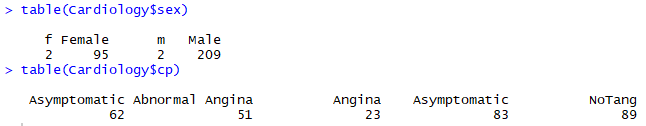
# Enterprise Database Technologies

# CA1

## Getting to know the Dataset using R

Part 1:

After receiving the dataset, I ran a few different commands to initially examine my dataset such as Summary and head. I then ran the table to command to check the columns in my dataset. By running the table command, I was able to get a more in depth look at the values being returned for each column and would be able to see any discrepancies in the data. This resulted in me finding out that two columns both had issues with data; the “Sex” and the “Cp” Column as shown below.



As you can see there are currently four values for “Sex”, when according to the documentation there should be only 2, Male and Female. The same can be seen in “CP” where Asymptomatic can be seen appearing twice. This is due to a white space in front of Asymptomatic meaning it is read as a separate variable. To more accurately evaluate the data, I decided to alter the values so that the extra values in “Sex” and “Cp” were moved back to where they were intended be.

Next, I moved on to find the percentage of missing values within the data. I ran a command on the dataset to query how many values were Na in the dataset. This returned me a result of 7 Na’s which gave me a percentage of 0.1517. I then ran the summary command on the dataset to quickly evaluate which columns had Na values. This showed me that the Cholesterol, Restecg and Class all had missing values and would need to have values imputed later.

Next, I found the max, min, mean, mode, median and the Standard deviation of the data to see if any information could be gained. The standard deviation was used to tell me how dispersed the data would be; A low standard deviation indicated it was closer to the mean and less dispersed, while a higher standard deviation indicated it was more dispersed.