Multiple Linear Regression

# import knitr package - Markdown purpose :  
setwd("C:\\Mugi\\R\\Concrete - Regression")  
  
library(readxl)

library(dplyr)

library(corrplot)

library(caTools)

library(car)

conc\_data = read\_excel('Concrete\_Data.xls')  
  
names(conc\_data) = c('cement','blast\_furnace\_slag','fly\_ash','water','superplasticizer','coarse\_ggregate', 'fine\_aggregate', 'age', 'compressive\_strength')  
  
nrow(conc\_data)

## [1] 1030

ncol(conc\_data)

## [1] 9

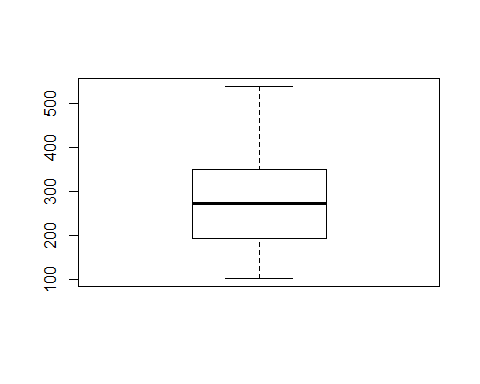
str(conc\_data)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 1030 obs. of 9 variables:  
## $ cement : num 540 540 332 332 199 ...  
## $ blast\_furnace\_slag : num 0 0 142 142 132 ...  
## $ fly\_ash : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ water : num 162 162 228 228 192 228 228 228 228 228 ...  
## $ superplasticizer : num 2.5 2.5 0 0 0 0 0 0 0 0 ...  
## $ coarse\_ggregate : num 1040 1055 932 932 978 ...  
## $ fine\_aggregate : num 676 676 594 594 826 ...  
## $ age : num 28 28 270 365 360 90 365 28 28 28 ...  
## $ compressive\_strength: num 80 61.9 40.3 41.1 44.3 ...

summary(conc\_data)

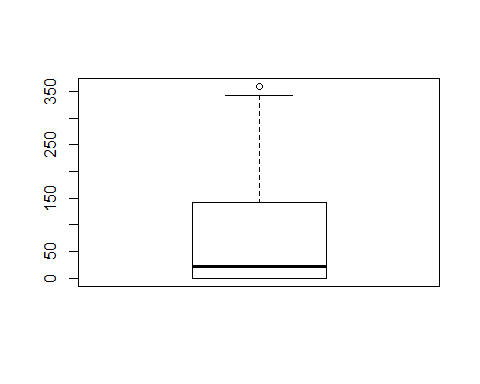
## cement blast\_furnace\_slag fly\_ash water   
## Min. :102.0 Min. : 0.0 Min. : 0.00 Min. :121.8   
## 1st Qu.:192.4 1st Qu.: 0.0 1st Qu.: 0.00 1st Qu.:164.9   
## Median :272.9 Median : 22.0 Median : 0.00 Median :185.0   
## Mean :281.2 Mean : 73.9 Mean : 54.19 Mean :181.6   
## 3rd Qu.:350.0 3rd Qu.:142.9 3rd Qu.:118.27 3rd Qu.:192.0   
## Max. :540.0 Max. :359.4 Max. :200.10 Max. :247.0   
## superplasticizer coarse\_ggregate fine\_aggregate age   
## Min. : 0.000 Min. : 801.0 Min. :594.0 Min. : 1.00   
## 1st Qu.: 0.000 1st Qu.: 932.0 1st Qu.:731.0 1st Qu.: 7.00   
## Median : 6.350 Median : 968.0 Median :779.5 Median : 28.00   
## Mean : 6.203 Mean : 972.9 Mean :773.6 Mean : 45.66   
## 3rd Qu.:10.160 3rd Qu.:1029.4 3rd Qu.:824.0 3rd Qu.: 56.00   
## Max. :32.200 Max. :1145.0 Max. :992.6 Max. :365.00   
## compressive\_strength  
## Min. : 2.332   
## 1st Qu.:23.707   
## Median :34.443   
## Mean :35.818   
## 3rd Qu.:46.136   
## Max. :82.599

#Finding Outlier - cement  
  
boxplot(conc\_data$cement)



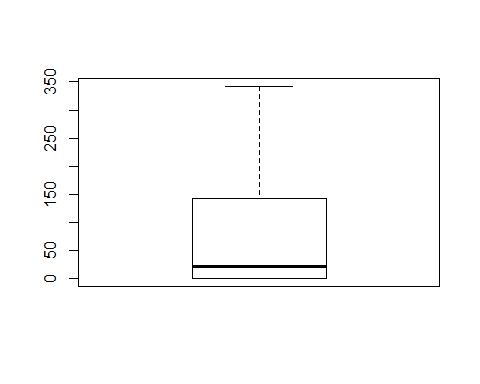
q = quantile(conc\_data$cement, probs=c(.25, .75))  
  
lower\_boundry = q[1] - (1.5 \* IQR(conc\_data$cement))  
  
upper\_boundry = q[2] + (1.5 \* IQR(conc\_data$cement))  
  
sum(conc\_data$cement > upper\_boundry | conc\_data$cement < lower\_boundry)

## [1] 0

#Finding Outlier - blast\_furnace\_slag  
  
boxplot(conc\_data$blast\_furnace\_slag)

q = quantile(conc\_data$blast\_furnace\_slag, probs=c(.25, .75))  
  
bfs\_lower\_boundry = q[1] - (1.5 \* IQR(conc\_data$blast\_furnace\_slag))  
  
bfs\_upper\_boundry = q[2] + (1.5 \* IQR(conc\_data$blast\_furnace\_slag))  
  
sum(conc\_data$blast\_furnace\_slag > bfs\_upper\_boundry | conc\_data$blast\_furnace\_slag < bfs\_lower\_boundry)

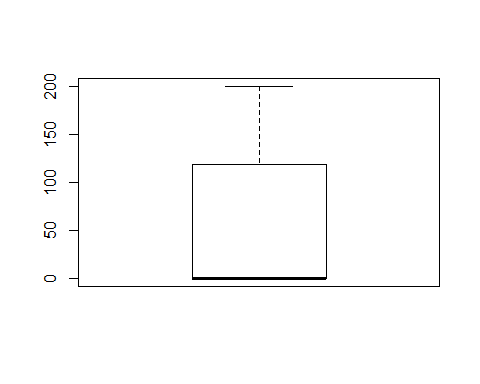
## [1] 2

boxplot(conc\_data$blast\_furnace\_slag[conc\_data$blast\_furnace\_slag < bfs\_upper\_boundry & conc\_data$blast\_furnace\_slag > bfs\_lower\_boundry])

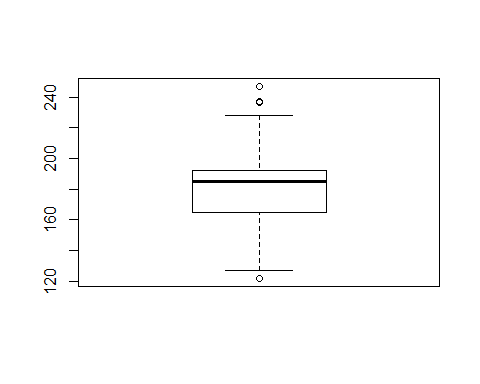
#Finding Outlier - fly\_ash  
  
boxplot(conc\_data$fly\_ash)  
  
q = quantile(conc\_data$fly\_ash, probs=c(.25, .75))  
  
lower\_boundry = q[1] - (1.5 \* IQR(conc\_data$fly\_ash))  
  
upper\_boundry = q[2] + (1.5 \* IQR(conc\_data$fly\_ash))  
  
sum(conc\_data$fly\_ash > upper\_boundry | conc\_data$fly\_ash < lower\_boundry)

## [1] 0

boxplot(conc\_data$fly\_ash[conc\_data$fly\_ash < upper\_boundry & conc\_data$fly\_ash > lower\_boundry])



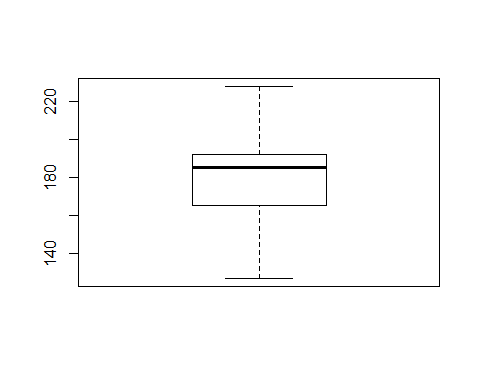
#Finding Outlier - water  
  
boxplot(conc\_data$water)



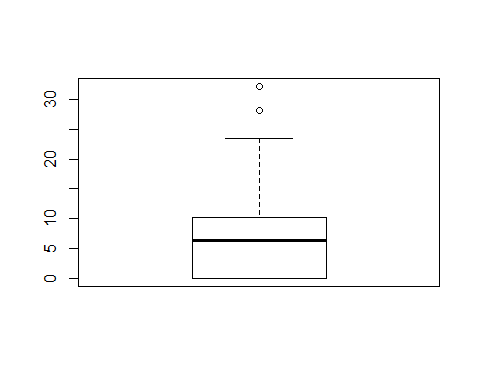
q = quantile(conc\_data$water, probs=c(.25, .75))  
  
w\_lower\_boundry = q[1] - (1.5 \* IQR(conc\_data$water))  
  
w\_upper\_boundry = q[2] + (1.5 \* IQR(conc\_data$water))  
  
sum(conc\_data$water > w\_upper\_boundry | conc\_data$water < w\_lower\_boundry)

## [1] 9

boxplot(conc\_data$water[conc\_data$water < w\_upper\_boundry & conc\_data$water > w\_lower\_boundry])



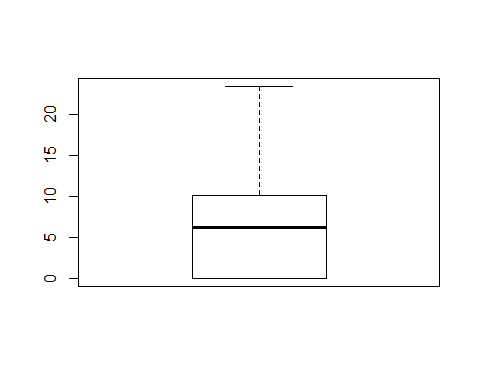
#Finding Outlier - superplasticizer  
  
boxplot(conc\_data$superplasticizer)



q = quantile(conc\_data$superplasticizer, probs=c(.25, .75))  
  
sp\_lower\_boundry = q[1] - (1.5 \* IQR(conc\_data$superplasticizer))  
  
sp\_upper\_boundry = q[2] + (1.5 \* IQR(conc\_data$superplasticizer))  
  
sum(conc\_data$superplasticizer > sp\_upper\_boundry | conc\_data$superplasticizer < sp\_lower\_boundry)

## [1] 10

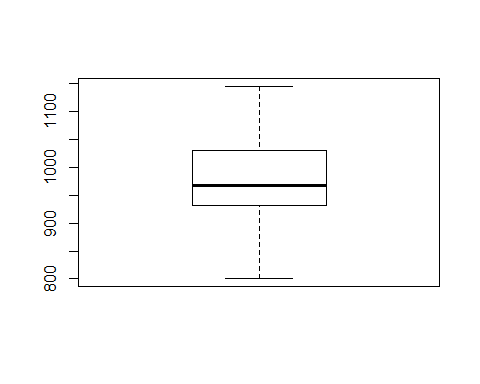
boxplot(conc\_data$superplasticizer[conc\_data$superplasticizer < sp\_upper\_boundry & conc\_data$superplasticizer > sp\_lower\_boundry])



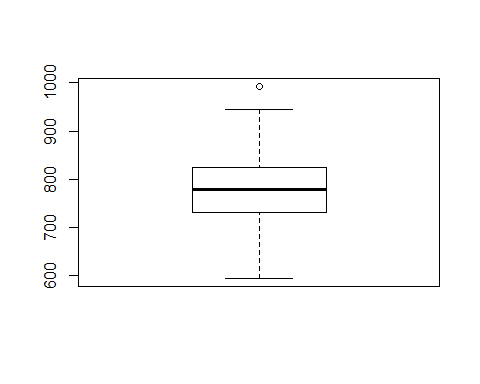
#Finding Outlier - coarse\_ggregate  
  
boxplot(conc\_data$coarse\_ggregate)  
  
q = quantile(conc\_data$coarse\_ggregate, probs=c(.25, .75))  
  
lower\_boundry = q[1] - (1.5 \* IQR(conc\_data$coarse\_ggregate))  
  
upper\_boundry = q[2] + (1.5 \* IQR(conc\_data$coarse\_ggregate))  
  
sum(conc\_data$coarse\_ggregate > upper\_boundry | conc\_data$coarse\_ggregate < lower\_boundry)

## [1] 0

boxplot(conc\_data$coarse\_ggregate[conc\_data$coarse\_ggregate < upper\_boundry & conc\_data$coarse\_ggregate > lower\_boundry])



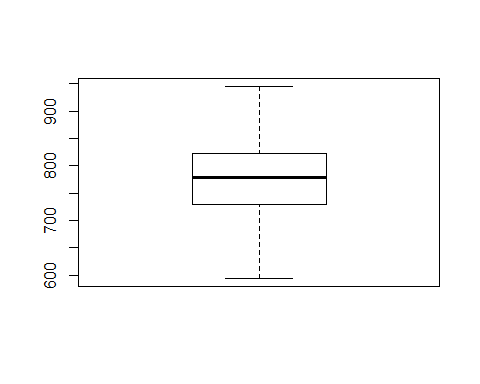
#Finding Outlier - fine\_aggregate  
  
boxplot(conc\_data$fine\_aggregate)



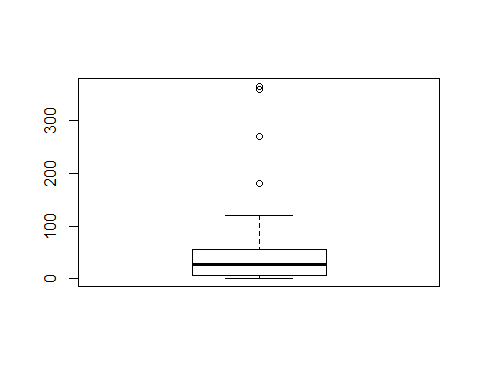
q = quantile(conc\_data$fine\_aggregate, probs=c(.25, .75))  
  
fa\_lower\_boundry = q[1] - (1.5 \* IQR(conc\_data$fine\_aggregate))  
  
fa\_upper\_boundry = q[2] + (1.5 \* IQR(conc\_data$fine\_aggregate))  
  
sum(conc\_data$fine\_aggregate > fa\_upper\_boundry | conc\_data$fine\_aggregate < fa\_lower\_boundry)

## [1] 5

boxplot(conc\_data$fine\_aggregate[conc\_data$fine\_aggregate < fa\_upper\_boundry & conc\_data$fine\_aggregate > fa\_lower\_boundry])



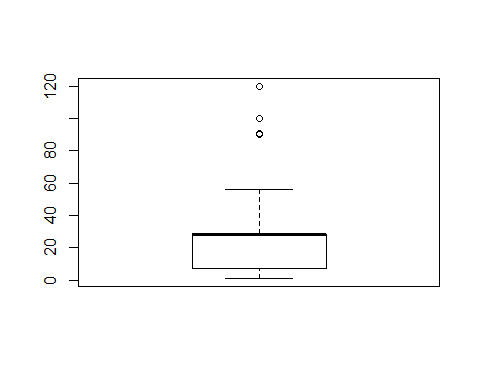
#Finding Outlier - age  
  
boxplot(conc\_data$age)



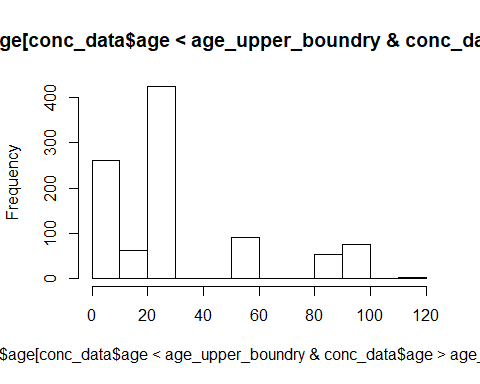
q = quantile(conc\_data$age, probs=c(.25, .75))  
  
age\_lower\_boundry = q[1] - (1.5 \* IQR(conc\_data$age))  
  
age\_upper\_boundry = q[2] + (1.5 \* IQR(conc\_data$age))  
  
sum(conc\_data$age > age\_upper\_boundry | conc\_data$age < age\_lower\_boundry)

## [1] 59

boxplot(conc\_data$age[conc\_data$age < age\_upper\_boundry & conc\_data$age > age\_lower\_boundry])



hist(conc\_data$age[conc\_data$age < age\_upper\_boundry & conc\_data$age > age\_lower\_boundry])



sum(conc\_data$age > 80)

## [1] 190

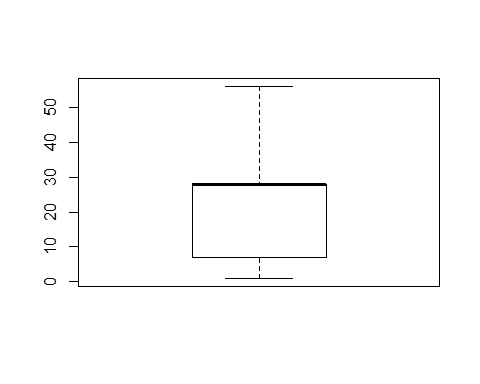
summary(conc\_data$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 7.00 28.00 45.66 56.00 365.00

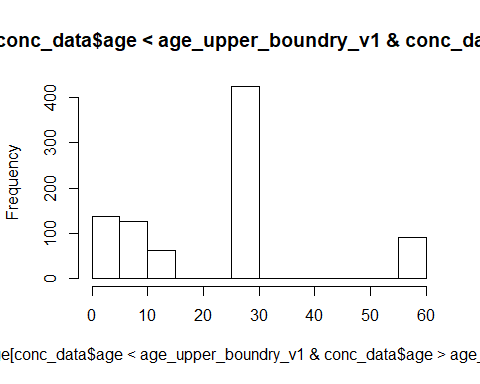
age\_v1 = conc\_data$age[conc\_data$age < age\_upper\_boundry & conc\_data$age > age\_lower\_boundry]  
  
  
q = quantile(age\_v1, probs=c(.25, .75))  
  
age\_lower\_boundry\_v1 = q[1] - (1.5 \* IQR(age\_v1))  
  
age\_upper\_boundry\_v1 = q[2] + (1.5 \* IQR(age\_v1))  
  
sum(conc\_data$age > age\_upper\_boundry\_v1 | conc\_data$age < age\_lower\_boundry\_v1)

## [1] 190

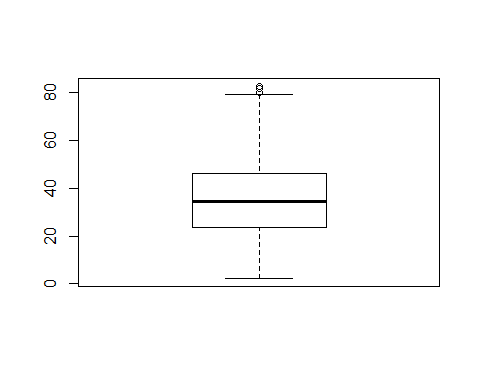
boxplot(conc\_data$age[conc\_data$age < age\_upper\_boundry\_v1 & conc\_data$age > age\_lower\_boundry\_v1])



hist(conc\_data$age[conc\_data$age < age\_upper\_boundry\_v1 & conc\_data$age > age\_lower\_boundry\_v1])



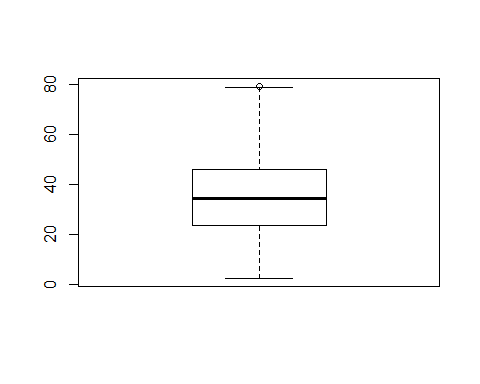
#Finding Outlier - Dependent Variable - compressive\_strength  
  
boxplot(conc\_data$compressive\_strength)



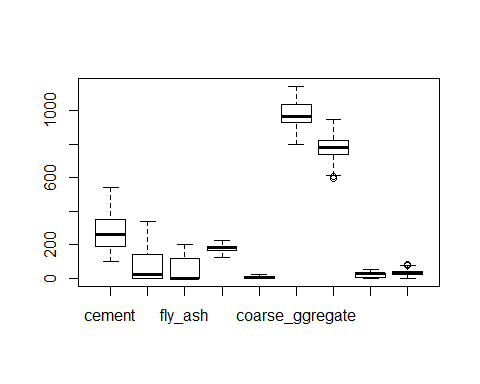
q = quantile(conc\_data$compressive\_strength, probs=c(.25, .75))  
  
lower\_boundry = q[1] - (1.5 \* IQR(conc\_data$compressive\_strength))  
  
upper\_boundry = q[2] + (1.5 \* IQR(conc\_data$compressive\_strength))  
  
sum(conc\_data$compressive\_strength > upper\_boundry | conc\_data$compressive\_strength < lower\_boundry)

## [1] 4

boxplot(conc\_data$compressive\_strength[conc\_data$compressive\_strength < upper\_boundry & conc\_data$compressive\_strength > lower\_boundry])



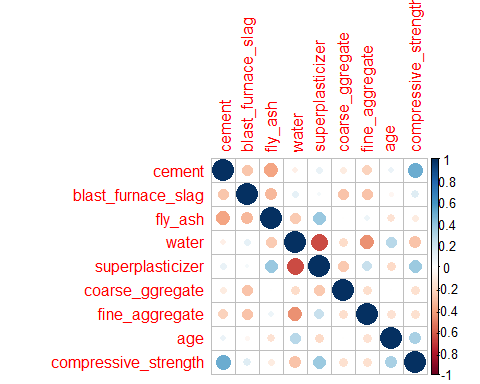
# Removing outliers from dataset - (only Independent variables)  
  
conc\_data\_clean = filter(conc\_data, (blast\_furnace\_slag > bfs\_lower\_boundry & blast\_furnace\_slag < bfs\_upper\_boundry)  
 & (water > w\_lower\_boundry & water < w\_upper\_boundry)  
 & (superplasticizer > sp\_lower\_boundry & superplasticizer < sp\_upper\_boundry)  
 & (fine\_aggregate > fa\_lower\_boundry & fine\_aggregate < fa\_upper\_boundry)  
 & (age > age\_lower\_boundry\_v1 & age < age\_upper\_boundry\_v1))  
  
  
  
boxplot(conc\_data\_clean)



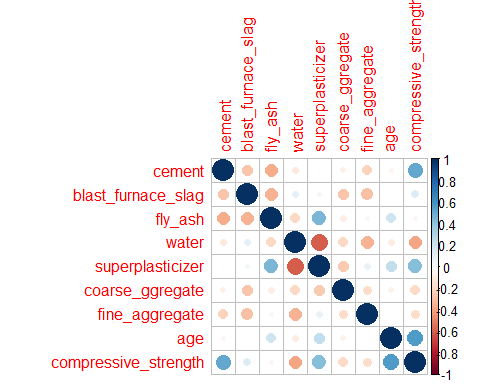
summary(conc\_data\_clean)

## cement blast\_furnace\_slag fly\_ash water   
## Min. :102.0 Min. : 0.00 Min. : 0.00 Min. :127.0   
## 1st Qu.:190.3 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.:164.9   
## Median :263.2 Median : 22.00 Median : 0.00 Median :183.0   
## Mean :276.5 Mean : 74.77 Mean : 59.53 Mean :180.2   
## 3rd Qu.:349.0 3rd Qu.:145.00 3rd Qu.:118.90 3rd Qu.:192.0   
## Max. :540.0 Max. :342.10 Max. :200.10 Max. :228.0   
## superplasticizer coarse\_ggregate fine\_aggregate age   
## Min. : 0.000 Min. : 801.0 Min. :594.0 Min. : 1.00   
## 1st Qu.: 0.000 1st Qu.: 931.0 1st Qu.:736.1 1st Qu.: 7.00   
## Median : 7.000 Median : 968.0 Median :778.5 Median :28.00   
## Mean : 6.431 Mean : 973.3 Mean :774.9 Mean :22.76   
## 3rd Qu.:10.370 3rd Qu.:1036.8 3rd Qu.:820.8 3rd Qu.:28.00   
## Max. :22.100 Max. :1145.0 Max. :945.0 Max. :56.00   
## compressive\_strength  
## Min. : 2.332   
## 1st Qu.:20.981   
## Median :31.730   
## Mean :33.088   
## 3rd Qu.:42.391   
## Max. :81.751

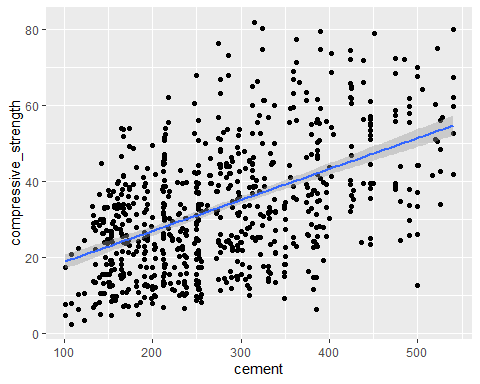
#EDA  
  
cor\_matrix = cor(conc\_data)  
  
corrplot(cor\_matrix)



cor\_matrix\_clean = cor(conc\_data\_clean)  
  
corrplot(cor\_matrix\_clean)

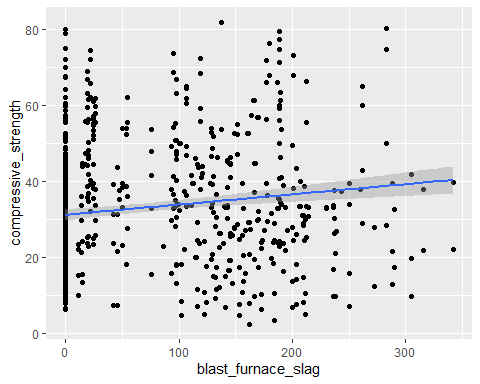


#Correlation between independent and dependent variables  
  
ggplot(conc\_data\_clean, aes(x=cement, y=compressive\_strength)) + geom\_point() + geom\_smooth(method=lm)



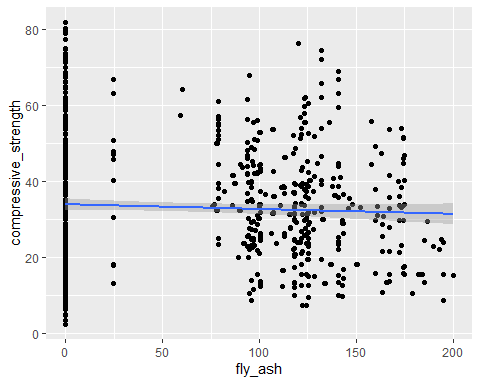
#Inference - Cement has positive correlation with comprehensive strengrh - when amount of cement increase, we can see the increase of strength

ggplot(conc\_data\_clean, aes(x=blast\_furnace\_slag, y=compressive\_strength)) + geom\_point() + geom\_smooth(method=lm)



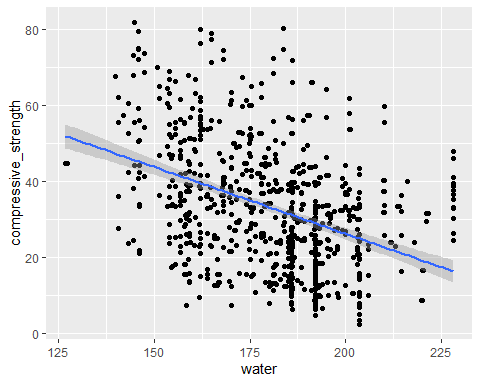
#Inference - no relation between blast\_furnace\_slag and compressive\_strength

ggplot(conc\_data\_clean, aes(x=fly\_ash, y=compressive\_strength)) + geom\_point() + geom\_smooth(method=lm)



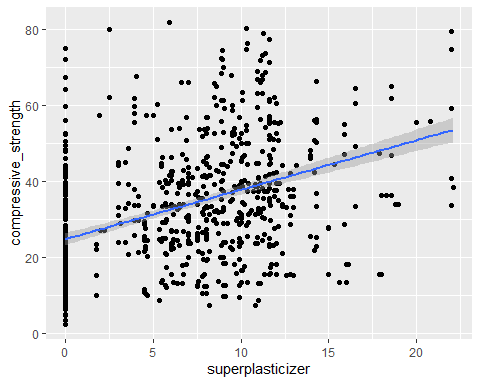
#Inference - no relation between fly\_ash and compressive\_strength

ggplot(conc\_data\_clean, aes(x=water, y=compressive\_strength)) + geom\_point() + geom\_smooth(method=lm)



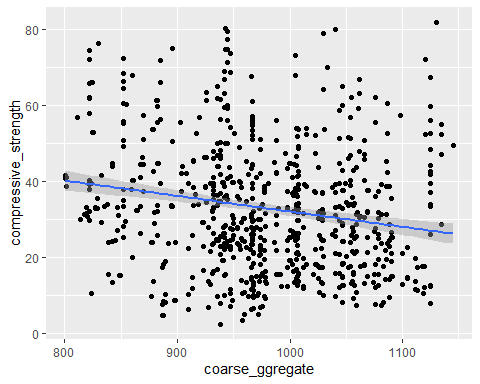
#Inference - There is a negative correlation between water and compressive\_strength

ggplot(conc\_data\_clean, aes(x=superplasticizer, y=compressive\_strength)) + geom\_point() + geom\_smooth(method=lm)



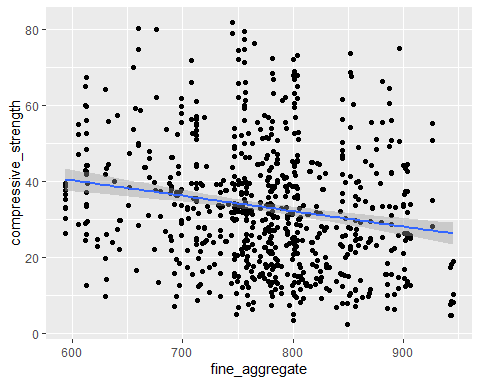
#Inference - We can see a positve correlation between superplasticizer and compressive\_strength

ggplot(conc\_data\_clean, aes(x=coarse\_ggregate, y=compressive\_strength)) + geom\_point() + geom\_smooth(method=lm)



#Inference - Very less negative correlation between coarse\_ggregate and compressive\_strength

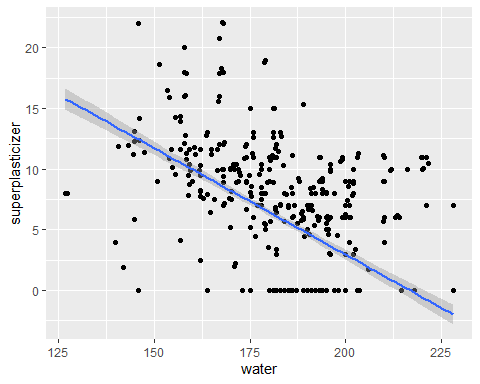
ggplot(conc\_data\_clean, aes(x=fine\_aggregate, y=compressive\_strength)) + geom\_point() + geom\_smooth(method=lm)



#Inference - Very less negative correlation between fine\_aggregate and compressive\_strength

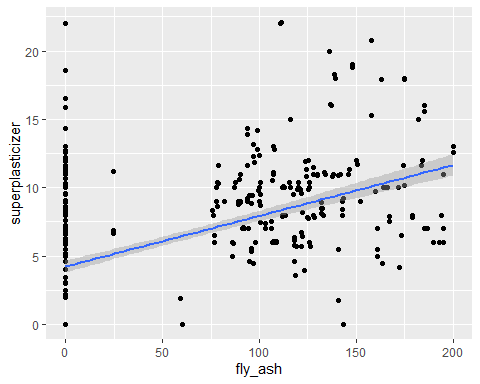
#Correlation between independent variables - to check multicollinearity

ggplot(conc\_data\_clean, aes(x=water, y=superplasticizer)) + geom\_point() + geom\_smooth(method=lm)



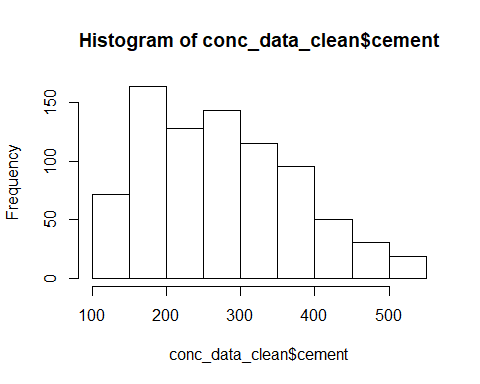
#Inference - water is negatively correlated with superplasticizer - This is as per the fact that superplasticizer is used to reduce water quantity in concrete

ggplot(conc\_data\_clean, aes(x=fly\_ash, y=superplasticizer)) + geom\_point() + geom\_smooth(method=lm)

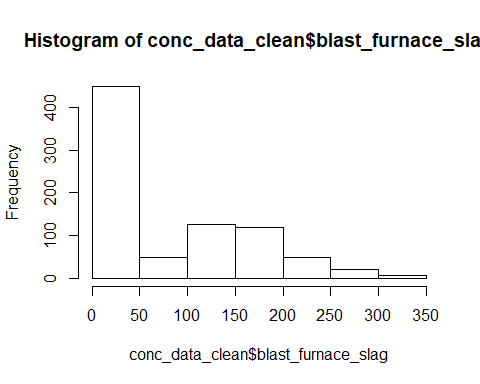


#Inference - fly\_ash has positive correlation with superplasticizer, which may result in multicollinearity

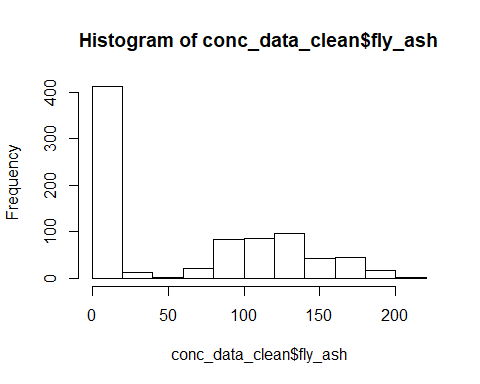
hist(conc\_data\_clean$cement)



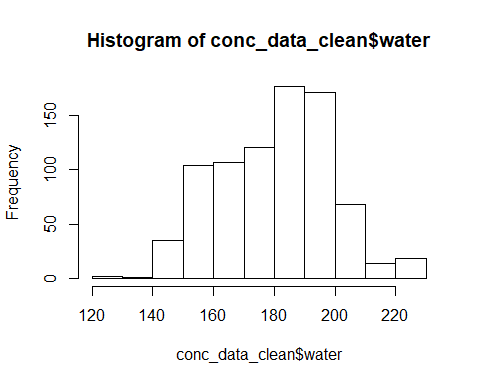
hist(conc\_data\_clean$blast\_furnace\_slag)



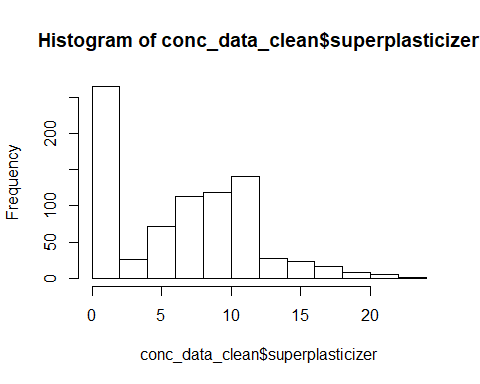
hist(conc\_data\_clean$fly\_ash)



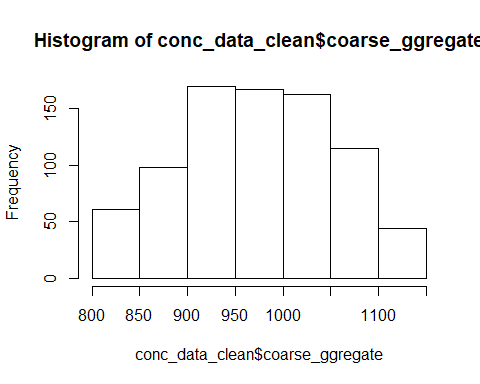
hist(conc\_data\_clean$water)



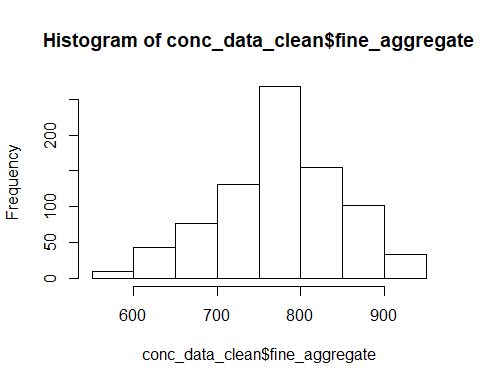
hist(conc\_data\_clean$superplasticizer)



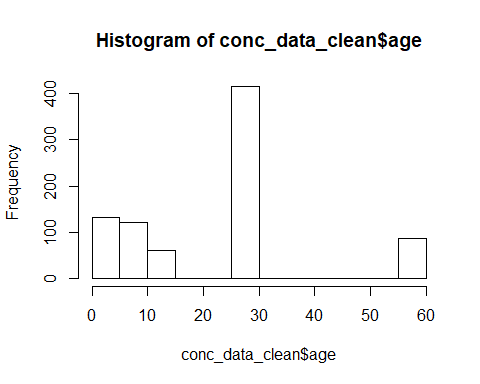
hist(conc\_data\_clean$coarse\_ggregate)



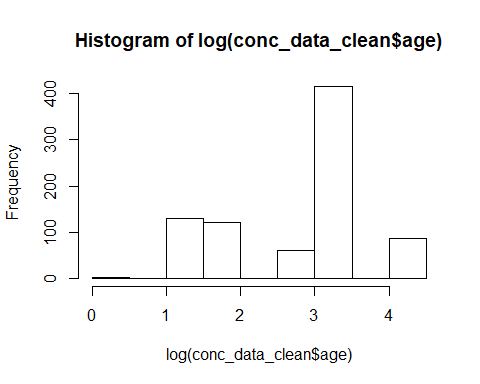
hist(conc\_data\_clean$fine\_aggregate)



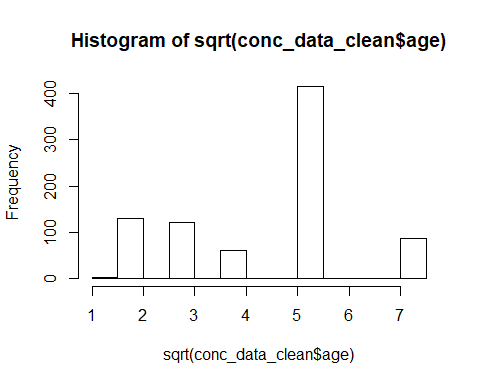
hist(conc\_data\_clean$age)



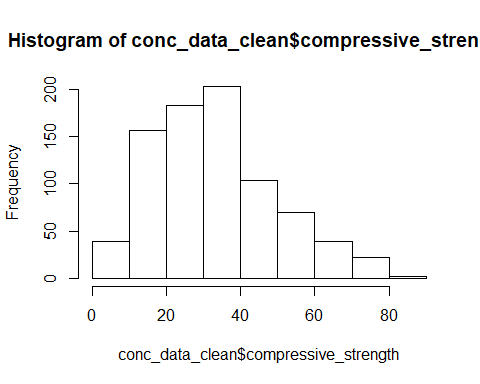
hist(log(conc\_data\_clean$age))



hist(sqrt(conc\_data\_clean$age))



hist(conc\_data\_clean$compressive\_strength)



names(conc\_data\_clean)

## [1] "cement" "blast\_furnace\_slag" "fly\_ash"   
## [4] "water" "superplasticizer" "coarse\_ggregate"   
## [7] "fine\_aggregate" "age" "compressive\_strength"

#Test and Train Data Split  
  
  
set.seed(123) # set seed to ensure you always have same random numbers generated  
  
sample = sample.split(conc\_data\_clean,SplitRatio = 0.80)   
  
train\_data =subset(conc\_data\_clean,sample ==TRUE)

## Warning: Length of logical index must be 1 or 818, not 9

test\_data =subset(conc\_data\_clean,sample ==FALSE)

## Warning: Length of logical index must be 1 or 818, not 9

linear\_eq = lm(compressive\_strength ~ cement + blast\_furnace\_slag + fly\_ash + water + superplasticizer + coarse\_ggregate + fine\_aggregate + age, data = train\_data)  
  
summary(linear\_eq)

##   
## Call:  
## lm(formula = compressive\_strength ~ cement + blast\_furnace\_slag +   
## fly\_ash + water + superplasticizer + coarse\_ggregate + fine\_aggregate +   
## age, data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.2350 -4.1800 -0.2122 3.6752 31.0660   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 19.6655530 24.5033848 0.803 0.423   
## cement 0.1109035 0.0075882 14.615 < 2e-16 \*\*\*  
## blast\_furnace\_slag 0.0810362 0.0091603 8.846 < 2e-16 \*\*\*  
## fly\_ash 0.0457349 0.0111808 4.090 4.87e-05 \*\*\*  
## water -0.2237279 0.0379983 -5.888 6.38e-09 \*\*\*  
## superplasticizer 0.0465466 0.0947029 0.492 0.623   
## coarse\_ggregate 0.0006942 0.0086228 0.081 0.936   
## fine\_aggregate 0.0009541 0.0097521 0.098 0.922   
## age 0.5590634 0.0196533 28.446 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.285 on 627 degrees of freedom  
## Multiple R-squared: 0.8048, Adjusted R-squared: 0.8023   
## F-statistic: 323.2 on 8 and 627 DF, p-value: < 2.2e-16

#VIF - Variance Inflation Factor to check Multicollinearity  
  
vif(linear\_eq)

## cement blast\_furnace\_slag fly\_ash water   
## 7.271422 7.609270 6.221441 5.778825   
## superplasticizer coarse\_ggregate fine\_aggregate age   
## 3.005247 5.661650 6.155691 1.086282

linear\_eq\_v1 = lm(compressive\_strength ~ cement + fly\_ash + water + superplasticizer + coarse\_ggregate + fine\_aggregate + age, data = train\_data)  
  
summary(linear\_eq\_v1)

##   
## Call:  
## lm(formula = compressive\_strength ~ cement + fly\_ash + water +   
## superplasticizer + coarse\_ggregate + fine\_aggregate + age,   
## data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26.142 -4.533 -0.616 4.435 33.113   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 205.142752 13.439230 15.264 < 2e-16 \*\*\*  
## cement 0.051167 0.003668 13.949 < 2e-16 \*\*\*  
## fly\_ash -0.039893 0.005931 -6.726 3.92e-11 \*\*\*  
## water -0.444505 0.030365 -14.639 < 2e-16 \*\*\*  
## superplasticizer 0.021811 0.100316 0.217 0.828   
## coarse\_ggregate -0.061522 0.005287 -11.637 < 2e-16 \*\*\*  
## fine\_aggregate -0.072781 0.005365 -13.565 < 2e-16 \*\*\*  
## age 0.551591 0.020808 26.509 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.72 on 628 degrees of freedom  
## Multiple R-squared: 0.7804, Adjusted R-squared: 0.778   
## F-statistic: 318.9 on 7 and 628 DF, p-value: < 2.2e-16

vif(linear\_eq\_v1)

## cement fly\_ash water superplasticizer   
## 1.513077 1.558731 3.286052 3.002627   
## coarse\_ggregate fine\_aggregate age   
## 1.895264 1.659050 1.084275

linear\_eq\_v1\_1 = lm(compressive\_strength ~ cement + fly\_ash + water + superplasticizer + coarse\_ggregate + fine\_aggregate + age, data = train\_data)  
  
summary(linear\_eq\_v1\_1)

##   
## Call:  
## lm(formula = compressive\_strength ~ cement + fly\_ash + water +   
## superplasticizer + coarse\_ggregate + fine\_aggregate + age,   
## data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26.142 -4.533 -0.616 4.435 33.113   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 205.142752 13.439230 15.264 < 2e-16 \*\*\*  
## cement 0.051167 0.003668 13.949 < 2e-16 \*\*\*  
## fly\_ash -0.039893 0.005931 -6.726 3.92e-11 \*\*\*  
## water -0.444505 0.030365 -14.639 < 2e-16 \*\*\*  
## superplasticizer 0.021811 0.100316 0.217 0.828   
## coarse\_ggregate -0.061522 0.005287 -11.637 < 2e-16 \*\*\*  
## fine\_aggregate -0.072781 0.005365 -13.565 < 2e-16 \*\*\*  
## age 0.551591 0.020808 26.509 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.72 on 628 degrees of freedom  
## Multiple R-squared: 0.7804, Adjusted R-squared: 0.778   
## F-statistic: 318.9 on 7 and 628 DF, p-value: < 2.2e-16

vif(linear\_eq\_v1\_1)

## cement fly\_ash water superplasticizer   
## 1.513077 1.558731 3.286052 3.002627   
## coarse\_ggregate fine\_aggregate age   
## 1.895264 1.659050 1.084275

linear\_eq\_v2 = lm(compressive\_strength ~ cement + fly\_ash + water + coarse\_ggregate + fine\_aggregate + age, data = train\_data)  
  
summary(linear\_eq\_v2)

##   
## Call:  
## lm(formula = compressive\_strength ~ cement + fly\_ash + water +   
## coarse\_ggregate + fine\_aggregate + age, data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26.246 -4.502 -0.619 4.412 33.046   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 207.076095 10.069271 20.565 < 2e-16 \*\*\*  
## cement 0.051075 0.003641 14.027 < 2e-16 \*\*\*  
## fly\_ash -0.039494 0.005635 -7.009 6.19e-12 \*\*\*  
## water -0.449265 0.021022 -21.371 < 2e-16 \*\*\*  
## coarse\_ggregate -0.062167 0.004374 -14.213 < 2e-16 \*\*\*  
## fine\_aggregate -0.073193 0.005015 -14.595 < 2e-16 \*\*\*  
## age 0.552197 0.020605 26.799 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.714 on 629 degrees of freedom  
## Multiple R-squared: 0.7804, Adjusted R-squared: 0.7783   
## F-statistic: 372.6 on 6 and 629 DF, p-value: < 2.2e-16

vif(linear\_eq\_v2)

## cement fly\_ash water coarse\_ggregate fine\_aggregate   
## 1.493213 1.409038 1.577335 1.299134 1.451750   
## age   
## 1.064825

linear\_eq\_v3 = lm(compressive\_strength ~ cement + fly\_ash + superplasticizer + coarse\_ggregate + fine\_aggregate + age, data = train\_data)  
  
summary(linear\_eq\_v3)

##   
## Call:  
## lm(formula = compressive\_strength ~ cement + fly\_ash + superplasticizer +   
## coarse\_ggregate + fine\_aggregate + age, data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.723 -5.131 -0.853 4.704 42.226   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 29.817347 7.054734 4.227 2.72e-05 \*\*\*  
## cement 0.070157 0.003970 17.670 < 2e-16 \*\*\*  
## fly\_ash -0.039095 0.006863 -5.697 1.88e-08 \*\*\*  
## superplasticizer 1.080738 0.080426 13.438 < 2e-16 \*\*\*  
## coarse\_ggregate -0.011847 0.004691 -2.525 0.0118 \*   
## fine\_aggregate -0.027845 0.005092 -5.468 6.56e-08 \*\*\*  
## age 0.548709 0.024078 22.789 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.934 on 629 degrees of freedom  
## Multiple R-squared: 0.7055, Adjusted R-squared: 0.7027   
## F-statistic: 251.2 on 6 and 629 DF, p-value: < 2.2e-16

vif(linear\_eq\_v3)

## cement fly\_ash superplasticizer coarse\_ggregate   
## 1.323821 1.558599 1.441288 1.114468   
## fine\_aggregate age   
## 1.115964 1.084178

#MAPE - Mean Absolute Percentage Error - linear\_eq\_v3  
  
model\_result = predict(linear\_eq\_v3, test\_data[,c('cement','fly\_ash','superplasticizer','coarse\_ggregate','fine\_aggregate','age')])  
  
  
model\_result= as.data.frame(model\_result)  
  
output <- cbind(test\_data, model\_result)  
  
output$diff = output$compressive\_strength - output$model\_result  
  
output$error = abs(output$diff/ output$compressive\_strength)\* 100  
  
MAPE<-mean(output$error)  
  
MAPE

## [1] 29.94019

#MAPE - Mean Absolute Percentage Error - linear\_eq\_v2  
  
model\_result = predict(linear\_eq\_v2, test\_data[,c('cement','fly\_ash','water','coarse\_ggregate','fine\_aggregate','age')])  
  
  
model\_result= as.data.frame(model\_result)  
  
output <- cbind(test\_data, model\_result)  
  
output$diff = output$compressive\_strength - output$model\_result  
  
output$error = abs(output$diff/ output$compressive\_strength)\* 100  
  
MAPE<-mean(output$error)  
  
MAPE

## [1] 23.61674

#MAPE - Mean Absolute Percentage Error - linear\_eq  
  
model\_result = predict(linear\_eq, test\_data[,c('cement','blast\_furnace\_slag','fly\_ash','water', 'superplasticizer','coarse\_ggregate','fine\_aggregate','age')])  
  
  
model\_result= as.data.frame(model\_result)  
  
output <- cbind(test\_data, model\_result)  
  
output$diff = output$compressive\_strength - output$model\_result  
  
output$error = abs(output$diff/ output$compressive\_strength)\* 100  
  
MAPE<-mean(output$error)  
  
MAPE

## [1] 21.11366