STA 106 Project 2

Han Bao, Junxi Chen

Professor: Maxime Pouokam

Hope you enjoy this report!!

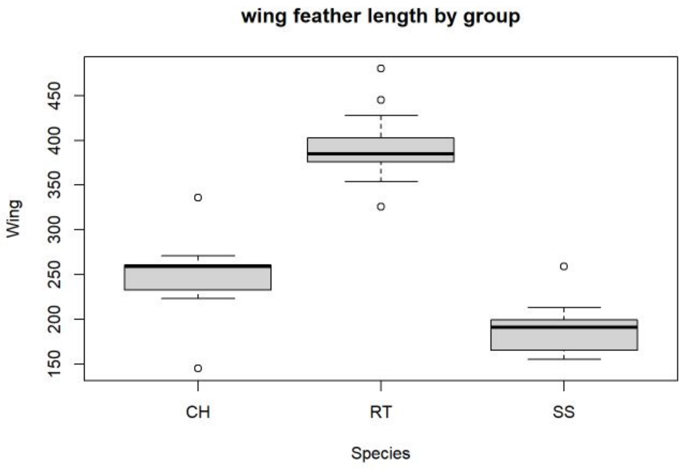
Part1 : Report for Topic 1

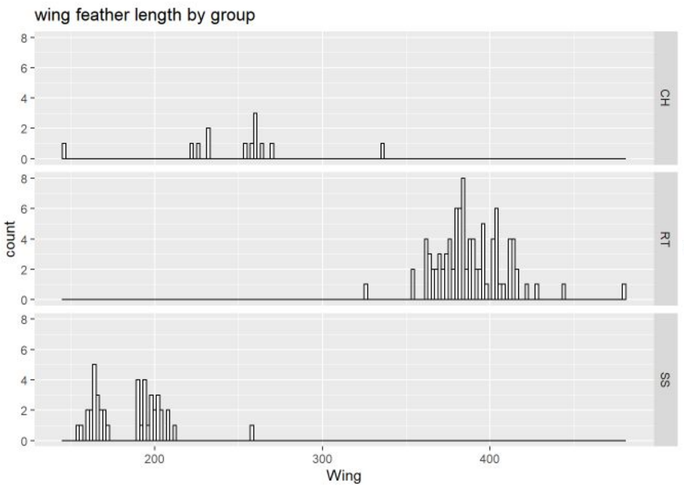
Topic1: Q2

A:Introduction:

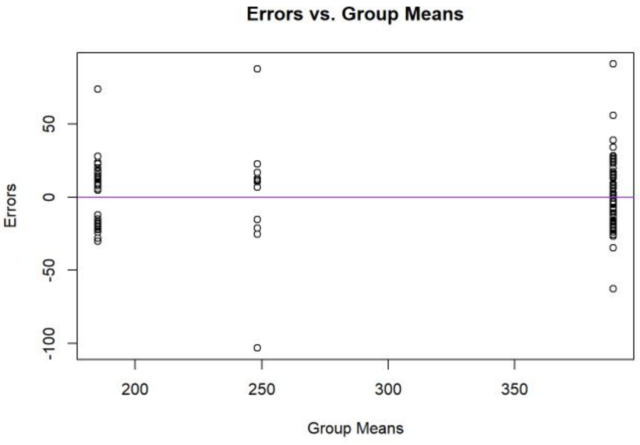
This Hawk data is related to the wings length and different species (Cooper’s, Red-tailed, and Sharp-Shinned) which are two columns.

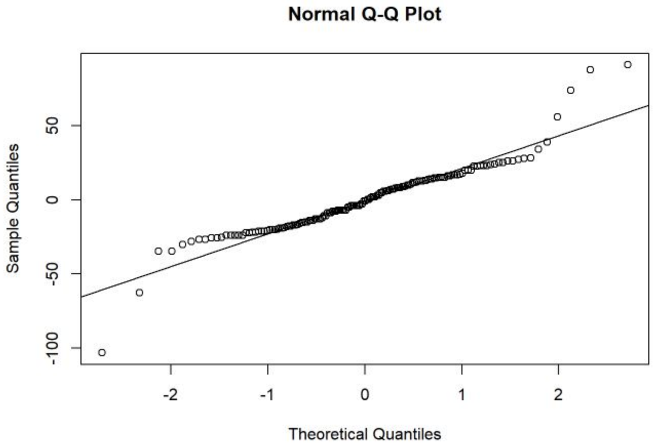
B: Plot and Interpretation

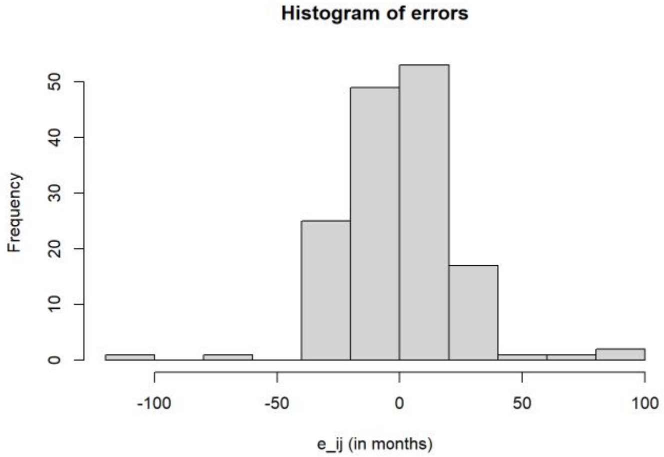

wing feather length by group


wing feather length by group

Plots

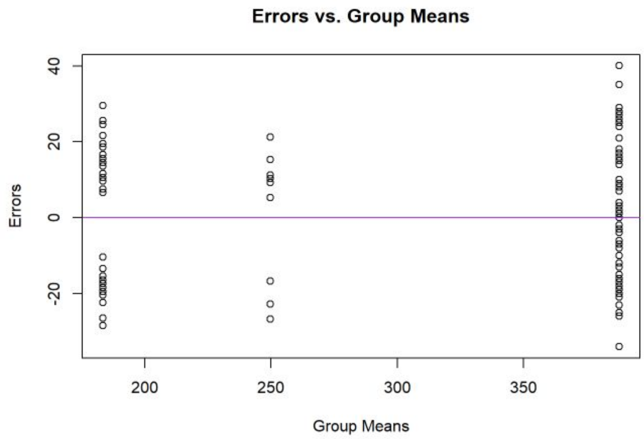

Errors vs. Group Means

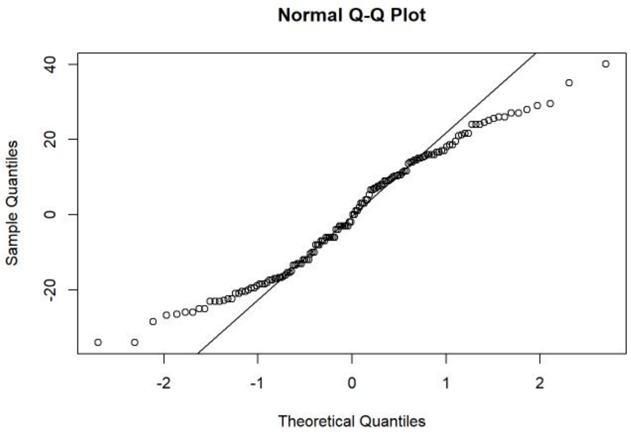

Normal Q-Q Plot

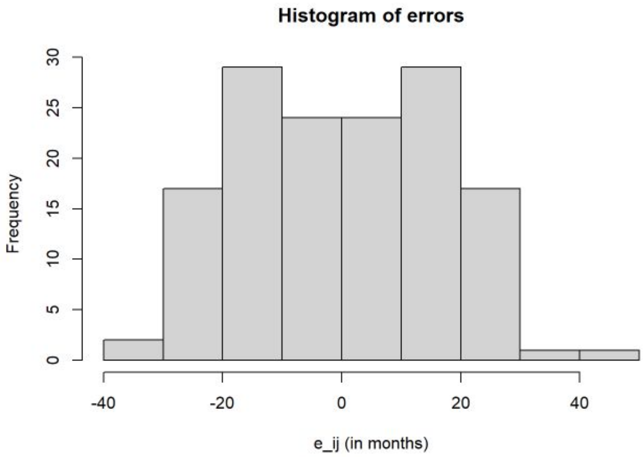

Histogram of errors

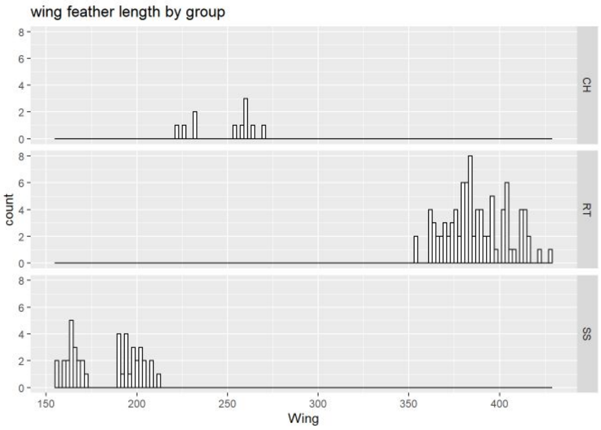
As we can see from the QQ plot of this data, there are some outliers included. In this way, before we transform our data, this is not normal distribution. And because of those outliers, there are also not equal variance. Besides those diagnostic plots, we also can see from the distribution of the data in boxplot and histogram. We also can realize that there are outliers, and the distribution for different groups are hugely different from each other.

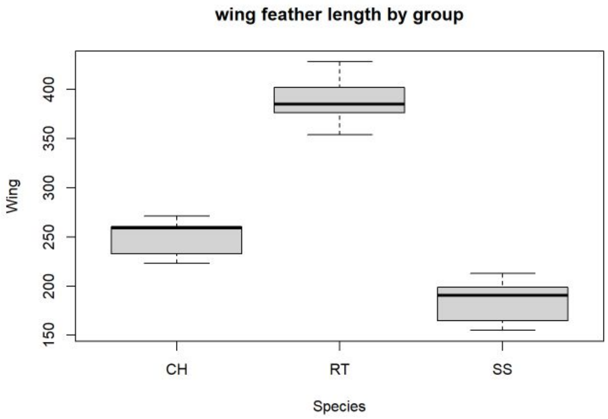
C. Plots











After we transform our data, we can see the outliers are removed, and the distribution for each group becomes more normal. And the error distribution also distributed versus the group mean becomes more symmetrical. And the vertical spread appears to be closer.

D: discuss

The transformation helps. It helps our data to be more normalized and remove the outliers. But it does have some downsides. iit makes interpretation, particularly of sample means, very difficult. Since we removed the outliers, the variance will be equal to each other. And the distribution will be more normal. Besides, we can see from those plots, the distribution of QQ plots is closer to the fit line, and the spread of error vs group mean is closer to each other. so we believe the transformed data will be better. And it satisfies the requirement of the ANOVA, the equal variance. Furthermore, I will also suggest clients to transform the data. This is because if a data set does not have the same variance, it does not meet the requirements to use ANOVA.

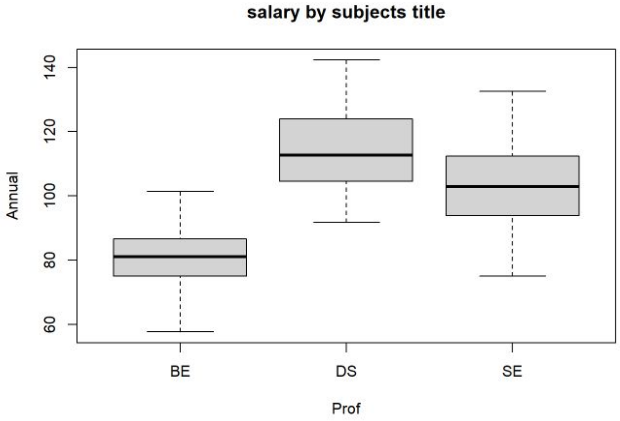
Part2 : Report for Topic 2

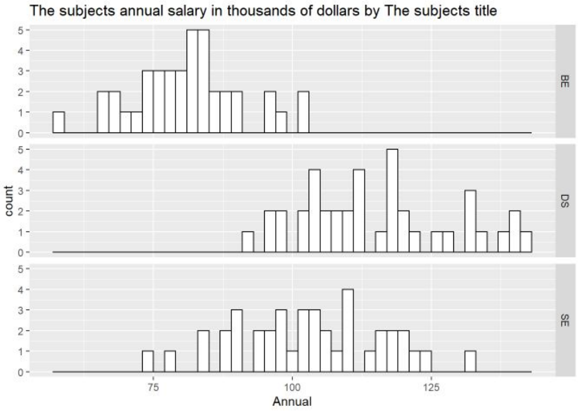
Topic2: Q1

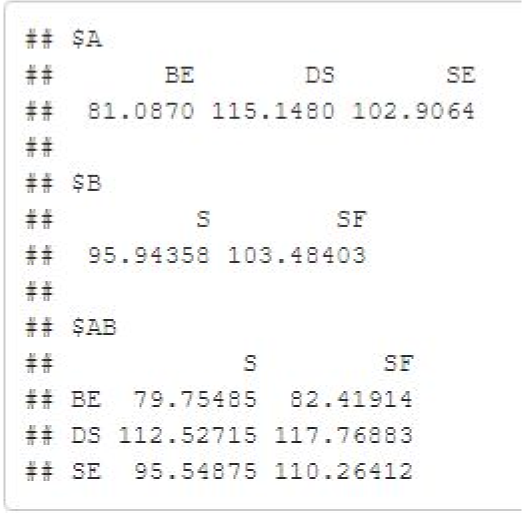
A. Introduction:

For this question, we are trying to test whether the interaction exists between each group or not. This is because we have the data from the annual salary for the different professors(Data Scientist),("Software Engineer"), and("Bioinformatics Engineer") with the different areas(San Francisco and Seattle). Hence, we want to determine the different areas and different professors will have some effects on the salary or not. And we will use this Two factor ANOVA to determine this interaction.

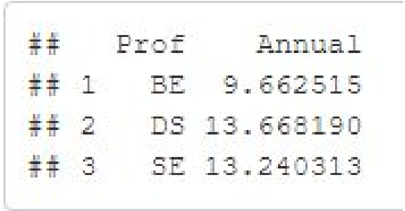
B. summary of data







```
## $A
##       BE       DS       SE
##  81.0870 115.1480 102.9064
##
## $B
##         S        SF
##  95.94358 103.48403
##
## $AB
##            S        SF
## BE  79.75485  82.41914
## DS 112.52715 117.76883
## SE  95.54875 110.26412
```



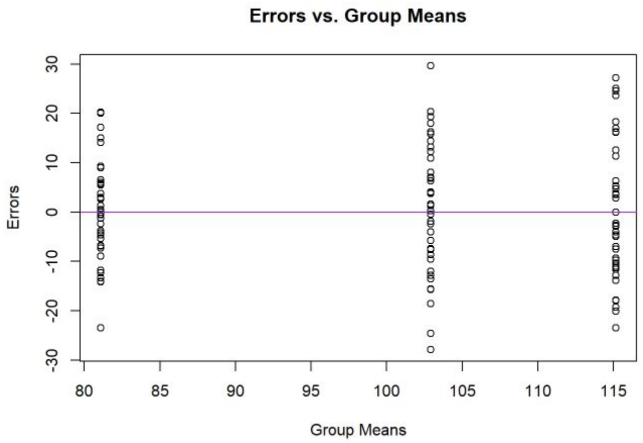
```
##   Prof    Annual
## 1   BE  9.662515
## 2   DS 13.668190
## 3   SE 13.240313
```

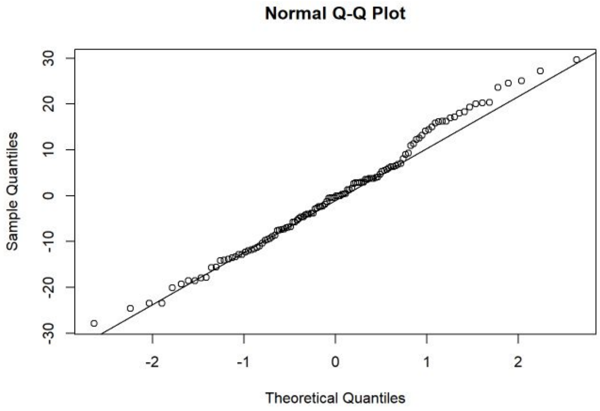
Sample means and Standard deviations for groups.

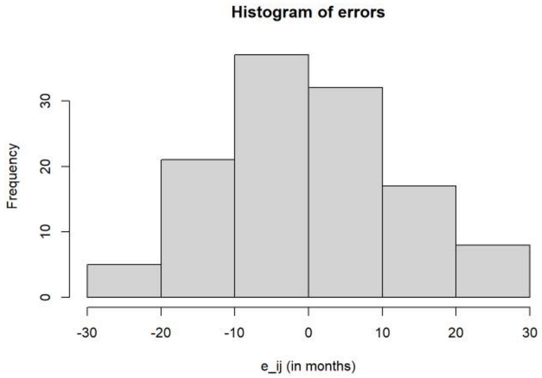
The salary of BE in two areas are mainly located in a low range with low level compared to other two professors. This mainly distributed in 60 to 100, with mean of 81.087. For salary in DS, this does have a high level compared to other two types of professors. This has a mean of 115.148, and SE has a mean of 102.9064. Hence, we can see from those plots that BE's salary is mainly located in median and right, DS's salary mainly located in median and left, adn SE's salary mainly distributed evenly.

C. Diagnostics

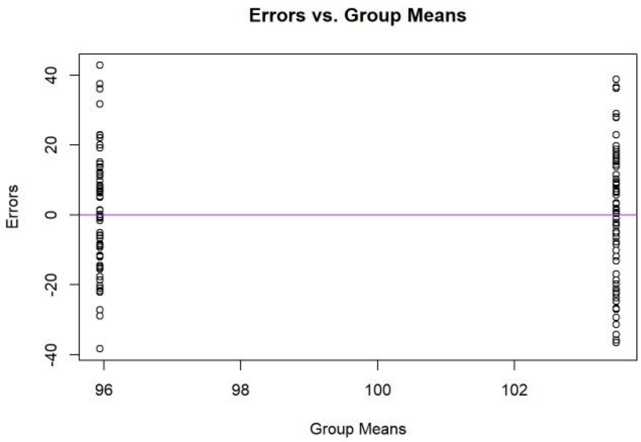
The relationship between salary and Prof

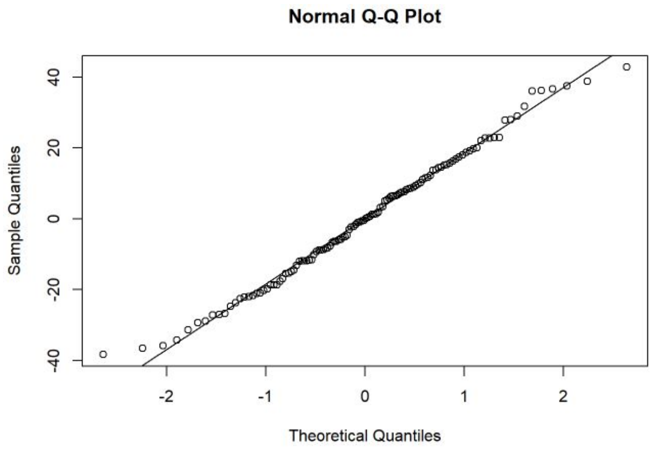


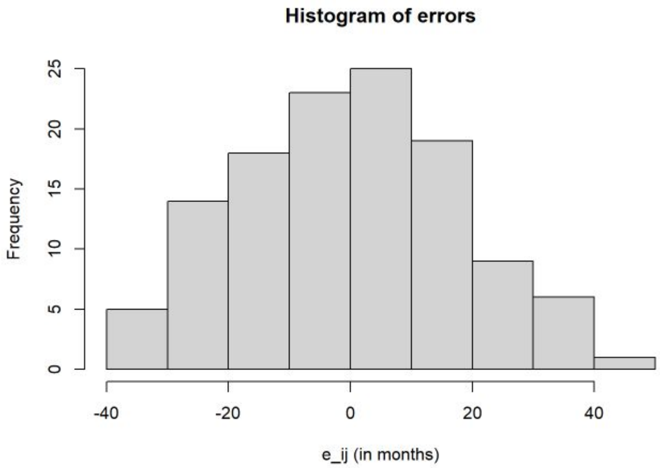




The relationship between Salary and Region



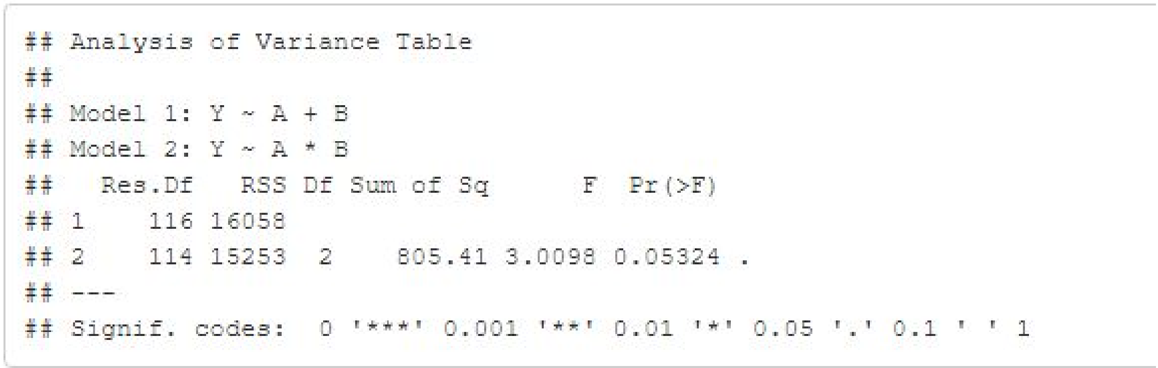




According to our diagnostic plots, we can see that there are no outliers in those data for comparing salary with professor or salary compared with different areas.

D. Analysis

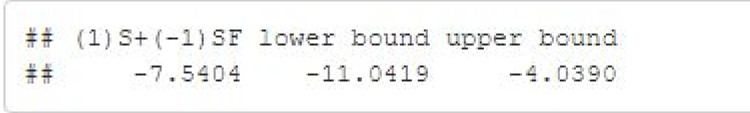
H0: The model with no interactions is a statistically better fit than the one with interactions (all (γδ)ij = 0)
HA: The model with no interactions is not a statistically better fit than the one with interactions (at least one (γδ)ij not equal 0)



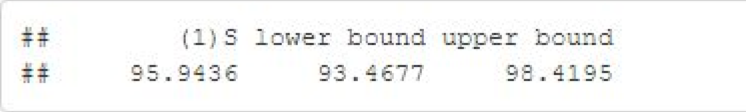
```
## Analysis of Variance Table
##
## Model 1: Y ~ A + B
## Model 2: Y ~ A * B
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    116 16058
## 2    114 15253  2    805.41 3.0098 0.05324 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the table, we can see that The test-statistic is: 3.0098, and the p-value is 0.05324. And we assume the alpha is 0.1(90% confidence). Because of this, we can determine that because our p value is less than the alpha.
And we choose 6 intervals by using tukey and scheffe to do the pairwise and contrast. For the first one that we calculated

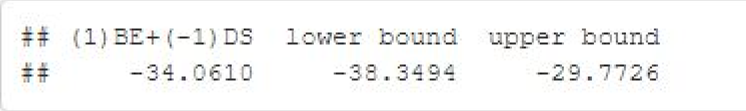
We are 90% confident that given all different professors, the true average subjects annual salary in the San Francisco region is less from 4.0390 to 11.0419 of Seattle.
We are 90% confident that the true average salary of professors in San Francisco, given all three possible professors, is between 93.467 and 98.4195.
We are 90% confident that given those two possible areas, the true average annual salary for BE is less 29.7726 to 38.3494 compared to DS.
We are 90% confident that given those two possible areas, the true average annual salary for DS is larger 7.9532 to 16.5299 compared to SE.



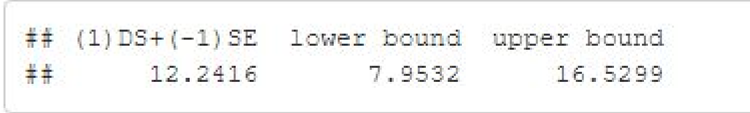
```
## (1)S+(-1)SF lower bound upper bound
##     -7.5404    -11.0419     -4.0390
```



```
##        (1)S lower bound upper bound
##     95.9436     93.4677     98.4195
```



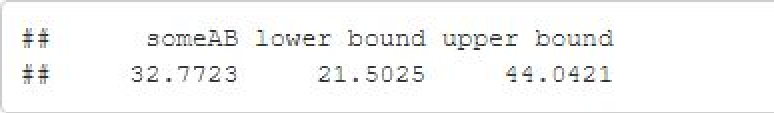
```
## (1)BE+(-1)DS  lower bound  upper bound
##     -34.0610     -38.3494     -29.7726
```



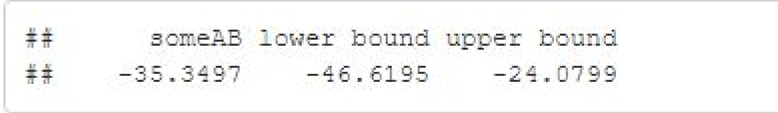
```
## (1)DS+(-1)SE  lower bound  upper bound
##      12.2416       7.9532      16.5299
```

We are 90% confident that the true average annual salary for SE in SF is larger than that of DS in SF by between 21.5025 to 44.0421 thousands of dollars

We are 90% confident that the true average annual salary for DS in S is less than that of SE in S by between 24.0799 to 46.6195 thousands of dollars.



```
##      someAB lower bound upper bound
##     32.7723     21.5025     44.0421
```



```
##      someAB lower bound upper bound
##    -35.3497    -46.6195    -24.0799
```

E. Interpretation:

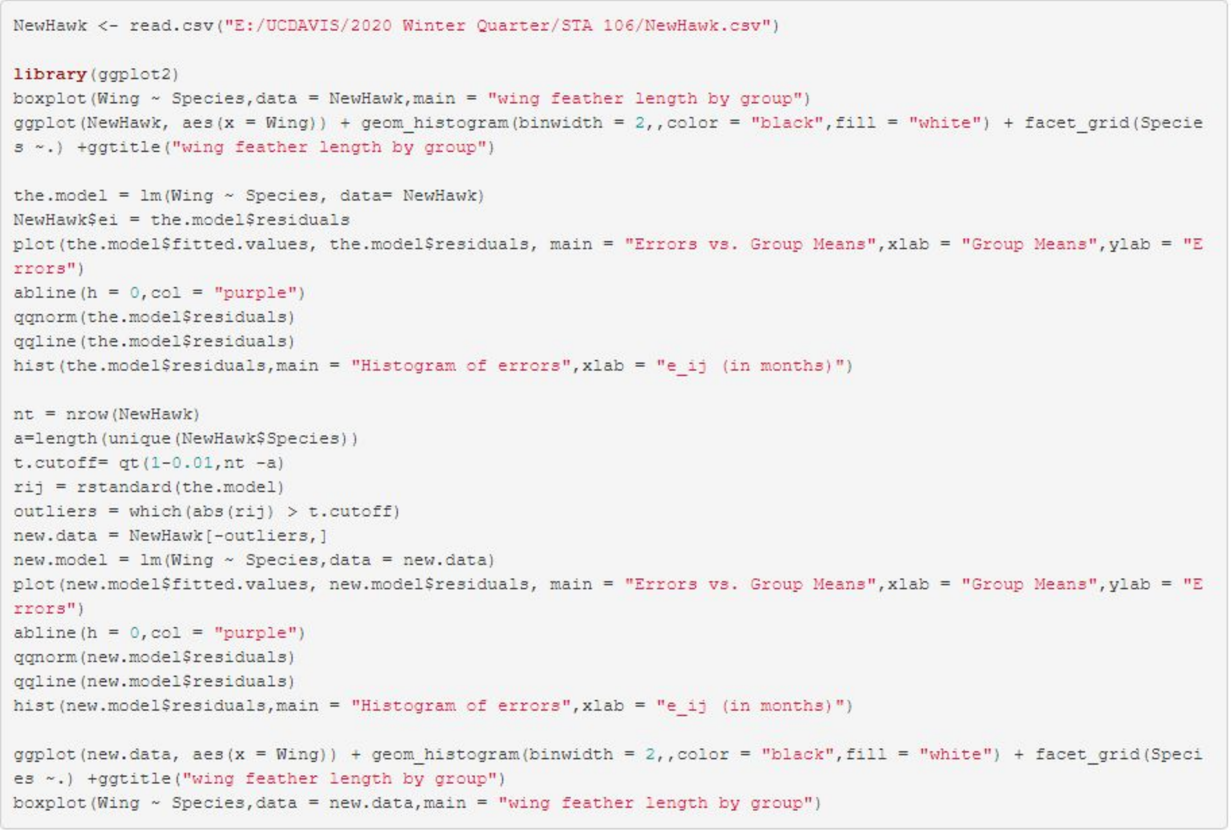
we can determine that because our p value is less than the alpha. So, we will reject the null hypothesis and conclude that there are some interaction effects between those variables. And from the confidence interval that we calculated for those different variables, we can see that none of them include zero, which means those all have significant different with each other. Hence, the interaction effect exists, which can also match and suggest our conclusion.

F:Conclusion:

Because of our test's p-value (0.05324) less than the alpha we decided (0.1), we conclude that an interaction term between the professor's subjects title and the region significantly improves the model fit compared to a model with only factor effects. Also, from our confidence intervals calculations, we found that from every interval, they do not contain zero, which stated that all have significant differences with each other between professors and regions. As a result, the interaction effects exist.

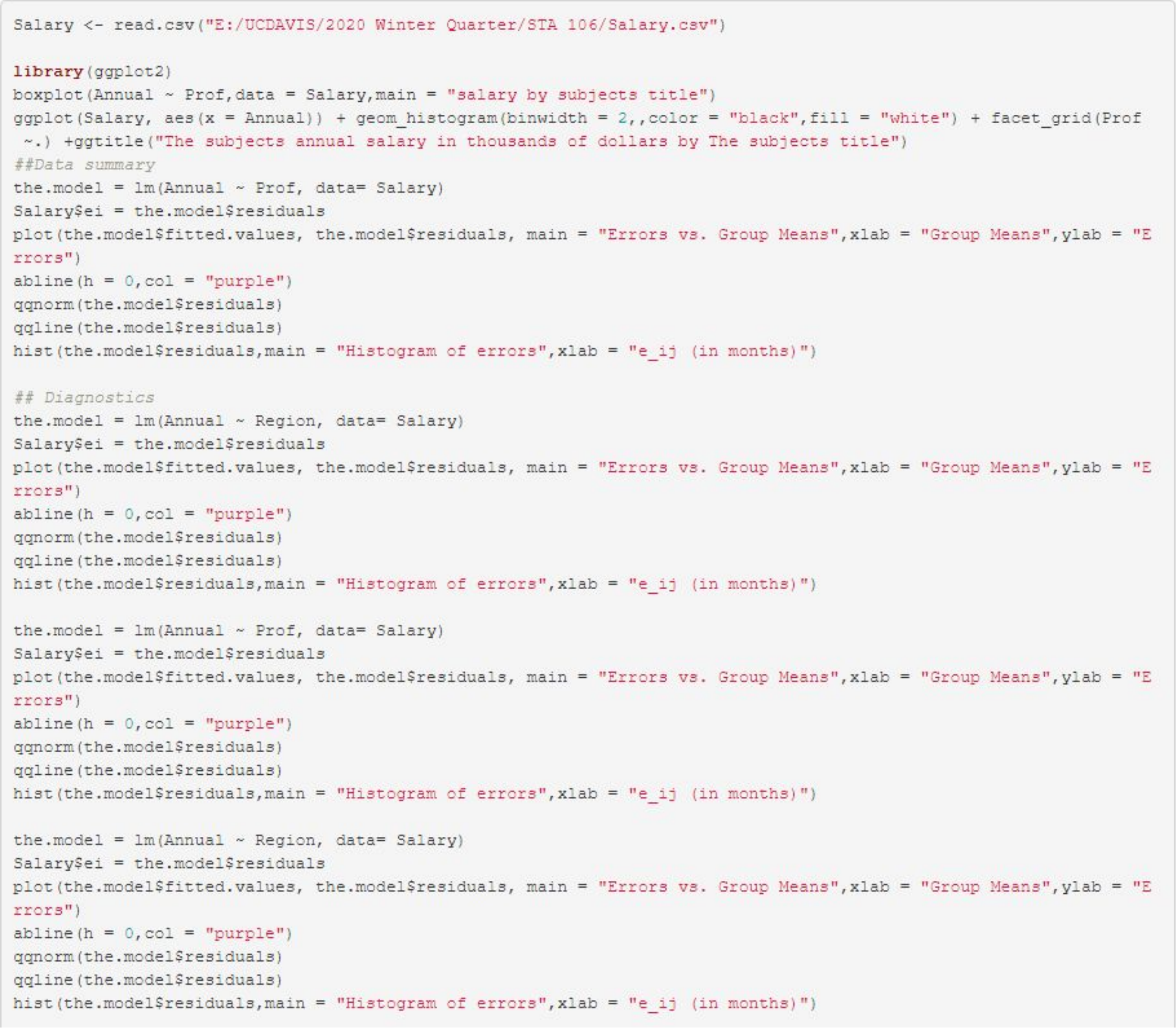
Appendix

Topic1 Question2:



```
NewHawk <- read.csv("E:/UCDAVIS/2020 Winter Quarter/STA 106/NewHawk.csv")

library(ggplot2)
boxplot(Wing ~ Species,data = NewHawk,main = "wing feather length by group")
ggplot(NewHawk, aes(x = Wing)) + geom_histogram(binwidth = 2,,color = "black",fill = "white") + facet_grid(Species ~.) +ggtitle("wing feather length by group")

the.model = lm(Wing ~ Species, data= NewHawk)
NewHawk$ei = the.model$residuals
plot(the.model$fitted.values, the.model$residuals, main = "Errors vs. Group Means",xlab = "Group Means",ylab = "Errors")
abline(h = 0,col = "purple")
qqnorm(the.model$residuals)
qqline(the.model$residuals)
hist(the.model$residuals,main = "Histogram of errors",xlab = "e_ij (in months)")

nt = nrow(NewHawk)
a=length(unique(NewHawk$Species))
t.cutoff= qt(1-0.01,nt -a)
rij = rstandard(the.model)
outliers = which(abs(rij) > t.cutoff)
new.data = NewHawk[-outliers,]
new.model = lm(Wing ~ Species,data = new.data)
plot(new.model$fitted.values, new.model$residuals, main = "Errors vs. Group Means",xlab = "Group Means",ylab = "Errors")
abline(h = 0,col = "purple")
qqnorm(new.model$residuals)
qqline(new.model$residuals)
hist(new.model$residuals,main = "Histogram of errors",xlab = "e_ij (in months)")

ggplot(new.data, aes(x = Wing)) + geom_histogram(binwidth = 2,,color = "black",fill = "white") + facet_grid(Species ~.) +ggtitle("wing feather length by group")
boxplot(Wing ~ Species,data = new.data,main = "wing feather length by group")
```

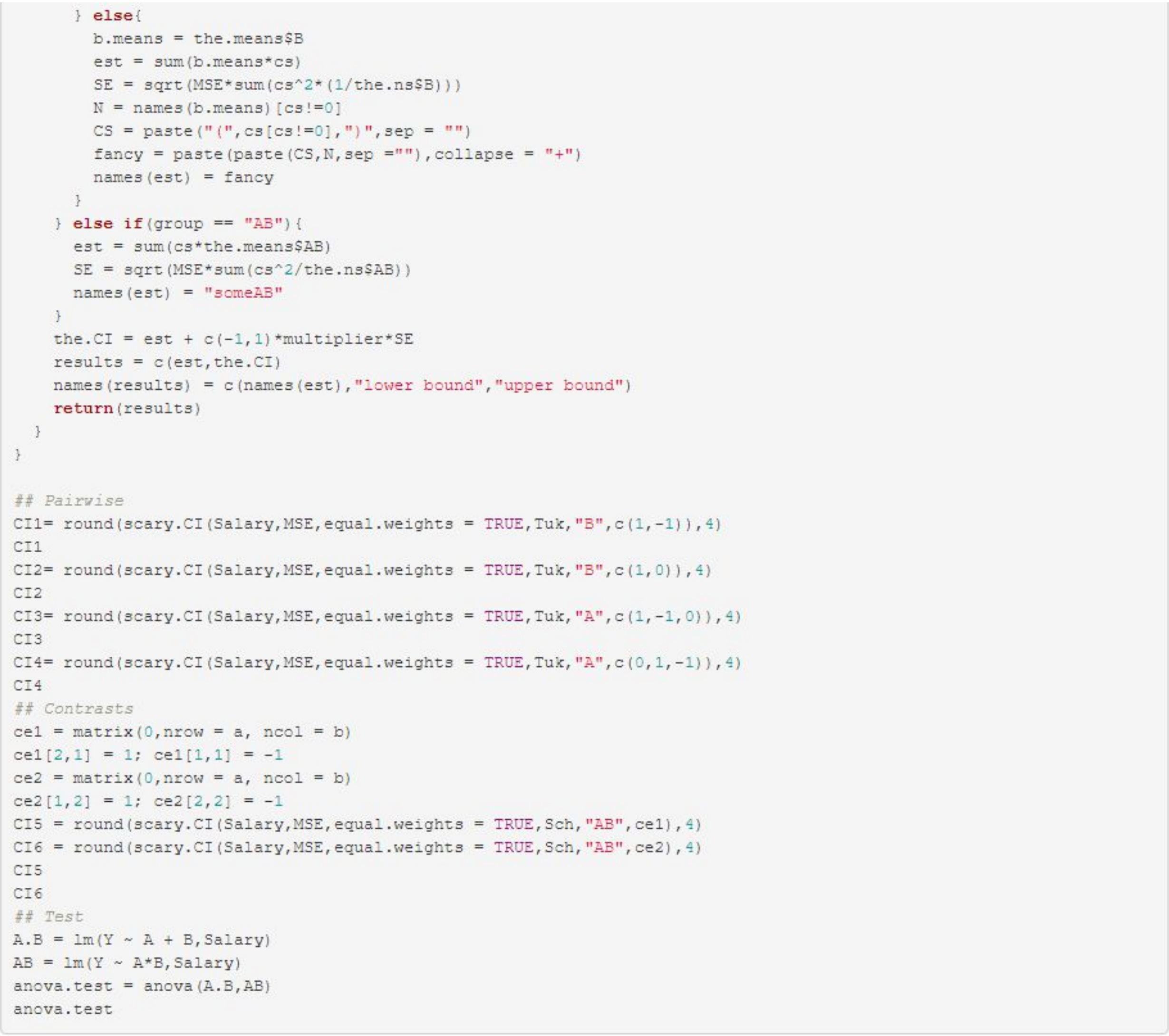
Topic2 Question1:



```r
Salary <- read.csv("E:/UCDAVIS/2020 Winter Quarter/STA 106/Salary.csv")

library(ggplot2)
boxplot(Annual ~ Prof,data = Salary,main = "salary by subjects title")
ggplot(Salary, aes(x = Annual)) + geom_histogram(binwidth = 2,,color = "black",fill = "white") + facet_grid(Prof
 ~.) +ggtitle("The subjects annual salary in thousands of dollars by The subjects title")
##Data summary
the.model = lm(Annual ~ Prof, data= Salary)
Salary$ei = the.model$residuals
plot(the.model$fitted.values, the.model$residuals, main = "Errors vs. Group Means",xlab = "Group Means",ylab = "E
rrors")
abline(h = 0,col = "purple")
qqnorm(the.model$residuals)
qqline(the.model$residuals)
hist(the.model$residuals,main = "Histogram of errors",xlab = "e_ij (in months)")

## Diagnostics
the.model = lm(Annual ~ Region, data= Salary)
Salary$ei = the.model$residuals
plot(the.model$fitted.values, the.model$residuals, main = "Errors vs. Group Means",xlab = "Group Means",ylab = "E
rrors")
abline(h = 0,col = "purple")
qqnorm(the.model$residuals)
qqline(the.model$residuals)
hist(the.model$residuals,main = "Histogram of errors",xlab = "e_ij (in months)")

the.model = lm(Annual ~ Prof, data= Salary)
Salary$ei = the.model$residuals
plot(the.model$fitted.values, the.model$residuals, main = "Errors vs. Group Means",xlab = "Group Means",ylab = "E
rrors")
abline(h = 0,col = "purple")
qqnorm(the.model$residuals)
qqline(the.model$residuals)
hist(the.model$residuals,main = "Histogram of errors",xlab = "e_ij (in months)")

the.model = lm(Annual ~ Region, data= Salary)
Salary$ei = the.model$residuals
plot(the.model$fitted.values, the.model$residuals, main = "Errors vs. Group Means",xlab = "Group Means",ylab = "E
rrors")
abline(h = 0,col = "purple")
qqnorm(the.model$residuals)
qqline(the.model$residuals)
hist(the.model$residuals,main = "Histogram of errors",xlab = "e_ij (in months)")
```

```
## CIS
names(Salary) = c("Y","A","B")
a = length(unique(Salary$A))
b = length(unique(Salary$B))
nt = nrow(Salary)
AB = lm(Y ~ A*B,Salary)
SSE = round(sum(AB$residuals^2),2)
the.ns = find.means(Salary,length)
B = round(AB$coefficients,3)
MSE = round(SSE/(nt-a*b),4)
find.mult = function(alpha,a,b,dfSSE,g,group){
  if(group == "A"){
    Tuk = round(qtukey(1-alpha,a,dfSSE)/sqrt(2),3)
    Bon = round(qt(1-alpha/(2*g), dfSSE ) ,3)
    Sch = round(sqrt((a-1)*qf(1-alpha, a-1, dfSSE)),3)
  }else if(group == "B"){
    Tuk = round(qtukey(1-alpha,b,dfSSE)/sqrt(2),3)
    Bon = round(qt(1-alpha/(2*g), dfSSE ) ,3)
    Sch = round(sqrt((b-1)*qf(1-alpha, b-1, dfSSE)),3)
  }else if(group == "AB"){
    Tuk = round(qtukey(1-alpha,a*b,dfSSE)/sqrt(2),3)
    Bon = round(qt(1-alpha/(2*g), dfSSE ) ,3)
    Sch = round(sqrt((a*b-1)*qf(1-alpha, a*b-1, dfSSE)),3)
  }
  results = c(Bon, Tuk,Sch)
  names(results) = c("Bonferroni","Tukey","Scheffe")
  return(results)
}
Bon = find.mult(0.1,a,b,nt-a*b,1,"A")[1]
Tuk = find.mult(0.1,a,b,nt-a*b,2,"B")[2]
Sch = find.mult(0.1,a,b,nt-a*b,2,"AB")[3]
```

```
scary.CI = function(the.data,MSE,equal.weights = TRUE,multiplier,group,cs){
   if(sum(cs) != 0 & sum(cs !=0 ) != 1){
    return("Error - you did not input a valid contrast")
  }else{
    the.means = find.means(the.data)
    the.ns =find.means(the.data,length)
    nt = nrow(the.data)
    a = length(unique(the.data[,2]))
    b = length(unique(the.data[,3]))
    if(group =="A"){
      if(equal.weights == TRUE){
        a.means = rowMeans(the.means$AB)
        est = sum(a.means*cs)
        mul = rowSums(1/the.ns$AB)
        SE = sqrt(MSE/b^2 * (sum(cs^2*mul)))
        N = names(a.means)[cs!=0]
        CS = paste("(",cs[cs!=0],")",sep = "")
        fancy = paste(paste(CS,N,sep =""),collapse = "+")
        names(est) = fancy
      } else{
        a.means = the.means$A
        est = sum(a.means*cs)
        SE = sqrt(MSE*sum(cs^2*(1/the.ns$A)))
        N = names(a.means)[cs!=0]
        CS = paste("(",cs[cs!=0],")",sep = "")
        fancy = paste(paste(CS,N,sep =""),collapse = "+")
        names(est) = fancy
      }
    }else if(group == "B"){
      if(equal.weights == TRUE){
        b.means = colMeans(the.means$AB)
        est = sum(b.means*cs)
        mul = colSums(1/the.ns$AB)
        SE = sqrt(MSE/a^2 * (sum(cs^2*mul)))
        N = names(b.means)[cs!=0]
        CS = paste("(",cs[cs!=0],")",sep = "")
        fancy = paste(paste(CS,N,sep =""),collapse = "+")
        names(est) = fancy
```



```
      } else{
        b.means = the.means$B
        est = sum(b.means*cs)
        SE = sqrt(MSE*sum(cs^2*(1/the.ns$B)))
        N = names(b.means)[cs!=0]
        CS = paste("(",cs[cs!=0],")",sep = "")
        fancy = paste(paste(CS,N,sep =""),collapse = "+")
        names(est) = fancy
      }
    } else if(group == "AB"){
      est = sum(cs*the.means$AB)
      SE = sqrt(MSE*sum(cs^2/the.ns$AB))
      names(est) = "someAB"
    }
    the.CI = est + c(-1,1)*multiplier*SE
    results = c(est,the.CI)
    names(results) = c(names(est),"lower bound","upper bound")
    return(results)
  }
}

## Pairwise
CI1= round(scary.CI(Salary,MSE,equal.weights = TRUE,Tuk,"B",c(1,-1)),4)
CI1
CI2= round(scary.CI(Salary,MSE,equal.weights = TRUE,Tuk,"B",c(1,0)),4)
CI2
CI3= round(scary.CI(Salary,MSE,equal.weights = TRUE,Tuk,"A",c(1,-1,0)),4)
CI3
CI4= round(scary.CI(Salary,MSE,equal.weights = TRUE,Tuk,"A",c(0,1,-1)),4)
CI4
## Contrasts
ce1 = matrix(0,nrow = a, ncol = b)
ce1[2,1] = 1; ce1[1,1] = -1
ce2 = matrix(0,nrow = a, ncol = b)
ce2[1,2] = 1; ce2[2,2] = -1
CI5 = round(scary.CI(Salary,MSE,equal.weights = TRUE,Sch,"AB",ce1),4)
CI6 = round(scary.CI(Salary,MSE,equal.weights = TRUE,Sch,"AB",ce2),4)
CI5
CI6
## Test
A.B = lm(Y ~ A + B,Salary)
AB = lm(Y ~ A*B,Salary)
anova.test = anova(A.B,AB)
anova.test
```

That’s all. Thank you