**WILDCAT STRIKES**

**#Loading the dataset**

data1<- read.csv(file.choose())

**#DATA EXPLORATION**

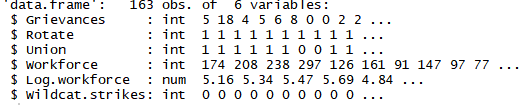
colnames(data1)

|  |
| --- |
| **"Grievances" "Rotate" "Union" "Workforce" "Log.workforce" "Wildcat.strikes"** |
|  |

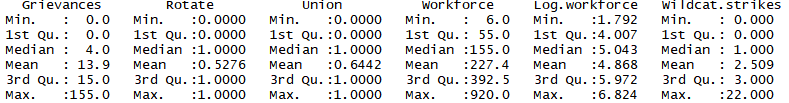
nrow(data1)

**163**

str(data1)



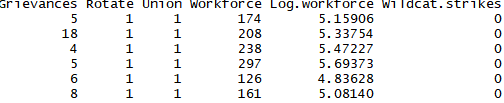
summary(data1)



sum(is.na(data1))

0 **#There are no missing values in the dataset**

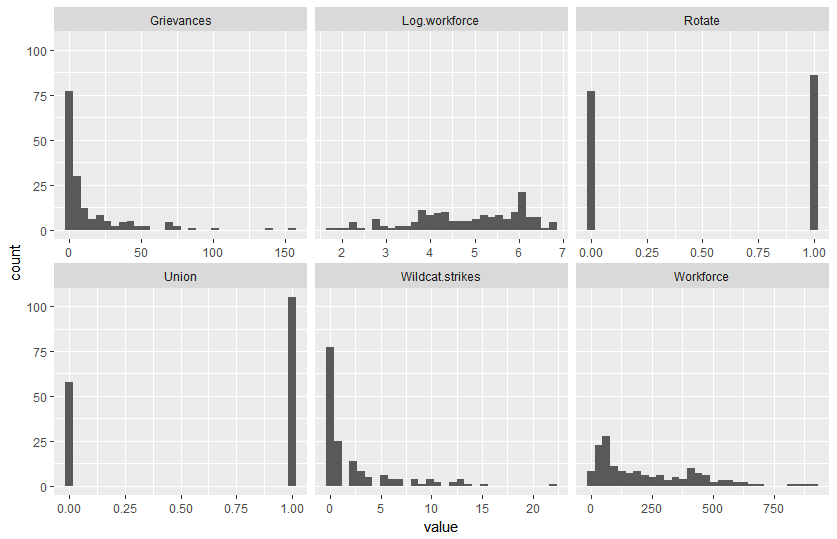
head(data1)



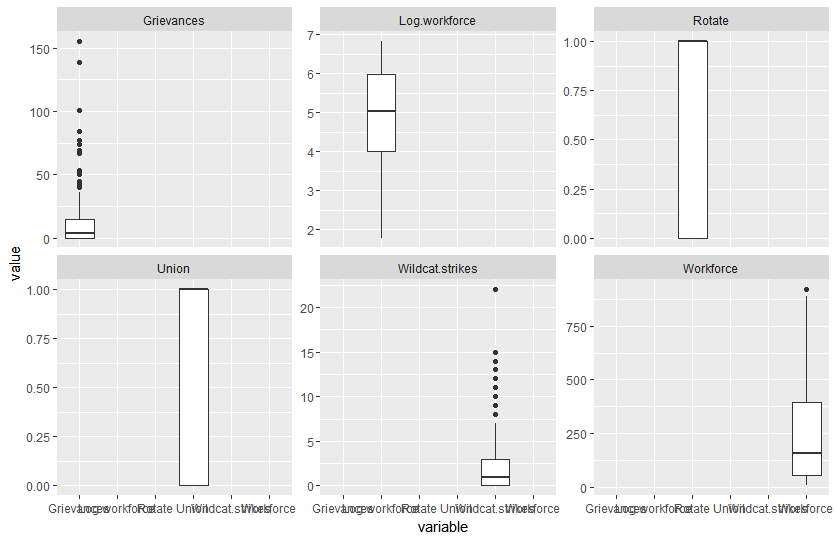
library(ggplot2)

library(tidyverse)

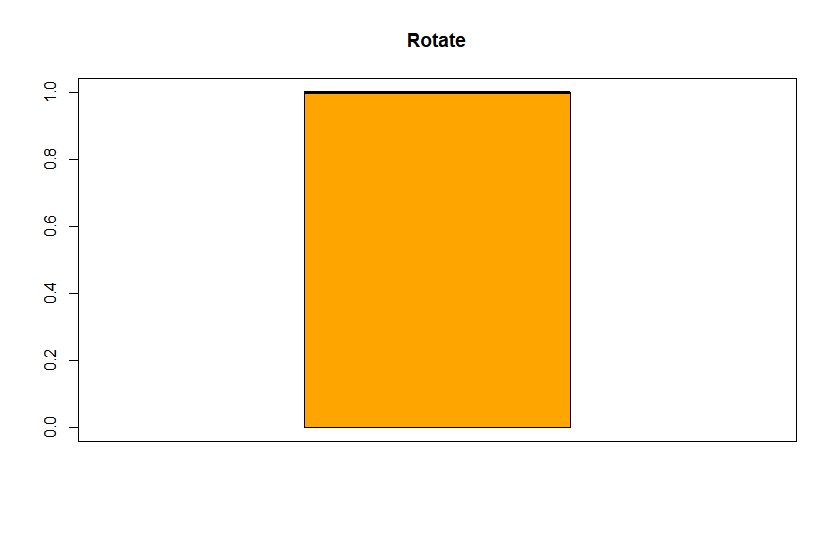
data1 %>% gather(Grievances:Wildcat.strikes, key = "variable", value = "value") %>% ggplot(aes(x = value))+ geom\_histogram(bins = 30) + facet\_wrap(~ variable, scales = 'free\_x')

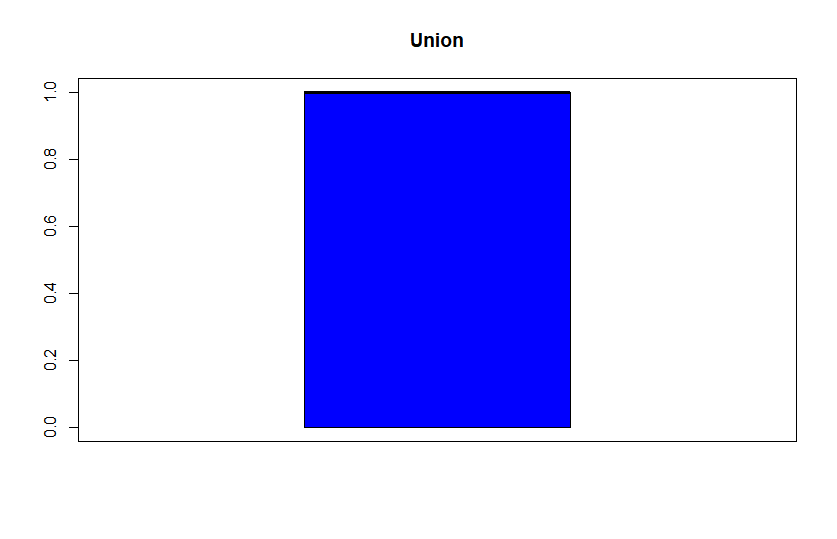


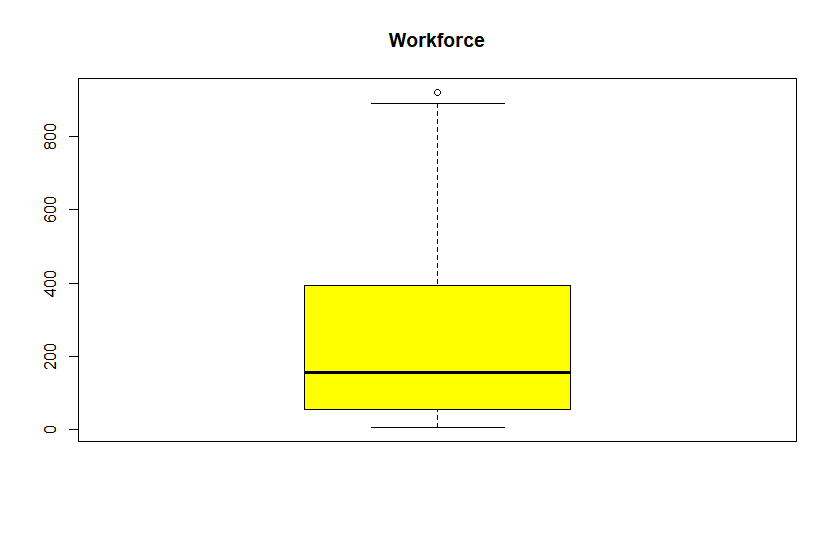
data1 %>% gather(Grievances:Wildcat.strikes, key = "variable", value = "value") %>% ggplot(aes(x=variable,y = value)) + geom\_boxplot() + facet\_wrap(~ variable, scales = 'free\_y')

boxplot(data1$Grievances, main="Grievances", col ="blue")

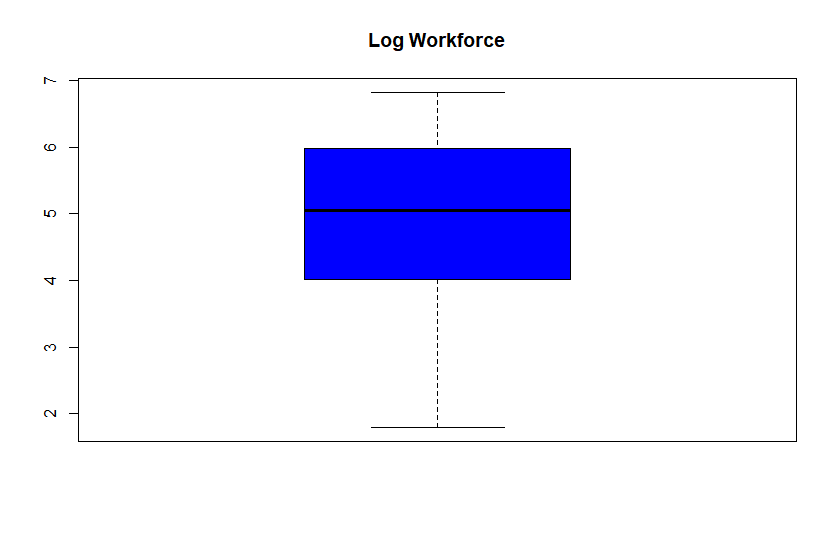
boxplot(data1$Rotate, main="Rotate", col ="orange")



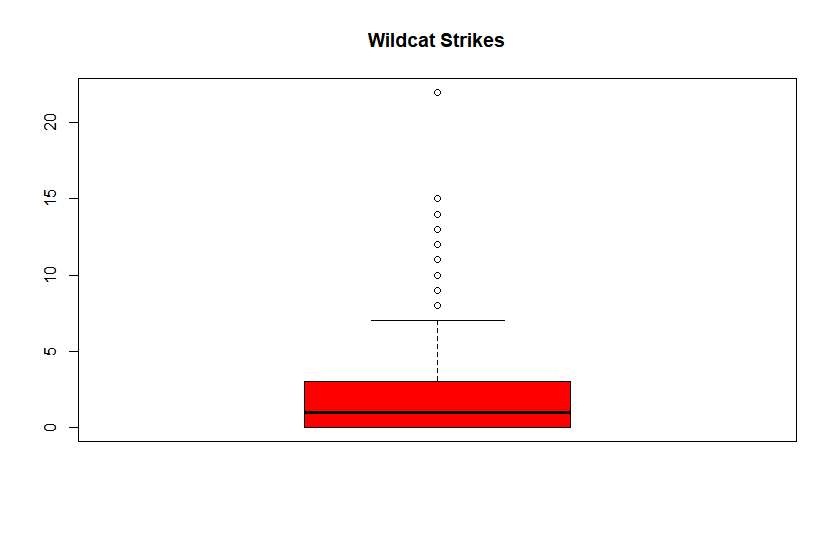
boxplot(data1$Union, main="Union", col ="blue")

boxplot(data1$Workforce, main="Workforce", col ="yellow")

boxplot(data1$Log.workforce, main="Log Workforce", col ="blue")



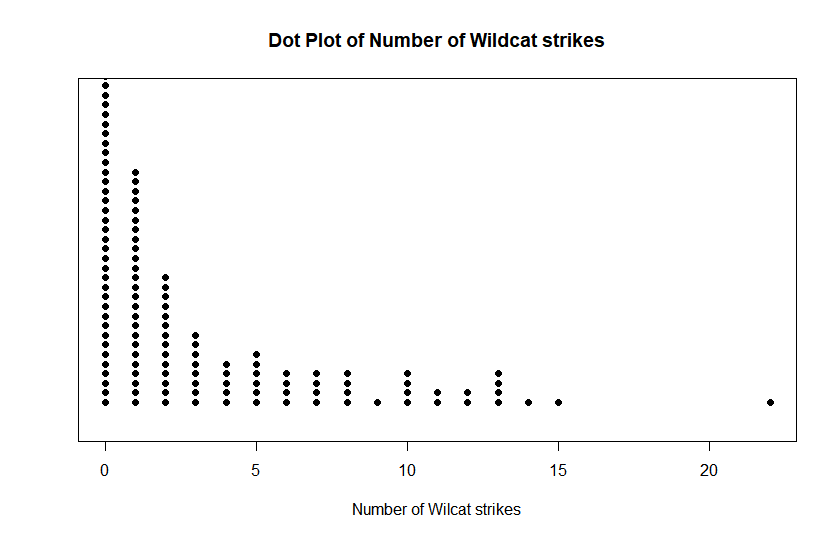
boxplot(data1$Wildcat.strikes, main="Wildcat Strikes", col ="red")



***The histograms and boxplots show the presence of few outliers but there is one mine (it would have easier to may be identify if we had the mines with their id) with an unusual similar behavior pattern across the attributes . This can be spotted in the boxplot of Strikes, Workforce.***

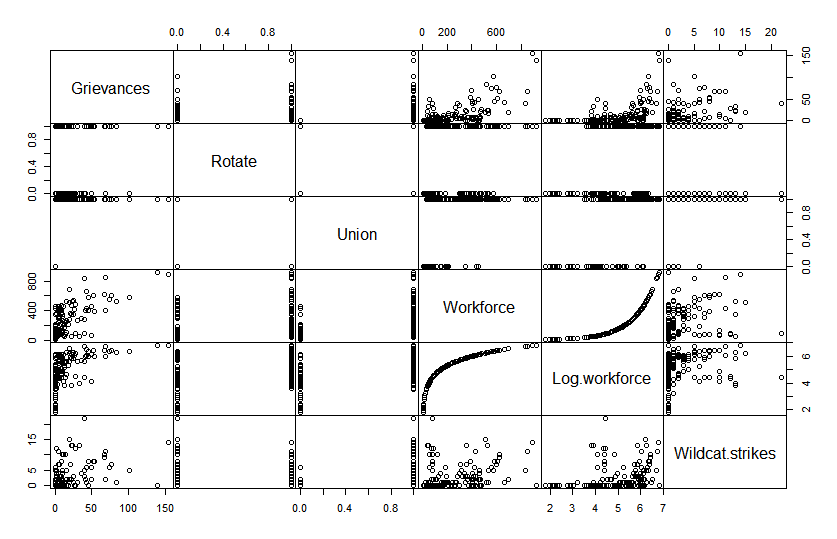
***Also, we do see the same thing in Dot Plot that shows an outlier with a value beyond 20(around 22), though the range is not very high (0-20)***

stripchart(data1$Wildcat.strikes, method="stack", offset=0.5, at =0.15,pch=19, main="Dot Plot of Number of Wildcat strikes",xlab="Number of Wilcat strikes")



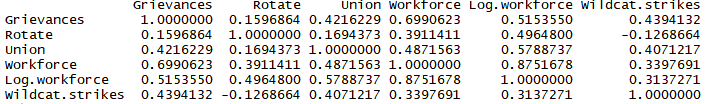
#Looking at pairwise scatterplots

pairs(data1, gap = 0, pch = 21)



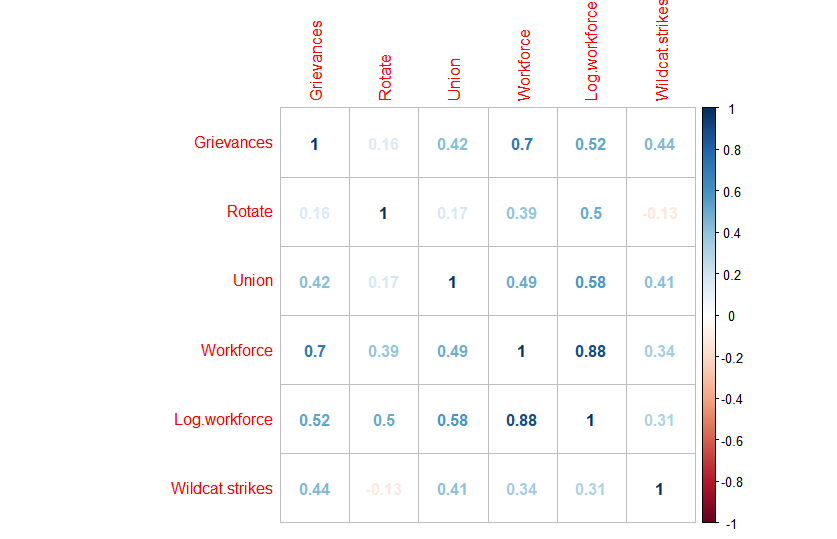
***We can again confirm the truth that an individual mine has an unusual behavior pattern by looking at the scatterplot between Log workforce and Wildcat Strikes or Workforce and Wildcat Strikes.***

cor(data1, method = c("pearson"))



library(corrplot)

corrplot(round(cor(data1),2), method = "number")



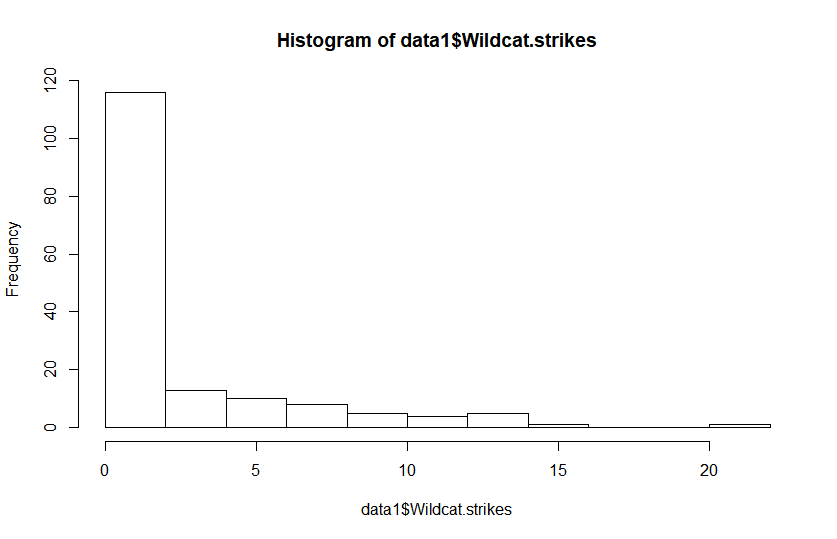
***#Only Grievance and Workforce seem to have a strong positive relationship***

data1$Wildcat.strikes

mean(data1$Wildcat.strikes)

**2.509202**

hist(data$Wildcat.strikes)



library(sqldf)

sqldf('select \* from data1 where "Wildcat.strikes" == 22')



mean(data$Grievances)

**13.90184**

**#Finding patterns in Union and Non-Union mines**

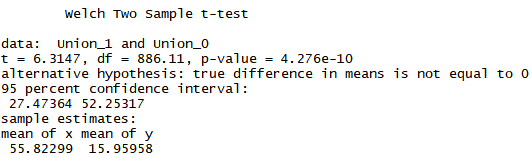
Union\_1<- sqldf('select \* from data1 where "Union" == 1')

Union\_1

Union\_0<- sqldf('select \* from data1 where "Union" == 0')

Union\_0

t.test(Union\_1,Union\_0)

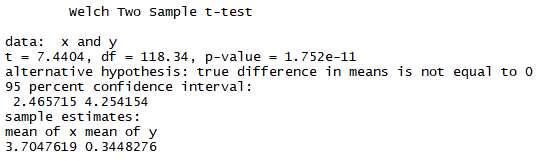


The t-test confirms that the Union and Non-Union mines should show different patterns as they have different means and there is a big difference in their means

x<-sqldf('select "Wildcat.strikes" from data1 where "Union" == 1')

y<-sqldf('select "Wildcat.strikes" from data1 where "Union" == 0')

t.test(x,y)



***The t-test here shows that the mean of the number of Wildcat strikes in Union and Non-Union mines are different. Hence, again a confirmation that both will show different patterns.***

sqldf('select avg("Wildcat.strikes") from data1 where "Union" == 1')



sqldf('select avg("Wildcat.strikes") from data1 where "Union" == 0')



sqldf('select avg("Wildcat.strikes") from data1 where "Rotate" == 1')



sqldf('select avg("Wildcat.strikes") from data1 where "Rotate" == 0')



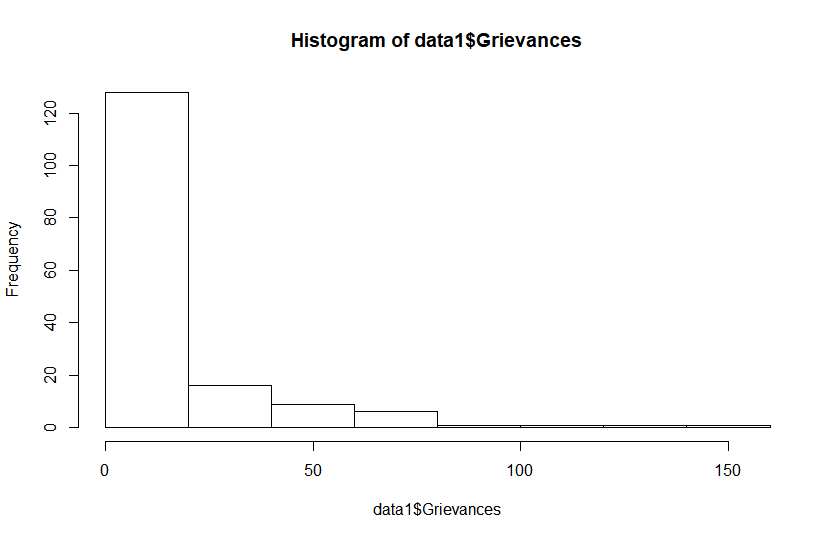
#This could be an obvious reason that generally Strikes would be more where there are no rotational shifts i.e. less flexibility.

mean(data1$Wildcat.strikes)

**2.509202**

head(arrange(Union, desc(Grievances)))

hist(data1$Grievances)

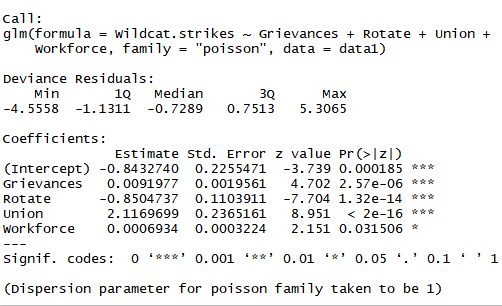


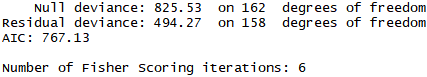
***Since, it is ample clear that the dataset can be split into two smaller datasets- Union mines dataset and Non-Union mines dataset as the tendencies of the two groups are very different. Hence, in order to study them further and model them, it would be a good idea to model them individually.***

***#As the predictor variable- Wildcat Strikes is a count data(discrete, random, non-negative), hence we shall start by fitting a Poisson model on the complete dataset.***

model = glm(Wildcat.strikes~Grievances+Rotate+Union+Workforce, data=data1, family = "poisson")

summary(model)



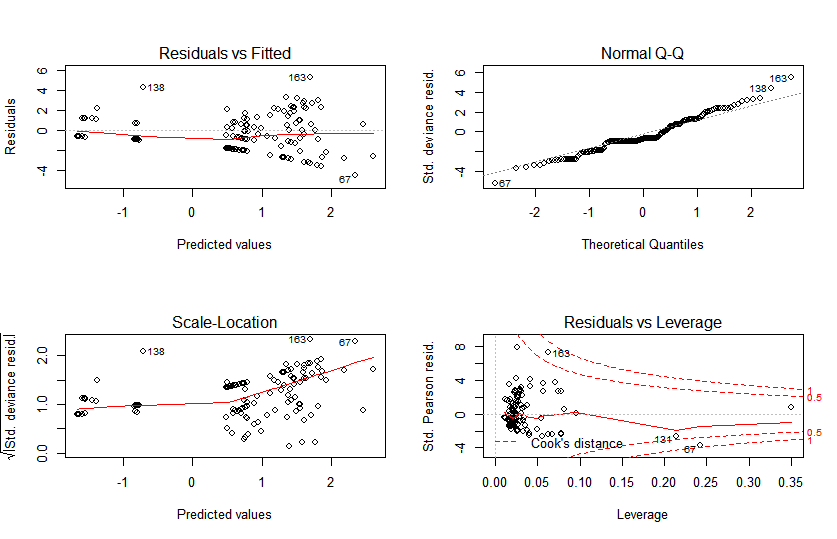


***#All the regressors show significant in the summary. The Residual Deviance though lower than Null Deviance is still very high as compares with its degree of freedom.***

par(mfrow=c(2,2))

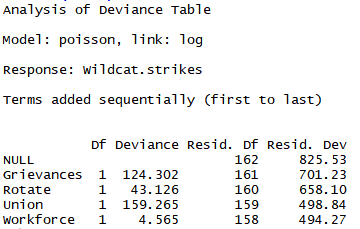
plot(model)

par(mfrow=c(1,1))



***# The plot shows that with the increase in mean, the variance is increasing.***

anova(model)



with(model, cbind(res.deviance=deviance, df=df.residual,p=pchisq(deviance, df.residual, lower.tail=FALSE)))

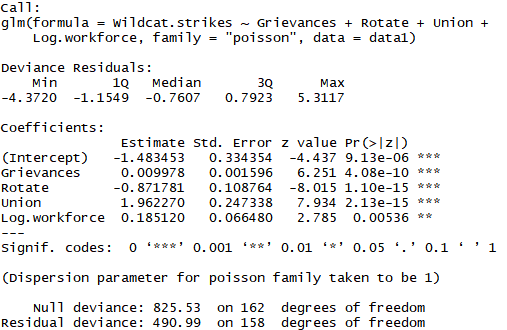


***# The significantly small value of p suggests that the model is not a good fit and we need to improvise it.***

***Let’s see if the model would improve if we replace Workforce by Log Workforce***

model1= glm(Wildcat.strikes~Grievances+Rotate+Union+Log.workforce, data=data, family = "poisson")

summary(model1)



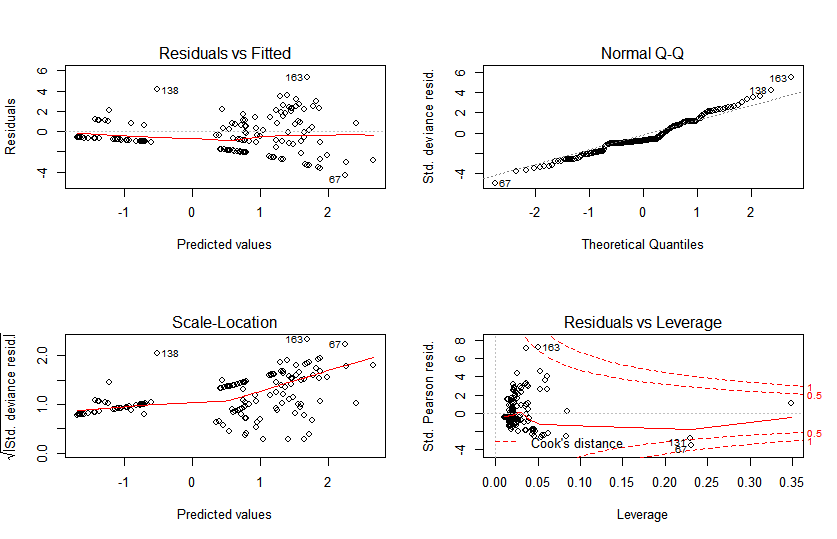
AIC: 763.84

Number of Fisher Scoring iterations: 6

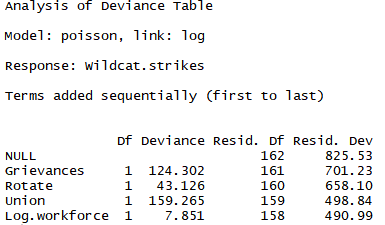
***There is not a significant improvement from the above basic model, however the Residual Deviance has further dropped and even the AIC is better. So, we shall use this model to improvise further.***

par(mfrow=c(2,2))

plot(model1)

par(mfrow=c(1,1))

anova(model1)



with(model1, cbind(res.deviance=deviance, df=df.residual,p=pchisq(deviance, df.residual, lower.tail=FALSE)))



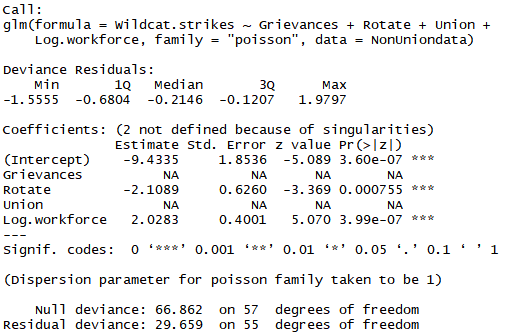
***# p-value suggests we need to incorporate changes in our model.***

**As discussed above , the Union and Non-Union mines need to be modelled differently.So, now let’s try to model the Non-Union mines.**

NonUniondata<-subset(data1, Union==0)

NonUnionmodel = glm(Wildcat.strikes~Grievances+Rotate+Union+Log.workforce, data=NonUniondata, family = "poisson")

summary(NonUnionmodel)



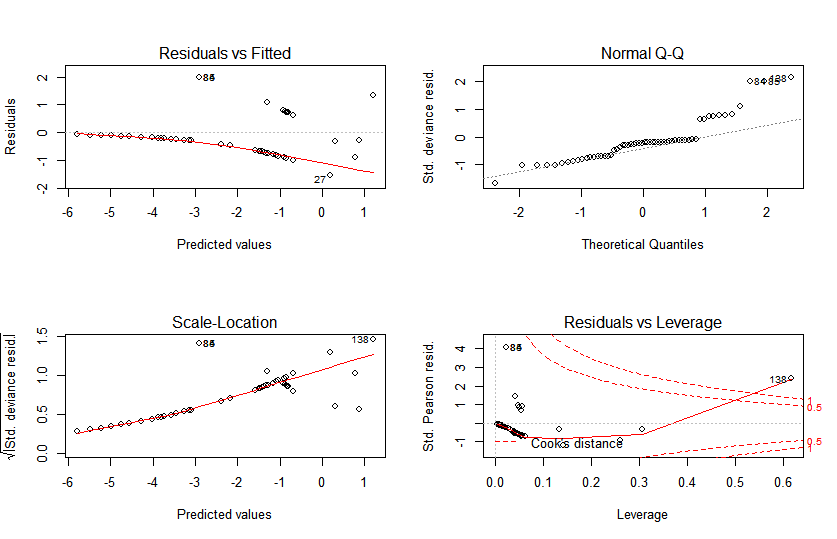
AIC: 65.93

Number of Fisher Scoring iterations: 6

***# This looks like a good fit because the Residual Deviance is closer to the degree of freedom and the AIC is much smaller than the previous models that we ran as a test on the complete dataset.***

par(mfrow=c(2,2))

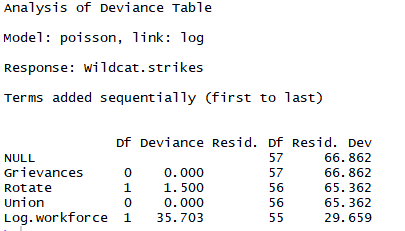
plot(NonUnionmodel)



par(mfrow=c(1,1))

***#The plots show that scatter or dispersion is increasing with increase in mean.***

**anova(NonUnionmodel)**



with(NonUnionmodel, cbind(res.deviance=deviance, df=df.residual,p=pchisq(deviance, df.residual, lower.tail=FALSE)))



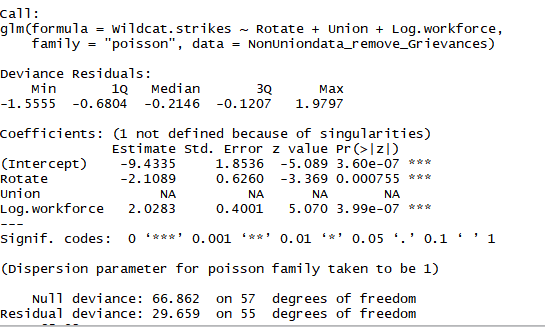
***#The large value of p suggest that we fail to reject the Null Hypothesis. Hence, Poisson model for Non-Union mines is a good fit.***

**#As per rules of regression, any column that has all the values as 0 because of non-claiming or non-filing can be removed. Hence, we may run another Poisson model after removing Grievances column.**

NonUniondata\_remove\_Grievances<-subset(data1, Union==0)

NonUniondata\_Grievances= glm(Wildcat.strikes~ Rotate+Union+Log.workforce, data= NonUniondata\_remove\_Grievances, family = "poisson")

summary(NonUniondata\_Grievances)



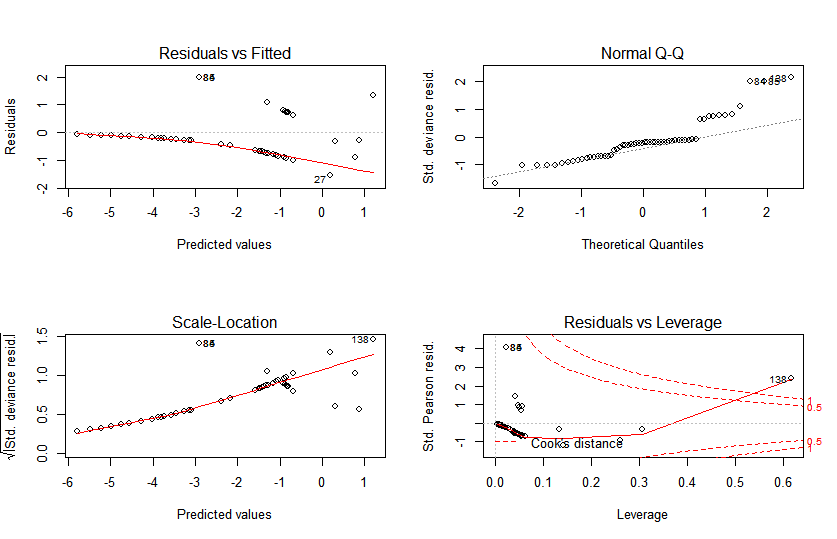
AIC: 65.93

Number of Fisher Scoring iterations: 6

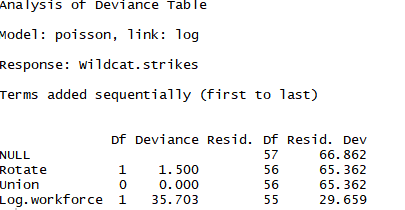
par(mfrow=c(2,2))

plot(NonUniondata\_Grievances)

par(mfrow=c(1,1))



anova(NonUniondata\_Grievances)



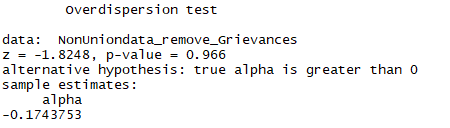
with(NonUniondata\_Grievances, cbind(res.deviance=deviance, df=df.residual,p=pchisq(deviance, df.residual, lower.tail=FALSE)))



***#The goodness of fit test gives a large p-value stating that our model is good fit.***

Still, if we want , we may check for overdispersion if at all present in our model

dispersiontest(NonUniondata\_remove\_Grievances,trafo=1)



**The value of alpha less than 0 suggests that there is no overdispersion**

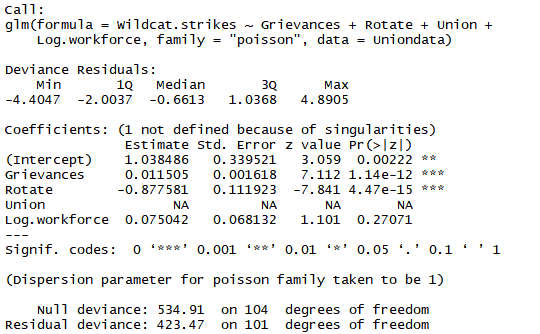
**The Non Union model data does not suffer from Overdispersion,hence the Poisson model may be deemed fit.**

**# Now, let’s model using Union==1**

Uniondata<-subset(data1, Union==1)

Unionmodel = glm(Wildcat.strikes~Grievances+Rotate+Union+Log.workforce, data=Uniondata, family = "poisson")

summary(Unionmodel)



AIC: 664.05

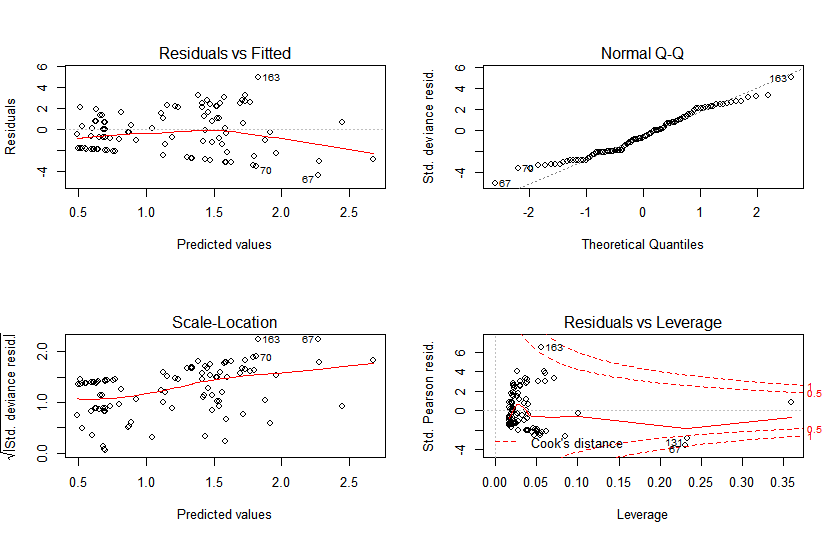
Number of Fisher Scoring iterations: 5

**# The Residual Deviance is almost 4 times of the degrees of freedom. It also has a high AIC.**

par(mfrow=c(2,2))

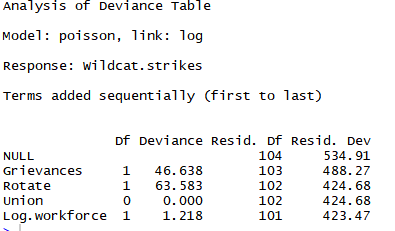
plot(Unionmodel)

par(mfrow=c(1,1))



**# The plots that scatter and the variance increasing with the increase in predicted values. *The residuals lie around the central line like a set of random observations without a pattern.***

**anova(Unionmodel)**



with(Unionmodel, cbind(res.deviance=deviance, df=df.residual,p=pchisq(deviance, df.residual, lower.tail=FALSE)))

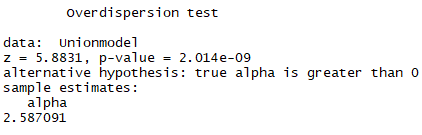


***# The p-value gives a strong statement about rejecting the Null Hypothesis.i.e. we need to improvise our model further.***

**Here, we shall run the Dispersion test to confirm the existence of Overdispersion**

library(AER)

dispersiontest(Unionmodel,trafo=1)



**The value of alpha>0 suggests that the data suffers from Overdispersion and needs an alternative model for the Union mines.**

**# As , it is confirmed that there is overdispersion, hence we shall try fitting Negative Binomial .**

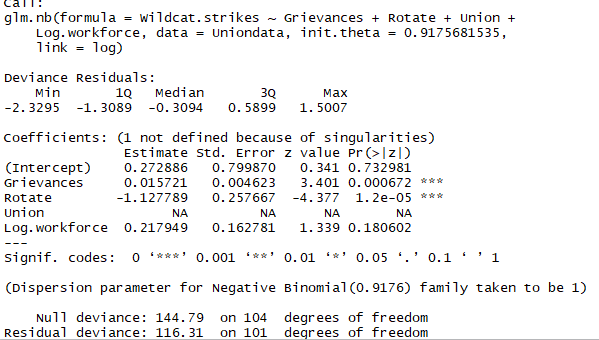
library(foreign)

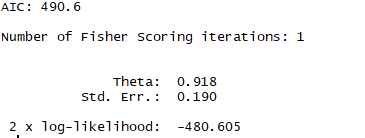
library(MASS)

Uniondata<-subset(data1, Union==1)

Unionmodel\_nb = glm.nb(Wildcat.strikes~Grievances+Rotate+Union+Log.workforce, data=Uniondata)

summary(Unionmodel\_nb)



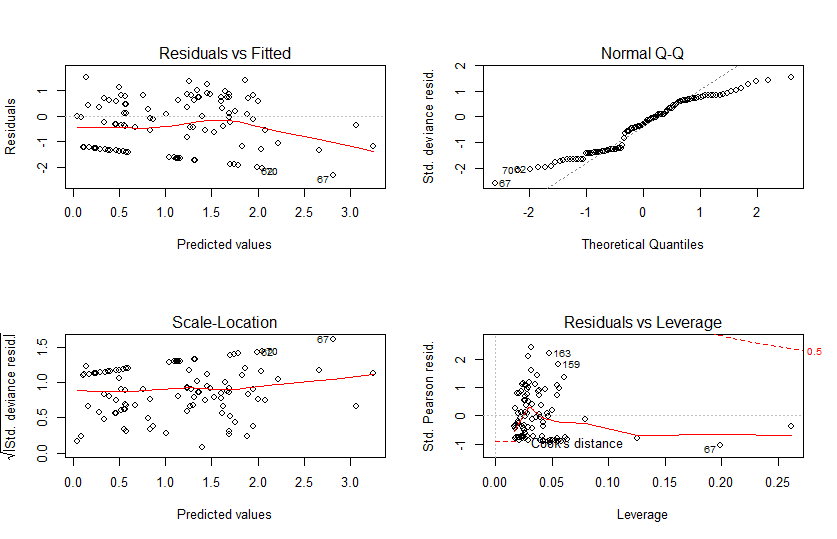


***# The Residual Deviance almost equals the degrees of freedom and the AIC is lower than the previous model.***

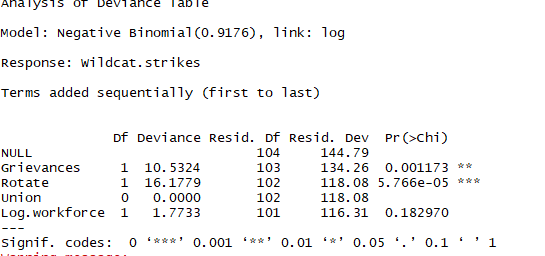
par(mfrow=c(2,2))

plot(Unionmodel\_nb)

par(mfrow=c(1,1))



anova(Unionmodel\_nb)



with(Unionmodel\_nb, cbind(res.deviance=deviance, df=df.residual,p=pchisq(deviance, df.residual, lower.tail=FALSE)))



**# The goodness of fit suggests that the p-value is large enough to fail to reject the Null Hypothesis. Hence the Negative Binomial for Union mines is a good fit. It looks relatively much better .**

**We may try transforming some attributes as next steps and try running the Negative Binomial.**

**It may lead to even better results.**

**Zero Inflated or Hurdle model can be made on the combined dataset. Like an ensemble**