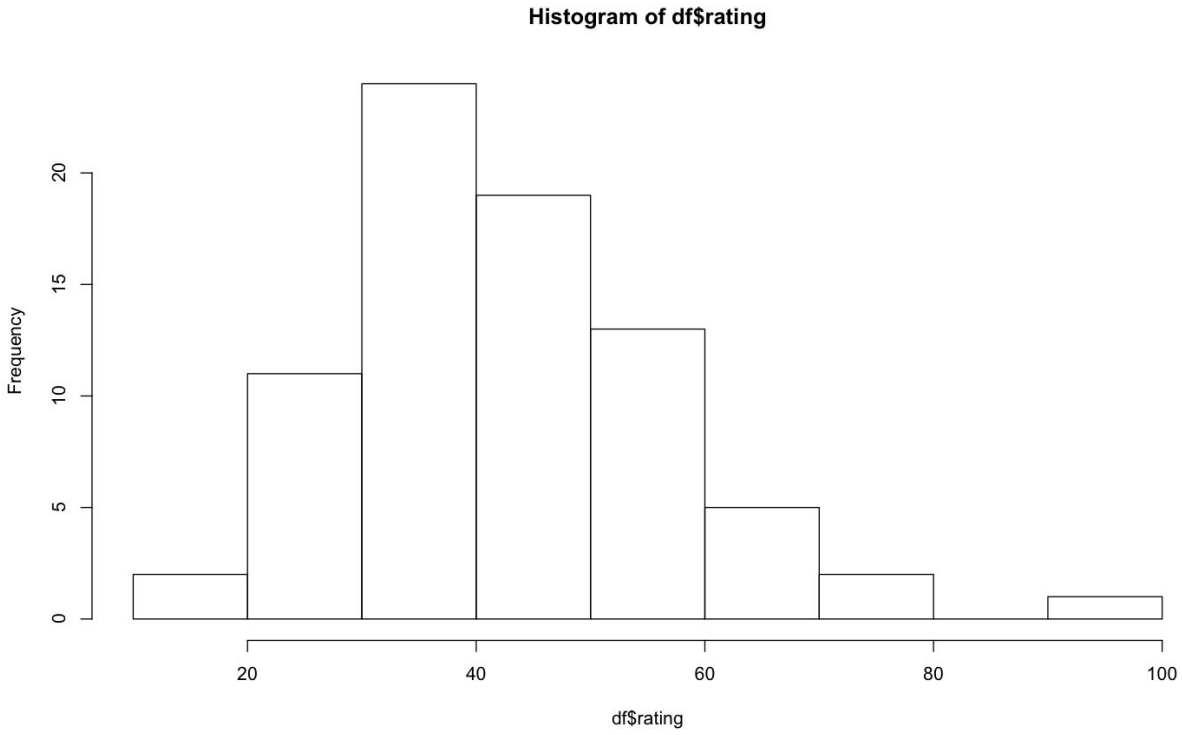
Statistical Data Mining 1 - Homework 1

Q1 A.

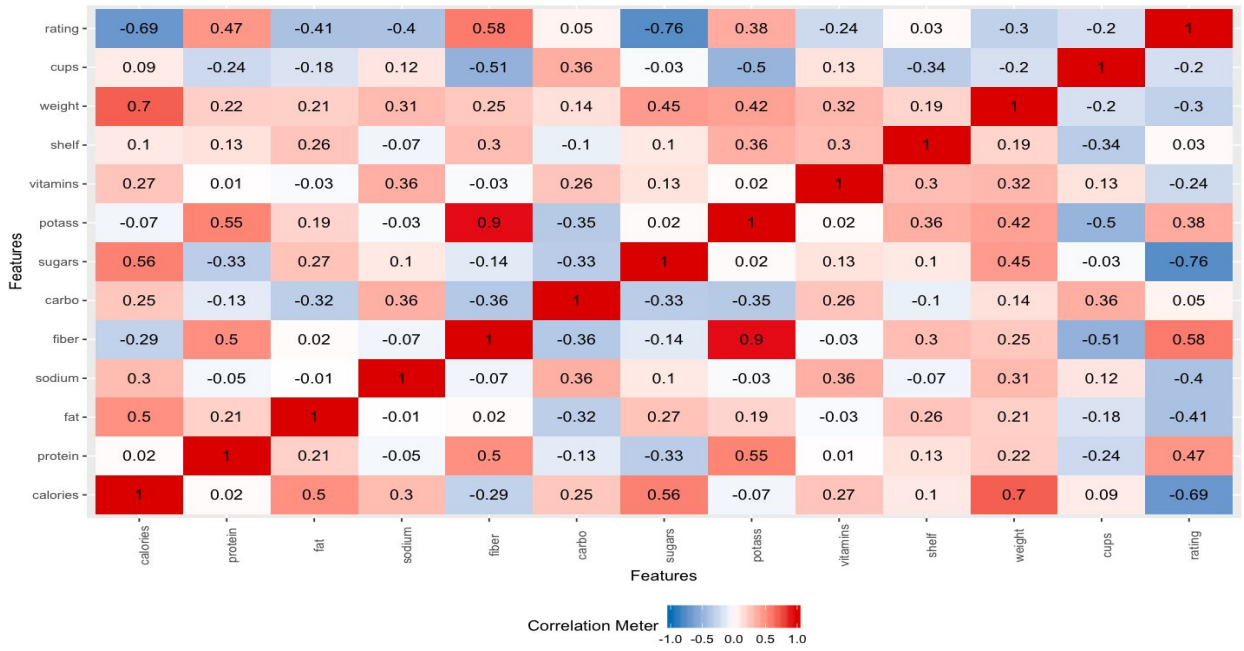
The assumption of linear regression is that the output is in a Gaussian distribution. The output rating here looks close to a Gaussian distribution so we can perform a linear regression. We can also see a possible outlier of rating which is lying separate from other values.



- The following is the output of the EDA of the dataset. There are 16 variables and the insights and findings from each variable are present as comments in the R code. The correlation matrix shows that six input variables are highly correlated with the output variable hence here in documentation those six will be described in detail along with the two other categorical variables mfr and type:

- Correlation matrix: Correlation shows the degree of association between each of the variables. We can see that calories and sugars are highly correlated with the output rating. The other variables which have moderate correlation are fiber, protein, fat, and sodium. These will be the variables that will be described here. Also, we see that fiber and potass are highly correlated, so we might drop potass as fiber has a higher correlation with rating.

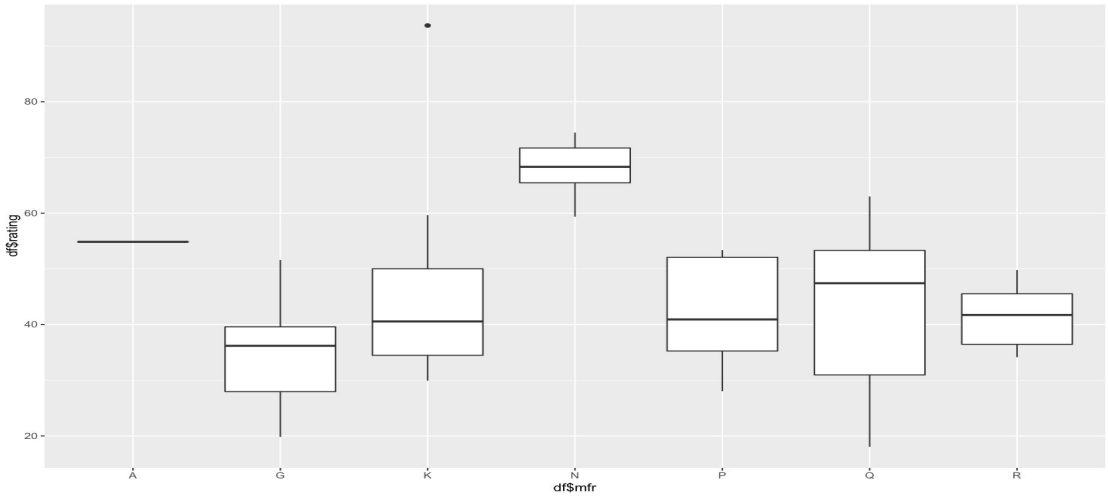
Figure:1



- **Name**: It is the primary key of the dataset so it will not be a part of the input variables

- **Mfr**: The manufacturer. We can see in Figure 2 that Nabisco has the highest mean rating of all the manufacturers. Kellog seems to have an outlier with a very high rating. We can also see that American food products have only one row of data for itself while Kelloggs and General Mills have more than 20 rows each. So there is a possibility that this dataset is biased towards the products of Kellogg or General Mills

Figure: 2



- **Type**: A categorical variable that takes values Hot(H) or Cold(C). Most of the products are Cold(C) and only 3 products out of 77 in the sample are Hot.

- **Calories**: We can see that calories are not linearly related to the rating in figure 3.1. Need to be transformed or bucketed into categorical variables such that it is **piecewise linearly** related to rating. We can see clearly that calories can be broken into 3 sub-categorical variables such that it is piecewise linearly related. Below are graphs for the relationship between calories and ratings before and after bucketing. We can clearly see that as calories increase then the rating decreases.

Figure 3.1:

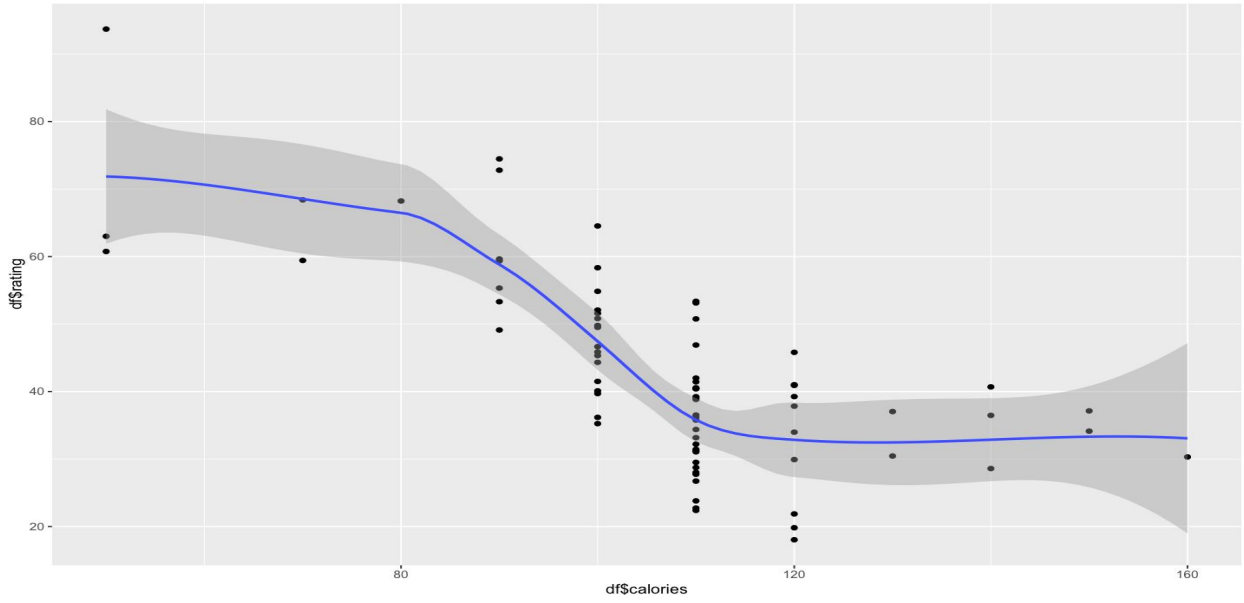
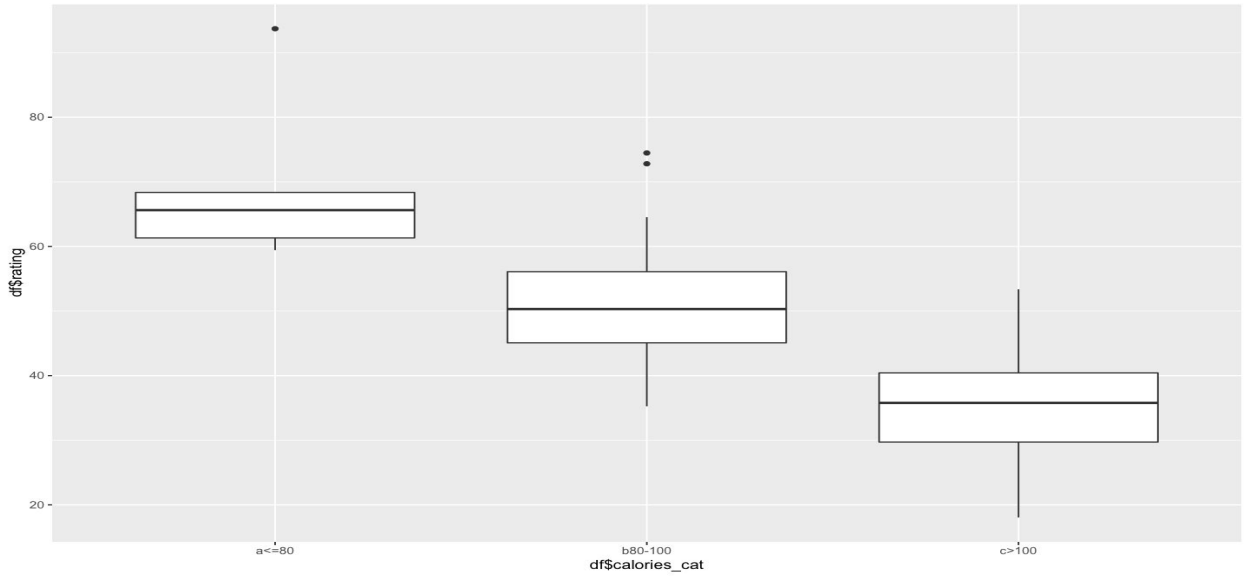
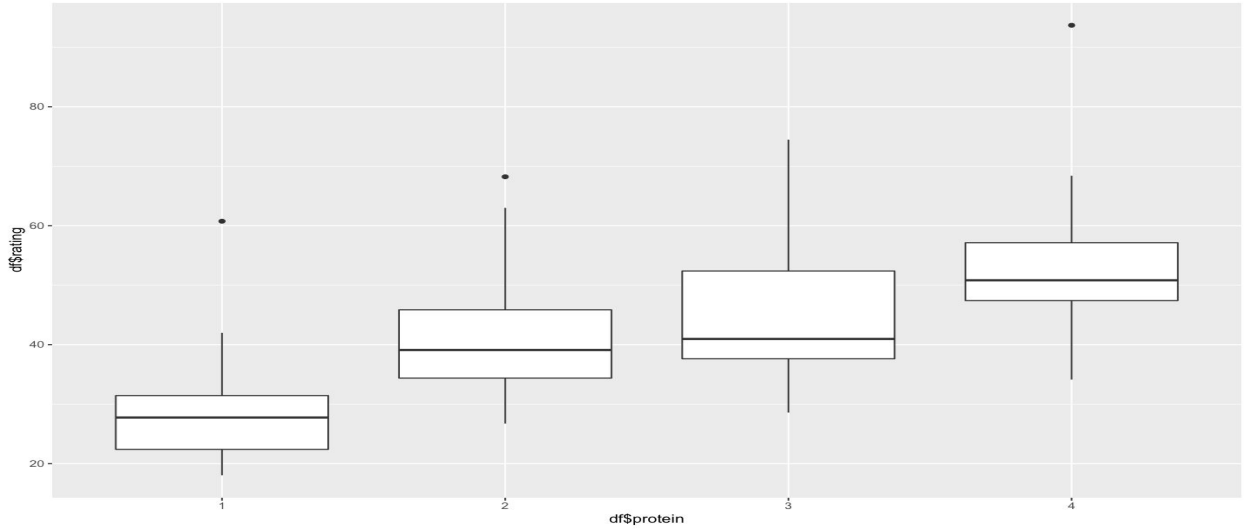


Figure 3.2:



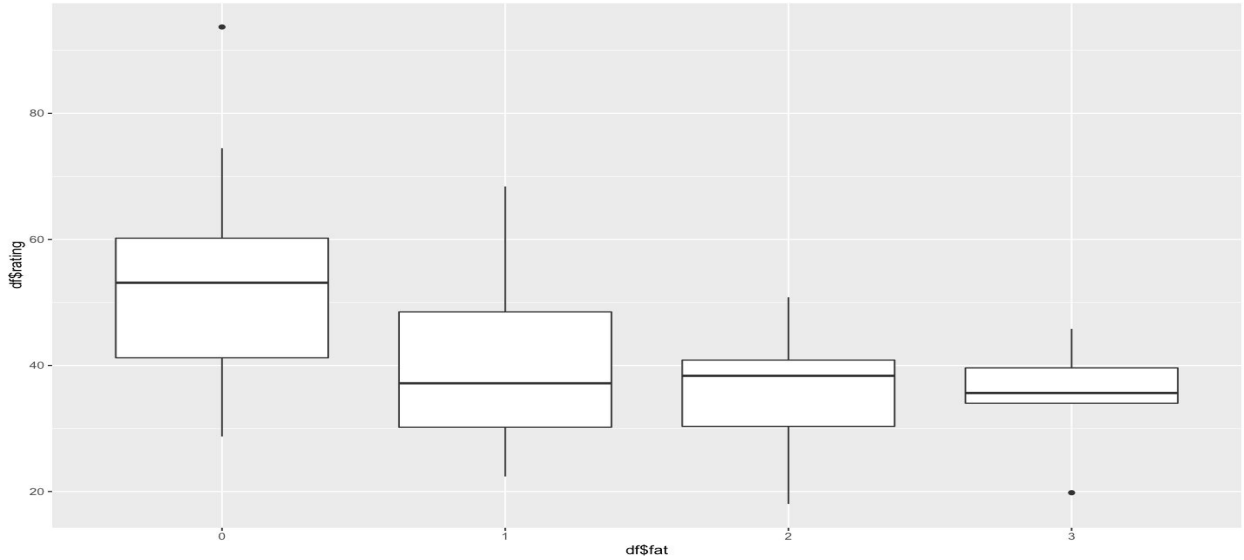
- **Protein:** The number of distinct values of protein is only 6 so it is better visualized as a categorical variable and most of its values are between the range (1, 4). So we can set a cut-off value of 4 for protein. The mean value of rating seems to increase with the increase in protein

Figure:4



- **Fat:** The number of distinct values of protein is five so it is also better visualized as a categorical variable than as a numeric variable. We can also see that only one row has the value 5 so we can cap the values at 3. The mean value of fat doesn't seem to impact fat beyond 0, so another way to categorize fat will be for values 0 and greater than 0

Figure:5



- **Sodium:** We can see that sodium is not linearly related to the rating based on the scatter plot. Since the correlation is quite low the values are scattered all over. So we can bucket the variable here similar to calories and see if we can find any relation. We can see that sodium value below 130 has a higher mean rating than those above it.

Figure:6.1

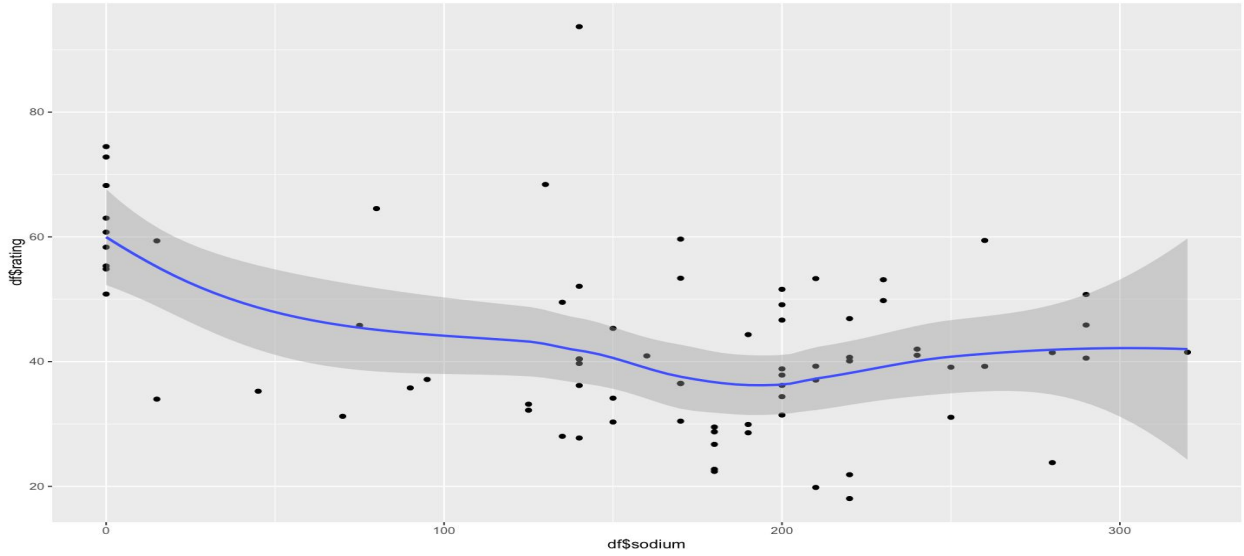
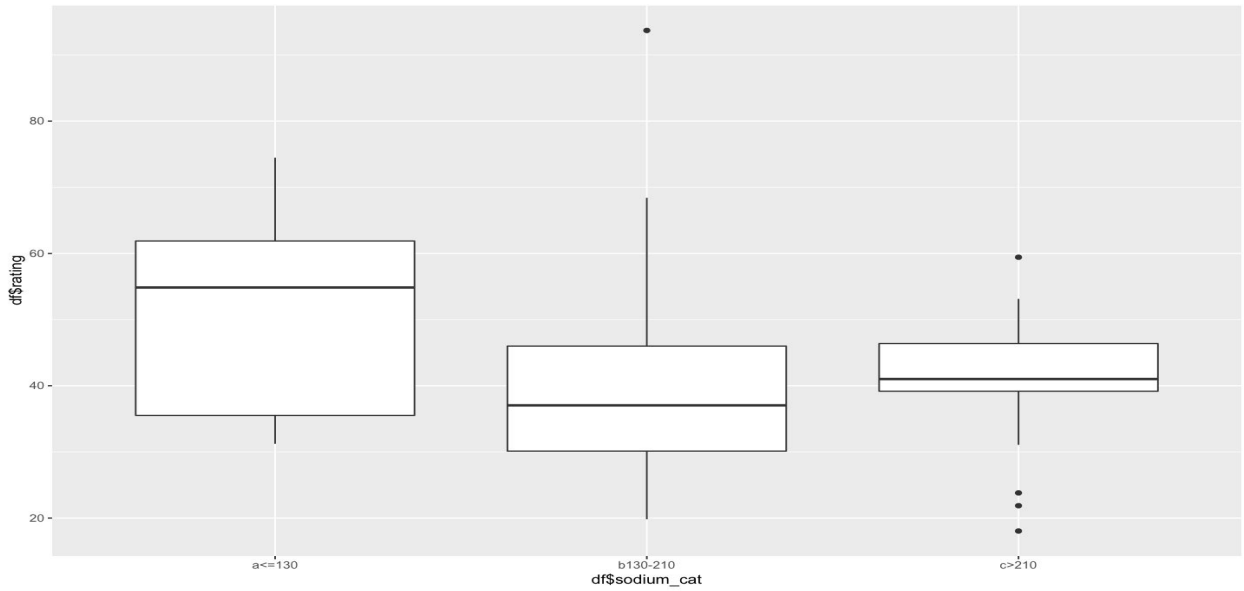
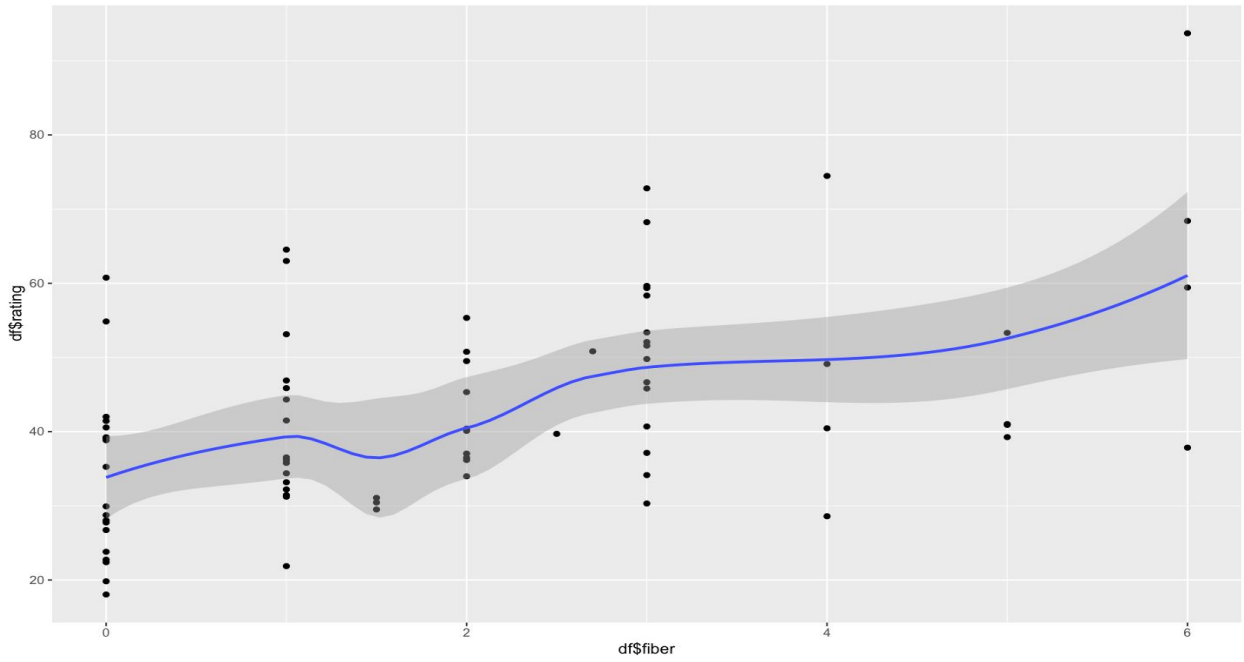


Figure:6.2



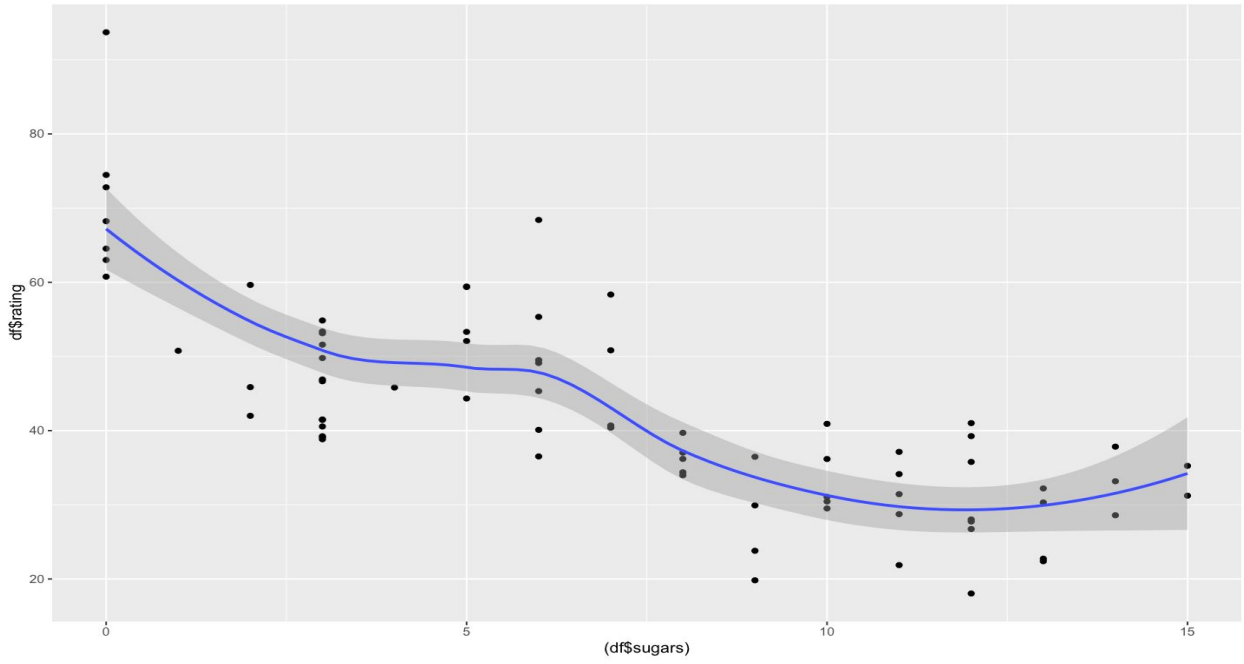
- **Fiber:** We can see in a scatter plot of fiber and rating that about 3 cereals have high fiber and can be considered outliers so we can cap the values of fiber at 6. The resultant plot as shown in Figure 7 below is almost linear and the increase in fiber slightly increases the rating

Figure 7:



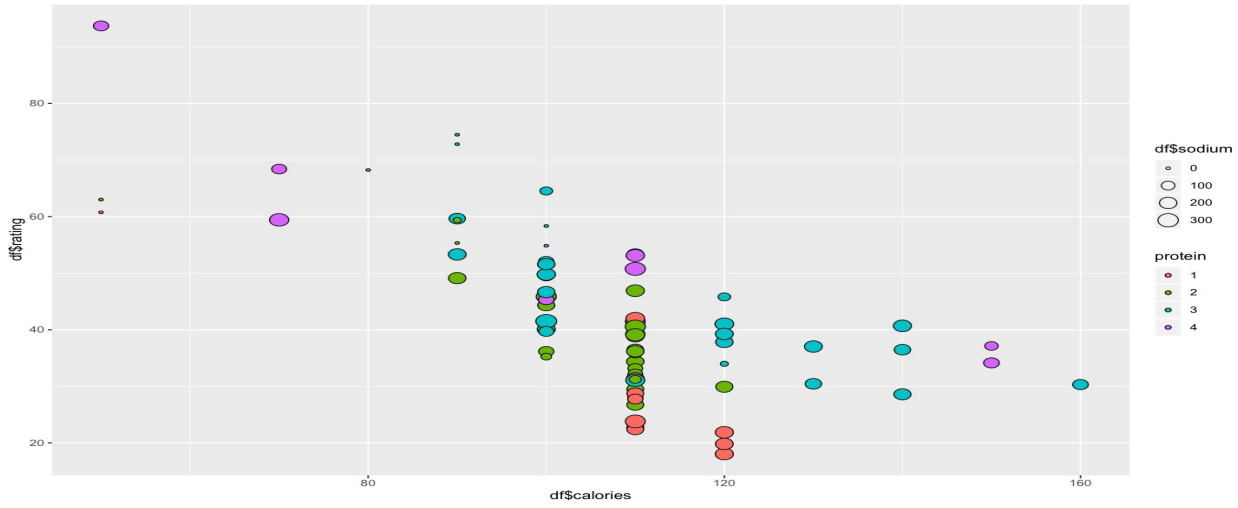
- **Sugars:** We can see that sugar is almost linearly related to rating with a negative slope explaining the strong negative correlation. There is a row whose value is less than 0 which has to be imputed. We impute this with the median of sugars. We could also think after looking at Figure 8 that sugars is related to rating as 1/x but this transform is only true for values of sugar less than 3.

Figure 8:

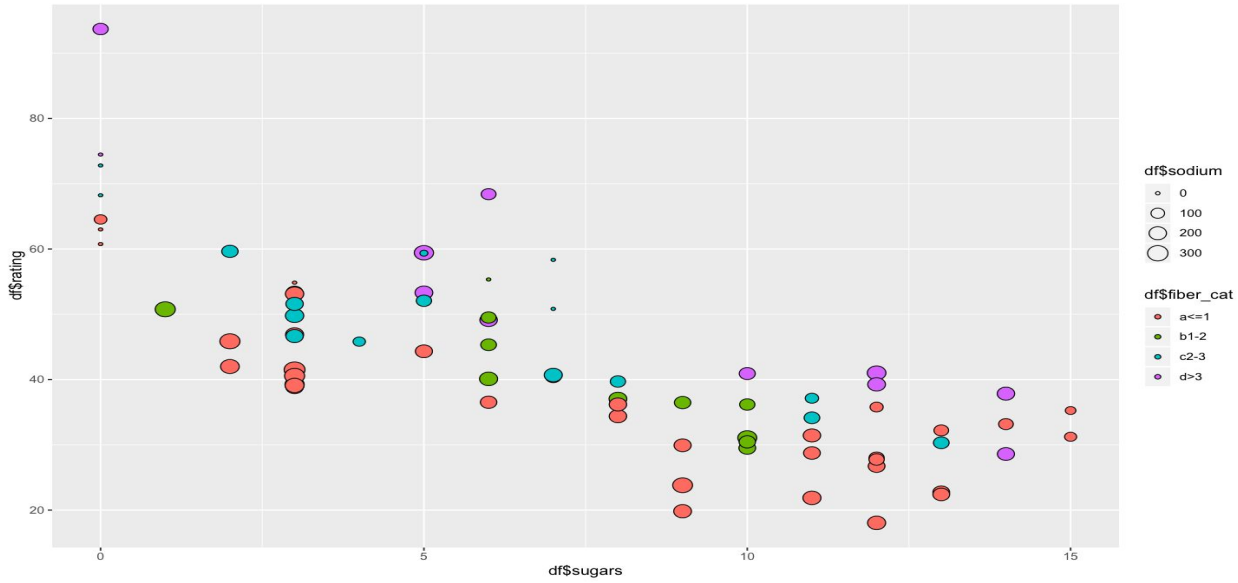


Other variables have been checked and similar analysis has been done in the R code. Not showing them here as the correlation was not strong with the response variable. We can look for interaction between these strongly correlated input variables and see if we can get any interesting insights.

The relation between rating, calories, protein, and sodium: We can see that for a higher rating, we have lower calories and higher protein and for a lower rating, we have high calories and lower protein so calorie: protein could be a good interaction variable. We also see that lower values of sodium along with lower calories and higher protein will give a high rating



The relation between rating, sugars, fiber, and sodium: We can see that for a lower value of sugar and lower value of sodium and medium values of fiber there will higher values of rating.



Q2:

Variables with outlier treatment:

Variables	**Outlier Treatment**
protein	If greater than 4 then cap them at 4
fat	If greater than 3 then cap the value at 3
fiber	If greater than 6 then cap the values at 6
carbo	If less than 10 then cap the values at 10

Some variables such as potass, carbo, and sugar which have missing values have been denoted by -1. The values for such variables have been imputed with the median value. The median has been chosen as the preferred measure of central tendency as it is least susceptible to outliers.

Also, the model has few categorical variables such as manufacturers, type, vitamins, and shelf. These variables have been one_hot encoded so that these categorical variables can be interpreted as numeric variables by the linear regression.

The various model outputs have been tabulated below. We can conclude that Model 2 is the best model as it has the lowest number of predictors and all the predictors of the input are highly significant($p<0.01$) with a very low standard error. The test R-squared for this model is the highest among all the models. It also has high train R^2 and high adj-R^2. The other models have higher train R^2 and train adj R^2 as they have overfitted. The residuals of all models are close to a normal distribution and F-stat for all models is less than 10^-16.

PS: p-value cut-offs have been taken as a thumb rule based on the p-values of the predictors. Std. error cut-offs have decided based on the distribution of standard errors of the predictors

Model	**Description**	**Input Variable List**	**P-value**	**Std.Error cut-off**	**Final_list**	**Train RMSE**	**Train R^2**	**Train Adj - R^2**	**Test RMSE**	**Test R^2**
Model 1	First iteration	(.-(potass+mfr_A+type_H+vit	0.05	1	calories+protein+fat+sodium+	2.33	0.9823	0.9741	3.0089	0.956

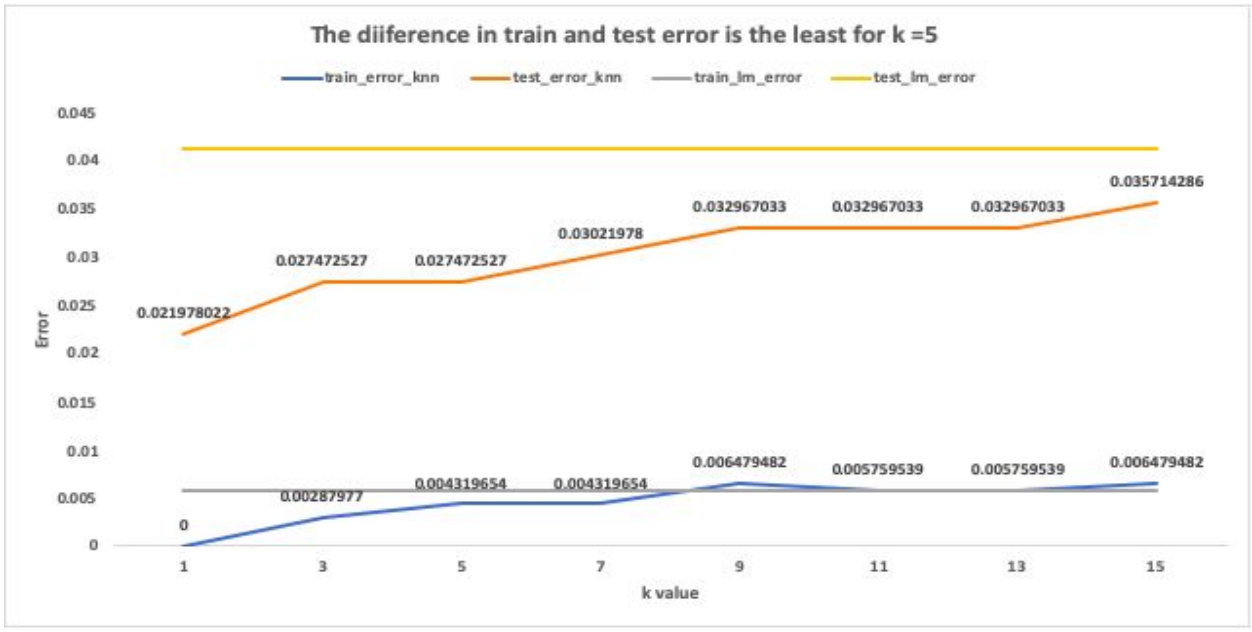
		amins_100+shelf_3))			fiber+sugars					
Model 2	**Final list of model 1 taken as input vars**	**calories+protein+fat+sodium+fiber+sugars**	**0.01**	**1**	**calories+protein+fat+sodium+fiber+sugars**	**2.968**	**0.9619**	**0.9579**	**2.22**	**0.9751**
Model 3	Considering 2nd order variables for interaction	(calories+protein+fat+sodium+fiber+sugars)^2	0.1	3	fiber+sugars+calories:fiber+protein:sugars+fat:sodium+fat:fiber+sodium:sugars	2.101	0.9859	0.9789	2.469	0.966
Model 4	Final list of model 3 taken as input	fiber+sugars+calories:fiber+protein:sugars+fat:sodium+fat:fiber+sodium:sugars	0.05	1	fiber+sugars+calories:fiber+fat:fiber+sodium:sugars	4.143	0.9271	0.918	2.549	0.952

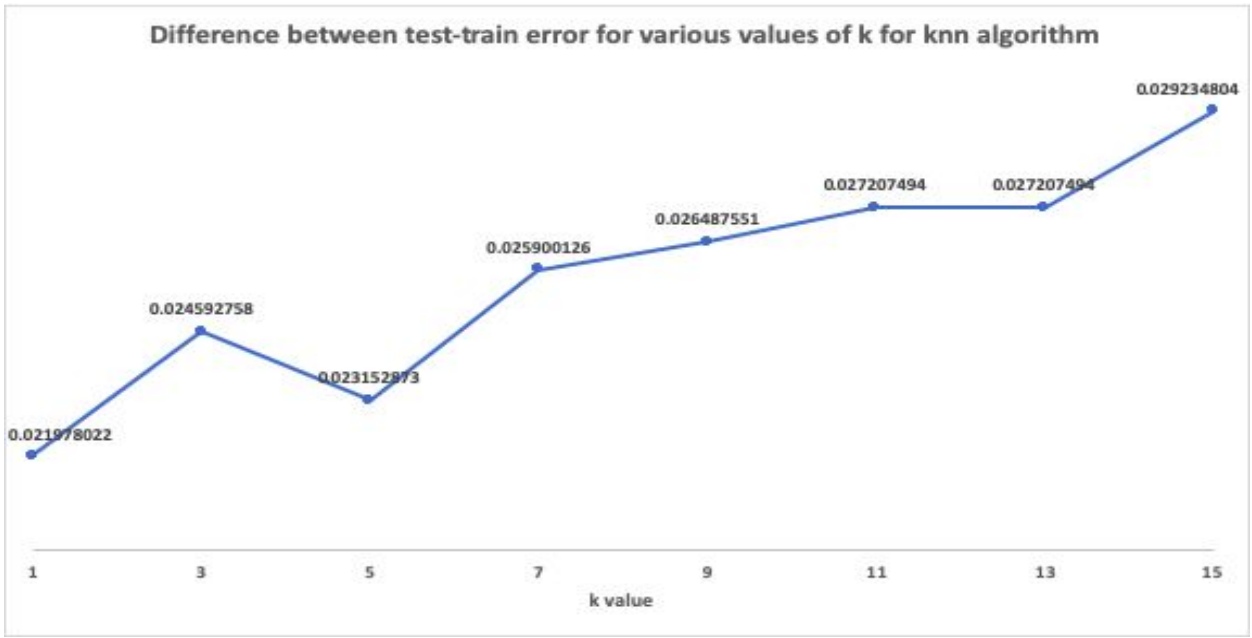
2(a) The predictors with significant relationship to the response variable are calories, protein, fat, sodium, fiber and sugars.

2(b) The co-efficient variable for sugars suggests that unit increase in sugars **decreases** the rating by about 1.574(B) when all other predictors are kept constant. The standard error(SE) is 0.11 and the p-value is <10^-16, so we can almost be 100% sure based on this sample that the beta estimate will be within B-SE and B+SE

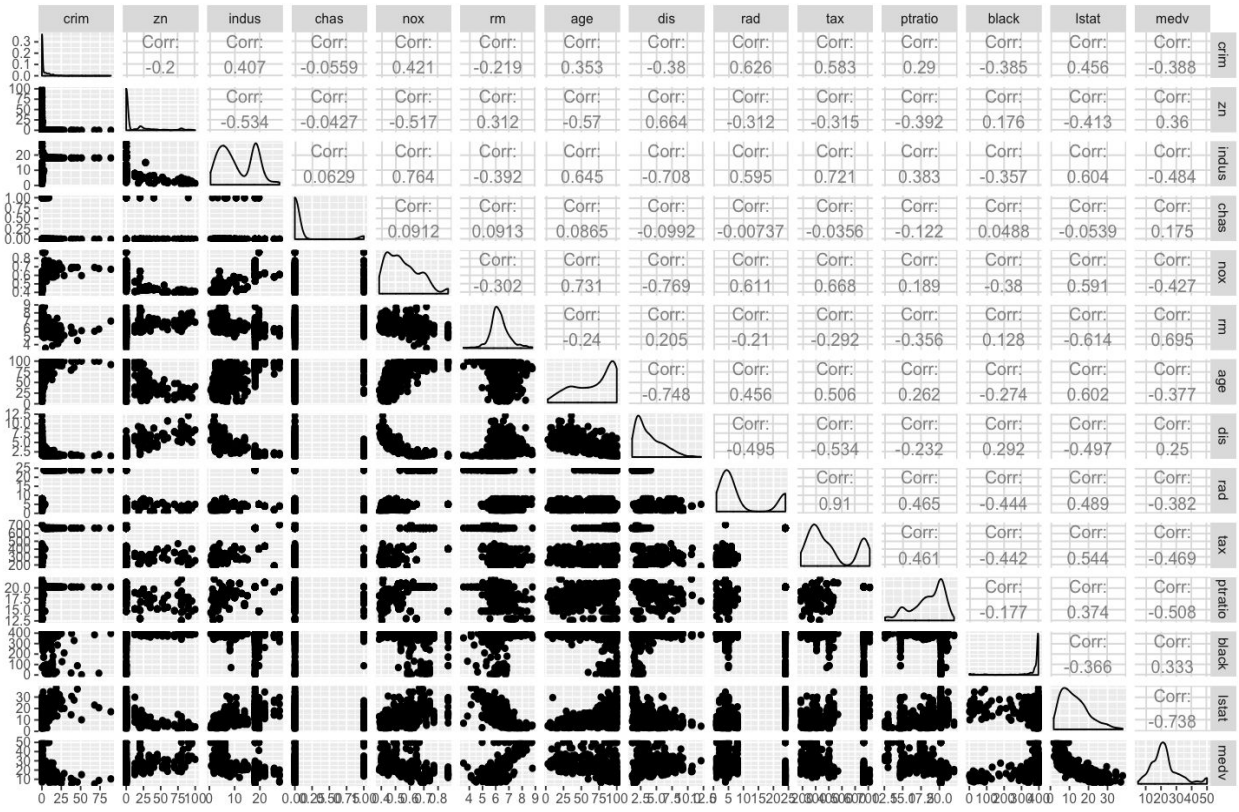
2(c) The significant interactions are sugars:sodium, fiber:fat and fiber:calories in the order of their p-value. The highest B estimate is that of fiber:fat.

3. We can see that the difference in train and test error for knn algorithm is lowest for k=5 with , hence that is the best value of k. The errors for knn is always lower than linear regression and hence it should be the preferred model over linear regression





4 (a).



The scatter plots combinations with high correlation(>0.7) and their respective medians are given below:

Variable 1	Variable 2	Median 1	Median 2	Correlation
Indus	Tax	9.69	330	0.72
Indus	Nox	9.69	0.538	0.76
Indus	Dis	9.69	3.207	-0.71
Nox	Age	0.538	77.5	0.73
Nox	Dis	0.538	3.207	-0.77
Age	Dis	77.5	3.207	-0.75

The following are the insights:

1. Strong positive correlation between high density of industries and higher tax shows that industrial suburbs are high the tax-paying entities. In fact, apart from one of the five Boston employment centers, all others pay tax greater than the median tax of Boston.
2. Strong positive correlation between high density of industries and higher nox shows that industrial suburbs are producing emissions of nitrogen oxides. The highest emissions are produced by suburbs with highest density of industries and only one suburb with industrial density greater than 15 has emissions less than 0.55
3. Strong negative correlation between high density of industries and distance shows that most of the suburbs are further away from the industrial centres. We can also see that there are a cluster of suburbs which are very close by to the five employment centres
4. The strong positive correlation between nitrogen oxide emissions and age shows that most of the suburbs with high nitrogen oxide emissions have been occupied for a long time and these people have the risk of health problems
5. The strong negative correlation suggests that people who live further away from employment centers have less exposure to nitrogen oxide emissions and can lead a healthier life
6. The strong negative correlation suggests that people who live in old houses are the closest to employment centers. The high median values of age and the low median value of dis indicate a high concentration of suburbs in that quadrant

4(b) There are no predictors which have a strong correlation with crime. Only tax and rad have moderate correlation with crim rate. High taxed employment suburbs are at slightly higher risk of crim that others while those suburbs with radial connectivity index greater than 4 have higher chance of crime

4(c)
(i) The distribution of a crim with suburbs is a pareto and it is skewed. More than 90% of population has crime less tha 11 while the max for a suburb is more than 80. Thus the about 10% of the suburbs have very high crim rates
(ii) There is a steep jump in number of suburbs with student population ratio at 20. About 10% of the suburbs have high student teacher ratio
(iii) 25% of the suburbs which are in high employment centers have very high tax rate.

4(d)
64 suburbs average more than seven rooms per dwelling
13 suburbs have more than 8 rooms per dwelling
For suburbs with more than 8 rooms per dwelling there is significant difference in medians for crim, lstat, indus and medv.