# Assignment Cover Sheet

**Computing, Engineering and Mathematics**

WESTERN SYDNEY UNIVERSITY

| | |
|---|---|
| **Student Name** | Grishma Mainali |
| **Student Number** | [redacted] |
| **Unit Name and Number** | COMP7006 Data Science |
| **Lecturer** | [redacted] |
| **Due Date** | 13th October, 2023 |
| **Date Submitted** | 10th October, 2023 |

**DECLARATION**

*I hold a copy of this assignment that I can produce if the original is lost or damaged.*

*I hereby certify that no part of this assignment/product has been copied from any other student's work or from any other source except where due acknowledgement is made in the assignment. No part of this assignment/product has been written/produced for me by another person except where such collaboration has been authorized by the subject lecturer/tutor concerned.*

*I am aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism (which may retain a copy on its database for future plagiarism checking))*

*I hereby certify that no part of this assignment or product has been submitted by me in another (previous or current) assessment, except where appropriately referenced, and with prior permission from the Lecturer/Tutor/ Unit Co-Ordinator for this unit*

*I hereby certify that I have read and understand what the School of Computing and Mathematics defines as minor and substantial breaches of misconduct as outlined in the learning guide for this unit.*

***Signature**:* 

**Note: An examiner or lecturer/tutor has the right not to mark this assignment if the above declaration has not been signed)**

# Analyzing Century-Long Trends in US Per Capita Kilocalorie Consumption: A Multimodal Approach Using Multiple Linear Regression, Decision Tree and Principal Component Analysis

Grishma Mainali
Western Sydney University, Sydney, NSW, Australia

**This report examines a century long trend of calorie consumption in the United States using data from United States Department of Agriculture (USDA) gathered from 1909 to 2010. For the analysis two supervised learning models, Multiple linear regression and Decision Tree, and one unsupervised learning model, Principal Component Analysis, are used. The dataset contains the information of average per capita consumption of kilocalories and other macronutrients by the population of US for that period. The population consumed different nutrients that are necessary for the body on a daily basis, including carbohydrates, fats, proteins, and fiber. But these nutrients also contribute to make up the population's total calorie intake. Therefore, predictive models are built using supervised learning methods, multiple linear regression and decision trees to predict "kilocalories" which is used as the target variable from the dataset. The value of target variable is predicted using other relevant variables, the one which directly or indirectly cause the increase or decrease in the average kilocalories intake per person. For the supervised methods, the original dataset is divided into training and testing sets. Then the model is built using only the training set and is verified using the testing set so that we can evaluate the model's accuracy with new data. Then the predicting ability of the models are tested and finally the best models of the two supervised methods are compared. The created predictive models are evaluated for accuracy, and the two supervised learning approaches are compared to see which one provided a more accurate and efficient prediction of the target variable, Kilocalories. In addition, the dataset is explored and examined for underlying structures and patterns using the unsupervised method Principal Component Analysis which aims in uncovering intriguing correlations and connections between the variables in the dataset except the target variable, kilocalories.**

***Index Terms*—kilocalories, supervised – unsupervised learning, nutrients, analysis.**

## I. Introduction

This report presents a comprehensive analysis of the century-long trends in the per-capita calorie and nutrient consumption in the United States. The analysis of long-term dietary trends helps in providing the evolution of dietary practices among the population.

The energy contained in food is measured in terms of calorie. Kilocalorie (Kcal) is a fundamental unit used to indicate higher levels of calorie consumption. Calories are essential for people to survive since without them, our body's cells would not be able to operate and would eventually die. We obtain this required energy from the food and drinks we consume. We can determine how much potential energy a food contains by counting its calories. For instance, it is estimated that 1g of carbohydrates contain 4 kcal, 1g of protein contains 4 Kcal and 1g of fat contains 9 Kcal. [1]

The nutritional pattern is always changing over the time and as is that rate of calories consumption. It is important to understand the complex relation between the nutrients content in a diet of the population and its corresponding calories intake, which is crucial for determining nutritional sufficiency or dietary deficiency as well as addressing different associated health issues such as malnutrition or obesity. Hence, it is important to be able to accurately analyze and anticipate the kilocalorie consumption by the population.

This report focuses on the fact that the other nutritional consumptions are associated with determining the overall calorie consumption. In this study, we explore the long-term trends in the per-capita calorie intake and nutritional composition in the United States. For this, we imply machine learning techniques: multiple linear regression, decision tree and principal component analysis to find influencing factors and model trends in kilocalorie consumption as well as try to examine hidden patterns and underlying relationships within the dataset.

The following parts of this report will include an in-depth analysis of the research conducted as well as the findings.

## II. Data description, exploration and pre-processing

The dataset used for this project is titled "Kilocalories and Macronutrients per Capita per Day in the U.S. Food Supply, 1909 - 2010". The United States Department of Agriculture (USDA) calculated the dataset, and they are also the dataset's source. According to the dataset's description, the data are based on estimates of the amount of food available for consumption per person made by the USDA Economic Research Service. Also, according to the dataset's description, before 1930, resident population only was considered, except for the war years (1917–19). But starting in 1930, residents and members of the armed forces stationed abroad made up the population.

The dataset consisted of 11 variables and 102 observations. The 11 variables are named as: "Year", "U.S. population, July 12", "Kilocalories", "Carbohydrate", "Fiber", "Protein", "Fat", "Saturated Fatty Acids", "Monounsaturated Fatty Acids", "Polysaturated Fatty Acids" and "Cholesterol". The data for the year variable covered a century, from 1909 to 2010. The corresponding US population for that specific year made up the US population variable. The other variables included, as their names indicates, average per-person calorie or macronutrient intake. "For detailed exploration of the dataset and codes implemented, please refer to **Appendix A**."

When the original dataset was explored in R, all the variables consisted of numerical values but was stored in the

form of character. Hence, the first step in the analysis was to convert the datatype of each variable to a respective suitable data type. The variables' datatypes were set to either numeric or integer depending on the data values they contained.

As mentioned in earlier section of the report, Kilocalorie is the target variable, and all other variables are the independent variables. To understand the dataset and the relationship between the variables, I tried looking into the correlation coefficients between the variables. According to analysis of the plot and calculated value, the target variable had a reasonably high association with year, population, carbohydrate, and saturated fatty acid. It had a significant association with fat, monounsaturated fat, polyunsaturated fat, and protein, which had the highest correlation of 0.9. "For plots and values of correlation between the variables, please refer to **Appendix A**."

Another important step that was done before performing the supervised analysis, was dividing the dataset into training set and testing set. An important objective of supervised learning is to build a model that performs well with new and unfamiliar data. [2] If we use the complete dataset for fitting the supervised model and then perform accuracy testing, it is obvious to give high accuracy rate. But the real test of the model would be if we test it with new dataset and try to analyze how accurately it performs and if the model that we built is capable enough to give correct result for any kind of data. For this purpose, a validation technique called train test split is used which enables us to test a model's performance on unseen and new data. [2] A ration of 70:30 is used to create the training set and testing set respectively. The dataset originally has 102 observations, hence around 70% of those observations would be required to build a good model and the remaining 30% would be used to test the accuracy of the built model. The process of splitting dataset into training and testing set is done through random sampling without replacement but with fixed seed number so that the results remain consistent throughout.

## III. Aims and objectives

The main aim of this research project is to perform a detailed analysis of century-long trends in American per capita calorie consumption and nutritional composition, using supervised and unsupervised learning methods. The following paragraphs list the objectives that must be attained to accomplish this aim.

The first objective of this project is to create predictive models for Kilocalorie consumption pattern in the United States over the past century. Multiple linear regression and decision tree are the two supervised learning approaches used to create the predictive models. Relevant predictor variables will be integrated into these models to predict the actual numeric value of average Kilocalorie consumption per capita. The models for each supervised approach will be developed using training sets and tested against testing sets. Each method will have multiple models developed for it, but only the best model after the validation tests will be employed as the final model.

The second objective of this project is to compare the two supervised learning methods, multiple linear regression and decision tree. These two methods will each have a best model that will make prediction of the target variable. As we make predictions using the test set, we will also have access to the real values, which will allow us to compare the predicted and actual values and identify differences. Through this we can see the performance of the two supervised methods in identifying and predicting patterns in calorie consumption. We compare both approaches to examine which approach makes more accurate predictions and insightful discoveries about the dynamics of dietary behavior.

The third objective focuses on identifying the dataset's internal layout and patterns, omitting the target variable, kilocalories. This will be done using the unsupervised learning approach principal component analysis (PCA). Through this we hope to recognize undiscovered connections, groups, or patterns among the variables, and try to figure out potential factors influencing food choices or eating behavior of the population.

Lastly, this report aims to explore and identify trends and advancements in the field of nutrition and calories, which has a direct link with the public health.

## IV. Method 1 (Multiple linear regression)

Multiple Linear Regression (MLR) is a type of regression model used to establish the association between a dependent variable and two or more independent variables. It assists in determining whether the independent variables have a significant impact on the dependent variable as well as in calculating the value of the dependent variable at a certain value of the independent variables. [3]

In this research, MLR is used to create a model for the trends in American calorie consumption over the previous century. Using this technique, we can look at how different predictor factors, including intake of nutrients, affect calorie intake. Hence, in this model, Kilocalorie is the dependent variable and all other are the independent variables.

Before using the MLR modelling, it's crucial to make sure that all the required variables can be accessed in numerical form. This thing was addressed at the very start, as mentioned in the data pre-processing section, all the variables' datatypes were checked and set to suitable format for analysis.

Different models were created using the MLR technique, and different relevant predictor variables were chosen to be included in the regression model. The regression model was applied initially with all other independent variables considered, and then variables were gradually added or removed based on statistical significance and taking the correlation coefficient with the dependent variable into consideration.

For the model evaluation, cross validation technique was used. There were total of 5 models built using MLR but only one model would be finally implemented as the predictive model. For that K-fold cross validation was used to find the best model among the 5. From the validation, below presented graph was obtained:

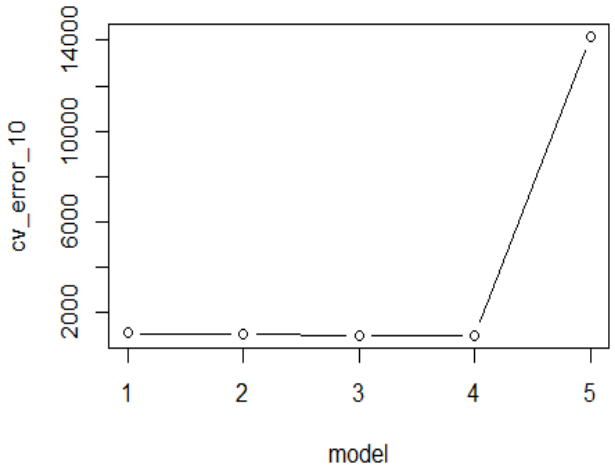



**Figure 1: Cross Validation to find the best model for MLR**

The obtained graph showed that there is no significant difference in error values between model 1, 2, 3 and 4. However, it was discovered after reviewing the model summaries that model 3 had integrated the fewest independent variables, all of which were significant. Additionally, among the other models, model 3 had the lowest AIC score. Through this information, model 3 was considered the best model for MLR. For the chosen model, model 3, the significant variable for predicting the target variable Kilocalories, were variables Carbohydrate, Protein and Fat.

"Please refer to **Appendix A** for a complete description of how the code was implemented, including data pre-processing, modelling, and evaluation."

## V. Method 2 (Decision tree)

Decision tree is another supervised technique that is employed to model Kilocalorie consumption trends in the United States. A decision tree is a type of flowchart that's useful in outlining probable outcomes, alternative courses of action, and ways to avoid undesirable ones. It is made up of a root node, branches, and leaf node, which when viewed together resembles a tree. It's a commonly used method for predictive framework that enables us to map out several options and choose the action that has the best possibility of succeeding. It creates a tree structure where each leaf node represents the projected outcome, and each internal node indicates a choice or split based on a certain predictor. [4] The prediction model for this study will be based on the same foundation.

Decision tree can be applied for any kind of datatypes within the dataset. Having all numeric variables in the dataset, regression decision tree was chosen to be implemented. There was no need to alter anything with the data values, however, the lengthier names of some variables were shortened to be viewed easily in the tree.

Initially the decision tree model was employed using target variable Kilocalories as dependent variable and all other variables as the independent variables. However, in the summary of the model only two variables Fiber and Year were considered by the mode as significant and were used to make up the decision tree. The tree has just 5 terminal nodes as a result. The next action was to evaluate the tree model using cross validation to determine the best tree size. After applying cross-validation evaluation below graph was obtained:

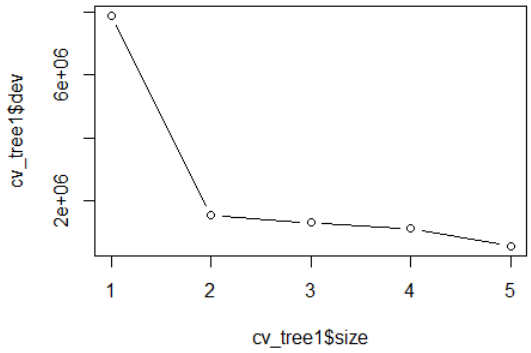



**Figure 2:Cross Validation to find the best size of decision tree**

As observed in the cross-validation figure above, the error for the tree size gradually decreases until it reaches the final size of 5. Hence, the best size for the tree model is 5, which is the same as the regression tree originally obtained. Therefore, the tree obtained above is the best model and there is no need to prune it further.

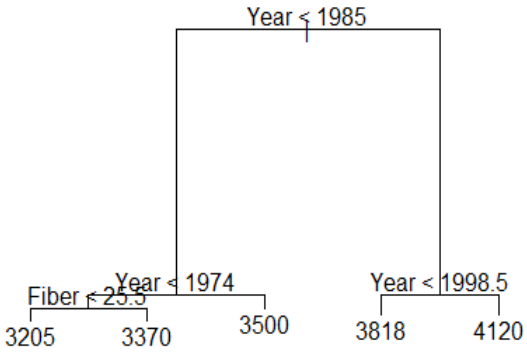



**Figure 3: Decision Tree**

The decision tree constructed for the dataset in its final form is shown above. It displays the considered variables and the nodes relating to it. From the obtained tree some information can be obtained about the impact of the considered variables

on the target variable such as:

- Year is the most important predictor in determining the Kilocalories and fiber is the other significant variable.
- For the year before 1974 and when the fiber intake was less than 25, the average calorie consumption was 3205, which is the least consumption as shown by the tree.
- For the year after 1998, the average calorie consumption was 4120, which is the highest consumption as shown by the tree.

Also, the tree shows that through the years the average consumption of calories per capita has been increasing.

"Please refer to **Appendix B** for a complete description of how the code was implemented, including data pre-processing, modelling, and evaluation."

## VI. Model comparision

In this section, the two supervised learning methods applied in this report namely Multiple Linear Regression (MLR) and Decision Tree Analysis are compared. These methods were applied to create the predictive models for target variable Kilocalories. MLR assumes that there is a linear relation between the dependent variable and independent variables and produces linear equations to express their relationship. Decision tree can be applied for regression as well as classification problems and produces hierarchical tree structure which display how independent variables affect dependent variables.

The objective of this section is to compare the two predictive models' levels of accuracy and identify which one was able to provide a more accurate prediction of the target variable. Both approaches were developed using training sets that were randomly derived from the original dataset with a ratio of 70%, and tested with testing sets that were randomly derived from the original dataset with a ratio of 30%.

Both the methods were compared based on how accurately or with how little error they were able to predict the target variable. As the target variable is numeric, the difference between the predicted value and the real value would be numeric as well. The best models from each approach were put to the test using testing sets, and for each value that was predicted, there was a reference to the actual value. Hence to calculate the accuracy of the models, Mean Squared Error (MSE) is calculated. MSE represents the average of the squared difference between the original and predicted values in the dataset. It also gives the measure of variance of the residuals from the model. Root Mean Squared Error (RMSE) is the square root of MSE and gives a measure of standard deviation of residuals. [5]

Below presented is a table containing the MSE and RMSE for MLR and Decision tree:

| | MSE | RMSE |
|---|---|---|
| MLR | 864.5705 | 29.40358 |
| Decision Tree | 9630.472 | 98.13497 |

The table above makes it evident that MLR has lower MSE and RMSE than Decision tree. The lower the MSE number, the better, as this indicates a smaller gap between the projected and actual values. The difference could be because of the fact that Decision tree only considered two variables whereas MLR considered more variables which were significant as well. Hence for this analysis, MLR proved to be a better approach for building predictive model.

## VII. Principal Component analysis

Principal Component Analysis (PCA) is a type of unsupervised technique which is also known as linear dimensionality reduction technique. It is utilized for transforming high dimensional data into lower dimensional representation without losing the important information. It works by retaining the sections within the dataset containing greater variation of data and removing the ones with lower variations. [6]

The objective to use PCA is for dimension reduction. In this project, the dataset contains different variables. Through PCA, all these complex relationships and information contained within the dataset is reduced into sets of Principal Component (PC), while still capturing all the significant variations of the data.

Before employing PCA into the original dataset, it is important to only include numeric variables, which is already the case of our dataset, and the target variable was removed. After modelling the dataset with PCA, 10 different PC were prepared. The variance captured by each component is presented in a plot below:

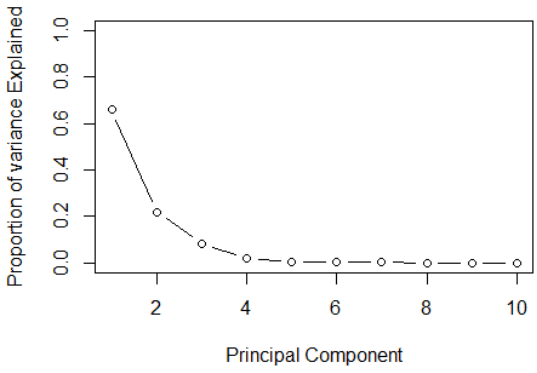



**Figure 4: Proportion of variance captured by each PC**

The first principal component explains about 66% of the variance in the data, the next principal component explains about 22% of the variance and so forth.

The cumulative proportion of variance captured by the components is presented below:

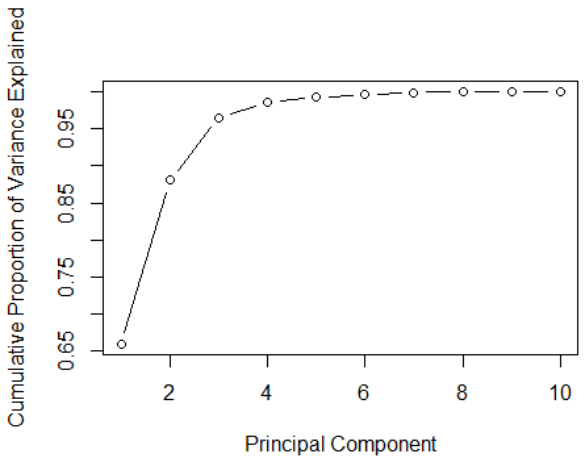



**Figure 5: Cumulative Proportion of variance explained by PC**

From the figure above, the first PC explains about 66% variation in the data set. The first and second PCs together explain about 88% of the variation in the data set. The first three PCs together explain about 96% of the variation in the data set and so forth.

Understanding the principal components is crucial for gaining insightful knowledge and an improved understanding of the patterns in the dataset. The reduced dimensionality of the dataset obtained through PCA can be viewed as:

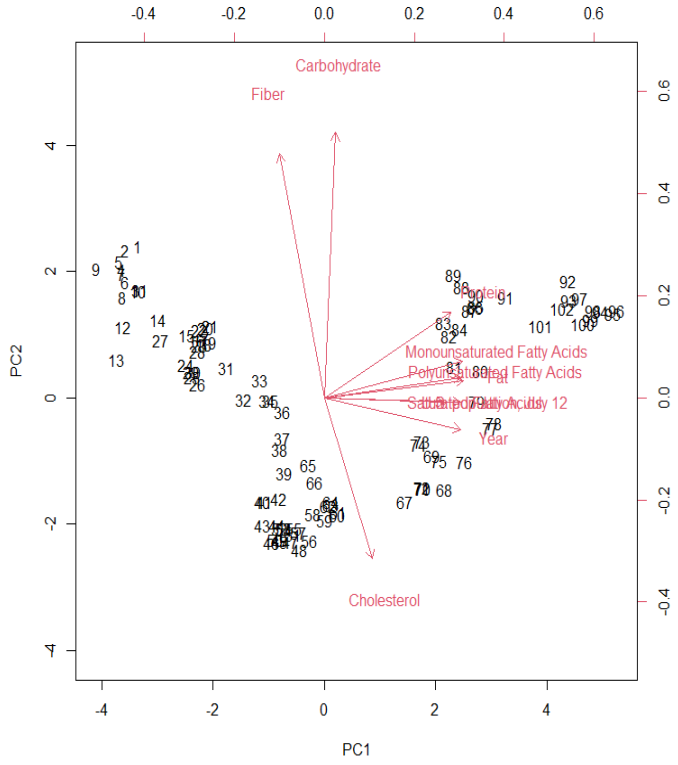



**Figure 6: PCA Biplot**

Some deductions made from the PCA biplot are as follows:

- First loading vector (PC1) places approximately equal weight on fat, protein, different fatty acids, population and year, with much less weight on fiber, carbohydrate and cholesterol.
- The second loading vector (PC2) places most of its weight on fiber, carbohydrate and cholesterol and much less weight on the other features
- We can see that with the increase in year the population has grown and along with that the average consumption of protein and fats too have increased.
- Interesting, Fats and protein are closely located (indicating that they are positively correlated). It is observed that people who consume more protein are likely to be consuming more fat as well.
- Cholesterol and Fiber are at 180-degree angle from each other indicating no correlation at all between them, which means that eating food containing high fiber contribute to no cholesterol intake.

"Please refer to **Appendix C** for a complete description of how the code was implemented, including data pre-processing, modelling, and analysis."

## VIII. Results and Recommendation

This report utilized multiple linear regression (MLR), decision tree analysis, and principal component analysis (PCA) to explore and analyze the patterns of kilocalorie and macronutrients consumption and this section will summarize the main conclusions derived from those analysis.

For supervised learning methods, models built using both MLR and decision tree were successfully able to predict the target variable based on its relationship with other relevant variables. With RMSE of 29 and 98 for MLR and Decision tree respectively, the model built using MLR was able to give more accurate predictions for the target variable. The fact that there were only 102 observations in the dataset should be taken into account because it may have a significant impact on the performance differences between the two supervised approaches. MLR, however, was proven to be more valuable for the dataset that was utilized, and it may be the best choice for future datasets with a similar pattern, category of variables, and amount of observations.

The unsupervised method, Principal Component Analysis (PCA) facilitated dimension reduction and alternative exploration of data structures and behavior. Through its graph, some hidden relations and patterns were found within the dataset which provided some new insights into the variables of the dataset.

In conclusion, this study provides a comprehensive understanding of kilocalorie consumption trends in the United States over a century through utilization of different analytical techniques. The insights gained through it can provide a foundation to understand public health, choice of food consumption and dietary practices of population in the 21st century.

## IX. References


[1] A. Kandola, "How many calories are burned while walking_," MedicalNewsToday, 30 05 2019. [Online]. Available: https://www.medicalnewstoday.com/articles/325323. [Accessed 02 10 2023].

[2] M. Galarnyk, "Train Test Split_ What it Means and How to Use It _ Built In," Builtin, 28 7 2022. [Online]. Available: https://builtin.com/data-science/train-test-split. [Accessed 3 10 2023].

[3] R. Bevans, "Multiple Linear Regression _ A Quick Guide (Examples)," Scribbr, 20 2 2020. [Online]. Available: https://www.scribbr.com/statistics/multiple-linear-regression/. [Accessed 4 10 2023].

[4] R. Cravit, "What is a Decision Tree & How to Make One [+ Templates]," Venngage, 2 2 2023. [Online]. Available: https://venngage.com/blog/what-is-a-decision-tree/. [Accessed 4 10 2023].

[5] A. Chugh, "MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better_ _ by Akshita Chugh _ Analytics Vidhya _ Medium," Medium, 8 12 2020. [Online]. Available: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e. [Accessed 4 10 2023].

[6] A. Sharma, "Principal Component Analysis (PCA) in Python Tutorial _ DataCamp," Datacamp, 1 2020. [Online]. Available: https://www.datacamp.com/tutorial/principal-component-analysis-in-python. [Accessed 4 10 2023].

## APPENDIX A (MULTIPLE LINEAR REGRESSION)

- Dataset

The dataset used for the analysis is called "Calories.xlsx" which is stored locally in F disk of my computer. The dataset is stored in an excel format. So I used the library "readxl" to import the dataset into a variable "ds" and viewed it.

```
library("readxl")
ds = read_excel("F:/WSU/Year 1/Semester 1/Data Science/Data Science/Assignment/Calories.xlsx")
head(ds)
```

```
## # A tibble: 6 × 11
##   Year  `U.S. population, July 12` Kilocalories Carbohydrate Fiber Protein Fat
##   <chr> <chr>                      <chr>        <chr>       <chr> <chr>   <chr>
## 1 <NA>  --- Millions ---           <NA>         <NA>        <NA>  <NA>    <NA>
## 2 1909  90.49                      3400         499         29    101     119
## 3 1910  92.406999999999996        3400         498         29    99      117
## 4 1911  93.863                     3400         492         28    98      118
## 5 1912  95.334999999999994        3400         493         28    98      115
## 6 1913  97.224999999999994        3400         492         28    97      116
## # ℹ 4 more variables: `Saturated Fatty Acids` <chr>,
## #   `Monounsaturated Fatty Acids` <chr>, `Polyunsaturated Fatty Acids` <chr>,
## #   Cholesterol <chr>
```

```
dim(ds)
```

```
## [1] 116  11
```

```
names(ds)
```

```
##  [1] "Year"
"U.S. population, July 12"
##  [3] "Kilocalories"
"Carbohydrate"
##  [5] "Fiber"
"Protein"
##  [7] "Fat"
"Saturated Fatty Acids"
##  [9] "Monounsaturated Fatty Acids"
"Polyunsaturated Fatty Acids"
## [11] "Cholesterol"
```

The dataset has 11 variables and 116 observations. However, there are some rows in the dataset which contains some information about the dataset but not the actual data values. These rows are not required for analysis and will cause problems. So, removing the unnecessary rows.

```
dt = ds[-c(1,104:116),]
dim(dt)
```

```
## [1] 102  11
```

"dt" is a new dataset which contains only the rows from original dataset which has data values.

```
str(dt)
```

```
## tibble [102 × 11] (S3: tbl_df/tbl/data.frame)
##  $ Year                          : chr [1:102] "1909" "1910" "1911" "1912" ...
##  $ U.S. population, July 12      : chr [1:102] "90.49" "92.406999999999996" "93.863" "95.334999999999994" ...
##  $ Kilocalories                  : chr [1:102] "3400" "3400" "3400" "3400" ...
##  $ Carbohydrate                  : chr [1:102] "499" "498" "492" "493" ...
##  $ Fiber                         : chr [1:102] "29" "29" "28" "28" ...
##  $ Protein                       : chr [1:102] "101" "99" "98" "98" ...
##  $ Fat                           : chr [1:102] "119" "117" "118" "115" ...
##  $ Saturated Fatty Acids         : chr [1:102] "50" "49" "49" "48" ...
##  $ Monounsaturated Fatty Acids: chr [1:102] "45" "44" "45" "44" ...
```

```
##  $ Polyunsaturated Fatty Acids: chr
[1:102] "13" "12" "12" "12" ...
##  $ Cholesterol                 : chr
[1:102] "440" "440" "460" "440" ...
```

While checking the structure of dataset, it can be seen that all the values are in numbers but are stored as character. So we need to change the datatype of variables into correct ones.

```
dt = type.convert(dt, as.is=FALSE)
str(dt)

## tibble [102 × 11] (S3:
tbl_df/tbl/data.frame)
##  $ Year                        : int
[1:102] 1909 1910 1911 1912 1913 1914
1915 1916 1917 1918 ...
##  $ U.S. population, July 12    : num
[1:102] 90.5 92.4 93.9 95.3 97.2 ...
##  $ Kilocalories                : int
[1:102] 3400 3400 3400 3400 3400 3400
3300 3300 3200 3300 ...
##  $ Carbohydrate                : int
[1:102] 499 498 492 493 492 486 483 473
472 468 ...
##  $ Fiber                       : int
[1:102] 29 29 28 28 28 27 28 27 28 27 ...
##  $ Protein                     : int
[1:102] 101 99 98 98 97 95 94 93 92 94
...
##  $ Fat                         : int
[1:102] 119 117 118 115 116 118 117 117
113 121 ...
##  $ Saturated Fatty Acids       : int
[1:102] 50 49 49 48 48 48 48 48 46 49 ...
##  $ Monounsaturated Fatty Acids: int
[1:102] 45 44 45 44 44 45 45 45 44 47 ...
##  $ Polyunsaturated Fatty Acids: int
[1:102] 13 12 12 12 12 13 13 13 12 14 ...
##  $ Cholesterol                 : int
[1:102] 440 440 460 440 430 430 430 430
410 420 ...

attach(dt)
```

Above we can see the variables associated with the dataset. “Kilocalories” is the target variable for this analysis and we will try to predict the value of kilocalories through other variables in the dataset. For that we will fit different models to the dataset and find the model which gives the best prediction.

```
pairs(dt, panel = panel.smooth)
```


Above in the plot, we can see the relationship between the variables in the dataset. The target variable kilocalories have positive linear regression with every variables other than cholesterol and fiber. One absolute strong positive corelation is present between Population and year indicating the constant rise in population. The consumption of fat by the population throughout the years too is constantly increasing.

```
cor(dt)

##
Year U.S. population, July 12
Kilocalories
## Year
1.00000000                0.99163900
0.70893950
## U.S. population, July 12
0.99163900                1.00000000
0.76905063
## Kilocalories
0.70893950                0.76905063
1.00000000
## Carbohydrate                    -
0.05336724                0.04189453
0.63376673
## Fiber                           -
0.45661885               -0.38256466
0.24838770
## Protein
0.83777164                0.87148623
0.91473206
```

```
## Fat
0.94972417               0.96129278
0.84597256
## Saturated Fatty Acids
0.76313189               0.74559801
0.71228244
## Monounsaturated Fatty Acids
0.93885371               0.95828518
0.87779681
## Polyunsaturated Fatty Acids
0.96310634               0.98253129
0.82738076
## Cholesterol
0.29469808               0.21385150
0.01541736
##
Carbohydrate        Fiber      Protein
Fat
## Year                          -
0.05336724 -0.456618849  0.837771639
0.9497242
## U.S. population, July 12
0.04189453 -0.382564661  0.871486227
0.9612928
## Kilocalories
0.63376673  0.248387696  0.914732058
0.8459726
## Carbohydrate
1.00000000  0.853922341  0.369852136
0.1372012
## Fiber
0.85392234  1.000000000 -0.002724935 -
0.2595543
## Protein
0.36985214 -0.002724935  1.000000000
0.9057630
## Fat
0.13720115 -0.259554254  0.905762995
1.0000000
## Saturated Fatty Acids
0.10117385 -0.183628693  0.713454911
0.8716169
## Monounsaturated Fatty Acids
0.20572260 -0.206496563  0.925271853
0.9924001
## Polyunsaturated Fatty Acids
0.12364245 -0.281498504  0.901700072
0.9807417
## Cholesterol                   -
0.41552779 -0.433044585  0.182842817
0.2969003
##                              Saturated
Fatty Acids Monounsaturated Fatty Acids
```

```
## Year
0.7631319                    0.9388537
## U.S. population, July 12
0.7455980                    0.9582852
## Kilocalories
0.7122824                    0.8777968
## Carbohydrate
0.1011738                    0.2057226
## Fiber
-0.1836287                   -0.2064966
## Protein
0.7134549                    0.9252719
## Fat
0.8716169                    0.9924001
## Saturated Fatty Acids
1.0000000                    0.8397072
## Monounsaturated Fatty Acids
0.8397072                    1.0000000
## Polyunsaturated Fatty Acids
0.7736701                    0.9750741
## Cholesterol
0.5150163                    0.2438962
##
Polyunsaturated Fatty Acids Cholesterol
## Year
0.9631063  0.29469808
## U.S. population, July 12
0.9825313  0.21385150
## Kilocalories
0.8273808  0.01541736
## Carbohydrate
0.1236425 -0.41552779
## Fiber
-0.2814985 -0.43304459
## Protein
0.9017001  0.18284282
## Fat
0.9807417  0.29690029
## Saturated Fatty Acids
0.7736701  0.51501630
## Monounsaturated Fatty Acids
0.9750741  0.24389620
## Polyunsaturated Fatty Acids
1.0000000  0.19342302
## Cholesterol
0.1934230  1.00000000
```

From above correlation values, we can confirm the relation as mentioned after graph. Kilocalories has moderately high correlation with year, population, carbohydrate and saturated fatty acid. It has strong correlation with polyunsaturated fatty acids,

monounsaturated fatty acids, fat and has the highest correlation of 0.9 with Protein.

- Training and Testing Sets

In order to conduct model building and model evaluation, the dataset is divided into training set and testing set. A ration of 70:30 is implied to build the sets. The training set will be used to build the model and the testing set will be used to evaluate the accuracy of the model.

```
set.seed(3)
trid = sample(1:nrow(dt), nrow(dt)*0.7)

trainset = dt[trid,]
testset = dt[-trid,]
```

- Multiple Linear Regression

```
m1 = glm(Kilocalories~., data = trainset)
summary(m1)

##
## Call:
## glm(formula = Kilocalories ~ ., data =
trainset)
##
## Coefficients:
##
Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -
427.4497  4274.0300  -0.100  0.92067
## Year
0.2349     2.3003   0.102  0.91901
## `U.S. population, July 12`      -
1.0178     1.3983  -0.728  0.46952
## Carbohydrate
4.4895     0.3740  12.003  < 2e-16 ***
## Fiber                          -
8.5084     5.2073  -1.634  0.10751
## Protein
4.1355     1.4117   2.929  0.00479 **
## Fat
11.9554     4.7628   2.510  0.01478 *
## `Saturated Fatty Acids`        -
6.8548     6.8052  -1.007  0.31784
## `Monounsaturated Fatty Acids`  -
1.8281     4.6545  -0.393  0.69589
## `Polyunsaturated Fatty Acids`
1.1612     6.3713   0.182  0.85600
## Cholesterol
0.1908     0.2825   0.676  0.50193
## ---
## Signif. codes:  0 '***' 0.001 '**'
0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian
family taken to be 871.7924)
##
##     Null deviance: 7253239  on 70
degrees of freedom
## Residual deviance:   52308  on 60
degrees of freedom
## AIC: 694.25
##
## Number of Fisher Scoring iterations: 2
```

In the first model (m1), plotting Kilocalories with every other variables on the testing set, we get carbohydrate, protein and fat as the significant variables. However, since the training set is randomly generated, it could affect the significance of other variables. The variables Year and US population would not contribute to predicting the kilocalories, so further models will not include these variables.

The second model (m2) will contain all the nutritional variables (i.e except year and population variables)

```
m2 =
glm(Kilocalories~Carbohydrate+Fiber+Prote
in+Fat+`Saturated Fatty
Acids`+`Monounsaturated Fatty
Acids`+`Polyunsaturated Fatty
Acids`+Cholesterol, data=trainset)
summary(m2)

##
## Call:
## glm(formula = Kilocalories ~
Carbohydrate + Fiber + Protein +
##     Fat + `Saturated Fatty Acids` +
`Monounsaturated Fatty Acids` +
##     `Polyunsaturated Fatty Acids` +
Cholesterol, data = trainset)
##
## Coefficients:
##                                Estimate
Std. Error t value Pr(>|t|)
## (Intercept)                   -99.7650
120.8995  -0.825 0.412428
## Carbohydrate                    4.4049
0.2925  15.060  < 2e-16 ***
## Fiber                          -2.7125
```

```
3.7036  -0.732 0.466689
## Protein                          3.0836
1.2529   2.461 0.016644 *
## Fat                             15.0038
4.2730   3.511 0.000838 ***
## `Saturated Fatty Acids`        -10.3832
5.9295  -1.751 0.084871 .
## `Monounsaturated Fatty Acids`   -4.9348
4.2631  -1.158 0.251478
## `Polyunsaturated Fatty Acids`   -6.1348
4.7085  -1.303 0.197415
## Cholesterol                      0.2929
0.2604   1.125 0.265025
## ---
## Signif. codes:  0 '***' 0.001 '**'
0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian
family taken to be 883.9339)
##
##     Null deviance: 7253239  on 70
degrees of freedom
## Residual deviance:   54804  on 62
degrees of freedom
## AIC: 693.56
##
## Number of Fisher Scoring iterations: 2
```

The third model will only contain the significant variables from above two models

```
m3 =
glm(Kilocalories~Carbohydrate+Protein+Fat
, data=trainset)
summary(m3)

##
## Call:
## glm(formula = Kilocalories ~
Carbohydrate + Protein + Fat, data =
trainset)
##
## Coefficients:
##              Estimate Std. Error t
value Pr(>|t|)
## (Intercept)   12.5556    42.4791
0.296    0.768
## Carbohydrate   3.9949     0.1146
34.856  < 2e-16 ***
## Protein        3.9655     0.8535
4.646 1.63e-05 ***
## Fat            8.6081     0.3924
21.938  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**'
0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian
family taken to be 873.7124)
##
##     Null deviance: 7253239  on 70
degrees of freedom
## Residual deviance:   58539  on 67
degrees of freedom
## AIC: 688.24
##
## Number of Fisher Scoring iterations: 2
```

The fourth model will contain all the variables that kilocalories has corelation with

```
m4 =
glm(Kilocalories~Carbohydrate+Protein+Fat
+`Saturated Fatty Acids`+`Monounsaturated
Fatty Acids`+`Polyunsaturated Fatty
Acids`, data=trainset)
summary(m4)

##
## Call:
## glm(formula = Kilocalories ~
Carbohydrate + Protein + Fat + `Saturated
Fatty Acids` +
##     `Monounsaturated Fatty Acids` +
`Polyunsaturated Fatty Acids`,
##     data = trainset)
##
## Coefficients:
##                                Estimate
Std. Error t value Pr(>|t|)
## (Intercept)                   -11.5229
102.3546  -0.113 0.910718
## Carbohydrate                    4.0959
0.1381  29.661  < 2e-16 ***
## Protein                         3.4588
1.0167   3.402 0.001158 **
## Fat                            14.7498
4.2654   3.458 0.000973 ***
## `Saturated Fatty Acids`        -8.4038
5.5464  -1.515 0.134649
## `Monounsaturated Fatty Acids`  -4.5146
4.1248  -1.095 0.277836
## `Polyunsaturated Fatty Acids`  -6.5586
4.6689  -1.405 0.164935
## ---
## Signif. codes:  0 '***' 0.001 '**'
0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for gaussian
family taken to be 882.9157)
##
##     Null deviance: 7253239  on 70
degrees of freedom
## Residual deviance:   56507  on 64
degrees of freedom
## AIC: 691.73
##
## Number of Fisher Scoring iterations: 2
```

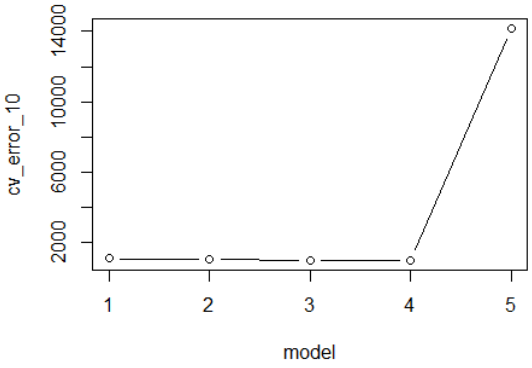
Using variables that have strong corelation values with kilocalories

```
m5 =
glm(Kilocalories~Protein+Fat+`Polyunsatur
ated Fatty Acids`+`Monounsaturated Fatty
Acids`+`Saturated Fatty Acids`,
data=trainset)
summary(m5)

##
## Call:
## glm(formula = Kilocalories ~ Protein +
Fat + `Polyunsaturated Fatty Acids` +
##     `Monounsaturated Fatty Acids` +
`Saturated Fatty Acids`,
##     data = trainset)
##
## Coefficients:
##                                Estimate
Std. Error t value Pr(>|t|)
## (Intercept)                     890.967
372.382   2.393 0.019628 *
## Protein                          21.727
3.082   7.049 1.42e-09 ***
## Fat                             -51.482
13.848  -3.718 0.000421 ***
## `Polyunsaturated Fatty Acids`    43.631
16.581   2.631 0.010610 *
## `Monounsaturated Fatty Acids`    52.731
13.891   3.796 0.000325 ***
## `Saturated Fatty Acids`          70.625
18.536   3.810 0.000311 ***
## ---
## Signif. codes:  0 '***' 0.001 '**'
0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian
family taken to be 12819.25)
##
##     Null deviance: 7253239  on 70
degrees of freedom
## Residual deviance:  833251  on 65
degrees of freedom
## AIC: 880.79
##
## Number of Fisher Scoring iterations: 2
```

- 10-fold cross-validation method to select the best model

From all above regression model built, we need to find the best model to predict the kilocalories values. 10 fold cross validation method is used to choose the best model

```
library(boot)

m = list(m1,m2,m3,m4,m5)
model = 1:length(m)
cv_error_10 = rep(0,length(m))

for (i in 1:length(m)) {
    error_10 = cv.glm(trainset, m[[i]],
K=10)$delta[1]
    cv_error_10[i] = error_10
}
cv_error_10

## [1]  1090.1590  1049.6255   989.4071
1010.5064 14169.5360

plot(model, cv_error_10, type = "b")
```




From above obtained graph and error values, the difference in error values between m1, m2, m3 and m4 is not that high.

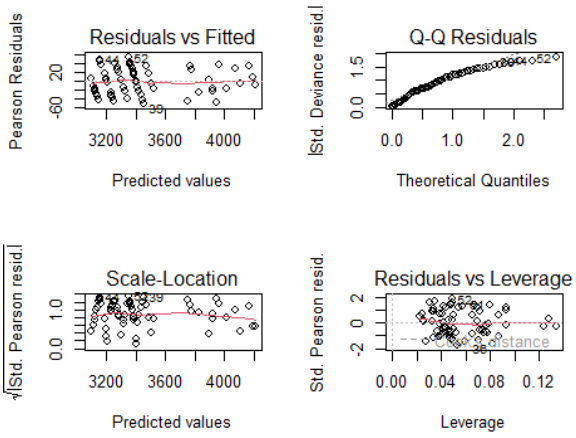
```
summary(m1)

##
## Call:
## glm(formula = Kilocalories ~ ., data =
```

```
trainset)
##
## Coefficients:
##
Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -
427.4497  4274.0300  -0.100  0.92067
## Year
0.2349     2.3003   0.102  0.91901
## `U.S. population, July 12`      -
1.0178     1.3983  -0.728  0.46952
## Carbohydrate
4.4895     0.3740  12.003  < 2e-16 ***
## Fiber                           -
8.5084     5.2073  -1.634  0.10751
## Protein
4.1355     1.4117   2.929  0.00479 **
## Fat
11.9554     4.7628   2.510  0.01478 *
## `Saturated Fatty Acids`         -
6.8548     6.8052  -1.007  0.31784
## `Monounsaturated Fatty Acids`   -
1.8281     4.6545  -0.393  0.69589
## `Polyunsaturated Fatty Acids`
1.1612     6.3713   0.182  0.85600
## Cholesterol
0.1908     0.2825   0.676  0.50193
## ---
## Signif. codes:  0 '***' 0.001 '**'
0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian
family taken to be 871.7924)
##
##     Null deviance: 7253239  on 70
degrees of freedom
## Residual deviance:   52308  on 60
degrees of freedom
## AIC: 694.25
##
## Number of Fisher Scoring iterations: 2

summary(m2)

##
## Call:
## glm(formula = Kilocalories ~
Carbohydrate + Fiber + Protein +
##     Fat + `Saturated Fatty Acids` +
`Monounsaturated Fatty Acids` +
##     `Polyunsaturated Fatty Acids` +
Cholesterol, data = trainset)
##
## Coefficients:
##                                Estimate
Std. Error t value Pr(>|t|)
## (Intercept)                    -99.7650
120.8995  -0.825 0.412428
## Carbohydrate                     4.4049
0.2925  15.060  < 2e-16 ***
## Fiber                           -2.7125
3.7036  -0.732 0.466689
## Protein                          3.0836
1.2529   2.461 0.016644 *
## Fat                             15.0038
4.2730   3.511 0.000838 ***
## `Saturated Fatty Acids`        -10.3832
5.9295  -1.751 0.084871 .
## `Monounsaturated Fatty Acids`   -4.9348
4.2631  -1.158 0.251478
## `Polyunsaturated Fatty Acids`   -6.1348
4.7085  -1.303 0.197415
## Cholesterol                      0.2929
0.2604   1.125 0.265025
## ---
## Signif. codes:  0 '***' 0.001 '**'
0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian
family taken to be 883.9339)
##
##     Null deviance: 7253239  on 70
degrees of freedom
## Residual deviance:   54804  on 62
degrees of freedom
## AIC: 693.56
##
## Number of Fisher Scoring iterations: 2

summary(m3)

##
## Call:
## glm(formula = Kilocalories ~
Carbohydrate + Protein + Fat, data =
trainset)
##
## Coefficients:
##              Estimate Std. Error t
value Pr(>|t|)
## (Intercept)   12.5556    42.4791
0.296    0.768
## Carbohydrate   3.9949     0.1146
34.856  < 2e-16 ***
## Protein        3.9655     0.8535
4.646 1.63e-05 ***
## Fat            8.6081     0.3924
21.938  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**'
0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian
family taken to be 873.7124)
##
##     Null deviance: 7253239  on 70
degrees of freedom
## Residual deviance:   58539  on 67
degrees of freedom
## AIC: 688.24
##
## Number of Fisher Scoring iterations: 2

summary(m4)

##
## Call:
## glm(formula = Kilocalories ~
Carbohydrate + Protein + Fat + `Saturated
Fatty Acids` +
##     `Monounsaturated Fatty Acids` +
`Polyunsaturated Fatty Acids`,
##     data = trainset)
##
## Coefficients:
##                                Estimate
Std. Error t value Pr(>|t|)
## (Intercept)                   -11.5229
102.3546  -0.113 0.910718
## Carbohydrate                    4.0959
0.1381  29.661  < 2e-16 ***
## Protein                         3.4588
1.0167   3.402 0.001158 **
## Fat                            14.7498
4.2654   3.458 0.000973 ***
## `Saturated Fatty Acids`        -8.4038
5.5464  -1.515 0.134649
## `Monounsaturated Fatty Acids`  -4.5146
4.1248  -1.095 0.277836
## `Polyunsaturated Fatty Acids`  -6.5586
4.6689  -1.405 0.164935
## ---
## Signif. codes:  0 '***' 0.001 '**'
0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian
family taken to be 882.9157)
##
##     Null deviance: 7253239  on 70
degrees of freedom
## Residual deviance:   56507  on 64
degrees of freedom
## AIC: 691.73
##
## Number of Fisher Scoring iterations: 2
```

However, from above model summaries we can see that m3 has the least number of independent variables and all those variables are significant. Also, m3 has the least AIC score among other models. Through these information, we conclude model m3 as the best model.

```
par(mfrow=c(2,2))
plot(m3)
```




In order to test the accuracy of our selected model, we calculate the Mean Squared Error (MSE), which is the mean of the squared distance between the actual value and the value predicted by our model.

```
actual = testset$Kilocalories

predicted = predict(m3, testset)

MSE = mean((actual-predicted)^2)
MSE

## [1] 864.5705

sqrt(MSE)

## [1] 29.40358
```

## Appendix B (Decision Tree)

In order to imply the decision tree model building, we install the "tree" library.

```
library(tree)
dt_new = dt
```

The decision tree obtained will contain the significant variables in the plot. But the variables in the dataset has lengthy names, so renaming some variables. "dt_new" is a new dataset which contains the old dataset with renamed variables.

```
colnames(dt_new)[2] ="Population"
colnames(dt_new)[8] ="SFA"
colnames(dt_new)[9] ="MFA"
colnames(dt_new)[10] ="PFA"
names(dt_new)

##  [1] "Year"         "Population"
"Kilocalories" "Carbohydrate" "Fiber"
##  [6] "Protein"      "Fat"
"SFA"          "MFA"          "PFA"
## [11] "Cholesterol"
```

The target variable "Kilocalories" is numeric. Hence, we are going to fit a regression tree model. But first, dividing the dataset into training and testing set in a ratio of 70:30 respectively.

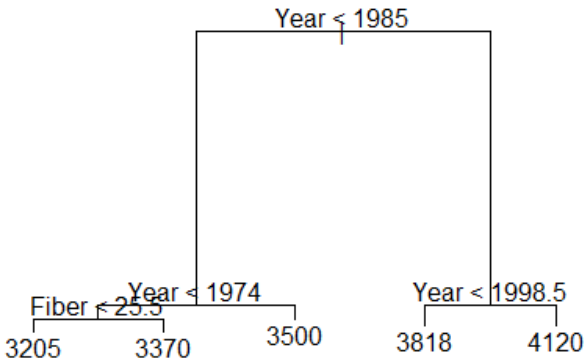
```
set.seed(2)
tid = sample(1:nrow(dt_new),
nrow(dt_new)*0.7)

training = dt_new[tid,]
testing = dt_new[-tid,]
```

Fitting a regression tree for the training data set

```
reg_tree1 = tree(Kilocalories~.,
data=training)
summary(reg_tree1)

##
## Regression tree:
## tree(formula = Kilocalories ~ ., data
= training)
## Variables actually used in tree
construction:
## [1] "Year"  "Fiber"
## Number of terminal nodes:  5
## Residual mean deviance:  5514 = 363900
/ 66
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean
3rd Qu.     Max.
## -170.000  -20.000   -4.545    0.000
30.000  195.500
```

From the summary, we can see that only "Year" and "Fiber" are the variables that has been considered by the performed regression tree. As a result, the resulting tree has only 5 terminal nodes. We can see the tree in the plot below:

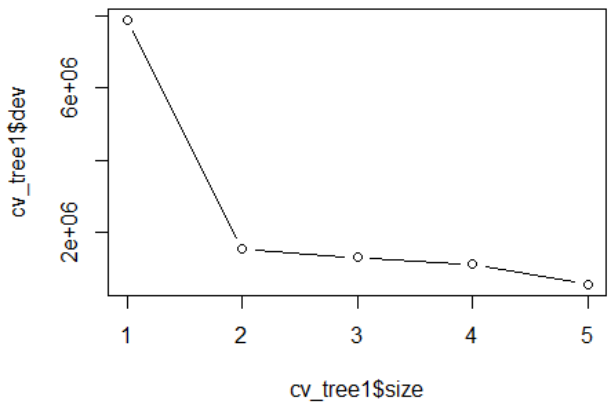
```
plot(reg_tree1)
text(reg_tree1, pretty = 0)
```




We need to further validate the best size of the tree. For that, we perform cross validation plot of the obtained regression tree.

```
set.seed(1)
cv_tree1 = cv.tree(reg_tree1)
cv_tree1

## $size
## [1] 5 4 3 2 1
##
## $dev
## [1]  567156.1 1121559.5 1311817.8
1547004.0 7879278.5
##
## $k
## [1]      -Inf  286787.9  315466.7
477160.2 6131605.8
```

```
##
## $method
## [1] "deviance"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"

plot(cv_tree1$size, cv_tree1$dev, type =
"b")
```




In the cross validation plot above, we can see that the error for the tree size is constantly dropping till the last size that is 5. Hence, the best size for the tree model is 5, which is the same as the regression tree obtained above. Therefore the tree obtained above is the best model and there is no need to prune it further.

Now we test the accuracy of the regression tree by predicting the target variable "Kilocalories" using the tree model for testing data set

```
p = predict(reg_tree1, testing)
kcl = testing$Kilocalories # actual
values
ms = mean((p-kcl)^2)
ms

## [1] 9630.472

sqrt(ms)

## [1] 98.13497
```

Describing the terminal nodes of the resulting decision tree

- Year is the most important predictor in determining the Kilocalories and fiber is the other significant variable.
- For the year before 1974 and when the fiber intake was less than 25, the average calorie consumption was 3205, which is the least consumption as shown in the tree
- For the year after 1998, the average calorie consumption was 4120, which is the highest consumption as shown in the tree.

Also, the tree shows that through the years the average consumption of calories per capita has been increasing.

```
attach(dt)

## The following objects are masked from
dt (pos = 5):
##
##     Carbohydrate, Cholesterol, Fat,
Fiber, Kilocalories,
##     Monounsaturated Fatty Acids,
Polyunsaturated Fatty Acids, Protein,
##     Saturated Fatty Acids, U.S.
population, July 12, Year

str(dt)

## tibble [102 × 11] (S3:
tbl_df/tbl/data.frame)
##  $ Year                         : int
[1:102] 1909 1910 1911 1912 1913 1914
1915 1916 1917 1918 ...
##  $ U.S. population, July 12    : num
[1:102] 90.5 92.4 93.9 95.3 97.2 ...
##  $ Kilocalories                 : int
[1:102] 3400 3400 3400 3400 3400 3400
3300 3300 3200 3300 ...
##  $ Carbohydrate                 : int
[1:102] 499 498 492 493 492 486 483 473
472 468 ...
##  $ Fiber                        : int
[1:102] 29 29 28 28 28 27 28 27 28 27 ...
##  $ Protein                      : int
[1:102] 101 99 98 98 97 95 94 93 92 94
...
##  $ Fat                          : int
[1:102] 119 117 118 115 116 118 117 117
113 121 ...
##  $ Saturated Fatty Acids       : int
[1:102] 50 49 49 48 48 48 48 48 46 49 ...
##  $ Monounsaturated Fatty Acids: int
[1:102] 45 44 45 44 44 45 45 45 44 47 ...
##  $ Polyunsaturated Fatty Acids: int
[1:102] 13 12 12 12 12 13 13 13 12 14 ...
##  $ Cholesterol                  : int
[1:102] 440 440 460 440 430 430 430 430
410 420 ...
```

In order to do an unsupervised learning, we have to remove all the categorical variables as well as the target variables. For our data set, all the variables are numeric however we have to remove the target variable

```
dt_us = dt[,-3]
```

Examining the mean and variance of the variables in the data set

```
sapply(dt_us, mean)

##                          Year    U.S.
population, July 12
##                   1959.50000
185.24797
##                 Carbohydrate
Fiber
##                    444.80392
23.40196
##                      Protein
Fat
##                    102.88235
147.66667
##       Saturated Fatty Acids
Monounsaturated Fatty Acids
##                     54.71569
58.46078
## Polyunsaturated Fatty Acids
Cholesterol
##                     23.21569
469.80392

sapply(dt_us, var)

##                          Year    U.S.
population, July 12
##                   875.500000
4277.631746
##                 Carbohydrate
Fiber
##                  1592.753252
8.836828
##                      Protein
Fat
##                   154.718695
564.382838
##       Saturated Fatty Acids
Monounsaturated Fatty Acids
##                    14.007474
146.686566
## Polyunsaturated Fatty Acids
Cholesterol
##                    83.616385
621.743351
```

It can be seen that the variables have vastly different mean and variance. This is due to the fact that the variables measure completely different things. Variable like year measure the years through the century and US Population

measure the population throughout those years. The remaining variables measure the average consumption of different micro-nutrients per capita.

We perform the Principal Component Analysis for the dataset by scaling the variables so that all the variables will have equal contibution despite of their range in data.

```
pr = prcomp(dt_us, scale=TRUE)
summary(pr)

## Importance of components:
##                             PC1    PC2
PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.5683 1.4860
0.91873 0.45450 0.25714 0.19864 0.16283
## Proportion of Variance 0.6596 0.2208
0.08441 0.02066 0.00661 0.00395 0.00265
## Cumulative Proportion  0.6596 0.8804
0.96483 0.98549 0.99210 0.99605 0.99870
##                             PC8     PC9
PC10
## Standard deviation     0.10666 0.03204
0.02504
## Proportion of Variance 0.00114 0.00010
0.00006
## Cumulative Proportion  0.99983 0.99994
1.00000
```

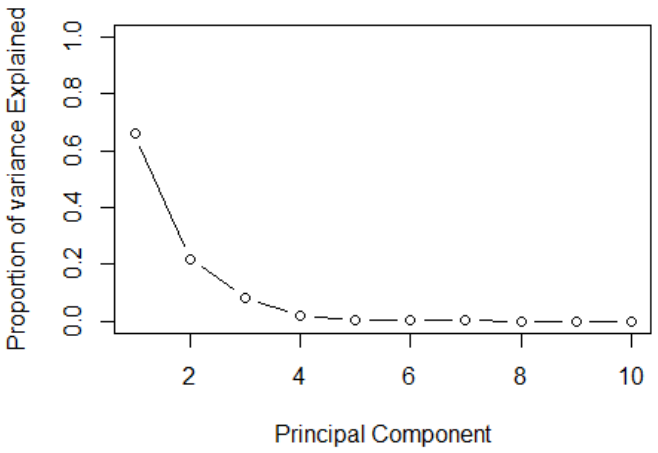
It can clearly be seen that there are 10 diffrent principal components.

```
pr$rotation

##
PC1         PC2          PC3         PC4
## Year
0.37822095 -0.07748650 -0.179767896 -
0.06524857
## U.S. population, July 12
0.37947131 -0.00989092 -0.219157292 -
0.08587808
## Carbohydrate
0.03155192  0.65048593  0.171384367 -
0.06855953
## Fiber                        -
0.12482160  0.59860850  0.308915119 -
0.01339562
## Protein
0.35239087  0.20927301  0.029048145 -
0.53054933
## Fat
0.38703840  0.04304441  0.003135458
0.12910462
## Saturated Fatty Acids
0.33463127 -0.01077906  0.418736488
0.72364229
## Monounsaturated Fatty Acids
0.38357719  0.08952529 -0.023048613
0.04568386
## Polyunsaturated Fatty Acids
0.38105732  0.05050703 -0.162355264 -
0.02980354
## Cholesterol
0.13447009 -0.39509997  0.769264643 -
0.39837634
##
PC5         PC6         PC7           PC8
## Year                        -
0.15412153 -0.32820310  0.56689407 -
0.1726372698
## U.S. population, July 12    -
0.20928852 -0.26057169  0.25222377 -
0.0065695180
## Carbohydrate                -
0.69400671  0.17317944 -0.01918429
0.1476562551
## Fiber
0.45473917 -0.53166669  0.09036394 -
0.1672101619
## Protein
0.42790022  0.54416478  0.22528490
0.1346456135
## Fat
0.08720448  0.00170966 -0.36038305 -
0.0065731571
## Saturated Fatty Acids
0.10788420  0.22079876  0.27101329
0.1641394951
## Monounsaturated Fatty Acids -
0.03391230  0.10949116 -0.39963952 -
0.7304764967
## Polyunsaturated Fatty Acids
0.08698621 -0.34647720 -0.43063807
0.5845344177
## Cholesterol                 -
0.18152954 -0.18659969 -0.08510160 -
0.0008873984
##
PC9          PC10
## Year                        -
0.568720502 -0.0910199962
## U.S. population, July 12
0.790452431  0.0180112285
## Carbohydrate                -
```

```
0.085672117 -0.0255533953
## Fiber
0.066982648 -0.0202654292
## Protein
0.004866153  0.0397339037
## Fat                          -
0.012796590 -0.8330547277
## Saturated Fatty Acids
0.046501897  0.1611329507
## Monounsaturated Fatty Acids -
0.044222317  0.3658331329
## Polyunsaturated Fatty Acids -
0.187839597  0.3673832876
## Cholesterol
0.018065764  0.0003752826
```

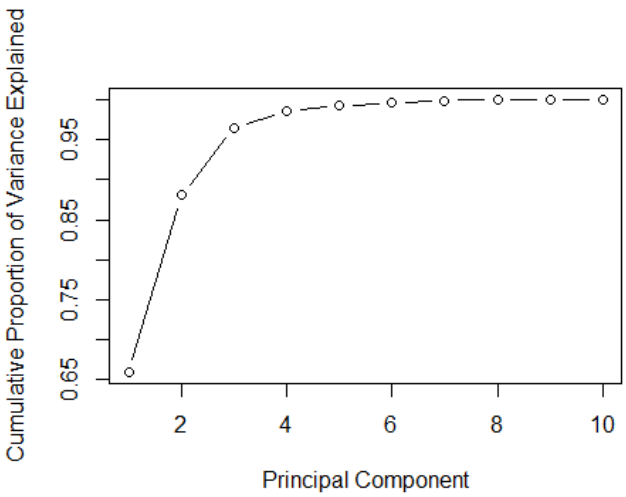
We have to calculate the variance captured by each PC.

```
prvar = pr$sdev^2

prvar # will give variance captured by
each pc

##  [1] 6.5960170647 2.2082329055
0.8440591372 0.2065679143 0.0661220089
##  [6] 0.0394569796 0.0265129906
0.0113773913 0.0010264610 0.0006271468

prve = prvar/sum(prvar)
prve # will give the proportion of
variance captured by each pc

##  [1] 6.596017e-01 2.208233e-01
8.440591e-02 2.065679e-02 6.612201e-03
##  [6] 3.945698e-03 2.651299e-03
1.137739e-03 1.026461e-04 6.271468e-05

plot(1:10, prve, type = "b", xlab =
"Principal Component", ylab="Proportion
of variance Explained", ylim = c(0,1))
```




It can be seen that the first principal component explains about 66% of the variance in the data, the next principal component explains about 22% of the variance and so forth.

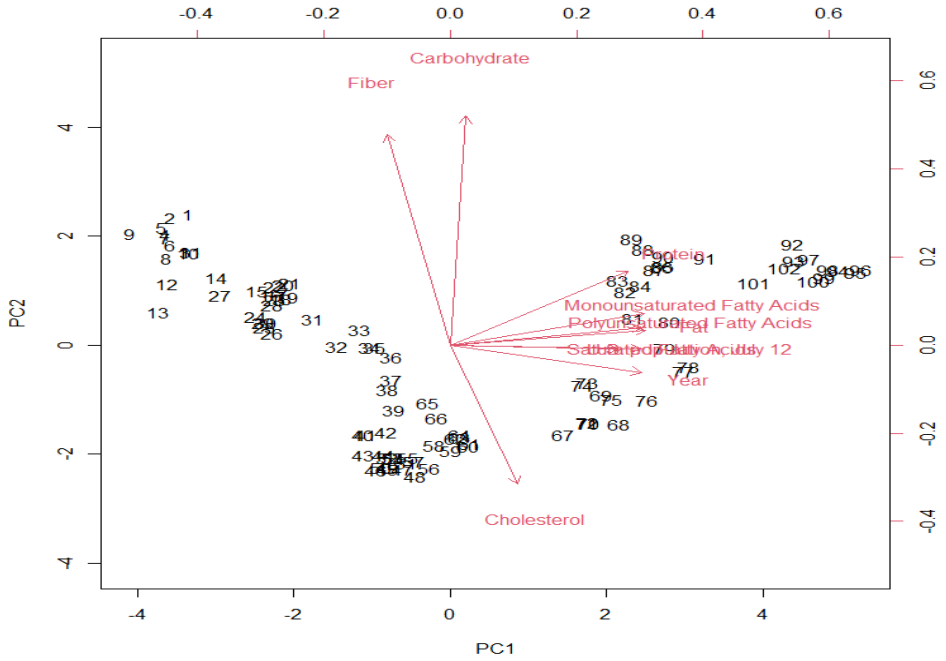
```
cumsum(prve)

##  [1] 0.6596017 0.8804250 0.9648309
0.9854877 0.9920999 0.9960456 0.9986969
##  [8] 0.9998346 0.9999373 1.0000000

plot(1:10, cumsum(prve), type = "b", xlab
= "Principal Component", ylab="Cumulative
Proportion of Variance Explained")
```




It can be seen clearly that the first PC explains about 66% variation in the data set. The first and second PCs together explain about 88% of the variation in the data set. The first three PCs

together explain about 96% of the variation in the data set and so forth.

```
biplot(pr, scale = 0)
```




We can see that the first loading vector (PC1) places approximately equal weight on fat, protein, different fatty acids, population and year, with much less weight on fiber, carbohydrate and cholesterol.

The second loading vector (PC2) places most of its weight on fiber, carbohydrate and cholesterol and much less weight on the other features.

We can see that with the increase in year the population has grown and along with that the average consumption of protein and fats too have increased.

Interesting, Fats and protein are closely located (indicating that they are correlated). It is observed that people who consume more protein are likely to be consuming more fat as well.

Cholesterol and Fiber are at 180-degree angle from each other indicating no correlation at all between them, which means that eating food containing high fiber contribute to no cholesterol intake.