**HMW2**

1a) By fitting each predictor individually with the response (per capita crime), we get the following results:

i) CRIM-ZN(proportion of residential land zoned for lots over 25,000 sqft): As evident from the ANOVA results, p-value is small, so this predictor shows a statistically significant relationship with the response variable. The scatterplot shows many outliers which means a low R² value and it is verified with a 0.04 value in ANOVA results.

ii) CRIM-INDUS(proportion of non-retail business acres per town): As evident from the ANOVA results, p-value is small, so this predictor shows a statistically significant relationship with the response variable. The scatterplot shows many outliers which means a low R² value and it is verified with a 0.165 value in ANOVA results.

iii) CRIM-CHAS(Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)): As evident from the ANOVA results, p-value is large, so this predictor does not show a statistically significant relationship with the response variable. The scatterplot shows many outliers which means a low R² value and it is verified with a 0.03 value in ANOVA results.

iv) CRIM-NOX(nitric oxides concentration (parts per 10 million)): As evident from the ANOVA results, p-value is small, so this predictor shows a statistically significant relationship with the response variable. The scatterplot shows many outliers which means a low R² value and it is verified with a 0.177 value in ANOVA results.

v) CRIM-RM(average number of rooms per dwelling): As evident from the ANOVA results, p-value is small, so this predictor shows a statistically significant relationship with the response variable. The scatterplot shows many outliers which means a low R² value and it is verified with a 0.048 value in ANOVA results.

vi) CRIM-AGE(proportion of owner-occupied units built prior to 1940): As evident from the ANOVA results, p-value is small, so this predictor shows a statistically significant relationship with the response variable. The scatterplot shows many outliers which means a low R² value and it is verified with a 0.124 value in ANOVA results.

vii) CRIM-DIS(weighted distances to five Boston employment centres): As evident from the ANOVA results, p-value is small, so this predictor shows a statistically significant relationship with the response variable. The scatterplot shows many outliers which means a low R² value and it is verified with a 0.144 value in ANOVA results.

viii) CRIM-RAD(index of accessibility to radial highways): As evident from the ANOVA results, p-value is small, so this predictor shows a statistically significant relationship with the response variable. The scatterplot shows relatively few outliers which means a better R² value and it is verified with a 0.391 value in ANOVA results.

ix) CRIM-TAX(full-value property-tax rate per $10,000): As evident from the ANOVA results, p-value is small, so this predictor shows a statistically significant relationship with the response variable. The scatterplot shows relatively few outliers which means a better R² value and it is verified with a 0.34 value in ANOVA results.

x) CRIM-PTRATIO(pupil-teacher ratio by town): As evident from the ANOVA results, p-value is small, so this predictor shows a statistically significant relationship with the response variable. The scatterplot shows many outliers which means a low R² value and it is verified with a 0.084 value in ANOVA results.

xi) CRIM-B(1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town): As evident from the ANOVA results, p-value is small, so this predictor shows a statistically significant relationship with the response variable. The scatterplot shows many outliers which means a low R² value and it is verified with a 0.148 value in ANOVA results.

xii) CRIM-LSTAT(% lower status of the population): As evident from the ANOVA results, p-value is small, so this predictor shows a statistically significant relationship with the response variable. The scatterplot shows many outliers which means a low R² value and it is verified with a 0.208 value in ANOVA results.

1b) When we fit the multivariate linear regression coefficient, we see significant interactions between the following predictor variables: DIS (p-value=0.009), RAD (p-value=0), B (p-value=0.009) & LSTAT (p-value=0.001)

1c) After fitting the multivariate linear regression model we observe that many of the variables which had significant interaction individually with the response variable were now insignificant in the multivariate model. This happens because of the effect of addition of other response variable and their combined effect on the response variable which makes many of the variables which were earlier significant now insignificant.

1d) By fitting a polynomial model of degree 3 for each predictor individually, we observe non linear observation between these response variables: INDUS (p-value=0), NOX (p-value=0), AGE (p-value=0.047 & 0.007), DIS (p-value=0) & PTRATIO (p-value=0.03, 0.04, 0.06).

1e) When we fit Ridge regression and Lasso to our model for regularization, we see that the number of non-zero features for ridge regression is 12 whereas for lasso, it is 1 (i.e., RAD with coeff of 0.23). While if we compare the R² values for two regularization methods, we notice that ridge has a higher value than lasso. This means that ridge imposes a smaller penalty on the coefficients and thus the model obtained by ridge regression is able to explain more variance in the dataset than the lasso model. By performing cross validation on Ridge regression, we notice that the highest mean cross validation score is obtained at alpha value of 10, so 10 is the optimal regularization parameter for ridge regression. Also, by performing cross validation on lasso, we notice that the highest mean cross validation score is obtained at alpha value of 0.5, so 0.5 is the optimal regularization parameter for lasso.

2 We observe that the median value for the response variable (i.e. per capita crime rate) is 0.25651. So, I assigned the dummy variable subject to the following condition: any value above the median crime rate is assigned the dummy variable of 1 and below the median value is assigned the dummy variable of 0. When we fit Logistic Regression, KNN and Naïve-Bayes Classifier, we observe that Logistic Regression model has the highest accuracy score among the three models of 0.854 which means that Logistic Regression is the best choice for this type of model. The mean cross validation score is the highest for KNN model which is 0.785.

Logistic Regression: This model fits the dataset with an accuracy score of 0.854 which is pretty good. When we apply 3-fold cross-validation for subset selection in this model, we observe that mean cross-validation score is 0.769 with the highest score of 0.93 for the model with the last subset selected as the testing data and the first 2/3 part of the dataset as the training data.

KNN: This model fits the dataset with an accuracy score of 0.847 which is also pretty good. When we apply 3-fold cross-validation for subset selection in this model, we observe that mean cross-validation score is 0.785 with the highest score of 0.94 for the model with the last subset selected as the testing data and the first 2/3 part of the dataset as the training data.

Naïve-Bayes Classifier: This model fits the dataset with an accuracy score of 0.83 which is also pretty good. When we apply 3-fold cross-validation for subset selection in this model, we observe that mean cross-validation score is 0.769 with the highest score of 0.93 for the model with the last subset selected as the testing data and the first 2/3 part of the dataset as the training data.