Spring 2017
44-566 Applied Data Analytics
Assignment 2: 8% (40 M)
Individual Submission ----------

1. The 'age' attribute is given below for a dataset.

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Calculate the measures of central tendency. And calculate the range and mid-range. **5M**

A. **Sum of all ages**

=13+15+16+16+19+20+20+21+22+22+25+25+25+25+30+33+33+35+35+ 35+35+35+36+40+45+46+52+70

=809

No. of attributes = 27

Measures of Central Tendency:

- Mean = 809/27 = 29.96
- Median = 25
- Mode = 25, 35

Range = 70-13 = 57

Mid- Range = (70+13)/2 = 41.5

2. A survey was done for the number of pets owned by your classmates, with the following results:

Number of pets	Frequency
0	4
1	12
2	8
3	2
4	1
5	2
6	1

Calculate the variance and standard deviation. **4M**

A. Sum of all the frequencies = 4+12+8+2+1+2+1 = 30
Size = 7
Mean = Sum of all the frequencies / Size = 30/7 = 4.28

Number of pets	Frequency	Mean	Difference from mean	Squared Difference from mean
0	4	4.28	-0.28	0.08
1	12	4.28	7.71	59.5
2	8	4.28	3.71	13.7
3	2	4.28	-2.28	5.22
4	1	4.28	-3.28	10.7
5	2	4.28	-2.28	5.22
6	1	4.28	-3.28	10.7
Total	30		Total	105.32

Variance = sum of the squared deviation / size -1

= 105.32/6

Variance = 17.55

Standard Deviation = Square root of Variance = 4.18

Standard Deviation = 4.18

3. Given two objects represented by the records: (22, 1, 42, 10) and (20, 0, 36, 8)
   a. Compute the Euclidean distance between the two records.
   b. Compute the Manhattan distance between the two records. **4M**

A. Euclidean Distance:

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

= squareroot of [$(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2$]
= squareroot of [4 + 1 + 36 + 4]
= squareroot of [45]
= 6.708

Euclidean Distance = 6.708

Manhattan Distance:

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+...+|x_{ip}-x_{jp}|$$

= |22-20|+|1-0|+|42-36|+|10-8|
=11

Manhattan Distance = 11

4. Suppose we have the following data set:

Document/Term	T1	T2	T3	T4	T5	T6	T7
D1	0	4	10	8	0	5	0
D2	5	19	7	16	0	0	32

Find the similarity between documents D1 and D2 using cosine similarity. **2M**

A. cos(d1,d2)=(d1*d2)/||d1||||d2||

Numerator = 0*5+4*19+10*7+8*16+0*0+5*0+0*32
= 276

|d1| = $0^2+4^2+10^2+8^2+0^2+5^2+0^2$
= 205
||d1|| = 14.31

|d2| = $5^2+19^2+7^2+16^2+0^2+0^2+32^2$
= 1715
||d2|| = 41.41

Denominator = 14.31*41.41 = 592.5

Cosine Similarity = 276 / 14.31*41.41 = 0.46

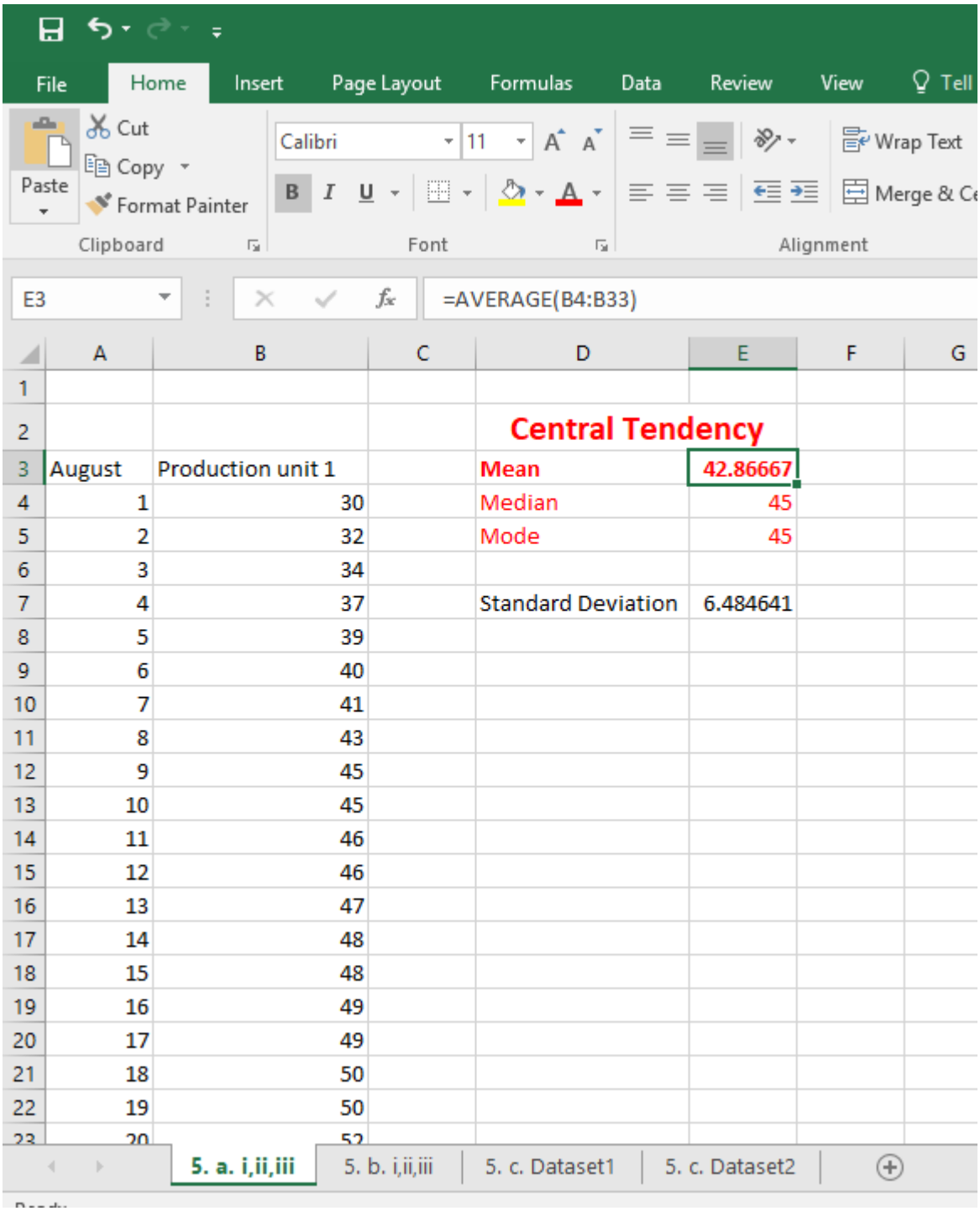
Similarity between documents D1 and D2 using cosine similarity = 0.46

5. Two production units' production datasets are given to you for August. Refer Assignment02.xlsx

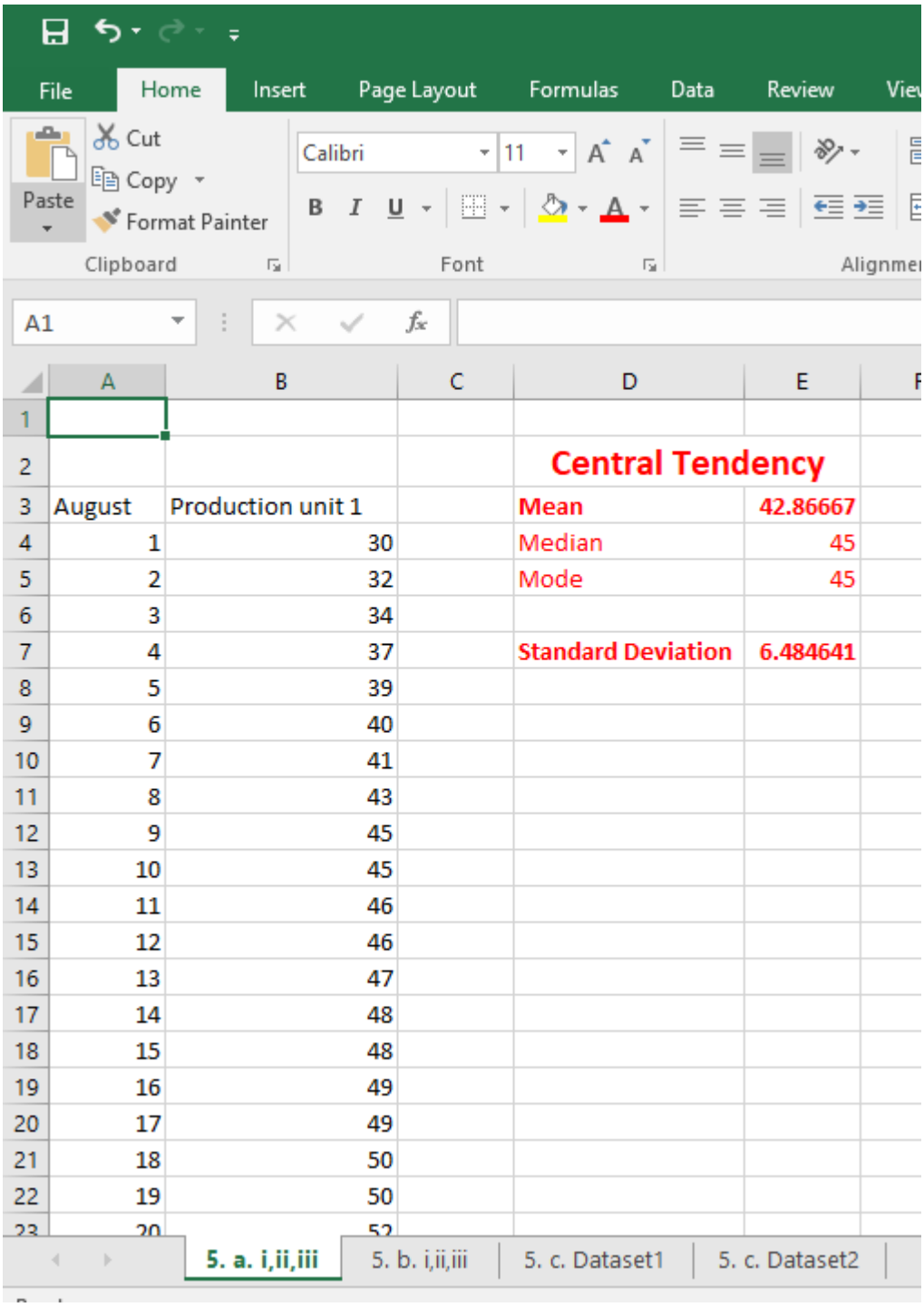
a. For dataset 1: For attribute 'Production Unit 1':
   a.i. Calculate the Central Tendency (in MS Excel).
   a.ii. From the data, what information can you observe after having the average?
   a.iii. What percentile of data is below median?

a.iv. Calculate the Standard Deviation on this attribute in MS Excel. **10M**

A. a.i.



a.ii. From the above data we can tell that on an average, 43 units are produced in a day. Median and mode of the data are same.
a.iii. 50% of data is below median.
a.iv.



	A	B	C	D	E
1					
2				Central Tendency	
3	August	Production unit 1		Mean	42.86667
4	1	30		Median	45
5	2	32		Mode	45
6	3	34			
7	4	37		Standard Deviation	6.484641
8	5	39			
9	6	40			
10	7	41			
11	8	43			
12	9	45			
13	10	45			
14	11	46			
15	12	46			
16	13	47			
17	14	48			
18	15	48			
19	16	49			
20	17	49			
21	18	50			
22	19	50			
23	20	52			

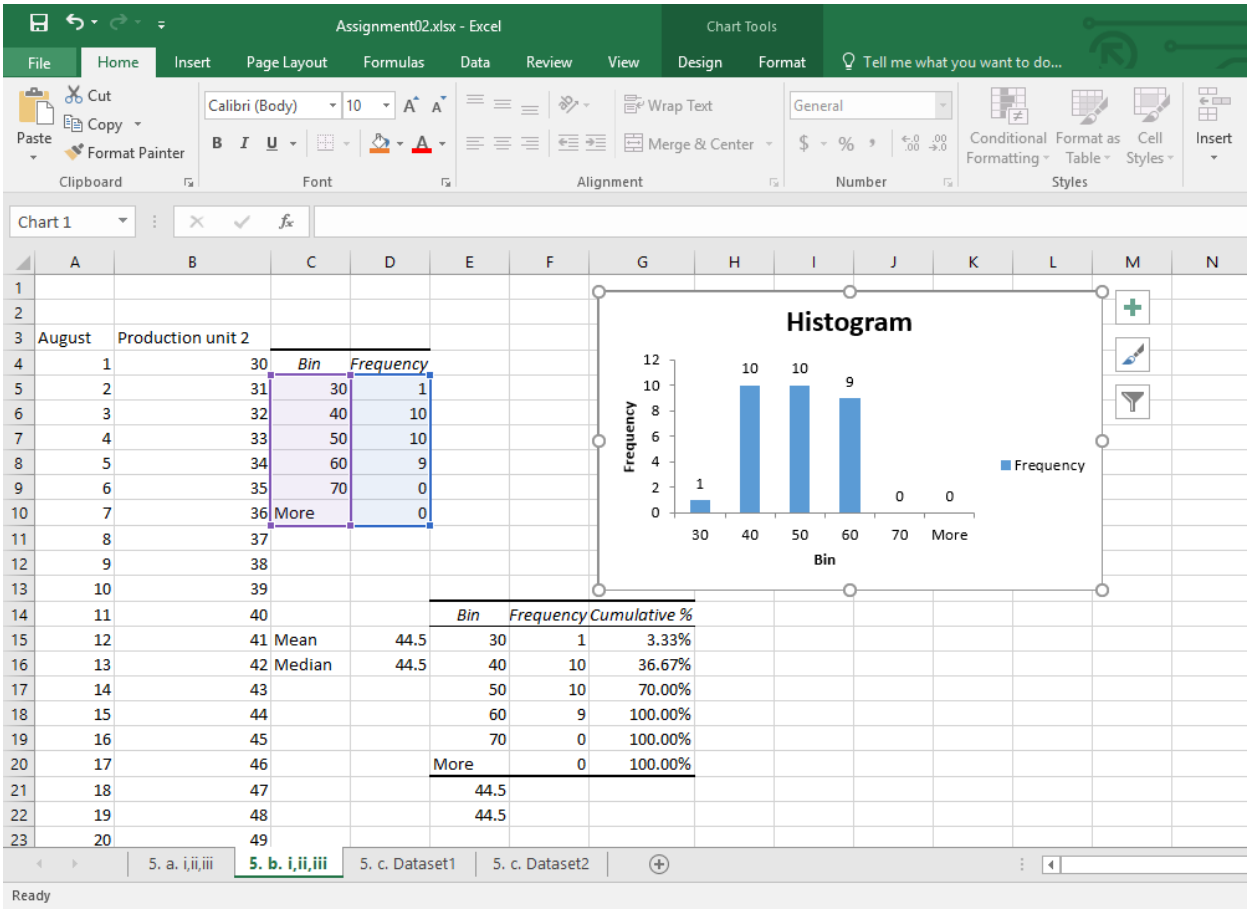
b. For dataset 2: For attribute ‘Production Unit 2’:

b.i. Draw a histogram (in MS Excel) for the given data

b.ii. Show the average on histogram

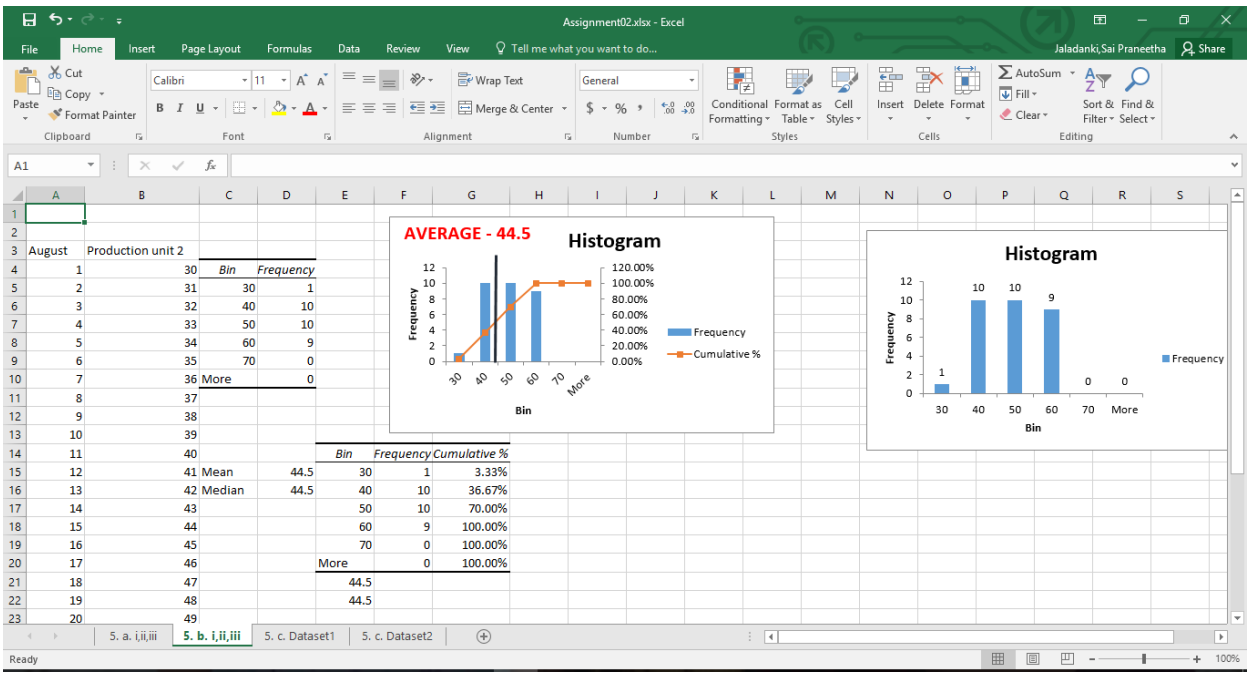
b.iii. Show in Histogram what percentile of data is below mean (use *Line* in Excel for this). **8M**

A. b.i.

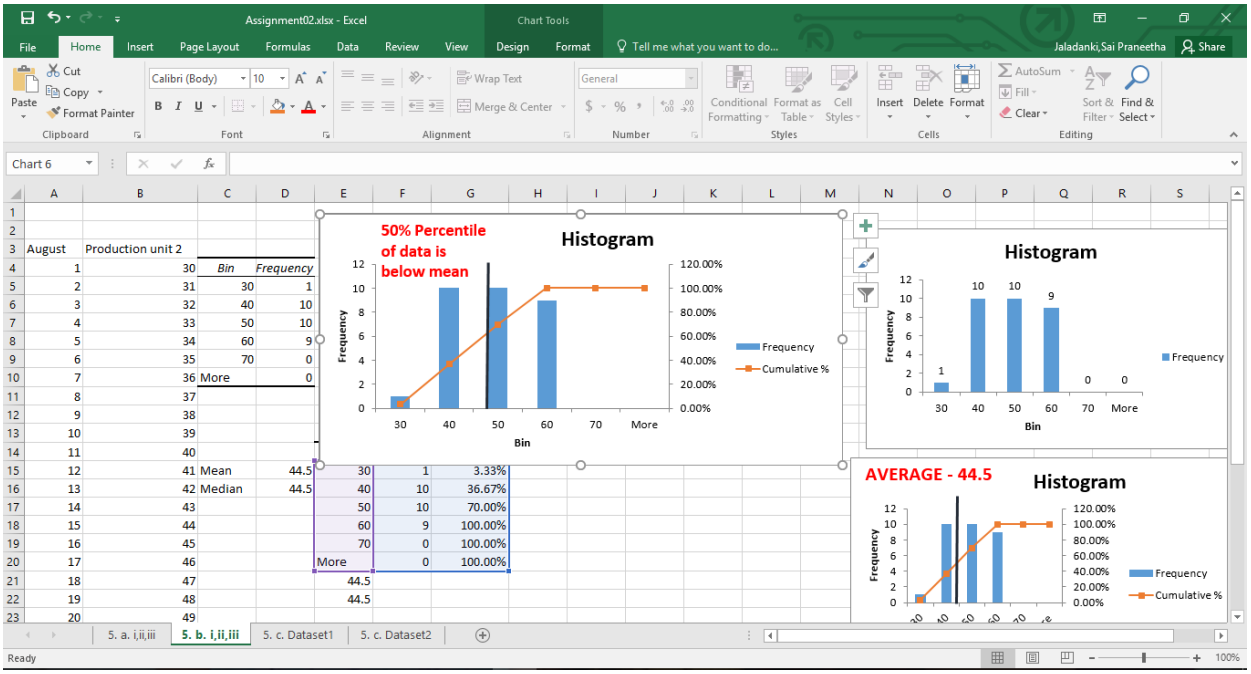


	A	B	C	D	E	F	G
3	August	Production unit 2					
4	1	30	*Bin*	*Frequency*			
5	2	31	30	1			
6	3	32	40	10			
7	4	33	50	10			
8	5	34	60	9			
9	6	35	70	0			
10	7	36	More	0			
14					*Bin*	*Frequency*	*Cumulative %*
15	12	41	Mean	44.5	30	1	3.33%
16	13	42	Median	44.5	40	10	36.67%
17	14	43			50	10	70.00%
18	15	44			60	9	100.00%
19	16	45			70	0	100.00%
20	17	46			More	0	100.00%
21	18	47			44.5		
22	19	48			44.5		

b.ii.

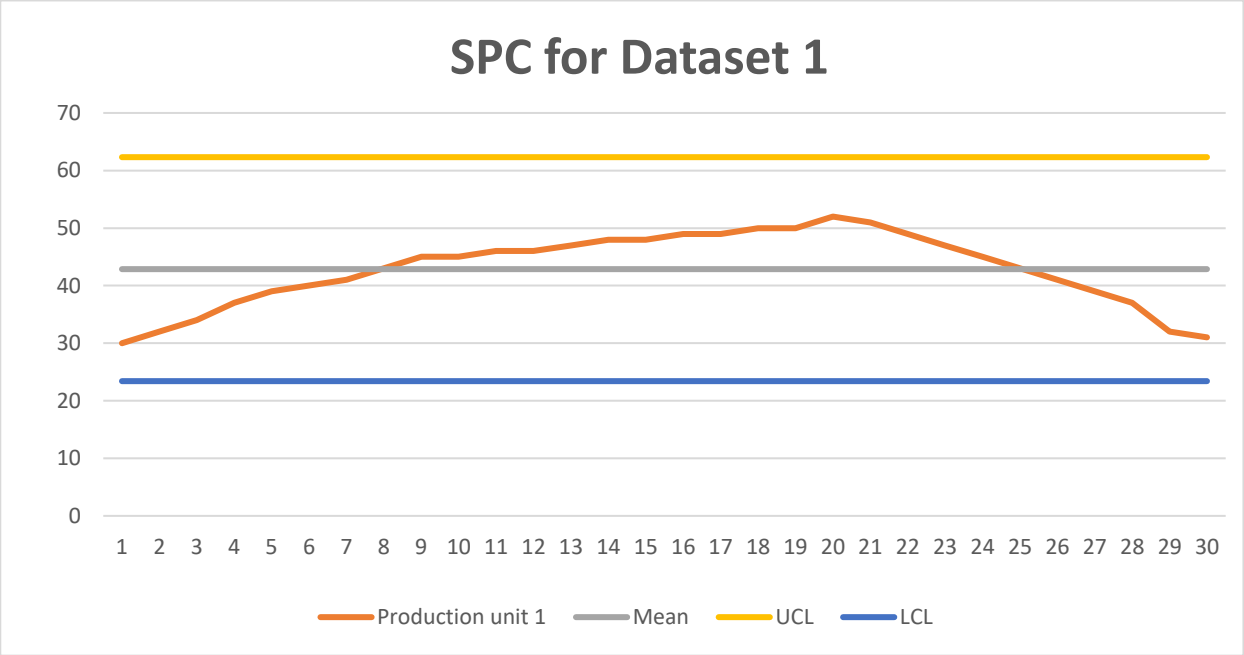


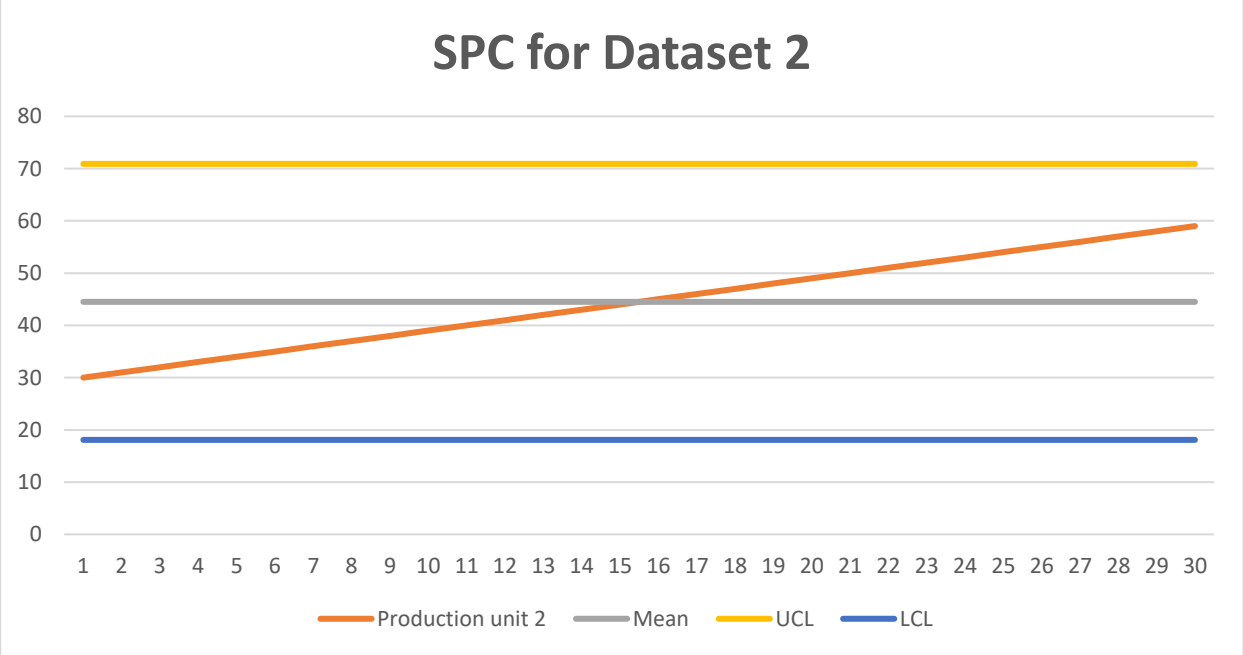
b.iii.



c. Make SPC charts for dataset 1 and 2 separately for august production.

You can use Minitab for SPC generation **5M**





d. Check if in any SPC the production process is stable? If not, mention the reason. **2M**

A. Both the SPCs are stable as all points lie within the upper control limit and lower control limit.

Submission:

1. Write all answers in MS Word file and convert it to pdf for submission
2. Paste all generated images in MS Word (including histograms)
3. For question 10. B (i, ii, iii), paste histograms in MS Word file as screenshots
4. Name your pdf as 'lastname_firstname.pdf'
5. Remember that you should submit only one pdf file which has answers for all the questions in a sequential order
6. **Deadline for this assignment is 11th Feb 23:59**