Statistical Data Mining I - Homework 1

Friday, September 13, 2019 6:10 PM

Vimal Kumarasamy
UBIT name: vimalkum
E-mail: Vimalkum@buffalo.edu

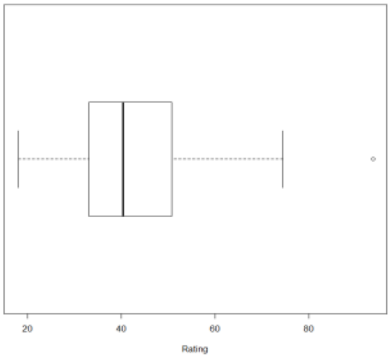
<u>Q 1 - Cereal Dataset - EDA, Pre-processing - Transformation, Elimination</u>

<u>Structure, Dimensions and Features</u>

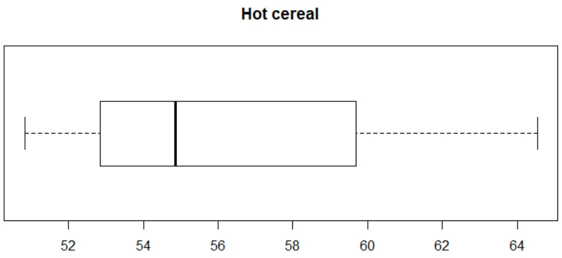
- There are 77 records in the dataset with "mfr" being the unique primary key
- Summary of the dataset shows that "name", "mfr" and "type" are categorical variables and other are numerical. "carbo", "sugar s" and "potass" variables have -1 as an observation, which seems incorrect as these values can't be negative
- Referring to the official reference for additional context about the dataset shows that -1 signifies missing data
- There are just 77 records so removing the entire datapoint is not a good idea, ignoring the feature completely is also not a good idea as we are not sure what is the value add from the other observations in the feature
- Let's convert these values to NA so that plots and models are not disturbed, however only the complete observations will be c onsidered for the analysis
- The nutritional metrics are all observed per serving, so there is no need to perform any normalization to compare across cere als

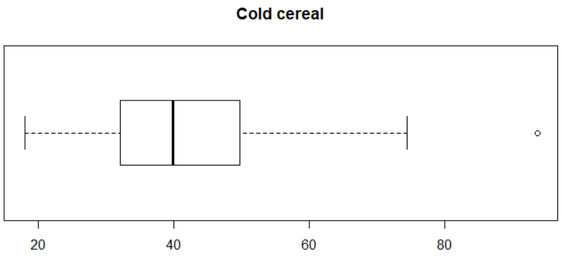
<u>Feature distribution study</u>

- Let's look at the distribution of the response variable - "rating" - boxplot



- There is one datapoint with a comparatively higher "rating", a closer look at it shows that it's a cold cereal type called 'A ll-Bran with Extra Fiber' manufactured by Kelloggs

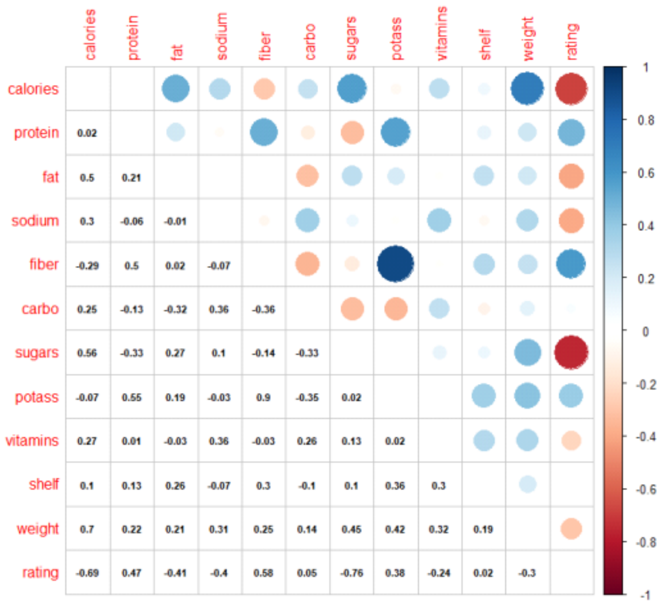




- Hot cereals seem to have higher median "ratings", it's evident that there is good distribution of the feature over a significantly wider range

Correlation Matrix Study

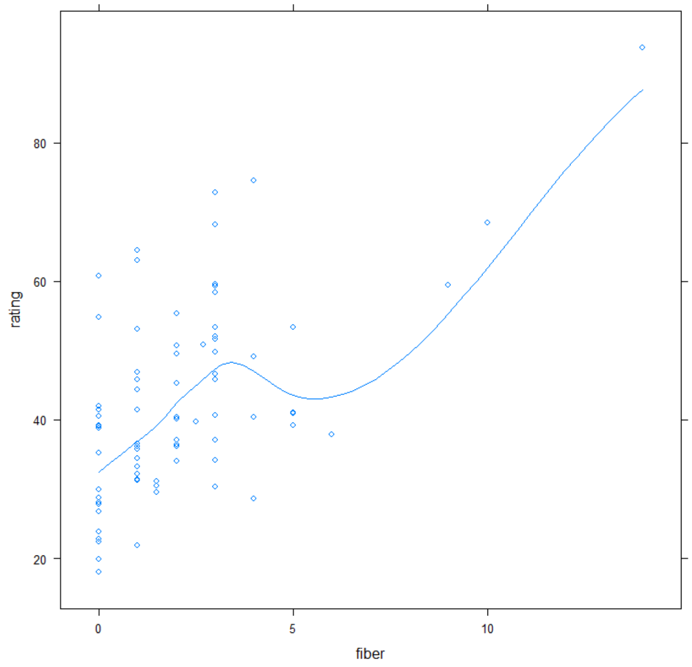
- Let's study the correlation matrix to understand any relationship between features and response variables and amongst features as well



- "rating" seems to have high positive correlation with "fiber", "protein" and "potass", and high negative correlation with "calories" and "sugars"
- Another point to be noted is that "potass" and "fiber" also seem to have high positive correlation. "sugar" and "calories" are having positive correlation, which is not very surprising

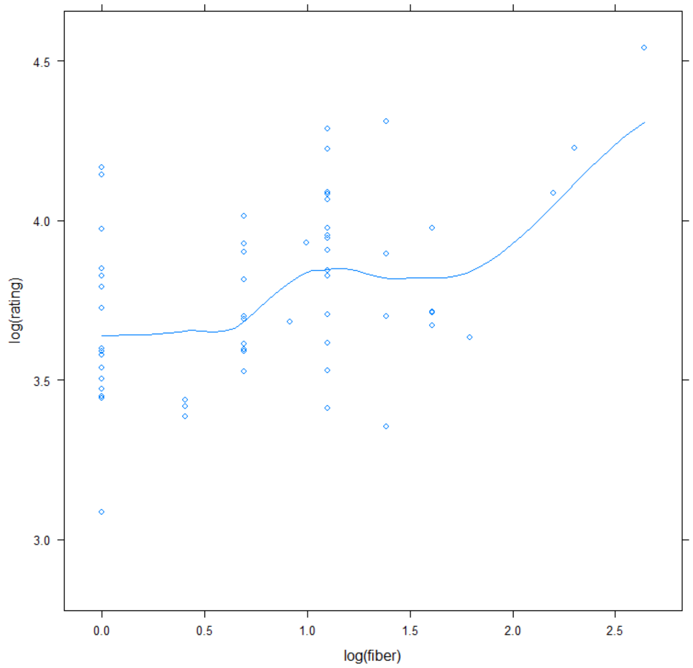
Predictor and Response variable relationship study

- Now that we have good understanding of what are the features and their relationship with "rating" let's dive in and study their relationship closer
- Plotting "fiber" against "rating" shows that there are few high values of "fiber" which are skewing the plot
- These high values are not outliers and certainly not to be removed, we can try different transformations to find a linear relationship between transformed feature and response variable
- Since Linear Regression provides helps us arrive at a single Beta Estimate for a predictor, its necessary for the predictor to have a linear relationship with the response variable

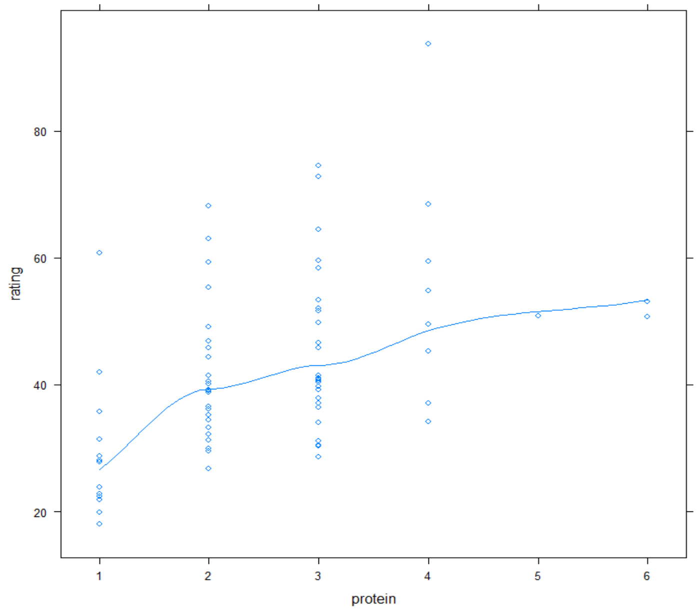


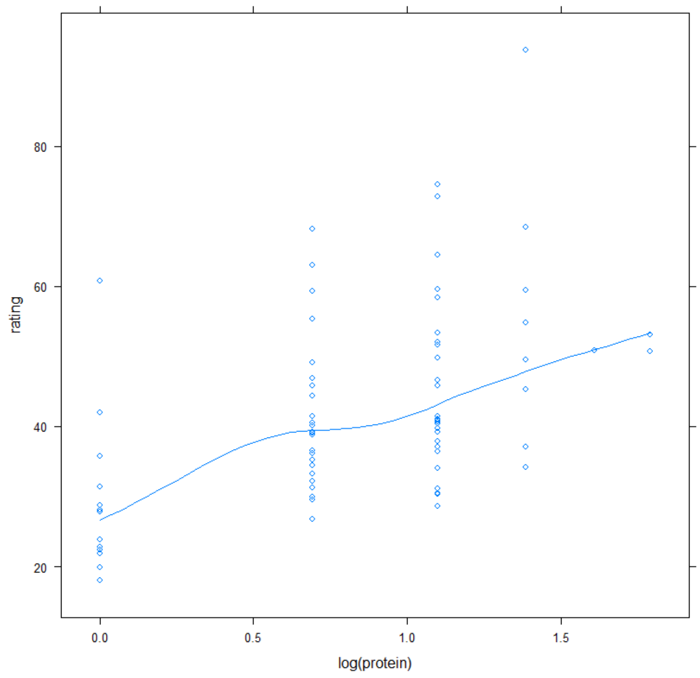
- Log transformation of "fiber" has helped in reducing the range of the variable and made the plot more linear. However let's look at percentile points and make the variable into 4 different buckets and introduce it as a categorical variable

- One-hot encoding is done, to ensure that the buckets are represented as 4 different 1 or 0 variables. This also helps in removing one of the variables (among 4 buckets) if its insignificant
- Making the variable as categorical will enable Linear Regression to give different Beta Estimates to the same variable

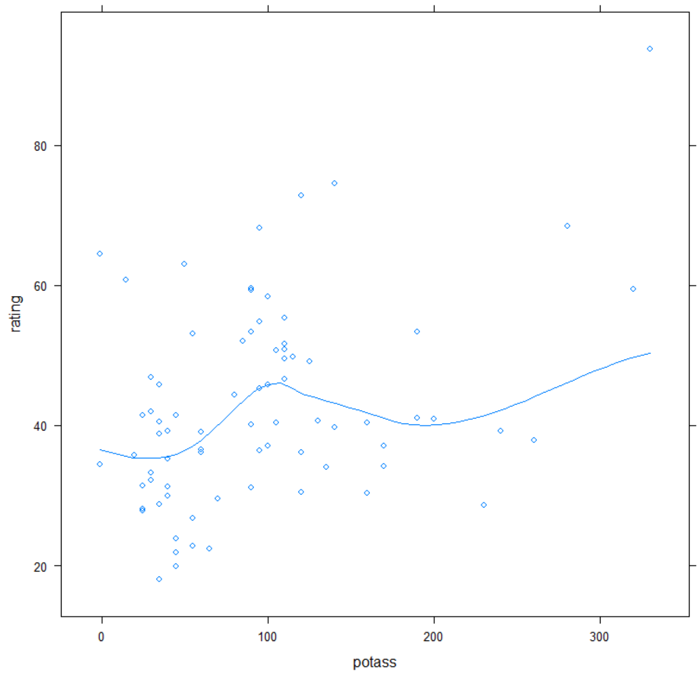


- Protein has a good linear relationship with rating, the increase in rating after a certain point in protein saturates and further increase in protein doesn't result in the increase of rating. Just like fiber, we can perform log transformation here

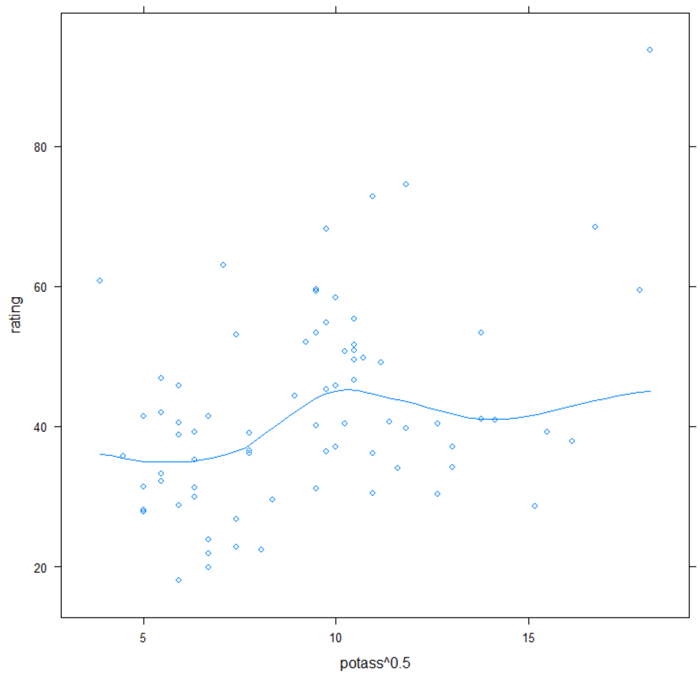




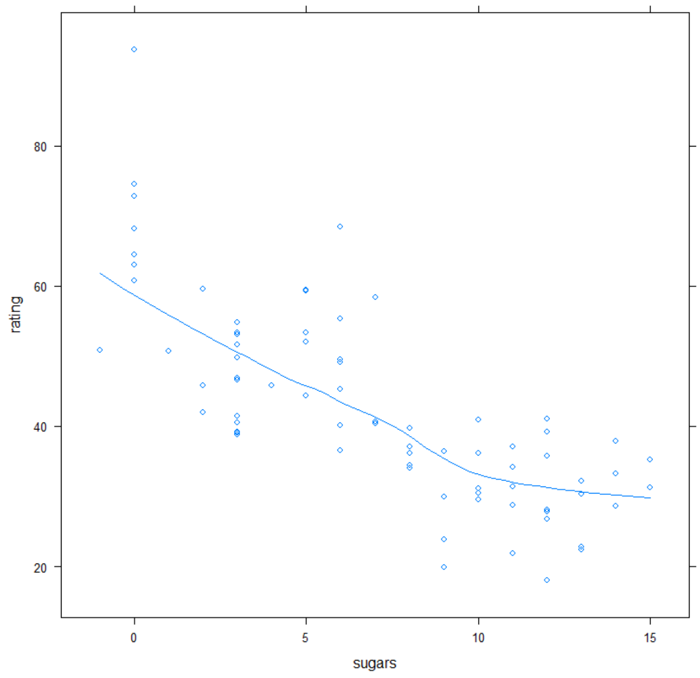
- Log transformation of "protein" has helped in making the plot linear and lets create a variable "protein_log" in the dataset
- Plotting "potass" against "rating" shows that there is a weak positive relationship, however there is some disturbance due to extreme values



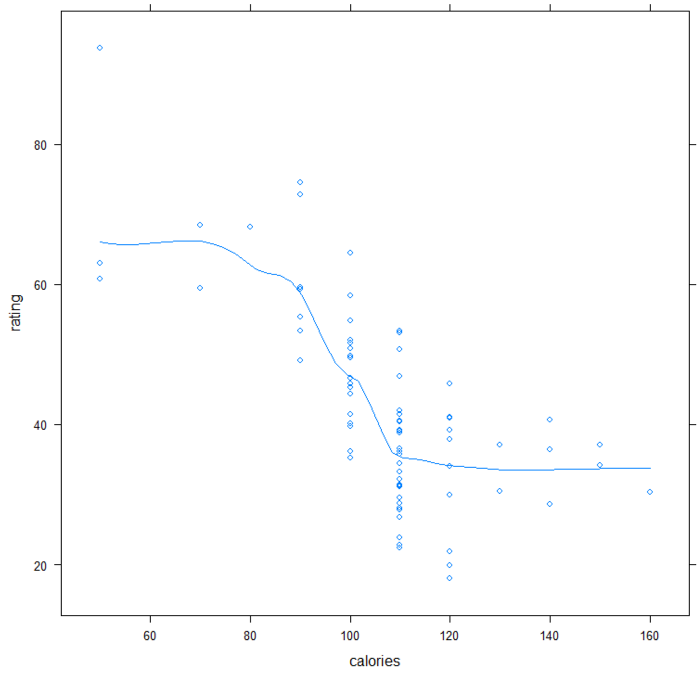
- There are few high values but still within the range, lets reduce the fluctuations by reducing the exponent of "potass", - taking square root



- Square root of "potass" has helped in reducing the fluctuations to an extent, lets create this variable in the dataset
- "Sugars" plotted against "rating" results in a clear linear relationship and can be used for modelling directly



- However "Calories" have an inverse sigmoid relationship with rating, and none of the transformations (log, exponential, squa re etc.) resulted in a linear plot. Since "calories" and "sugars" have high correlation, they might signify a similar information in explaining the response vari able, so we can experiment if the variable turns out to be significant

- 

- Created a composite score as a metric - ("Protein * Vitamins / (Fat * Calories) and this variable had a nice linear relationship with rating. We can experiment with this variable in the model
- "vitamins" just has 3 unique values in the dataset - 0,25 and 100. And there seems to be no linear relationship directly. After a saturation point, increasing "vitamins" in the cereal may not help in increasing rating.
- So it's better to convert this variable into categorical and one-hot encode this into 3 different 0 or 1 variable. As mentioned earlier, this will help us remove one of the 3 variables in the model if it turns out to insignificant rather than removing the entire "vitamins" variable
- "sodium" seems to have a parabolic relationship with "rating", so "sodium" is split up into 3 1 or 0 variables based on percentile points
- "shelf" being a numerical variable, it signifies the position of the cereal, so it can’t be considered as continuous variable (cereal with 3 as shelf will get 3X the impact compare with another cereal with shelf 1, which is incorrect). Making a categorical variable and one-hot encoding it
- Quick check on other variables such as "fat", "carbo", "mfr" didn’t result in strong linear relationship with the rating, we can experiment with these variables during the modelling phase

Q 2 - Multiple Regression, predictor significance, interaction variables

Least Squared model

- Based on the EDA performed above, we can start with building a least squared model with all the meaningful variables as predictors to explain the variability of the response variable
- The adjusted R squared is at 0.9488 with few variables insignificant, if there are 4 one-hot encoded variables, upon introducing all 4 to the model, one of the variable will not get a beta estimate, as that would form the base class for prediction

Further Iterations

- Removing the "fiber" flags, and swapping insignificant "vitamins" with other "vitamins" flags to check for significance
- All "fiber" flag variable combinations are insignificant, as its correlated with "protein" which is represented as "protein_log" we can let go of "fiber"
- Dropping "fiber" flags has not caused any significant drop in the adj. R squared, so we are not losing any information due to this drop
- Ensured that the residual error is also not increasing while iterating further
- Further experimentation by introducing "carbo", "weight" has not turned out to be fruitful. Both are insignificant
- "Shelf" flags have also turned out to be insignificant in all combinations
- Comparing the current parsimonious model against the former least squared model shows that there is not much increase in residual error and no major drop in adj. R squared
- Also the beta estimates of the current variables are intuitive towards rating, and significant

2.a Significant Predictors

- vitamin_25, sugars, fat, calories, sodium_1, sodium_2, protein_log, potassium_sq_rt

2.b Coefficient of "Sugars"

- -1.88748 Signifies that "sugars" has a negative relationship with rating
- Inference: For every gram increase of sugar in the cereal, there is a -1.88748 units drop in the rating

2.c Interaction variables

- Experimenting with interaction between 2 variables that are significant already and few non-significant variables as well
- (*) considers the product of the 2 variables and also the individual variables, whereas (:) considers only the product
- Introducing "fiber" flags with "potass", and tried few other interactions which turned out to be insignificant
- "sugars" and "calories" when combined together, it turns out to be significant and the rest of the model remains significant as well

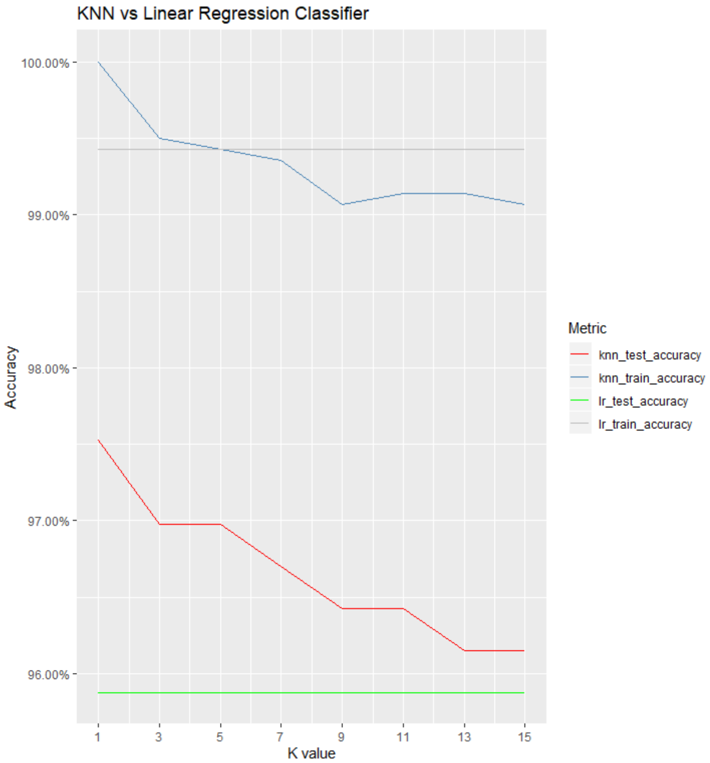
Q 3 - Zip code data - KNN vs Linear Regression

Understanding the Data

- Dimensions and column names of the dataset shows that there are 7291 datapoints for train, first column is the label and there are 256 predictors (16 X 16 pixels)
- Referring to the official reference for the additional context about the dataset
- The entire dataset is subset only for labels = 2 or 3 and the model is trained
- Looked at the range of every variable (other than the labels) and they are in the range of -1 to +1. So the datapoints don't require any scaling or normalization
- A function is defined, which is capable of fitting a KNN model for the K value which is passed in a loop and it saves the train and test accuracy in a matrix
- Instead of error rate, I am showing accuracy as it's the inverse and easy to comprehend and explain it to someone who is new to the project
- Linear Regression for classification is not generally preferred as the predictions are not bound, and it can mathematically be anywhere in -Infinity to + infinity
- The result obtained from LR is having a median around ~2.4, in order to convert the continuous value predicted through LR to either 2 or 3, I am setting a cutoff of 2.5 If the prediction is greater than 2.5 then its 3 else 2

Performance of KNN vs Linear Regression

- Train and Test accuracies are calculated for KNN across different values of K and plotted
- Linear Regression train and test accuracies won't change across K, as K is not a parameter for Linear Regression, however to compare against how it fares out against KNN, I have plotted Linear Regression Train and Test accuracies across K values



Inference

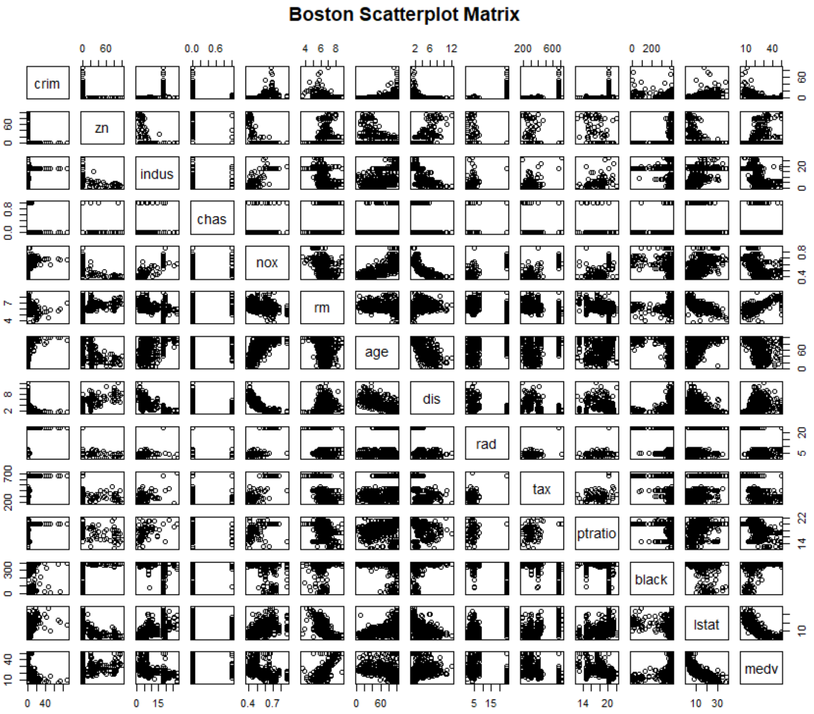
- Test accuracy would be the ideal metric based on which we can decide which algorithm performs better in classifying 2s and 3s
- KNN test accuracy for the given range (1 to 15) is higher than Linear Regression accuracy
- The accuracy is best when K=1, this shows that one closest neighbor decides on what could be the current datapoint (2 or 3) when its represented in 256 dimensions
- However the reason why there is a steady fall in the accuracy when k increases is not completely analyzed, but the total number of datapoints on the test dataset is 364 and the number of incorrect predictions when k=15 is just 14 whereas the number of incorrect predictions when k=1 is 12
- Due to insufficiency of test datapoints, it looks like a drastic drop in test accuracy, however the drop is exaggerated due to lack of large test dataset
- Looking at the performance, We should definitely consider a bigger dataset for test and stick with smaller values of K (say 1 or 3 for this classification use case)

Q 4 - Boston Housing Dataset Analysis

Understanding the dataset

- Boston dataset contains 506 datapoints and 14 variables
- Adding another variable (Serial number) to uniquely identify and label the suburbs - naming the variable as suburb_id
- Checked if there are any NA in the dataset and there are no NA

4 a Pairwise scatterplot and findings



- crim and age have positive relationship
- Inference: Higher the proportion of units built prior to 1940, higher the crime in the suburb

- crim and medv has negative relationship
- Inference: Higher the median value of owner-occupied homes, lower the crime rate

- zn vs indus and zn vs nox - Positive relationship
- Inference: Higher the nox - nitrogen oxide concentration, higher the proportion of land zoned for lots
- Similarly higher the nox, higher the proportion of non-retail business acres per town

- Chas vs other variables
- Inference: Chas river doesn't seem to have any strong realtion with any variable
- There suburbs with Charles river seem to be similar to those without

- rm vs lstat and medv - negative and positive relationship respectively
- Infernce: As expected, rm - number of rooms increases with decrease in lstat - percentage of lower status of population in the suburb
- higher the number of the rooms in dwelling - residential setup, higher then median value of homes in the suburb

- nox vs dis - negative relationship
- Inference: Higher the nitrous oxide concentration in the suburb, farther the boston eomployment centres

- lstat vs medv
- Inference: Stating the obvious, higher goes the median value of the homes, percent lower status of the population

4 b Predictors associated with per capita crime rate

- age - positive association
- Inference: Higher the proportion of the buildings built before 1940, higher the crime rate in the suburb

- dis - negative association
- Inference: Higher the crime rate per capita, lower the mean distance of the suburb from 5 boston employment centers

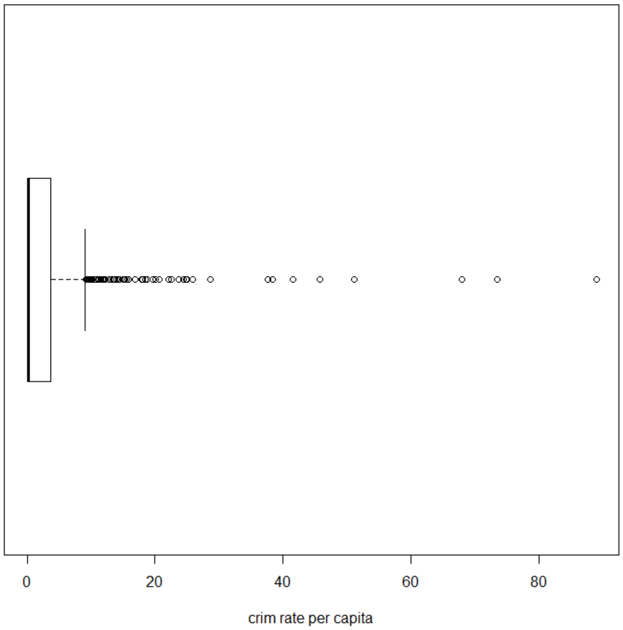
- lstat - positive association

- Inference: Higher the crime rates, higher the percentage of population with lower status

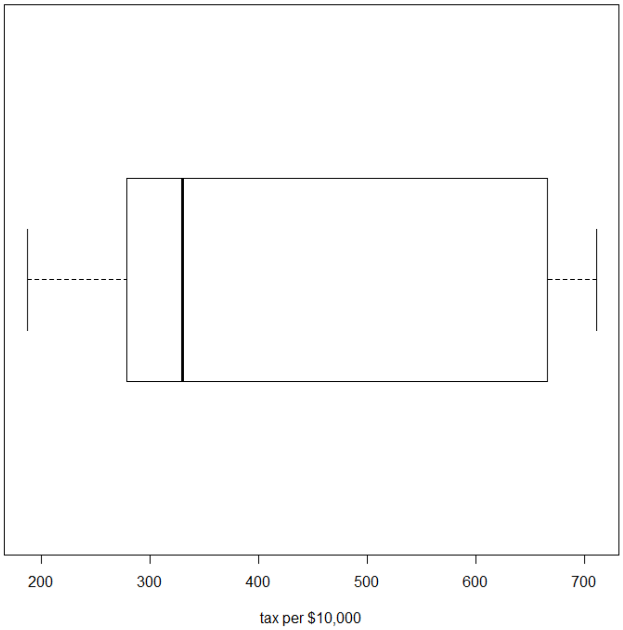
- medv - negative association
- Inference: Higher the crime rate in the suburb, lower the median price of the homes

4 C - high crime rates? Tax rates? Pupil-teacher ratios?

- Let's study the crime rates across suburbs through a boxplot

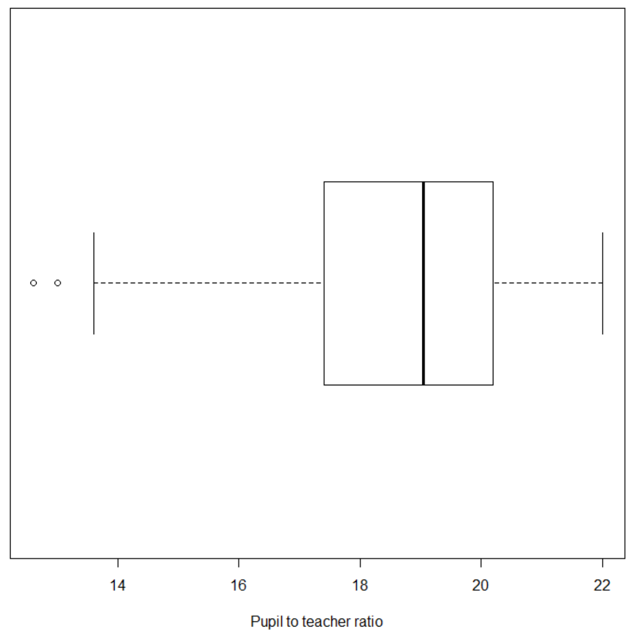


- Inference: The box plot signifies that majority of the suburbs have lower crim rate, and there are few suburbs that are infested with crime
- Looking at the percentile points, we decide that anything more than 99th percentile as high crime rate
  0% 25% 50% 90% 95% 99% 100%
  0.006320 0.082045 0.256510 10.753000 15.789150 41.370330 88.976200
- There are 4 suburbs with crime rate per capita higher than 41.37



- Looking at the distribution of tax per $10,000 we observe that median is in the left side of the range, signifying that the distribution is left skewed
- Percentile points
  0% 25% 50% 90% 95% 99% 99.9% 100%

  187   279   330   666   666   666   711   711
- There are 5 suburbs with tax of $711 per $10,000 or higher



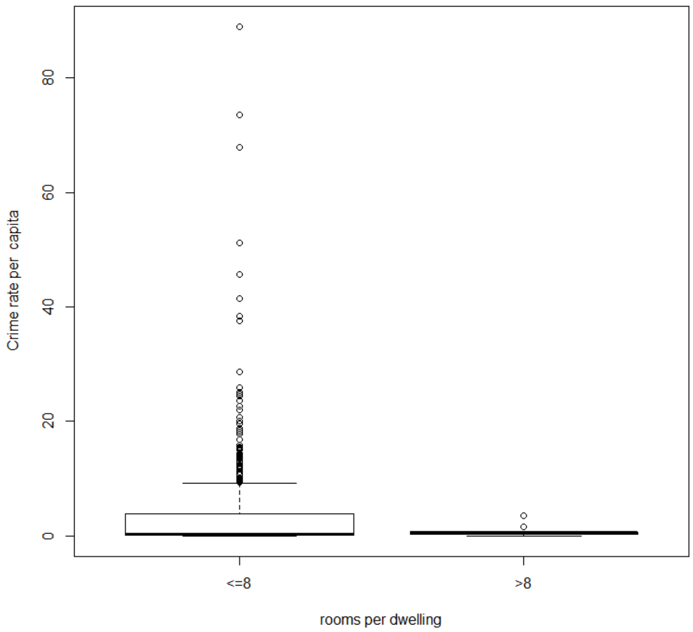
- There are few suburbs with pretty low ptraio
- Percentile points
  0%   25%   50%   90%   95%   99% 99.9% 100%
  12.60 17.40 19.05 20.90 21.00 21.20 22.00 22.00
- There are 2 suburbs with pratio higher than 21 which is the 99th percentile

4 D - Suburs with avg. rooms per dwelling more than 7 / 8

- There are 64 suburbs with avg. rooms per dwelling more than 7
- There are 13 suburbs with avg. rooms per dwelling more than 8

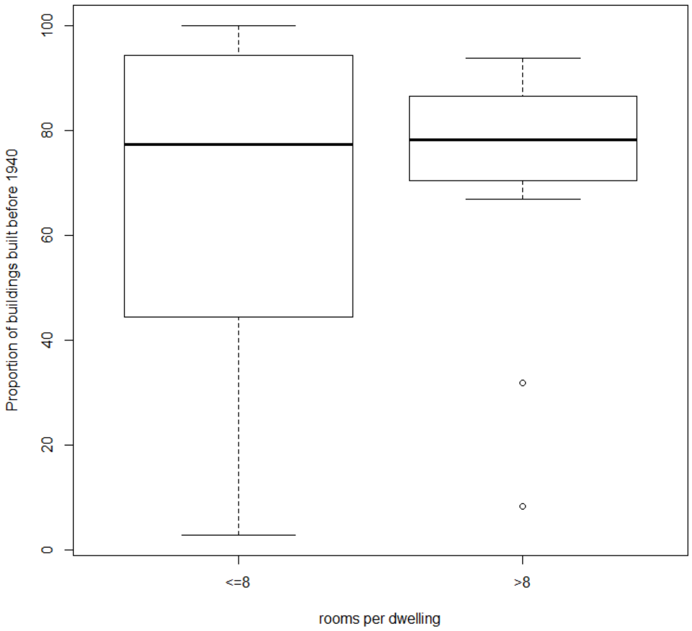
Findings

- Comparing the crime rates amongst suburbs with avg. rooms per dwelling more than 8 and the rest
- Crime rates are far lesser in the suburbs where avg. rooms per dwelling is higher than 8

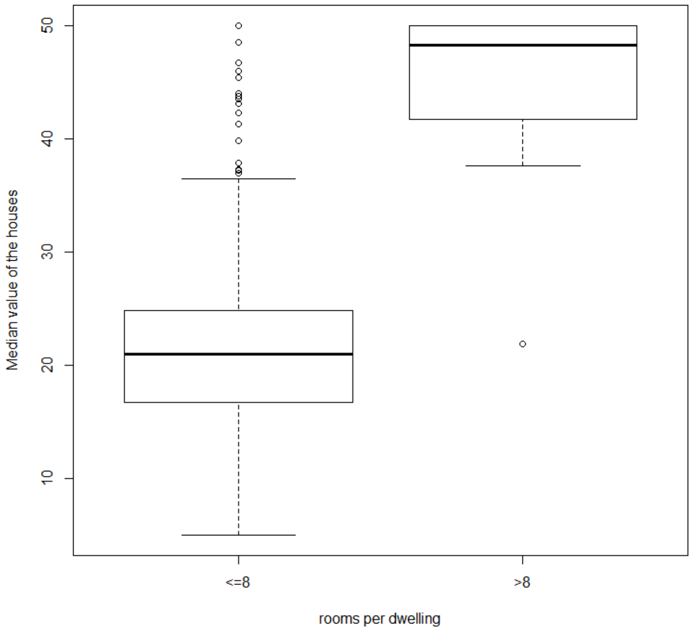


- Checking if the bigger houses belong to the previous generations, does the modern houses tend to be smaller?
- The box plot shows that the suburbs where the proportion of the houses that were built before 1940 is high, are those with hi gher rooms per dwelling. This

proves our hypothesis that older houses tend to be bigger



- It's obvious, we expect the suburbs with avg. rooms per dwelling higher than 8 should be of higher value
- The below boxplot confirms the above expectation


Median value of the houses
<=8
>8
rooms per dwelling