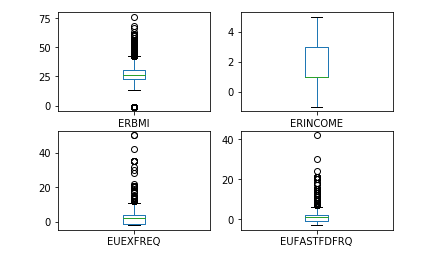
Machine learning models for the data ehresp\_2015

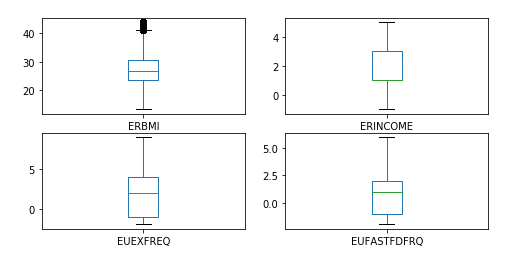
**Data exploration and cleanings**

The assignment required to use only four features; ERBMI, ERINCOME, EUEXFREQ, EUFASTFDFRQ from the EH\_RESP2015 dataset and to create an indicator variable OVERWEIGHT with binary cases of 0 representing respondent with ERBMI <25.0 and 1 representing respondents with ERBMI >25.0. the data had to be clean by removing all the missing values and the outliers. This is to ensure that the involved features attain a gaussian distribution for accurate prediction and classification of the previously unseen features by the models. Before cleaning the data there were 10626 rows and after the cleaning the rows reduced to 9309 for the feature. This are the cases that were used in training the regression and classification machine learning models. Quick look on the distribution of the feature before and after removing outliers using box plots.

Before removing outliers



After removing outliers, the data look normal (gaussian distribution)



**Regression algorithms**

Part 2 of the assignment required to build prediction models that will be predicting body mass index (ERBMI) using the variables ERINCOME, EUEXFREQ, and EUFASTFDFRQ. Regression models employs statistical techniques to predicting continuous data from other continuous numerical features. In this task, both multiple regression and KNN from KNeighborRegressor class from scikit-learn module in python were used.

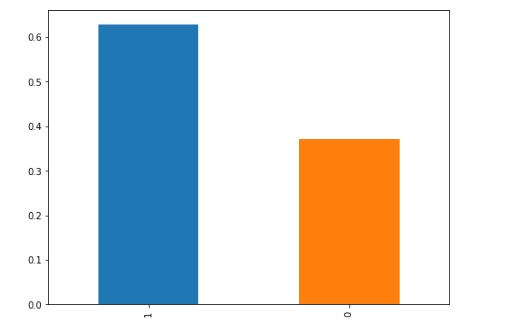
On training the KNN model and the multiple regression both on 10 cross validation splits to determine how well is ERBMI is predicted the variable mentioned above it was found to have low accuracies of 0.167 and 0.031 respectively based on the R-squared criteria. Converting into percentage, only 16.7% and 3.1% of predictions were being accounted by the both models respective. Further, I trained support vector machines for regression and decision tree regressor algorithms to predict ERBMI using the three features, the accuracies attained were still very low. 0.011 for decision tree regressor and 0.010 for support vector machines.

**Finding from the regression algorithms**

The features ERINCOME, EUEXFREQ, and EUFASTFDFRQ are not good enough to predict ERBMI using regression methods. As seen above, their accuracy level are so low to adopt the models

**Classification algorithms**

The assignment specified to create an indicator variable OVERWEIGHT that will contain binary values 1, if ERBMI >25.0 and 0 if ERBMI <25.0. on determining the distribution of the overweight variable, it was found that about 63% of cases had ERBMI >25.0 while about 37% of cases had ERBMI < 25.0. the chart below shows.



I trained classification models for overweight using ERINCOME, EUEXFREQ, and EUFASTFDFRQ as the predictor and obtained the results as follows. First, KNN classifier model trained on 10-fold cross validation splits attained an accuracy of 58.2%. Second Logistic Regression model trained on 10 folds cross validation split attained an accuracy of 62.9%, and third, Gaussian naïve Bayes model trained on 10 folds cross validation splits scored an accuracy of 62.6%. further I decided to use boosting method for classification using Adaboost classifier algorithm and Gradient boosting algorithm both on 10 folds cross validation splits. Both of the boosting classification algorithms attained same accuracy of 63.1% which is higher than the three classification algorithms used.

**Findings from classification models**.

Classification models on predicting whether a person is overweight performs fairly well with average accuracy of 61.23% using ERINCOME, EUEXFREQ, and EUFASTFDFRQ as the predictors.

**Recommendation based on the findings**

Although the prediction of the classification algorithms performs fairly well more than 36% of miss classification is still a big concern that need to be looked at. A research should be conducted that will bring more features that are significant enough, with less pattern complexity within the data that machine learning will be able to generalize and bring high accuracy when predicting. The algorithms to predict ERBMI performs poorly and they should not be used at all