PROJECT REPORT

DATA ANALYTICS - CS40003

TOPIC -5

Roll no: 16CS30033   Name: Shaikh Moin Dastagir
Mob no: 7709400624

- The language used for completing this project is R.

- The software used for completing this project is RStudio.

- For Labeled Data Encoding I have first converting the non-numeric columns to factors, then used the match function to map the argument 1 with the argument 2 , so that I will get a vector with distinct numbers for each categorical value(Please see additional explanation provided in the comment in the .R file regarding this).

- For one Hot Encoding basically I made new columns for each value possible under the labeled columns and assigned value 1 if the corresponding value is there and assigned 0 otherwise.

- The sorting operation with respect to Hours_Worked is applied on the preprocessed Data and the result is stored in the corresponding .csv file.I have used the order function to achieve that operation.

- Only rows with United_States operation is applied on the preprocessed Data Set and the corresponding result is stored in the .csv file(The method is self explanatory)

- The subset division with respect to Male and Female is applied on the Preprocessed Data set.