**Data Analytics (Assignment 1)**

**Name: Kavish Bhathija**

**Student number: 18302335**

**Description of dataset:**

The dataset consists of 944 columns and 1501 rows. This makes a total of 1416944 data points.

**Aim:**

Our task is to impute the variable (Column) “rearn” with the help of linear and logistic regression. Data imputation is done in Rstudio.

**Stepwise approach to Data Imputation:**

1. The dataset is imported in the Rstudio. After that, the dataset is visualized using a standard scatter plot. Due to space restrictions, the graph is not attached here.
2. Then, the correlation matrix is generated to find out the variables which are the most dependent on the variable “rearn”. In order to do so, the data has to be cleaned i.e. the number of columns which have greater than 1000 NA values are rejected. Then, each column is converted to integer. This is done due to syntax limitation. Finally, a correlation matrix is generated and according to this matrix, five features are selected; 'men', 'ag65m', 'white ', 'immig2' ,'educ\_r', 'zwelfare' and 'rearn'.
3. The second step is to impute all the missing values in the selected feature matrix. For the concerned purpose, MICE library is used. MICE can use different methods to impute the data. For our purpose, ‘cart’ method is used which is an acronym for classification and regression trees. This method was chosen due to research done online.
4. After this, Logistic regression is applied. To successfully apply it, first a new dataset with only the selected features along with our dependent variable is taken. A new column is generated which replaces positive values in ‘rearn’ with 1, 0 and negative values as 0 while keeping NA values constant. With the help of this new variable, Logistic regression is applied, and predictions are taken on the NA values.
5. At the end of the step 3, the ‘rearn’ dependent variable is updated with the newly predicted values from logistic regression algorithm; negative values are updated as 0, positive values as well as NA values are retained as it is.
6. Further, the newly added column is removed, and Linear regression model is finally applied. The predicted values are updated on the original dataset.
7. Standard deviation and Mean of the data is calculated from this imputed dependent variable

**Results:**

Standard Deviation = 45267

Mean = 29246