Andrew Resnikoff
Data Analysis Exam 2

Abstract

In today's society, diamonds are seen as a luxury item. They are a very important part of popular jewelry. Probably because of how difficult it is to mine diamonds, there is only a small group of companies that participate in most of the diamond trade. These companies can then set diamond whatever prices they like, making diamonds incredibly expensive. Through this investigation, we hoped to get a better understanding of how diamonds are priced. From 285 diamond samples we used multivariate linear regression to find a relationship between 10 different predictor variables and the price of a diamond. In our conclusions, we found that not all of these predictors were significant. Interestingly, while the weight in carats of a diamond was significant we concluded that increasing the weight of a diamond corresponds to a decrease in price. Based on our analysis, increasing the width of a diamond will correspond to an increase in price.

Understanding the diamond trade can be very important to limiting the power that the few companies involved in the market have. Because our society places such a high value on diamond jewelry, we must be sure that diamonds are being priced fairly and that we are not being taken advantage of by the companies in the industry who set the prices. In this analysis, we will explore different qualities of diamonds and see what their affect on the price of the diamond is. We expect that a diamond with more carats will be more expensive and that wider diamonds will be more expensive.

Exploratory Data Analysis

We are given a dataset with the final price of the diamond in dollars, the weight of the diamond in carats, the graded quality of the cut of the diamond, the color of the diamond, the graded measurement of how clear the diamond is, the length of the diamond in millimeters, the width of the diamond in millimeters, the height of the diamond in millimeters, the width of the top part of the diamond relative to the widest point in percent and the depth of the diamond from the widest point relative to total depth in percent. For this set of 10 variables, we are given 285 observations of sample diamonds.

Table 1: Summary Statistics for Continuous Variables

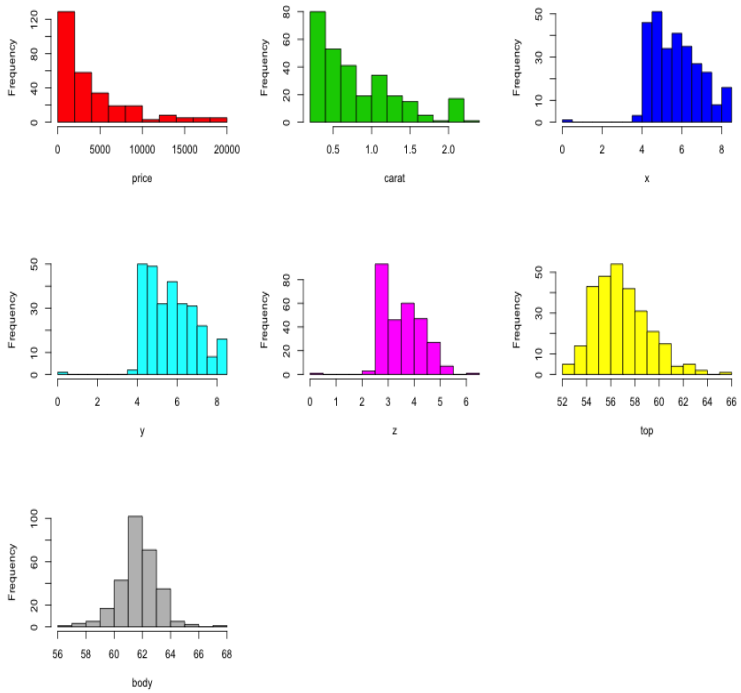
	price	carat	x	y	z	top	body
Min	425	.23	0	0	0	52	56.9
Max	18795	2.28	8.5	8.43	6.16	66	67.5
Median	2276	.7	5.67	5.66	3.5	57	61.8
Mean	4015.905	0.8069123	5.711263	5.708	3.529509	57.352632	61.765614
SD	4295.598	0.5052202	1.21275	1.203136	0.7585242	2.264648	1.330259

Table 2: Summary Statistics for Categorical Variables

cut	percent	color	percent	clarity	percent
Fair	2.105263%	D	12.982456%	I1	1.403509%
Good	10.175439%	E	15.438596%	SI1	23.157895%
Very Good	21.754386%	F	18.596491%	SI2	17.192982%
Ideal	41.403509%	G	23.157895%	VS1	17.543860%
Premium	24.561404%	H	14.736842%	VS2	19.649123%
		I	10.877193%	VVS1	5.614035%
		J	4.210526%	VVS2	12.280702%
				IF	3.157895%

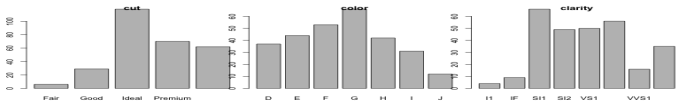
In Table 1 above, summary statistics are shown for each continuous variable. From the histograms in Figure 1, we can get a better look at the distributions of these variables. The distribution of *price*, the response variable, looks skewed right. The distributions of x, y and z all appear to be skewed right but each of them has one observation with a value of 0. It turns out there is one observation that is missing values for x, y and z (as it seems implausible that a diamond with no width, height or depth could be worth anything). The variable *top* seems to have a skewed right distribution and the variable *body* seems to have a skewed left distribution, but these skews are slight so its possible either one could really have a symmetrical distribution.

Figure 1:



From Table 2, we have the distribution of the observation among each of the categorical variables. For *cut*, most of the observations are from the Very Good, Ideal and Premium categories (about 90%). Ideal is the most common cut by about 17%. In *color*, the largest proportion of diamond observations have color grade G, which is halfway between the best and worst rating. As expected, more of the diamonds have a middle color grade and for more extreme color grades (better or worse) there are less observations. Lastly, in *clarity*, very few of the diamonds were of clarity grade I1 (1.5%) or IF (3%), the worst and best grade respectively. The categories SI1,SI2 VS1 and VS2 each contained about 20% of the observations.

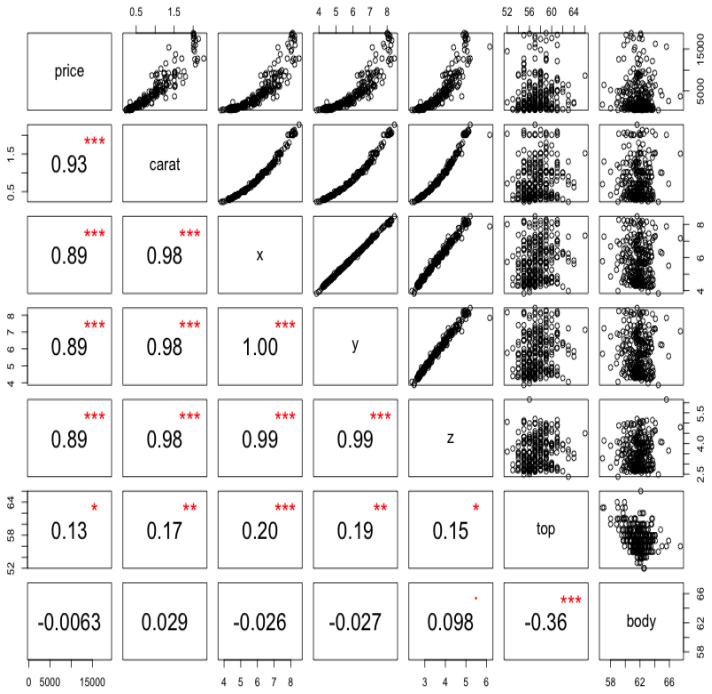
Figure 2:



Before performing multivariate exploratory data analysis, the observation missing x, y and z values is removed from the dataset because the data is not accurate. From the pairs plot in Figure, we can see positive linear relationships between x and y, x and z, and y and z (The correlation for each pair is greater than .99). We also see a strong, positive correlation between *price* and each of *carat*, x, y and z (correlations .93, .89, .89 and .89) respectively. *Top* has the next highest

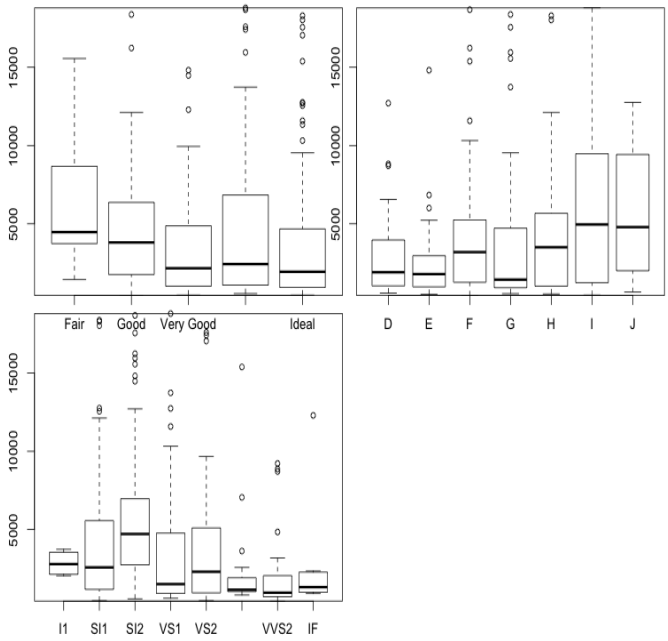
correlation with *price*, at .13 (a weak, positive correlation). Finally, *body* and *price* have a correlation of .006, indicating the potential lack of a relationship there. In fact, there appears to be very little correlation between *body* and any of the other continous predictor variables (very weak correlations). *top* has a weak positive correlation with x, y, z and *carat*.

Figure 3:



Using the boxplots in Figure 4, we can analyze the relationships between each of the categorical variables and *price*. It appears to be that diamonds with a "Fair" cut are worth more money than other cuts (The 25th percentile for the "Fair" cut is greater than the median for all other cuts). The similar boxplots for diamonds with "Very Good" and "Ideal" cuts implies that these cuts have close to the same *price* distribution. For *color*, it looks like, in general, a decrease in color grade (moving in the D to J direction) is associated with an increase in *price*. A *clarity* grade of "SI2" seems to be associated with a greater *price* value. Diamonds with a *clarity* grade of "I1" is in general associated with a greater *price* than diamonds with "IF", "VVS1" or "VVS2" (which all have very similar distributions of *price*). Observations with *clarity* values of "SI1", "VS1" and "VS2" (which all have very similar distributions of *price*) have a similar median *price* value as "I1" but have a much larger interquartile range.

Figure 4:



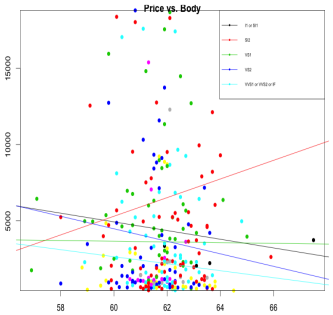
Model

We will begin with a basic model, using *price* as the response and the remaining variables as predictors.

To appropriately incorporate *cut*, *color* and *clarity* into the model, we must decide how they will be coded into the model. For *cut*, "Fair" and "Good" will be collapsed into one category to even out the number of observations in each category. For the same reason, in *clarity*, "I1" and "SI1" will be collapsed into "poorClarity" and "VVS1", "VVS2" and "IF" will be collapsed into "bestClarity". We can collapse these variables the way we have because these variables are ordered categorical, so we can still see how increasing and decreasing a grade for a category will affect the *price*. We use indicator variables (with "poorClarity" as a reference) for *clarity* because it does not seem like there is a trending relation ship in *price* based on moving from better to worse grades (or vice-versa) in *clarity*. It looks like there is a possibility that there is a negative relationship between *cut* and *price* and also between *color* and *price*. These will be coded as continuous variables so that we can see what the change in *price* from one grade to the next best grade.

Looking at a Figure 5, there is no obvious interaction between *body* and *clarity*

Figure 5:



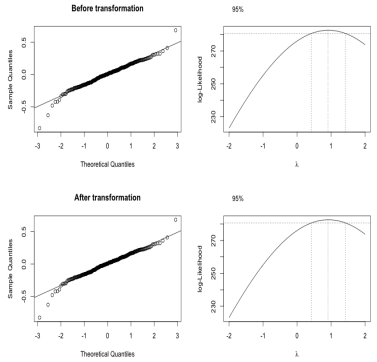
Notice that *carat* is a confounding variable in this analysis because not only are diamonds sold by the number of carats they have but also the weight of the diamond should be related to all of the size measurements (x, y and z) which is supported by the correlation of .98 *carat* shares with each of these variables. Despite the high correlation, the plot of *carat* against each of these other size measurements looks non linear so we cannot exclude all of these measurements. Thus, *carat*, and at least one of x, y and z should be included in the model. But since x, y and z have an almost perfectly positive, linear correlation, we do not need to include all of these variables in our model. The fact that x and y have a correlation of 1 implies that we definitely do not need both of these in our model (as one is just a scalar multiple of the other). This also means neither will predict z better than the other. Since the relationship between x (or y) and z is linear and has a correlation very close to 1, we will arbitrarily include x (instead of y) and exclude y and z from the model.

Now our model is initialized. Our model predicts the *price* of a diamond from the weight (*carat*), length (x), graded cut quality (*cut*, using a collapsed category of "Fair" and "Good" as a reference), graded color (*color*, as an ordered categorical variable), graded clarity (*clarity*, using a collapsed category of "I1" and "SI1" as a reference), the width of the top part of the diamond (*top*) and the depth of the diamond at the widest point (*body*). We now must check to make sure the assumptions of our model are met by testing the diagnostics.

Diagnostics

First we will test to see if our normality assumption is met using the qqnorm plot. From this plot, it appears that the normality assumption is violated. Using the box-cox transformation plot, we can see that a transform with $\lambda = 0$. This implies that taking the logarithm of the model will correct the invalidated assumption. The new qqnorm reflects the corrected normality assumption and the new boxcox plot gives the value of 1 for λ, which implies further transformations are not necessary.

Figure 6:



Looking at the residual plot, the assumption that the errors have a mean of zero with a constant variance appears to be met. Additionally, there do not appear to be any patterns in the residuals. From Figure 8, we see that our categorial variables look good for our assumptions and or independence assumption is met.

Figure 7:

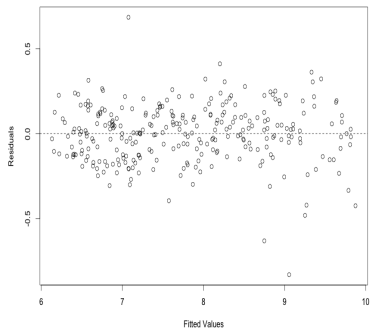
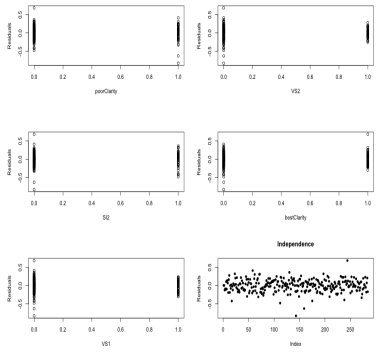


Figure 8:



Model Inference and Results

Since the assumptions in our model have been met, we can now test the appropriateness of our final model. Looking at the P-values in this model, it appears that *top* and *Cut* may not be significant predictors for *price*. We can test this idea using a partial F test. After performing this test,

we fail to reject the null at the .05 level of significance and thus we cannot conclude that *top* and Cut have non-zero coefficients. Thus we can remove these from the model. A table of the summary of our final model is shown below.

Table 3: Final Model Summary

Variable	β (Std. Error)	t-score	P-value
(Intercept)	-1.490437 (0.573143)	-2.600	0.00981
$carat$	-0.861499 (0.106779)	-8.068	2.23e-14
$Color$	-0.085503 (0.006448)	-13.260	$<$ 2e-16
$SI2$	-0.134360 (0.032715)	-4.107	5.29e-05
$VS1$	0.256643 (0.032007)	8.018	3.09e-14
$VS2$	0.159062 (0.030531)	5.210	3.71e-07
$bestClarity$	0.401217 (0.031805)	12.615	$<$ 2e-16
x	1.316891 (0.046716)	28.189	$<$ 2e-16
$body$	0.040585 (0.007887)	5.146	5.08e-07

From this model, we can conclude that a heavier diamond is actually less expensive (β = -0.861499), contrary to our initial belief. Because x has a coefficient of (β = 1.316891) and y had a positive linear relationship with x, this implies that an increase in the width of a diamond does correspond to an increase in price. We also found that the cut of a diamond is not able to significantly predict the price of the diamond. Lastly, we did not find a significant interaction between *body* and *clarity*.

Ultimately, we found that $carat$, $Color$ (as an ordered categorical variable), *clarity* (as a collapsed set of indicator variables), x (the length of the diamond) and *body* were significant predictors for the *price* of a diamond.

R Code

```
# import libraries
library(MASS)
source("~/Desktop/36401/panelfxns.R")

# import data
data<-read.table("~/Desktop/36401/exam2-87.txt")

# get vars
price<-data[,1]
carat<-data[,2]
cut<-data[,3]
color<-data[,4]
clarity<-data[,5]
x<-data[,6]
y<-data[,7]
z<-data[,8]
top<-data[,9]
body<-data[,10]

n = nrow(data)

# Univariate EDA

# Continuous EDA
get_eda<-function(data){
  result<-c(min(data),max(data),median(data),mean(data),sd(data))
  return(result)
}
get_eda(price)
get_eda(carat)
get_eda(x)
get_eda(y)
get_eda(z)
get_eda(top)
get_eda(body)

par(mfrow = c(3,3))

# Histograms
hist(price,col=2,breaks=10,main="")
hist(carat,col=3,breaks=10,main="")
hist(x,col=4,breaks=20,main="")
hist(y,col=5,breaks=20,main="")
hist(z,col=6,breaks=20,main="")
hist(top,col=7,breaks=10,main="")
hist(body,col=8,breaks=10,main="")
```

```
# Categorical EDA
getCategoricalEDA<-function(vector,categories)
{
  index = 1
  result<-c(0)
  for (c in categories)
  {
    num<-length(which(vector==c))
    perc<-100*num/n
    result[index]=perc
    index = index + 1
  }
  return(result)
}

getCategoricalEDA(cut,c("Fair", "Good", "Very Good", "Ideal", "Premium"))
getCategoricalEDA(color,c("D","E","F","G","H","I","J"))
getCategoricalEDA(clarity,c("I1","SI1","SI2","VS1","VS2","VVS1","VVS2","IF"))

# Bar Graphs
par(mfrow = c(3,3))
barplot(table(cut),main="cut")
barplot(table(color),main="color")
barplot(table(clarity),main="clarity")

# Remove bad data
badIndex<-which(x==0 & y==0 & z==0)
data<-data[-badIndex,]

# get vars again
price<-data[,1]
carat<-data[,2]
cut<-data[,3]
color<-data[,4]
clarity<-data[,5]
x<-data[,6]
y<-data[,7]
z<-data[,8]
top<-data[,9]
body<-data[,10]

cut = factor(cut,c("Fair","Good","Very Good","Premium","Ideal"))
clarity = factor(clarity,c("I1","SI1","SI2","VS1","VS2","VVS1","VVS2","IF"))

# adjust n
n<-n-1

# Continuous Bivariate EDA
vars<-cbind(price,carat,x,y,z,top,body)
```

```
pairs(vars,lower.panel=panel.cor)

# Categorical Bivariate EDA
par(mfrow=c(2,2))
boxplot(price~cut,ylab="Price ($)")
boxplot(price~color,ylab="Price ($)")
boxplot(price~clarity,ylab="Price ($)")

# Model

# coding categorical variables

model.cut<-lm(price~cut)
summary(model.cut)

fairAndGood<-as.numeric(cut=="Fair" | cut=="Good")
veryGood<-as.numeric(cut=="Very Good")
premium<-as.numeric(cut=="Premium")
ideal<-as.numeric(cut=="Ideal")

Cut<-rep(NA,length(cut))
Cut[cut=="Fair"] = 0; Cut[cut=="Good"] = 0
Cut[cut=="Very Good"] = 1; Cut[cut=="Premium"] = 2
Cut[cut=="Ideal"] = 3;

Color<-rep(NA,length(color))
Color[color=="D"] = 0; Color[color=="E"] = 1
Color[color=="F"] = 2; Color[color=="G"] = 3
Color[color=="H"] = 4; Color[color=="I"] = 5
Color[color=="J"] = 6

poorClarity<-as.numeric(clarity=="I1" | clarity=="SI1")
SI2<-as.numeric(clarity=="SI2")
VS1<-as.numeric(clarity=="VS1")
VS2<-as.numeric(clarity=="VS2")
bestClarity<-as.numeric(clarity=="VVS1" | clarity=="VVS2" | clarity=="IF")


model.size<-lm(price~x+y+z)
aov(model.size)
model.size<-lm(price~x+z+y)
aov(model.size)
model.size<-lm(price~y+x+z)
aov(model.size)
model.size<-lm(price~y+z+x)
aov(model.size)
model.size<-lm(price~z+x+y)
aov(model.size)
model.size<-lm(price~z+y+x)
```

```
aov(model.size)

model<-lm(price~carat+Cut+Color+SI2+VS1+VS2+bestClarity+x+top+body)
summary(model)

model.2<-lm(price~carat+veryGood+premium+ideal+Color+SI2+VS1+VS2+bestClarity+x+y+z+top+body)
summary(model.2)

par(mfrow = c(1,1))
plot(body,price,pch=16,col=as.numeric(clarity),main="Price vs. Body")

abline(lm(price[poorClarity==1]~body[poorClarity==1]),col=1)
abline(lm(price[SI2==1]~body[SI2==1]),col=2)
abline(lm(price[VS1==1]~body[VS1==1]),col=3)
abline(lm(price[VS2==1]~body[VS2==1]),col=4)
abline(lm(price[bestClarity==1]~body[bestClarity==1]),col=5)
legend("topright",c("I1 or SI1","SI2","VS1","VS2","VVS1 or VVS2 or IF"),col=c(1,2,3,4,5),lwd=1

model<-lm(price.t~carat+Cut+Color+SI2+VS1+VS2+bestClarity+x+top+body)
par(mfrow=c(2,2))
qqnorm(model$res,main="Before transformation"); qqline(model$res)
boxcox(model)

price.t<-log(price)
model.t<-lm(price.t~carat+Cut+Color+SI2+VS1+VS2+bestClarity+x+top+body)
qqnorm(model.t$res,main="After transformation"); qqline(model.t$res)
boxcox(model.t)

plot(model.t$fitted,model.t$res,xlab="Fitted Values",ylab="Residuals",main=""); abline(h=0,lty=

layout(matrix(c(1,2,3,4,5,6),ncol=2))
title("Poor Clarity \n vs. Residuals")
plot(poorClarity,model.t$res,ylab="Residuals")
plot(SI2,model.t$res,ylab="Residuals")
plot(VS1,model.t$res,ylab="Residuals")
plot(VS2,model.t$res,ylab="Residuals")
plot(bestClarity,model.t$res,ylab="Residuals")
plot(model.t$res,pch=16,ylab="Residuals",main="Independence")
abline(h=0, lty=2)

summary(model.t)
aov(lm(price.t~carat+Cut+Color+SI2+VS1+VS2+bestClarity+x+body+top))
ssr.top = .00536
aov(lm(price.t~carat+Color+SI2+VS1+VS2+bestClarity+x+body+top+Cut))
ssr.Cut = .07959
aov(lm(price.t~carat+Color+SI2+VS1+VS2+bestClarity+x+top+Cut+body))
ssr.body = .61961
sse = 7.53497
```

```
f = ((ssr.top + ssr.Cut)/2)/(sse/(n-10))
f
f.star = qf(.95,df1=2,df2=n-10)
f.star
f > f.star

model.final<-lm(price.t~carat+Color+SI2+VS1+VS2+bestClarity+x+body)
summary(model.final)
```