PROG8430

Assignment02

Fei Yun

8680643

Data Transformation  
1. As demonstrated in class, transform any variables that are required to conduct the regression analysis.

names(Diamond)<-c('Price\_FY','Carat.Size\_FY','Source\_FY','Year\_FY','Clar\_FY','Col\_FY','Cut\_FY','Val\_FY')

Descriptive Data Analysis

1. Create numeric and graphical summaries of the data (as demonstrated in class).

summary(Diamond)

A close up of text on a white background

Description automatically generated

par(mfrow=c(3,3))

sapply(names(Diamond), function(cname){

if(is.numeric(Diamond[[cname]]))

print(hist(Diamond[[cname]],main=cname))

})

par(mfrow=c(1,1))

A close up of a organ

Description automatically generated

1. Comment on anything noteworthy or unusual. You are looking for

distributions that seem reasonable and reflective of the data you are analysing.

From the Histgram we can see the price of Diamond is not following normal distribution.

Outliers

1. Create boxplots of all relevant variables to determine outliers.

par(mfrow=c(3,3))

sapply(names(Diamond), function(cname){

if(is.numeric(Diamond[[cname]]))

print(boxplot(Diamond[[cname]],main=cname))

})

par(mfrow=c(1,1))

Box and whisker chart

Description automatically generated

1. Comment on any outliers you see.

From Cut and Clar and Carat.Size and Val outliers we can see they have extreme values out of max and min values in this variables.

Exploratory Analysis

1. Create QQNorm plots and numeric tests for normality of data and

identify data that seems to be normal and not anything else that seems

remarkable. If none do, state that.

par(mfrow=c(3,3))

sapply(names(Diamond), function(cname){

if(is.numeric(Diamond[[cname]]))

print(qqnorm(Diamond[[cname]],main=cname))

})

par(mfrow=c(1,1))

Shape

Description automatically generated

NumDiamd<-Diamond[-c(3)]

DiaNrm<-lapply(NumDiamd,shapiro.test)

DiaNrm

str(DiaNrm[[4]])

DiaRes<-sapply(DiaNrm, '[',c("statistic","p.value"))

DiaRest<-t(DiaRes)

DiaRest

Text

Description automatically generated

From this statistic and p.value we can see none of those variables is following normal disturibute.

2 Correlations: Create both numeric and graphical correlations (as demonstrated) and comment on noteworthy correlations you observe. Are these surprising? Do they make sense?

library(corrgram)

corrgram(NumDiamd,order=TRUE,lower.panel=panel.shade,upper.panel=panel.pie,

text.panel=panel.txt,main="Diamond stats")

A picture containing diagram

Description automatically generated

res<-cor(NumDiamd,method = "spearman")

round(res,2)

Table

Description automatically generated

From correlation we can see Price have strong positive relation with Carat.size. And carat size has strong positive relation with its value. The year of first cut almost has nothing relation with anything. But Clarity has negative relation with price and caratsize. And we can see the price the diamond sold for almost euqual its value for insurance.