MONASH University

Information Technology

FIT1006
Business Information Analysis

Lecture 6
Statistical Analysis and Report Writing

Topics covered:

- Statistical Analysis
  - Visualising data
  - Using appropriate statistics
  - Describing data
- Report Writing
  - The structure of a report
  - Figures and tables
  - Citing references

Motivating problem…

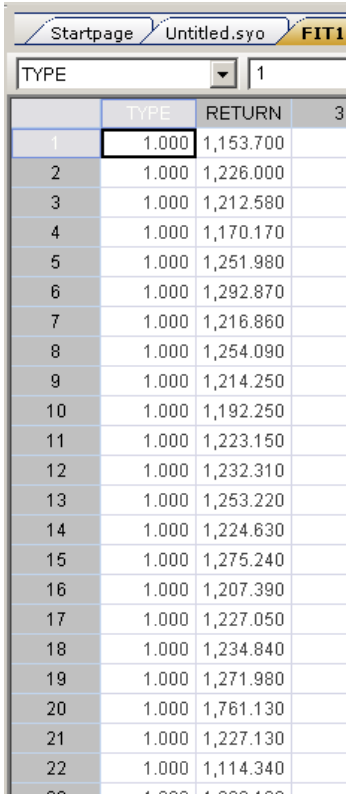
- A grocery store wants you to analyse the amount spent by their customers. They also think there might be different types of customers. They have given you the sales history of 10 randomly sampled customers.

- The same problem…

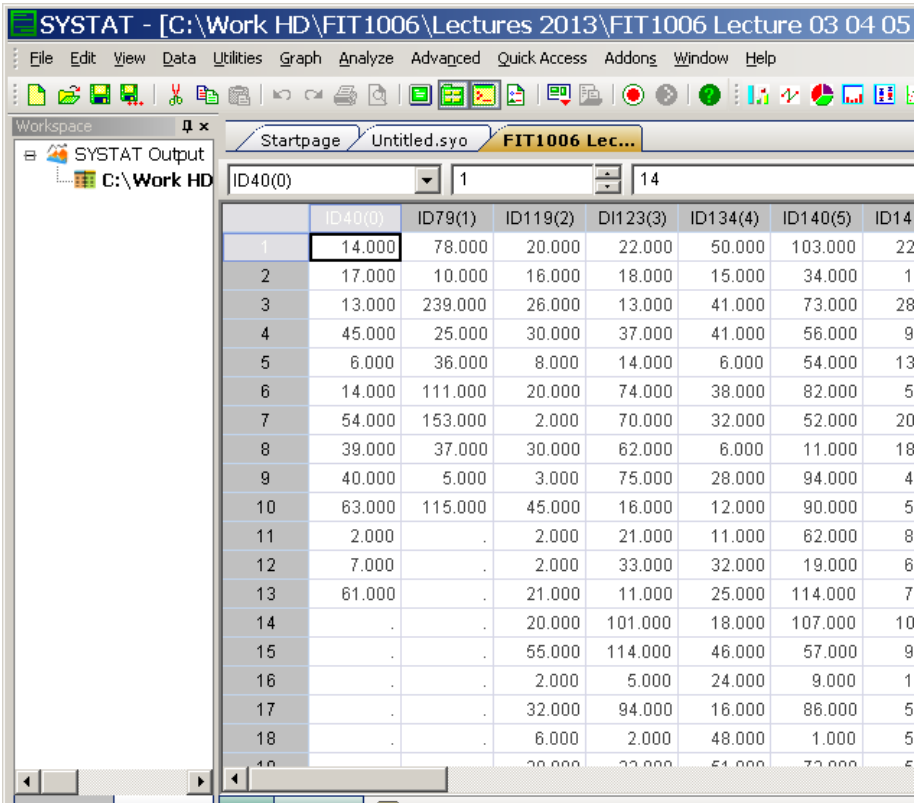
  … but now you can use the computer…

Data

Generate data. You can leave as indexed rows (LHS) but easier if you put in individual rows (RHS).

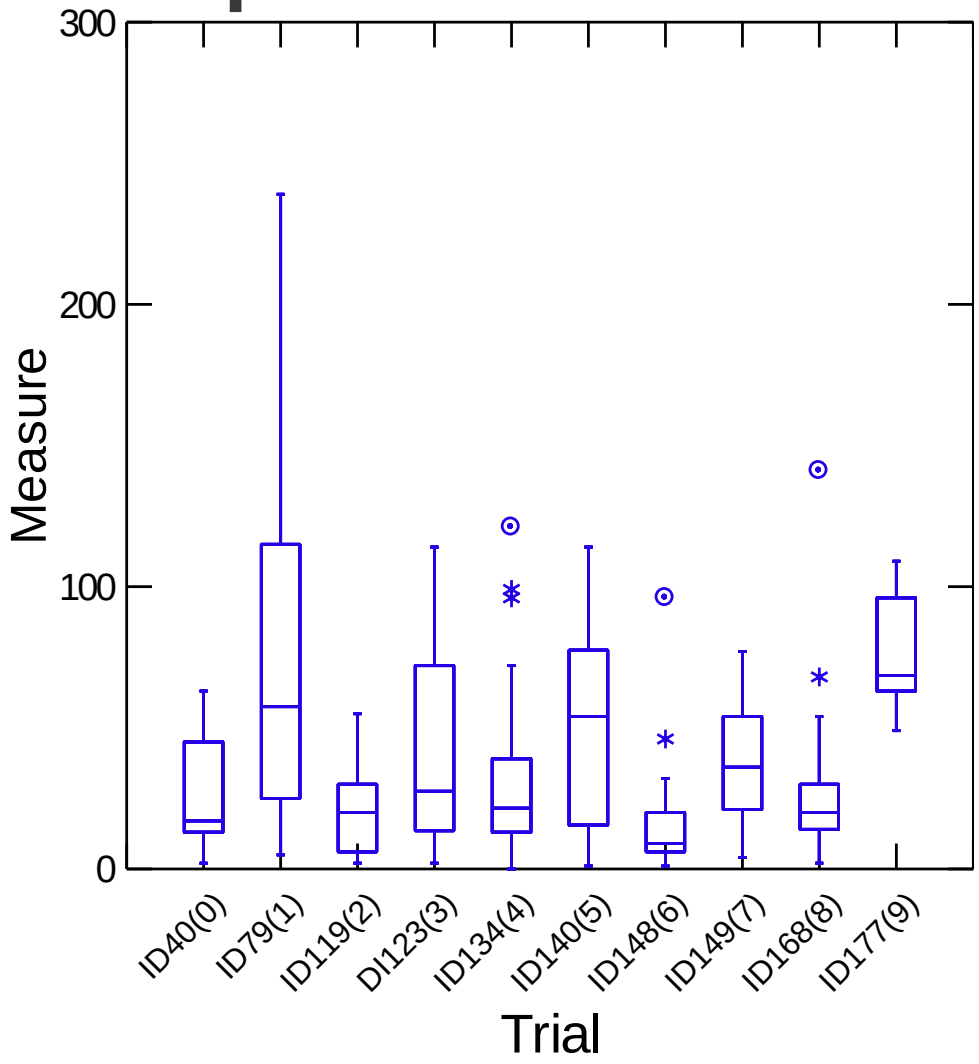


	TYPE	RETURN	3
1	1.000	1,153.700	
2	1.000	1,226.000	
3	1.000	1,212.580	
4	1.000	1,170.170	
5	1.000	1,251.980	
6	1.000	1,292.870	
7	1.000	1,216.860	
8	1.000	1,254.090	
9	1.000	1,214.250	
10	1.000	1,192.250	
11	1.000	1,223.150	
12	1.000	1,232.310	
13	1.000	1,253.220	
14	1.000	1,224.630	
15	1.000	1,275.240	
16	1.000	1,207.390	
17	1.000	1,227.050	
18	1.000	1,234.840	
19	1.000	1,271.980	
20	1.000	1,761.130	
21	1.000	1,227.130	
22	1.000	1,114.340	

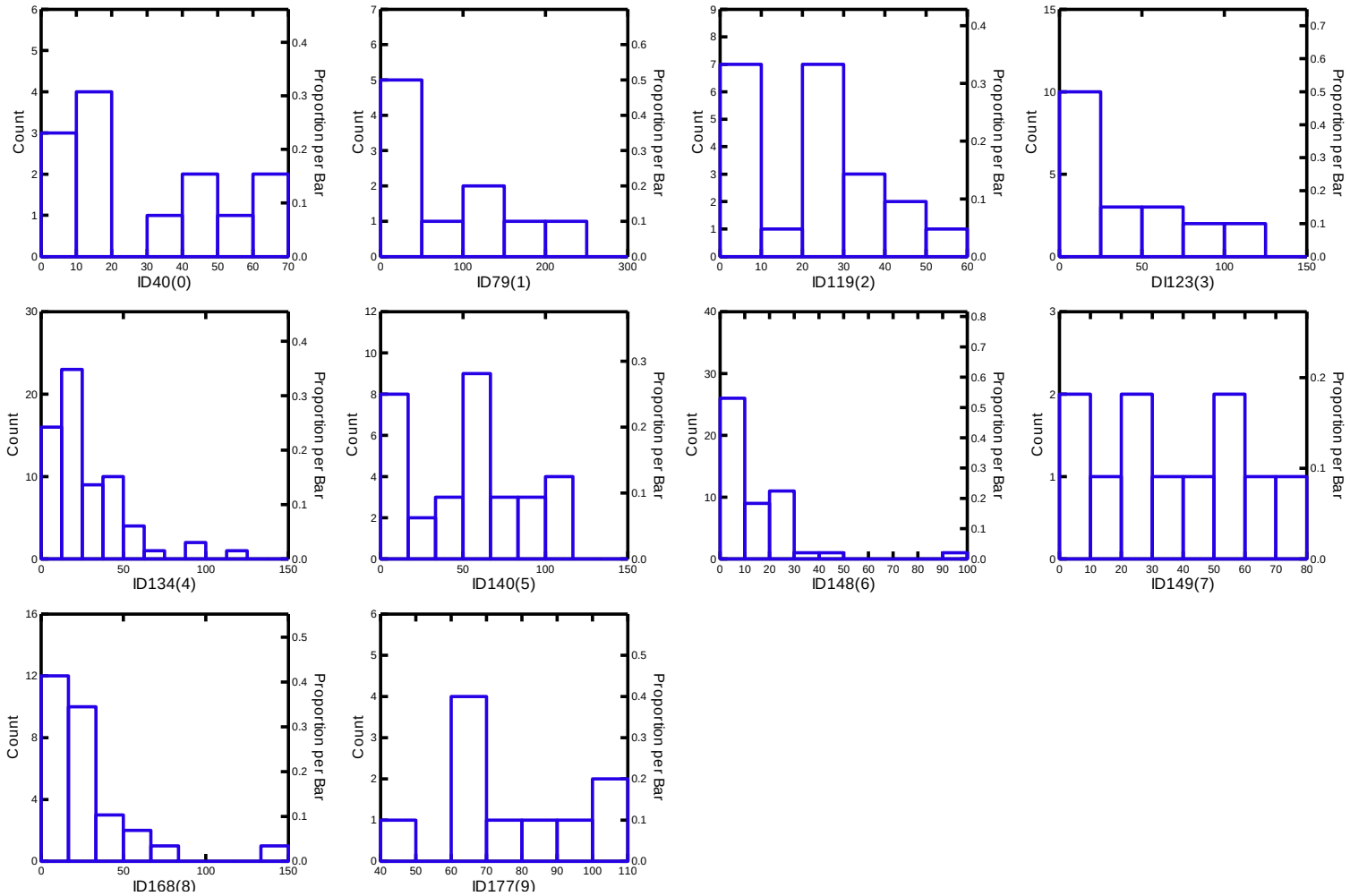


	ID40(0)	ID79(1)	ID119(2)	DI123(3)	ID134(4)	ID140(5)	ID14
1	14.000	78.000	20.000	22.000	50.000	103.000	22
2	17.000	10.000	16.000	18.000	15.000	34.000	1
3	13.000	239.000	26.000	13.000	41.000	73.000	28
4	45.000	25.000	30.000	37.000	41.000	56.000	9
5	6.000	36.000	8.000	14.000	6.000	54.000	13
6	14.000	111.000	20.000	74.000	38.000	82.000	5
7	54.000	153.000	2.000	70.000	32.000	52.000	20
8	39.000	37.000	30.000	62.000	6.000	11.000	18
9	40.000	5.000	3.000	75.000	28.000	94.000	4
10	63.000	115.000	45.000	16.000	12.000	90.000	5
11	2.000	.	2.000	21.000	11.000	62.000	8
12	7.000	.	2.000	33.000	32.000	19.000	6
13	61.000	.	21.000	11.000	25.000	114.000	7
14	.	.	20.000	101.000	18.000	107.000	10
15	.	.	55.000	114.000	46.000	57.000	9
16	.	.	2.000	5.000	24.000	9.000	1
17	.	.	32.000	94.000	16.000	86.000	5
18	.	.	6.000	2.000	48.000	1.000	5

Graphs – Boxplot



Graphs – Histogram



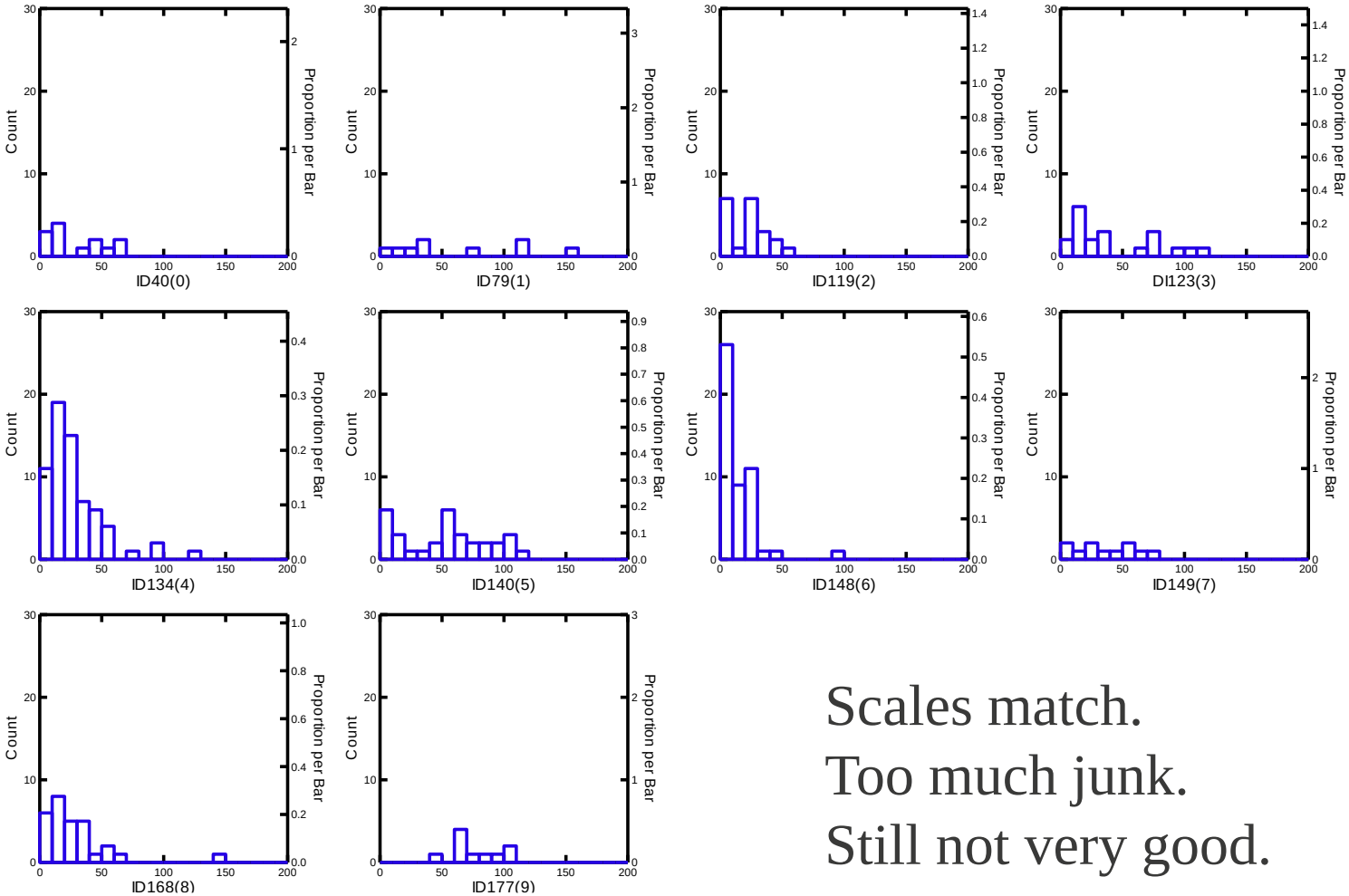
Summary Statistics

	ID40(0)	ID79(1)	ID119(2)	DI123(3)	ID134(4)	ID140(5)	ID148(6)	ID149(7)	ID168(8)	ID177(9)
N of Cases	13	10	21	20	66	32	49	11	29	10
Minimum	2.000	5.000	2.000	2.000	0.000	1.000	1.000	4.000	2.000	49.000
Maximum	63.000	239.000	55.000	114.000	121.000	114.000	96.000	77.000	141.000	109.000
Median	17.000	57.500	20.000	27.500	21.500	54.000	9.000	36.000	20.000	68.500
Arithmetic Mean	28.846	80.900	20.143	41.250	27.379	51.938	14.184	37.273	27.414	76.600
Standard Deviation	22.124	74.630	15.193	35.179	22.820	34.847	15.451	24.491	27.014	20.860
Method = CLEVELAND										
1 of 4	11.500	25.000	5.250	13.500	13.000	15.500	5.750	17.500	14.000	63.000
2 of 4	17.000	57.500	20.000	27.500	21.500	54.000	9.000	36.000	20.000	68.500
3 of 4	47.250	115.000	30.000	72.000	39.000	77.500	20.250	54.000	31.000	96.000

Discussion 1

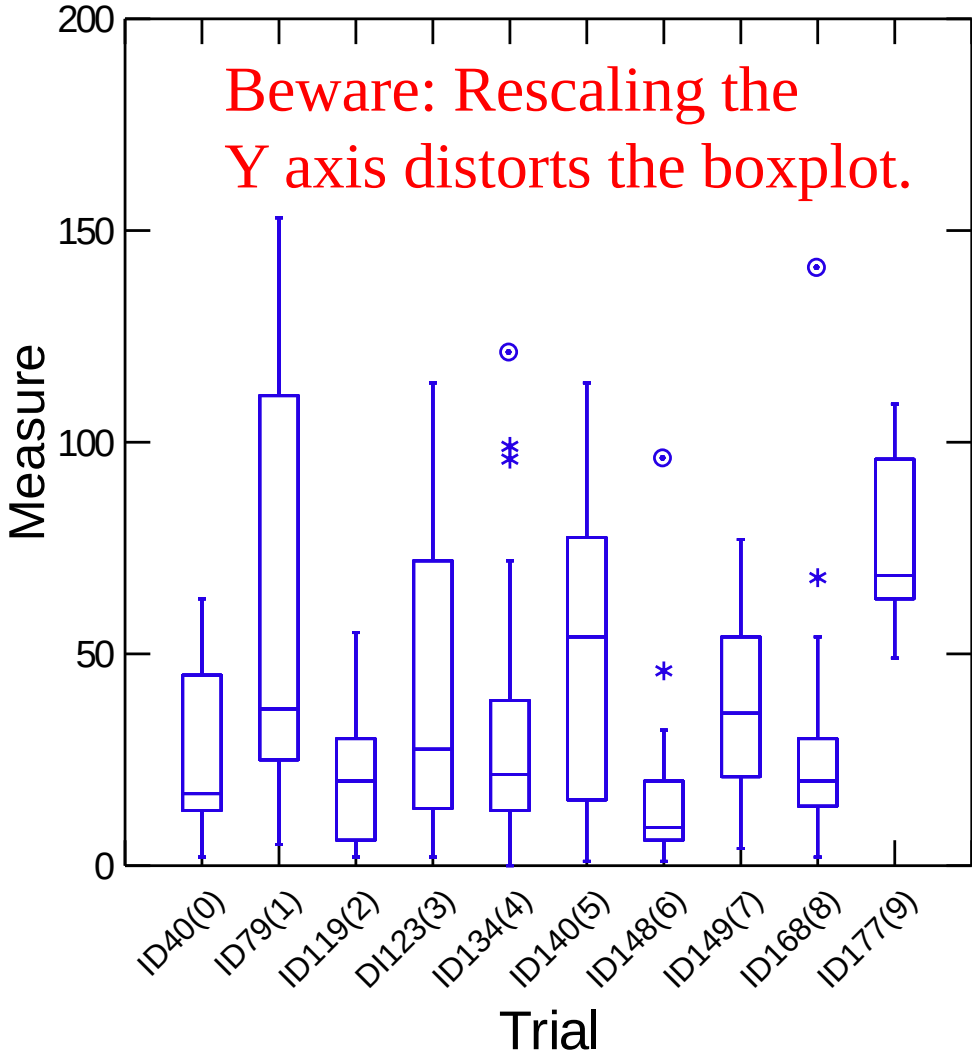
- Data and graphs straight from SYSTAT look pretty terrible. How do you improve them?

Graphs – Histogram



Scales match.
Too much junk.
Still not very good.

Graphs – Boxplot



Discussion 2

- How would you argue the case that Customer #79 was the ‘best’ customer?

- Mean or median
- Spread IQR or standard deviation
- Effect of outliers
- Other observations

Assignment …

- Statistical Analysis and Report Writing
- Here's some guidance on IT report structure:

https://www.monash.edu/rlo/assignment-samples/information-technology/it-report-structure

Tips:

- The table should be written to make comparisons easy. Keep table simple. Use as few columns as possible.
- Table should have a clear title. Columns (and rows) should have clear titles. If a unit of measurement is needed put this with the title (not with each entry).
- Figure captions go below, Table captions go above (by convention).
- Use as few decimal places or significant figures as possible. One or two decimal places should suffice. Show counts as integers.
- Numbers should be right justified (or could be centered if this looks ok).

Tips…

- Get your tables looking as good as possible. They should be designed to make comparisons between groups easy. For Dunhumby data you might do something like…

	ID40(0)	ID79(1)	ID119(2)	DI123(3)	ID134(4)	ID140(5)	ID148(6)	ID149(7)	ID168(8)	ID177(9)
Observations	13	10	21	20	66	32	49	11	29	10
Mean	28.8	80.9	20.1	41.3	27.4	51.9	14.2	37.3	27.4	76.6
Standard Deviation	22.1	74.6	15.2	35.2	22.8	34.8	15.5	24.5	27.0	20.9
Maximum	63.0	239.0	55.0	114.0	121.0	114.0	96.0	77.0	141.0	109.0
Q3	47.3	115.0	30.0	72.0	39.0	77.5	20.3	54.0	31.0	96.0
Median	17.0	57.5	20.0	27.5	21.5	54.0	9.0	36.0	20.0	68.5
Q1	11.5	25.0	5.3	13.5	13.0	15.5	5.8	17.5	14.0	63.0
Minimum	2.0	5.0	2.0	2.0	0.0	1.0	1.0	4.0	2.0	49.0

Tips: Interpretation

- Comment on the shape of the distribution; measures of centre and spread.
- Interesting, unusual or noteworthy values; Any other relevant statistics.
- What the reader should observe from tables and figures. Do not literally describe the appearance of graphs.
- What can you conclude from the evidence presented. You should justify statements with evidence from your table of results.
- Compare groups. Don’t discuss each investment individually.

Tips: Report

- Design your report as best you can.
- All graphs should have a consistent appearance and should be the same size etc.
- Use a graphics package if you need to clean up graphs etc.
- Similar thinking applies to tables.
- Use a consistent template for typefaces and consistent margin widths and layout style.

Structure of Report

1. **Title Page**
2. **Body of Report**

   2-3 written pages. Include tables and graphs that you will refer to in your text. Beware: posted examples are not perfect! Look at marker's comment.

3. **References**

   Harvard or APA is fine

4. **Appendix**

   Include any working/calculations/graphs here that are too lengthy to put in the body of the report. Don't list your data.