Foundation of Statistical Inference

Question 1:

Each student in a large statistics class of 600 students is asked to toss a fair coin 100 times, count the resulting number of Heads, and construct a 0.95-level confidence interval for the probability of Heads. Assume that each student uses a fair coin and constructs the confidence interval correctly. True or False: We would expect approximately 570 of the confidence intervals to contain the number 0.5.

The question asking whether it’s true that approximately 570 of the confidence intervals contain the number 0.5 means that we are looking for 0.95% confidence interval (570/600=0.95) containing the number 0.5, or estimated probability for a fair coin. Since the focus of the test is to check whether the coin is fair or not, we can create a simulator to conduct a Bernoulli trial to test the following hypothesis:

Ho: The coin is fair (p=0.5)

H1: The coin is not fair (p≠0.5)

In order to replicate the trial conducted by students, we need to perform total 60,000 tosses of a fair coin (600 students *100 tosses each) and compare resulting probability of successes (heads) with the expected probability of 0.5 with 95% confidence interval.

```
> #total number of tosses
> tosses = 600*100
> tosses
[1] 60000
```

I used rbinom() function to generate a binomial distribution for 100 (number of observations for each student) fair (p=0.5) coin tosses performed 600 times (number of students) using the following code:

```
> #Use rbinom() function to generate binomial distribution
> #100 - number of observations, 600 - number of trials(students), p =0.5 probability for a fair
coin
> set.seed(123)
> results1 <-rbinom(600, 100, 0.5)
```

Below are the results of this simulation presented as a table indicating how many successes (heads) were received by the corresponding number of students in the simulation:

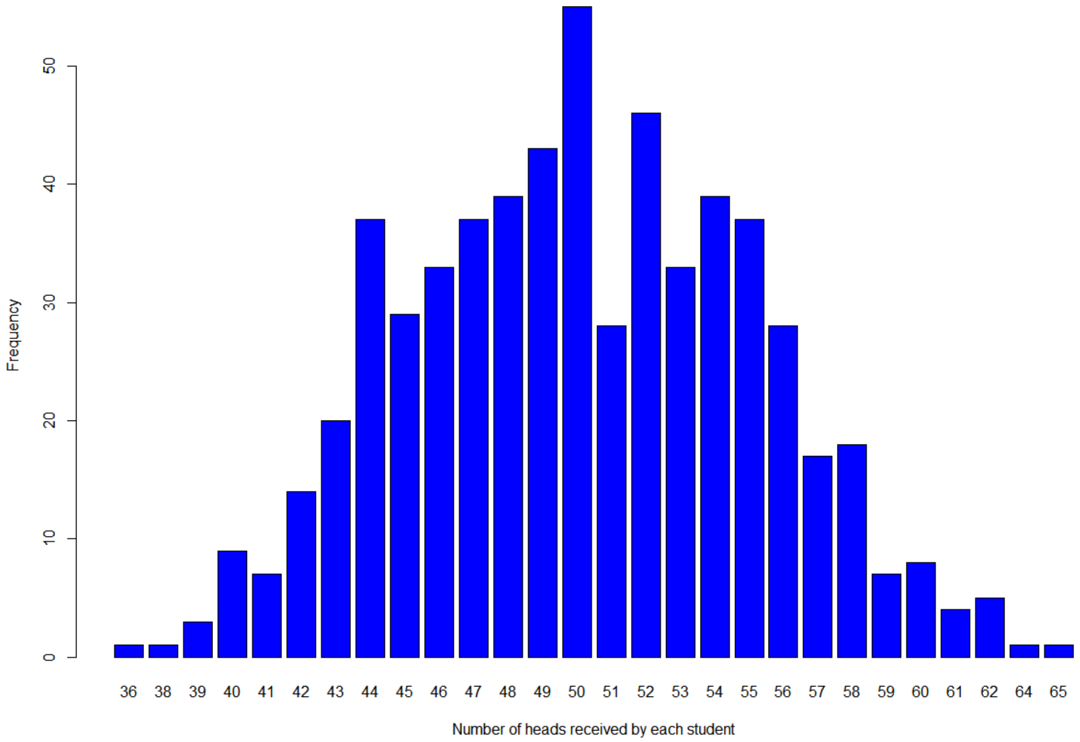
```
> #Simulation results
> table(results1) #frequency distribution for the number of heads received by each student in the
simulation
results1
36 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 64 65 
 1  1  3  9  7 14 20 37 29 33 37 39 43 55 28 46 33 39 37 28 17 18  7  8  4  5  1  1 
```

The mean number of heads was equal to 50.12 with the standard deviation of 5.036985:

```
> mean(results1) #mean number of heads
[1] 50.12
> sd(results1) #standard deviation
[1] 5.036985
```

The following plot illustrates the results:

```
> #Plot of the frequency distribution
> barplot(table(results1), xlab = 'Number of heads received by each student',
ylab = 'Frequency' ,col = 'blue')
```



As the histogram shows, the distribution of the successes overall follows a bell shape, but it is not a completely normal distribution.

The total number of successes (heads) received in 60000 simulated trials was 30072.

```
> #total number od successes (heads) in the simulated trial
> heads <- sum(results1)
> heads
[1] 30072
```

Next, I used binom.test() function to test the previously stated hypothesis about the fairness of the coin using data provided by the simulated trial:

```
>  #Test hypothesis using simulation results
> #Ho: coin is fair (probability=0.5)
> #Ha: coin is not fair (probability != 0.5)
> binom.test(heads, tosses, p=0.5, alternative = "two.sided", conf.level = 0.95)

        Exact binomial test

data:  heads and tosses
number of successes = 30072, number of trials = 60000, p-value = 0.5594
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4971909 0.5052089
sample estimates:
probability of success 
                0.5012
```

As the output above shows, the resulting p-value for the two-sided test for the simulated trial is equal to 0.5594, which is well above the significance level of 0.05. It means that we fail to reject the null hypothesis, stating that the coin in the experiment was fair (with the probability of success = 0.5).

Question 2.

A company that manufactures light bulbs has advertised that its 75- watt bulbs burn an average of 800 hours before failing. In reaction to the company’s advertising campaign, several dissatisfied customers have complained to a consumer watchdog organization that they believe the company’s claim to be exaggerated. The consumer organization must decide whether or not to allocate some of its financial resources to countering the company’s advertising campaign. So that it can make an informed decision, it begins by purchasing and testing 100 of the disputed light bulbs. In this experiment, the 100 light bulbs burned an average of $\bar{x}$= 745.1 hours before failing, with a sample standard deviation of s = 238.0 hours. Formulate null and alternative hypotheses that are appropriate for this situation. Calculate a significance probability. Do these results warrant rejecting the null hypothesis at a significance level of α = 0.05?

In order to estimate the validity of the claims made the bulb manufacturer, we need to establish if the mean value of 745.1 hours for the sample is statistically different from the mean value of 800 hours, claimed by the bulbs’ manufacture.

So, the hypothesises are as follow:

H0: There is no statistical difference between the estimated population mean and the sample mean;

Ha: There is a statistical difference between the population mean and the sample mean.

In order to test these hypotheses, we need to calculate the critical value of the test statistic at the 0.05 significance level.

Because the population variance is unknown, I used the following code to calculate the test statistic:

```
> #input data for the experiment
> sample_mean <- 745.1 #mean number of hours
> n <- 100 #sample size
> s <- 238.0 #sample standard deviation
>
> mu <- 800.0 #estimated population
> teststat <- (sample_mean - mu)/(s * 1/sqrt(n))
> teststat
[1] -2.306723
```

The test statistic is equal to -2.306723. Next step is to calculate the significance probability at 0.05 significance level. Since consumers are most likely concerned with the bulbs burning for a shorter period of time than the advertised mean, I used a one-tailed p-value:

```
> #Calculate significance probability at 0.05 significance level
> p_value <- pnorm(teststat, lower.tail = TRUE)
> p_value
[1] 0.01053514
```

The above output shows that the p_value or probability of population mean light bulbs lifespan being equal to 800 hours given the sample mean for them 100 light bulbs tested is 0.01053514. It is less than the significance level of 0.05. It means that we have to reject the null hypothesis in favor of the alternative hypothesis. According to the test results, the manufacturer’s claims are not valid.

Question 3.

http://lib.stat.cmu.edu/datasets/1993.expo/cereal # cereal data set

http://lib.stat.cmu.edu/datasets/1993.expo/ # description

Analyze the cereal data set. Write a report of your findings, including commands, corresponding plots/tables, and interpretations.

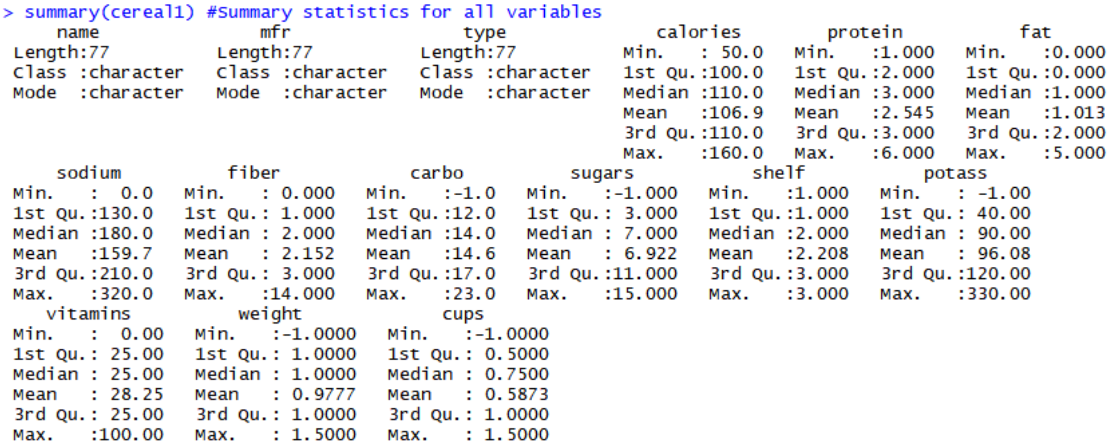
For ideas, see sample questions on the cereal description page or create your own. Your group should answer at least 5 questions (among these, you should include at least 2 challenging problems).

First, I loaded the dataset and added the column names:

```
> #Load Data into a table
> cereal1 <- read.table("cereal.csv", header = FALSE, sep=" ", quote="", stringsAsFactors = FALSE
)
> cereal1
> #Add column names
> names(cereal1) <- c("name", "mfr", "type", "calories", "protein", "fat", "sodium", "fiber", "ca
rbo", "sugars", "shelf", "potass", "vitamins", "weight", "cups")
> str(cereal1)
'data.frame':	77 obs. of  15 variables:
 $ name    : chr  "100%_Bran" "100%_Natural_Bran" "All-Bran" "All-Bran_with_Extra_Fiber" ...
 $ mfr     : chr  "N" "Q" "K" "K" ...
 $ type    : chr  "C" "C" "C" "C" ...
 $ calories: int  70 120 70 50 110 110 110 130 90 90 ...
 $ protein : int  4 3 4 4 2 2 2 3 2 3 ...
 $ fat     : int  1 5 1 0 2 2 0 2 1 0 ...
 $ sodium  : int  130 15 260 140 200 180 125 210 200 210 ...
 $ fiber   : num  10 2 9 14 1 1.5 1 2 4 5 ...
 $ carbo   : num  5 8 7 8 14 10.5 11 18 15 13 ...
 $ sugars  : int  6 8 5 0 8 10 14 8 6 5 ...
 $ shelf   : int  3 3 3 3 3 1 2 3 1 3 ...
 $ potass  : int  280 135 320 330 -1 70 30 100 125 190 ...
 $ vitamins: int  25 0 25 25 25 25 25 25 25 25 ...
 $ weight  : num  1 1 1 1 1 1 1 1.33 1 1 ...
 $ cups    : num  0.33 -1 0.33 0.5 0.75 0.75 1 0.75 0.67 0.67 ...
```

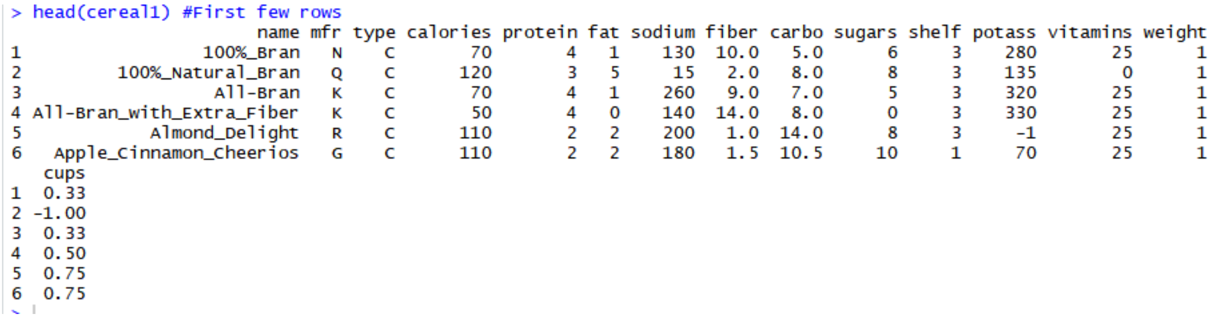
The cereal dataset contains 77 observations of 15 variables describing various breakfast cereals. In order to come up with questions, I performed some exploratory data analysis.

I checked the summary statistics for all variables:



```
> summary(cereal1) #Summary statistics for all variables
     name                mfr                type              calories        protein           fat       
 Length:77          Length:77          Length:77          Min.   : 50.0   Min.   :1.000   Min.   :0.000  
 Class :character   Class :character   Class :character   1st Qu.:100.0   1st Qu.:2.000   1st Qu.:0.000  
 Mode  :character   Mode  :character   Mode  :character   Median :110.0   Median :3.000   Median :1.000  
                                                          Mean   :106.9   Mean   :2.545   Mean   :1.013  
                                                          3rd Qu.:110.0   3rd Qu.:3.000   3rd Qu.:2.000  
                                                          Max.   :160.0   Max.   :6.000   Max.   :5.000  
     sodium          fiber            carbo          sugars           shelf           potass      
 Min.   :  0.0   Min.   : 0.000   Min.   :-1.0   Min.   :-1.000   Min.   :1.000   Min.   : -1.00  
 1st Qu.:130.0   1st Qu.: 1.000   1st Qu.:12.0   1st Qu.: 3.000   1st Qu.:1.000   1st Qu.: 40.00  
 Median :180.0   Median : 2.000   Median :14.0   Median : 7.000   Median :2.000   Median : 90.00  
 Mean   :159.7   Mean   : 2.152   Mean   :14.6   Mean   : 6.922   Mean   :2.208   Mean   : 96.08  
 3rd Qu.:210.0   3rd Qu.: 3.000   3rd Qu.:17.0   3rd Qu.:11.000   3rd Qu.:3.000   3rd Qu.:120.00  
 Max.   :320.0   Max.   :14.000   Max.   :23.0   Max.   :15.000   Max.   :3.000   Max.   :330.00  
    vitamins          weight            cups        
 Min.   :  0.00   Min.   :-1.0000   Min.   :-1.0000  
 1st Qu.: 25.00   1st Qu.: 1.0000   1st Qu.: 0.5000  
 Median : 25.00   Median : 1.0000   Median : 0.7500  
 Mean   : 28.25   Mean   : 0.9777   Mean   : 0.5873  
 3rd Qu.: 25.00   3rd Qu.: 1.0000   3rd Qu.: 1.0000  
 Max.   :100.00   Max.   : 1.5000   Max.   : 1.5000  
```

Looked at the first few rows of data:

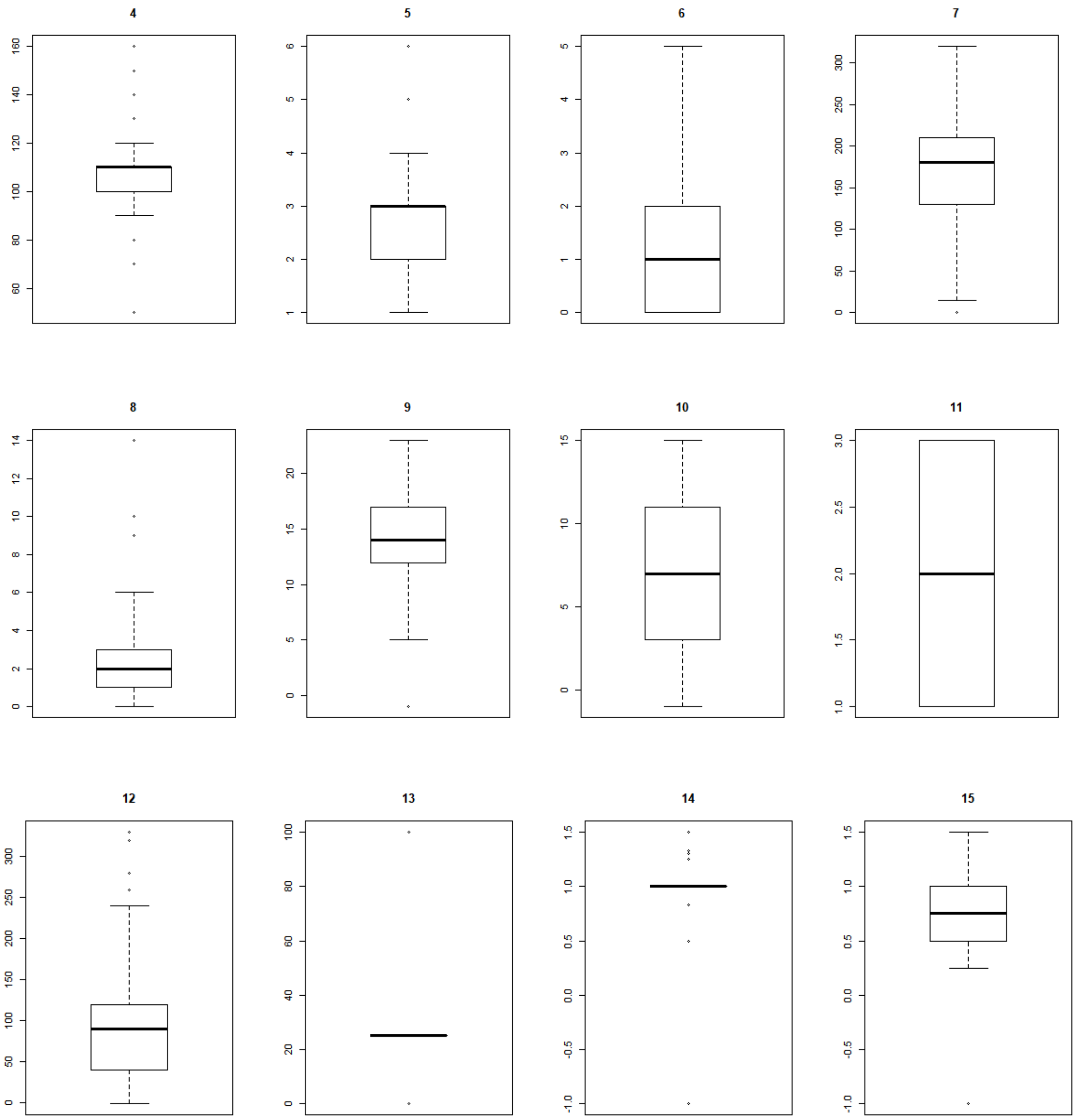


```
> head(cereal1) #First few rows
                       name mfr type calories protein fat sodium fiber carbo sugars shelf potass vitamins weight
1                 100%_Bran   N    C       70       4   1    130  10.0   5.0      6     3    280       25      1
2         100%_Natural_Bran   Q    C      120       3   5     15   2.0   8.0      8     3    135        0      1
3                  All-Bran   K    C       70       4   1    260   9.0   7.0      5     3    320       25      1
4 All-Bran_with_Extra_Fiber   K    C       50       4   0    140  14.0   8.0      0     3    330       25      1
5            Almond_Delight   R    C      110       2   2    200   1.0  14.0      8     3     -1       25      1
6   Apple_Cinnamon_Cheerios   G    C      110       2   2    180   1.5  10.5     10     1     70       25      1
   cups
1  0.33
2 -1.00
3  0.33
4  0.50
5  0.75
6  0.75
```

```
> sum(is.na(cereal1)) #Missing values?
[1] 0
```

While there was no missing data in the data set(see above), the summary statistics and the boxplots below suggested that there are values = -1 for the columns in which we would expect only positive numbers.

```
> #Boxplots for all numerical variables in the dataset
> par(mfrow=c(2,4))
> for (i in 4:15){
+     boxplot(cereal1[i], main=i)
+     }
```

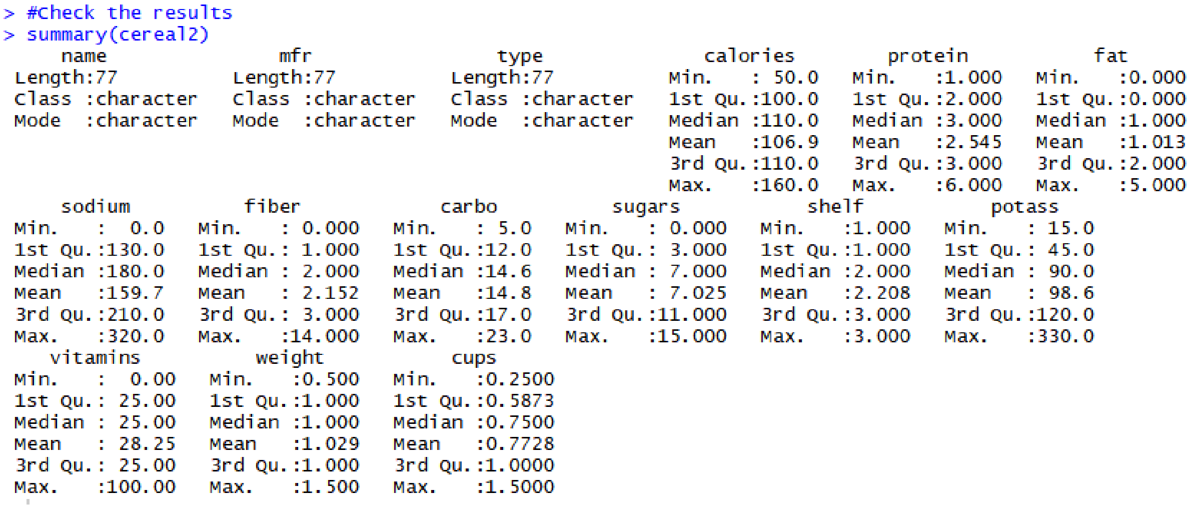


It is likely that the -1 values were intended as indicators of missing data. There were no other abnormalities in the data.

Due to the size of the data set and the number of values=-1, I decided not to discard of the incomplete record, and impute unknown values with the mean values for the appropriate columns using the following:

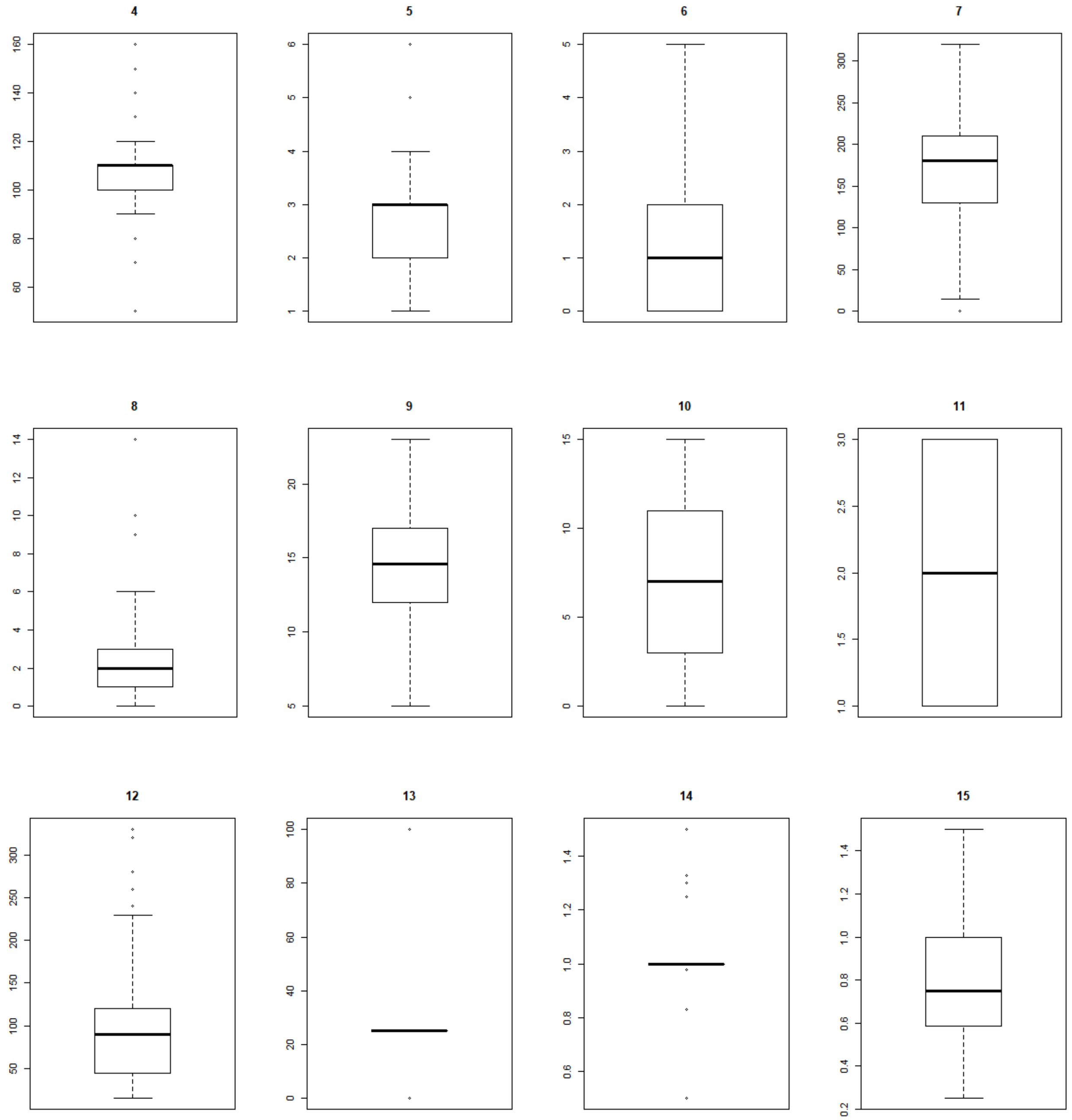
```
> #For impute values = -1  (N/A) with mean values for the carbo, sugars, potass, weight and cups
variables
> cereal2 <- cereal1
> cereal2$carbo <- ifelse(cereal2$carbo == -1.00, mean(cereal2$carbo), cereal2$carbo) #impute -1
(N/A) with mean values
> cereal2$sugars <- ifelse(cereal2$sugars == -1, mean(cereal2$sugars), cereal2$sugars) #impute -1
(N/A) with mean values
> cereal2$potass <- ifelse(cereal2$potass == -1, mean(cereal2$potass), cereal2$potass) #impute -1
(N/A) with mean values
> cereal2$weight <- ifelse(cereal2$weight == -1, mean(cereal2$weight), cereal2$weight) #impute -1
(N/A) with mean values
> cereal2$cups <- ifelse(cereal2$cups == -1, mean(cereal2$cups), cereal2$cups) #impute -1 (
```

The resulting data did not have any unusual values:



```
> #Check the results
> summary(cereal2)
     name               mfr                type              calories        protein           fat       
 Length:77          Length:77          Length:77          Min.   : 50.0   Min.   :1.000   Min.   :0.000  
 Class :character   Class :character   Class :character   1st Qu.:100.0   1st Qu.:2.000   1st Qu.:0.000  
 Mode  :character   Mode  :character   Mode  :character   Median :110.0   Median :3.000   Median :1.000  
                                                          Mean   :106.9   Mean   :2.545   Mean   :1.013  
                                                          3rd Qu.:110.0   3rd Qu.:3.000   3rd Qu.:2.000  
                                                          Max.   :160.0   Max.   :6.000   Max.   :5.000  
     sodium          fiber            carbo          sugars           shelf           potass     
 Min.   :  0.0   Min.   : 0.000   Min.   : 5.0   Min.   : 0.000   Min.   :1.000   Min.   : 15.0  
 1st Qu.:130.0   1st Qu.: 1.000   1st Qu.:12.0   1st Qu.: 3.000   1st Qu.:1.000   1st Qu.: 45.0  
 Median :180.0   Median : 2.000   Median :14.6   Median : 7.000   Median :2.000   Median : 90.0  
 Mean   :159.7   Mean   : 2.152   Mean   :14.8   Mean   : 7.025   Mean   :2.208   Mean   : 98.6  
 3rd Qu.:210.0   3rd Qu.: 3.000   3rd Qu.:17.0   3rd Qu.:11.000   3rd Qu.:3.000   3rd Qu.:120.0  
 Max.   :320.0   Max.   :14.000   Max.   :23.0   Max.   :15.000   Max.   :3.000   Max.   :330.0  
    vitamins          weight           cups       
 Min.   :  0.00   Min.   :0.500   Min.   :0.2500  
 1st Qu.: 25.00   1st Qu.:1.000   1st Qu.:0.5873  
 Median : 25.00   Median :1.000   Median :0.7500  
 Mean   : 28.25   Mean   :1.029   Mean   :0.7728  
 3rd Qu.: 25.00   3rd Qu.:1.000   3rd Qu.:1.0000  
 Max.   :100.00   Max.   :1.500   Max.   :1.5000  
```

```
> #Boxplots for all numerical variables in the dataset
> par(mfrow=c(2,4))
> for (i in 4:15){
+    boxplot(cereal2[i], main=i)
+  }
```

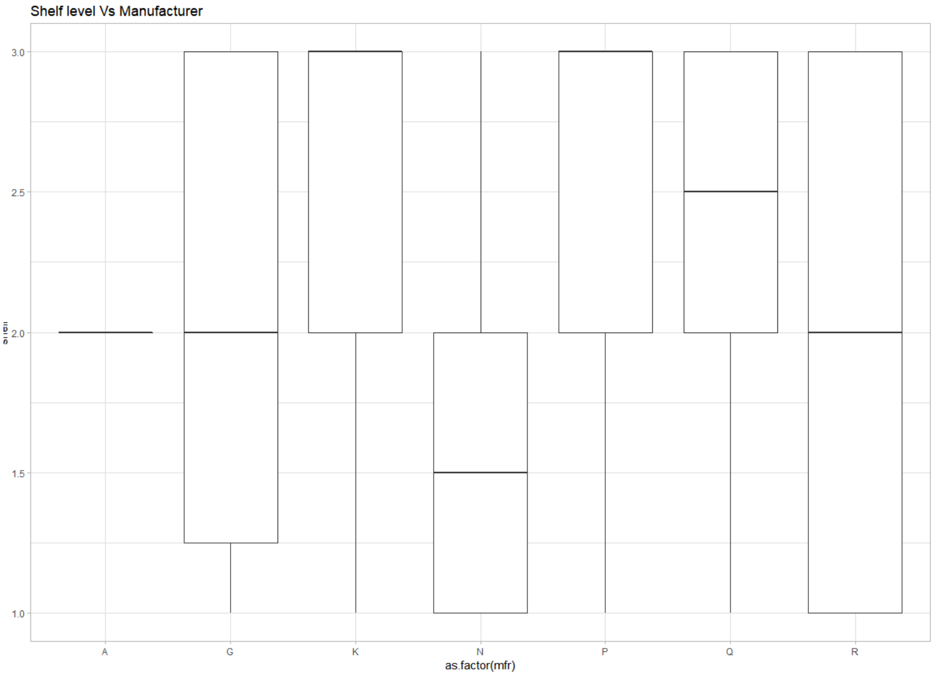


The data set is ready for analysis.

Research question 1: Is there a link between the manufacturer and the location shelf?

I would like to find if shelf placement depends on a manufacturer (e.g., some companies consistently get more profitable location at the eye level of the customers).

```
> qplot(as.factor(mfr), shelf, data = cereal2, geom = "boxplot", main="Shelf level Vs Manufacturer") +
+ theme_light()
>
```



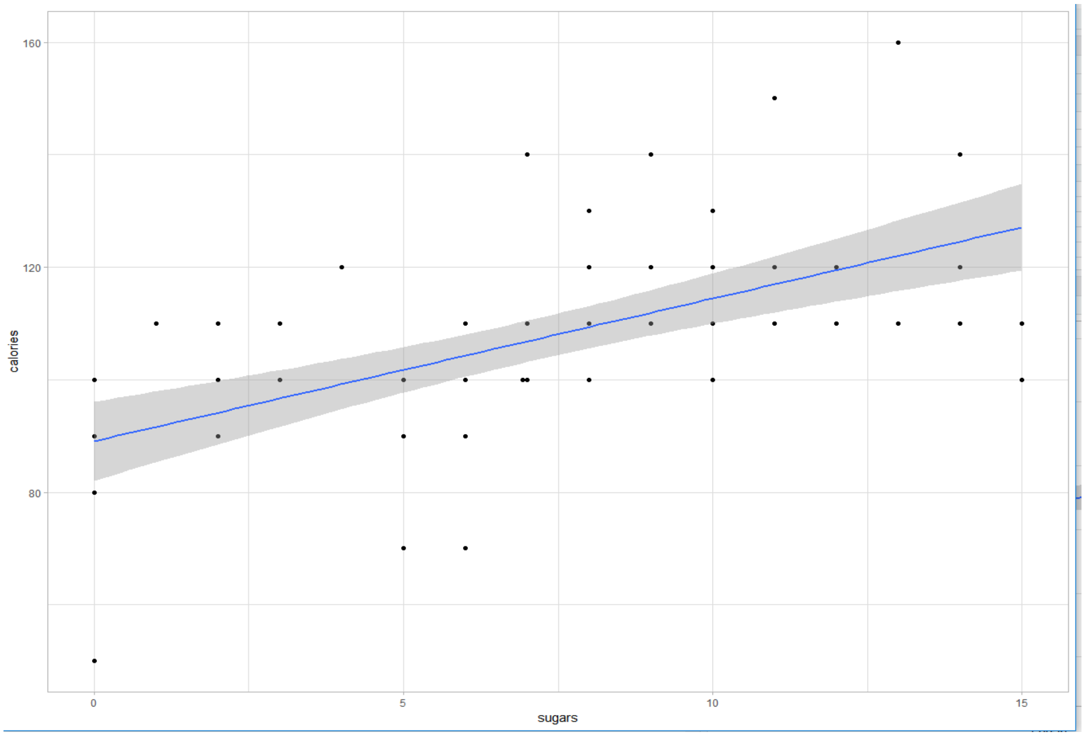
The above plot suggests, that there are manufacturers that use all three levels of shelves and there some that find their brands on two of the shelves. One of the manufacturers displays cereal only on the medium shelf. Overall, these graphical analysis does not yield consistent results. So, I used the ANOVA model to check if there is a relationship between manufacturer and shelf placement.

```
> #Use ANOVA to check for relationship between the manufacturer and the level of shelf
> shelf_mfr <- aov(shelf ~ as.factor(mfr), data = cereal2)
> summary(shelf_mfr)
               Df Sum Sq Mean Sq F value Pr(>F)
as.factor(mfr)  6   3.44  0.5727   0.814  0.562
Residuals      70  49.24  0.7034
```

These results show that there is no statistically significant link between the manufacturer and shelf placement.

Research question 2: Is there a linear relationship between the level of sugar and the total amount of calories per serving?

```
> #Q2: Relationship between sugar and calories
> qplot(sugars, calories, data=cereal2) + geom_smooth(method = "lm") + theme_light()
```



The above graph suggests that while there is a trend line characterizing the relationship between the number of calories and sugar content (which is to be expected), there are outliers on this graph. Some cereals have fewer calories that we could expect judging by their sugar content. The question is where the linear relationship is statistically significant or not.

H0: there is no significant linear relationship between the sugar content and the total calories;

Ha: there is a significant linear relationship between the sugar content and the overall number of calories.

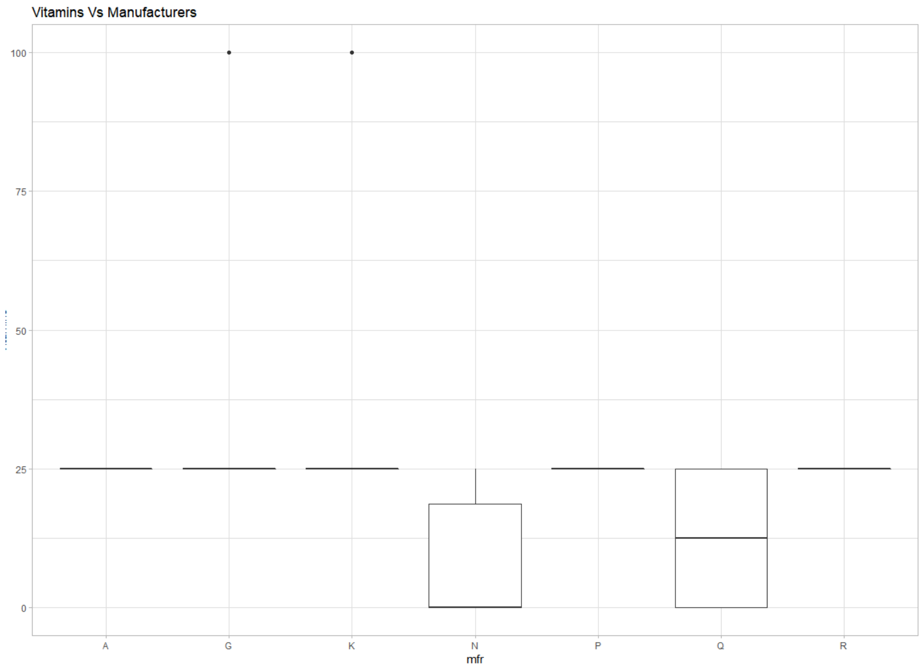
Next, I fit a linear regression model:

```
> #Fit a linear regression model
> sugar_calories_model = lm(calories ~ sugars, data = cereal2)
> summary(sugar_calories_model)

Call:
lm(formula = calories ~ sugars, data = cereal2)

Residuals:
   Min     1Q Median     3Q    Max 
-39.07  -9.50   0.50  10.93  37.96 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  89.0673     3.5164  25.329  < 2e-16 ***
sugars        2.5361     0.4263   5.948 8.03e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.17 on 75 degrees of freedom
Multiple R-squared:  0.3205,	Adjusted R-squared:  0.3115 
F-statistic: 35.38 on 1 and 75 DF,  p-value: 8.029e-08
```

The above output with the p-values smaller than the significance level of 0.05 shows that there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis. In fact, there is a statistically significant relationship between sugars and overall calories in breakfast cereals.

Research question 3: Do any of the manufacturers consistently offer healthier cereals in terms of vitamin content?

```
> #Q3: Do any of the manufacturers consistnetly offer healthier cereals (higher vitamin content)?
> qplot(mfr,vitamins, data = cereal2, geom = "boxplot", main="Vitamins Vs Manufacturers") + theme
_light()
```



Preliminary visual inspection suggests, that while vitamin content is approximately equal for most of the manufacturers, there were two observations with unusually high vitamin content and two manufacturers offering cereals with overall lower vitamins and minerals content than their competitors. At this point, it not possible to tell whether this impression is caused by inconsistent data, or there is a statistical significance there.

First, I used ANOVA:

```
> #Use ANOVA to find out if there is a relationhip between manufacturer and vitalmins
> vitamins_mfr <- aov(vitamins ~ as.factor(mfr), data = cereal2)
> summary(vitamins_mfr)
               Df Sum Sq Mean Sq F value Pr(>F)  
as.factor(mfr)  6   6607  1101.2    2.46 0.0323 *
Residuals      70  31331   447.6                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> #Use ANOVA to find out if there is a relationhip between manufacturer and vitalmins
> vitamins_mfr <- aov(vitamins ~ as.factor(mfr), data = cereal2)
> summary(vitamins_mfr)
               Df Sum Sq Mean Sq F value Pr(>F)  
as.factor(mfr)  6   6607  1101.2    2.46 0.0323 *
Residuals      70  31331   447.6                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The above output showed that there is a statistically significant difference between the mean vitamin content depending on a manufacturer. In order to find a particular manufacturer potentially offering healthier cereal choices, I used pairwise comparisons using TukeyHSD test:

```
> TukeyHSD(vitamins_mfr, conf.level = 0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = vitamins ~ as.factor(mfr), data = cereal2)

$`as.factor(mfr)`
             diff       lwr       upr     p adj
G-A  1.022727e+01 -55.44068 75.895223 0.9991120
K-A  9.782609e+00 -55.82324 75.388461 0.9993072
N-A -1.666667e+01 -86.03708 52.703746 0.9902777
P-A  1.421085e-14 -67.69859 67.698590 1.0000000
Q-A -1.250000e+01 -80.62039 55.620393 0.9977606
R-A  3.197442e-14 -68.12039 68.120393 1.0000000
K-G -4.446640e-01 -19.59745 18.708119 1.0000000
N-G -2.689394e+01 -56.47358  2.685704 0.0988149
P-G -1.022727e+01 -35.63987 15.185329 0.8833665
Q-G -2.272727e+01 -49.24310  3.788554 0.1411853
R-G -1.022727e+01 -36.74310 16.288554 0.9026428
N-K -2.644928e+01 -55.89080  2.992252 0.1066579
P-K -9.782609e+00 -35.03431 15.469096 0.9007812
Q-K -2.228261e+01 -48.64427  4.079055 0.1525682
R-K -9.782609e+00 -36.14427 16.579055 0.9177640
P-N  1.666667e+01 -17.18263 50.515962 0.7469673
Q-N  4.166667e+00 -30.51854 38.851873 0.9998010
R-N  1.666667e+01 -18.01854 51.351873 0.7677485
Q-P -1.250000e+01 -43.70751 18.707508 0.8856632
R-P  1.776357e-14 -31.20751 31.207508 1.0000000
R-Q  1.250000e+01 -19.61226 44.612261 0.8987241
```

Unfortunately, it showed that there is no one manufacturer that is offering cereals than have statistically significantly different vitamin content that any other manufacturer in the group.

Research question 4: Can the type of the cereals (hot, called) be predicted based on their nutritional content?

In order to answer this question, I had to prepare the dataset first – remove unrelated columns (name and manufacturer), and change the data type for the target variable (type) to a factor.

```
> #Data preparation
> cereal3<-cereal2
> cereal3 <- cereal3[-1] #Drop name variable
> cereal3 <- cereal3[-1] #Drop mfr variable
> cereal3$type <-as.factor(cereal3$type) #Change type variable to a factor
> str(cereal3)
'data.frame':	77 obs. of  13 variables:
 $ type    : Factor w/ 2 levels "C","H": 1 1 1 1 1 1 1 1 1 1 ...
 $ calories: int  70 120 70 50 110 110 110 130 90 90 ...
 $ protein : int  4 3 4 4 2 2 2 3 2 3 ...
 $ fat     : int  1 5 1 0 2 2 0 2 1 0 ...
 $ sodium  : int  130 15 260 140 200 180 125 210 200 210 ...
 $ fiber   : num  10 2 9 14 1 1.5 1 2 4 5 ...
 $ carbo   : num  5 8 7 8 14 10.5 11 18 15 13 ...
 $ sugars  : num  6 8 5 0 8 10 14 8 6 5 ...
 $ shelf   : int  3 3 3 3 3 1 2 3 1 3 ...
 $ potass  : num  280 135 320 330 96.1 ...
 $ vitamins: int  25 0 25 25 25 25 25 25 25 25 ...
 $ weight  : num  1 1 1 1 1 1 1.33 1 1 ...
 $ cups    : num  0.33 0.587 0.33 0.5 0.75 ...
```

Then, I split the data set into the training and testing parts:

```
> #Split the dataset into 80% training and  20% testing data
> set.seed(123)
> split = 0.08
> trainIndex <- createDataPartition(cereal3$type, p=split, list = FALSE)
> cereal_test<- cereal3[trainIndex, ]
> cereal_train <- cereal3[-trainIndex, ]
```

The resulting training set contains 70 observations and testing set is composed of 7 observations.

It should be noted that the initial dataset is very imbalances by type.

```
> table(cereal3$type)

 C  H 
74  3 
```

It contains 74 observations for cold cereals and only 3 observations for hot cereals. As a result, these extreme imbalance is present in both training and testing sets:

```
> table(cereal_train$type)

 C  H 
68  2 
> table(cereal_test$type)
```

```
C H
6 1
```

This imbalance is likely to affect predictive power of any model. So, I used the Naïve Bayes algorithm that is less sensitive in such situations.

I used the following code to fit the model, generate prediction and evaluate results using a confusion matrix.

```
> #Test dataset without the type column
> cereal_test_notype <- cereal_test[-1]
>
> #Use Naive Bayes to fit the model
> set.seed(123)
> nb_model <- naiveBayes(type ~ ., data=cereal_train)
> #Generate predictions
> nb_prediction <- predict(nb_model,newdata = cereal_test_notype)
> #generate classification table
> nb_table1 <- table(nb_prediction, cereal_test$type)
> nb_table1

nb_prediction C H
            C 4 0
            H 2 1
> #Evaluate model performance
> confusionMatrix(nb_table1)
Confusion Matrix and Statistics


nb_prediction C H
            C 4 0
            H 2 1

               Accuracy : 0.7143
                 95% CI : (0.2904, 0.9633)
    No Information Rate : 0.8571
    P-Value [Acc > NIR] : 0.9348

                  Kappa : 0.3636
 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 0.6667
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.3333
             Prevalence : 0.8571
         Detection Rate : 0.5714
   Detection Prevalence : 0.5714
      Balanced Accuracy : 0.8333

       'Positive' Class : C
```

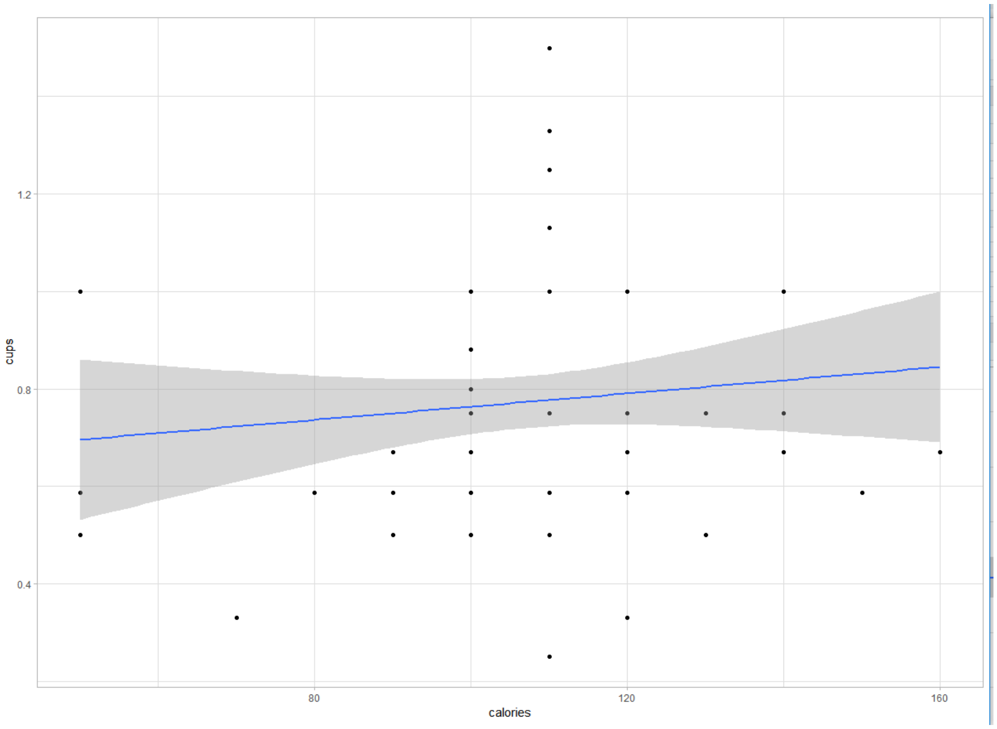
On a limited test set of 7 observations, the model demonstrated 71.43% accuracy. It did detect the lone hot cereal in the test set, but mistakenly classified 2 of the cold cereals as hot. Balanced accuracy taking into consideration probabilities in the imbalanced test set was 83.33%.

It means, that while the accuracy of these particular model is not very high, there is a possibility to predict a type of breakfast cereal, especially if there is a possibility of collecting a more balanced training data set.

Research question 5: Do manufacturers use smaller serving sizes in order to decrease the number of calories indicated on mandatory labels?

In other words, I would like to find out if there is a correlation between a serving size and calorie content for the cereals in the dataset.

```
qplot(calories, cups, data=cereal2) +  geom_smooth(method = "lm") + theme_light()
```



Visual analysis again is inconclusive as there are samples that have the same calorie content, but varying serving sizes, that would suggest that some manufactured try to limit the number of calories by decreasing serving sizes. At the same time, some cereals display linear dependency between the calorie content and the number of cups in a serving size.

H0: there is no statistically significant relationship between the serving size and the calorie content (the variables are independent);

Ha: there is a statistically significant relationship between the serving size and calorie content (the variables are not independent).

```
> cor.test(cereal2$calories, cereal2$cups)

         Pearson's product-moment correlation

data:  cereal2$calories and cereal2$cups
t = 0.98626, df = 75, p-value = 0.3272
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1137089  0.3287976
sample estimates:
      cor 
0.1131517 
```

The above output shows that the p-value for this test is 0.3272 which is higher than the significance level of 0.05. It means that we do not have sufficient evidence to reject the null hypothesis that there is no statistically significant link between the two variables.

In the case of cereals, one might assume that with an increase in the amount of food (number of cups) in one serving, the calorie content for this serving should also increase. However, the above test indicated that there is no evidence to reject the idea that these two variables are independent. It means that manufacturers do intentionally manipulate the amount of serving they use to calculate the calorie content for the purposes of mandatory labeling. It makes the task of comparing breakfast cereals before purchasing more difficult for consumers.