Statistical Data Mining I
Homework1

Abhishek Kumar

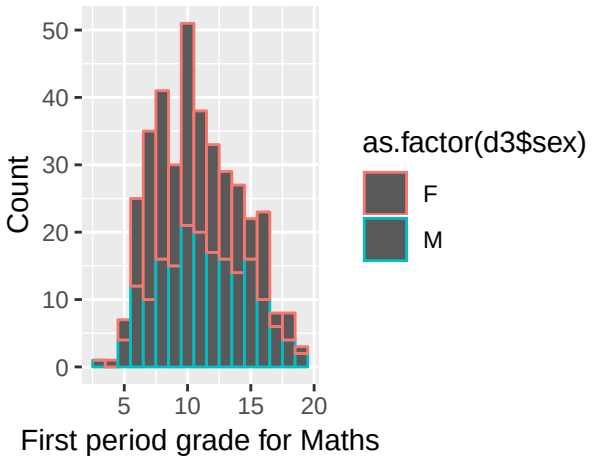
September 14, 2018

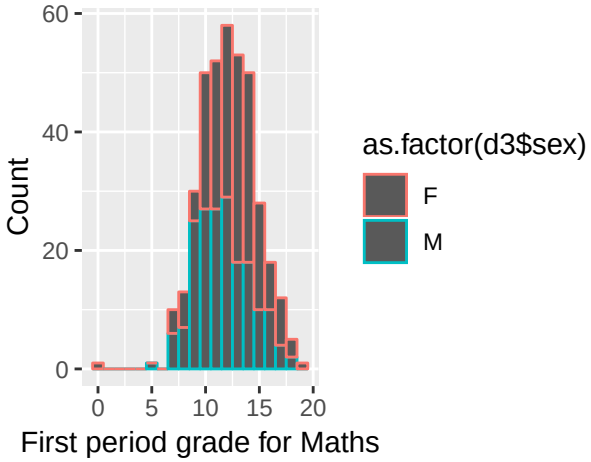
Solution 1

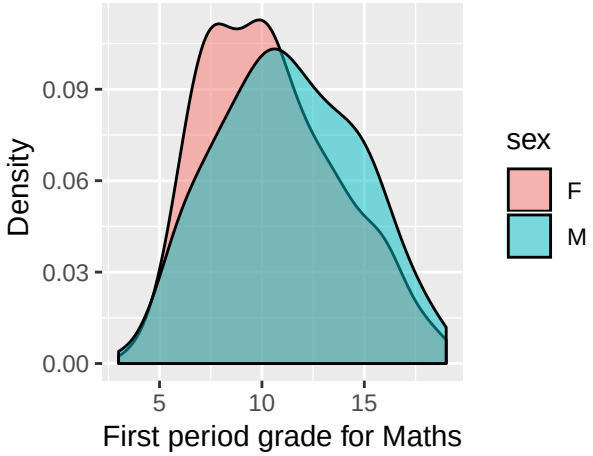
The Student Performance Data Set is a highly researched open-source dataset in the education sector. This data set was put together by collecting data about grades from two portuguese schools and complementing it with survey from students. The motivation is use the available data to make better decisions about changes to improve student performance. Here we will explore the data and try to find parameters/predictors that may effect student grades.

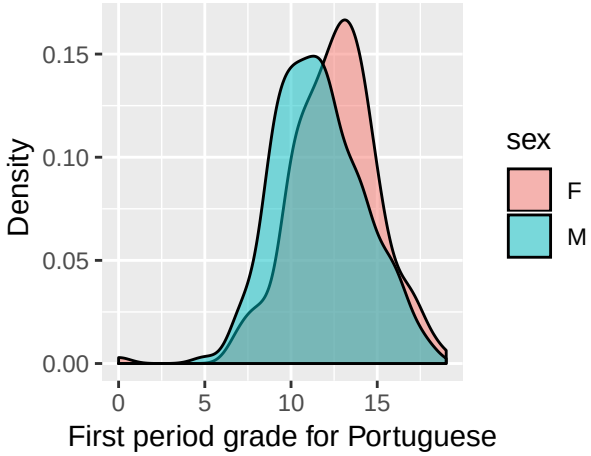
We have two datasets 'd1' and 'd2' for Mathematics and Portuguese respectively. To start, first we merge the two datasets into one to get a dataframe with 382x53. All of the variables we have are factor variables except absences which range from 0-93. Our target variable is G1, the first period grade.

Let's look at some distribution of the first period grade.

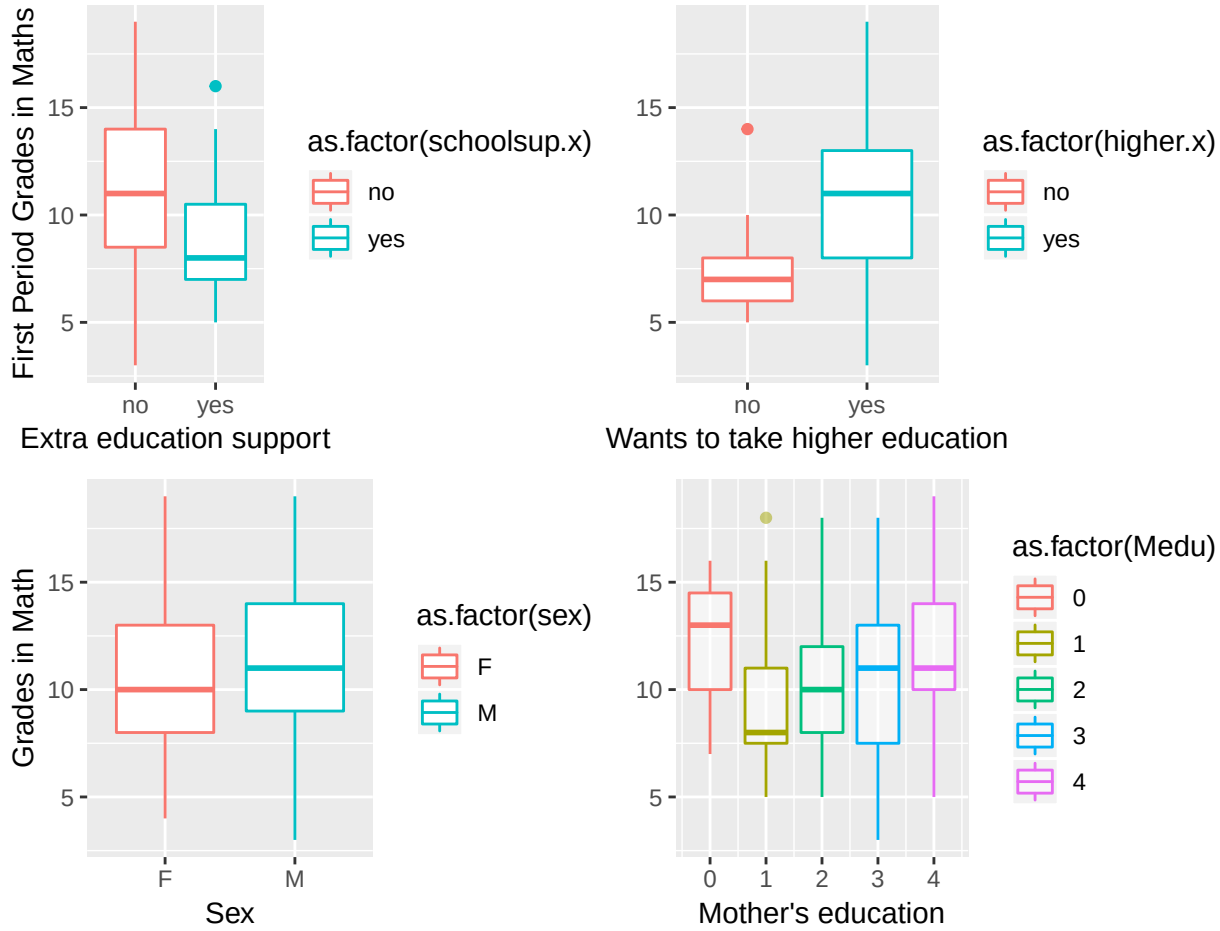


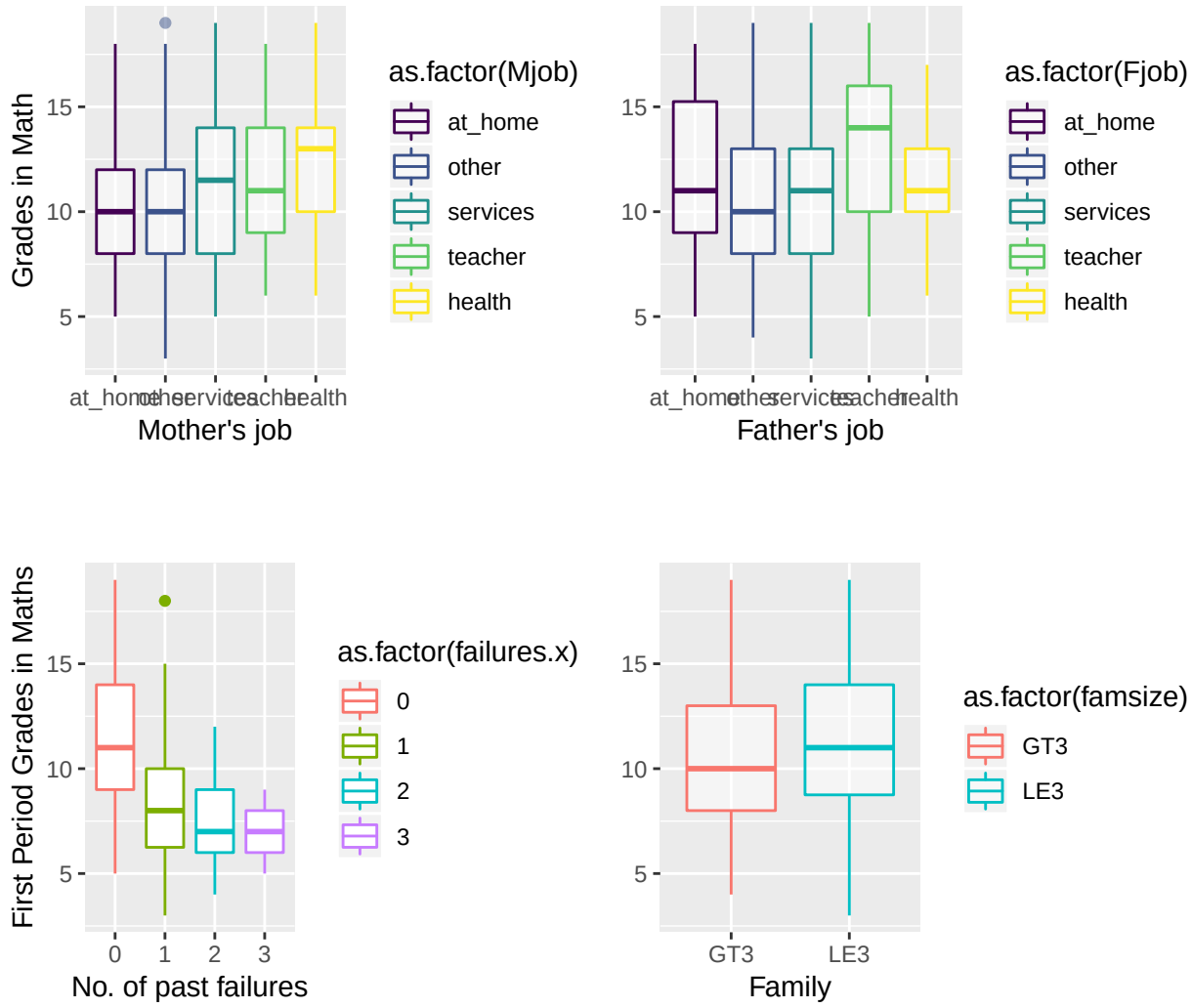




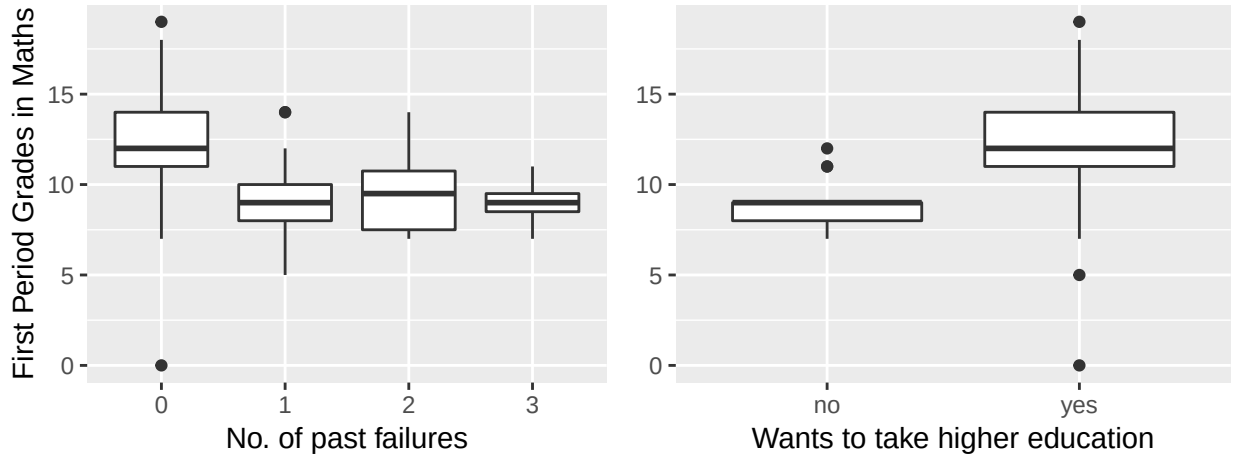


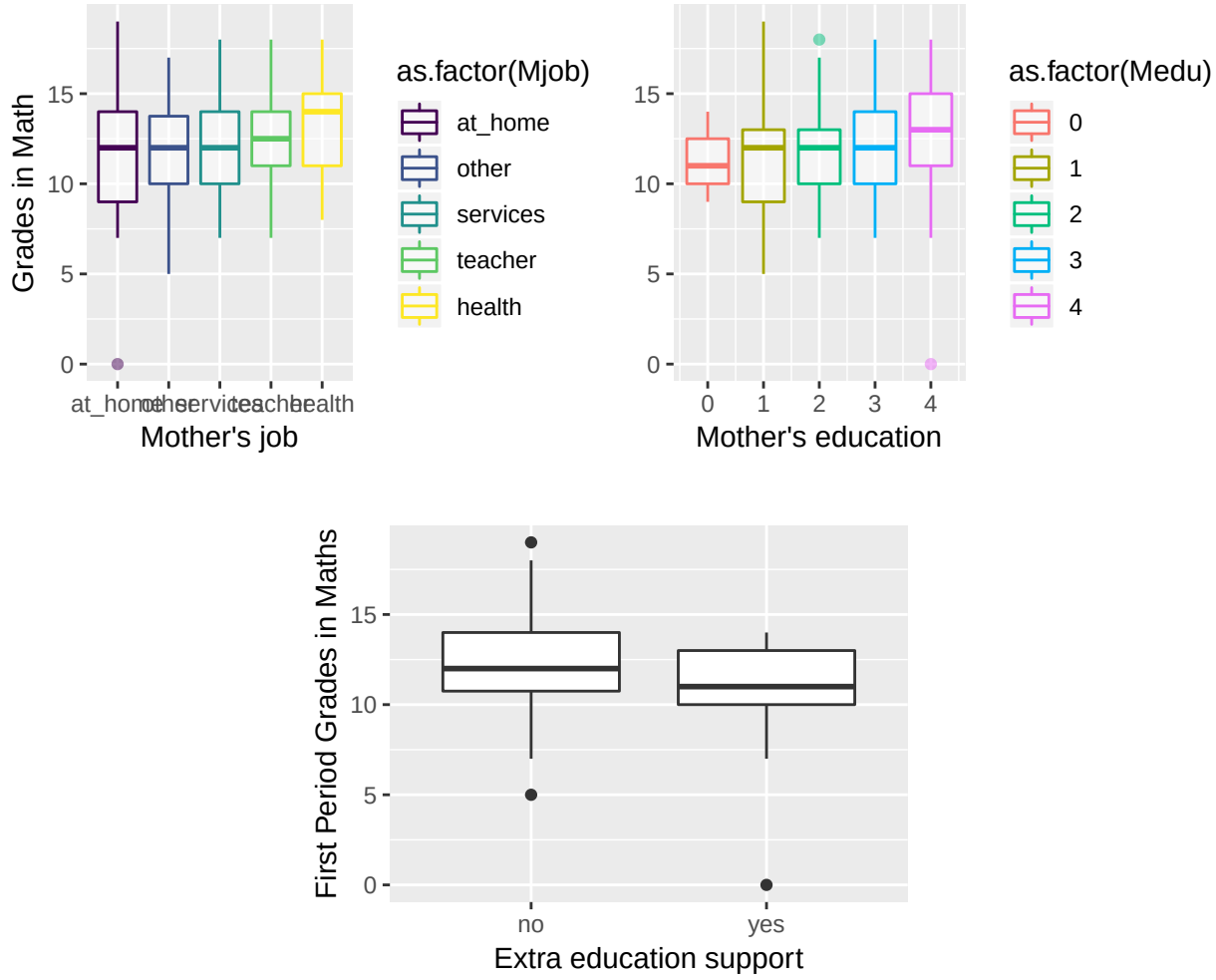
After analysis and exploration, I found some important inferences from the plots. From the histogram and density plot above it is evident that the male population performed better in maths while females did perform better in Portuguese. Other predictors such as schoolsup, sex(male), Medu(higher education), Mjob(health) and Fjob(teacher) positively affected the first period grades for maths while failures and familysize(>3) negatively affected it. As these are factor variables, it will be easy to visualize using boxplot. Let's look at some of those.





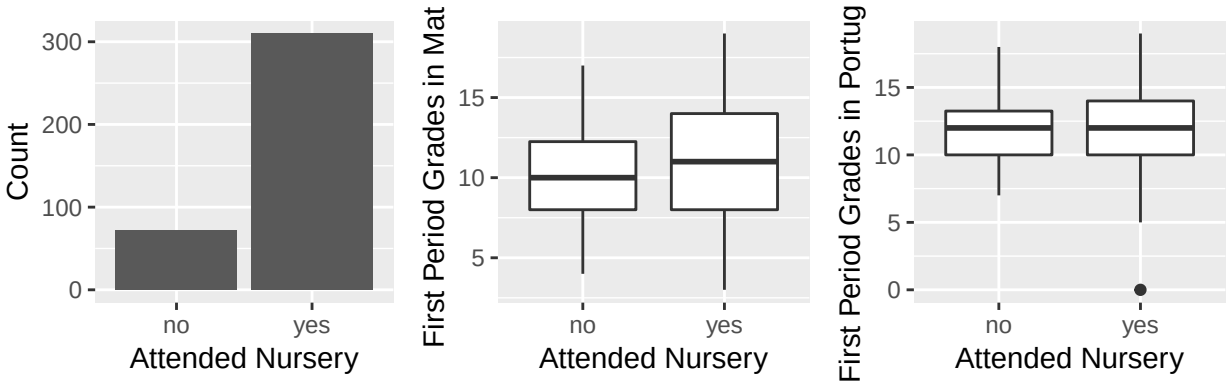
While for Portuguese, parameters like failures, higher, Mjob, Medu, schoolsup were major factors that affected students grades.

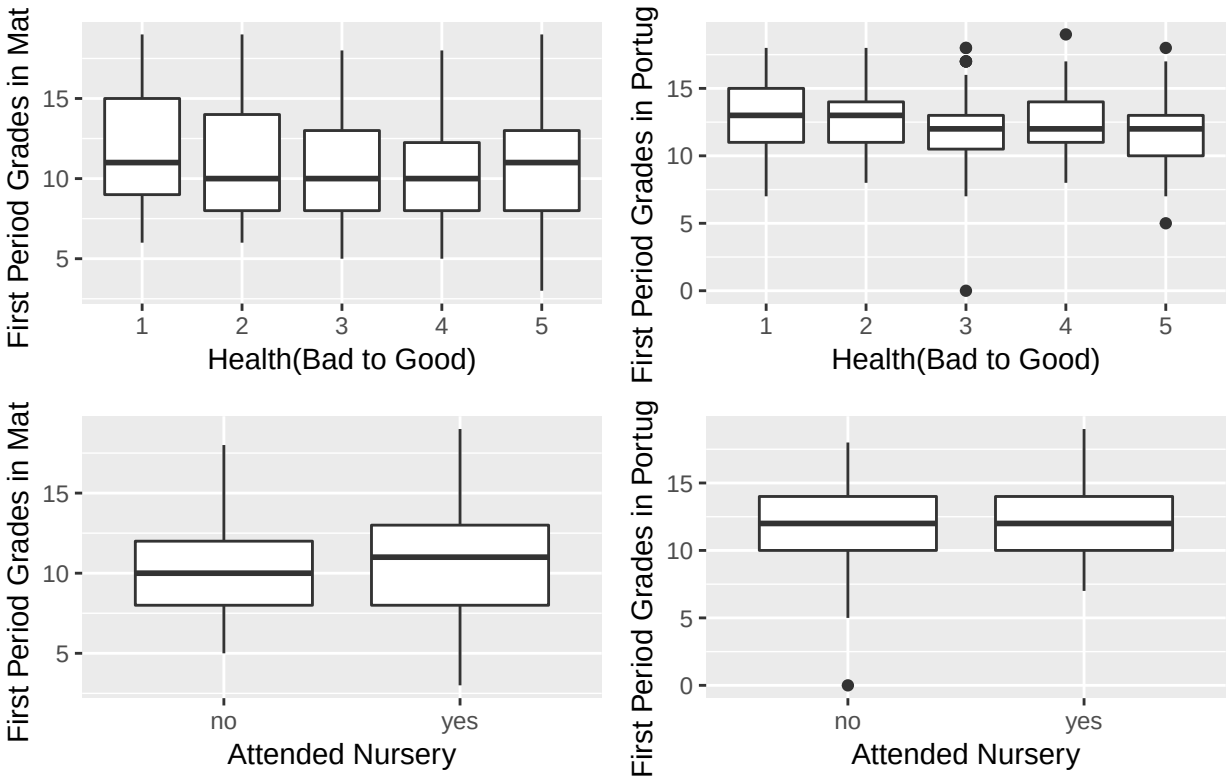




I also stumbled upon situations where the target variable, the first period grade seemed completely independent of some variables. Like internet, health and nursery that does not seem to affect students performance in either Maths or Portuguese. Even if they vary slightly, there doesn't seem to be a pattern. Its likely some noise. So, to may want to delete these variables from our training data. Let's look at these variables.

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



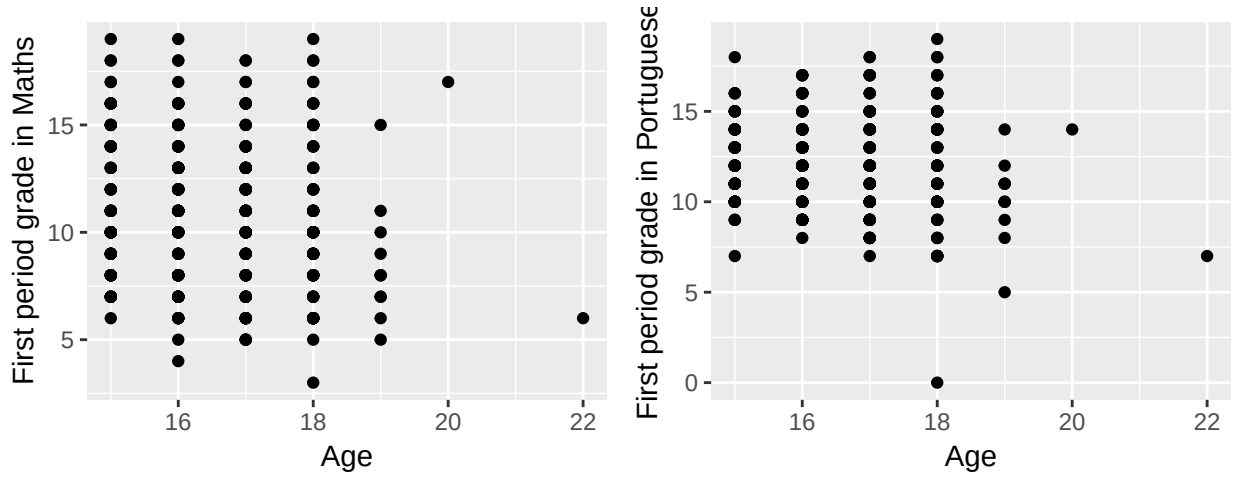


Outliers

If we look at the stem-leaf plot for age, we find that there is just one student with ages 20 and 22 and their grades in Maths and Portuguese are also in the oulines. Due to lack of more data in the same age group, I feel it may influence my model but grades for these age group conforms to my imagination and I would like to keep these outliers in my training data. I feel that as students with age 20 and 22 are quite older with respect to the average population of the class they may find it difficult to interact, discuss and collaborate with their fellow mates which would influence their grades. Also when I individually see these two students, I find the the one with age 20 had no absences, no failures in Maths, wanted to pursue higher education, was non-alcoholic, having good qualities of a student. While the other with age 22 was alcoholic with poor health, frequently outgoing, had previous failures in both the subjects and had more than average number of absences. Hence, I'll be keeping them in my training dataset.

```
##
##   The decimal point is at the |
##
##    15 | 000000000000000000000000000000000000000000000000000000000000000000000+1
##    15 |
##    16 | 000000000000000000000000000000000000000000000000000000000000000000000+27
##    16 |
##    17 | 000000000000000000000000000000000000000000000000000000000000000000000+20
##    17 |
##    18 | 000000000000000000000000000000000000000000000000000000000000000000000+1
##    18 |
##    19 | 00000000000
##    19 |
##    20 | 0
##    20 |
##    21 |
```

```
##    21 |
##    22 | 0
```



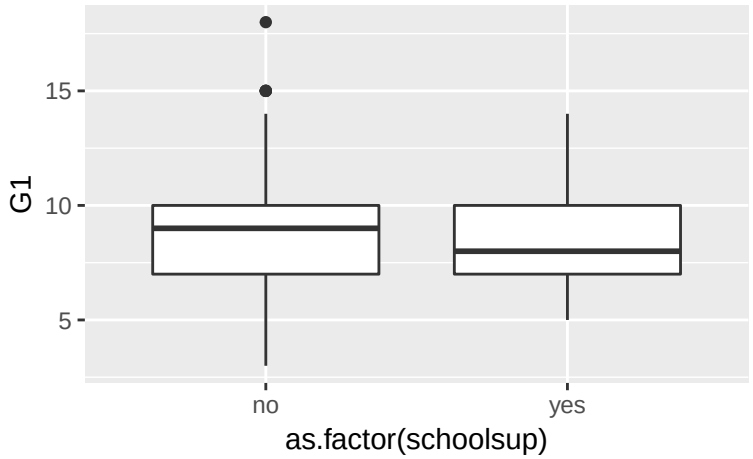
Finally,I would go ahead with deleting the non-influential factors(internet, nursery, health) and delete G2 and G3 as well to create a controlled dataset and use it to train my model and verify the performance by comparing its Mean-Squared Error with the original dataset which contains every variables except G2 and G3.

Additional Inferences :

1)

We made a new subset of students who have failed earlier and made a box plot to see the performance of the students who took schoolsup Vs those who didn't. We found that among most of the students who have failed earlier and didn't take school support did better than those who did.

```
failed_stud.x <- subset(d3_controlled, failures>0)
ggplot(failed_stud.x, aes(x = as.factor(schoolsup), y = G1))+ geom_boxplot()
```




```
by(failed_stud.x$G1, failed_stud.x$schoolsup, summary) #check median
```

```
## failed_stud.x$schoolsup: no
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   3.000   7.000   9.000   8.759  10.000  18.000
## ------------------------------------------------------------
## failed_stud.x$schoolsup: yes
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.000   7.000   8.000   8.714  10.000  14.000
```

2)Finding einsteins !! Most of their parents are either both teachers or one of them is.

```
##         Mjob    Fjob schoolsup G1
## 287 services teacher        no 18
## 481 services teacher        no 19
## 491  teacher teacher        no 18
## 501  teacher   other        no 18
## 502  teacher   other        no 18
## 522  teacher teacher        no 18
## 533    other   other        no 18
## 753  teacher teacher        no 18
## 922  teacher teacher        no 18
```

Solution 2

a)

First let's check the Multiple R-squared of the model with the original data but excluding G2 and G3. The Multiple R-squared error is 0.2973 and accuracy is 0.5452. The major predictors in increasing order is failures(-), schoolsupyes(-), higheryes(+), studytime(+) and paidyes(-).

For the controlled model we get a Multiple R-squared of 0.2945, almost what we got with the original data and the accuracy is almost equal, 0.5427. This also confirms that the variables removed were not important predictors. And the influencial factors we get with the controlled data are failures(-), schoolsupyes(-), higheryes(+), studytime(+), paidyes(-), SchoolMS(-), Fjobteacher(+), famsupyes(-) and famsizeLE3(+).

```
##                       actual_grade_orig predicted_grade_orig
## actual_grade_orig             1.0000000            0.5452187
## predicted_grade_orig          0.5452187            1.0000000
```

```
##                       actual_grade_cont predicted_grade_cont
## actual_grade_cont             1.0000000            0.5426873
## predicted_grade_cont          0.5426873            1.0000000
```

b)

From the summary of the model we can see the direction of coefficients and decide if the influencial predictors are positive or negative. If I were to recommend a first year student to achieve a good grades would be: i) Work hard and try to get into a better school. Always set your ambitions high. Even if you don't have money to take up extra educational support from school or family, don't worry! My data suggests its not always good for your good grades. Sometimes you have a learn to self-learn. It'll make you independent in the long run.

Good Luck!!

c)

After trying most of the possible interaction, the best model was with "Mjob", where the Multiple R-squared is 0.4293 and the accuracy improved from 0.5427 to 0.5919.

Solution 3

Let's start with the first step to analyze the dataset, the famous Boston Housing Dataset with just checking its head and summary. It has 14 variables and our target variable is "medv", which is the median value of owner-occupied homes in $1000s. The variables are numerical and not factor variables and their range also varies so we need to normalize them before feeding into our model.

Exploring data

Let's check the correlation of our target variable with other variables. I'll be plotting a correlation intensity plot. Here we can see that for our target variable median prices of the houses are positively correleted with number of rooms per dwelling(rm, cor = 0.69) and negatively correlated with the lower status of the population (lstat, cor = -0.74). There are other variables like proportion of non-retail business acres per town (indus, cor = -0.48), pupil-teacher ratio (ptratio, cor = -0.51), full-value property-tax rate per $10,000 (tax, cor = -0.47) and nitrogen oxides concentration (nox, cor = -0.42) that are correlated with the mean prices of house.

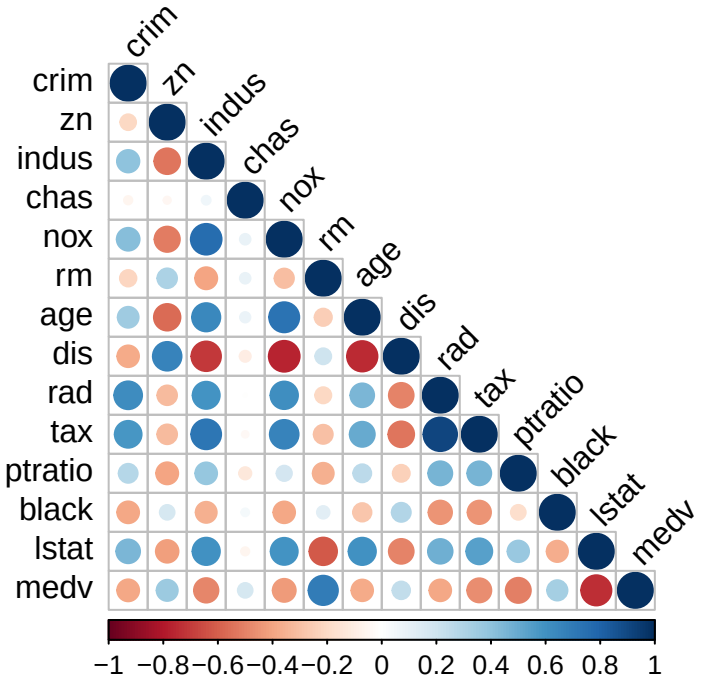


Figure 1: Correlation Matrix Intensity

We can also draw some clear outliers for the plot. The charles river dummy variable (chas) has almost no correlation with house prices or any other variables, so we may want to eliminate it from our data. Next, we can observe that 'nox' and 'indus' one of the possible major factors are highly correleted (cor = 0.76), so we may want to use just one of them in our training data.

a)

Now, lets visualize how each variables are associated with other variables using a pairwise scatterplot. We have removed 'nox' and 'chas' from this plot to focus on other important variables. With this scatterplot we can infer some interesting relationships.
i) medv VS lstat: We can see the inverse relation between the two
ii) medv VS rm: Positive linear relation between these; as expected

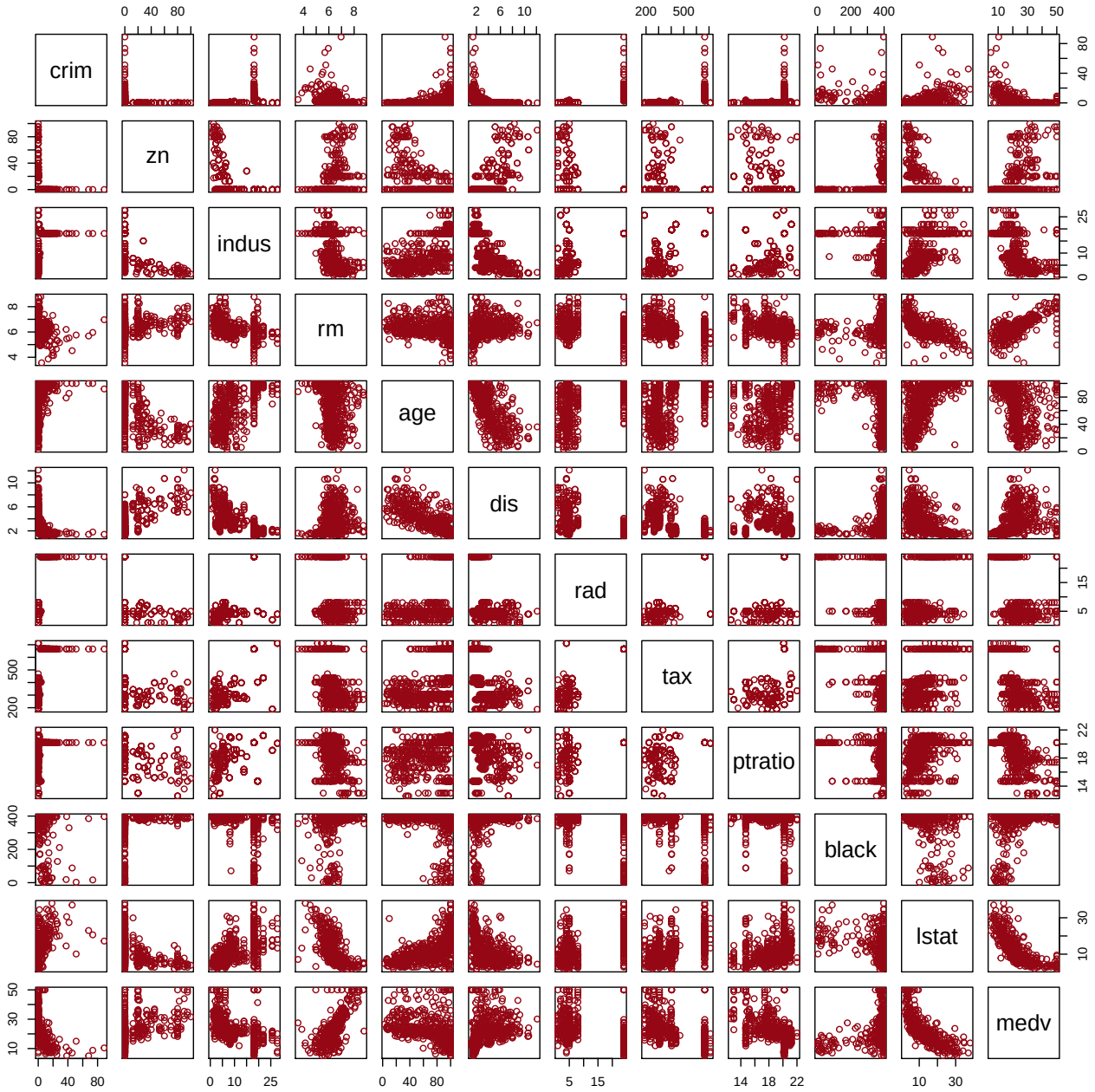


Figure 2: Pairwise Scatterplot

b)

Yes, if we look at the correlation-matrix intensity plot we find that the index of accessibility to radial highways (rad, cor = 0.63) and full-value property-tax rate (tax, cor = 0.58) are highly related to the per-capita crime rate. Also “rm” and “lstat” are inversely related.

c) To get an idea of these distribution, lets look at some plots.

There are few suburbs with high crime rates ranging upto 89%. But most suburbs have less than 1% per capita crime rates. When we look at the histogram for tax rates we see that 132 suburbs have tax of ~$670 per $10,000 and in some suburbs tax rate is as high as $710. We can also notice that there is no suburb whose property-tax is between $470-$670. In general the tax-rates among suburbs range between $190-$710. And the pupil-teacher ratio varies between 12.6-22.0 while most suburbs have pupil-teacher ratio of 20.2

Summary of crim, tax, ptratio.

```
##       crim              tax           ptratio
##  Min.   : 0.00632   Min.   :187.0   Min.   :12.60
##  1st Qu.: 0.08204   1st Qu.:279.0   1st Qu.:17.40
##  Median : 0.25651   Median :330.0   Median :19.05
##  Mean   : 3.61352   Mean   :408.2   Mean   :18.46
##  3rd Qu.: 3.67708   3rd Qu.:666.0   3rd Qu.:20.20
##  Max.   :88.97620   Max.   :711.0   Max.   :22.00
```

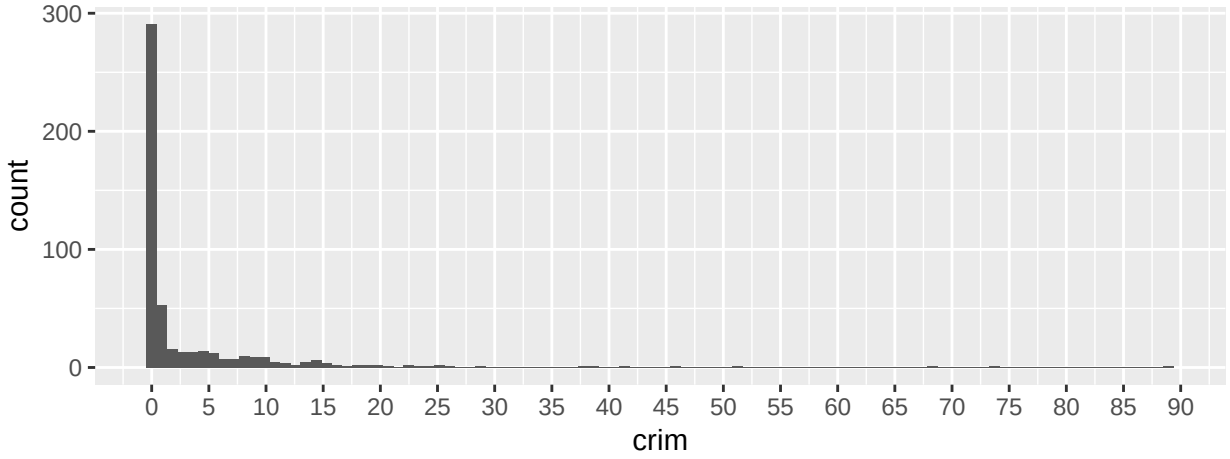


Figure 3: per capita crime-rate

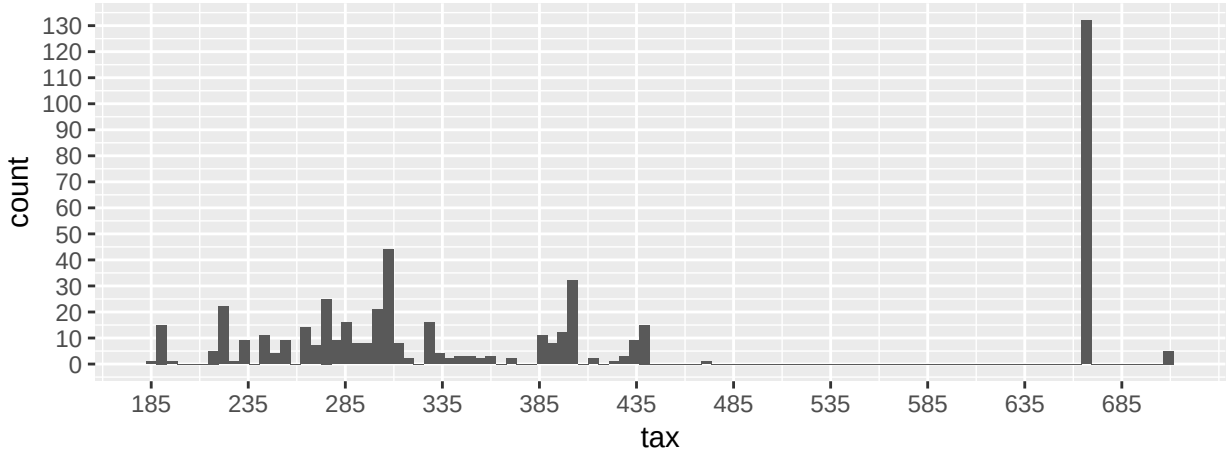


Figure 4: tax rates

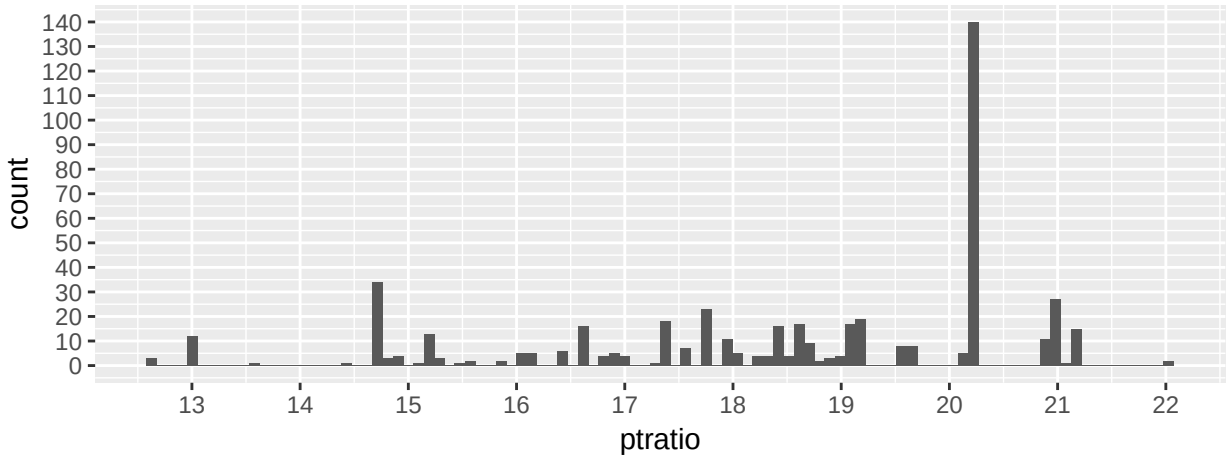


Figure 5: pupil-teacher ratio

d)

There are 333 suburbs with average more than 6 rooms per dwelling, 64 with more than 7 rooms per dwelling and 13 suburbs with more than 8 rooms per dwellings. For suburbs with more than 8 rooms per dwelling we can infer from the graphs below that they mostly costly and age over 60 years. Also the tax rates and crime rates are below the average.

```
nrow(subset(Boston, Boston$rm>6))
```

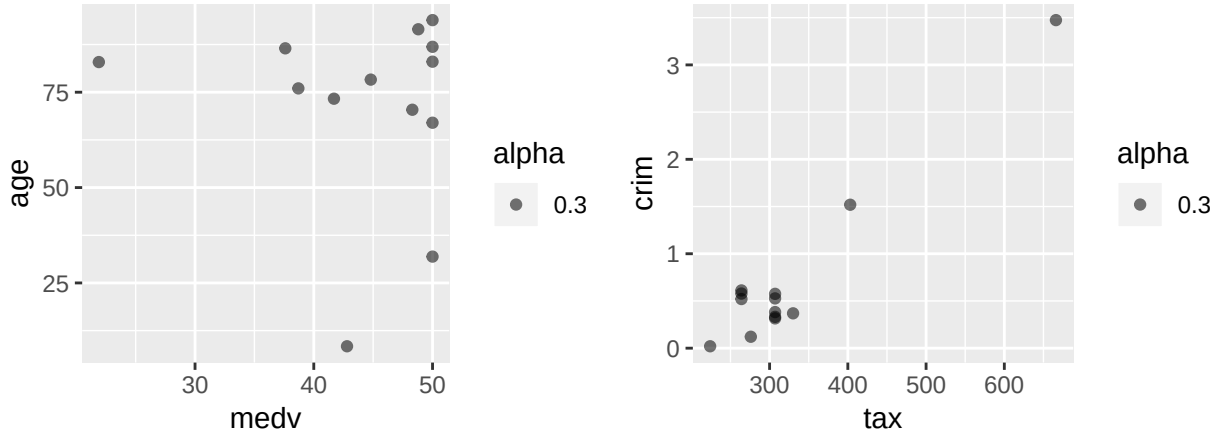
```
## [1] 333
```

```
nrow(subset(Boston, Boston$rm>7))
```

```
## [1] 64
```

```
nrow(subset(Boston, Boston$rm>8))
```

```
## [1] 13
```



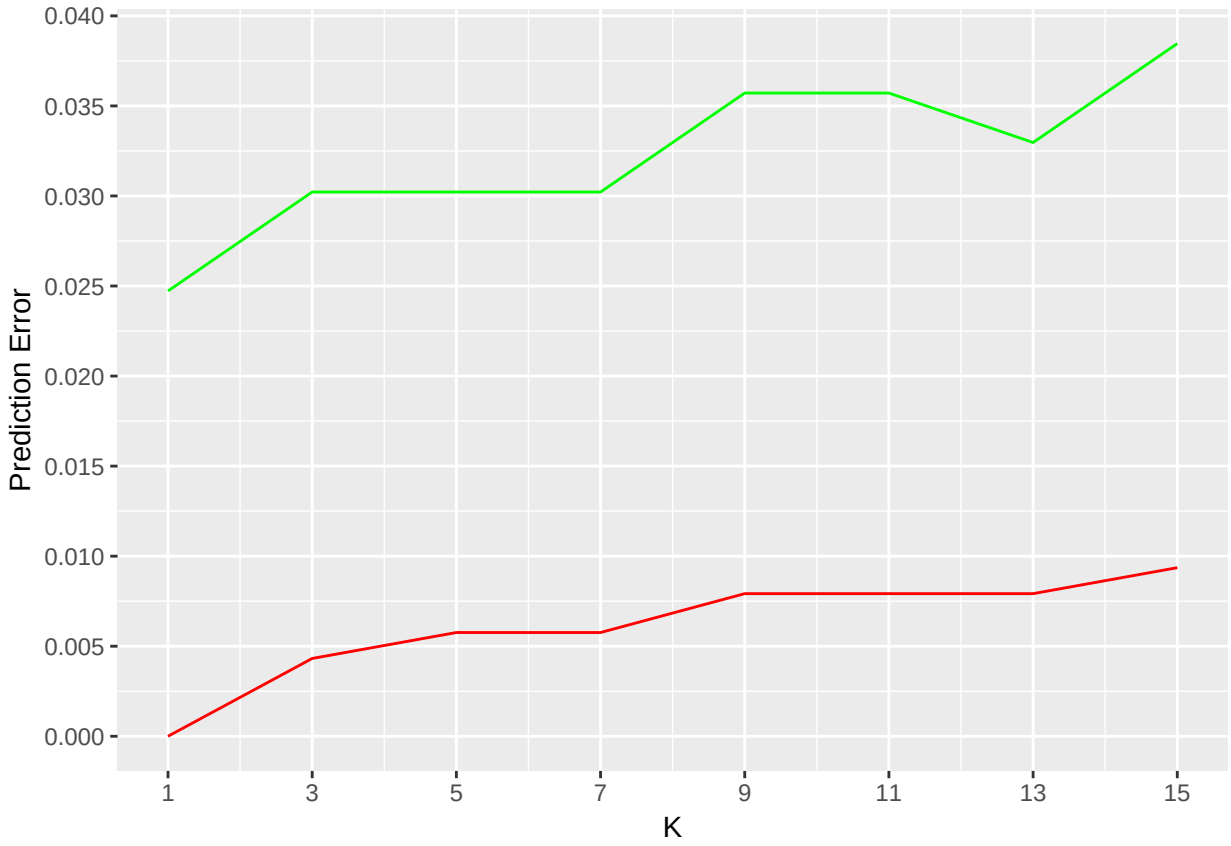
Solution 4.

The zipcode data is the image of digits from 0-9, each image of size 16x16 pixels and our dataset contains has one row with 256 pixels of an image and one column, the first one the classification of the image from 0-9. Here we have been asked to train our model to classify just two digits 2 and 3. So, the first step would be select the rows whose 1st rows has either 2's or 3's. Then change 2's to 0 and 3's to 1 for better classification when using linear model.

Let's look at some of the data to get a feel.

KNN Implementation

```
## Loading required package: class
```



Linear Model Implementation

The training error is 0.005760 and the testing error is 0.0412. Its highly accurate as it is a binary classification with lots of training data but I am very sure that if we use the same model to classify all the 10 digits the performance will degrade.

In case of KNN the minimum error is at k=3,5 and 7 almost ~3.02% while in the linear model our error is ~4.12%. KNN is working better here!!