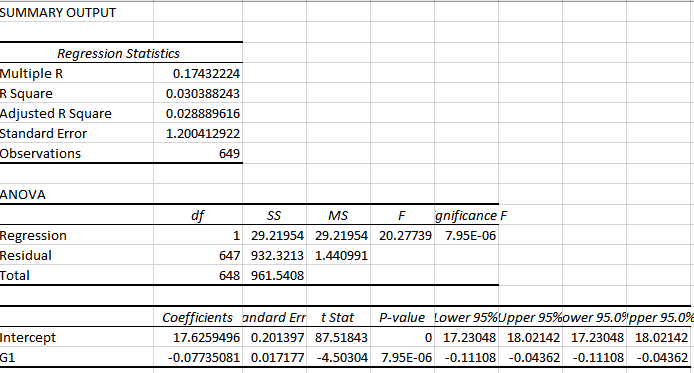
Now that you are familiar with correlation, regression, and ANOVA analysis, use R, Python, or Excel to determine which two or three attributes in your dataset are closely correlated.  Then answer the following questions on 1 single-spaced page:

1. How did you decide which two or three attributes to pick for your analysis?

Taking student grade on basis, and what best features relates to fluctuate the grade of the student is making us to identify the best three attributes in the dataset. For this we are identifying that DALC, G1, age are the best three attributes which we consider as the independent attributes because by using these attributes, it derives to a conclusion of the grade of the student how it is getting fluctuating. As a matter of fact, In Nut shell, it is been observed using common sense that student how is having age less than mandatory age required, it is effecting the grade value, absence of the student. Students those who are drinking on daily basis are related to Grade down grade in fact health of the student also decreases which means by using one attribute we can anticipate a relation with other attribute to identify the accurate results from the data. We identify the age as a primary attribute where it get inter relates to other attribute using common sense. Also not only using common sense, we can identify the relation using R Code and MS Excel to identify how best the attributes get related.   
  
  
Example:  
  
  


1. What correlation do the results of your analysis reveal?

Results derived for the above three Attributes, taking age as a class attribute, and comparing with Dalc and G1 results in this format.

R Code:   
  
chisq.test(GKR$age,GKR$Dalc)

cor(GKR$age,GKR$Dalc,use="complete.obs",method=c("pearson"))

reg1<- lm(GKR$age~GKR$Dalc)

plot(GKR$age~GKR$Dalc)

abline(reg1)

anova(reg1)

chisq.test(GKR$age,GKR$G1)

cor(GKR$age,GKR$G1,use="complete.obs",method=c("pearson"))

reg2<- lm(GKR$age~GKR$G1)

plot(GKR$age~GKR$G1)

abline(reg2)

anova(reg2)  
  
Output:  
  
> chisq.test(GKR$age,GKR$Dalc)

Pearson's Chi-squared test

data: GKR$age and GKR$Dalc

X-squared = 84.728, df = 28, p-value = 1.293e-07

Warning message:

In chisq.test(GKR$age, GKR$Dalc) :

Chi-squared approximation may be incorrect

> cor(GKR$age,GKR$Dalc,use="complete.obs",method=c("pearson"))

[1] 0.1347683

> reg1<- lm(GKR$age~GKR$Dalc)

> plot(GKR$age~GKR$Dalc)

> abline(reg1)

> anova(reg1)

Analysis of Variance Table

Response: GKR$age

Df Sum Sq Mean Sq F value Pr(>F)

GKR$Dalc 1 17.46 17.4640 11.969 0.0005766 \*\*\*

Residuals 647 944.08 1.4592

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> chisq.test(GKR$age,GKR$G1)

Pearson's Chi-squared test

data: GKR$age and GKR$G1

X-squared = 162.05, df = 112, p-value = 0.0014

Warning message:

In chisq.test(GKR$age, GKR$G1) : Chi-squared approximation may be incorrect

> cor(GKR$age,GKR$G1,use="complete.obs",method=c("pearson"))

[1] -0.1743222

> reg2<- lm(GKR$age~GKR$G1)

> plot(GKR$age~GKR$G1)

> abline(reg2)

> anova(reg2)

Analysis of Variance Table

Response: GKR$age

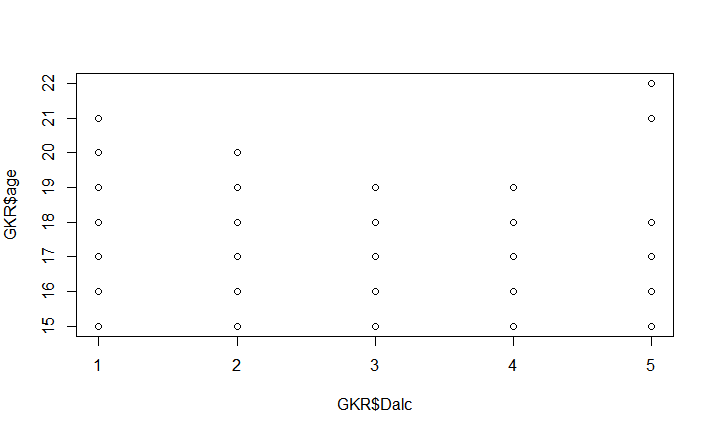
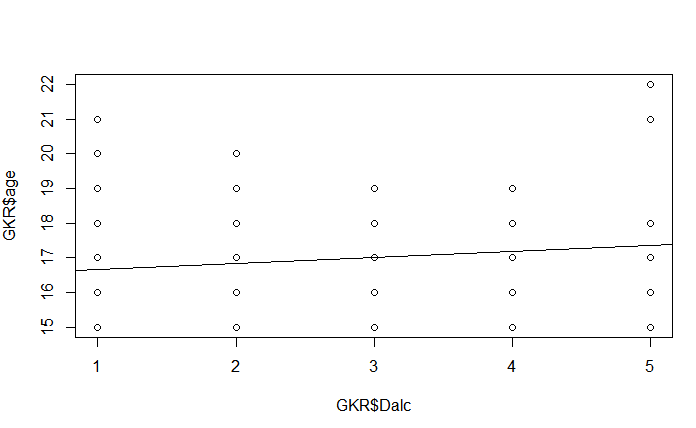
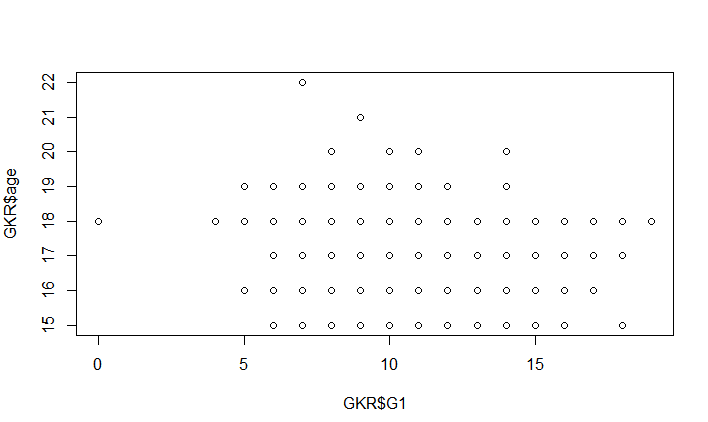
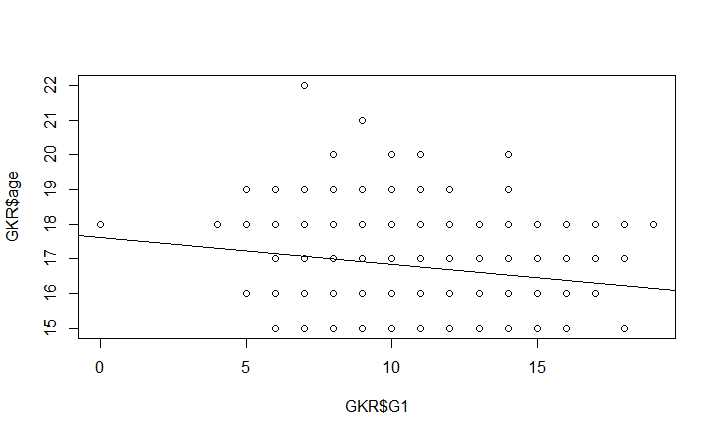
Df Sum Sq Mean Sq F value Pr(>F)

GKR$G1 1 29.22 29.220 20.277 7.946e-06 \*\*\*

Residuals 647 932.32 1.441

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

1. Were there any surprises?  Did you have to go back and pick different attributes and re-run your correlation exercise?  
     
   Yes, We expected first of all to have age, famrel, G1 to be related to each other as per common sense, and then we had tried to run the correlation and we identify with famrel to have a negative relation with each other so make it more near to defined results, so we started with p value to keep the anxiety to trust the data, and we had identified the value is more than 0.05 which made to work with Dalc, G1 and age to drive the results towards reducing the alcohol consumption and make a alcohol free society. Taking a new values we moved back taking new attributes but kept the class value as same of age because keeping age as restriction, would always help the society for alcohol free society.
2. Can you predict future values of one attribute from a regression on one or two others?

Include screenshots of your R or MS Excel analysis (including how you set up your regression) and of your results and include your R code and your results.

R Code:   
  
chisq.test(GKR$age,GKR$Dalc)

cor(GKR$age,GKR$Dalc,use="complete.obs",method=c("pearson"))

cor(GKR$famrel,GKR$Walc,use="complete.obs",method=c("pearson"))

reg1<- lm(GKR$age~GKR$Dalc)

plot(GKR$age~GKR$Dalc)

abline(reg1)

anova(reg1)

chisq.test(GKR$age,GKR$G1)

cor(GKR$age,GKR$G1,use="complete.obs",method=c("pearson"))

reg2<- lm(GKR$age~GKR$G1)

plot(GKR$age~GKR$G1)

abline(reg2)

anova(reg2)

chisq.test(GKR$age,GKR$Dalc)

cor(GKR$age,GKR$Dalc,use="complete.obs",method=c("pearson"))

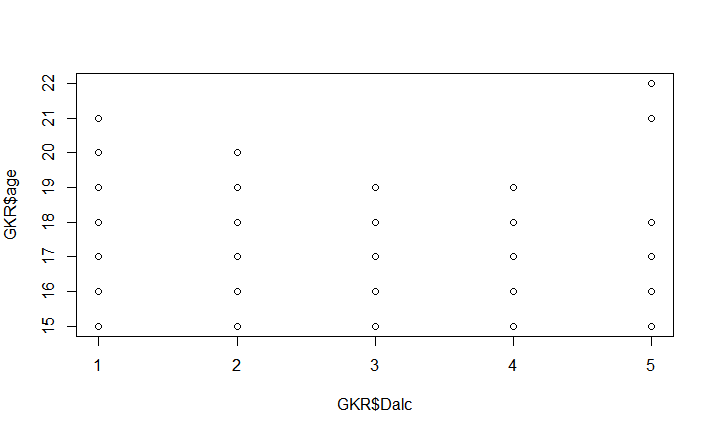
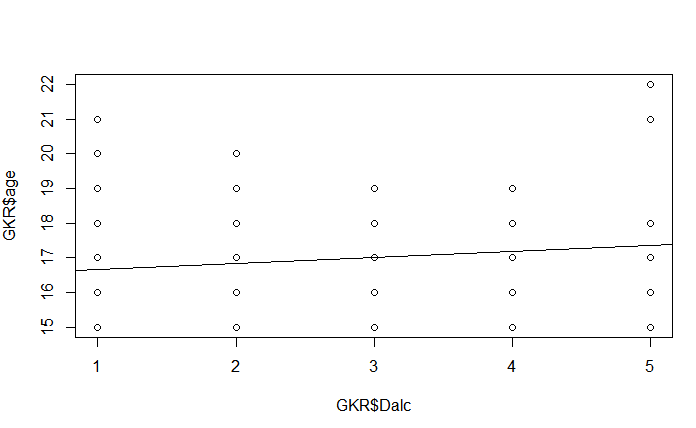
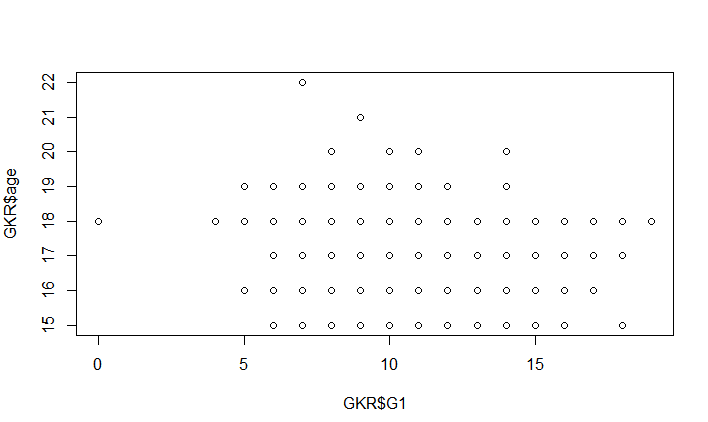
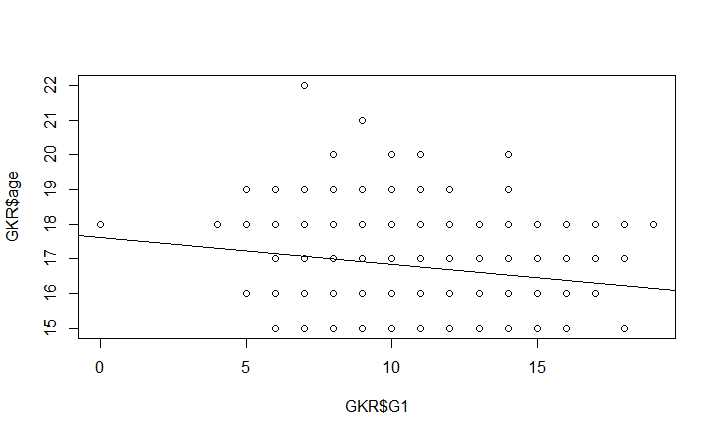
reg3<- lm(GKR$age~GKR$Dalc)

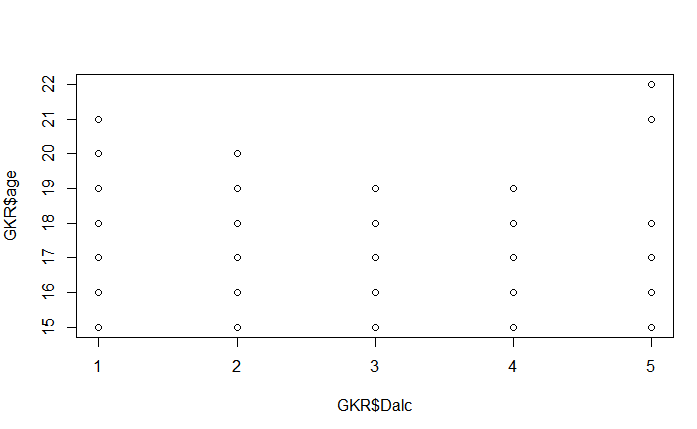
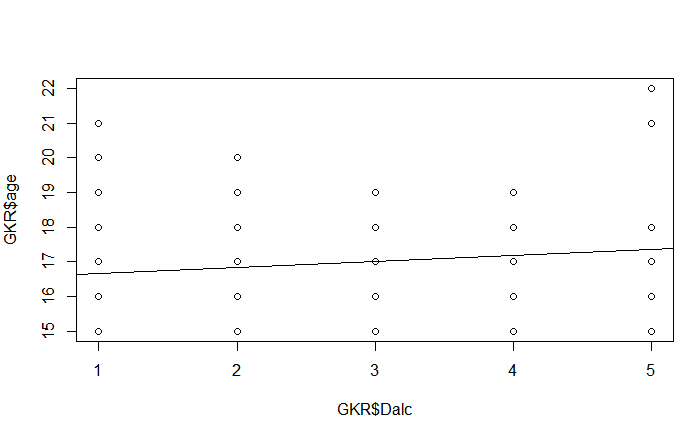
plot(GKR$age~GKR$Dalc)

abline(reg3)

anova(reg3)  
  
  
Output:

|  |
| --- |
| > cor(GKR$age,GKR$Dalc,use="complete.obs",method=c("pearson"))  [1] 0.1347683  > chisq.test(GKR$age,GKR$Dalc)  Pearson's Chi-squared test  data: GKR$age and GKR$Dalc  X-squared = 84.728, df = 28, p-value = 1.293e-07  Warning message:  In chisq.test(GKR$age, GKR$Dalc) :  Chi-squared approximation may be incorrect  > cor(GKR$age,GKR$Dalc,use="complete.obs",method=c("pearson"))  [1] 0.1347683  > cor(GKR$famrel,GKR$Walc,use="complete.obs",method=c("pearson"))  [1] -0.09351081  > reg1<- lm(GKR$age~GKR$Dalc)  > plot(GKR$age~GKR$Dalc)  > abline(reg1)  > anova(reg1)  Analysis of Variance Table  Response: GKR$age  Df Sum Sq Mean Sq F value Pr(>F)  GKR$Dalc 1 17.46 17.4640 11.969 0.0005766 \*\*\*  Residuals 647 944.08 1.4592  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  > chisq.test(GKR$age,GKR$G1)  Pearson's Chi-squared test  data: GKR$age and GKR$G1  X-squared = 162.05, df = 112, p-value = 0.0014  Warning message:  In chisq.test(GKR$age, GKR$G1) : Chi-squared approximation may be incorrect  > cor(GKR$age,GKR$G1,use="complete.obs",method=c("pearson"))  [1] -0.1743222  > reg2<- lm(GKR$age~GKR$G1)  > plot(GKR$age~GKR$G1)  > abline(reg2)  > anova(reg2)  Analysis of Variance Table  Response: GKR$age  Df Sum Sq Mean Sq F value Pr(>F)  GKR$G1 1 29.22 29.220 20.277 7.946e-06 \*\*\*  Residuals 647 932.32 1.441  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  > chisq.test(GKR$age,GKR$Dalc)  Pearson's Chi-squared test  data: GKR$age and GKR$Dalc  X-squared = 84.728, df = 28, p-value = 1.293e-07  Warning message:  In chisq.test(GKR$age, GKR$Dalc) :  Chi-squared approximation may be incorrect  > cor(GKR$age,GKR$Dalc,use="complete.obs",method=c("pearson"))  [1] 0.1347683  > reg3<- lm(GKR$age~GKR$Dalc)  > plot(GKR$age~GKR$Dalc)  > abline(reg3) |
|  |
| |  | | --- | | > | |

  
  
  
Taking it with above specified results, we can identify that as on the graph display that as on the value as on the both the value increases, where they become directly proportional to each other and the value identifies they are dependent with each other with class attribute, and in order to disprove H0 and make an alternative hypothesis to make and near to results derived, and keep the confidence value of the data to be 95% , and the variance of the data should be less 5% which resembles with less than 0.05 value.   
  
In order to predict the future values in the selected attributes , where each attribute would be considered as interdependent in this data set, we consider the Df Sum Sq Mean Sq F value Pr(>F) values in consider where it can identify the trust of the data if the p value is than 0.05 , and we can have r square value to be used to predict the future values.  
  
  
Taking the R square value of the desired attributes we can have future value can be predicted.

R Code:  
  
anova(reg1)  
anova(reg2)  
anova(reg3)  
  
Output:

|  |
| --- |
| > anova(reg1)  Analysis of Variance Table  Response: GKR$age  Df Sum Sq Mean Sq F value Pr(>F)  GKR$Dalc 1 17.46 17.4640 11.969 0.0005766 \*\*\*  Residuals 647 944.08 1.4592  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  > anova(reg2)  Analysis of Variance Table  Response: GKR$age  Df Sum Sq Mean Sq F value Pr(>F)  GKR$G1 1 29.22 29.220 20.277 7.946e-06 \*\*\*  Residuals 647 932.32 1.441  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  > anova(reg3)  Analysis of Variance Table  Response: GKR$age  Df Sum Sq Mean Sq F value Pr(>F)  GKR$Dalc 1 17.46 17.4640 11.969 0.0005766 \*\*\*  Residuals 647 944.08 1.4592  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| |  | | --- | | > | |

Now it is the time of the manager to identify whether to take the prediction in consideration or not as there would be a situation that it might go over the expectation, nor it might go down.

|  |
| --- |
| > reg1  Call:  lm(formula = GKR$age ~ GKR$Dalc)  Coefficients:  (Intercept) GKR$Dalc  16.4775 0.1775  > reg2  Call:  lm(formula = GKR$age ~ GKR$G1)  Coefficients:  (Intercept) GKR$G1  17.62595 -0.07735  > reg3  Call:  lm(formula = GKR$age ~ GKR$Dalc)  Coefficients:  (Intercept) GKR$Dalc  16.4775 0.1775 |
|  |
| |  | | --- | | > | |