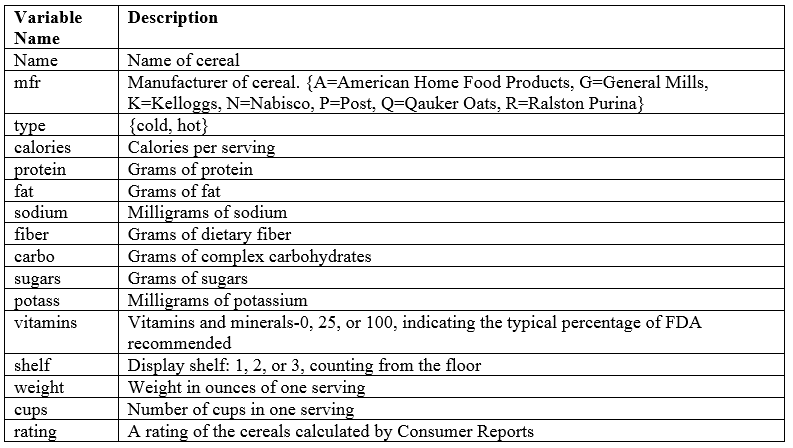
ITM 818 Data Management and Visualization in Analytics Homework 5: Data Pre-Processing and Dimensionality Reduction (60 Points)

In this homework, we will first clean a dataset and create some simple visualizations. Then, we will conduct a principal component analysis to reduce dimensionality. You must use R (NOT Tableau) to answer the questions. The data were collected on the nutritional information and consumer rating of 77 breakfast cereals. For each cereal we have included 13 numerical variables and 2 categorical variables (i.e. mfr and type).



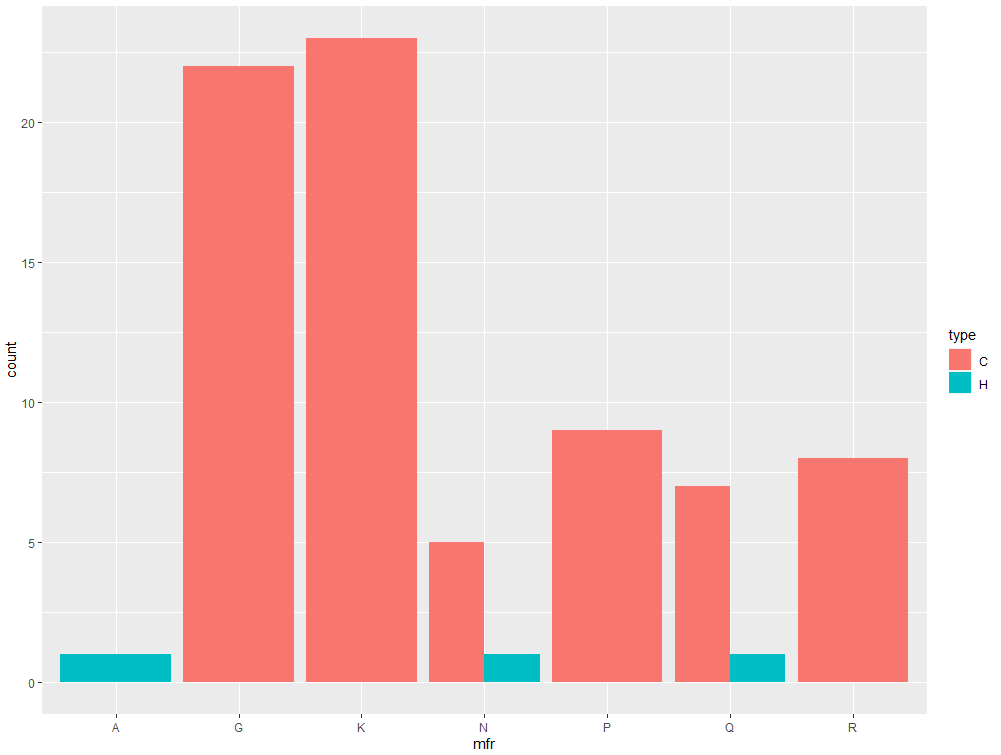
Please answer the following questions. Make sure that you include all your R code and screenshots in the submission.

1. Regarding two categorical variables (mfr and type), use a single proper visualization to show the distributions of both variables. (6 points)

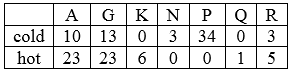
**Code and output:**

# Q1----------------------------------------------------------------------------------------

ggplot(data=Cereals)+geom\_bar(mapping=aes(x=mfr,fill=type),position="dodge")



2. Calculate the average consumer rating for each combination of manufacturer (mfr) and type. Obtain a data frame (table) that follows the structure below. Use “0” to represent missing value if there is no any cereal for a certain combination of manufacturer and type. Take a screenshot of the data frame. Example data frame (suppose there are average values, “0” means missing value): (6 points)



**Code and output:**

#Q2-----------------------------------------------------------------------------------------

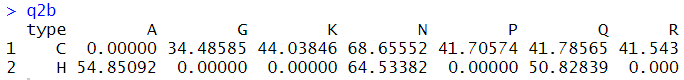
q2a = Cereals %>% group\_by(mfr,type) %>% summarise(mean = mean(rating))

q2b = spread(q2a,mfr,mean)

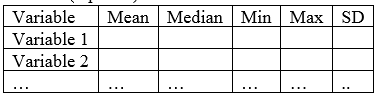
q2b[is.na(q2b)] <- 0

q2b = as.data.frame(q2b)

q2b



3. Focus on the numerical attributes and calculate the mean, median, min, max, and standard deviation for each of the numerical attributes. Store these summary statistics in a data frame having the structure below. Take a screenshot of the data frame. (6 points)



**Code and output:**

d3 = Cereals[,-c(1,2,3)]

d3 = as.data.frame(t(d3))

q3 = data.frame(Means=apply(d3, MARGIN = 1, FUN = mean,na.rm = TRUE),

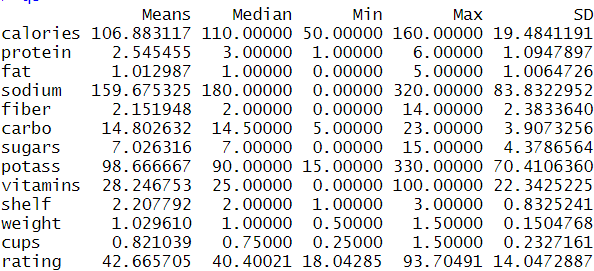
Median = apply(d3, MARGIN = 1, FUN = median,na.rm = TRUE),

Min = apply(d3, MARGIN = 1, FUN = min,na.rm = TRUE),

Max = apply(d3, MARGIN = 1, FUN = max,na.rm = TRUE),

SD = apply(d3, MARGIN = 1, FUN = sd,na.rm = TRUE))

q3



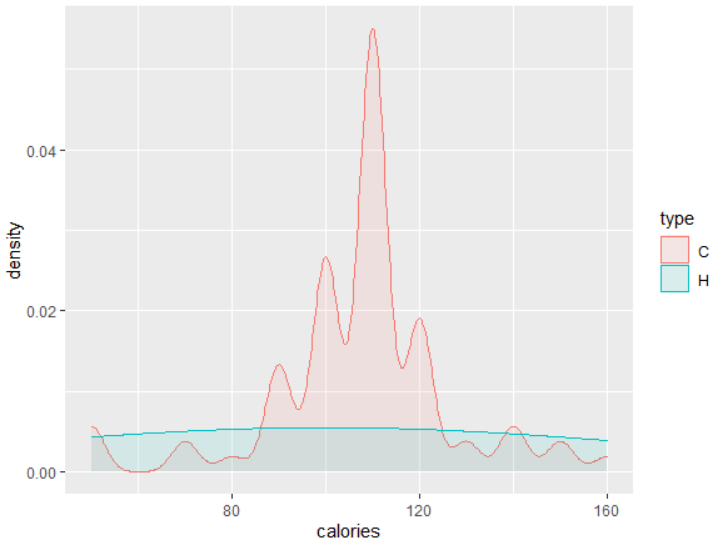
4. Use a proper visualization to compare the distribution of calories by hot vs. cold cereals. What does this plot show us regarding mean, variance, and skewness? (6 points)

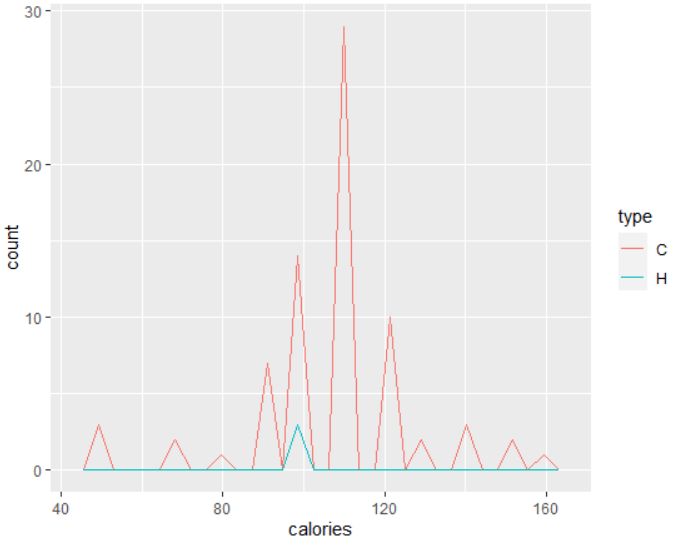
**Code and output:**

ggplot(data=Cereals)+

geom\_area(aes(x=calories,fill=type,color=type),stat="density",alpha=0.1,position="identity")

ggplot(data=Cereals)+geom\_freqpoly(aes(x=calories,color=type),position="identity")





Mean for both the types for Calories is around 105. More for cold compared to hot type.

Variance of calories is negligible for hot whereas is substantial for Cold type.

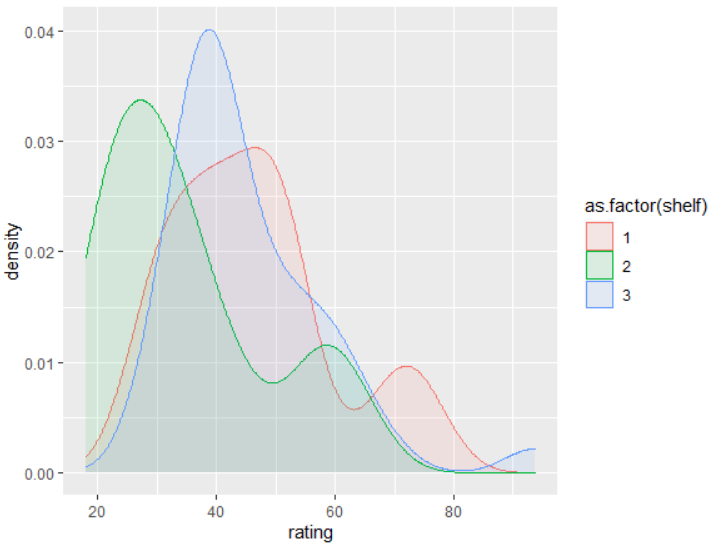
There seems to be no or negligible negative skewness for the cold type (almost normal distribution). No comment can be made for Hot type

5. Use a proper visualization to show the distribution of consumer rating by different shelf height. If we were to predict consumer rating from shelf height, does it appear that we need to keep all three categories of shelf height or we could combine two of the categories as a new category? (6 points)

**Code and output:**

ggplot(data=Cereals)+

geom\_area(aes(x=rating,fill=as.factor(shelf),color=as.factor(shelf)), stat="density",alpha=0.1,position="identity")



Though all 3 shelf distributions are somewhat different, we can combine shelf 1 and shelf 3 as their area under the curve is a lot common.

6. Get the correlation table for the numerical attributes. In addition, generate a scatter plot array for these variables. Find one pair of variables with strongest positive correlation and find one pair of variables with strongest negative correlation. (6 points)

**Code and output:**

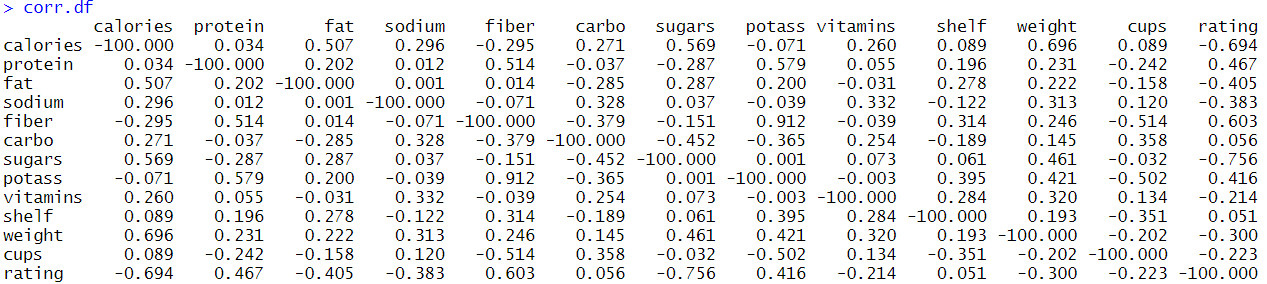
corr.df = round(cor(Cereals[,!colnames(Cereals) %in% c("name","mfr","type")],use = "complete.obs"),3)

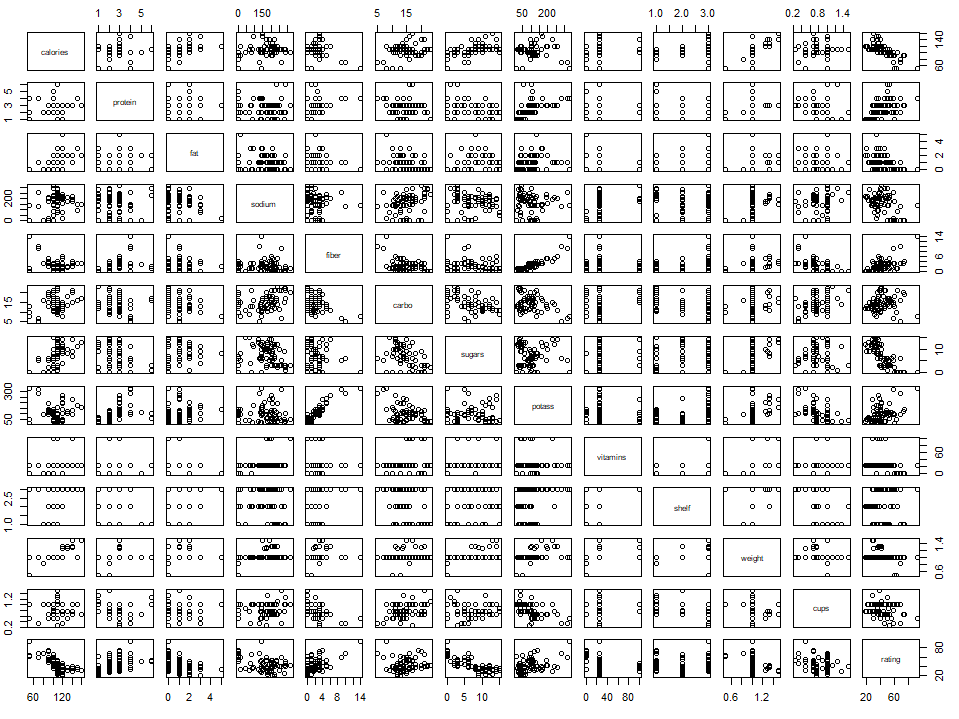
plot(Cereals[,!colnames(Cereals) %in% c("name","mfr","type")])

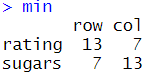
which(corr.df == min(corr.df), arr.ind = TRUE)

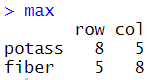
corr.df[corr.df == 1] = -100

which(corr.df == max(corr.df), arr.ind = TRUE)









7. Standardize the numerical attributes and then conduct the correlation analysis (as question 6). How would the correlations change if we standardize the data first? (6 points)

**Code and Output:**

d7 = Cereals[,-c(1,2,3)]

d7$calories = scale(d7$calories)

d7$protein = scale(d7$protein)

d7$fat = scale(d7$fat)

d7$sodium = scale(d7$sodium)

d7$fiber = scale(d7$fiber)

d7$carbo = scale(d7$carbo)

d7$sugars = scale(d7$sugars)

d7$potass = scale(d7$potass)

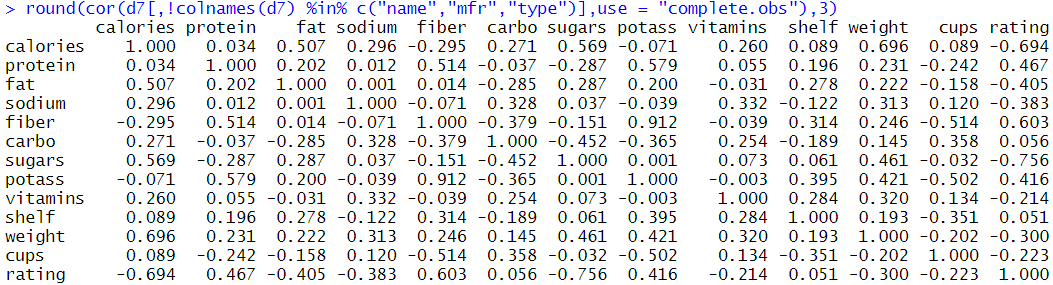
d7$vitamins = scale(d7$vitamins)

d7$weight = scale(d7$weight)

d7$cups = scale(d7$cups)

d7$rating = scale(d7$rating)

round(cor(d7[,!colnames(d7) %in% c("name","mfr","type")],use = "complete.obs"),3)

plot(d7[,!colnames(d7) %in% c("name","mfr","type")])

Even if we scale the data the relationship between 2 variables do not change.

8. Conduct a principal component analysis for the numerical variables except consumer rating to derive 4 principal components. Obtain the cumulative proportion of variance (of the numerical variables) captured by the 4 principal components? (6 points)

**Code and Output:**

d8 = Cereals

d8= na.omit(d8)

d8.pr <- prcomp(d8[c(4:15)], retx = TRUE, scale = TRUE)

summary(d8.pr)

screeplot(d8.pr, type = "l", npcs = 15, main = "Screeplot of the first 12 PCs")

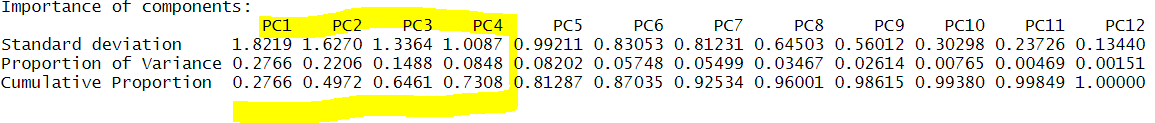
abline(h = 1, col="red", lty=5)

legend("topright", legend=c("Eigenvalue = 1"),

col=c("red"), lty=5, cex=0.6)

pca\_4 = d8.pr$x[,1:4]

pca\_4 = as.data.frame(pca\_4)



the cumulative proportion of variance (of the numerical variables) captured by the 4 principal components is **0.7308**

9. Create proper visualizations to show distributions of 4 principal components (one chart for each component). Which principal component has the largest variance? (6 points)

**Code and Output:**

p1 = ggplot(pca\_4, aes(x=PC1)) +

geom\_histogram(binwidth=.5, colour="black", fill="white")

p2 = ggplot(pca\_4, aes(x=PC2)) +

geom\_histogram(binwidth=.5, colour="black", fill="white")

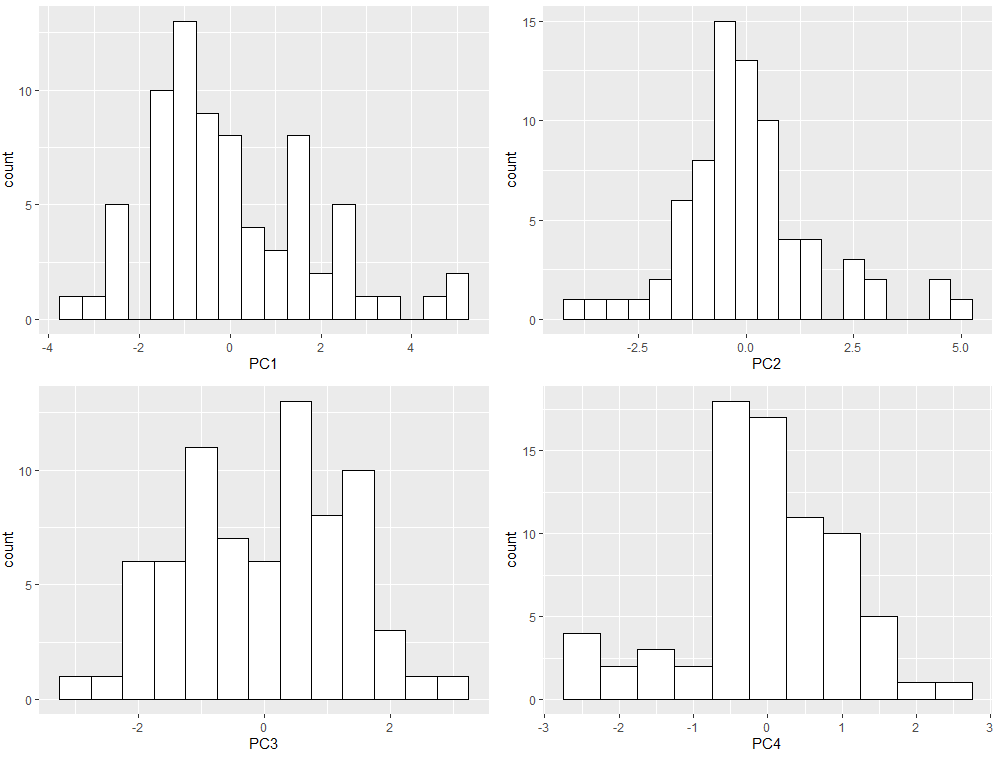
p3 = ggplot(pca\_4, aes(x=PC3)) +

geom\_histogram(binwidth=.5, colour="black", fill="white")

p4 = ggplot(pca\_4, aes(x=PC4)) +

geom\_histogram(binwidth=.5, colour="black", fill="white")

grid.arrange(p1,p2,p3,p4,ncol = 2)



principal component having the largest variance is **PC1**

10. Build scatter plots to show the relationships between 4 principal components and consumer rating (Note: you need to create 4 scatter plots. In each scatter plot, one principal component as X, and rating as Y). In each scatter plot, show the linear regression line with confidence intervals. Which principal component is mostly closely related to consumer rating? (6 points)

**Code and Output:**

d10=as.data.frame(cbind(PC1 = pca\_4$PC1, PC2=pca\_4$PC2, PC3=pca\_4$PC3,PC4 = pca\_4$PC4, Rate =d8$rating))

p1 = ggplot(data=d10,mapping=aes(x=PC1,y=Rate))+

geom\_point()+

geom\_smooth(method="lm")

p2 = ggplot(data=d10,mapping=aes(x=PC2,y=Rate))+

geom\_point()+

geom\_smooth(method="lm")

p3 = ggplot(data=d10,mapping=aes(x=PC3,y=Rate))+

geom\_point()+

geom\_smooth(method="lm")

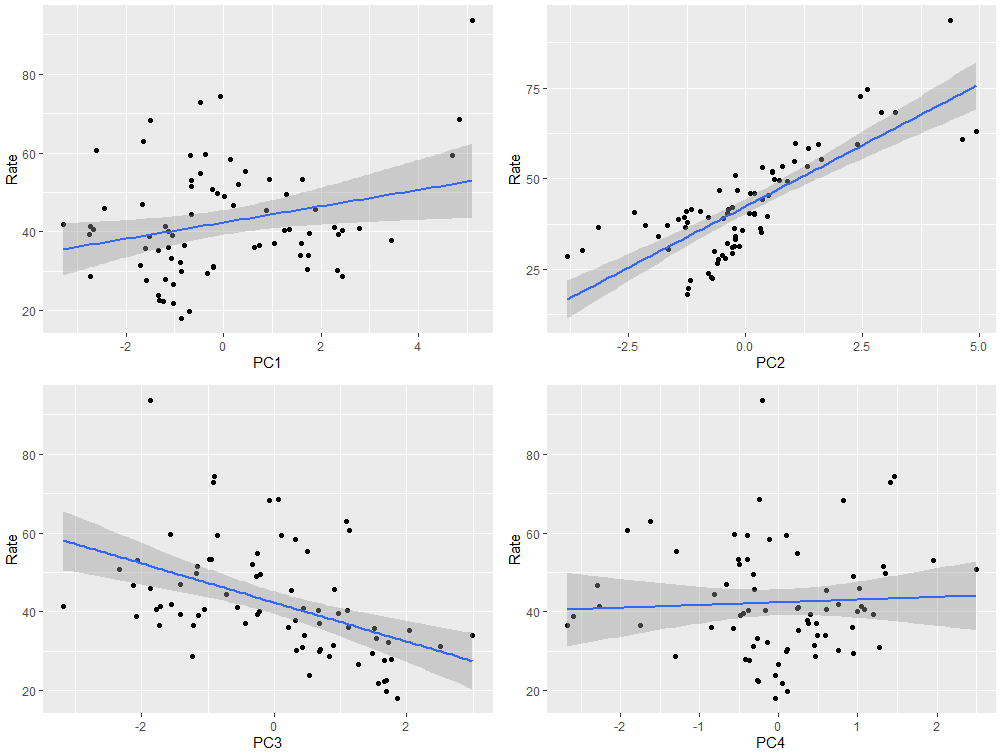
p4 = ggplot(data=d10,mapping=aes(x=PC4,y=Rate))+

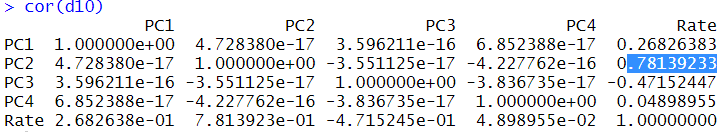
geom\_point()+

geom\_smooth(method="lm")

grid.arrange(p1,p2,p3,p4,ncol = 2)

cor(d10)





**principal component 2** is mostly closely related to consumer rating. We can check this visually and confirm even statistically.