**ANALYSIS:**

* + - This type of regression is used in binary classification. It uses the sigmoid function: f(x) = 1/(1+exp(-x))
    - This function outputs a value between 0 and 1 logistically proportionate to any value of input from negative infinity to infinity.
    - Basically a linear regression expression is used as an input to this function (AKA sigmoid function) and the output is taken as the probabilistic likelihood towards any binary choice.
    - 0 to 0.5 is attributed towards a tendency towards a particular binary classification and 0.5 to 1 is attributed to a tendency towards the other.
    - Importing the data:
      * The data is imported from the ‘statsmodels’ package using their own method of importing into a dataframe.
    - Exploratory analysis and feature engineering:
      * The dataframe is then analysed.
      * A new column ‘had\_affair’ is created and takes binary values of 1 and 0 based on whether the affairs column is non zero or zero.
      * The dataframe is grouped by the ‘had\_affair’ column and the mean is calculated for every other feature, distinguished on whether the ‘had\_affair’ value is 0 or 1.
      * Then a series of plots using seaborn is done to analyse the effect of factors like age, education, number of children etc. on the extramarital affair tendency.
      * Once the dataframe is set up, we realise that the ‘occupation’ and ‘occupation\_husb’ are categorical variables. Namely, the values hold qualitative discrete values instead of along a spectrum, and hence using them as is into the regression formula would distort the prediction. Therefore, first these variables are split into more columns in the form of dummy variables. A dataframe of these dummy variable is created using the ‘get\_dummies()’ in the pandas library . then these are concatenated into the original dataframe while dropping the redundant ‘occupation’ and ‘occupation\_husb’ features.
      * Now, we prepare the data for fitting by separating the ‘had\_affair’ column into the Y set (target set) and the rest into the X set. (we drop the ‘affairs’ column in X so that the prediction is meaningful)
    - Multicollinearity check and Logistic regression :
      * Now that we have the dummy variables as columns. But this poses a problem. These dummy variables can cause one of them to be expressed as a linear combination of the other causing some of the features to be more weighted then the others. This is called Multicollinearity. This is also called as the dummy variable trap. Therefore to prevent this we take away one column from every set of dummy variables.
      * Once that is done, we start our Logistic regression model by frat creating a LogisticRegression object.
      * Then we split the (X,Y) into training and testing set using the ‘train\_test\_split()’ function imported from ‘sklearn.model\_selection’.
      * Once that is done we fit the training sets, predict using the test set then store the predicted set and then measure the score of the prediction. The predicted set is compared with the Y\_test for evaluation of accuracy.
      * the accuracy comes out to be approximately 73%
      * we can further increase the accuracy by :
        + regularization
        + non-linear modelling