Long questions:

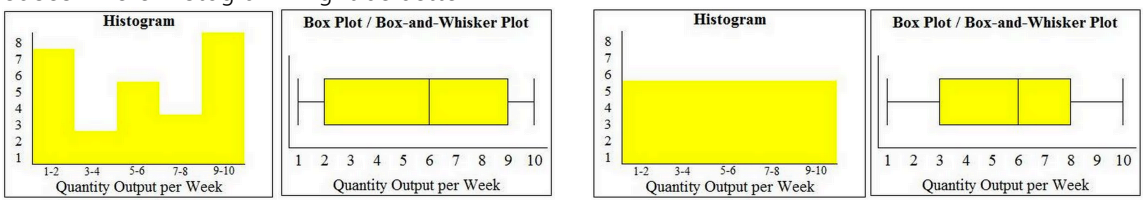
Q1: Give a situation where you would prefer a histogram to box-plot and vice-versa.

Ans:

Histogram is helpful when you want to compare frequencies between different variables. A box-plot is helpful when you want to measure and compare different descriptive values of a set (minimum, first quartile, median, third quartile and maximum). A histogram is useful when you have either wide or low range of variance between the sets. The histogram helps us to determine the peaks within the data and make decision based on that. A simple box plot averages the values and the data appears to be normal. This feature of box plot is helpful to see if the data is normally distributed and not skewed. So we can plot both types of graph and see the different types of result that these present.

For eg:

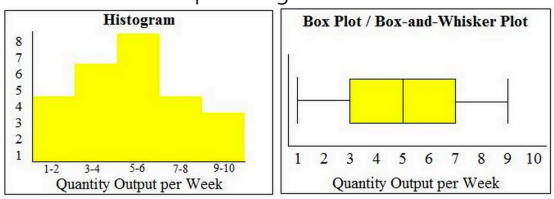
Cases where Histogram might be better.



Case 1 Case 2

The histogram is more helpful because it shows the variance between the different sets which is not visible in the box plot.

Case where box plot might be useful.



The Histogram shows that the data might be skewed to the left but the box plot helps us visualize the even distribution.

Image source: http://www.brighthubpm.com/six-sigma/58254-box-plots-vs-histograms-in-project-management/#imgn_5

Q2: Given the following statistical data for a sample, state the null hypothesis and find out if we are 95% confident that the mean age of the population is 55 years. Use the data mentioned below:

1. Mean age of the sample: 53
2. Standard error: 0.92
3. Sample size: 30
4. The table of critical values below.

degrees of freedom	significance level					
	20%	10%	5%	2%	1%	0·1%
1	3·078	6·314	12·706	31·821	63·657	636·619
2	1·886	2·920	4·303	6·965	9·925	31·598
3	1·638	2·353	3·182	4·541	5·841	12·941
4	1·533	2·132	2·776	3·747	4·604	8·610
5	1·476	2·015	2·571	3·365	4·032	6·859
6	1·440	1·943	2·447	3·143	3·707	5·959
7	1·415	1·895	2·365	2·998	3·499	5·405
8	1·397	1·860	2·306	2·896	3·355	5·041
9	1·383	1·833	2·262	2·821	3·250	4·781
10	1·372	1·812	2·228	2·764	3·169	4·587
11	1·363	1·796	2·201	2·718	3·106	4·437
12	1·356	1·782	2·179	2·681	3·055	4·318
13	1·350	1·771	2·160	2·650	3·012	4·221
14	1·345	1·761	2·145	2·624	2·977	4·140
15	1·341	1·753	2·131	2·602	2·947	4·073
16	1·337	1·746	2·120	2·583	2·921	4·015
17	1·333	1·740	2·110	2·567	2·898	3·965
18	1·330	1·734	2·101	2·552	2·878	3·922
19	1·328	1·729	2·093	2·539	2·861	3·883
20	1·325	1·725	2·086	2·528	2·845	3·850
21	1·323	1·721	2·080	2·518	2·831	3·819
22	1·321	1·717	2·074	2·508	2·819	3·792
23	1·319	1·714	2·069	2·500	2·807	3·767
24	1·318	1·711	2·064	2·492	2·797	3·745
25	1·316	1·708	2·060	2·485	2·787	3·725
26	1·315	1·706	2·056	2·479	2·779	3·707
27	1·314	1·703	2·052	2·473	2·771	3·690
28	1·313	1·701	2·048	2·467	2·763	3·674
29	1·311	1·699	2·043	2·462	2·756	3·659
30	1·310	1·697	2·042	2·457	2·750	3·646
40	1·303	1·684	2·021	2·423	2·704	3·551
60	1·296	1·671	2·000	2·390	2·660	3·460
120	1·289	1·658	1·980	2·158	2·617	3·373
∞	1·282	1·645	1·960	2·326	2·576	3·291

Ans:

α = .05

H_0 = Mean age of the population is 55 years with 95% confidence.

$t = \frac{\bar{X}-\mu}{s_{\bar{X}}}$ = (55 – 53)/.92 = 2.17

From the table we can find the critical value of $t_{.05}$ for (N-1) i.e. for 29 = 2.043

Since $|t| > |t_{.}|$ → H_0 is not true.

Hence we cannot say that the mean age of the population is 55 years with 95% confidence.

Short Questions:

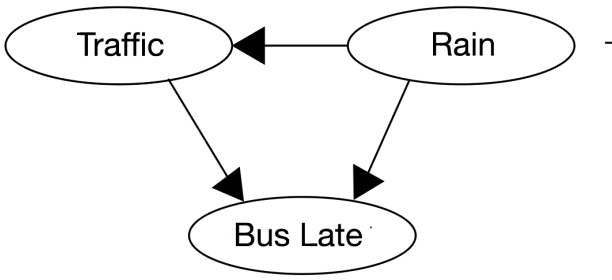
Q3: What is Pruning and what are the different ways of pruning?

Ans:

Pruning is a method to check overfitting in a decision tree. Pruning helps us to remove the unnecessary nodes and reduce the depth of the nodes, which increase the noise in the decision trees. We can do prepruning or postpruning on the tree. It is more accurate to do postpruning as you can see the effect of pruning at each step. When you perform the reduced error pruning, you basically start with the leaf node, replace it with the most popular class and then see the affect on the accuracy. If the accuracy is not affected then we can keep the change otherwise we revert back.

Q4: Consider the following Bayesian network

	Traffic	
Rain	T	F
F	0.4	0.6
T	0.01	0.99



	Rain	
T	F	
0.2	0.8	

		Bus Late	
Traffic	Rain	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Use the network to calculate the probability that if the bus is late then it must be raining?

Ans:

P(Rain) = 0.2

P(Late|Rain) = Probability of the bus being late when rain is true = 0.8*0.2 + 0.99*0.2

P(Late) = Probability that the bus is late = Bus late = true from table 3 and combination of rain and traffic from table 1 = (0.0*0.6 + 0.8*0.99 + 0.9*0.4 + 0.99*0.01)

P(Rain|Late) = P(Late|Rain)*P(Rain) / P(Late)
= (0.8*0.2 + 0.99*0.2)*0.2 / (0.0*0.6 + 0.8*0.99 + 0.9*0.4 + 0.99*0.01)
= 0.06

Q5: What is significance test of correlation and why do you need it?

Ans:

Pearson correlation represents the strength of a relationship between two independent variables. It can be misleading and cause incorrect inferences. For example there could be a very high correlation between rain and foreign exchange rate, but that does not mean that one affects the other. Therefore we should test the significance of the correlation as well. We can do this by doing a t-test for two variables. A t-test will measure how likely is this correlation to exist. We state the confidence level (say 95%) and test if the data can predict such level of probability for that correlation.