Data Mining Practical 2

Q1 Data Transformation
Data normalization is adjusting the data to similar scale across dataset. The two data normalization methods that are used are min-max normalization and z-score normalization.

The question is asking us to normalize the data for two columns and calculate the mean of the inputs.

Input 3 column is normalized using the z-score normalization and Input 12 column is normalized using min-max normalization

The mean of all the columns is calculated as

```
#mean of inputs
data['Average Input'] = data1.mean(axis=1)
```

axis=1 will perform the mean operation across rows rather than column.

The output to the file is written as

```

#write the output to output file
data.to_csv('./output/question1_out.csv',index=False)
```

index=False will exclude the index while writing to the csv file

Q2 Data reduction and Discretisation
Principal component Analysis is a procedure that performs orthogonal transformation to a set of data that are possibility correlated into a set which are linearly uncorrelated called principal components.

```
#Apply PCA for 95% variance
pca = PCA(n_components=0.95)
```

n_compnents is set to 0.95 as 95% variance is expected from the data set.

```
pca_generated_data = pca.transform(df)
```

pca.tranform(dataframe) will reduce the dataframe into a dataset with the variance that has been set to.

cut and **qcut** are the two pandas functions that are used to create the bins of equal with and equal frequency respectively.

Both these functions require a data set column and the number of bins to be passed as arguments.