Question 2

1)

We have clusters of datasets of salary that are far from each other.We should use.Binning by means was used to properly set the values properly in the right bins.

Bins are decided as:
Bin 1: [1600-2000] contents ->1800,1800,1800,1800 (emp ids: 125,200,205,216)
Bin2: [2000-5000] ->3500 (emp ids: 220)
Bin3: [5000-11000] ->8000 (emp ids: 100,301)

This properly classifies values in the right bin for its value. For example we can infer that technicians don't get more than 2000 so we can set that as the boundary for the first bin. Directors values are close to 8000 and hence we can set bin 3 properly accommodating directors' salaries. Hence the data is properly discretized.

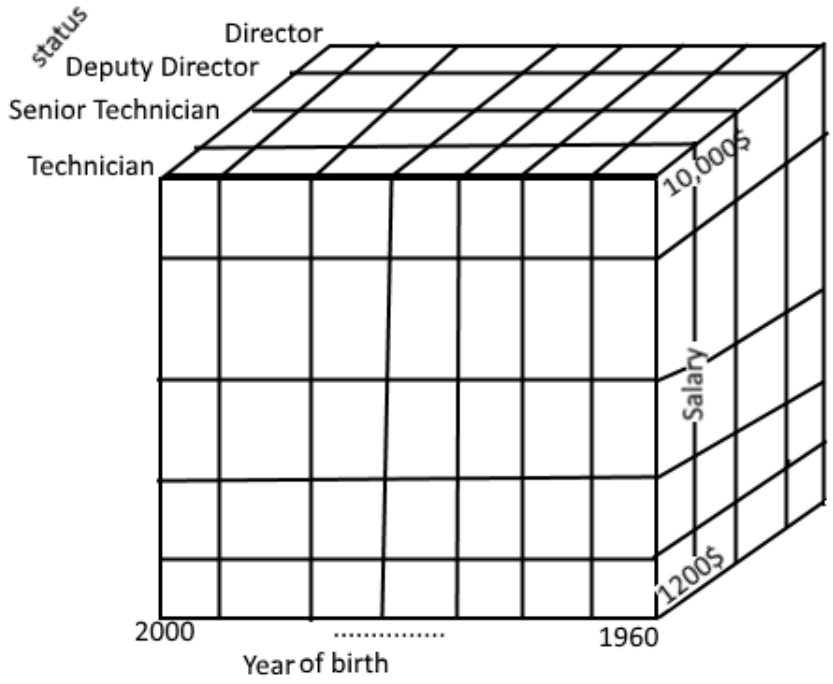
2)

We can set Miss Davis' salary close to 3300 as that is the mean of the value of that bin.One can further infer that status and salary have a very high positive relation and form very distinct clusters of values.
Therefore one can say that since Miss Davis' status is of Senior Technician her salary will be somewhere close to the 'senior technician' clusters centre. We have assumed that to be 3300 hence we can put values from 3400-3600 for the unknown value.

3)

We can set emp id 100 as an outlier as it is the most farthest from the other values. With more datapoints about senior technicians and statuses we can accurately determine the outliers in this situation.

4)

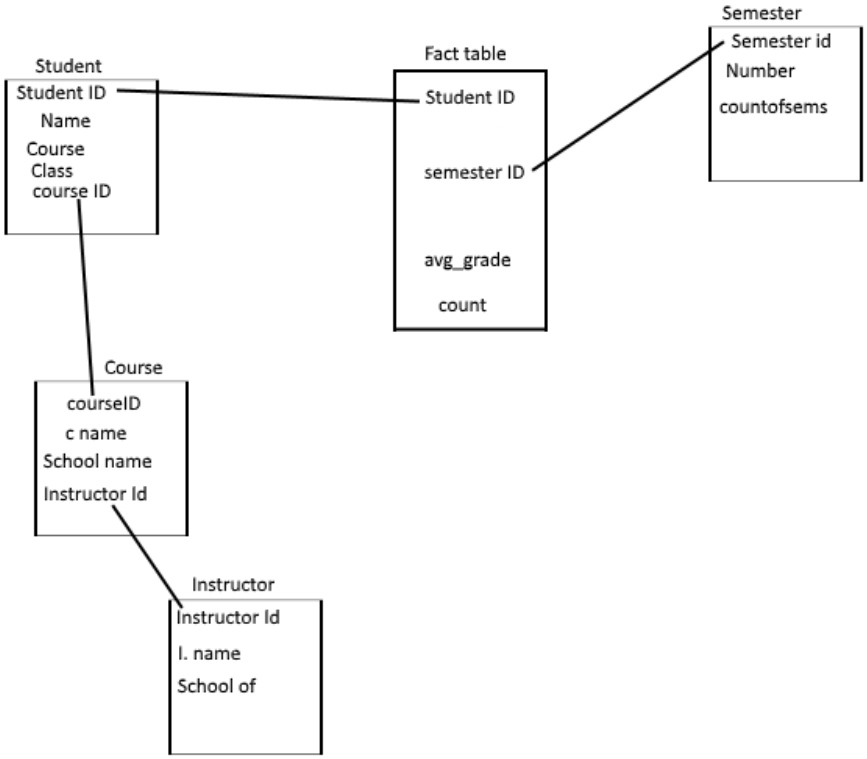


5)
The points in the data cube refer to the one pair of actual data with the corresponding data values. For example for
Year of birth 1960-64,Salary 10,000$ and status director cube refers to all the people with the values of the given attributes will be represented by that cube,.ie Mr Smith and other directors satisfying the conditions

6)
You can slice the cube to just show the attributes we want or to reorient the data for summarization or to make the data more detailed.

- You can pivot the cube to rotate the cube to see it from different view.
- You can drill down the cube by splitting the attributes to show finer detail. This makes the inner cubes smaller and makes the data more detailed
- You can roll up by showing less data by aggregating some particular attributes. This makes the inner cubes bigger and makes the data more summarized as the small cube contain more data items.
- You can slice the cube to show data along only one/two features by slicing the cube. This results in a smaller cube. This leads to highly summarized data view.

Question 4
1)Snowflake schema for university



2)
Slice the instructor out of the cube.
Drill down the details per semester
Drill down course to computer science
Pivot appropriately to show cube with each students grades of computer science course.

3)
There are four dimensions each having 5 levels. The number of cubes will be 4^4. If we drilldown to levels we will get 4*5=20 features, each dimension with 5 features which leads to 20^4.