Name: Chaitanya

Email_Id : sai02220@gmail.com

Phone: 7904479679

PROJECT REPORT

ON

CEREALS ANALYSIS

PROBLEM STATEMENT

The problem statement is **classification problem. Which products of Cereal will an customer buy again?**

INTRODUCTION

In this competition, Cereals is challenging as analyst community to use this anonymized data on customer orders over time to predict which previously purchased products will be in a user's next order.

The dataset for this competition is a relational set of files describing customers' orders over time. The goal of the competition is to predict which products will be in a user's next order. The dataset is anonymized and contains a sample of over 75 Cereal orders from more than 200,000 users. For each user, they provide between 4 and 100 of their orders and rating, with the sequence of rating in each order.

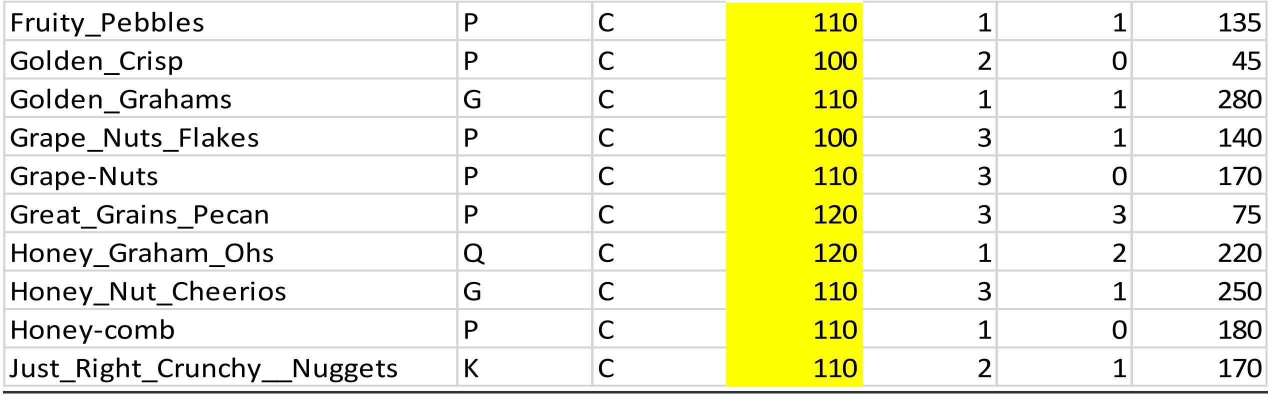
Each entity (calories, name, rating, shelf, etc.) has an associated unique id.

It consist of Two files

1. Cereals.xlsx

2. Cereals_practice.xlsx

Data



Fruity_Pebbles	P	C	110	1	1	135
Golden_Crisp	P	C	100	2	0	45
Golden_Grahams	G	C	110	1	1	280
Grape_Nuts_Flakes	P	C	100	3	1	140
Grape-Nuts	P	C	110	3	0	170
Great_Grains_Pecan	P	C	120	3	3	75
Honey_Graham_Ohs	Q	C	120	1	2	220
Honey_Nut_Cheerios	G	C	110	3	1	250
Honey-comb	P	C	110	1	0	180
Just_Right_Crunchy__Nuggets	K	C	110	2	1	170

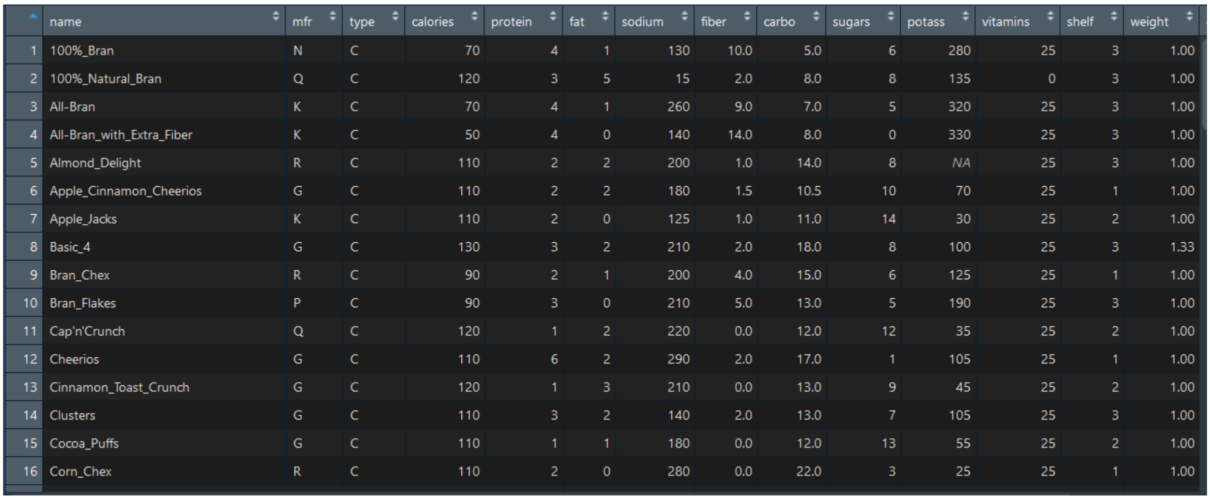
1. Reading the data

```
library(psych)

library(corrplot)

library(readxl)

cereals_practice <- read_excel("cereals_practice.xlsx")

View(cereals_practice)

#removal of Na

a=cereals_practice

str(a)

a=na.omit(a)

View(a)

summary(a)

#converting characters into factors

a$name=as.factor(a$name)

a$mfr=as.factor(a$mfr)

a$type=as.factor(a$type)

summary(a)
```

```
str(a)
#finding correlation
library(corrplot)
corfull=cor(a)
corrplot(corfull)
shel=cor(a[c("calories","shelf")])
corrplot(shel)
b=a[-c(1,2,3)]
View(b)
calorie=cor(b)
corrplot(calorie)
calrating=cor(a[c("calories","rating")])
corrplot(calrating)
#--------------------------------
sale=lm(a$calories~protein+fat+sodium+fiber+carbo+sugars+
        potass+vitamins+rating, data = a)
sale
summary(sale)
rating=lm(a$rating~calories+protein+fat+sodium+fiber+carbo+sugars+
potass+vitamins, data=a)
rating
summary(rating)
#----------------------------------
```

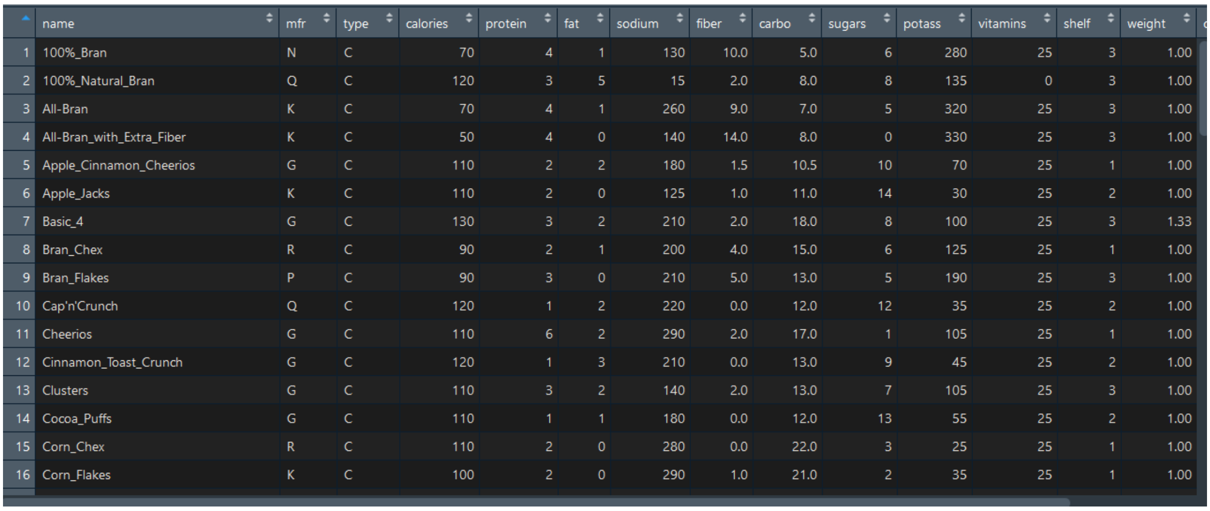
2. Datasets

a. Cereals data



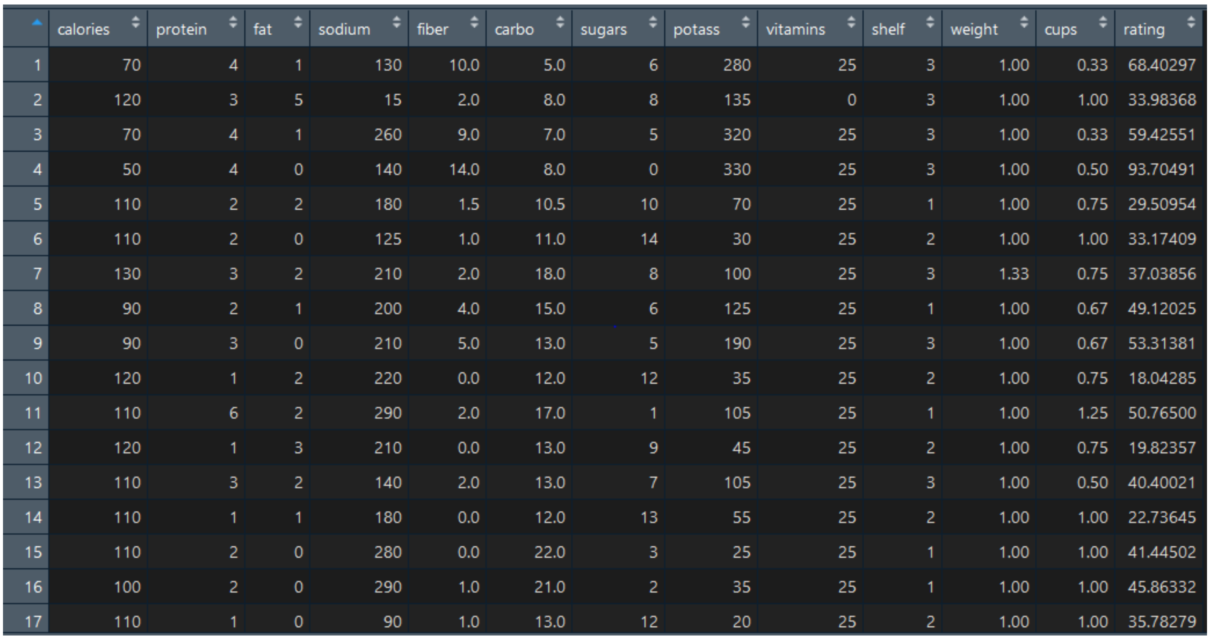
	name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight
1	100%_Bran	N	C	70	4	1	130	10.0	5.0	6	280	25	3	1.00
2	100%_Natural_Bran	Q	C	120	3	5	15	2.0	8.0	8	135	0	3	1.00
3	All-Bran	K	C	70	4	1	260	9.0	7.0	5	320	25	3	1.00
4	All-Bran_with_Extra_Fiber	K	C	50	4	0	140	14.0	8.0	0	330	25	3	1.00
5	Almond_Delight	R	C	110	2	2	200	1.0	14.0	8	NA	25	3	1.00
6	Apple_Cinnamon_Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25	1	1.00
7	Apple_Jacks	K	C	110	2	0	125	1.0	11.0	14	30	25	2	1.00
8	Basic_4	G	C	130	3	2	210	2.0	18.0	8	100	25	3	1.33
9	Bran_Chex	R	C	90	2	1	200	4.0	15.0	6	125	25	1	1.00
10	Bran_Flakes	P	C	90	3	0	210	5.0	13.0	5	190	25	3	1.00
11	Cap'n'Crunch	Q	C	120	1	2	220	0.0	12.0	12	35	25	2	1.00
12	Cheerios	G	C	110	6	2	290	2.0	17.0	1	105	25	1	1.00
13	Cinnamon_Toast_Crunch	G	C	120	1	3	210	0.0	13.0	9	45	25	2	1.00
14	Clusters	G	C	110	3	2	140	2.0	13.0	7	105	25	3	1.00
15	Cocoa_Puffs	G	C	110	1	1	180	0.0	12.0	13	55	25	2	1.00
16	Corn_Chex	R	C	110	2	0	280	0.0	22.0	3	25	25	1	1.00

b. Assigned to A removing of all NA(Cleaned data)



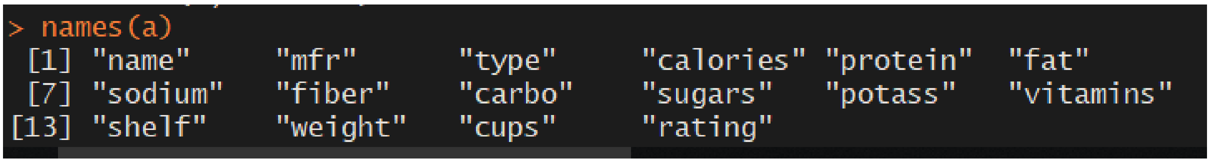
	name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight
1	100%_Bran	N	C	70	4	1	130	10.0	5.0	6	280	25	3	1.00
2	100%_Natural_Bran	Q	C	120	3	5	15	2.0	8.0	8	135	0	3	1.00
3	All-Bran	K	C	70	4	1	260	9.0	7.0	5	320	25	3	1.00
4	All-Bran_with_Extra_Fiber	K	C	50	4	0	140	14.0	8.0	0	330	25	3	1.00
5	Apple_Cinnamon_Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25	1	1.00
6	Apple_Jacks	K	C	110	2	0	125	1.0	11.0	14	30	25	2	1.00
7	Basic_4	G	C	130	3	2	210	2.0	18.0	8	100	25	3	1.33
8	Bran_Chex	R	C	90	2	1	200	4.0	15.0	6	125	25	1	1.00
9	Bran_Flakes	P	C	90	3	0	210	5.0	13.0	5	190	25	3	1.00
10	Cap'n'Crunch	Q	C	120	1	2	220	0.0	12.0	12	35	25	2	1.00
11	Cheerios	G	C	110	6	2	290	2.0	17.0	1	105	25	1	1.00
12	Cinnamon_Toast_Crunch	G	C	120	1	3	210	0.0	13.0	9	45	25	2	1.00
13	Clusters	G	C	110	3	2	140	2.0	13.0	7	105	25	3	1.00
14	Cocoa_Puffs	G	C	110	1	1	180	0.0	12.0	13	55	25	2	1.00
15	Corn_Chex	R	C	110	2	0	280	0.0	22.0	3	25	25	1	1.00
16	Corn_Flakes	K	C	100	2	0	290	1.0	21.0	2	35	25	1	1.00

c. Data set without name, mfr, type



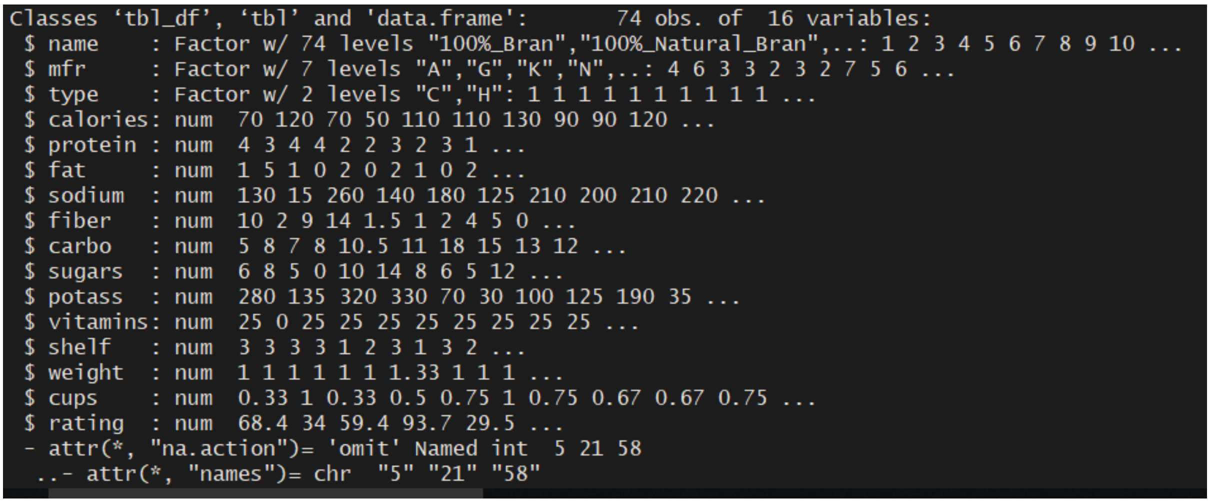
	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
1	70	4	1	130	10.0	5.0	6	280	25	3	1.00	0.33	68.40297
2	120	3	5	15	2.0	8.0	8	135	0	3	1.00	1.00	33.98368
3	70	4	1	260	9.0	7.0	5	320	25	3	1.00	0.33	59.42551
4	50	4	0	140	14.0	8.0	0	330	25	3	1.00	0.50	93.70491
5	110	2	2	180	1.5	10.5	10	70	25	1	1.00	0.75	29.50954
6	110	2	0	125	1.0	11.0	14	30	25	2	1.00	1.00	33.17409
7	130	3	2	210	2.0	18.0	8	100	25	3	1.33	0.75	37.03856
8	90	2	1	200	4.0	15.0	6	125	25	1	1.00	0.67	49.12025
9	90	3	0	210	5.0	13.0	5	190	25	3	1.00	0.67	53.31381
10	120	1	2	220	0.0	12.0	12	35	25	2	1.00	0.75	18.04285
11	110	6	2	290	2.0	17.0	1	105	25	1	1.00	1.25	50.76500
12	120	1	3	210	0.0	13.0	9	45	25	2	1.00	0.75	19.82357
13	110	3	2	140	2.0	13.0	7	105	25	3	1.00	0.50	40.40021
14	110	1	1	180	0.0	12.0	13	55	25	2	1.00	1.00	22.73645
15	110	2	0	280	0.0	22.0	3	25	25	1	1.00	1.00	41.44502
16	100	2	0	290	1.0	21.0	2	35	25	1	1.00	1.00	45.86332
17	110	1	0	90	1.0	13.0	12	20	25	2	1.00	1.00	35.78279

Cereals Variables



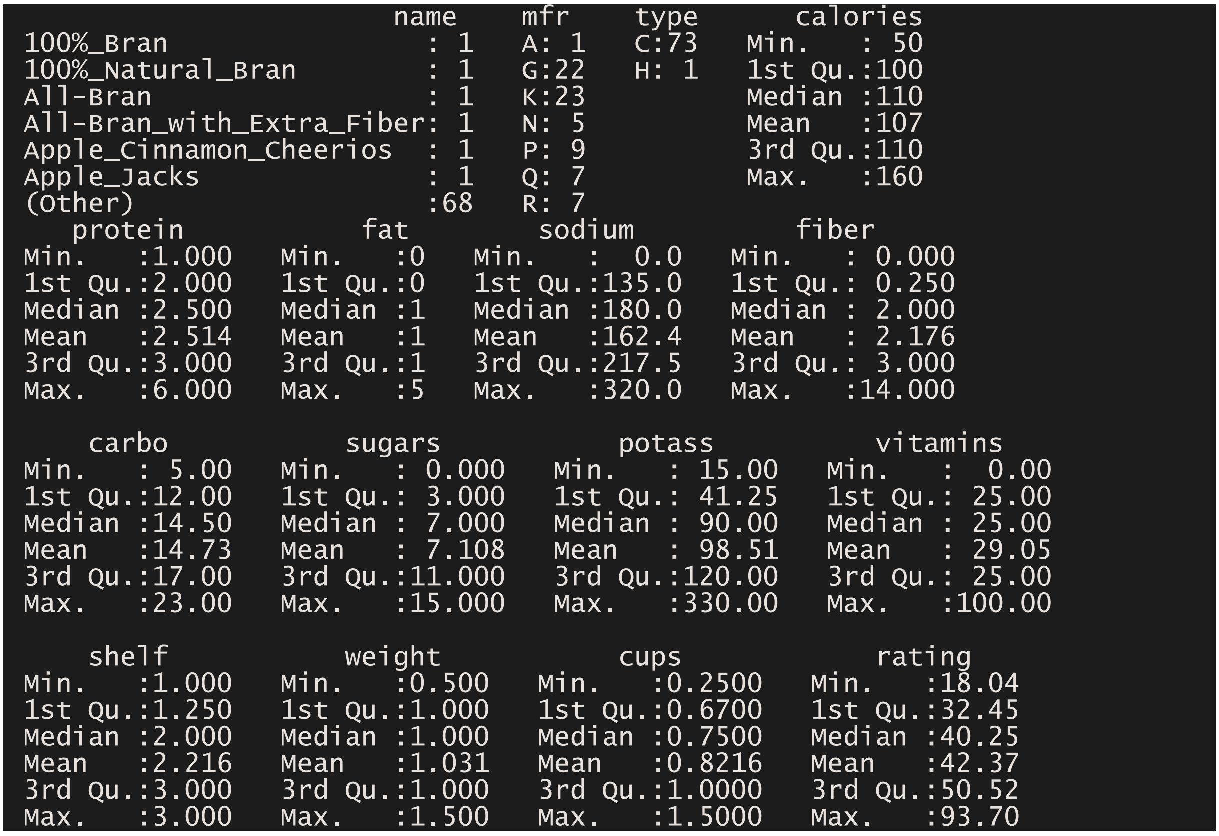
```
> names(a)
 [1] "name"     "mfr"      "type"     "calories" "protein"  "fat"
 [7] "sodium"   "fiber"    "carbo"    "sugars"   "potass"   "vitamins"
[13] "shelf"    "weight"   "cups"     "rating"
```

Structure of A=cereal data (cleaned data)



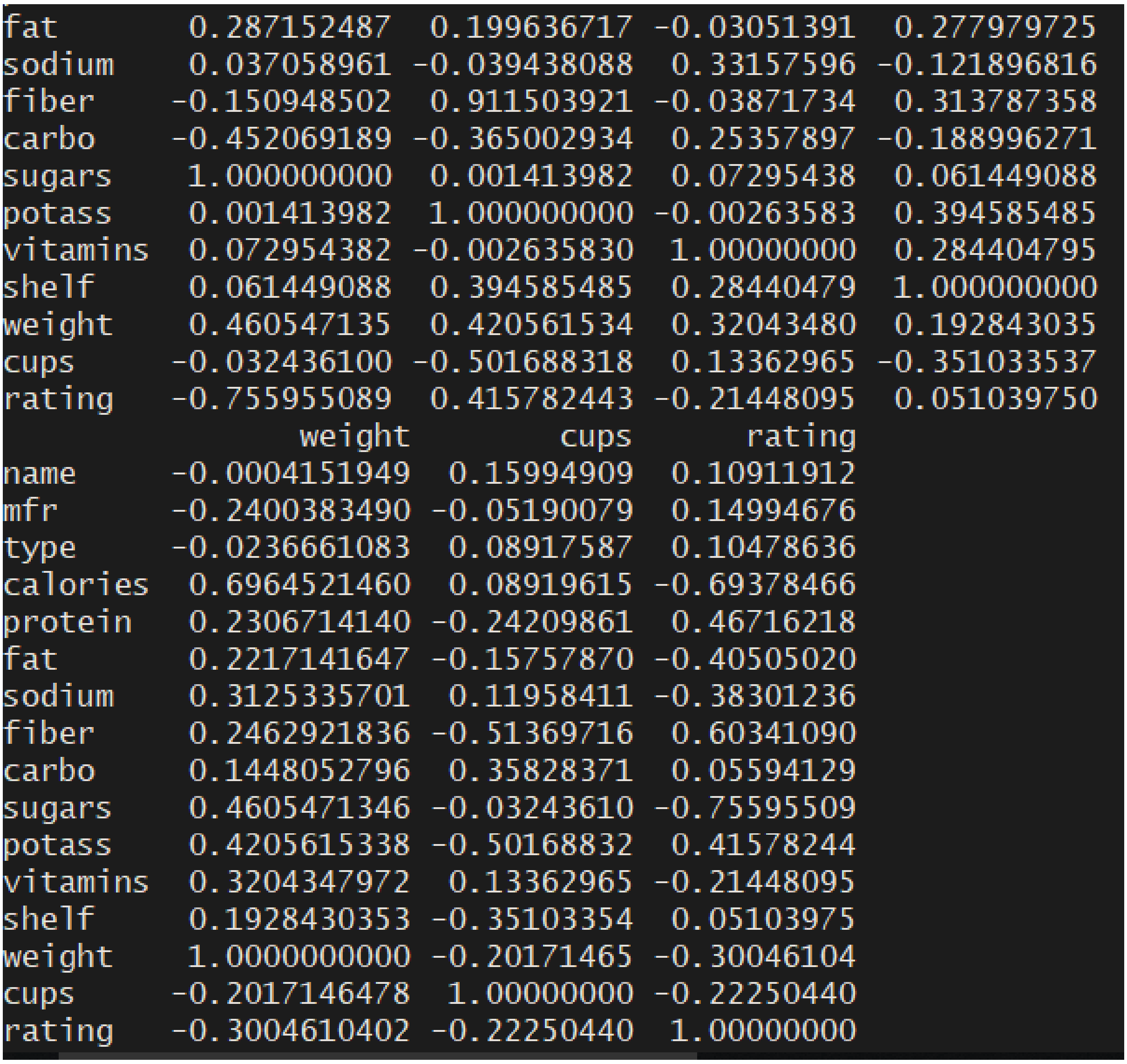
```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':	74 obs. of  16 variables:
 $ name    : Factor w/ 74 levels "100%_Bran","100%_Natural_Bran",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ mfr     : Factor w/ 7 levels "A","G","K","N",..: 4 6 3 3 2 3 2 7 5 6 ...
 $ type    : Factor w/ 2 levels "C","H": 1 1 1 1 1 1 1 1 1 1 ...
 $ calories: num  70 120 70 50 110 110 130 90 90 120 ...
 $ protein : num  4 3 4 4 2 2 3 2 3 1 ...
 $ fat     : num  1 5 1 0 2 0 2 1 0 2 ...
 $ sodium  : num  130 15 260 140 180 125 210 200 210 220 ...
 $ fiber   : num  10 2 9 14 1.5 1 2 4 5 0 ...
 $ carbo   : num  5 8 7 8 10.5 11 18 15 13 12 ...
 $ sugars  : num  6 8 5 0 10 14 8 6 5 12 ...
 $ potass  : num  280 135 320 330 70 30 100 125 190 35 ...
 $ vitamins: num  25 0 25 25 25 25 25 25 25 25 ...
 $ shelf   : num  3 3 3 3 1 2 3 1 3 2 ...
 $ weight  : num  1 1 1 1 1 1 1.33 1 1 1 ...
 $ cups    : num  0.33 1 0.33 0.5 0.75 1 0.75 0.67 0.67 0.75 ...
 $ rating  : num  68.4 34 59.4 93.7 29.5 ...
 - attr(*, "na.action")= 'omit' Named int  5 21 58
  ..- attr(*, "names")= chr  "5" "21" "58"
```

3. Summary of Dataset



```
                        name    mfr    type      calories
 100%_Bran                : 1   A: 1   C:73   Min.   : 50
 100%_Natural_Bran        : 1   G:22   H: 1   1st Qu.:100
 All-Bran                 : 1   K:23          Median :110
 All-Bran_with_Extra_Fiber: 1   N: 5          Mean   :107
 Apple_Cinnamon_Cheerios  : 1   P: 9          3rd Qu.:110
 Apple_Jacks              : 1   Q: 7          Max.   :160
 (Other)                  :68   R: 7
    protein           fat         sodium          fiber
 Min.   :1.000   Min.   :0   Min.   :  0.0   Min.   : 0.000
 1st Qu.:2.000   1st Qu.:0   1st Qu.:135.0   1st Qu.: 0.250
 Median :2.500   Median :1   Median :180.0   Median : 2.000
 Mean   :2.514   Mean   :1   Mean   :162.4   Mean   : 2.176
 3rd Qu.:3.000   3rd Qu.:1   3rd Qu.:217.5   3rd Qu.: 3.000
 Max.   :6.000   Max.   :5   Max.   :320.0   Max.   :14.000

     carbo           sugars           potass          vitamins
 Min.   : 5.00   Min.   : 0.000   Min.   : 15.00   Min.   :  0.00
 1st Qu.:12.00   1st Qu.: 3.000   1st Qu.: 41.25   1st Qu.: 25.00
 Median :14.50   Median : 7.000   Median : 90.00   Median : 25.00
 Mean   :14.73   Mean   : 7.108   Mean   : 98.51   Mean   : 29.05
 3rd Qu.:17.00   3rd Qu.:11.000   3rd Qu.:120.00   3rd Qu.: 25.00
 Max.   :23.00   Max.   :15.000   Max.   :330.00   Max.   :100.00

     shelf           weight          cups            rating
 Min.   :1.000   Min.   :0.500   Min.   :0.2500   Min.   :18.04
 1st Qu.:1.250   1st Qu.:1.000   1st Qu.:0.6700   1st Qu.:32.45
 Median :2.000   Median :1.000   Median :0.7500   Median :40.25
 Mean   :2.216   Mean   :1.031   Mean   :0.8216   Mean   :42.37
 3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:50.52
 Max.   :3.000   Max.   :1.500   Max.   :1.5000   Max.   :93.70
```

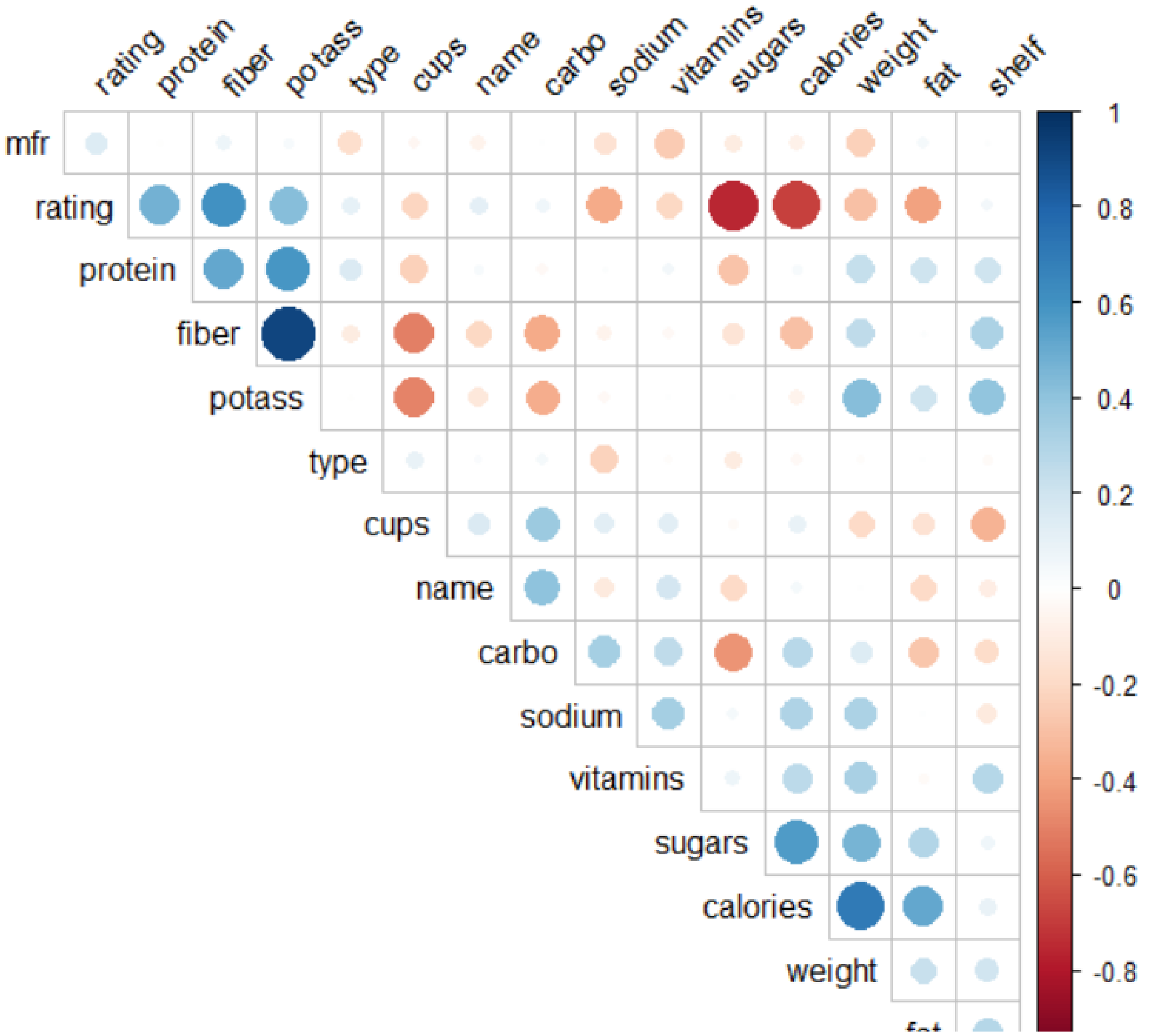
Correlation of A=cereal data (cleaned data set)



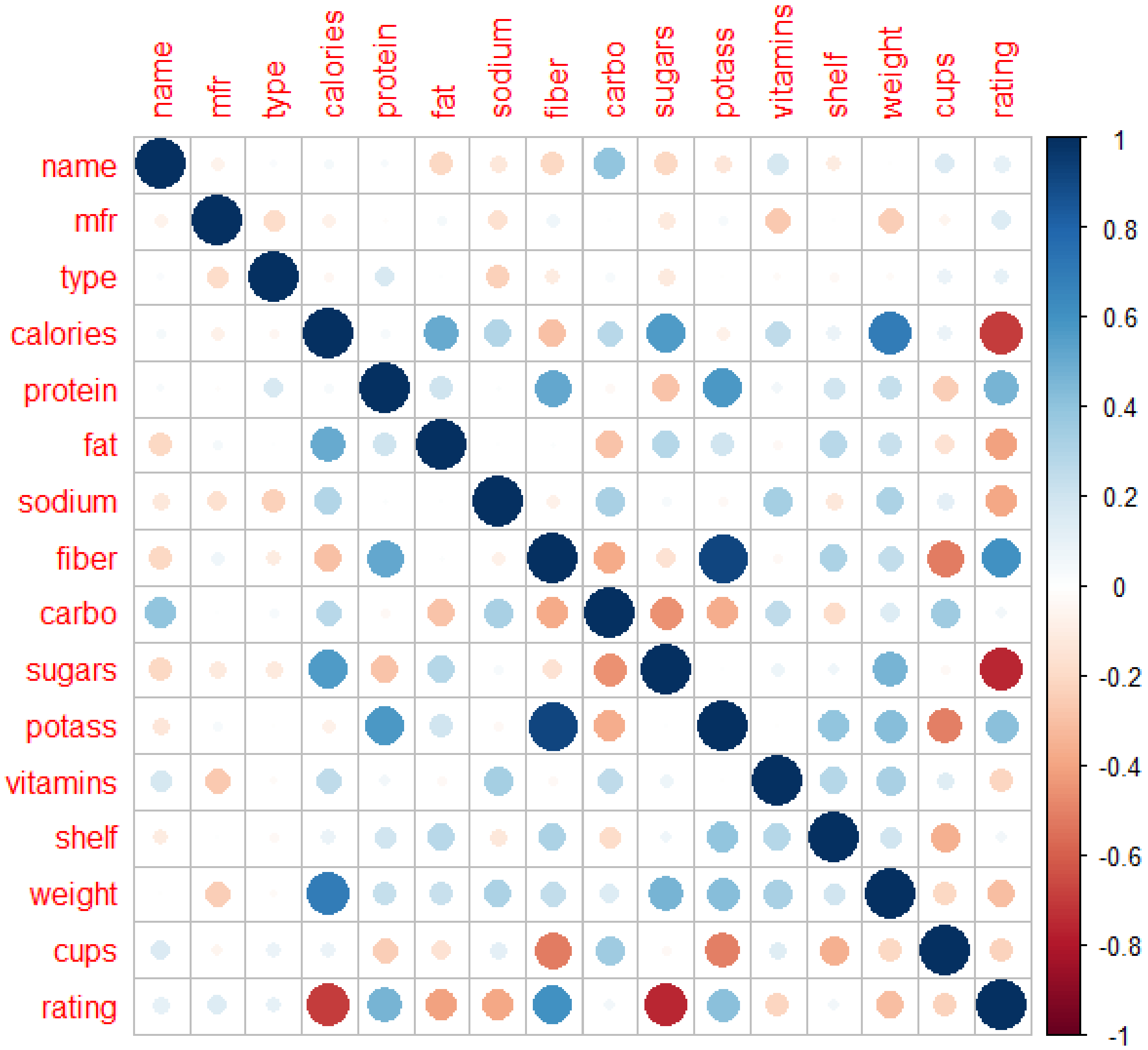
```
fat        0.287152487  0.199636717 -0.03051391  0.277979725
sodium     0.037058961 -0.039438088  0.33157596 -0.121896816
fiber     -0.150948502  0.911503921 -0.03871734  0.313787358
carbo     -0.452069189 -0.365002934  0.25357897 -0.188996271
sugars     1.000000000  0.001413982  0.07295438  0.061449088
potass     0.001413982  1.000000000 -0.00263583  0.394585485
vitamins   0.072954382 -0.002635830  1.00000000  0.284404795
shelf      0.061449088  0.394585485  0.28440479  1.000000000
weight     0.460547135  0.420561534  0.32043480  0.192843035
cups      -0.032436100 -0.501688318  0.13362965 -0.351033537
rating    -0.755955089  0.415782443 -0.21448095  0.051039750
                weight         cups       rating
name      -0.0004151949  0.15994909  0.10911912
mfr       -0.2400383490 -0.05190079  0.14994676
type      -0.0236661083  0.08917587  0.10478636
calories   0.6964521460  0.08919615 -0.69378466
protein    0.2306714140 -0.24209861  0.46716218
fat        0.2217141647 -0.15757870 -0.40505020
sodium     0.3125335701  0.11958411 -0.38301236
fiber      0.2462921836 -0.51369716  0.60341090
carbo      0.1448052796  0.35828371  0.05594129
sugars     0.4605471346 -0.03243610 -0.75595509
potass     0.4205615338 -0.50168832  0.41578244
vitamins   0.3204347972  0.13362965 -0.21448095
shelf      0.1928430353 -0.35103354  0.05103975
weight     1.0000000000 -0.20171465 -0.30046104
cups      -0.2017146478  1.00000000 -0.22250440
rating    -0.3004610402 -0.22250440  1.00000000
```

Representing Correlational graph of a=cereal data (cleaned data)

FIG(1.0)



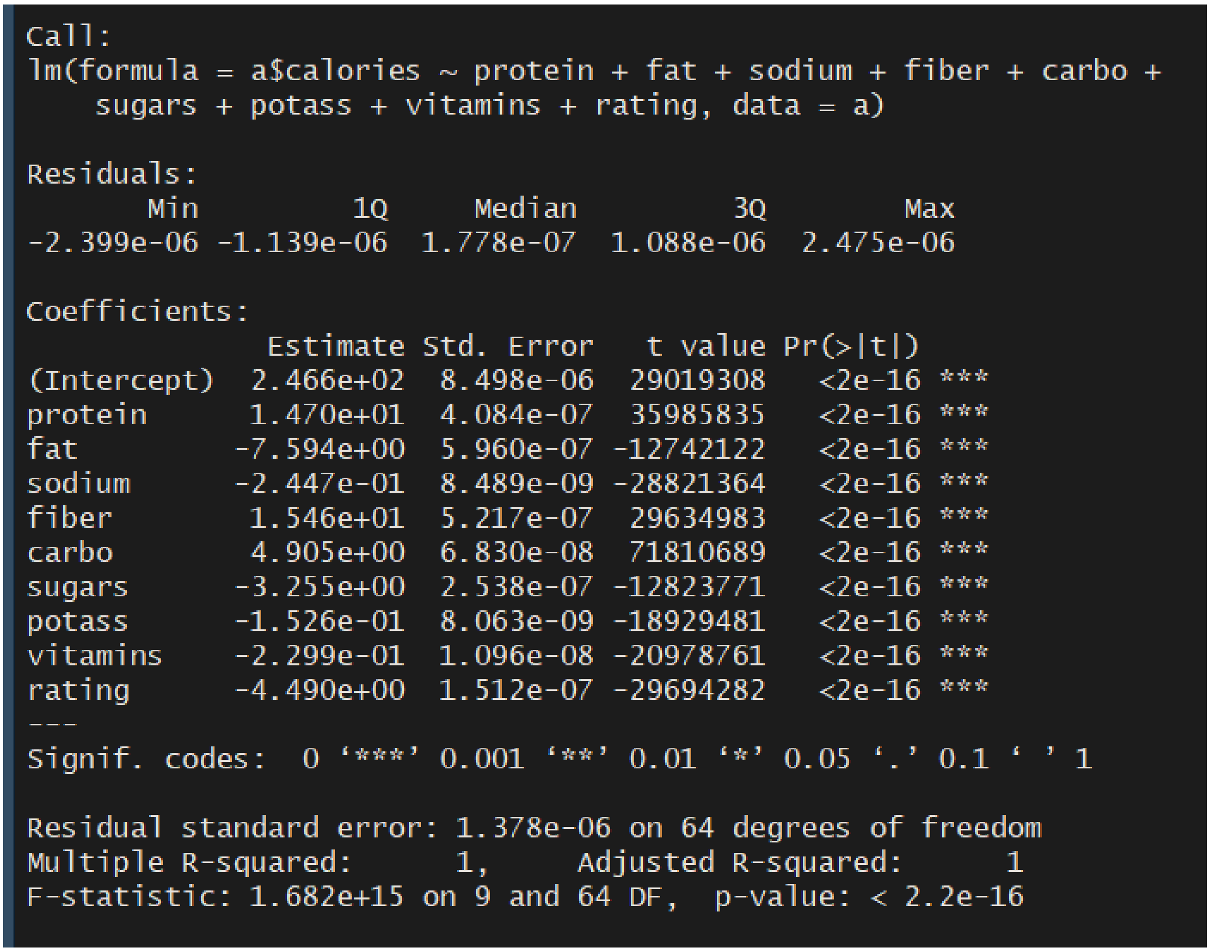
FIG(1.1)



As we can observe how all the variables are co-related to eachother , which means if an

One objective is sold there are chances of other being sold . where the blue indicates the co-relation between the objectives and the red indicates no co-relation

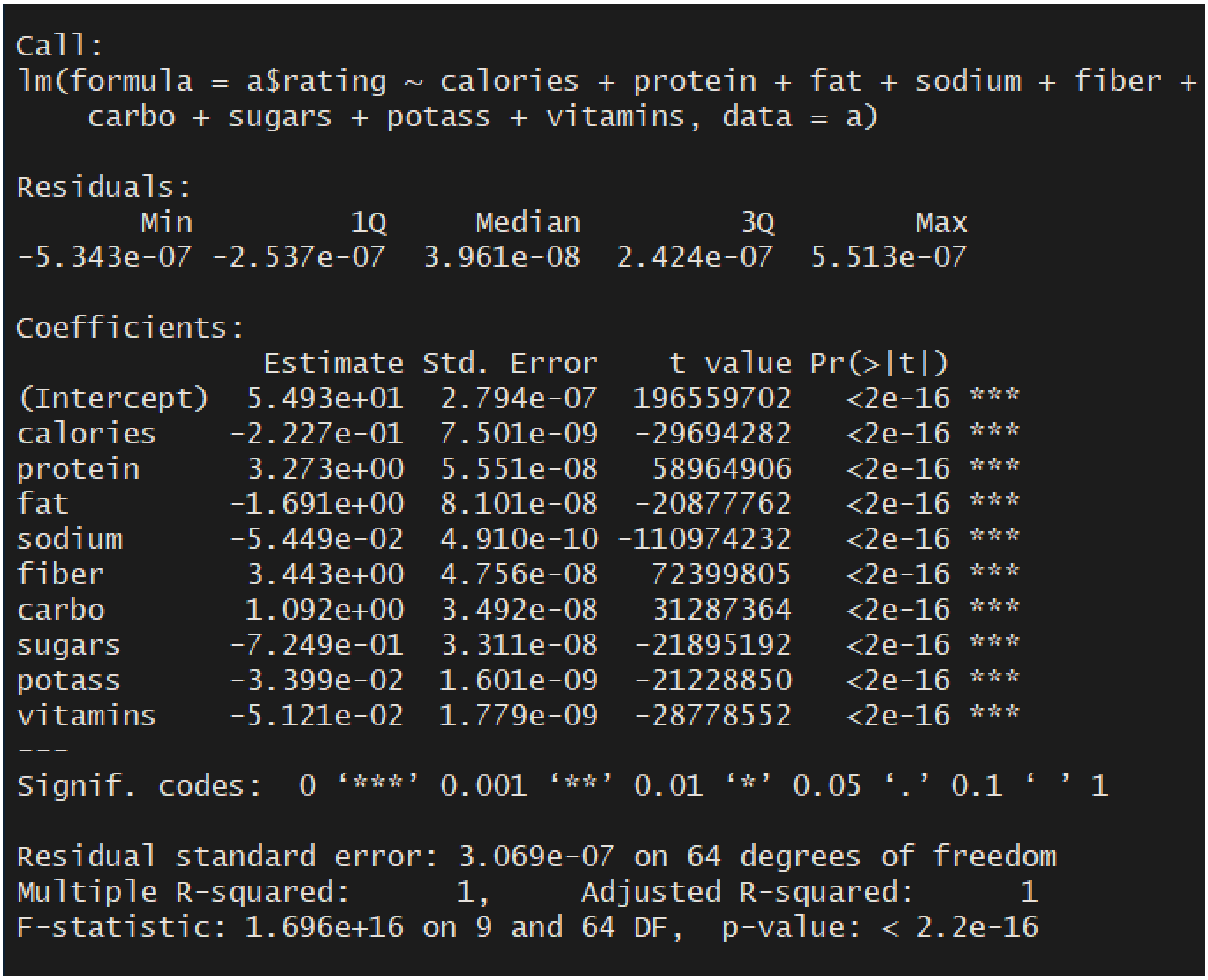
Taking calories and rating into consideration



```
Call:
lm(formula = a$calories ~ protein + fat + sodium + fiber + carbo +
    sugars + potass + vitamins + rating, data = a)

Residuals:
       Min         1Q     Median         3Q        Max 
-2.399e-06 -1.139e-06  1.778e-07  1.088e-06  2.475e-06 

Coefficients:
              Estimate Std. Error   t value Pr(>|t|)    
(Intercept)  2.466e+02  8.498e-06  29019308   <2e-16 ***
protein      1.470e+01  4.084e-07  35985835   <2e-16 ***
fat         -7.594e+00  5.960e-07 -12742122   <2e-16 ***
sodium      -2.447e-01  8.489e-09 -28821364   <2e-16 ***
fiber        1.546e+01  5.217e-07  29634983   <2e-16 ***
carbo        4.905e+00  6.830e-08  71810689   <2e-16 ***
sugars      -3.255e+00  2.538e-07 -12823771   <2e-16 ***
potass      -1.526e-01  8.063e-09 -18929481   <2e-16 ***
vitamins    -2.299e-01  1.096e-08 -20978761   <2e-16 ***
rating      -4.490e+00  1.512e-07 -29694282   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.378e-06 on 64 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1 
F-statistic: 1.682e+15 on 9 and 64 DF,  p-value: < 2.2e-16
```

In the above , we can see that it is significant which means it’s a good factor to increase the chances of sales and by the sample provided we can say 100% of variance in calories is explained by the variance mentioned in the above pic

Rating of the products

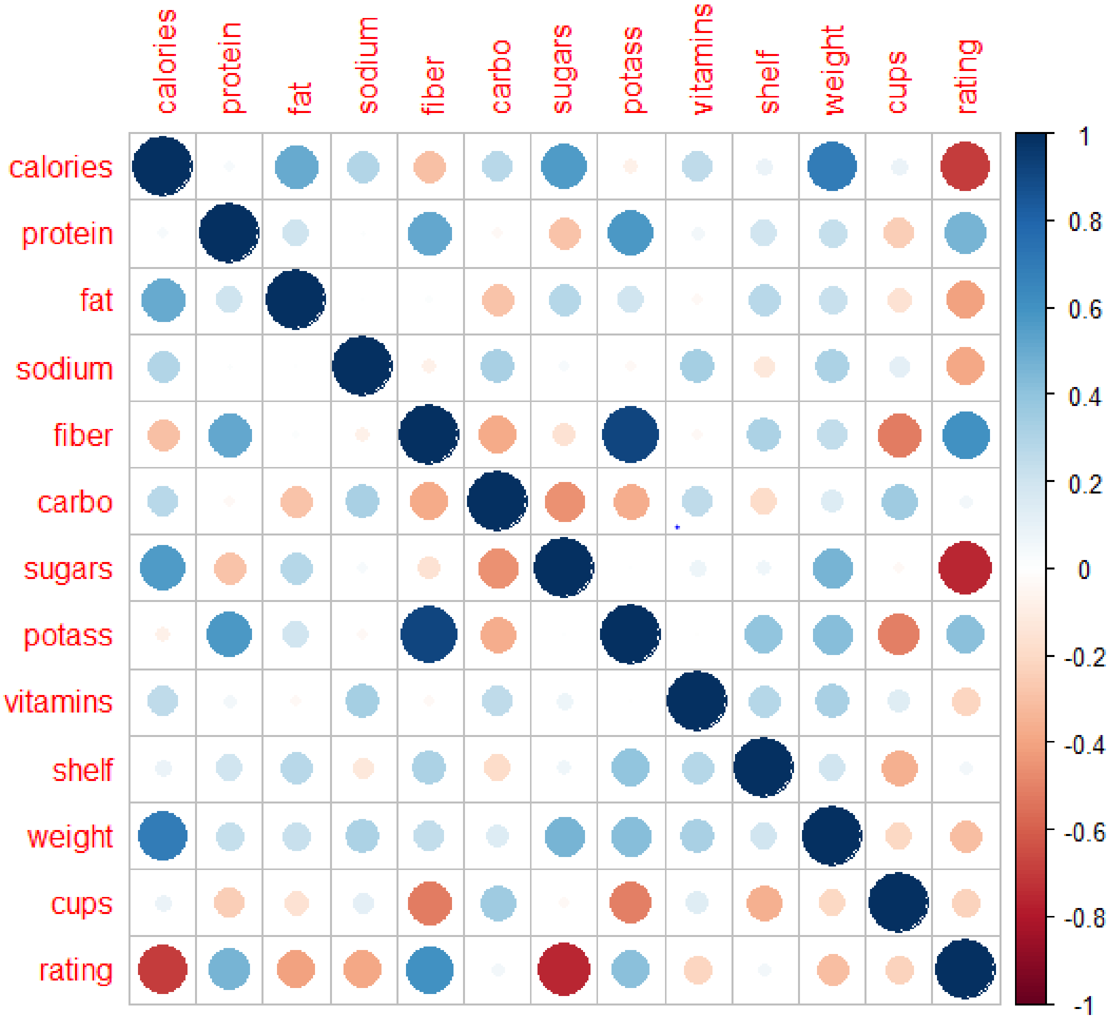


```
Call:
lm(formula = a$rating ~ calories + protein + fat + sodium + fiber +
    carbo + sugars + potass + vitamins, data = a)

Residuals:
       Min         1Q     Median         3Q        Max
-5.343e-07 -2.537e-07  3.961e-08  2.424e-07  5.513e-07

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept)  5.493e+01  2.794e-07  196559702   <2e-16 ***
calories    -2.227e-01  7.501e-09  -29694282   <2e-16 ***
protein      3.273e+00  5.551e-08   58964906   <2e-16 ***
fat         -1.691e+00  8.101e-08  -20877762   <2e-16 ***
sodium      -5.449e-02  4.910e-10 -110974232   <2e-16 ***
fiber        3.443e+00  4.756e-08   72399805   <2e-16 ***
carbo        1.092e+00  3.492e-08   31287364   <2e-16 ***
sugars      -7.249e-01  3.311e-08  -21895192   <2e-16 ***
potass      -3.399e-02  1.601e-09  -21228850   <2e-16 ***
vitamins    -5.121e-02  1.779e-09  -28778552   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.069e-07 on 64 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 1.696e+16 on 9 and 64 DF,  p-value: < 2.2e-16
```

In the above picture rating also plays an significant role in selling of products ,hence we can say that 100% of variance in rating can be explained by the variance in entities .

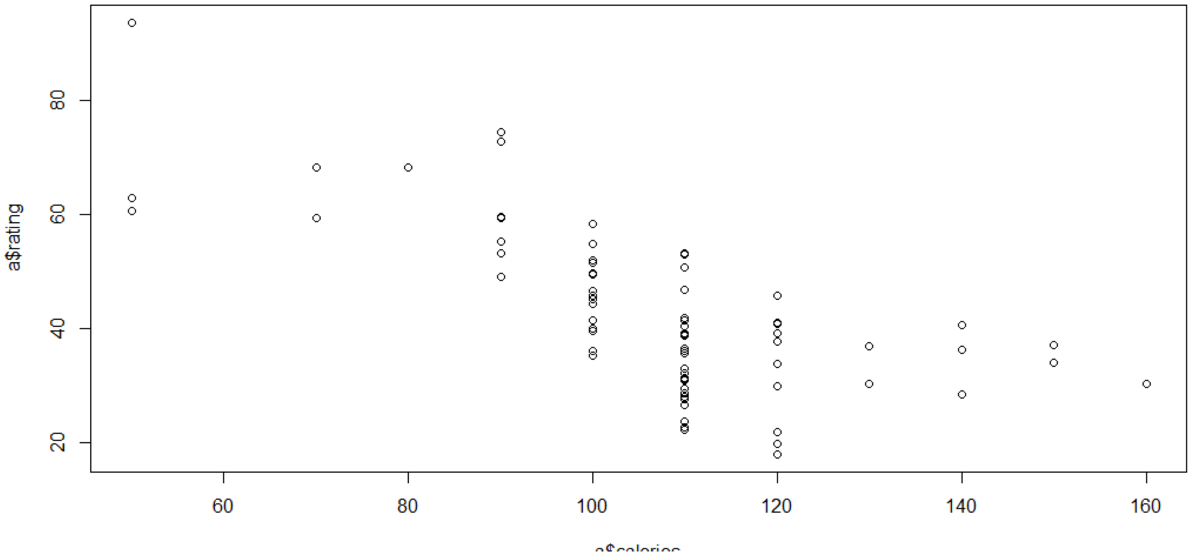
B data sets without (name, mfr,type)

FIG(1.2)



relation of rating and calories

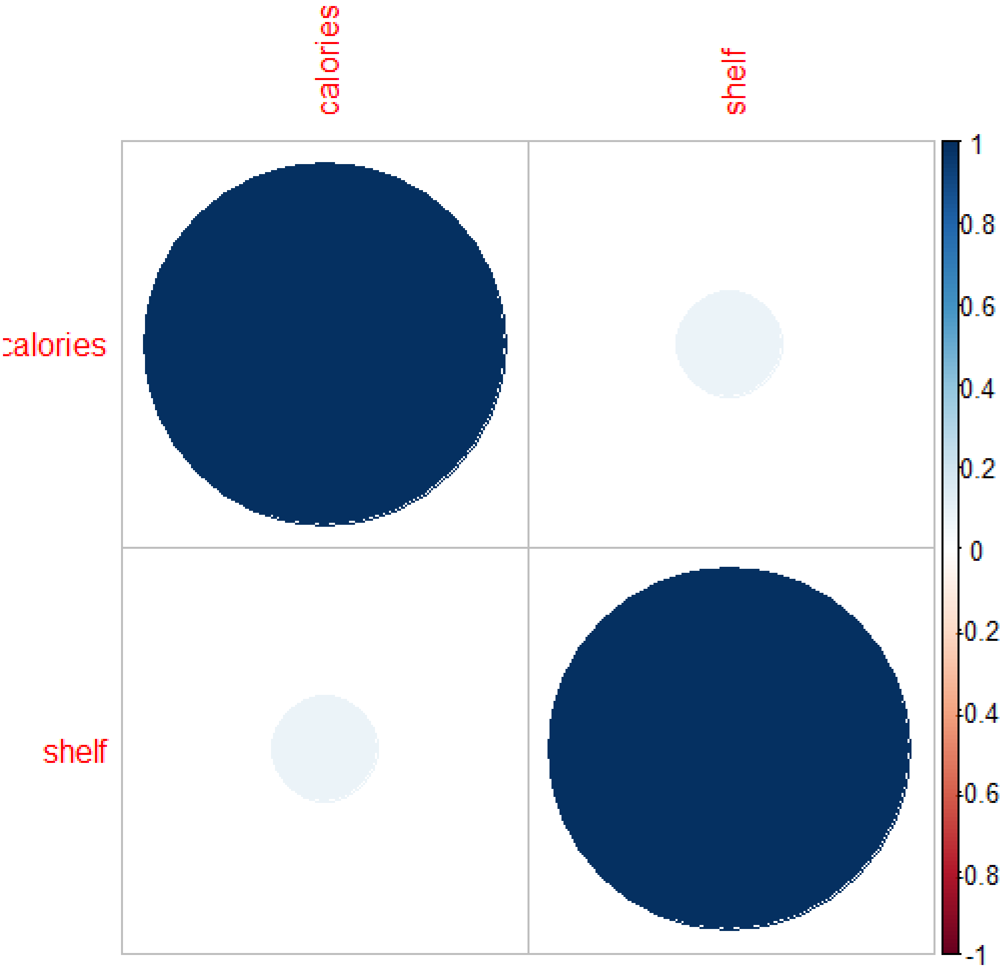
FIG=2.0



In fig=2.0 we can see that from calorie80-120 the rating are compact , which we can interpret that customers are buying the product which are with high calories which ranges from 80-120, so it is advised to keep the products which are high in calories which falls in a range of 80-120

Why placements of products are Important

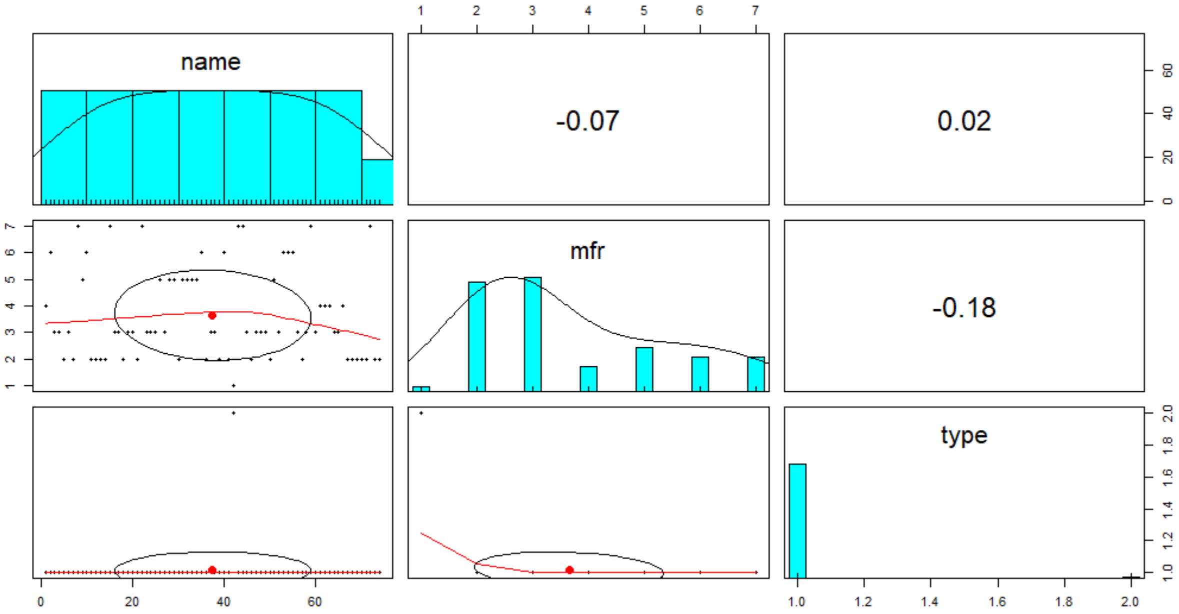
FIG=3.0



In the fig=3.0 it shows the correlation of the shelf and calories , where in the changes in placements of products in shelf brings changes in the products being sold, in the above the placement of products which might be 0.2 chances of being sold of other products as well.

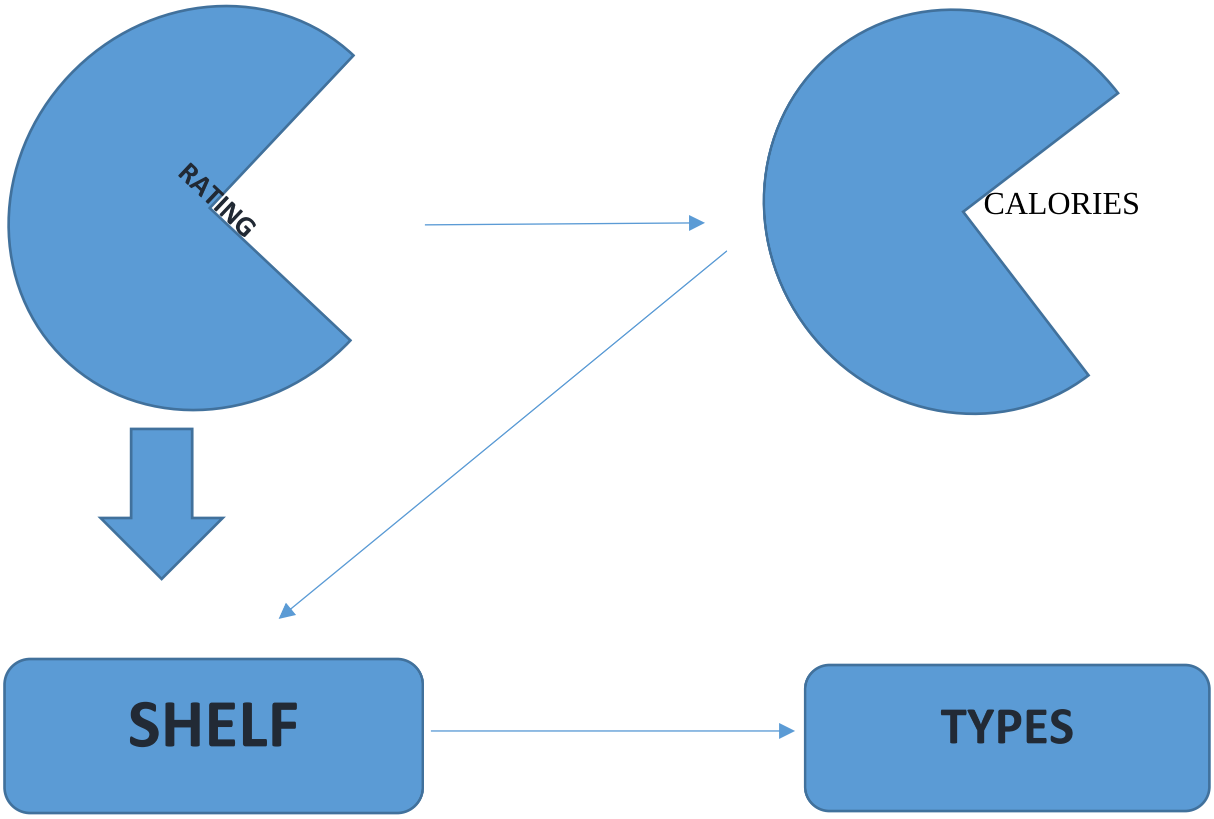
Does the name really matters for huge production of a store

Fig=3.1



In the above provided sample , the analysis show that the name of the product really doesn’t matters since the type has 0.02 relation , so it is advised that to keep the types which are *cold and hot* .

CONCLUSION FIG=4.0



*)The final conclusion is that , rating and calories of the product has more significant which has the chances of 70% of the products being sold,

*)Arranging the products which has more ratings and calories in the first shelf which gives chances of product being sold is 20%

*) where in the 2% of the people who likes different types which are cold and hot , so the chances of product being sold are of 1.5%

*) following all those instruction which are the chances of being sold are 91.5%,

So we can say that 85%-91.5% chances of the business development.