help of Statistic and sampling data. In recent days, the analogy used by statistic experts predicts the Election results most accurately. Recent exit polls are almost 90-95% accurate. And also used in share markets to predict the direction of stocks flow.

3) Weather Forecast:

   The Meteorology Department all over the world have become much more sophisticated using scientific Statistics. The data is retrieved from various sources all over the places, satellites, ground sensors, radars etc and all are handled statistically in a structured way and predictions nowadays are completely flawless.

4) Air pollution:

   Different AQI monitoring tools are located all over the places or zones exposed to Air Pollution. And from the data collected, the Statistical tools were used to interpret and model air pollution data. The prediction includes, overall different component of air particles, like No2, So2, PM10, Pm2.5 etc..and respective effects on human health hazard and global environment. Necessary precautions and cautionary advices are formulated well ahead of critically bad Air Quality Index, using statistical tools. And awareness is created.

5) Covid-19 Vaccine Trials & pandemic situation Analysis.

   Using the statistical data, the propagation or status of Pandemic in all over the world are analysed time to time. And in the lines of preparation & administrating of vaccines, the different combinations of component structures and its effects are also numerically evaluated based on the statistics of the behaviour of certain drugs over the period of time.