PREDICT 412 - Assignment 2

Alain Bonacossa - Oct 19, 2014

1 Introduction and Modeling Problem

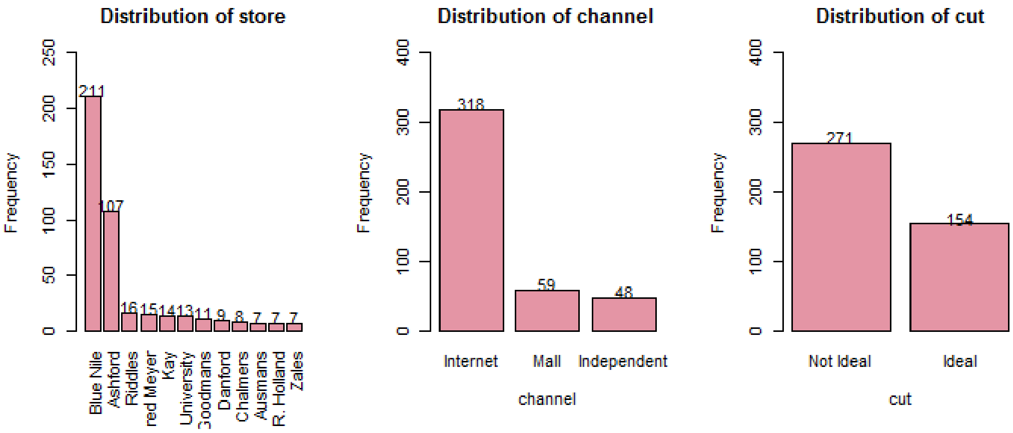
Variable	Description	Coding
carat	Diamond's weight in carats	
color	Color grade	10 levels; D = 1 to M = 10
clarity	Purity of diamond	11 levels; FL = 1 to I3 = 11
cut	Quality of cut	2 levels; Not Ideal = 0, Ideal = 1
channel	Type of Jeweler	3 levels; Mall = 0, Independent = 1, Internet = 2
store	Store	12 levels
price	Price in U.S. dollars	

Table 1: Variable Description

2 Data

2.1 Data Quality Check

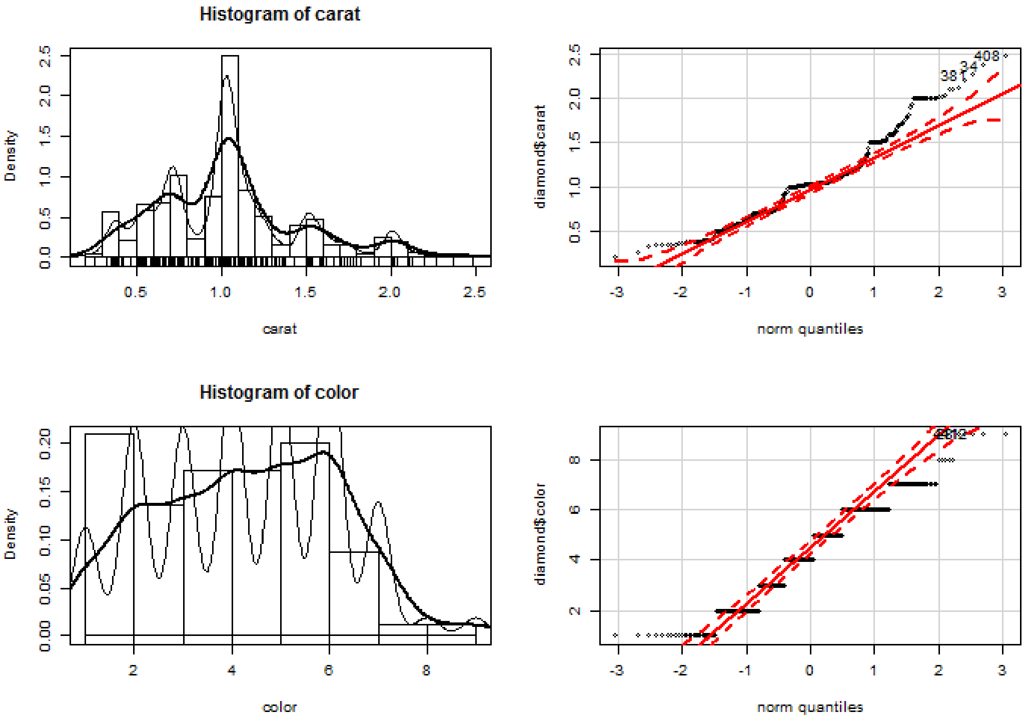
Figure 1: Frequency Distribution for Store, Channel and Cut

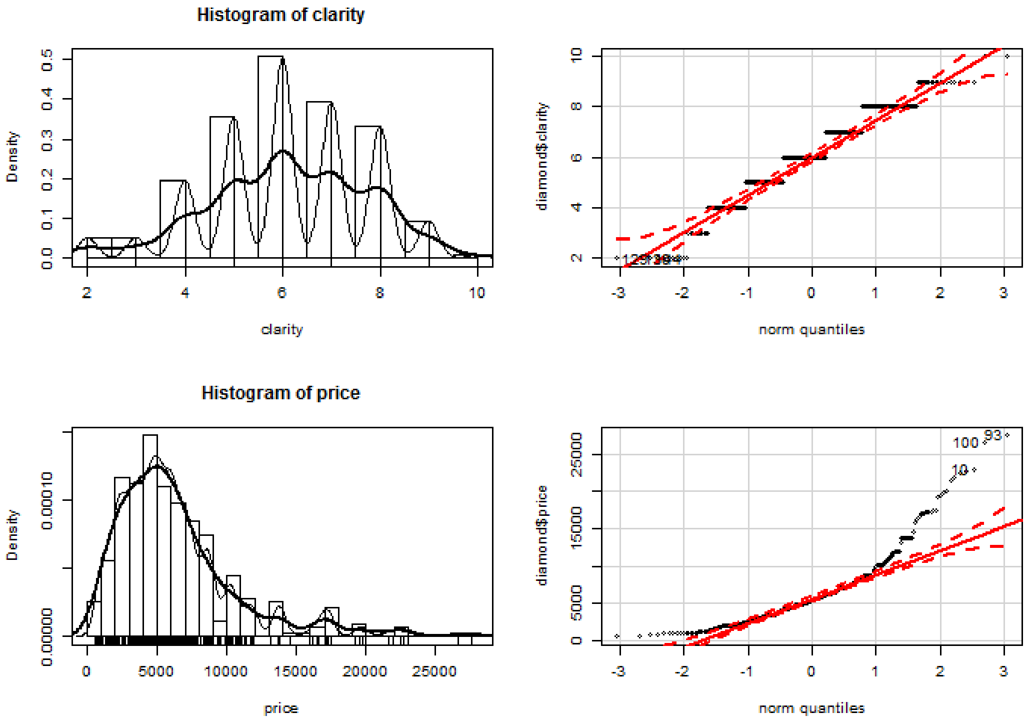


	carat	color	clarity	price
nbr.val	425.00	425.00	425.00	425.00
nbr.null	0.00	0.00	0.00	0.00
nbr.na	0.00	0.00	0.00	0.00
min	0.20	1.00	2.00	497.00
max	2.48	9.00	10.00	27575.00
range	2.28	8.00	8.00	27078.00
sum	442.29	1833.00	2607.00	2701297.00
median	1.02	4.00	6.00	5476.00
mean	1.04	4.31	6.13	6355.99
SE.mean	0.02	0.09	0.08	213.64
CI.mean.0.95	0.04	0.18	0.15	419.92
var	0.18	3.47	2.57	19397306.87
std.dev	0.42	1.86	1.60	4404.24
coef.var	0.41	0.43	0.26	0.69
skewness	0.70	-0.01	-0.30	1.71
skew.2SE	2.97	-0.06	-1.28	7.23
kurtosis	0.43	-0.77	-0.17	3.77
kurt.2SE	0.91	-1.63	-0.35	7.98
normtest.W	0.95	0.95	0.96	0.86
normtest.p	0.00	0.00	0.00	0.00

Table 2: Descriptive Statistics

Figure 2: Histograms and Q-Q plots with Univariate Outliers





2.2 Exploratory Data Analysis

	carat	color	clarity	cut	channel	store
carat						
color	-0.19					
clarity	-0.57	-0.24				
cut	0.02	-0.10	-0.13			
channel	-0.62	0.03	0.18	-0.33		
store	-0.89**	0.10	0.52	-0.13	0.38	
price	0.96***	-0.36	-0.60	-0.04	-0.54	-0.80*

Table 3: Variable Mean by Class

Figure 3: Scatterplot Matrix with Univariate Diastribution Displays and Multivariate Outliers

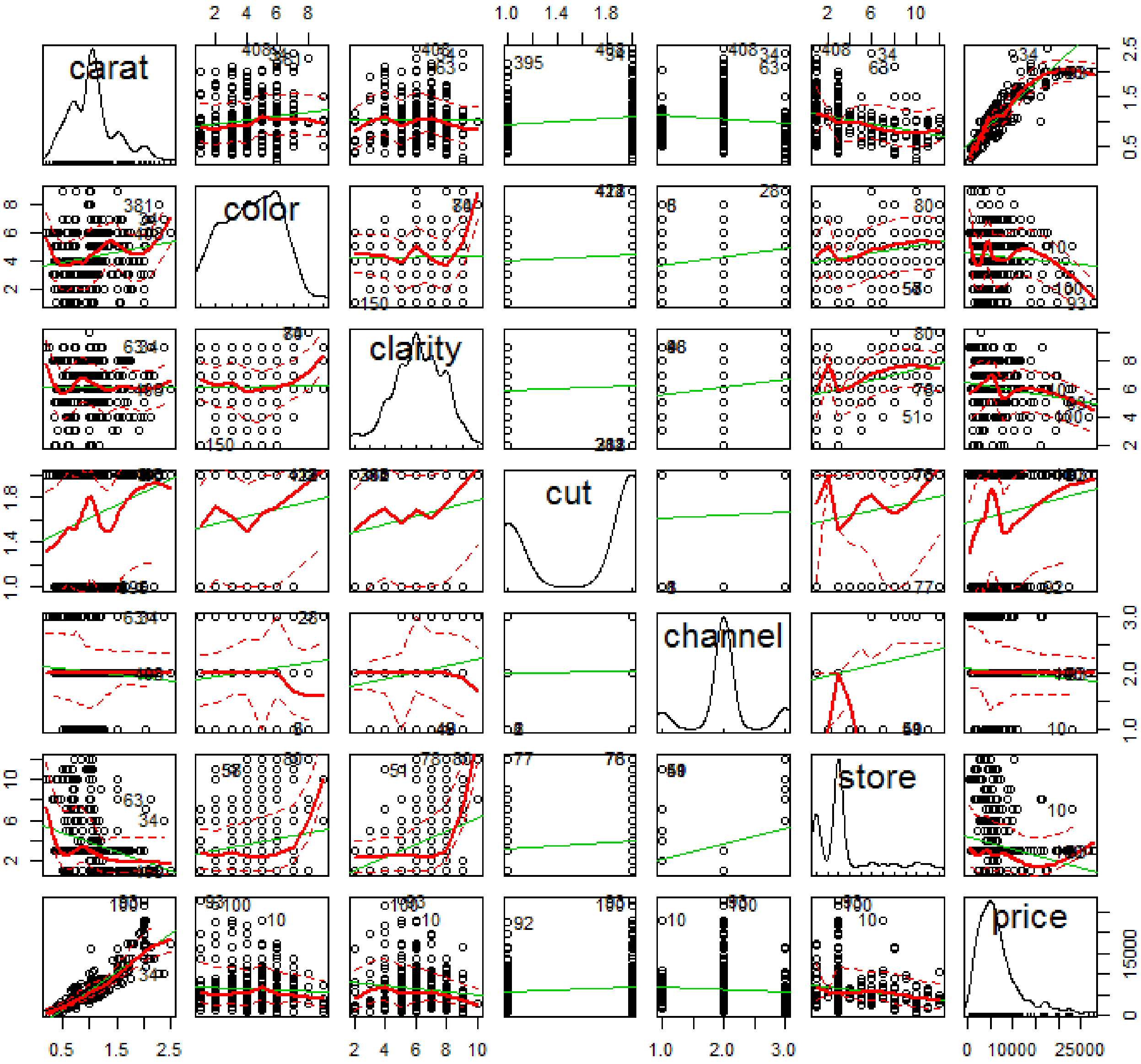
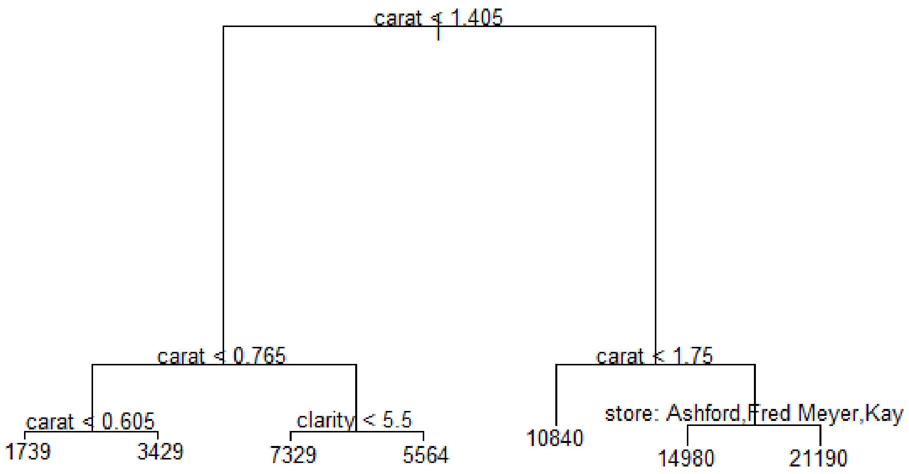


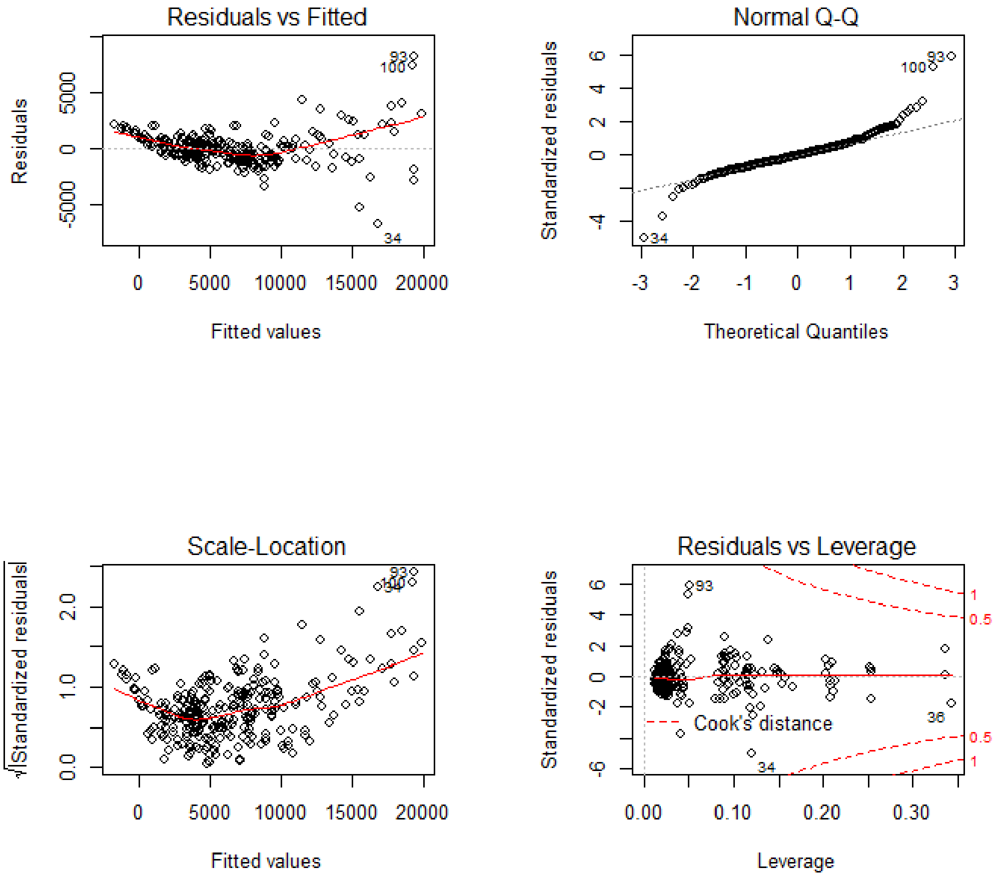
Figure 4: EDA tree plot



3 The Model Build

3.1 Baseline Model

Figure 5: Regression Diagnostics - Baseline Model



Appendix - R code

```r
#Create a 70/30 training-test data split
set.seed(3456)
trainindex=sample(1:nrow(diamond), 298)
train=diamond[trainindex,]
test=diamond[-trainindex,]
```

```r
#Table 2 - Descriptive statistics
library(pastecs)
xtable(stat.desc(diamond[,-c(4:6)], basic=TRUE, desc=TRUE, norm=TRUE))
```

```r
#Frequency of store, channel and cut
par(mfrow=c(1,3))
ds <- rbind(summary(diamond$store))
ord <- order(ds[1,], decreasing=TRUE)
bp <-  barplot(ds[,ord], beside=TRUE, ylab="Frequency", las=3, ylim=c(0, 250),
               col=colorspace::rainbow_hcl(1))
text(bp, ds[,ord]+6, ds[,ord])
title(main="Distribution of store")
```

```r
#Histograms and Box plots for Carat, Color, Clarity and Price
par(mfrow=c(2,2), cex=0.6)
hist(carat, main="Carat")
with(diamond, {
  hist(carat, breaks="FD", freq=FALSE, ylab="Density")
  lines(density(carat), lwd=2)
  lines(density(carat, adjust=0.5), lwd=1)
rug(carat)
box()
})
plot20<-Boxplot(diamond$carat, id.n=5, notch=TRUE, ylab="Carat",
                cex.axis=0.85, col=c("turquoise3"))
```

```r
#Table 4 - Correlation matrix
#corstarsl: http://myowelt.blogspot.com/2008/04/beautiful-correlation-tables-in-r.html
xtable(xtabs(~store+channel, data=diamond))
diamond.matrix<-data.matrix(diamond, rownames.force = NA)
cor(diamond.matrix)
cor_diamond=cor(diamond.matrix, use="complete.obs")
xtable(corstarsl(cor_diamond))
```

```r
#Figure 1 - Scatterplot with variable distribution and outliers
library(car)
scatterplotMatrix(diamond, id.n=3)
```

```r
#Figure 2 - EDA tree plot
library(tree)
tree.data=tree(diamond$price~., diamond)
plot(tree.data)
text(tree.data, pretty=0, cex=0.8)
```