Week 12 Tutorial exercises (May 26th 2006)
Outlier Detection

Exercise 1. ***Z-score, Box-plot and Scatter-plot***

The doctor of a school has measured the height of pupils in a 5th grade class. The result (in cm) is as follows:

130	132	138	136	131	153	131	133	129	133	110	132	129	134	135

132	135	134	133	132	130	131	134	135	135	134	136	133	133	130

a- Which ones are outliers and why?
b- The weight of those pupils was measured in kg and the results is as follows. Draw the box-plot for weight.

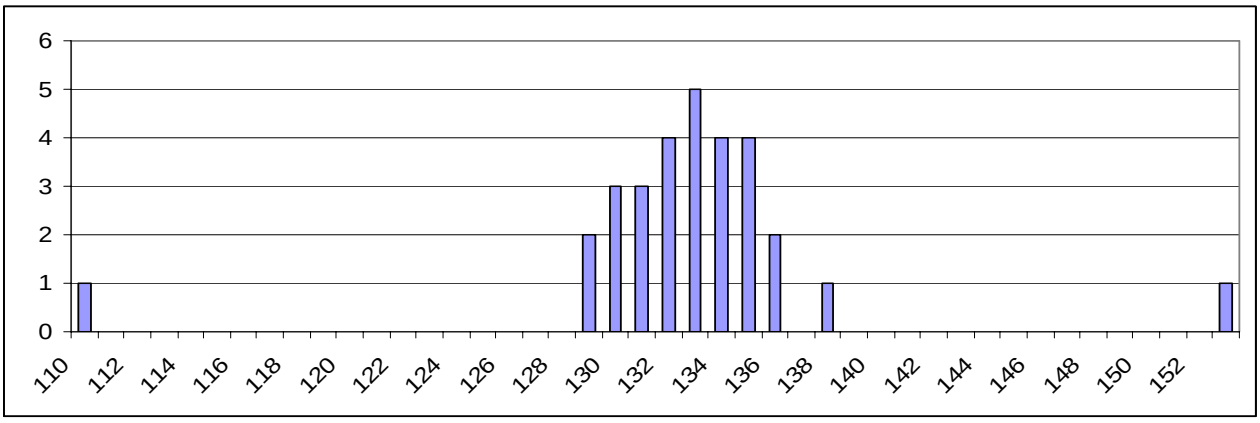
37	40	39	40.5	42	51	41.5	39	41	30	40	42	40.5	39.5	41

40.5	37	39.5	40	41	38.5	39.5	40	41	39	40.5	40	38.5	39.5	41.5

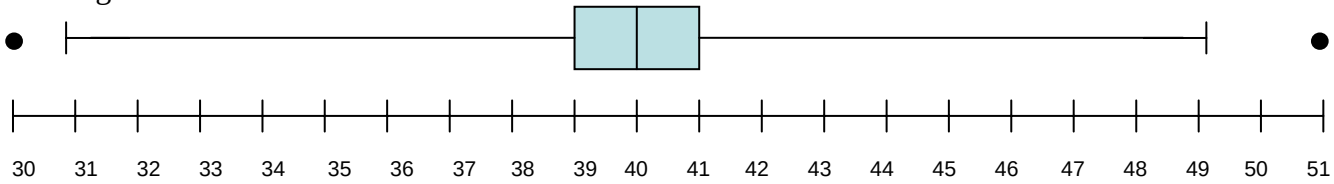
c- Draw the scatter-plot for both variables height and weight.

Solution

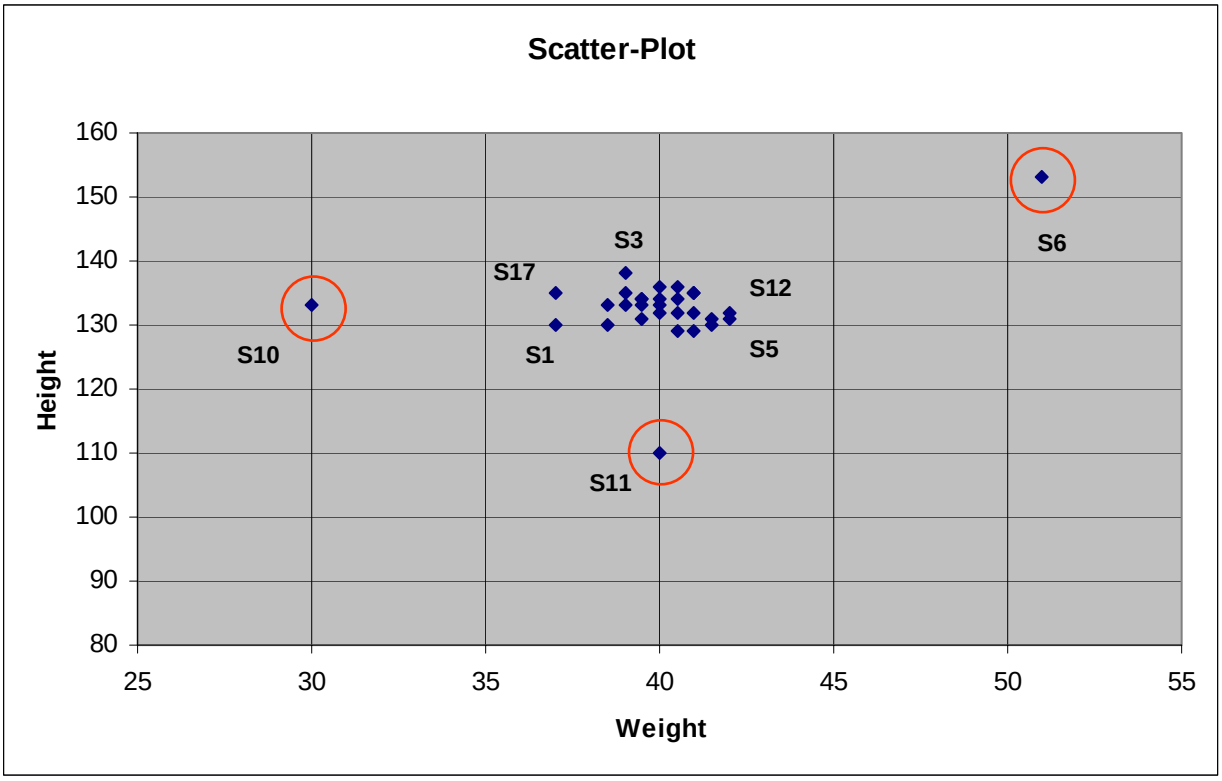
a- For the data set in the table the mean=132.77, s= 6.06, 3s=18.18,
z-score of the observation of 153 is (153-132.77)/6.06=3.34,
z-score of 110 is (110-132.77)/6.06=-3.76.
Since the absolute values of z-score of 153 and 110 are more than 3, the height of 153 cm and the height of 110 cm are outliers in the data set.



b- The mean for the weight is 40 kg. The Standard deviation s=3.02 and 3s=9.06. The median is 40kg while the quartiles at 25% and 75% are 39.125 and 41 respectively. The normal distribution would range from mean ± 3s = [30.93, 49.07]. The box-plot for weight would look like this and shows outliers 30kg and 51kg:



c- The scatter plot for weight and height would look like this and shows three outliers:
S6 (**H=153, W=51**); S11 (**H=110**, W=40) and S10 (H=133, **W=30**).



<u>Exercise 2.</u> k-Nearest neighbor approach

The data from the previous exercise is organized in a table as follows. Use the k-nearest neighbor to rank the pupils by most outlier to least outlier and give the top 4 outliers. Use k=3 and the Euclidian distance.

Pupil	Height	Weight
S1	130	37
S2	132	40
S3	138	39
S4	136	40.5
S5	131	42
S6	153	51
S7	131	41.5
S8	133	39
S9	129	41
S10	133	30
S11	110	40
S12	132	42
S13	129	40.5
S14	134	39.5
S15	135	41

Pupil	Height	Weight
S16	132	40.5
S17	135	37
S18	134	39.5
S19	133	40
S20	132	41
S21	130	38.5
S22	131	39.5
S23	134	40
S24	135	41
S25	135	39
S26	134	40.5
S27	136	40
S28	133	38.5
S29	133	39.5
S30	130	41.5

Solution

The Euclidian distance between two pupils' records A and B is $((H_a-H_b)^2 + (W_A-W_b)^2)^{½}$
We need to find the three closest pupils to each pupil and calculate the average distance between them.
First, here is the distance matrix:

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20
S1	0	3.60	8.24	6.94	5.10	26.92	4.61	3.60	4.12	7.61	20.22	5.38	3.64	4.71	6.40	4.03	5.00	4.71	4.24	4.47
S2		0	6.08	4.03	2.24	23.71	1.80	1.41	3.16	10.05	22.00	2.00	3.04	2.06	3.16	0.50	4.24	2.06	1.00	1.00
S3			0	2.50	7.62	19.21	7.43	5.00	9.22	10.30	28.02	6.71	9.12	4.03	3.61	6.18	3.61	4.03	5.10	6.32
S4				0	5.22	19.98	5.10	3.35	7.02	10.92	26.00	4.27	7.00	2.24	1.12	4.00	3.64	2.24	3.04	4.03
S5					0	23.77	0.50	3.61	2.24	12.17	21.10	1.00	2.50	3.91	4.12	1.80	6.40	3.91	2.83	1.41
S6						0	23.32	23.96	26.00	29.00	44.38	22.85	26.20	22.21	20.59	23.48	22.80	22.21	22.83	23.26
S7							0	3.20	2.06	11.67	21.05	1.12	2.24	3.61	4.03	1.41	6.02	3.61	2.50	1.12
S8								0	4.47	9.00	23.02	3.16	4.27	1.12	2.83	1.80	2.83	1.12	1.00	2.24
S9									0	11.70	19.03	3.16	0.50	5.22	6.00	3.04	7.21	5.22	4.12	3.00
S10										0	25.08	12.04	11.24	9.55	11.18	10.55	7.28	9.55	10.00	11.05
S11											0	22.09	19.01	24.01	25.02	22.01	25.18	24.01	23.00	22.02
S12												0	3.35	3.20	3.16	1.50	5.83	3.20	2.24	1.00
S13													0	5.10	6.02	3.00	6.95	5.10	4.03	3.04
S14														0	1.80	2.24	2.69	0.00	1.12	2.50
S15															0	3.04	4.00	1.80	2.24	3.00
S16																0	4.61	2.24	1.12	0.50
S17																	0	2.69	3.61	5.00
S18																		0	1.12	2.50
S19																			0	1.41
S20																				0
S21																				
S22																				
S23																				
S24																				
S25																				
S26																				
S27																				
S28																				
S29																				
S30																				

	S21	S22	S23	S24	S25	S26	S27	S28	S29	S30
S1	1.50	2.69	5.00	6.40	5.39	5.32	6.71	3.35	3.91	4.50
S2	2.50	1.12	2.00	3.16	3.16	2.06	4.00	1.80	1.12	2.50
S3	8.02	7.02	4.12	3.61	3.00	4.27	2.24	5.02	5.02	8.38
S4	6.32	5.10	2.06	1.12	1.80	2.00	0.50	3.61	3.16	6.08
S5	3.64	2.50	3.61	4.12	5.00	3.35	5.39	4.03	3.20	1.12
S6	26.18	24.82	21.95	20.59	21.63	21.71	20.25	23.58	23.07	24.88
S7	3.16	2.00	3.35	4.03	4.72	3.16	5.22	3.61	2.83	1.00
S8	3.04	2.06	1.41	2.83	2.00	1.80	3.16	0.50	0.50	3.91
S9	2.69	2.50	5.10	6.00	6.32	5.02	7.07	4.72	4.27	1.12
S10	9.01	9.71	10.05	11.18	9.22	10.55	10.44	8.50	9.50	11.88
S11	20.06	21.01	24.00	25.02	25.02	24.01	26.00	23.05	23.01	20.06
S12	4.03	2.69	2.83	3.16	4.24	2.50	4.47	3.64	2.69	2.06
S13	2.24	2.24	5.02	6.02	6.18	5.00	7.02	4.47	4.12	1.41
S14	4.12	3.00	0.50	1.80	1.12	1.00	2.06	1.41	1.00	4.47
S15	5.59	4.27	1.41	0.00	2.00	1.12	1.41	3.20	2.50	5.02
S16	2.83	1.41	2.06	3.04	3.35	2.00	4.03	2.24	1.41	2.24
S17	5.22	4.72	3.16	4.00	2.00	3.64	3.16	2.50	3.20	6.73
S18	4.12	3.00	0.50	1.80	1.12	1.00	2.06	1.41	1.00	4.47
S19	3.35	2.06	1.00	2.24	2.24	1.12	3.00	1.50	0.50	3.35
S20	3.20	1.80	2.24	3.00	3.61	2.06	4.12	2.69	1.80	2.06
S21	0	1.41	4.27	5.59	5.02	4.47	6.18	3.00	3.16	3.00
S22		0	3.04	4.27	4.03	3.16	5.02	2.24	2.00	2.24
S23			0	1.41	1.41	0.50	2.00	1.80	1.12	4.27
S24				0	2.00	1.12	1.41	3.20	2.50	5.02
S25					0	1.80	1.41	2.06	2.06	5.59
S26						0	2.06	2.24	1.41	4.12
S27							0	3.35	3.04	6.18
S28								0	1.00	4.24
S29									0	3.61
S30										0

	3 nearest neighbours	Average distance
S1	{S21, S22, S28}	2.52
S2	{S16, S19, S20}	0.83
S3	{S27, S4, S25}	**2.58** ←
S4	{S27, S15, S24}	0.91
S5	{S7, S12, S30}	0.87
S6	{S3, S4, S27}	**19.81** ←
S7	{S5, S30, S12}	0.87
S8	{S28, S29, S19}	0.67
S9	{S13, S30, S7}	1.23
S10	{S17, S1, S28}	**7.80** ←
S11	{S13, S9, S21}	**19.36** ←
S12	{S5, S20, S7}	1.04
S13	{S9, S30, S7}	1.38
S14	{S18, S23, S26}	0.50
S15	{S24, S4, S26}	0.75
S16	{S2, S20, S19}	0.71
S17	{S25, S28, S14}	2.40
S18	{S14, S23, S26}	0.50
S19	{S29, S2, S8}	0.83
S20	{S16, S2, S12}	0.83
S21	{S22, S1, S13}	1.72
S22	{S2, S16, S21}	1.32
S23	{S14, S18, S26}	0.50
S24	{S15, S4, S26}	0.75
S25	{S14, S18, S23}	1.22
S26	{S23, S14, S18}	0.83
S27	{S4, S15, S24}	1.11
S28	{S8, S29, S14}	0.97
S29	{S8, S19, S14}	0.67
S30	{S7, S5, S9}	1.08

Top 4 outliers are those that have the highest average distance to their 3 nearest neighbours, and they are:
S6 (**H=153, W=51**) with average distance 19.81;
S11 (**H=110**, W=40) with average distance 19.36 and
S10 (H=133, **W=30**) with average distance 7.80, and with less extent
S3 (H=138, W=39) with an average distance 2.58 from which S1 and S17 are very close in rank with respectively 2.52 and 2.40 average distance to their nearest neighbours.

Exercise 3. ***Density-based outliers*** ***(to be done at your own time, not in class)***

Use the previous dataset and calculate the LOF for each pupil data point and give the top 4 outliers. Use k=3. Use the same distance matrix you calculated in the previous exercise.

Solution:

We first get the k-distance neighbourhood for each pupil. The furthest among the 3 nearest neighbours gives the k-distance for each pupil. To calculate the reachability distance for each pupil p with regard to each of its 3 nearest neighbours, for each neighbours o we take the maximum between the distance d(p,o) and k-distance of o. The reachability density for each pupil is the average between the reachability distances with regard to its nearest neighbours.

	3 nearest neighbours	k-distance	Reach-dist$_3$(p)	1/lrd$_3$(p)	lrd$_3$(p)	**LOF**	**Top4**
S1	{S21, S22, S28}	3.35	{2.24, 2.69, 3.35}	2.76	0.36	1.77	← (6)
S2	{S16, S19, S20}	1	{1.12, 1, 1}	1.04	0.96	1.00	
S3	{S27, S4, S25}	3	{2.24, 2.50, 3}	2.58	0.38	**2.11**	← 4
S4	{S27, S15, S24}	1.12	{1.41, 1.12, 1.12}	1.21	0.83	1.02	
S5	{S7, S12, S30}	1.12	{1.12, 1.12, 1.12}	1.12	0.89	0.94	
S6	{S3, S4, S27}	20.25	{19.21, 19.98, 20.25 }	**19.81**	**0.05**	**13.13**	← 1
S7	{S5, S30, S12}	1.12	{1.12, 1.12, 1.12}	1.12	0.89	0.94	
S8	{S28, S29, S19}	1	{1.41, 1, 1}	1.14	0.88	1.08	
S9	{S13, S30, S7}	2.06	{2.24, 1.12, 2.16}	1.84	0.54	1.30	
S10	{S17, S1, S28}	8.5	{7.28, 7.61, 8.50}	**7.80**	**0.13**	**4.17**	← 3
S11	{S13, S9, S21}	20.05	{19.01, 19.03, 20.06}	**19.37**	**0.05**	**10.00**	← 2
S12	{S5, S20, S7}	1.12	{1.12, 1, 1.12}	1.08	0.93	0.97	
S13	{S9, S30, S7}	2.24	{2.06, 1.41, 2.24}	1.90	0.53	1.34	
S14	{S18, S23, S26}	1	{1, 0.5, 1}	0.83	1.20	0.94	
S15	{S24, S4, S26}	1.12	{1.12, 1.12, 1.12}	1.12	0.89	1.09	
S16	{S2, S20, S19}	1.12	{1, 1, 1.12}	1.04	0.96	1.00	
S17	{S25, S28, S14}	2.69	{2, 2.5, 2.69}	2.40	0.42	2.07	← (5)
S18	{S14, S23, S26}	1	{1, 0.5, 1}	0.83	1.20	0.87	
S19	{S29, S2, S8}	1	{1, 1, 1}	1	1	0.95	
S20	{S16, S2, S12}	1	{1.12, 1, 1.12}	1.08	0.93	1.02	
S21	{S22, S1, S13}	2.24	{1.41, 3.35, 2.24}	2.33	0.43	1.48	← (7)
S22	{S2, S16, S21}	1.41	{1.12, 1.41, 2.24}	1.59	0.63	1.24	
S23	{S14, S18, S26}	0.5	{1, 1, 1}	1	1	1.20	
S24	{S15, S4, S26}	1.12	{1.12, 1.12, 1.12}	1.12	0.89	1.09	
S25	{S14, S18, S23}	1.41	{1.12, 1.12, 1.41}	1.22	0.82	1.38	
S26	{S23, S14, S18}	1	{0.5, 1, 1}	0.83	1.20	0.94	
S27	{S4, S15, S24}	1.41	{1.12, 1.41, 1.41}	1.31	0.76	1.14	
S28	{S8, S29, S14}	1.41	{1, 1, 1.41}	1.17	0.85	1.21	
S29	{S8, S19, S14}	1	{1, 1, 1}	1	1	1.03	
S30	{S7, S5, S9}	1.12	{1.12, 1.12, 2.06}	1.43	0.70	1.10	

The final step is to calculate the LOF for each pupil.
Sorting in descending order of LOF, the top 4 outliers are:
S6 (**H=153, W=51**) (LOF=13.13),
S11 (**H=110**, W=40) (LOF=10.00),
S10 (H=133, **W=30**) (LOF=4.17) and
S3 (H=138, W=39) (LOF=2.11).

The next outliers would be S17, S1 and S21 according to the LOF ranking.

Exercise 4:* Resolution-based outliers *(to be done at your own time, not in class)

Use the previous dataset and calculate the ROF for each pupil data point and give the top 4 outliers. Use the same distance matrix you calculated in the previous exercise. You can use a resolution step of 1.

Solution:

Initially all pupils are clustered alone in 30 different clusters. We first iteratively cluster the pupils starting with a minimal distance (resolution) d=1 until all pupils are grouped in one cluster. This happens when the distance d=20. We count the size of the found clusters at each iteration and assign to each pupil the size of the cluster in which they belong. From d=8 to d=20 nothing happens so may skip the steps.

	d=0	d=1	d=2	d=3	D=4	d=5	d=6	d=7	d=8	d=20	**ROF**	**Top4**
S1	1	1	26	27	27	27	27	27	28	30	6.60	←
S2	1	13	26	27	27	27	27	27	28	30	7.07	
S3	**1**	**1**	**1**	**27**	27	27	27	27	28	30	5.68	← 4
S4	1	1	26	27	27	27	27	27	28	30	6.60	←
S5	1	13	26	27	27	27	27	27	28	30	7.07	
S6	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**30**	0	← 1
S7	1	1	26	27	27	27	27	27	28	30	6.60	←
S8	1	13	26	27	27	27	27	27	28	30	7.07	
S9	1	2	26	27	27	27	27	27	28	30	6.64	
S10	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**28**	30	0.9	← 3
S11	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**30**	0	← 2
S12	1	13	26	27	27	27	27	27	28	30	7.07	
S13	1	2	26	27	27	27	27	27	28	30	6.64	
S14	1	13	26	27	27	27	27	27	28	30	7.07	
S15	1	2	26	27	27	27	27	27	28	30	6.64	
S16	1	13	26	27	27	27	27	27	28	30	7.07	
S17	1	2	26	27	27	27	27	27	28	30	6.64	
S18	1	13	26	27	27	27	27	27	28	30	7.07	
S19	1	13	26	27	27	27	27	27	28	30	7.07	
S20	1	13	26	27	27	27	27	27	28	30	7.07	
S21	1	1	26	27	27	27	27	27	28	30	6.60	←
S22	1	1	26	27	27	27	27	27	28	30	6.60	←
S23	1	13	26	27	27	27	27	27	28	30	7.07	
S24	1	2	26	27	27	27	27	27	28	30	6.64	
S25	1	1	26	27	27	27	27	27	28	30	6.60	←
S26	1	13	26	27	27	27	27	27	28	30	7.07	
S27	1	1	26	27	27	27	27	27	28	30	6.60	←
S28	1	13	26	27	27	27	27	27	28	30	7.07	
S29	1	13	26	27	27	27	27	27	28	30	7.07	
S30	1	2	26	27	27	27	27	27	28	30	6.64	

d=1: s2,s5,s8,s12, s14,s16,s18,s19,s20, s23,s26,s28,s29 } {s9,s13} {s15,s24} {s17,s30}
d=2: s1,s2,s4,s5,s7,s8,s9,s12,s13,s14,s15,s16,s17,s18,s19,s20,s21,s22,s23,s24,s25,s26,s27,s28,s29,s30}
d=3:{s1,s2,s3,s4,s5,s7,s8,s9,s12,s13,s14,s15,s16,s17,s18,s19,s20,s21,s22,s23,s24,s25,s26,s27,s28,s29,s30}
at d=8 s10 is added and at d=20 s11 and s6 are added.

The final step is to calculate the ROF for each pupil.
Sorting in ascending order of ROF, the top 4 outliers are:
S6 (**H=153, W=51**) (ROF=0),
S11 (**H=110**, W=40) (ROF=0),
S10 (H=133, **W=30**) (ROF=0.9) and
S3 (H=138, W=39) (ROF=5.68).

Many pupils are in the 5th place (S1, S4, S7, S21, S22, S25, S27). A lower step increase in the resolution change, for instance 0.5 rather than 1, would distinguish between them without altering the ranking of the top 4..