CIS31041 - Practical for Data Mining

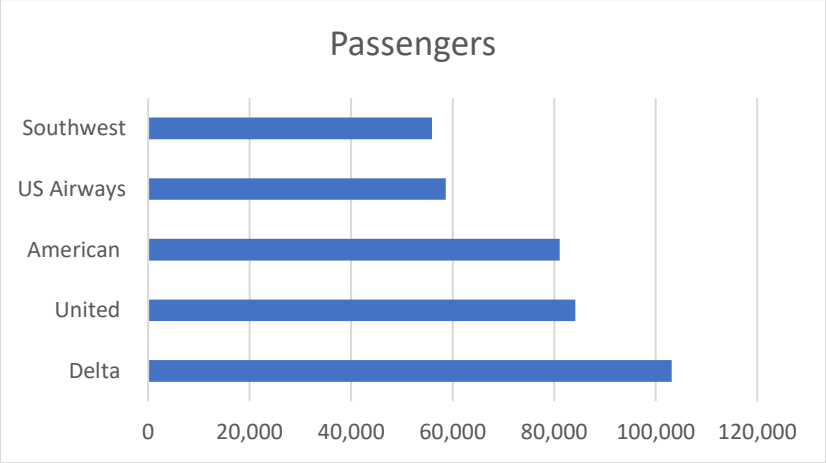
Reg.No: SEU.IS.20.ICT.084 Lab Sheet 01

Question 01

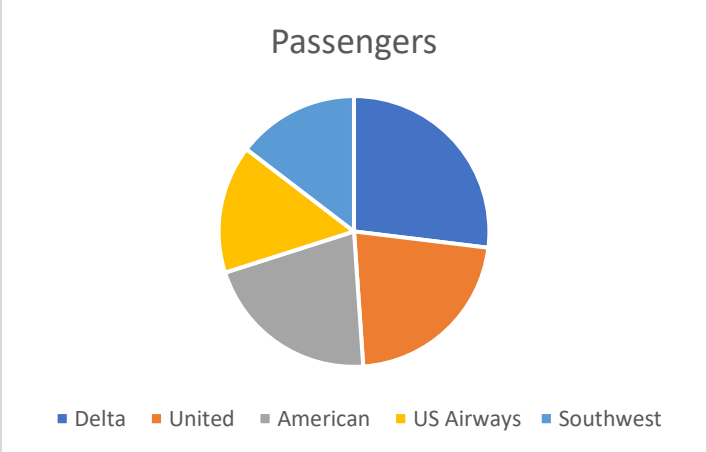
The data set shown below the number of passengers flown by top five Airlines. **Construct Bar Chart, Line Chart and Pie Chart to depict the following information**.

Airlines	Delta	United	American	US Airways	Southwest
Passengers	103,133	84,203	81,083	58,659	55,946

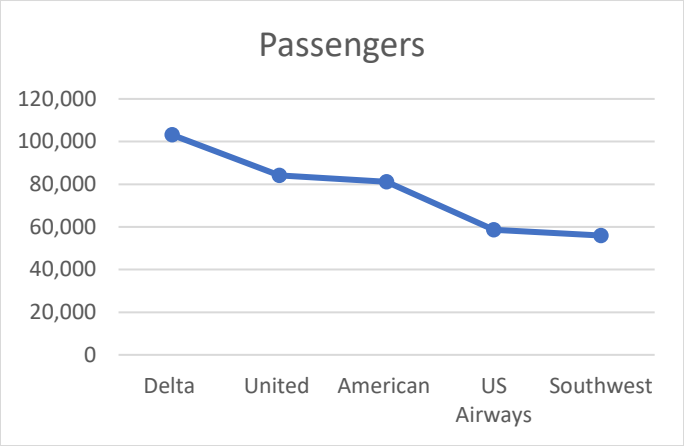
Construct Bar



Pie Chart

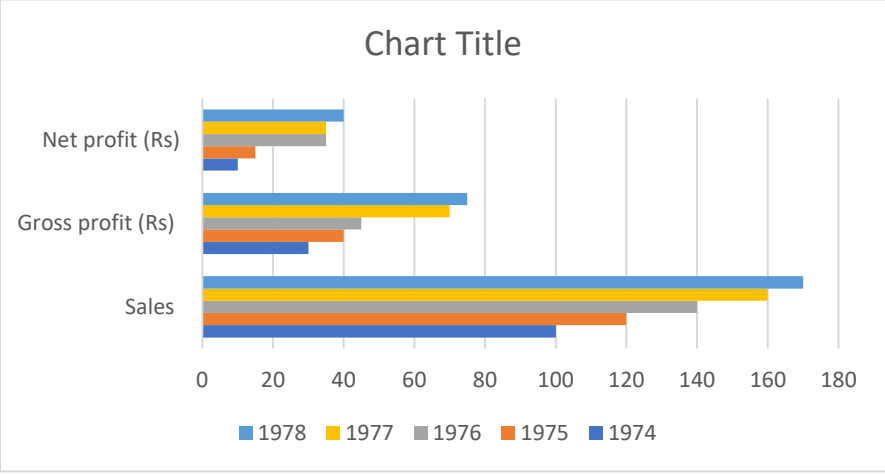


Line Chart

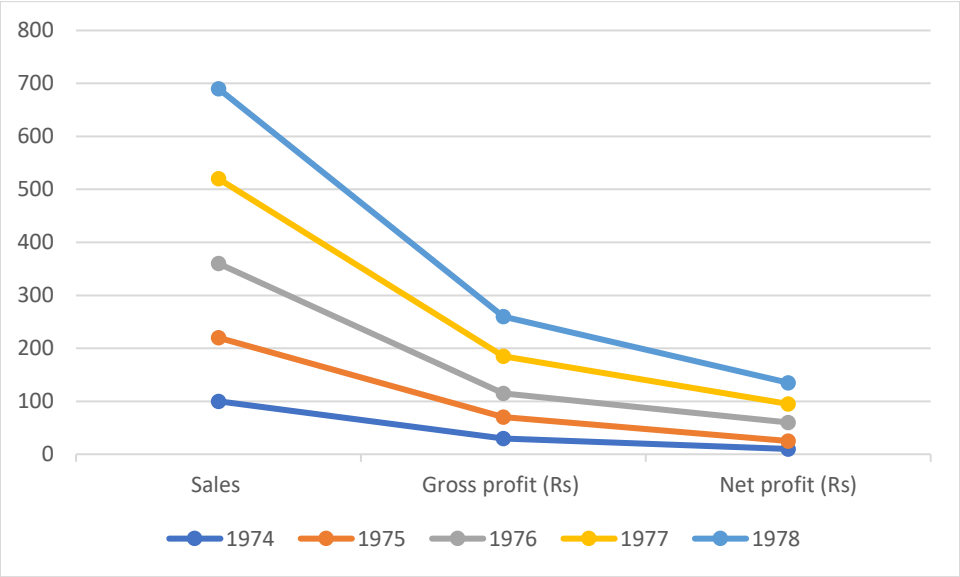


Year	Sales	Gross profit (Rs)	Net profit (Rs)
1974	100	30	10
1975	120	40	15
1976	140	45	35
1977	160	70	35
1978	170	75	40

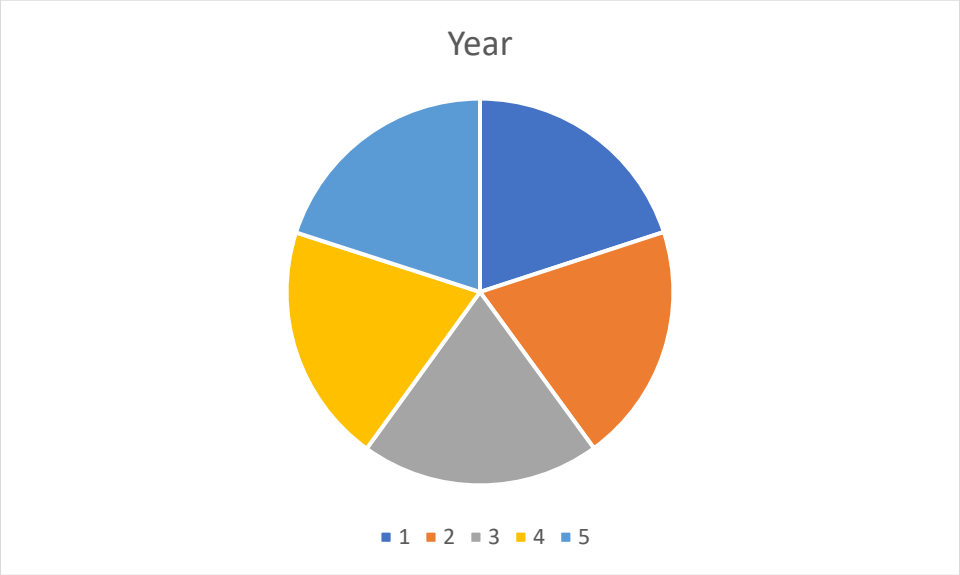
Construct Bar Chart



Line Chart



Pie Chart



Question 02

Calculate **Mean, Median, Mode, Quartiles, Inter Quartile Range (IQR)**

i. 36 , 48, 90, 67, 81, 41, 28, 34, 90, 77, 90

36	28			
48	34	Mean	(28+34+36+41+48+67+77+81+90+90+90)/11	62
90	36	Median	(n+1)/2 = (11+1)/2=6th	67
67	41	Mode		90
81	48	Quartiles 1	(n+1)/4=(11+1)/4=3th	36
41	67	Quartiles 2	(n+1)/2=(11+1)/2=6th	67
28	77	Quartiles 3	3(n+1)/4=3(11+1)/4=9th	90
34	81	Inter Quartile Range (IQR)	Q3-Q1=90-36	54
90	90			
77	90			
90	90			

ii. 6 , 7 , 2 , 5 , 6 , 0 , 3 , 6, 7 , 4 , 5 , 4 , 5 , 7, 7

6	0	Mean	(0+2+3+4+4+5+5+5+6+6+6+7+7+7+7)/15	4.933333
7	2	Median	(n+1)/2=(15+1)/2=8th	5
2	3	Mode		7
5	4	Quartiles 1	(n+1)/4=(15+1)/4 =4th	4
6	4	Quartiles 2	(n+1)/2=(15+1)/4 =8th	5
0	5	Quartiles 3	3(n+1)/4=(15+1)/4 =12th	7
3	5	Inter Quartile Range (IQR)	Q3-Q1=7-4	3
6	5			
7	6			
4	6			
5	6			
4	7			
5	7			
7	7			
7	7			

Question 03

i. Estimate the following for the below grouped frequency distribution table
ii. Measures the mean, Var, Std, CV for the following data.

Class	1.2 - 1.6	1.6 - 2.0	2.0 - 2.4	2.4 - 2.8	2.8 - 3.2
Frequency	220	150	90	110	280

Class	Frequency	xi	f*x
1.2 - 1.6	220	1.4	308
1.6 - 2.0	150	1.8	270
2.0 - 2.4	90	2.2	198
2.4 - 2.8	110	2.6	286
2.8 - 3.2	280	3	840
Total	850		1902

Mean Value	∑fi*xi/∑fi	1902-850	2.237647059
Variance Value	[∑fi*xi*xi-((∑fi*xi)/∑fi)*((∑fi*xi)/∑fi)]/∑fi	[4616.4-((1902/850)*(1902/850))]850	5.42516816
Std Value	√Variance	√5.425168	2.329199038
Cv Value	(standard Deviation/Mean)	2.237647059/2.329199038	0.960693793

Mean=$\frac{\sum fixi}{\sum fi}$ variance=$\frac{\sum fi(xi-\bar{x})^2}{\sum fi}$ standard deviation $=\sqrt{\boldsymbol{\sigma}}$ coefficient of variation (CV) $=\left(\frac{\boldsymbol{\sigma}}{\bar{x}}\right) x100\%$

Class	1.2 - 1.6	1.6 – 2.0	2.0 - 2.4	2.4 - 2.8	2.8 - 3.2	3.2 – 3.6	3.6 – 4
Frequency	3	7	40	50	30	15	5

Class	'equency(f	Middle(xi)	Fi*Xi	fi*xi*xi
1.2-1.6	3	1.4	4.2	5.88
1.6-2.0	7	1.8	12.6	22.68
2.0-2.4	40	2.2	88	193.6
2.4-2.8	50	2.6	130	338
2.8-3.2	30	3	90	270
3.2-3.6	15	3.4	51	173.4
3.6-4.0	5	3.8	19	72.2
Total	150	18.2	394.8	1075.76

Mean Value	∑fi*xi/∑fi	394.8/150	2.6324
Variance Value	[∑fi*xi*xi-((∑fi*xi)/∑fi)*((∑fi*xi)/∑fi)]/∑fi	[1075-((394.8/150)*(394.8/150)]150	35.8864
Std Value	√Variance	√35.8864	5.990525853
Cv Value	(standard Deviation/Mean)	5.99052585337881/2.6324	2.275689809

Question 04

Calculate the correlation coefficient of the following data set and find the linear regression model

i. (1,30), (2,45), (3,51), (4,57), (5,60), (6,65), (7,70), (8,71)

	xi	yi	xi*yi	xi*xi	yi*yi
	1	30	30	1	900
	2	45	90	4	2025
	3	51	153	9	2601
	4	57	228	16	3249
	5	60	300	25	3600
	6	65	390	36	4225
	7	70	490	49	4900
	8	71	568	64	5041
Total	36	449	2249	204	26541

n=8			
m	[n*∑xi*yi-∑xi*∑yi]/n*∑xi*xi-(∑xi)^2	[8*2249-36*449]/8*204-(36)^2	5.440476
c	(∑yi-m*∑xi)/n	(449-5.440476*36)/8	31.642858
Correlation coefficient	[n*∑xi*yi-∑xi*∑yi]/√(n*∑xi*xi-(∑xi^2)*√(n*∑yi*yi-(∑yi^2)	[8*2249-36*449]/√(8*204-(36^2)*√(8*26541-(449^2)	0.962869

ii. (100,75), (120,90), (140,115), (160,140), (170,155)

	xi	yi	xi*yi	xi*xi	yi*yi
	100	75	7500	10000	5625
	120	90	10800	14400	8100
	140	115	16100	19600	13225
	160	140	22400	25600	19600
	170	155	26350	28900	24025
Total	690	575	83150	98500	70575

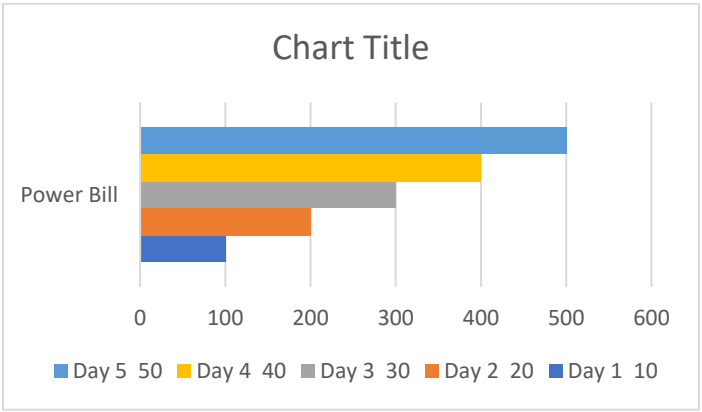
n=8			
m	[n*∑xi*yi-∑xi*∑yi]/n*∑xi*xi-(∑xi)^2	[8*2249-36*449]/8*204-(36)^2	5.440476
c	(∑yi-m*∑xi)/n	(449-5.440476*36)/8	31.642858
Correlation coefficient	[n*∑xi*yi-∑xi*∑yi]/√(n*∑xi*xi-(∑xi^2)*√(n*∑yi*yi-(∑yi^2)	[8*2249-36*449]/√(8*204-(36^2)*√(8*26541-(449^2)	0.962869

Exercises:

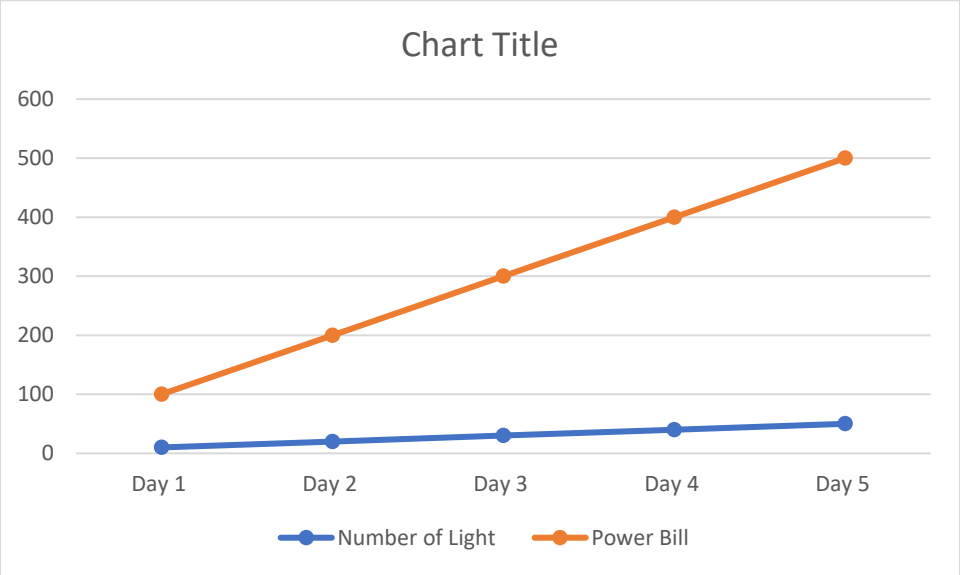
Construct Bar Chart, Line Chart and Pie Chart to depict the following information

Days	Number of Light	Power Bill
Day 1	10	100
Day 2	20	200
Day 3	30	300
Day 4	40	400
Day 5	50	500

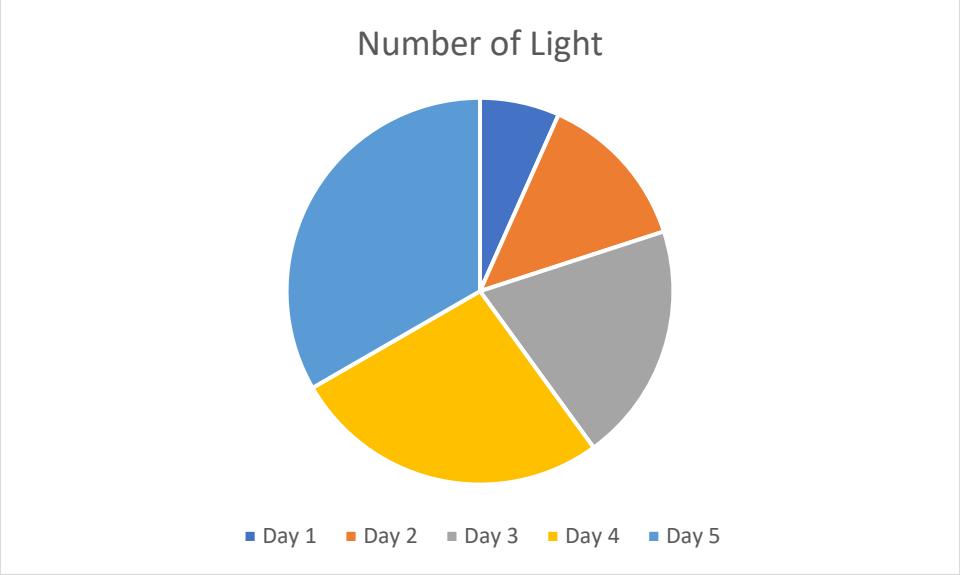
Bar Chart



Line Chart



Pie Chart



Calculate Mean, Median, Mode, Quartiles, Inter Quartile Range (IQR)

I. 8, 15, 18, 13, 15, 28, 30, 17, 41, 27, 17, 17, 44, 31, 24

8	8			
15	13	Mean	(8+13+15+15+17+17+17+18+24+27+28+30+31+41+44)/15	23
18	15	Median	(n+1)/2=(15+1)/2=8th	18
13	15	Mode		17
15	17	Quartiles 1	(n+1)/4=(15+1)/4 =4th	15
28	17	Quartiles 2	(n+1)/2=(15+1)/4 =8th	17
30	17	Quartiles 3	3(n+1)/4=(15+1)/4 =12th	30
17	18	Inter Quartile Range (IQR)	Q3-Q1=30-15	15
41	24			
27	27			
17	28			
17	30			
44	31			
31	41			
24	44			

Calculate the correlation coefficient of the following data set and find the linear regression model

i. (32,30), (36,30), (39,31), (40,39), (41,37), (41,34), (44,37), (45,41)

	xi	yi	xi*yi	xi*xi	yi*yi
	32	30	960	1024	900
	36	30	1080	1296	900
	39	31	1209	1521	961
	40	39	1560	1600	1521
	41	37	1517	1681	1369
	41	34	1394	1681	1156
	44	37	1628	1936	1369
	45	41	1845	2025	1681
Total	318	279	11193	12764	9857

n=8			
m	[n*∑xi*yi-∑xi*∑yi]/n*∑xi*xi-(∑xi)^2	(8 * 11193 - 318 * 279) / (8 * 12764 - 318^2)	0.831984
c	(∑yi-m*∑xi)/n	(279 - (0.831 * 318)) / 8	1.84275
Correlation coefficient	[n*Σxi*yi - Σxi*Σyi] / sqrt[(n*Σxi*xi - (Σxi)^2)*(n*Σyi*yi - (Σyi)^2)]	(8 * 11193 - 318 * 279) / SQRT((8 * 12764 - 318^2) * (8 * 9857 - 279^2))	0.820843

Discussion: