Homework 1

Freya Genesis D Mello

September 14, 2018

1 Problem Statement 1

Consider the Student Performance Data Set on the UCI machine learning repository. Suppose that you are getting this data in order to build a predictive model for First Period Grades. Using the full dataset, investigate the data using exploratory data analysis such as scatterplots, and other tools we have discussed in class. Preprocess this data and justify your choices (elimination of outliers, elimination of variables, variable transformations, etc.) in your write up. Submit the cleaned dataset as an *.RData file.

1.1 Merging datasets

The datasets provided were separate; containing the grades of the students from two different subjects - Math and Portuguese and the two of them were merged to form a resultant dataset of 382 students.

1.2 Second and Final Year Grades not considered

In the mission to build a predictive model for First Period Grades, the columns G2 and G3 containing the second and final year grades are not considered. Other columns such as guardian, travel time are common across both subjects - Math and Portuguese and hence, the ".y" (one common variables on Portuguese data set) corresponding variables are ignored.

1.3 Elimination of Outliers

On observing the data, we first see the student count in both schools - GP and MS. The number of students in MS are comparatively lesser than the ones in GP but it cannot be ignored as they belong to a safe age range. (See fig. 1 and fig. 2)

The grades across both subjects were considered and plotted to realize the density. The number of students scoring higher grades were more in the Math subject than in Portuguese. (See fig. 3)

When the grades were compared against the age of the students, we see that there's no effect of it on the grades. However, we do spot an Outlier which is a student of the age 22. We see that the student of age 22 has 3 failures and 16 absences, 1 hour of travel and 1 hour of study time but instead has a free time of 4 hours. He goes out more than he is required to and thus, is distant from all other observation points i.e students. (Refer fig. 4)

1.4 Data Investigation

Students invest more free time than study or travel time as the distribution suggests in fig. 5

On observing the Quantity of family relationships versus the average grades on each level, we see that the ones with good to excellent levels (4 and 5) have better average grading than the rest. (Refer fig. 6)

Students across Math and Portuguese subjects, after investing similar study time perform better in Portuguese. (Refer 7)

Students across Math and Portuguese subjects, after investing similar going out time perform better in Portuguese. (Refer 8)

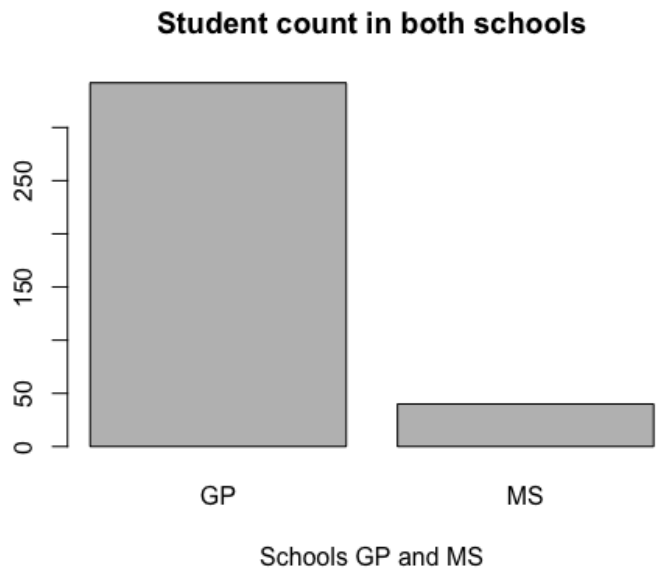

Figure 1: Histogram - Student count in both schools

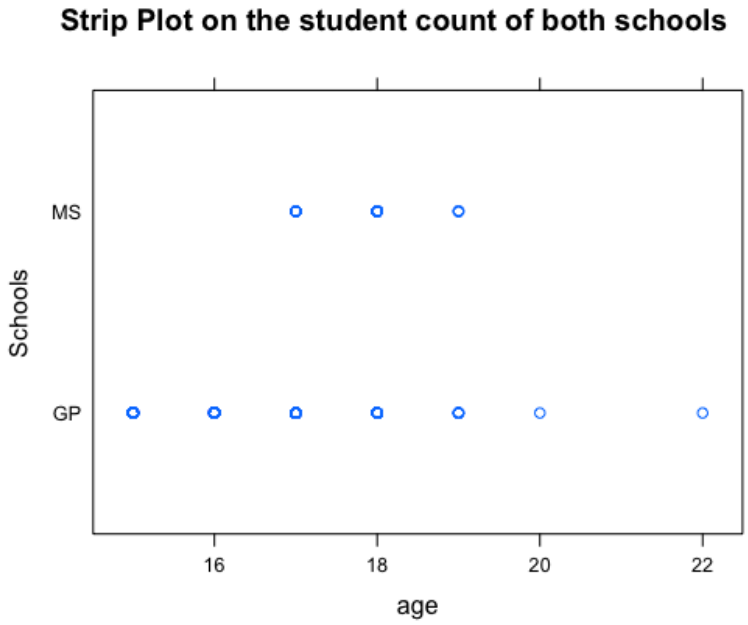

Figure 2: Strip Plot on the student count of both schools

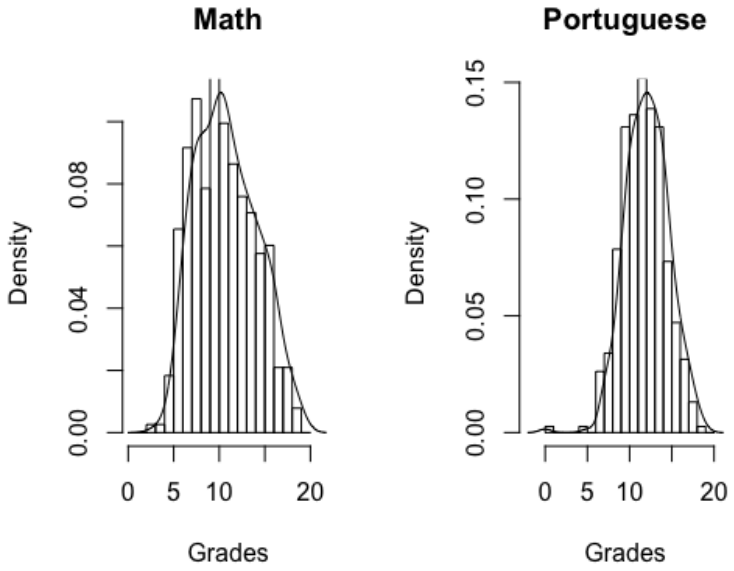

Figure 3:

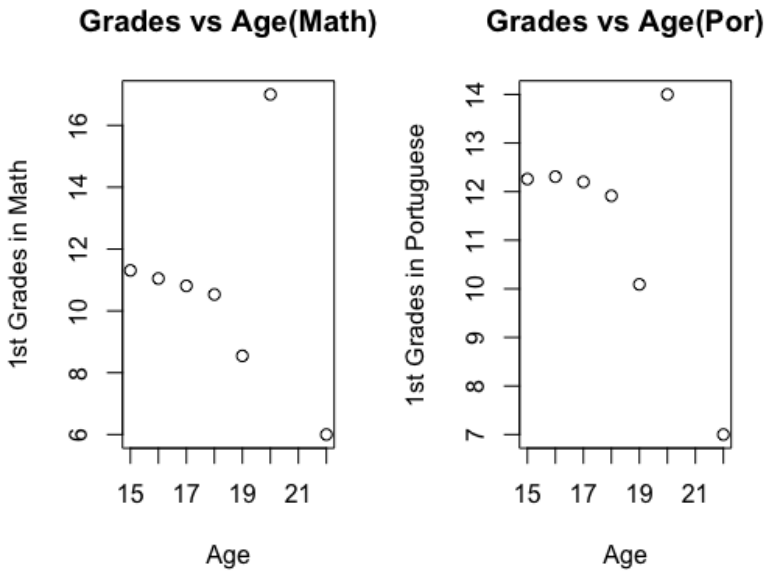

Figure 4: Scatter plot - Grades vs Age

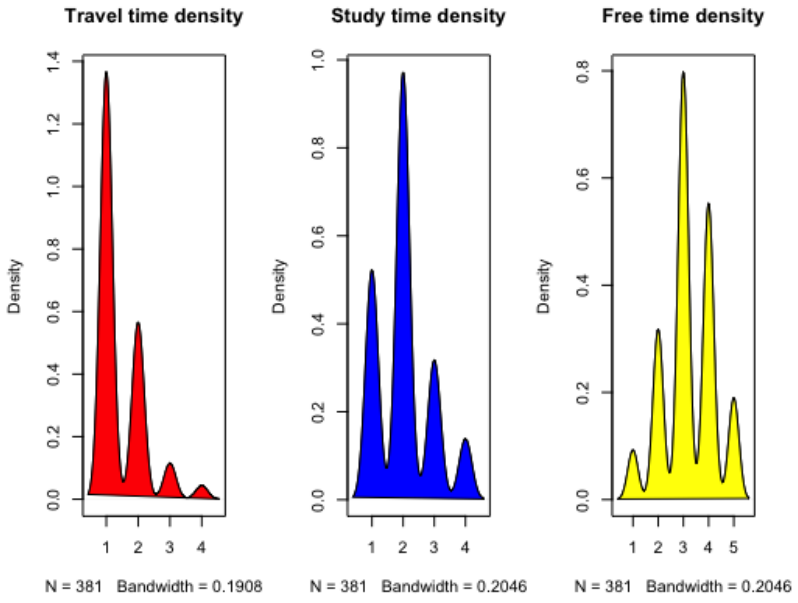

Figure 5: Travel vs Study vs Free time

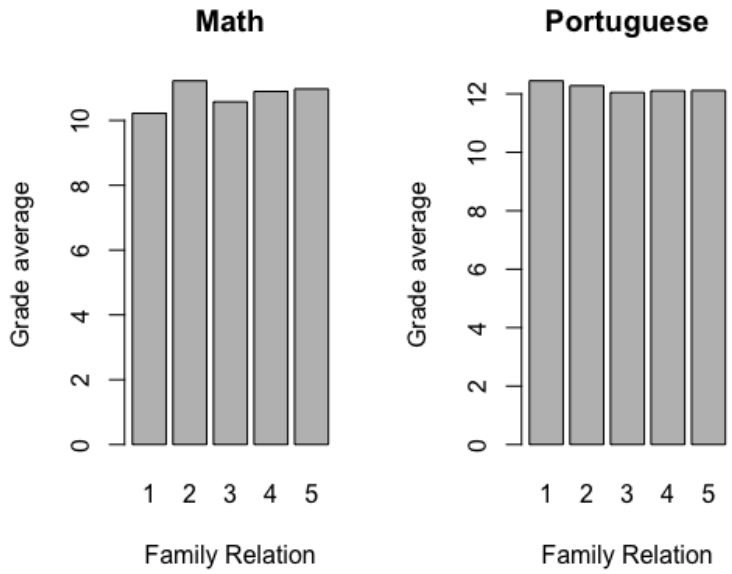

Figure 6: Quantity of family relationships versus the average grades

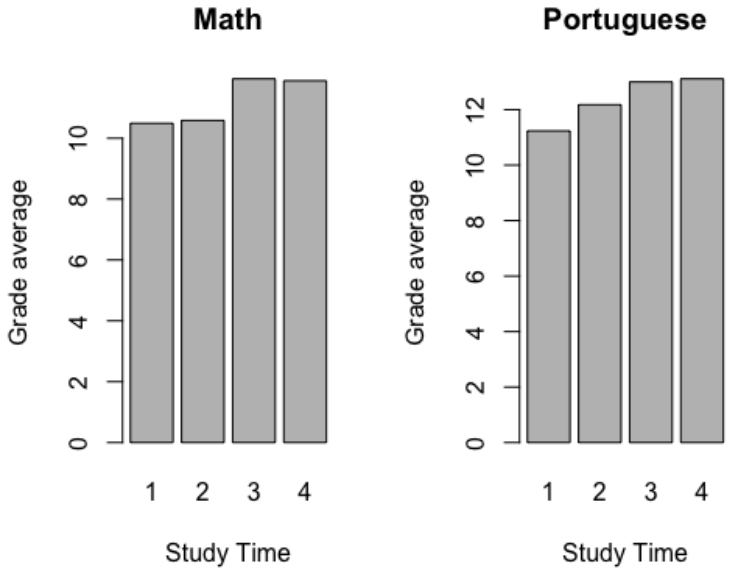

Figure 7: Quantity of Study time vs grades

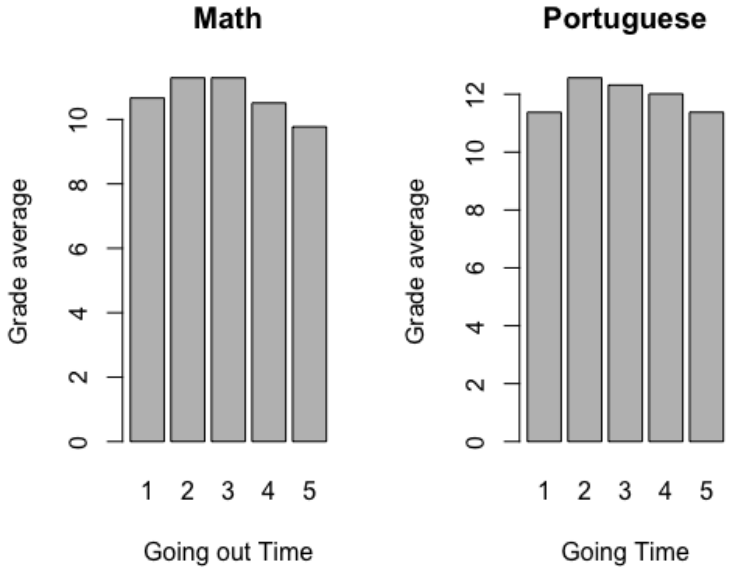

Figure 8: Quantity of Going out time vs grades

1.5 Elimination of Variables

In the mission to build a predictive model for First Period Grades, the columns G2 and G3 containing the second and final year grades are not considered. Other columns such as guardian, travel time are common across both subjects - Math and Portuguese and hence, the ".y" (one common variables on Portuguese data set) corresponding variables are ignored.

2 Problem Statement 2

Perform a multiple regression on the dataset you pre-processed in question one. The response are the first period grades. Use the lm() function in R. a) Which predictors appear to have a significant relationship to the response. b) What suggestions would you make to a first-year student trying to achieve good grades. c) Use the * and : symbols to fit models with interactions. Are there any interactions that are significant?

2.1 Predictors related to first period grades

Using lm(), two linear regression models with respect to first period grades and the two subjects - Math and Portuguese is built. On summarizing the linear model, thus created, it is observed that lesser the PR(>|t|), more significant the variables are to the Grades. Variables thus chosen as predictors are failures, school support, study time, family support, sex, going out, higher education, Father's job (services), Father's job (other).

2.2 Suggestions to improve first period grades

2.2.1 Failures

The relation between the failure data vs Grades is such that the ones with least number of failures (0) were of the majority and secured higher grades than the rest. The students must manage to reduce past failures in order to secure higher grades.

2.2.2 Study Time

If more number of study hours are invested, the grades will come higher. The students majorly invest 1-2 hours comparative to lesser number of students who spend more hours studying.

2.2.3 Going out

This suggests a balance between study time and leisure time. If we observe the stats, the students who go out for an hour or five hours, which are the 2 extreme cases, do not manage to score as well as the ones who invest 2-4 hours on leisure time.

2.2.4 Higher Education

The students who have aimed for higher education have managed to secure higher grades than the others. The median value of the grades of the students who have plans for higher education is about 12 whereas the latter have a median value of about 7.

2.3 Interactions

Considering the summary values on the linear regression models made over Math and Portuguese data, we have our adjusted R squared values of 0.287 and 0.2473 respectively.

An interaction model is made on the Math data taking the grades as an output vs intercepts that have a more significant coefficient value - failures, school support, family support, study time, going out. The model summarizes the different interactive variables to make compute summaries. One where the interaction considers grades in math vs family support, study time, going out and failures and school support. The other has the interaction that considers grades in math vs family support, study time, going out,failures, school support.

In the case of Portuguese data, an interaction model is made on the Portuguese data taking the grades as an output vs intercepts that have a more significant coefficient value - health, family

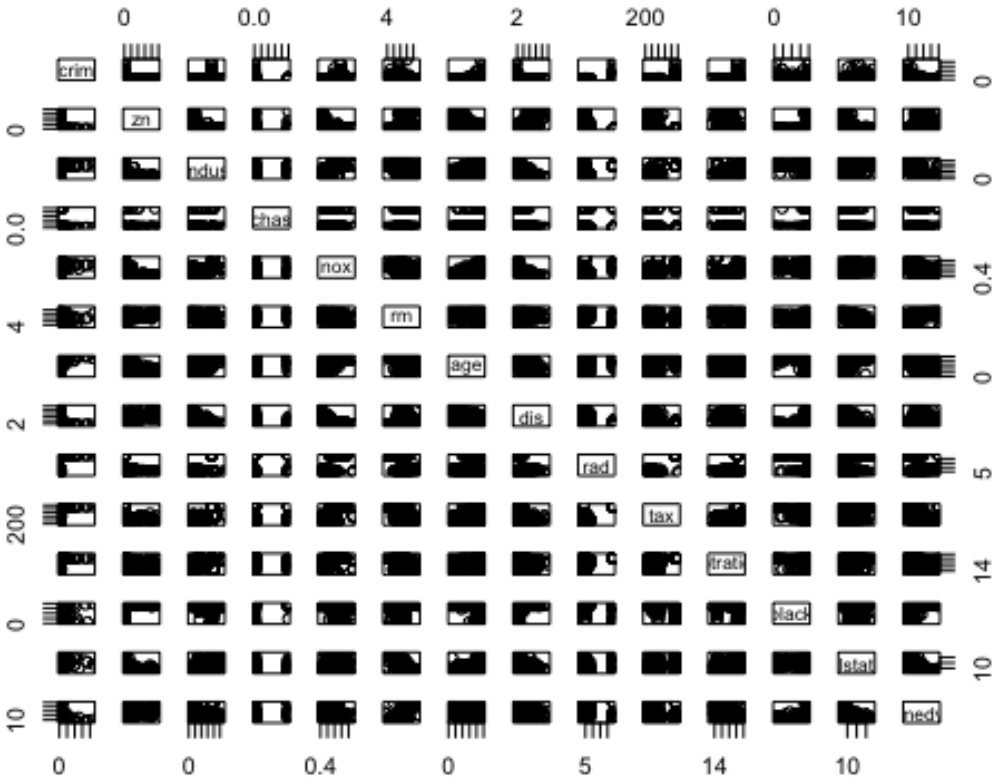

Figure 9: Pairwise plotting between all the variables of the data set.

size, school, sex, failures, school support. The model summarizes the different interactive variables to make compute summaries. One where the interaction considers grades in Portuguese vs health, family size, school, sex and failures and school support. The other has the interaction that considers grades in Portuguese vs health, family size, school, sex, failures, school support.

3 Problem Statement 3

ISL textbook exercise 2.10 modified: This exercise concerns the Boston housing data in the MASS library (>library(MASS) >data(Boston)). a) Make pairwise scatter plots of the predictors, and describe your findings. b) Are any of the predictors associated with per capita crime rate? c) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor. d) In this data set, how many of the suburbs average more than seen rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

3.1 Loading data

Housing Values in Suburbs of Boston are loaded into a data frame "BostonHousDF" and pairwise scatter plots of the predictors are plotted. (See fig. 9)

3.2 Study on the pairwise plottings

3.2.1 Crime rate vs Median Value of Houses

Here, a scatter plot is plotted considering Crime rate vs Median Value of Houses. A negative correlation is observed - as the price of the houses keep increasing, the crime rate decreases.

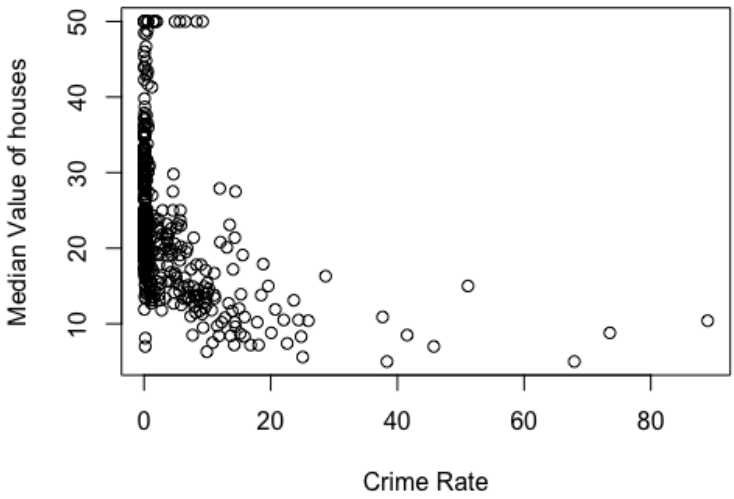

Figure 10: Crime rate vs Median Value of Houses

Majority of the crime rates are less than 0. (See fig. 10)

3.2.2 Lower status population vs Median Value of Houses

Here, a scatter plot is plotted considering Lower status population vs Median Value of Houses. A negative correlation is observed - As the sections of lower status population broaden, the median value of the dwellings decrease. Majority of lower status of the population live in houses worth lower than 20000. (See fig. 11)

3.2.3 Nitrogen Oxide Concentration vs Distances to employment centres

Here, a scatter plot is plotted considering Nitrogen Oxide Concentration vs Distances to employment centres. A negative correlation is observed - As the house to Boston employment centers distance increases i.e. away from the industrial areas, the nitrogen oxides concentration decreases. (See fig. 12)

3.2.4 Age vs Tax

Here, a scatter plot is plotted considering the proportion of owner-occupied units built prior to 1940. vs full-value property-tax rate per USD10,000. A higher proportion of the owner-occupied units built prior to 1940 pay comparatively lesser full value property tax. The older the building, the lesser the tax paid (See fig. 13)

3.3 Predictors associated with per capita crime rate

Refer to the table 1. Here, the correlation values are computed against each of the variables with respect to per capita crime rate.

3.4 Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor

3.4.1 Crime Rates

On observing the crime rate frequencies, it appears that the crime rates 20 and above are at the higher range. Out the total 506 suburbs, 18 are of higher crime rate 3.55percent. (See Fig. 14)

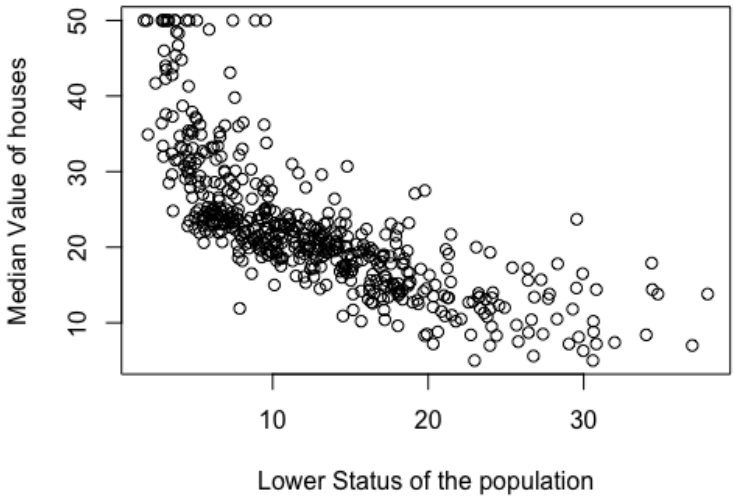

Figure 11: Lower status population vs Median Value of Houses

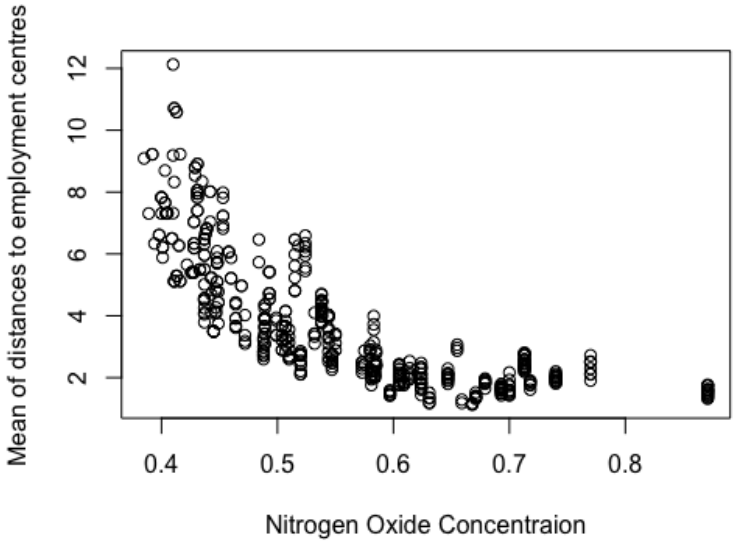

Figure 12: Nitrogen Oxide Concentration vs Distances to employment centres

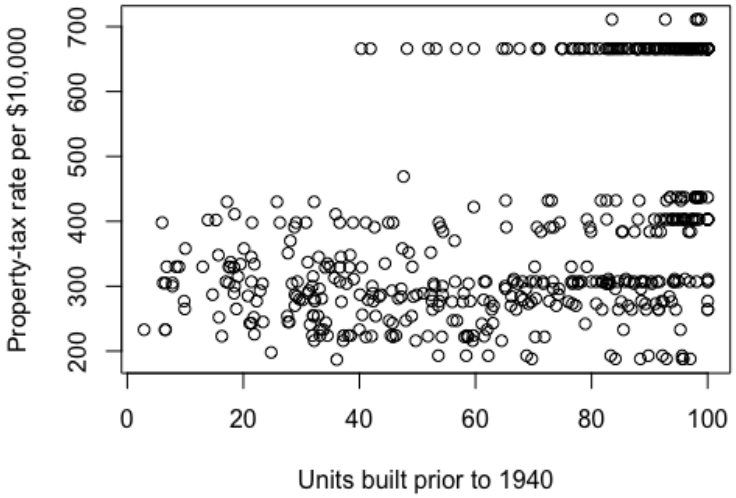

Figure 13: Age(built prior to 1940) vs Property Tax

Variable	Coefficient
rad	0.62550515
tax	0.58276431
lstat	0.45562148
nox	0.42097171
indus	0.40658341
age	0.35273425
ptratio	0.28994558
chas	-0.05589158
zn	-0.20046922
rm	-0.2192467
dis	-0.37967009
black	-0.38506394
medv	-0.38830461

Table 1: Predictors associated with per capita crime rate

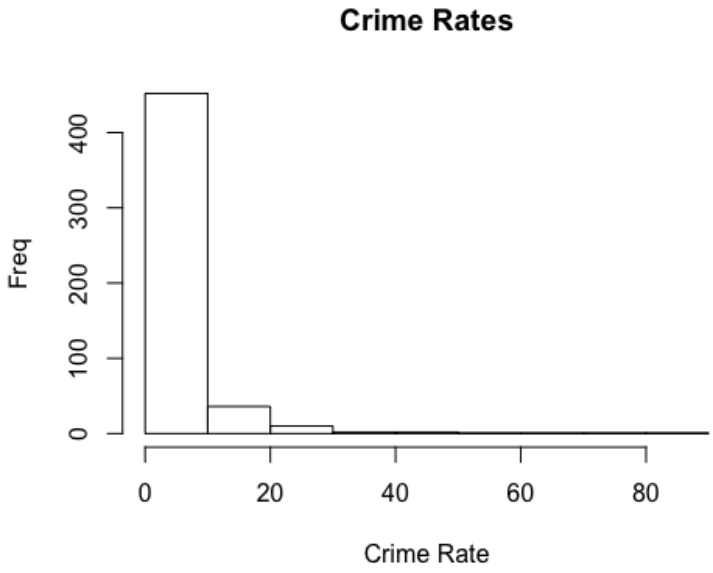

Figure 14: Crime Rates Frequency

3.4.2 Tax Rates

On observing the tax rate frequencies, it appears that the tax rates 650 and above are at the higher range as per the value at the third quartile of the histogram. Out the total 506 suburbs, 137 are of higher tax rate 27.07percent. (See Fig. 15)

3.4.3 Pupil Teacher Ratio

On observing the Pupil Teacher Ratio frequencies, it appears that the Pupil Teacher Ratio 20 and above are at the higher range. Out the total 506 suburbs, 201 are of higher Pupil Teacher Ratio 39.72percent. (See Fig. 16)

3.5 The suburbs that average more than 7/8 rooms per dwelling

The variable readings were plotted on a histogram to see the distribution of the number of dwellings vs its frequency among the suburbs. (See Fig. 17)

Its observed that dwellings that have rooms greater than 7 are 64 out of the 506 dwellings which makes it 12.6percent of the total number of dwellings.

Also, dwellings that have rooms greater than 8 are 13 out of the 506 dwellings which makes it 2.5percent of the total number of dwellings.

4 Problem Statement 4

ESL textbook exercise 2.8 modified: Compare the classification performance of linear regression and k-nearest neighbor classification on the zipcode data. In particular, consider only the 2's and 3's for this problem, and k = 1,3,5,7,9,11, 13,15. Show both the training and the test error for each choice of k.

4.1 Loading data

The zipcode data is available in the ElemStatLearn package. The training and test data were already provided.
>?zip.test
>?zip.train The zip code data considered had only 2's and 3's normalized pixel values

Figure 15: Tax Rates Frequency

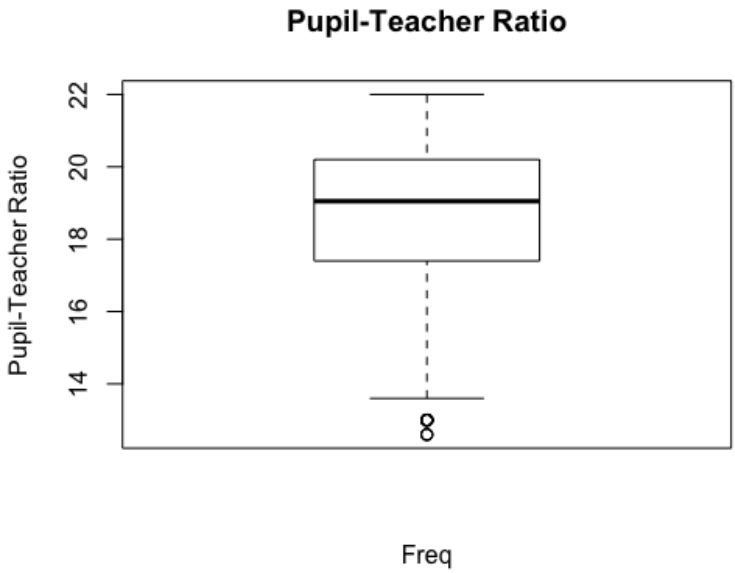

Figure 16: Pupil Teacher Ratio Frequency

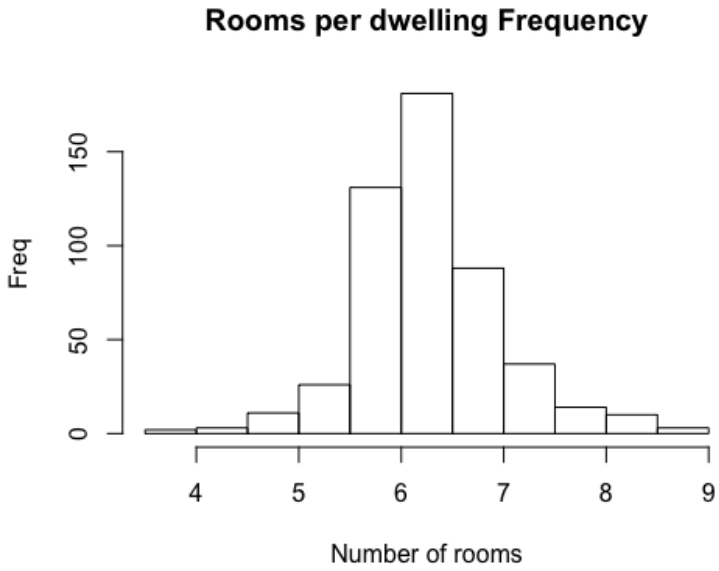

Figure 17: The suburbs that average more than 7 or 8 rooms per dwelling

4.2 Building the linear regression model

The linear regression model was built on the training data set considering the 2's and 3's values against the rest of the variables that defined it's normalized pixel values. We could see the different correlation values and it's respective coefficients against each of the variables.

The predicted values of the testing and training data set was individually observed and noted and rounded off to the nearest factor. Like example, if it were higher than 2.5, then it'd be 3 otherwise 2.

The mean square errors were computed considering the predicted values and actual values of training and test data - thus computing the misclassification rate.

4.3 k-nearest neighbor classification

For values 1,3,5,7,9,11, 13,15 of k, the k-nearest neighbor classification is built to compute the training and test errors across various k values - thus computing the misclassification rate.

4.4 Plotting the errors and the comparison

The Test and train data set errors versus the different KNN values were plotted along with the training data set error. (See Fig. 18)

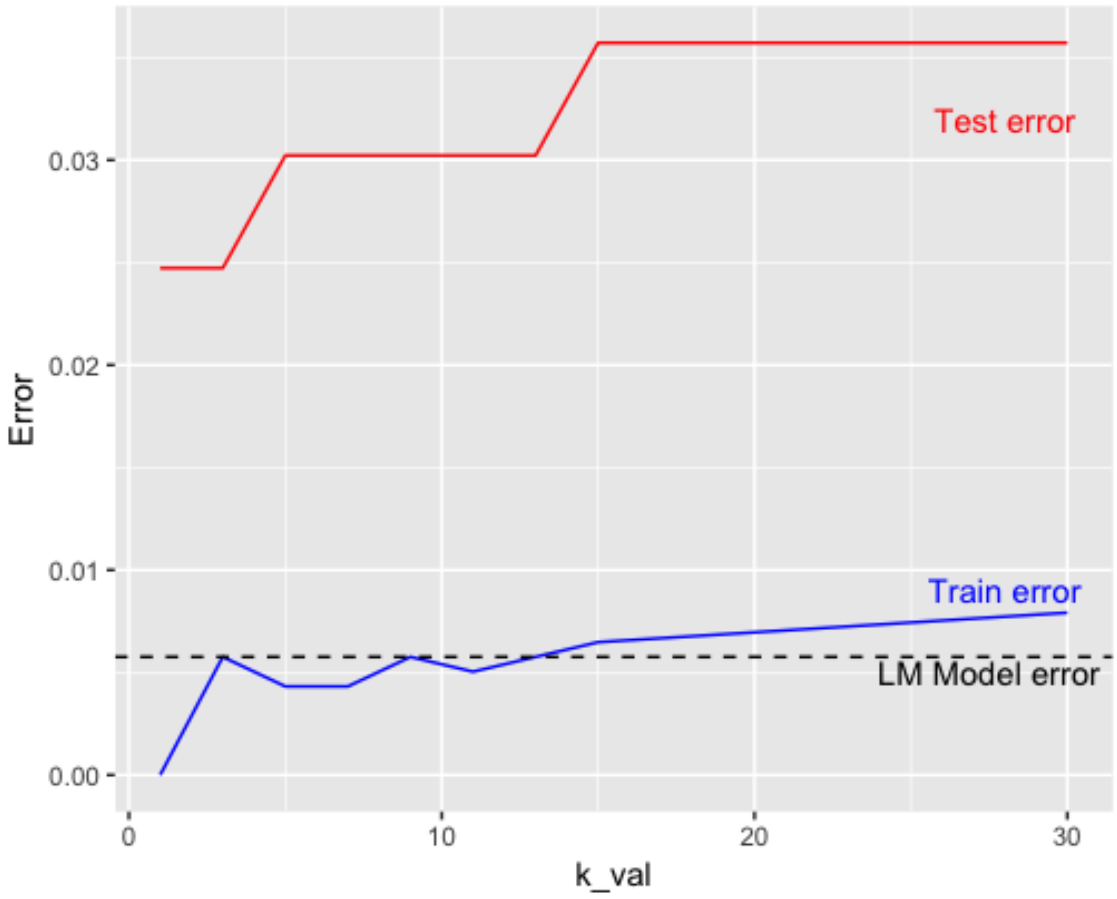

Figure 18: Test and train data set errors vs KNN values