STAT394 Group Project Final Report

Ken MacIver, Tom Tribe, Jundi Yang, Mei Huang

2022-10-09

Contents

1 Abstract

An exploratory data analysis (EDA) of the 'diamonds' dataset was undertaken. Based on the results of the EDA, the techniques of Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Factor Analysis (FA) and Cluster Analysis (CA) were carried out. As these techniques were new to the project team and everyone was keen to learn, it was decided to experiment with each of them (with varying degrees of success).

The primary aim was to find the best predictor variables for 'price'. A secondary aim was to use the new techniques to see if classifications could be made. Other techniques used were single and multiple linear regression and testing to find the best regression model using the Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) tests.

2 Introduction

One of the key skills required of statisticians is the ability to explore a new and unfamiliar dataset and allow that exploration to guide their next steps. A second key skill is knowing which analytical methods to subsequently apply to the dataset. As students enrolled in the Victoria University of Wellington year three Multivariate Statistics paper (STAT394), we were tasked with undertaking a group project designed to give us practice in developing these skills.

We selected the 'diamonds' dataset, which was unfamiliar to us all. 'Diamonds' is available in the ggplot2 package in RStudio and is also freely available on Kaggle ("Diamonds Dataset, Kaggle.com" 2016). It contains information on ten different variables for 53940 diamonds.

Of these ten variables three are categorical while seven are numerical. A full breakdown of the variables is given below.

This project provided the opportunity to face the challenges of working with a new and unfamiliar dataset and to gain experience in trouble-shooting the many issues that were encountered.

2.1 The variables

- Carat: a measure of the diamond's weight. One carat equals 1/5 gram. In this data set there are diamonds with carat values ranging from 0.2 - 5.01.
- Cut: A diamond's cut defines its proportions and its ability to reflect light. This variable has five levels: Fair (lowest quality), Good, Very Good, Premium and Ideal (highest quality).
- Color: A diamond's color refers to how clear/colorless it is. This variable has seven levels: J (lowest quality), I, H, G, F, E, D (highest quality).
- Clarity: measures small imperfections on the surface and within the stone. This variable has eight levels: I1 (lowest clarity), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (highest clarity).
- x: length in mm. In this data set there are diamonds with a length ranging from 0 - 10.74.
- y: width in mm. In this data set there are diamonds with a width ranging from 0 - 58.9.
- z: depth in mm. In this data set there are diamonds with a depth ranging from 0 - 31.8.
- Depth: total depth percentage. This is calculated by dividing the total width by the total depth ($depth = \frac{2 \times z}{(x+y)}$). Depth percentage impacts how light reflects of the diamond. In this data set we have diamonds with a depth percentage ranging from 43% - 79%.
- Table: The flat facet on the top of a diamond is called its table. Table is calculated by dividing table width by the total width. Table percentage impacts how light reflects of the diamond. In this data set there are diamonds with a table percentage ranging from 43% - 95%.
- Price: price of the diamond in United States dollars (USD). In this data set there are diamonds with a price ranging from $326 - $18823.

2.1.1 Why 'diamonds'?

We selected the diamonds dataset because it was easy to understand what the variables were measuring. This was in contrast to some of the other ones we looked at which required domain specific familiarity, such as knowledge about biology or chemistry (or even worse,

bio-chemistry!). This data set also allows us to undertake an investigation from which we glean insights about diamonds that have real-world application and use.

2.1.2 Aims

Our primary aim was to identify which variables best predictor the *price* of a given diamond. We hypothesised that increased diamond size ('x', 'y' and 'z') and weight ('carat') will be positively correlated with diamond price. We also expected to see increased prices for diamonds with higher levels of the categorical variables, all of which reflect the quality of the diamond. We were unsure how diamond depth and table percentage will relate to diamond price.

We also sought to explore the techniques of Principal Component Analysis (PCA), Factor Analysis (FA), Linear Discriminant Analysis (LDA) and Cluster Analysis (CA) that we have learned in STAT394. The diamonds dataset has 280 unique possible interactions (5 x 7 x 8) between different levels of the categorical variables. Our secondary aim was to discover whether one of these techinques (PCA, FA, LDA or CA) could classify the data in a simpler manner, or predict into which level of a category a new observation would fall. Encouraged by our lecturer, we adopted an approach of 'playing' with the techniques with the aim of learning as much as possible, even if a more experienced statistician could see that they were not the correct tools to achieve our aims. We wanted to learn by doing, regardless of whether the results of these methods were successful for our project or not.

3 Methodology

The diamonds data set was accessed via the kaggle.com website ("Diamonds Dataset, Kaggle.com" 2016) while all statistical analysis and reporting was completed using the computer software package R (R Core Team 2017), in RStudio (RStudio Team 2022) and using RMarkdown (Allaire et al. 2020). Github ("Github.com" 2022) was used a repository for storing all relevant documents and code during this project.

Upon loading the data set into R we first used the following code to delete unnecessary columns, ensure categorical variables were treated as factors with set levels and created a subset data set containing only the numeric variables.

```
# Read the data set into R
diamonds <- read.csv("./diamonds.csv", encoding = "UTF-8")
# Remove the index column
diamonds$X <- NULL
# Set categorical variables as factors and set levels
diamonds$cut <- factor(diamonds$cut,
                       levels = c("Fair","Good","Very Good","Premium","Ideal"))
diamonds$color <- factor(diamonds$color,
                         levels = c("J","I","H","G","F","E","D"))
diamonds$clarity <- factor(diamonds$clarity,
                   levels = c("I1","SI2","SI1","VS2","VS1","VVS2","VVS1","IF"))
```

```
# Make data frame of just the numerical variables
diamonds_num <- subset(diamonds, select = c(carat,depth,table,price,x,y,z))
```

When scaling numerical values in our data set we used the `scale()` function.

When taking smaller samples from our data set to improve the interpretability of visual plots we used a seed consisting of either the student identification number of the team member who wrote the code or 1234567890, and the pseudo random number generator Mersenne Twister.

To begin with, we undertook an in-depth Exploratory Data Analysis (EDA) of the data. The key results from the EDA can be found in the results section below.

4 Results

4.1 Consolidated Exploratory Data Analysis (EDA)

The initial EDA was submitted previously as part of intermediary Milestones 3 and 4 for STAT34 and is not reproduced in full here. Key findings included:

- most numerical variables did not follow a Normal distribution. Some followed the beta distribution and some were undefined (according to the Cullen-Frey plots).
- strong positive correlations between a number of the numeric variables ('price', 'carat', 'x', 'y' and 'z'), and a strong negative correlation between 'depth' and 'table'.
- some of the density plots of the numerical variables showing multiple modes (peaks) suggesting that the effects of one or more of the levels of the categorical variables were having a strong influence on the distribution. These multiple modes were even more apparent in the density plots of the log-transformed variables.
- there were more extreme values for the Mahalanobis distances than would be expected for Normally distributed data.

In this section we present again some of the key findings from the consolidated Exploratory Data Analysis as well as findings discovered by team members subsequent to the previous submission of the EDA in Milestones 3 and 4.

4.2 Summaries of Data

Table 1: Summary statistics for 'diamonds' (2 d.p.)

	carat	depth	table	price	x	y	z
sample size	53940.00	53940.00	53940.00	53940.00	53940.00	53940.00	53940.00
minimum	0.20	43.00	43.00	326.00	0.00	0.00	0.00
first quartile	0.40	61.00	56.00	950.00	4.71	4.72	2.91
median	0.70	61.80	57.00	2401.00	5.70	5.71	3.53
mean	0.80	61.75	57.46	3932.80	5.73	5.73	3.54
third quartile	1.04	62.50	59.00	5324.25	6.54	6.54	4.04
maximum	5.01	79.00	95.00	18823.00	10.74	58.90	31.80
IQR	0.64	1.50	3.00	4374.25	1.83	1.82	1.13
standard deviation	0.47	1.43	2.23	3989.44	1.12	1.14	0.71
skewness	1.12	-0.08	0.80	1.62	0.38	2.43	1.52
kurtosis	4.26	8.74	5.80	5.18	2.38	94.21	50.08

A brief look at the summary statistics of the numerical variables (table 1) shows that there are a range of different scales. We also see high skewness values for carat, price, y and z and high kurtosis values for all variables, particularly y and z. Both of these indicate non-normality of our numeric variables. It is also interesting to note that at least one diamond has a zero value for x, y and z. These values are further investigated in the "Outliers and Unusual Points" section below.

4.2.1 Summaries of the Categorical Variables

These summaries are included to emphasise the fact that the categorical variables do not have an even distribution across their different levels. In some cases, this unevenness is quite pronounced.

4.2.2 Tables of counts for the categorical variables

Table 2: Count of 'color'

Level	Count
J	2808
I	5422
H	8304
G	11292
F	9542
E	9797
D	6775

Table 2 shows the number of diamonds in each level of the 'color' variable. As a reminder, 'J' is the lowest level and 'D' is the highest. Note that the most of the diamonds are in the middle of the range of levels.

Table 3: Count of 'cut'

Level	Count
Fair	1610
Good	4906
Very Good	12082
Premium	13791
Ideal	21551

Table 3 gives a breakdown of the how many diamonds are in each level of the 'cut' variable. As a reminder, 'Fair' is the lowest level and 'Ideal' is the highest. We can see that most are in the 'Ideal', with a substantial number also in the 'Premium' and 'Very Good'. The distribution increases markedly as the quality of 'cut' increases.

Table 4: Count of 'clarity'

Level	Count
I1	741
SI2	9194
SI1	13065
VS2	12258
VS1	8171
VVS2	5066
VVS1	3655
IF	1790

Table 4 shows the number of diamonds in each level of the 'clarity' variable. As a reminder, 'I1' is the lowest level and 'IF' is the highest. Note that most of the diamonds are toward the middle of the range of levels.

4.2.3 Distribution of Numeric Variables

While examining histograms, Cullen and Frey Plots, skewness and kurtosis of the numerical variables, we observed evidence that the numerical variables were not normally distributed. The assumption of normality underlies many statistical procedures so we investigated this in more depth using Normal Quantile plots and goodness-of-fit tests. The code for these tests is included below.

```
par(mfrow=c(3,3))
qqnorm(diamonds$carat, xlab = "Observations",
       ylab = "Normal Quantiles", main = "Carat", col = "red")
qqline(diamonds$carat, col = "blue", lwd =2)
qqnorm(diamonds$depth, xlab = "Observations",
       ylab = "Normal Quantiles", main = "Depth", col = "red")
qqline(diamonds$depth, col = "blue", lwd =2)
qqnorm(diamonds$table, xlab = "Observations",
       ylab = "Normal Quantiles", main = "Table", col = "red")
qqline(diamonds$table, col = "blue", lwd =2)
qqnorm(diamonds$price, xlab = "Observations",
       ylab = "Normal Quantiles", main = "Price", col = "red")
qqline(diamonds$price, col = "blue", lwd =2)
qqnorm(diamonds$x, xlab = "Observations",
       ylab = "Normal Quantiles", main = "x", col = "red")
qqline(diamonds$x, col = "blue", lwd =2)
qqnorm(diamonds$y, xlab = "Observations",
       ylab = "Normal Quantiles", main = "y", col = "red")
qqline(diamonds$y, col = "blue", lwd =2)
qqnorm(diamonds$depth, xlab = "Observations",
       ylab = "Normal Quantiles", main = "z", col = "red")
qqline(diamonds$depth, col = "blue", lwd =2)
```

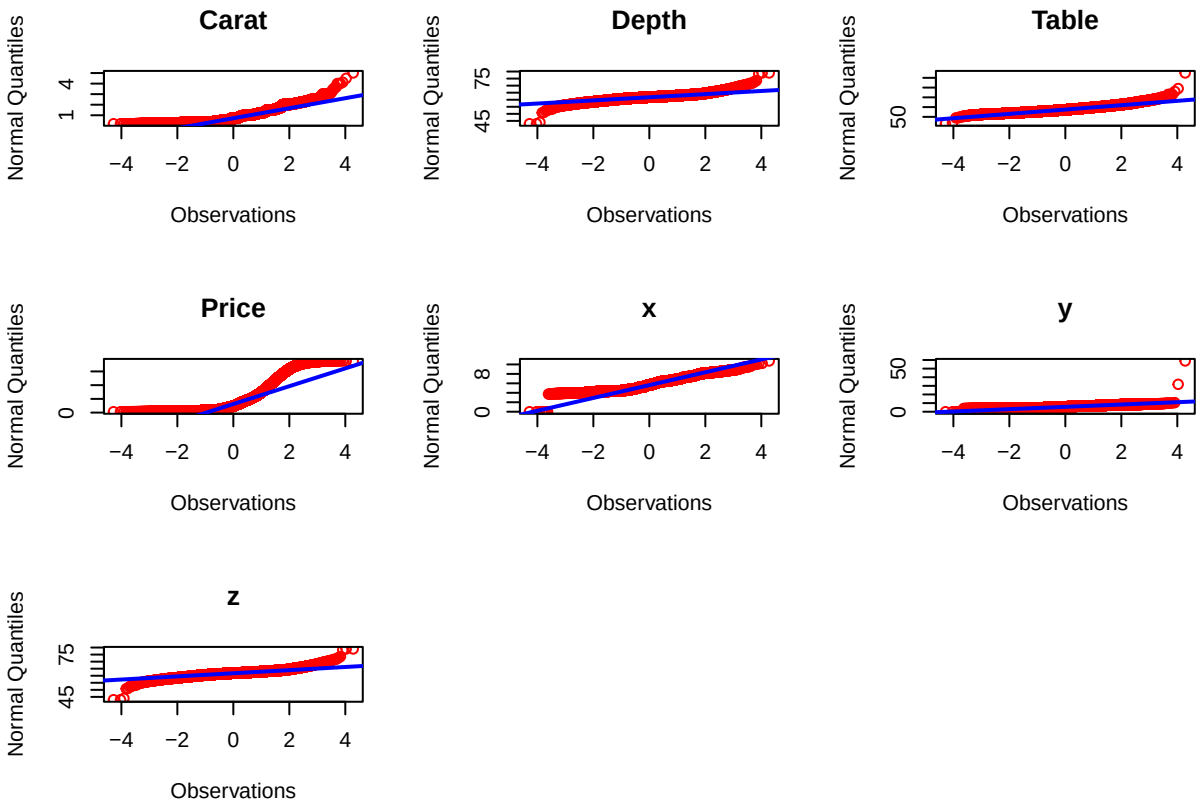


Figure 1: Normal QQ Plots of Numeric Variables

Despite the small size of the Normal QQ plots in figure 1, it is clear that each numeric variable deviates significantly from a normal distribution. We tested this using Kolmogorov-Smirnov goodness-of-fit test. The results from these tests are shown in the table below.

Table 5: Kolmogorov-Smirnov GOF results

Variable	Test_Statistic	p_value
Carat	0.122740	<2.2e-16
Depth	0.075871	<2.2e-16
Table	0.132250	<2.2e-16
Price	0.184670	<2.2e-16
x	0.093545	<2.2e-16
y	0.088528	<2.2e-16
z	0.089273	<2.2e-16

Table 5 shows that for all seven of the numeric variables, the Kolmogorov-Smirnov goodness-of-fit tests provided strong evidence for non-normality, as evidenced by the p-values which are all machine precision zero.

These tests supported our earlier observation that the data does not follow a normal distribution. We note this and proceed with our investigation anyway.

4.2.4 Correlation Plot

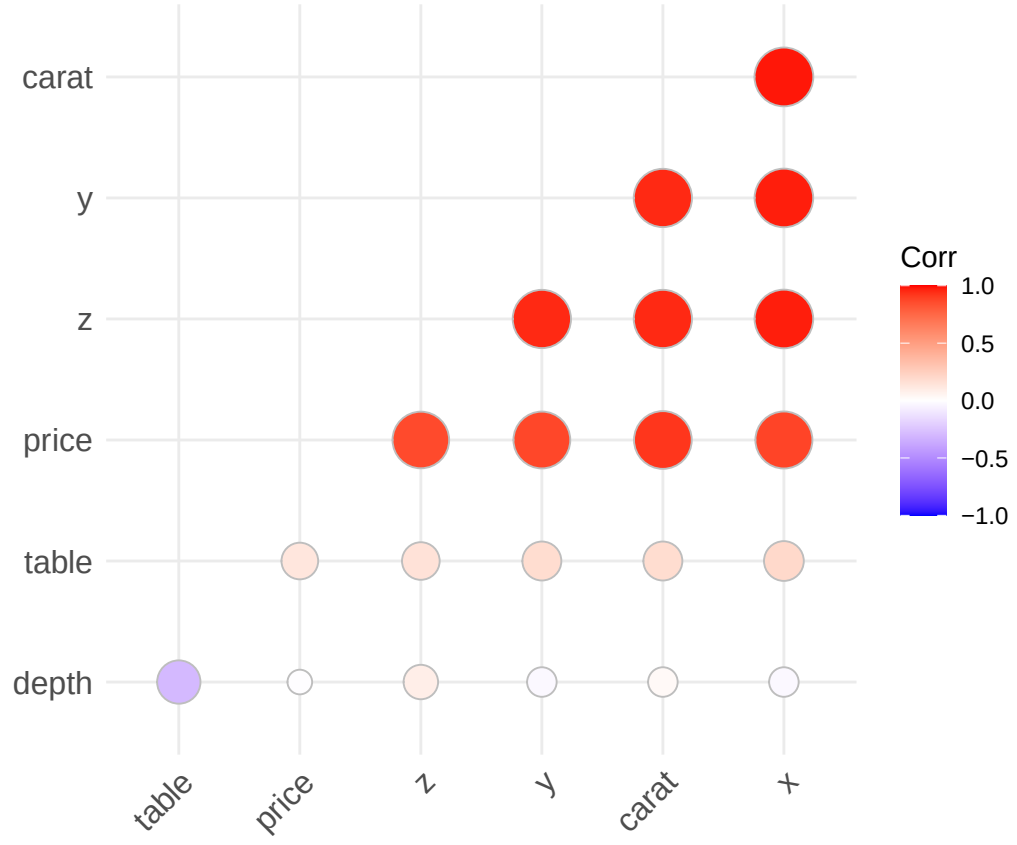


Figure 2: A Visualization of the Correlation Matrix

Figure 2 shows the correlation plot for the numerical variables in the diamonds data. 'Carat', 'x', 'y', 'z' and 'price' all show very strong correlations with each other, as evidenced by the large red dots. As "x", "y" and "z" are all measures of size we should expect this and there may be some redundancy in these predictors. The 'table' variable is relatively uncorrelated with any of the others. 'Depth' and 'table' are negatively correlated (large purple dot), while depth is not correlated with any other variable. The strongest predictor of price is carat with length, width, depth also strongly correlated with price. Table is only very weakly correlated with price while depth is negatively correlated with price.

4.2.5 Correlation Matrix

Table 6: Correlation matrix for the numerical variables

	carat	depth	table	price	x	y	z
carat	1.000	0.028	0.182	0.922	0.975	0.952	0.953
depth	0.028	1.000	-0.296	-0.011	-0.025	-0.029	0.095
table	0.182	-0.296	1.000	0.127	0.195	0.184	0.151
price	0.922	-0.011	0.127	1.000	0.884	0.865	0.861
x	0.975	-0.025	0.195	0.884	1.000	0.975	0.971
y	0.952	-0.029	0.184	0.865	0.975	1.000	0.952
z	0.953	0.095	0.151	0.861	0.971	0.952	1.000

Table 6 shows the correlation matrix. The strong positive correlations between price and 'carat' (0.922), price and 'x' (0.884), price and 'y' (0.865) and price and 'z' (0.861) supported one of our starting hypotheses, namely that these dimension variables would be positively correlated with 'price'. We noticed that they were also highly correlated with each other (all values greater than 0.9), a fact we used subsequently to combine them into one of the factors for the Factor Analysis.

4.2.6 Price differences across Categorical Variables

As our leading question is investigating which variables are most predictive or price we decided to investigate if there are significant differences in price across levels of each categorical variable.

4.2.6.1 Cut

Here we get the mean price of diamonds of different levels of the 'cut' variable.

Table 7: Mean prices for the five levels of 'cut'

	Mean Price
Fair	4359
Good	3929
Very Good	3982
Premium	4584
Ideal	3458

The mean price of diamonds differs for different levels of 'cut' (table 7). However, the higher levels of cut do not necessarily lead to higher prices.

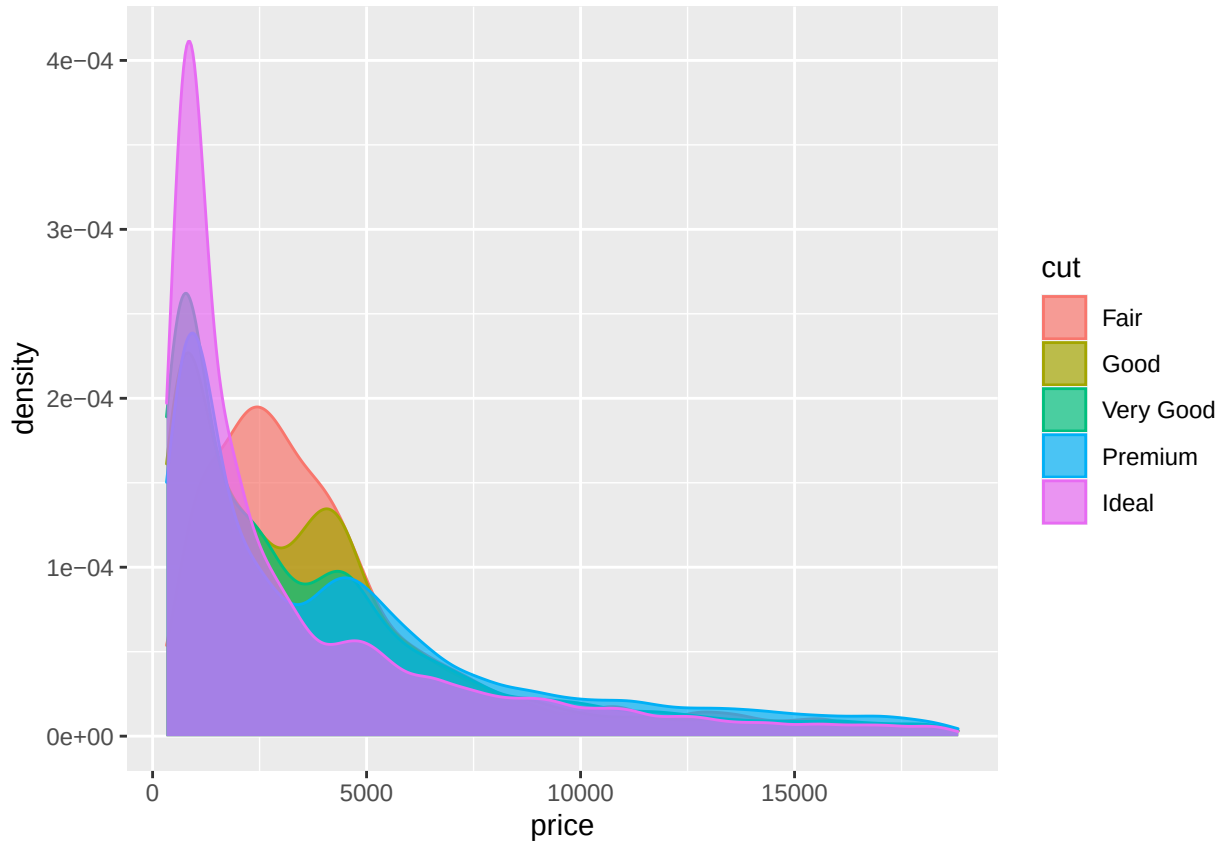


Figure 3: Densities of 'price' for different levels of 'cut'

Figure 3 shows the density plots for the different levels of the 'price' variable. The legend on the right shows the level names and their order ('fair' = poorest, 'ideal' = best). Of note is the fact that the two best leves (premium and ideal) have significant peaks near the lower end of the price range compared with the other three. This is somewhat surprising, as intuitively one would imagine price to increase as the quality of cut increases.

We conducted a Analysis of Variance (ANOVA) for all three of the categorical variables to see if the observed differences in mean price across the various levels were significant. We also conducted a Levene's test to verfiy if the assumption of homogeneity of variance was violated. If these tests showed violations of the ANOVA assumptions, we then implemented a non-parametric Kruskal Wallis Test. From the graph of the distributions of price for different levels of 'cut' we can see that not all of them have a shape consistent with being normally distributed. A one-way ANOVA is reasonably robust to departures from normality, particularly as we have a very large sample. The code for these tests may be found in the appendices.

Table 8: ANOVA of Price by Cut

Test	Test_Statistic	p_value
ANOVA	175.70	<2.2e-16
Levene's Test	123.60	<2.2e-16
Kruskal Wallis	978.62	<2.2e-16

Table 8 shows the results of the ANOVA, Levene's and Kruskal Wallis tests for the 'cut' variable. The tests are to check whether there are differences in price between the different levels of 'cut'. The ANOVA test returned a p-value of $<$ 2.2e-16 (machine precision zero), meaning that there is strong evidence to suggest that mean price differs across levels of "cut".

The Tukey Test indicated that at the 5% significance level, the only pairs between which we do not see a significant difference in mean price are 'Very Good' and 'Good' as well as 'Premium' and 'Fair'. Both the Levene's test and Kruskal Wallis Test return significant results indicating that: a) the assumption of equal variance is violated and; b) that we have evidence of a significant difference in median prices for different levels of cut.

4.2.6.2 Clarity

Here we get the mean price of diamonds of different levels of the 'clarity' variable.

Table 9: Mean prices for the 8 levels of 'clarity'

	Mean Price
I1	3924
SI2	5063
SI1	3996
VS2	3925
VS1	3839
VVS2	3284
VVS1	2523
IF	2865

Table 9 shows that the mean price of diamonds varies across different levels of 'clarity'. Interestingly the diamonds with the two highest clarities show the lowest mean price.

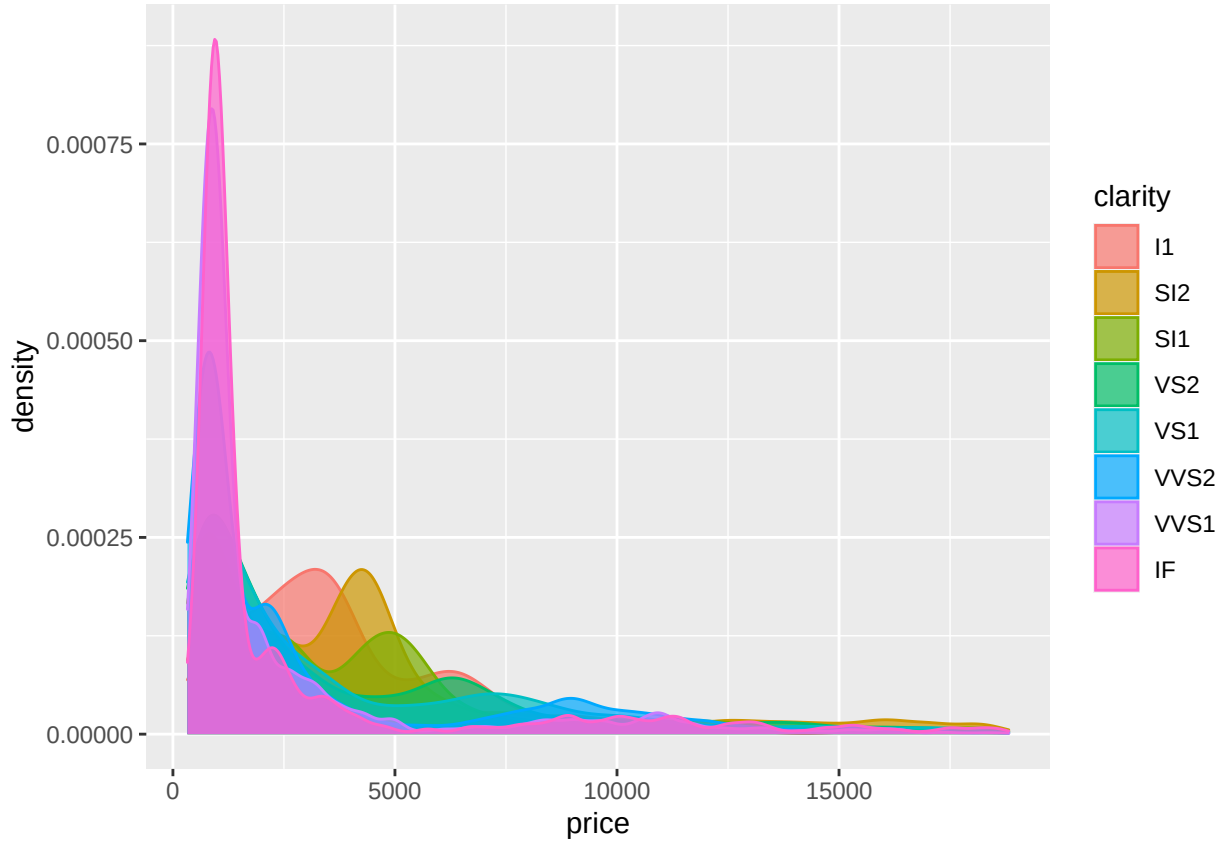


Figure 4: Densities of 'price' for different levels of 'clarity'

Figure 4 shows the densities of the different levels of 'clarity' for price. A prominent peak is visible near the left for the better quality levels (IF, best, pink; WS1 second best, purple; WS2, third best, blue).

Table 10: ANOVA of Price by Clarity

Test	Test_Statistic	p_value
ANOVA	215.000	<2.2e-16
Levene's Test	77.809	<2.2e-16
Kruskal Wallis	2718.200	<2.2e-16

Table 10 shows the table of results from the three tests on the variable 'clarity'. The output from the ANOVA returns a p-value of $< 2.2e\text{-}16$, meaning there is strong evidence to suggest that mean price differs across levels of 'clarity'. The output of the Levene's test and Kruskal Wallis tests show that there is not equal variances between the levels, and that there is a significant difference between different levels of clarity. A Tukey test showed that most pairwise combinations show a significant difference. There are only six that do not show a difference and they are: SI1-I1; VS2-I1; VS1-I1; VS2-SI1; VS1-VS2; and IF-VVS1.

4.2.6.3 Color

Here we get the mean price of diamonds of different levels of the 'color' variable.

Table 11: Mean prices for the 8 levels of 'color'

	Mean Price
J	5324
I	5092
H	4487
G	3999
F	3725
E	3077
D	3170

Table 11 shows the mean prices for the different levels of the 'color' variable and it is clear that they vary. The diamonds with the two highest color quality show the lowest mean price.

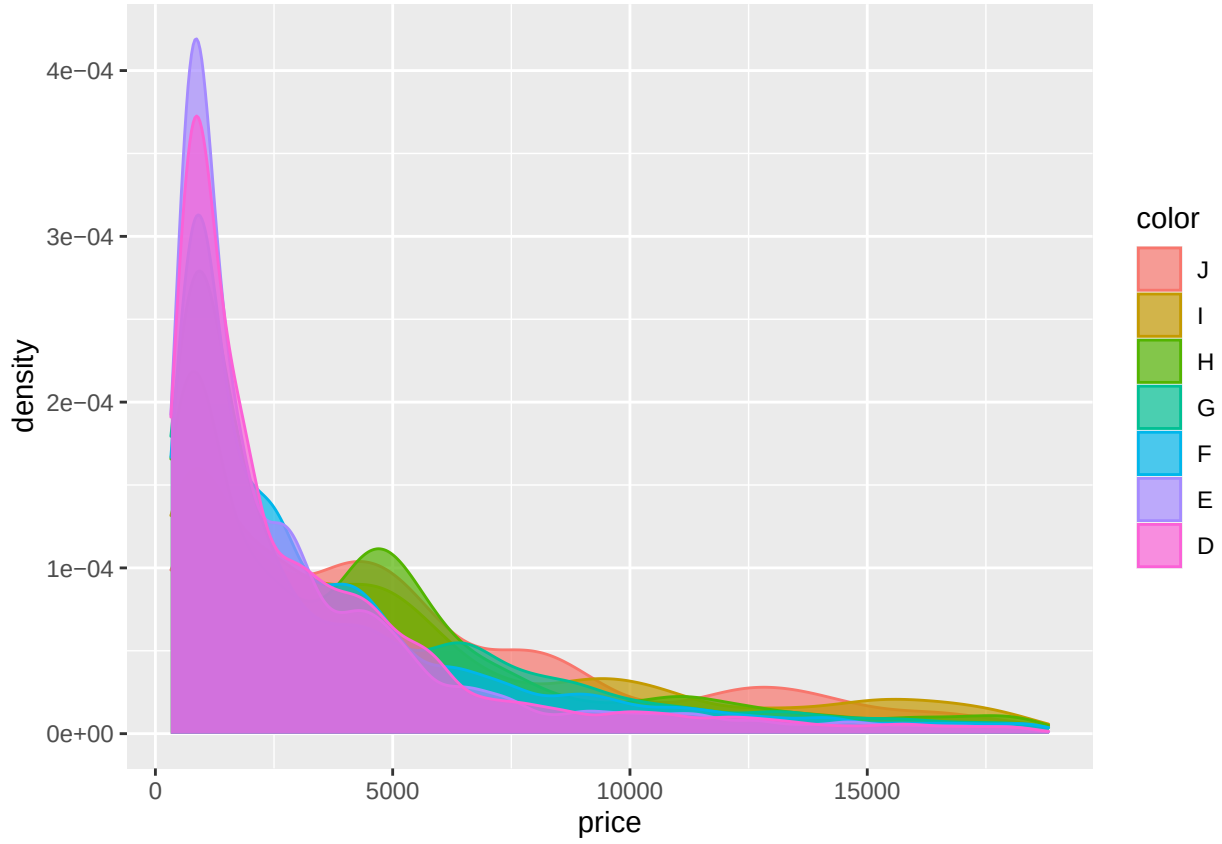


Figure 5: Densities of 'price' for the different levels of 'color'

Figure 5 shows the density plots of 'price' for the different levels of the 'color' variable. There is a prominent peak on the left for the better quality levels of color (D in pink, E in purple, and F in blue), but otherwise the densities appear to be fairly similar.

Table 12: ANOVA of Price by Color

Test	Test_Statistic	p_value
ANOVA	175.70	<2.2e-16
Levene's Test	219.12	<2.2e-16
Kruskal Wallis	1335.60	<2.2e-16

Table 12 shows the results of the tests for the 'color' variable. We have strong evidence to suggest that mean price differs across levels of 'color'. The output of the Levene's test and Kruskal Wallis test above show that the variances are not equal between the levels, and that there is a significant difference in median price across different levels of 'color'. The output from the Tukey Test showed significant differences in mean price for nearly all pairwise comparisons of diamond colors.

4.2.7 Outliers and Unusual Points

4.2.7.1 Mahalanobis Distance

We decided to investigate outliers and unusual points in our dataset. These points may have increased leverage or influence when we come to using variables to predict price.

We began by using the Mahalanobis Distance to identify surprising and very surprising points. A 'somewhat surprising' point is one that is equal or greater than the 90th percentile, a 'surprising' point is one that is equal or greater than the 95th percentile, while a 'very surprising' point is one that is equal or greater than the 99th percentile. A 'typical' point is one that lies within the 90th percentile.

Table 13: Mahalanobis distances surprising values

Mahal Dist	Count
Typical	49611
Somewhat Surprising	1001
Surprising	1489
Very Surprising	1839

We see from table 13 while the vast majority of diamonds are typical in terms of their Mahalanobis distances, there are a reasonable amount of "Surprising" and "Very Surprising" points in this dataset. We can see how many very surprising points (as identified with the

Mahalanobis Distance) are members of each level of our categorical variables. This might help to indicate if any classes contain more surprising points than others.

Table 14: Mahalanobis distances for 'cut'

Mahal Dist	Count
Fair	517
Good	216
Very Good	286
Premium	412
Ideal	408

Table 14 shows the 'very surprising' Mahalanobis distances for the different levels of the variable 'cut'.

Table 15: Mahalanobis distances for 'clarity'

Mahal Dist	Count
I1	209
SI2	570
SI1	280
VS2	251
VS1	231
VVS2	103
VVS1	105
IF	90

Table 15 shows the 'very surprising' Mahalanobis distances for the different levels of the variable 'cut'.

Table 16: Mahalanobis distances for 'color'

Mahal Dist	Count
J	213
I	284
H	338
G	328
F	290
E	216
D	170

Table 16 shows the 'very surprising' Mahalanobis distances for the different levels of the variable 'color'.

4.2.7.2 Zero Values

In this section we identify any zero values and determine whether or not they are errors. The code below shows the lowest values for each of the seven numerical variables. The number of elements to display was tweaked until the output showed all the zero values (if any). This was partially to discover whether the extreme values in the dataset were having undue influence on results.

```
sort(decreasing=F, diamonds$carat)[1:10]
```

```
##  [1] 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2
```

```
sort(decreasing=F, diamonds$x)[1:10]
```

```
##  [1] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 3.73 3.73
```

```
sort(decreasing=F, diamonds$y)[1:10]
```

```
##  [1] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 3.68 3.71 3.71
```

```
sort(decreasing=F, diamonds$z)[1:22]
```

```
##  [1] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
## [16] 0.00 0.00 0.00 0.00 0.00 1.07 1.41
```

```
sort(decreasing=F, diamonds$depth)[1:10]
```

```
##  [1] 43.0 43.0 44.0 50.8 51.0 52.2 52.3 52.7 53.0 53.1
```

```
sort(decreasing=F, diamonds$table)[1:10]
```

```
##  [1] 43.0 44.0 49.0 49.0 50.0 50.0 50.1 51.0 51.0 51.0
```

```
sort(decreasing=F, diamonds$price)[1:10]
```

```
##  [1] 326 326 327 334 335 336 336 337 337 338
```

The output above shows that x, y and z all have a number of zero values, which presumably are errors.

4.3 Idendifying whether the zero observations are errors

We firstly make a dataframe version of the dataset to allow easy subsetting. Note that this version includes the 'surprising' Mahalanobis distance values (column name 'surprise'). This is useful as it allows us to see which erroneous values have generated very large Mahalanobis distances.

```
diamonds.df <- data.frame(diamonds)
```

We now display the zero values for each of the variable, noting whether the other variables are consistent with zero readings, or whether they indicate that this is impossible. For example,

the diamond in row 1 below with zero mm length ('x'), but 6.62 mm width ('y').

```
# display the zero values for 'x'
diamonds.df[diamonds$x==0,]
```

```
##       carat       cut color clarity depth table price x    y z surprise
## 11183  1.07     Ideal     F     SI2  61.6    56  4954 0 6.62 0     Very
## 11964  1.00 Very Good     H     VS2  63.3    53  5139 0 0.00 0     Very
## 15952  1.14      Fair     G     VS1  57.5    67  6381 0 0.00 0     Very
## 24521  1.56     Ideal     G     VS2  62.2    54 12800 0 0.00 0     Very
## 26244  1.20   Premium     D    VVS1  62.1    59 15686 0 0.00 0     Very
## 27430  2.25   Premium     H     SI2  62.8    59 18034 0 0.00 0     Very
## 49557  0.71      Good     F     SI2  64.1    60  2130 0 0.00 0     Very
## 49558  0.71      Good     F     SI2  64.1    60  2130 0 0.00 0     Very
```

```
# display the zero values for 'y'
diamonds.df[diamonds$y==0,]
```

```
##       carat       cut color clarity depth table price x y z surprise
## 11964  1.00 Very Good     H     VS2  63.3    53  5139 0 0 0     Very
## 15952  1.14      Fair     G     VS1  57.5    67  6381 0 0 0     Very
## 24521  1.56     Ideal     G     VS2  62.2    54 12800 0 0 0     Very
## 26244  1.20   Premium     D    VVS1  62.1    59 15686 0 0 0     Very
## 27430  2.25   Premium     H     SI2  62.8    59 18034 0 0 0     Very
## 49557  0.71      Good     F     SI2  64.1    60  2130 0 0 0     Very
## 49558  0.71      Good     F     SI2  64.1    60  2130 0 0 0     Very
```

```
# display the zero values for 'z'
diamonds.df[diamonds$z==0,]
```

```
##       carat       cut color clarity depth table price    x    y z surprise
## 2208   1.00   Premium     G     SI2  59.1    59  3142 6.55 6.48 0     Very
## 2315   1.01   Premium     H      I1  58.1    59  3167 6.66 6.60 0     Very
## 4792   1.10   Premium     G     SI2  63.0    59  3696 6.50 6.47 0     Very
## 5472   1.01   Premium     F     SI2  59.2    58  3837 6.50 6.47 0     Very
## 10168  1.50      Good     G      I1  64.0    61  4731 7.15 7.04 0     Very
## 11183  1.07     Ideal     F     SI2  61.6    56  4954 0.00 6.62 0     Very
## 11964  1.00 Very Good     H     VS2  63.3    53  5139 0.00 0.00 0     Very
## 13602  1.15     Ideal     G     VS2  59.2    56  5564 6.88 6.83 0     Very
## 15952  1.14      Fair     G     VS1  57.5    67  6381 0.00 0.00 0     Very
## 24395  2.18   Premium     H     SI2  59.4    61 12631 8.49 8.45 0     Very
## 24521  1.56     Ideal     G     VS2  62.2    54 12800 0.00 0.00 0     Very
## 26124  2.25   Premium     I     SI1  61.3    58 15397 8.52 8.42 0     Very
## 26244  1.20   Premium     D    VVS1  62.1    59 15686 0.00 0.00 0     Very
## 27113  2.20   Premium     H     SI1  61.2    59 17265 8.42 8.37 0     Very
## 27430  2.25   Premium     H     SI2  62.8    59 18034 0.00 0.00 0     Very
## 27504  2.02   Premium     H     VS2  62.7    53 18207 8.02 7.95 0     Very
```

```
## 27740  2.80      Good     G     SI2  63.8    58 18788 8.90 8.85 0     Very
## 49557  0.71      Good     F     SI2  64.1    60  2130 0.00 0.00 0     Very
## 49558  0.71      Good     F     SI2  64.1    60  2130 0.00 0.00 0     Very
## 51507  1.12   Premium     G      I1  60.4    59  2383 6.71 6.67 0     Very
```

The outputs above show that the zero values for 'x', 'y' and 'z' are all errors. We can tell this because these diamonds have carat, depth and price values, meaning that they cannot have zero length (x), width (y) and depth (z). Note that for a lot of the observations with zero values in one of these variables also have zero values in the others. The variable 'y' is a good example; the output shows that every observation with a zero value for 'y' also has zero values for 'x' and 'z'. Note also that all of the zero values have 'very surprising' Mahalanobis distance results. This suggests that they are exerting undue influence or leverage on our results. However, this might be counterbalanced to a degree by the very large number of observations in the dataset.

4.3.1 Upper-value outliers

The outputs below show the upper values for the seven numerical variables. This output guides further investigation as to which are likely to be genuine and which are probably errors.

```
# show ten largest values for each of the seven numerical variables
sort(decreasing=T, diamonds$carat)[1:10]
```

```
##  [1] 5.01 4.50 4.13 4.01 4.01 4.00 3.67 3.65 3.51 3.50
```

```
sort(decreasing=T, diamonds$x)[1:10]
```

```
##  [1] 10.74 10.23 10.14 10.02 10.01 10.00  9.86  9.66  9.65  9.54
```

```
sort(decreasing=T, diamonds$y)[1:10]
```

```
##  [1] 58.90 31.80 10.54 10.16 10.10  9.94  9.94  9.85  9.81  9.63
```

```
sort(decreasing=T, diamonds$z)[1:10]
```

```
##  [1] 31.80  8.06  6.98  6.72  6.43  6.38  6.31  6.27  6.24  6.17
```

```
sort(decreasing=T, diamonds$depth)[1:10]
```

```
##  [1] 79.0 79.0 78.2 73.6 72.9 72.2 71.8 71.6 71.6 71.3
```

```
sort(decreasing=T, diamonds$table)[1:10]
```

```
##  [1] 95 79 76 73 73 73 73 71 70 70
```

```
sort(decreasing=T, diamonds$price)[1:10]
```

```
##  [1] 18823 18818 18806 18804 18803 18797 18795 18795 18791 18791
```

The output above shows the ten largest values for each of the seven numerical variables. At

a glance (informal inference) it appears that 'carat', 'y', 'z' and 'table' all have one or more values that are considerably higher than the rest, with y and z having maximums that are so extreme it is worth considering whether they are erroneous.

4.4 Determining whether the outliers are errors

In this section we go through the variables one at a time to determine whether the upper values are likely to be erroneous or genuine.

4.4.1 The 'x' variable: probably no upper value errors

```
# display the values of 'x' that are greater than or equal to 10
diamonds.df[diamonds$x>=10,]
```

```
##       carat       cut color clarity depth table price     x     y    z surprise
## 25999  4.01   Premium     I      I1  61.0    61 15223 10.14 10.10 6.17     Very
## 26000  4.01   Premium     J      I1  62.5    62 15223 10.02  9.94 6.24     Very
## 26445  4.00 Very Good     I      I1  63.3    58 15984 10.01  9.94 6.31     Very
## 27131  4.13      Fair     H      I1  64.8    61 17329 10.00  9.85 6.43     Very
## 27416  5.01      Fair     J      I1  65.5    59 18018 10.74 10.54 6.98     Very
## 27631  4.50      Fair     J      I1  65.8    58 18531 10.23 10.16 6.72     Very
```

The output above shows that the largest 'x' values are not outrageously larger than the rest. Additionally, the values of the other variables for the largest 'x' value are reasonably aligned in terms of magnitude, suggesting that these large 'x' values are genuine.

4.4.2 The 'y' variable: probably 2 upper value errors

```
# display the values of 'x' that are greater than or equal to 11
diamonds.df[diamonds$y>=10,]
```

```
##       carat     cut color clarity depth table price     x     y    z surprise
## 24068  2.00 Premium     H     SI2  58.9    57 12210  8.09 58.90 8.06     Very
## 25999  4.01 Premium     I      I1  61.0    61 15223 10.14 10.10 6.17     Very
## 27416  5.01    Fair     J      I1  65.5    59 18018 10.74 10.54 6.98     Very
## 27631  4.50    Fair     J      I1  65.8    58 18531 10.23 10.16 6.72     Very
## 49190  0.51   Ideal     E     VS1  61.8    55  2075  5.15 31.80 5.12     Very
```

The output above suggests that the two extreme values for 'y' are almost certainly errors. They are 58.90 and 31.80; both are far larger than the next highest value, which is 10.54. It is very unlikely that diamonds with such enormous width values do not also have extreme length and depth values, or that they did not sell for more money.

4.4.3 The 'z' variable: probably 2 upper value errors

```
diamonds.df[diamonds$z>=8,]
```

```
##       carat       cut color clarity depth table price    x     y     z surprise
## 24068  2.00   Premium     H     SI2  58.9  57.0 12210 8.09 58.90  8.06     Very
## 48411  0.51 Very Good     E     VS1  61.8  54.7  1970 5.12  5.15 31.80     Very
```

Similarly with the extreme value of 'z' (31.80). The other dimensions of this diamond do not tally with the extreme depth value, nor does the relatively low price. The next highest value has been included for comparison (8.06). This is almost certainly an error.

4.4.4 The 'carat' variable: probably no upper value errors

```
diamonds.df[diamonds.df$carat>4,]
```

```
##       carat     cut color clarity depth table price     x     y    z surprise
## 25999  4.01 Premium     I      I1  61.0    61 15223 10.14 10.10 6.17     Very
## 26000  4.01 Premium     J      I1  62.5    62 15223 10.02  9.94 6.24     Very
## 27131  4.13    Fair     H      I1  64.8    61 17329 10.00  9.85 6.43     Very
## 27416  5.01    Fair     J      I1  65.5    59 18018 10.74 10.54 6.98     Very
## 27631  4.50    Fair     J      I1  65.8    58 18531 10.23 10.16 6.72     Very
```

The output above shows that the highest value for 'carat' (5.01) is not outrageously higher than the next highest (4.50), suggesting that this is a genuine value. Also, the other variables are comparible between the highest and second highest, so it is likely that this value can be trusted.

4.4.5 The 'table' variable: probably no upper value errors

```
# display the values of 'table' that are greater than or equal to 73
diamonds.df[diamonds.df$table>=73,]
```

```
##       carat  cut color clarity depth table price    x    y    z surprise
## 24933  2.01 Fair     F     SI1  58.6    95 13387 8.32 8.31 4.87     Very
## 49376  0.70 Fair     H     VS1  62.0    73  2100 5.65 5.54 3.47     Very
## 50774  0.81 Fair     F     SI2  68.8    79  2301 5.26 5.20 3.58     Very
## 51343  0.79 Fair     G     SI1  65.3    76  2362 5.52 5.13 3.35     Very
## 51392  0.71 Fair     D     VS2  55.6    73  2368 6.01 5.96 3.33     Very
## 52861  0.50 Fair     E     VS2  79.0    73  2579 5.21 5.18 4.09     Very
## 52862  0.50 Fair     E     VS2  79.0    73  2579 5.21 5.18 4.09     Very
```

The output above for the highest values of 'table' show that the diamond with the largest value (95) was also considerably larger in other ways and was much more expensive. This suggests that this is a genuine value, rather than an error.

Note that all the upper outliers above, whether genuine or erroneous, have generated 'very

surprising' values for their Mahalanobis distances.

4.4.6 'depth' and 'price': probably no upper value errors

```
sort(diamonds.df$depth, decreasing = TRUE)[1:10]
```

```
##  [1] 79.0 79.0 78.2 73.6 72.9 72.2 71.8 71.6 71.6 71.3
```

```
sort(diamonds.df$price, decreasing = TRUE)[1:10]
```

```
##  [1] 18823 18818 18806 18804 18803 18797 18795 18795 18791 18791
```

The output above shows that neither 'depth' nor 'price' have any extreme values. We can probably safely conclude that these are all genuine upper values.

4.5 Colour-coded scatterplots

In these scatterplots we see the three categorical variables of 'cut', 'color' and 'clarity' all versus 'price', but further colour-coded so we can see the price of diamonds at the different levels of these variables. This exercise is useful to see why our subsequent linear regression model using the single predictor of 'carat' did not work as well as hoped. This is expanded upon in the commentary below under the heading "Why prediction without the categorical variables is not working well".

4.5.1 Scatterplot of 'Carat' vs 'Price', colour-coded by 'Cut'

```
# scatterplot of carat vs price, but also colouring the observations
# by 'cut'
ggplot(data=diamonds, aes(x=carat, y=price, color=cut))+
  geom_point()+
  guides(colour=guide_legend(reverse = T))
```

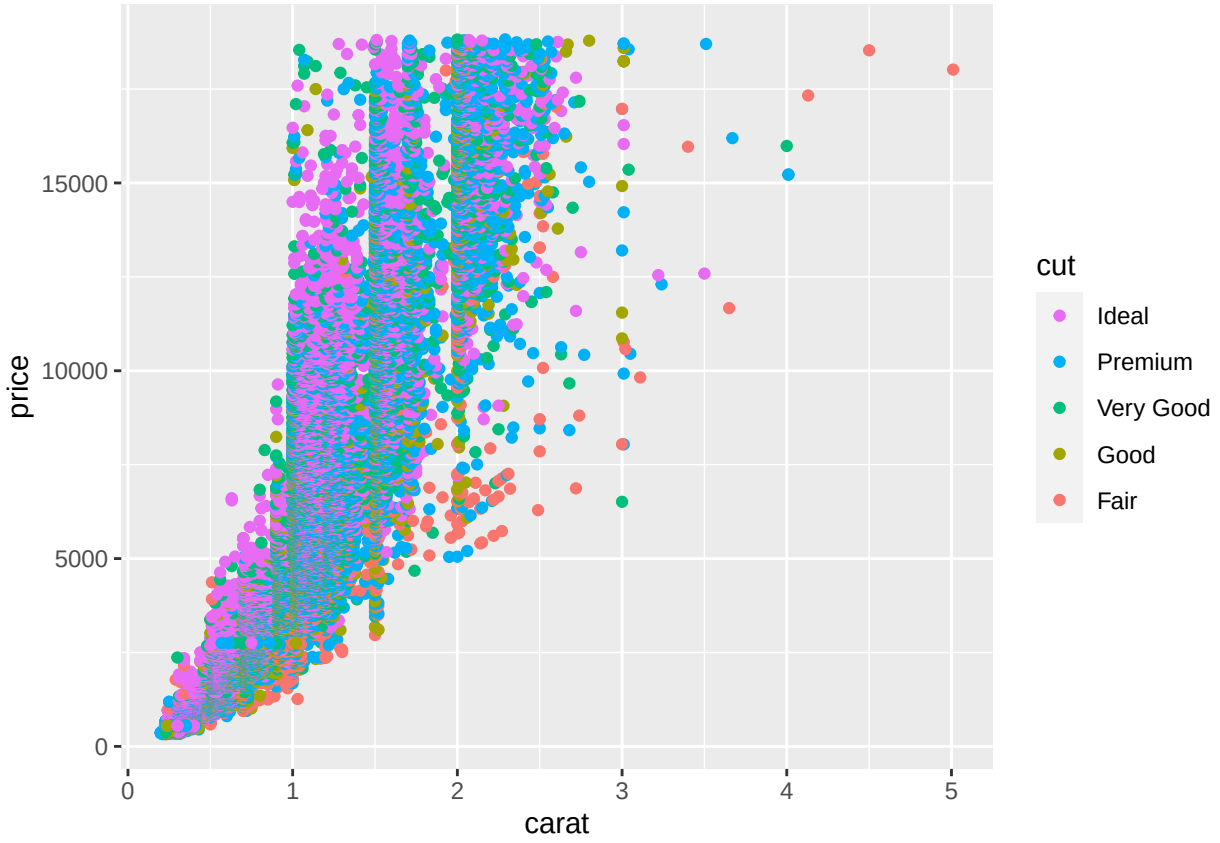


Figure 6: Carat vs Price, coloured by Cut

Figure 6 shows the scatterplot of 'carat' versus 'price' and coloured by 'cut'. Notice the amount of overlap between diamonds from different levels of 'cut'. There is no clear separation, meaning that a discriminant analysis or cluster analysis will struggle to distinguish between the levels.

4.5.2 Scatterplot of 'Carat' vs 'Price', colour-coded by 'Clarity'

```
# scatterplot of carat vs price, but also colouring the observations
# by 'cut'
ggplot(data=diamonds, aes(x=carat, y=price, color=clarity))+
  geom_point()+
  guides(colour=guide_legend(reverse = T))
```

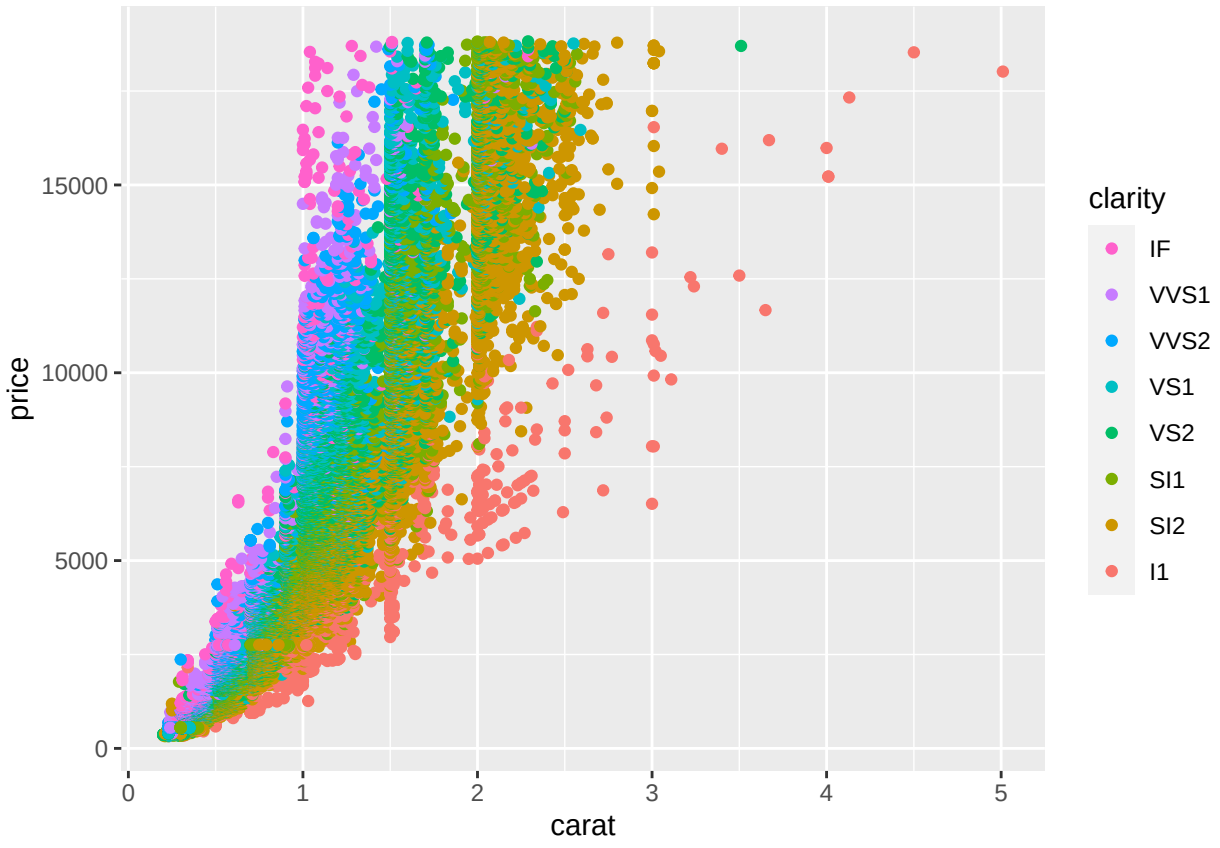


Figure 7: Carat vs Price coloured by Clarity

Figure 7 shows the scatterplot of 'carat' versus 'price' and coloured by 'clarity'. In terms of separation of the different levels, this appears more promising. Note the clear colour bands running from the bottom left to the upper right. While not perfectly separated, we can see a clear orange band at the bottom, then a clearish yellow/brown band above that, and then a clear green band etc. However, these bands still run diagonally through a large range of 'price' values (for example, the band of crimson dots at the very left runs upward from a few hundred dollars all the way up past the 17,000 dollar mark), meaning that predicting price based solely on these levels will be problematic.

4.5.3 Scatterplot of 'Carat' vs 'Price', colour coded by 'Color'

```
# scatterplot of carat vs price, but also colouring the observations
# by 'cut'
ggplot(data=diamonds, aes(x=carat, y=price, color=color))+
  geom_point()+
  guides(colour=guide_legend(reverse = T))
```

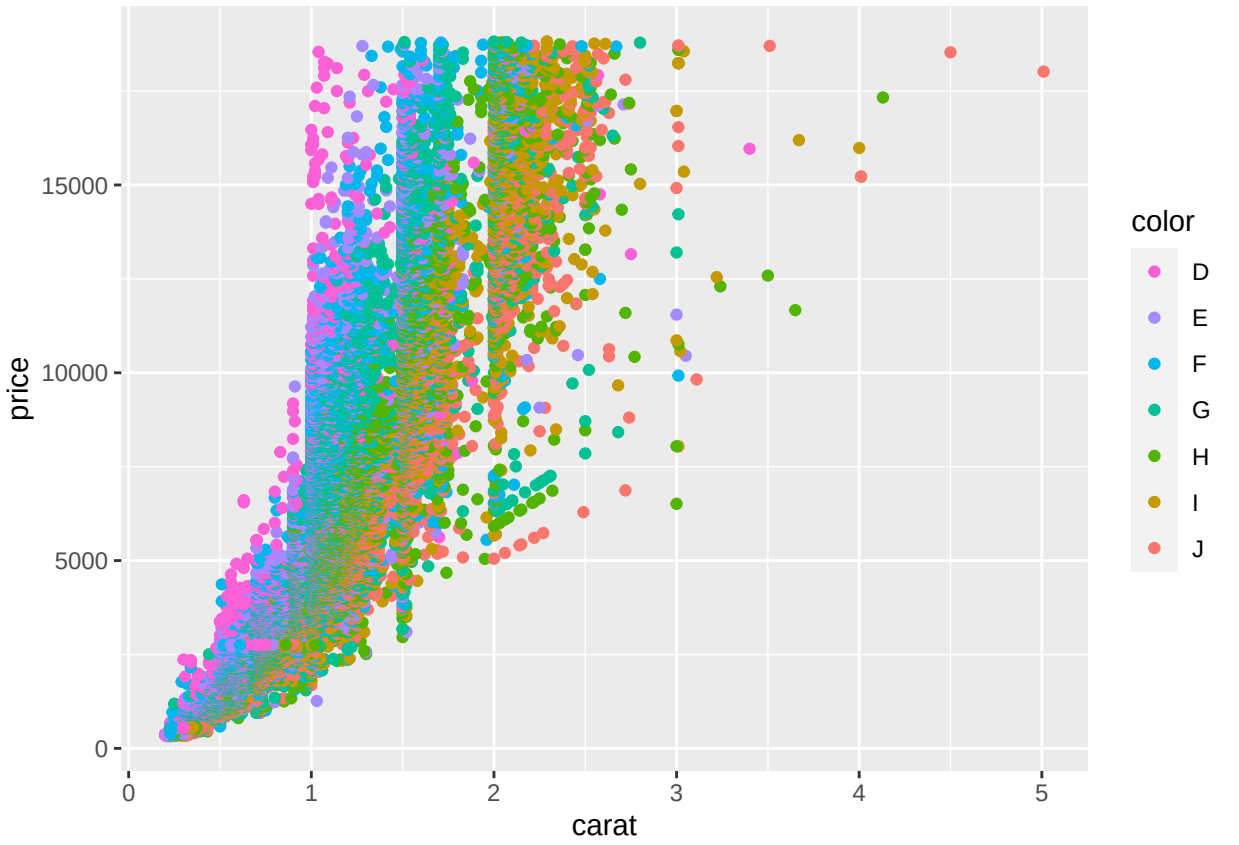


Figure 8: Carat vs Price coloured by Color

Figure 8 shows the scatterplot of 'carat' versus 'price' and coloured by 'color'.

In the three scatterplots figures 6, 7 and 8 we see a clear trend of lighter diamonds (lower 'carat' values) that are most expensive (so, in the upper left of the plot) having the highest quality of 'cut', 'clarity' and 'color'. This is no surprise, as we would expect the most expensive lighter diamonds to be of the best quality.

Conversely, we can also see some heavier diamonds which are of lower quality and lower price. Note in figure 6 that some of the lowest quality diamonds in terms of 'cut' (orange dots in the centre of the plot, on the vertical value of 2) are relatively cheap, despite being heavier than some of the lighter more expensive diamonds. But eventually the weight of the diamonds drives the price up regardless of the quality of 'cut', as we can see from the three orange dots in the top right of the plot. These are of the lowest quality cut, but are the three heaviest diamonds in the dataset, and so end up being expensive.

4.5.4 Interactions

As we have 3 categorical variables: color, clarity and cut with 7, 8 and 5 levels respectively we have 280 separate interactions in our data set. This makes it likely that there will be unknown classes in our data.

5 Simple and Multiple Regression

As we are interested in predicting price we thought it would be appropriate to do an initial multiple regression including all variables to see which are thought to be significant in predicting price when all other variables are included in the model. We have done this firstly with the raw variables and then with scaled versions of the variables.

5.1 Simple linear regression using 'carat'

We theorised that 'carat' would be a very good predictor of 'price' based on the fact that 'carat' was the variable most highly correlated with 'price', with a correlation of 0.9216 between the two variables. Here we fit a simple linear regression model using 'carat' as the sole predictor variable.

```
# fit the linear regression model
carat.lm <- lm(price~carat, data = diamonds, x=T)

# display summary
summary(carat.lm)
```

```
##
## Call:
## lm(formula = price ~ carat, data = diamonds, x = T)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18585.3   -804.8    -18.9    537.4  12731.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2256.36      13.06  -172.8   <2e-16 ***
## carat        7756.43      14.07   551.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1549 on 53938 degrees of freedom
## Multiple R-squared:  0.8493, Adjusted R-squared:  0.8493
## F-statistic: 3.041e+05 on 1 and 53938 DF,  p-value: < 2.2e-16
```

```
# get anova table of linear regression model
anova(carat.lm)
```

```
## Analysis of Variance Table
##
## Response: price
##               Df     Sum Sq    Mean Sq F value    Pr(>F)
## carat          1 7.2913e+11 7.2913e+11  304051 < 2.2e-16 ***
```

```
## Residuals 53938 1.2935e+11 2.3980e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output from the linear regression model above shows that 'carat' is a good predictor of 'price'. The t-test statistic is 3.0405091×10^5, which is an enormous number, and the p-value is 0. So, with a p-value of zero, we have very strong evidence to say that 'carat' is a good predictor of 'price'.

5.2 Testing to find the best regression model

5.2.1 The Akaike Information Criterion (AIC) test

We now use the Akaike Information Criterion (AIC) stepwise regression test in R to find the best regression model that has 'price' as the response variable. We are particularly interested in whether any other models outperform the simple 'carat' model. The output for these tests is pages long, so for brevity the `results='hide'` command has been used in the code chunk to allow viewing of the code while not displaying the output.

```
# create version with the 'surprising' variable (Mahalanobis distance) removed
diamonds_reg <- diamonds[,-11]
```

```
# show code, but don't display results as they run to pages long
# fit the linear regression model with all predictors
full.lm <- lm(price ~ .,data = diamonds_reg)
```

```
# show code, but don't display results as they run to pages long
# display the model summary
summary(full.lm)

# use the step() function to perform the stepwise AIC test
step(full.lm, direction = "both")
```

Based on the output from the stepwise AIC regression comparison above, the best model which balances accuracy against parsimony is the one with predictors carat + cut + color + clarity + depth + table + x + z. In other words, the only predictor that was dropped is 'y'. The AIC value for this model is 758425, which is the same for the model that additionally drops 'z'. Therefore, following the principle of parsimony, we select the simplest model, which means selecting the one that drops the variables 'y' and 'z'. This model is: price = carat + cut + color + clarity + depth + table + x.

5.2.2 Bayes Information Criterion (BIC) model comparison

As a check, we also perform a model comparison using the BIC criteria.

```
# show code, but don't display results as they run to pages long
# store the sample size in a variable
n <- length(diamonds$price)
```

```
# repeat the step test with the added 'k' argument to perform the BIC
step(full.lm, direction = "both", k=log(n))
```

The model comparison using the Bayes Information Criterion (BIC) confirms that the best model is the one with variables carat + cut + color + clarity + depth + table + x. The BIC value for this model is 758621, compared with 758629 for the next best (which includes 'z').

5.3 Fitting the best model

The 'best' model (as found by the AIC test above) is fitted below.

```
best.lm <- lm(price ~ carat + cut + color + clarity +
                depth + table + x, data = diamonds)
summary(best.lm)
```

```
##
## Call:
## lm(formula = price ~ carat + cut + color + clarity + depth +
##     table + x, data = diamonds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21385.0   -592.4   -183.7    376.5  10694.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3.418    391.056  -0.009    0.993
## carat        11256.968     48.600 231.626   <2e-16 ***
## cutGood        580.240     33.572  17.283   <2e-16 ***
## cutVery Good   726.820     32.212  22.564   <2e-16 ***
## cutPremium     762.759     32.225  23.670   <2e-16 ***
## cutIdeal       833.260     33.396  24.951   <2e-16 ***
## colorI         903.322     26.337  34.299   <2e-16 ***
## colorH        1389.382     24.890  55.820   <2e-16 ***
## colorG        1887.561     24.313  77.637   <2e-16 ***
## colorF        2096.670     24.813  84.497   <2e-16 ***
## colorE        2160.267     24.922  86.682   <2e-16 ***
## colorD        2369.504     26.131  90.678   <2e-16 ***
## claritySI2    2702.077     43.812  61.674   <2e-16 ***
## claritySI1    3664.905     43.627  84.005   <2e-16 ***
## clarityVS2    4266.612     43.847  97.308   <2e-16 ***
## clarityVS1    4577.589     44.535 102.786   <2e-16 ***
## clarityVVS2   4950.168     45.847 107.972   <2e-16 ***
## clarityVVS1   5007.061     47.152 106.190   <2e-16 ***
## clarityIF     5344.338     51.015 104.761   <2e-16 ***
```

```
## depth          -66.769      4.091 -16.322   <2e-16 ***
## table          -26.457      2.911  -9.089   <2e-16 ***
## x            -1029.478     20.549 -50.098   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1130 on 53918 degrees of freedom
## Multiple R-squared:  0.9198, Adjusted R-squared:  0.9198
## F-statistic: 2.944e+04 on 21 and 53918 DF,  p-value: < 2.2e-16
```

The summary output for the 'best' model above shows how the categorical variables, all of which measure diamond quality, add a lot to the accuracy of the model. For example, 'Clarity', which has 8 levels, adds from 2702.077 (lowest quality clarity) to 5344.338 (highest quality clarity) to price depending on which level the diamond is categorised at. This is a wide range and cannot be obtained with these categorical variables dropped from the model.

5.4 Check Regression Assumptions

Here we check the regression assumptions by examining the diagnostic plots.

```
par(mfrow=c(2,2))
plot(best.lm, 1)
plot(best.lm, 2)
plot(best.lm, 3)
plot(best.lm, 4)
```

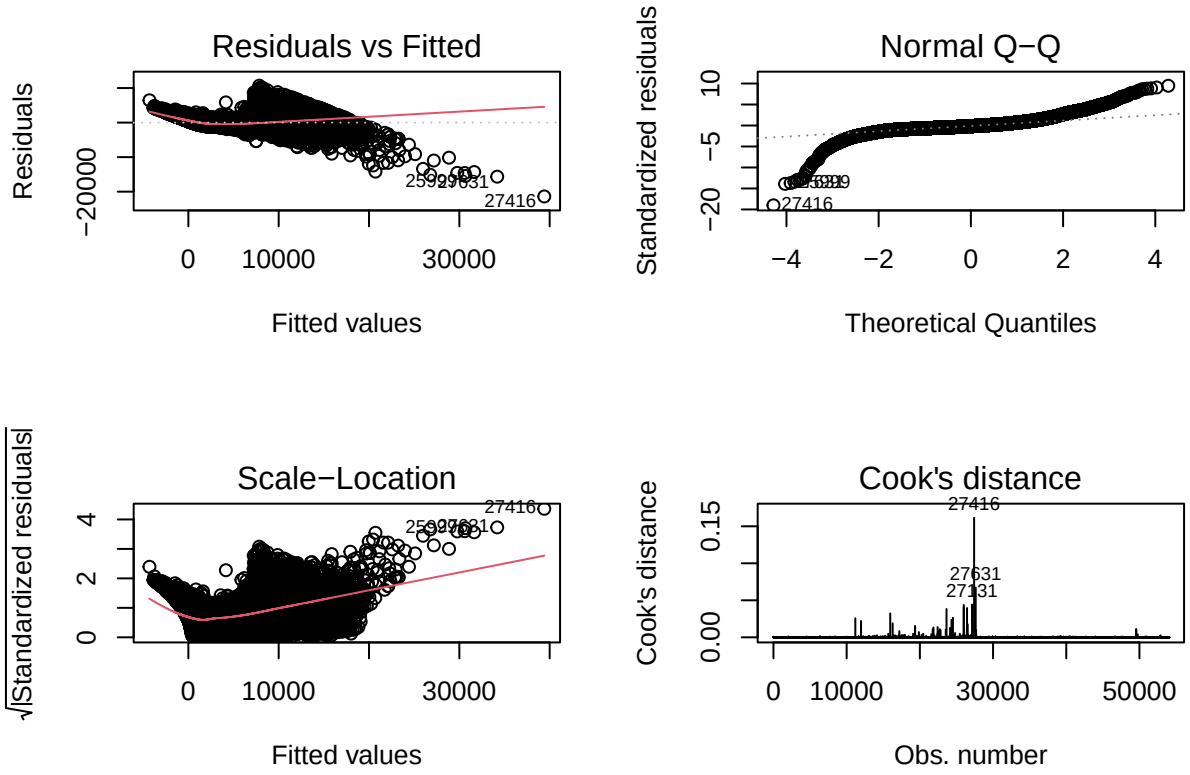


Figure 9: Diagnostic plots to check regression assumptions

By examining the residual plots (figure 9) we see that it is likely that the assumptions underlying a multiple linear regression such as linearity, homogeneity of the variance and normality of the residuals are violated. There also appears to be a number of points with high leverage and influence. We should therefore be cautious in concluding much from this regression model.

5.5 Scaled multiple regression model

In this version we scale the numerical variables and include the three categorical variables in the model.

```
diamondslm <- lm(price ~ scale(carat) + scale(x) + scale(y) +
                   scale(z) + scale(depth) + scale(table) +
                   cut + clarity + color, data = diamonds)
summary(diamondslm)
```

```
##
## Call:
## lm(formula = price ~ scale(carat) + scale(x) + scale(y) + scale(z) +
##     scale(depth) + scale(table) + cut + clarity + color, data = diamonds)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21376.0   -592.4   -183.5    376.4  10694.2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2564.403     53.433 -47.993   <2e-16 ***
## scale(carat)  5335.934     23.050 231.494   <2e-16 ***
## scale(x)     -1131.028     36.903 -30.648   <2e-16 ***
## scale(y)        10.975     22.081   0.497    0.619
## scale(z)       -35.369     23.631  -1.497    0.134
## scale(depth)   -91.410      6.496 -14.071   <2e-16 ***
## scale(table)   -59.156      6.506  -9.092   <2e-16 ***
## cutGood        579.751     33.592  17.259   <2e-16 ***
## cutVery Good   726.783     32.241  22.542   <2e-16 ***
## cutPremium     762.144     32.228  23.649   <2e-16 ***
## cutIdeal       832.912     33.407  24.932   <2e-16 ***
## claritySI2    2702.586     43.818  61.677   <2e-16 ***
## claritySI1    3665.472     43.634  84.005   <2e-16 ***
## clarityVS2    4267.224     43.853  97.306   <2e-16 ***
## clarityVS1    4578.398     44.546 102.779   <2e-16 ***
## clarityVVS2   4950.814     45.855 107.967   <2e-16 ***
## clarityVVS1   5007.759     47.160 106.187   <2e-16 ***
## clarityIF     5345.102     51.024 104.757   <2e-16 ***
## colorI         903.154     26.337  34.292   <2e-16 ***
## colorH        1389.131     24.891  55.809   <2e-16 ***
## colorG        1887.359     24.313  77.628   <2e-16 ***
## colorF        2096.544     24.813  84.492   <2e-16 ***
## colorE        2160.280     24.922  86.683   <2e-16 ***
## colorD        2369.398     26.131  90.674   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1130 on 53916 degrees of freedom
## Multiple R-squared:  0.9198, Adjusted R-squared:  0.9198
## F-statistic: 2.688e+04 on 23 and 53916 DF,  p-value: < 2.2e-16
```

This multiple regression indicates that all variables but y and z are significant in the model (once the effect of the other predictors has been accounted for). This model accounts of 91.98% of the variation in the data.

6 Principal Component Analysis

We will perform a PCA with the aim of being able to explain the variation in the data set with fewer dimensions. We also hope to then create a parsimonious model for predicting diamond price through principal components regression.

```
pca.diamonds <- prcomp(subset(diamonds.df, select = c(1,5:10)),
                       center=TRUE, scale. = TRUE, retx=TRUE)
```

```
pca.var <- pca.diamonds$sdev^2
pca.var.per <- round(pca.var/sum(pca.var)*100,1)
```

```
barplot(pca.var.per, names.arg = pca.var.per,
        xlab = "Principal Components 1 to 7",
        ylab = "% information retained by each PC",
        main="Screeplot of the Principal Components")
```

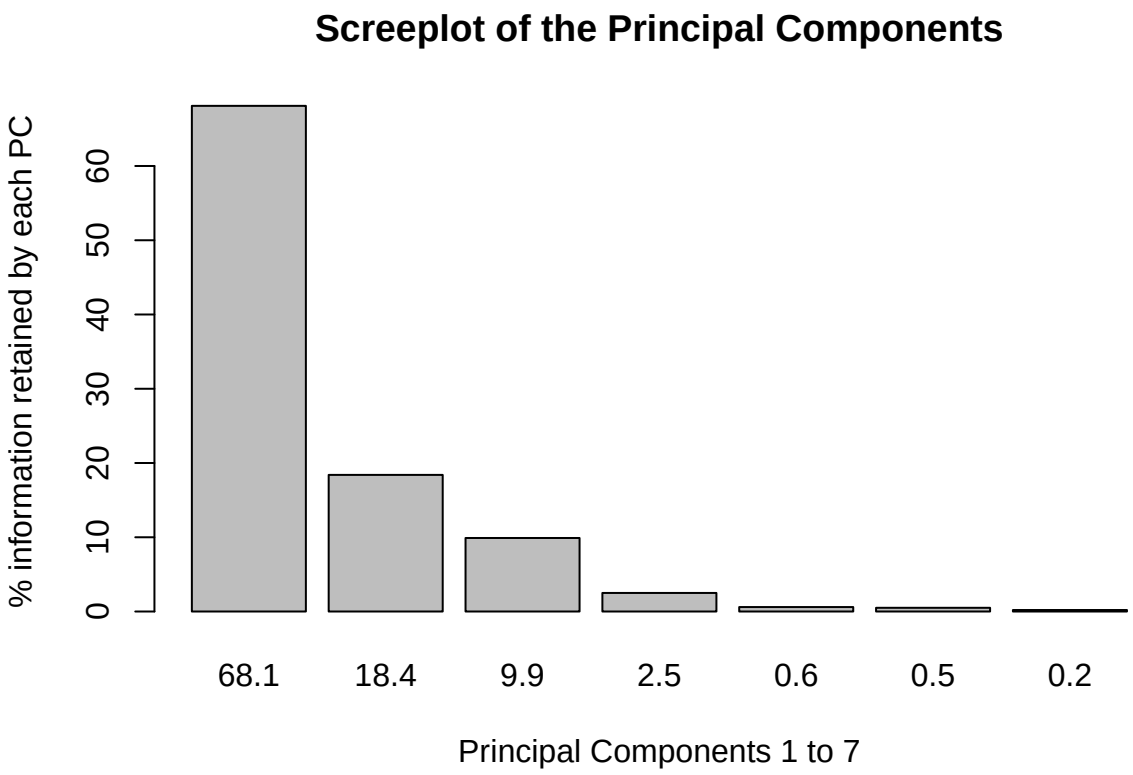


Figure 10: Scree plot of the Principal Components

Figure 10 shows the scree plot for the Principal Components of the diamonds dataset. PC1 contains 68.1% of the variance, PC2 contains 18.4%, while PC3 contains 9.9%. Between them they contain 96.4%, which is clearly very good. Components beyond the third principal

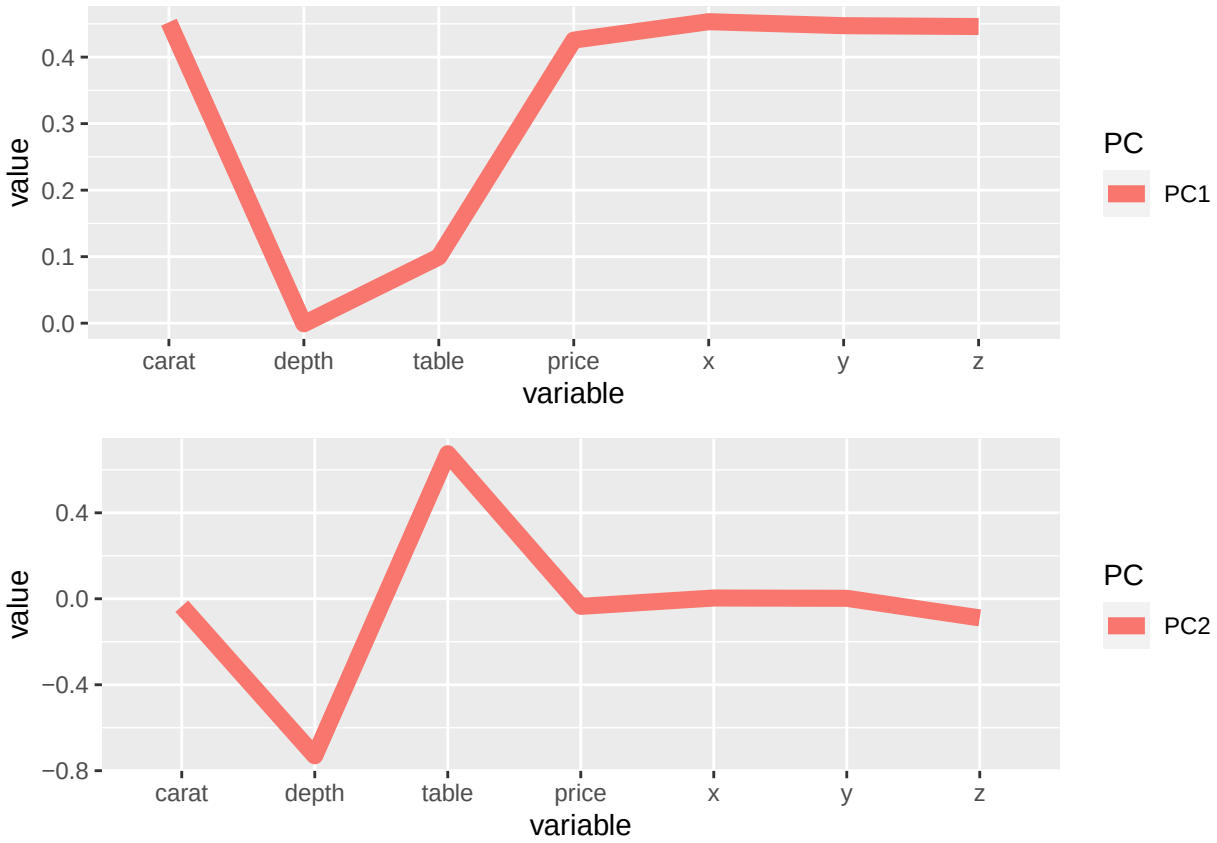
component explain minimal variance (less than 4%). This suggests we can describe the data using only the first two principal components without losing much information. We will now examine which variables contribute most strongly to each Principal Component with a particular focus on the first two.

```
pca.diamonds$rotation
```

```
##                  PC1          PC2          PC3         PC4         PC5
## carat  0.4524454941 -0.034696011  0.005494814 -0.06835945  0.13399948
## depth -0.0009161301 -0.730679714 -0.672829294 -0.04724800 -0.08873829
## table  0.0995160875  0.675067376 -0.728069469 -0.05954060 -0.01037614
## price  0.4255192667 -0.035257945  0.105449477 -0.84977817 -0.05377206
## x      0.4532125054  0.003512550  0.039508824  0.24299509  0.08898016
## y      0.4472649035  0.002157912  0.054188788  0.32846061 -0.77405793
## z      0.4459536619 -0.089035176 -0.039603439  0.31700727  0.60339656
##                PC6          PC7
## carat  0.76815114  0.425880295
## depth  0.01445027 -0.055600264
## table -0.02526831 -0.002049255
## price -0.27330947 -0.082814286
## x      0.19846061 -0.828658219
## y     -0.21526655  0.208857094
## z     -0.49867040  0.279957944
```

Looking at the eigenvectors we see that PC1 is primarily defined by carat, price, x, y and z. PC1 seems to be defined by dimension variables and, as we saw in the correlation matrix, price is closely associated with these variables. The variables with the most weight in PC2 are depth and table. We see a similar pattern in PC3 with depth and table having the largest weighting. This makes sense as in our examination of correlations between variables we saw that price, x, y, z and carat were all strongly correlated. The opposite loadings of depth and table in PC2 and PC3 reflect there negative correlation. A visualization of the eigenvectors of the first two principal components is shown in figure **??**

```
Rotation.df <- data.frame(t(pca.diamonds$rotation))
Rotation.df$PC <- c("PC1", "PC2", "PC3", "PC4", "PC5", "PC6", "PC7")
Rotation.melt <- melt(Rotation.df, id.vars = "PC")
PC1eigen <- ggplot(Rotation.melt[c(1,8, 15,22,29,36,43),]) +
  geom_line(aes(x = variable, y = value, group = PC, col = PC), size = 3)
PC2eigen <- ggplot(Rotation.melt[c(2,9, 16,23,30,37,44),]) +
  geom_line(aes(x = variable, y = value, group = PC, col = PC), size = 3)
ggarrange(PC1eigen, PC2eigen, ncol =1, nrow =2)
```



A biplot of the data projected over the first two principal components is shown below.

```
ggbiplot(pca.diamonds, obs.scale =1, var.scale = 1)
```

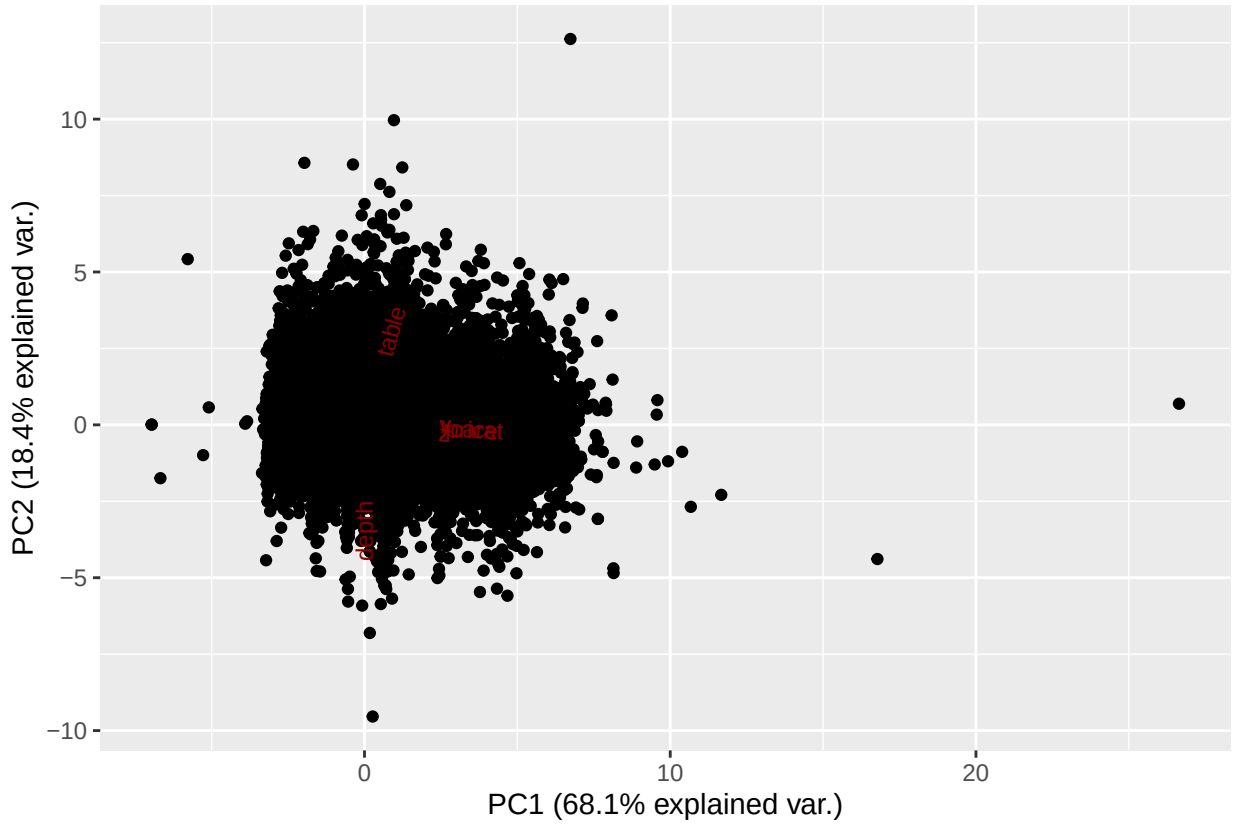


Figure 11: Biplot of Principal Components

The large number of observations in the biplot of PC1 versus PC2 in figure 11 and consequent overlapping of the dots makes interpretation difficult. We will take a smaller sample of 1000 from the diamonds data and redo the PCA after first checking that the smaller sample has similar properties to the full sample. Also of note in this biplot are two significant outliers to the far right, which are extreme values of PC1. We shall identify those before creating the smaller sample.

6.1 Identify the extreme PC1 values

```
sort(decreasing=T,(pca.diamonds$x[,1]))[1:10]
```

```
##     24068     48411     27416     27631     49190     27131     25999     26000
## 26.641670 16.776354 11.669245 10.673276 10.389346  9.928569  9.581991  9.558664
##     26445     27680
##  9.491827  8.919222
```

6.2 PCA with Smaller Sample

```
set.seed(300525287, kind = "Mersenne-Twister")
smallersample <- diamonds[sample(nrow(diamonds), "1000"), ]
reducedpca <- prcomp(smallersample[,-c(2:4,11)], center = TRUE, scale = TRUE)
```

```
pca.var_red <- reducedpca$sdev^2
pca.var.per_red <- round(pca.var_red/sum(pca.var_red)*100,1)
```

```
barplot(pca.var.per_red, names.arg = pca.var.per_red,
        xlab = "Principal Components 1 to 7",
        ylab = "% information retained by each PC",
        main="Screeplot of the Principal Components (reduced dataset)")
```

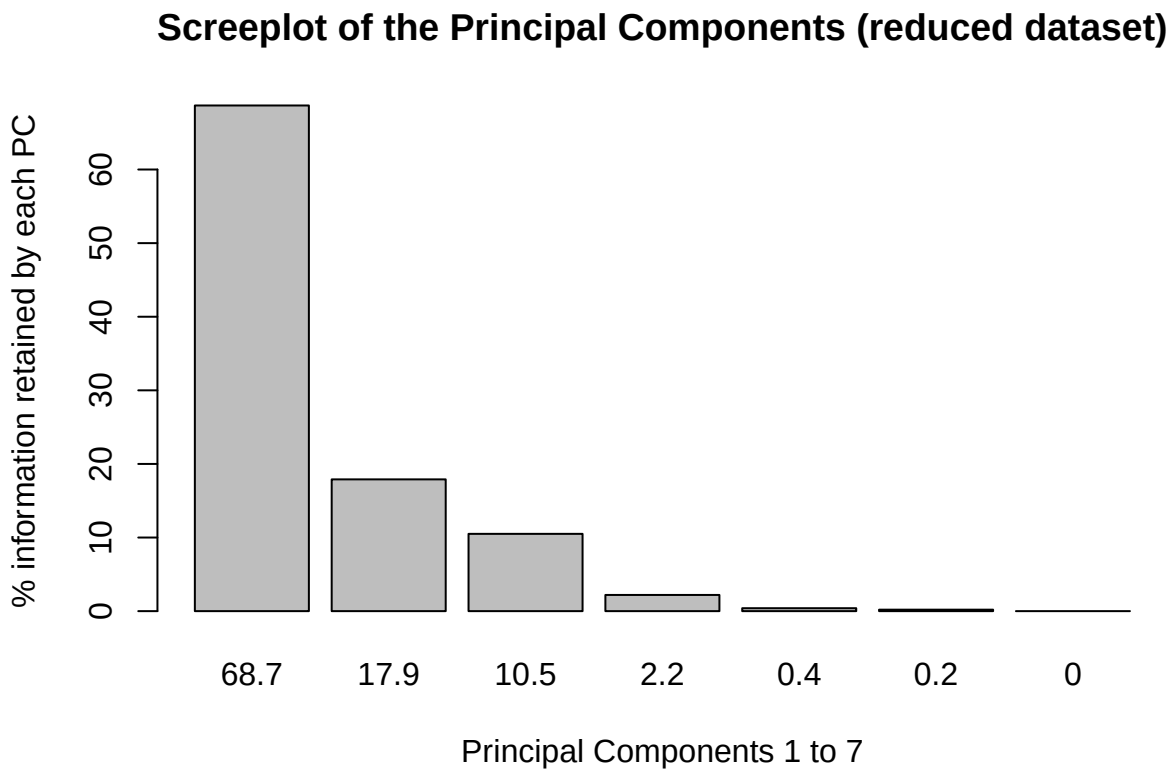


Figure 12: Scree plot of the Principal Components (reduced dataset)

Figure 12 shows the scree plot of the Principal Components for the reduced dataset. PC1 contains 68.7% of the variance, PC2 contains 17.9%, while PC3 contains 10.5%. Between them they contain 97.1%.

```
summary(reducedpca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.1937 1.1191 0.8565 0.39676 0.16404 0.12767 0.03195
## Proportion of Variance 0.6875 0.1789 0.1048 0.02249 0.00384 0.00233 0.00015
## Cumulative Proportion  0.6875 0.8664 0.9712 0.99368 0.99753 0.99985 1.00000
```

We see from the summary output above that the PCA with a sample size of 1000 has almost identical properties to the PCA of the whole dataset, which allows us to proceed.

```
ggbiplot(reducedpca, obs.scale =1, var.scale = 1, varname.size = 5)
```

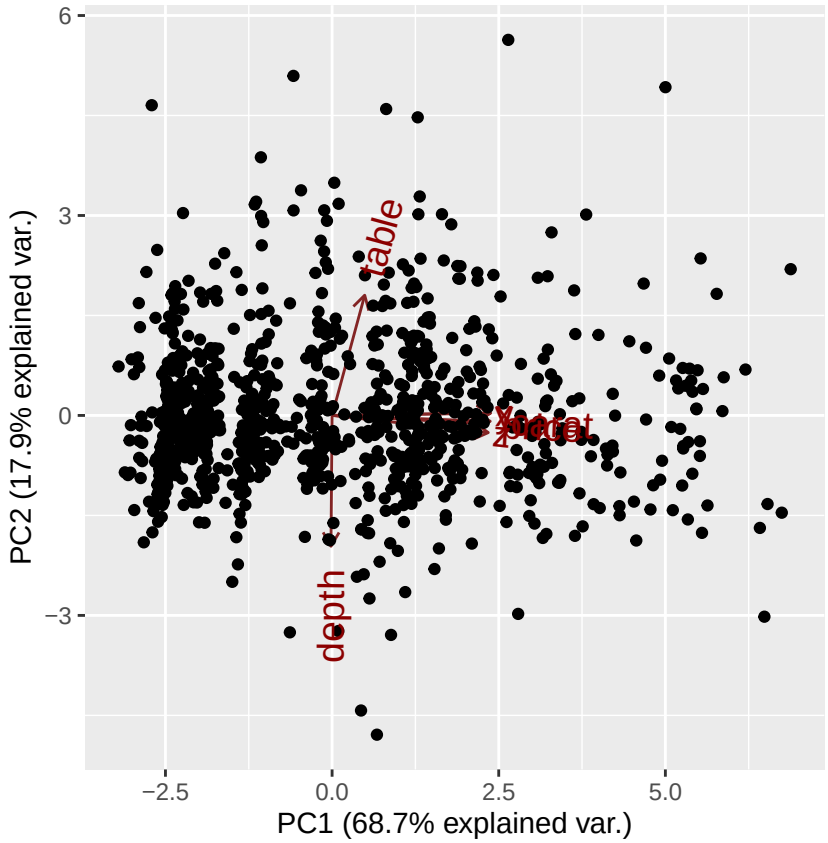


Figure 13: Biplot of the reduced dataset PC1 vs PC2

Producing a biplot using the reduced dataset allows us to interpret the biplot much better (figure 13). We see that carat, x, y, z and price are strongly influence the first PC. While depth and table strongly (and reasonably equally) influence the second PC. We see evidence of the negative correlation between table and depth and a close to zero correlation between these two variables and the variables that strongly influence PC1. We also have an indication that there is redundancy between carat, x, y, z and price. The percentage of variation retained by each Principal Component is included in the axis labels.

In the above analysis we have shown we are able to us principal components analysis to reduce the number of dimensions in the dataset while still retaining a large proportion of its variation. In this investigation we are interested in predicting the price of diamonds. To do this we will perform a regression with principal components with price as the dependent variable and the other numerical variables as explanatory variables.

6.3 Principal Components Regression

We will now perform a principal regression analysis with price as the response variable and and all other numerical variables as the explanatory variables. We Will again use the smaller sample to increase the ease of interpretation of the visual plots. Below we see an initial multiple regression using only the numeric variables

```
numericlm <- lm(scale(price) ~ scale(carat) + scale(table) + scale(depth) + scale(x) + s
summary(numericlm)
```

```
##
## Call:
## lm(formula = scale(price) ~ scale(carat) + scale(table) + scale(depth) +
##     scale(x) + scale(y) + scale(z), data = diamonds_num)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9853 -0.1542 -0.0127  0.0872  3.1982
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.359e-14  1.616e-03   0.000  1.00000
## scale(carat)  1.270e+00  7.509e-03 169.085  < 2e-16 ***
## scale(table) -5.738e-02  1.727e-03 -33.216  < 2e-16 ***
## scale(depth) -7.295e-02  1.976e-03 -36.910  < 2e-16 ***
## scale(x)     -3.699e-01  1.211e-02 -30.547  < 2e-16 ***
## scale(y)      1.899e-02  7.307e-03   2.599  0.00937 **
## scale(z)      7.364e-03  7.837e-03   0.940  0.34744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3752 on 53933 degrees of freedom
## Multiple R-squared:  0.8592, Adjusted R-squared:  0.8592
## F-statistic: 5.486e+04 on 6 and 53933 DF,  p-value: < 2.2e-16
```

In this model all but one (z) of the seven numeric predictors are found to be significant in the model. This model explains 86% of variation in price. We will now attempt to create a more parsimonious model using the principal components.

```
PCAprice <- prcomp(smallersample[,-c(2:4,7,11)], center = TRUE, scale = TRUE)
```

```
pca.varpri <- PCAprice$sdev^2
pca.varpri.per <- round(pca.varpri/sum(pca.varpri)*100,1)
```

```
barplot(pca.varpri.per, names.arg = pca.varpri.per,
        xlab = "Principal Components 1 to 7",
        ylab = "% information retained by each PC",
        main="Screeplot of the regression Principal Components")
```

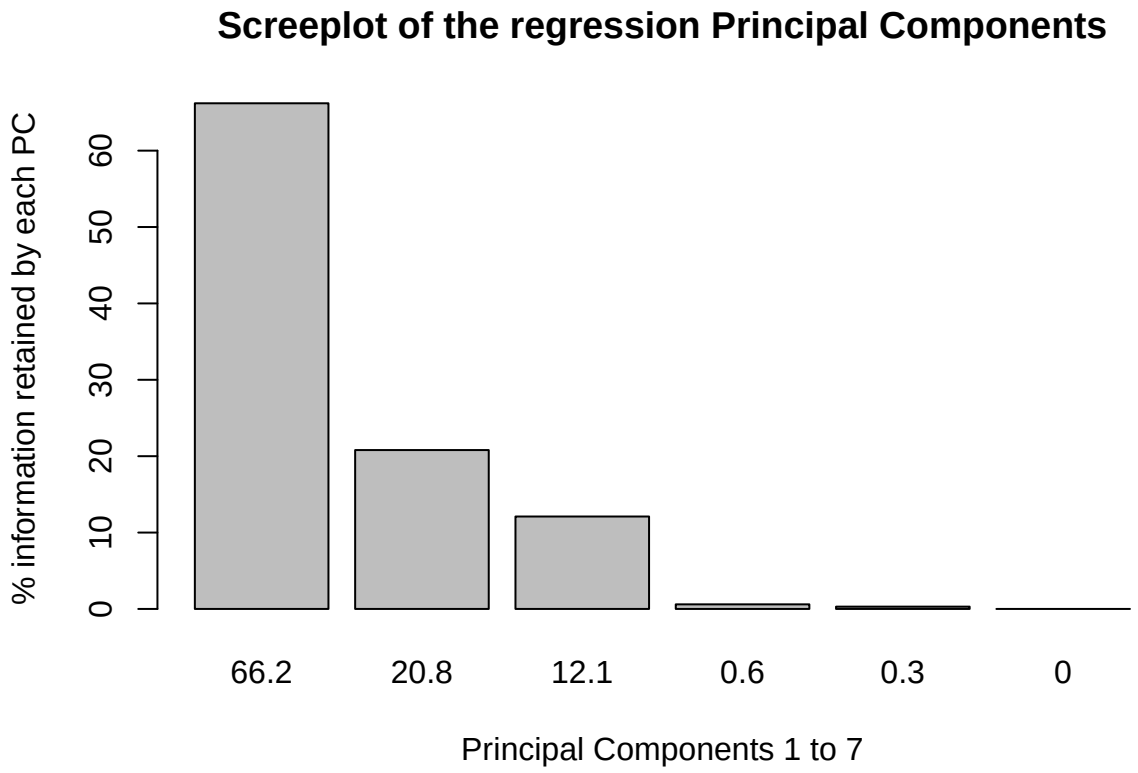


Figure 14: Scree plot of the regression Principal Components

Figure 14 shows the scree plot of the proportion of variance explained by each of the Principal Components of the PC regression. The first three principal components explain nearly all (99%) the variation in the data.

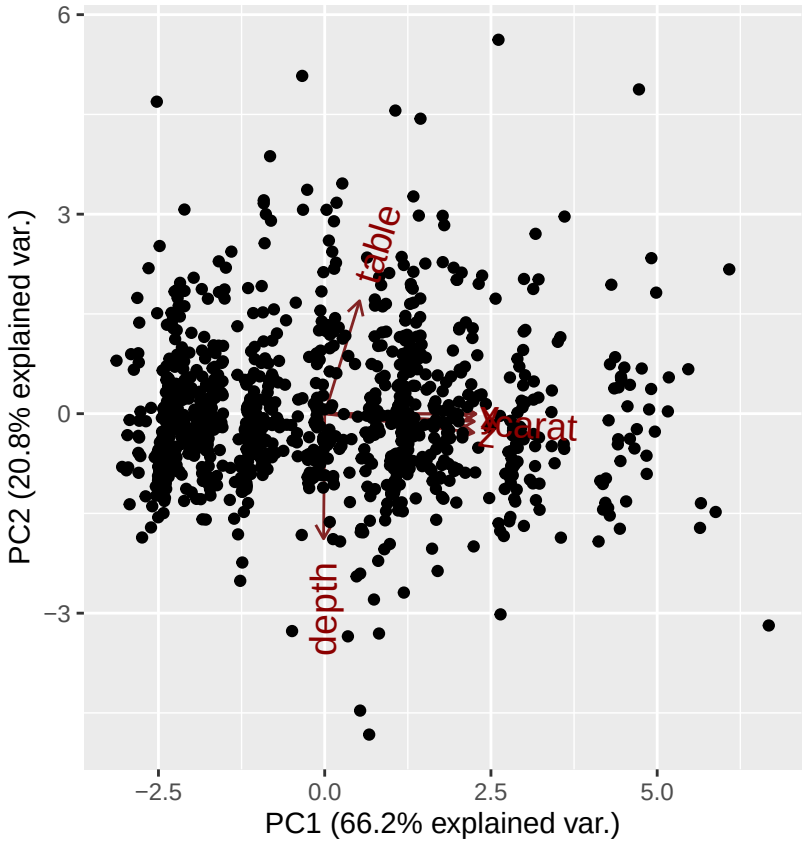
```
PCAprice$rotation
```

```
##                PC1           PC2         PC3          PC4          PC5
## carat  0.494554831 -0.0461150758  0.04445202 -0.806136953 -0.318086228
## depth -0.003656238 -0.7370594608 -0.66518876 -0.040970793  0.112128623
## table  0.115148753  0.6649799315 -0.73788433 -0.006134332 -0.004793799
```

```
## x      0.499625318 -0.0008101947  0.07042242  0.107199954  0.477863033
## y      0.499442121 -0.0022545940  0.07443970  0.134779378  0.482475245
## z      0.493026395 -0.1114123072 -0.02395964  0.564596906 -0.651989189
##                PC6
## carat  0.016652366
## depth  0.001569924
## table  0.002926009
## x     -0.711034219
## y      0.702896417
## z     -0.008867829
```

We see that in the first PC carat, x, y and z all have approximately equal weighting. Again this reflects the strong positive correlation between these two variables. In the second PC depth and table have the strongest weighting. We see that removing price from the PCA does not change much in terms of how much variation the first few components explain as well as which variables contribute most strongly to each component. This is because of how strongly price is correlated with the dimension variables.

```
ggbiplot(PCAprice, obs.scale =1, var.scale = 1, varname.size = 5)
```



Even with price not included we still see the same general structure as the PCA with price.

```
PCApricevars <- prcomp(smallersample[,-c(2:4,7,11)],
                       center = TRUE, scale = TRUE)$x
diamondpricePCA <- lm(smallersample$price ~ PCApricevars[,1] + PCApricevars[,2]
                      + PCApricevars[,3] + PCApricevars[,4] + PCApricevars[,5] +
                        PCApricevars[,6])
summary(diamondpricePCA)
```

```
##
## Call:
## lm(formula = smallersample$price ~ PCApricevars[, 1] + PCApricevars[,
##     2] + PCApricevars[, 3] + PCApricevars[, 4] + PCApricevars[,
##     5] + PCApricevars[, 6])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14748.2   -604.6   -109.9    401.3   7965.2
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3967.32      46.55  85.229  < 2e-16 ***
## PCApricevars[, 1]   1835.78      23.37  78.555  < 2e-16 ***
## PCApricevars[, 2]   -208.03      41.66  -4.993 7.01e-07 ***
## PCApricevars[, 3]    444.25      54.69   8.122 1.35e-15 ***
## PCApricevars[, 4] -4228.97     249.04 -16.981  < 2e-16 ***
## PCApricevars[, 5] -2619.38     349.95  -7.485 1.58e-13 ***
## PCApricevars[, 6] 10444.39    1418.67   7.362 3.79e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1472 on 993 degrees of freedom
## Multiple R-squared:  0.8703, Adjusted R-squared:  0.8695
## F-statistic:  1110 on 6 and 993 DF,  p-value: < 2.2e-16
```

We see that every single principal component is found to be significant in this model when the effects of the other principal components is taken into account. Therefore initially it seems we have failed to create a more parsimonious model for predicting price through using regression with principal components. However if we refit the model with just the first two principal components (see below) we still create a model that is able to explain over 81% of the variance in price with only two components. While this is not as good as the model with all the principal components (which explains 85%) it is still reasonably good and is much more simple. The original numeric multiple regression model explained 86% of the variation in price with 6 explanatory variables while our PCR model explains 81% with just the first two principal components. Principal components regression has allowed us to explain nearly the same amount of variation in price in a much more parsimonius way.

```
diamondpricePCA2 <- lm(smallersample$price ~ PCApricevars[,1] + PCApricevars[,2])
summary(diamondpricePCA2)
```

```
##
## Call:
## lm(formula = smallersample$price ~ PCApricevars[, 1] + PCApricevars[,
##     2])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8846.0 -1162.0  -139.3   877.8  8890.8
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3967.32      56.28   70.49  < 2e-16 ***
## PCApricevars[, 1]   1835.78      28.26   64.97  < 2e-16 ***
## PCApricevars[, 2]   -208.03      50.37   -4.13 3.94e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1780 on 997 degrees of freedom
## Multiple R-squared:  0.8096, Adjusted R-squared:  0.8092
## F-statistic:  2119 on 2 and 997 DF,  p-value: < 2.2e-16
```

It is likely that the remaining, unexplained variation could be explained by the categorical variables in the dataset. As Principal Components Analysis/Regression can only be used for numerical variables we had to leave out the variables of cut, clarity and color.

7 Factor Analysis

In our investigation of the principal components of the dataset we observed that the first two principal components explained the the vast majority of the variation in the data. The first was dominated by the strong positive correlation between price and the dimension variables (x, y, z, carat) which we may call price + dimension. The second was the negative correlation between depth and table which we might combine into light performance as both variables are crucial in giving a diamond its "sparkle". We hypothesise that these two factors (1. "Price + Dimension" 2. "Light Performance") may explain the unobserved variable of "overall diamond quality". To investigate this we will conduct an factor analysis with two factors.

```
factanal(diamonds[,-c(2:4,11)], factors =2)
```

```
##
## Call:
## factanal(x = diamonds[, -c(2:4, 11)], factors = 2)
##
```

```
## Uniquenesses:
## carat depth table price     x     y     z
## 0.041 0.005 0.877 0.205 0.006 0.045 0.039
##
## Loadings:
##       Factor1 Factor2
## carat  0.976
## depth  0.114   0.991
## table  0.152  -0.316
## price  0.885  -0.113
## x      0.987  -0.140
## y      0.967  -0.141
## z      0.980
##
##                Factor1 Factor2
## SS loadings      4.641   1.141
## Proportion Var   0.663   0.163
## Cumulative Var   0.663   0.826
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 22188.2 on 8 degrees of freedom.
## The p-value is 0
```

As we hypothesized, 'carat', 'price', 'x', 'y' and 'z' are well explained by factor 1. 'Depth' is very well explained by factor 2, but 'table' is explained only to a lesser degree. We see some evidence of the factor structure in our hypothesis with the 'price + dimension' variables strongly explained in Factor 1 and 'Light Performance' variables most strongly explained by factor 2. However, we also see other original variables we did not expect in each of our factors. If our hypothesis was correct we would see a negligible loading of depth and table in factor 1 and a negligible loading for the price + dimension variables in factor 2.

Factor 1 explains 66% of the variance in the data and Factor 2 explains an additional 16%. Together are two factors manage to explain 83% of the variation in the data. We also see high uniqueness ratings for 'table' indicating that the two factors do not well account for it's variance.

Overall the hypothesis test gives a highly significant result meaning that two factors are not sufficient to capture the full dimensionality of the data set.

8 Linear Discriminant Analysis (LDA)

As discussed previously the diamonds data set has three categorical variables each with many levels. This means we have the possibility of 280 unique interactions between the levels of these categorical variables. This many factors is unlikely to be that useful as a classification tool. Nevertheless we will begin by attempting a Linear Discriminant Analysis using the factors.

```
diamonds$Interaction <- interaction(diamonds$cut, diamonds$color, diamonds$clarity)
set.seed(123456789, kind = "Mersenne-Twister")
ind <- sample(c("Train", "Test"),
              nrow(diamonds),
              replace = TRUE,
              prob = c(.8, .2))
diamonds.Train <- diamonds[ind == "Train",]
diamonds.Test <- diamonds[ind == "Test",]
LDAinteraction <- lda(Interaction ~ carat + depth + table + price + x + y + z, data = di
```

We will not include the output from the LDA using the interactions as it is very long given there are 280 classes. Instead we will report the key finding; the overall accuracy of using the interactions as a classifier.

```
RPinteraction <- predict(LDAinteraction, diamonds.Test)$class
RCMinteraction <- table(RPinteraction, actual = diamonds.Test$Interaction)
sum(diag(RCMinteraction))/sum(RCMinteraction)
```

```
## [1] 0.06546961
```

An overall accuracy of 0.065 is very low and therefore we can conclude that using the factors as classes was not effective. It's likely the data is not linearly separable into 280 classes. We will now instead attempt to use Linear Discriminant Analysis to classify observations more simply using levels of each of the categorical variables. We will review how each of these categorical variables perform as classifiers.

```
LDAcut <- lda(cut ~ carat + depth + table + price
              + x + y + z, data = diamonds.Train)
LDAcolor <- lda(color ~ carat + depth + table + price
                + x + y + z, data = diamonds.Train)
LDAclarity <- lda(clarity ~ carat + depth + table + price
                  + x + y + z, data = diamonds.Train)
```

```
g1 <- ggord(LDAcut, diamonds.Train$cut, alpha = 0.5, )
g2 <- ggord(LDAcolor, diamonds.Train$color, alpha = 0.5 )
g3 <- ggord(LDAclarity, diamonds.Train$clarity, alpha = 0.5)
ggarrange(g1, g2, g3, ncol = 2, nrow =2)
```

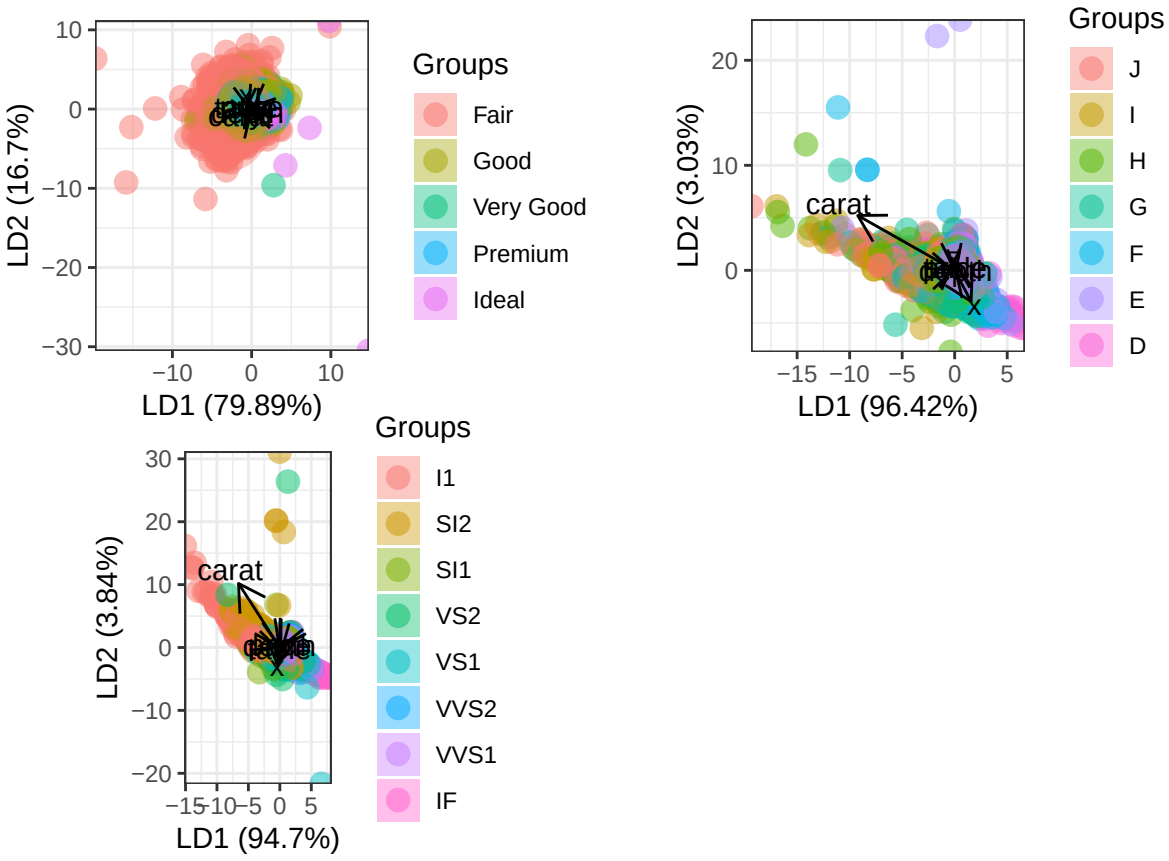


Figure 15: Ordination plots

All three ordination plots (figure 15) show a lot of overlap between different levels of the categorical variables. It seems unlikely that a linear discriminant function would have much success in partitioning the data according to these classes. We shall compute the confusion matrix for each LDA and assess the overall accuracy of each model.

```
RPcut <- predict(LDAcut, diamonds.Test)$class
RCMcut <- table(RPcut, actual = diamonds.Test$cut)
RPcolor <- predict(LDAcolor, diamonds.Test)$class
RCMcolor <- table(RPcut, actual = diamonds.Test$color)
RPclarity <- predict(LDAclarity, diamonds.Test)$class
RCMclarity <- table(RPcut, actual = diamonds.Test$clarity)
```

```
Cutacc <- sum(diag(RCMcut))/sum(RCMcut)
Coloracc <- sum(diag(RCMcolor))/sum(RCMcolor)
Clarityacc <- sum(diag(RCMclarity))/sum(RCMclarity)
```

Table 17: Classification Accuracy for Different LDAs

Class	Accuracy
Cut	0.6258
Color	0.1738
Clarity	0.2039

Table 17 displays the accuracy of the classification achieved by the LDA for the different categorical variables. The only variable that was classified with any success was 'cut'. For 'color' and 'clarity' the LDA performs poorly. As "cut" was the only class to perform moderately well as classifier we will investigate its performance further.

```
LDAcut
```

```
## Call:
## lda(cut ~ carat + depth + table + price + x + y + z, data = diamonds.Train)
##
## Prior probabilities of groups:
##       Fair       Good  Very Good    Premium      Ideal
## 0.02973538 0.09083101 0.22437326 0.25680130 0.39825905
##
## Group means:
##               carat    depth    table    price        x        y        z
## Fair      1.0378845 64.06230 59.02217 4324.327 6.236620 6.172709 3.976776
## Good      0.8474930 62.35461 58.68827 3921.449 5.835173 5.846637 3.637005
## Very Good 0.8047383 61.82327 57.94319 3976.481 5.736677 5.765891 3.558380
## Premium   0.8920031 61.26609 58.75017 4585.801 5.974316 5.941453 3.646883
## Ideal     0.7026392 61.71047 55.94881 3456.299 5.506723 5.519916 3.400944
##
## Coefficients of linear discriminants:
##                 LD1           LD2           LD3           LD4
## carat -1.1399770736 -0.7654500021 -1.6344087398  6.1979429288
## depth -0.4313836891 -0.6258113096 -0.0222994014 -0.3772585315
## table -0.5535698046  0.1099286561  0.1453488495 -0.0892954001
## price  0.0001287086  0.0001352239  0.0001759807 -0.0001005275
## x     -0.5721476535  1.2446100820 -6.7713856573 -4.3794312244
## y      0.4971894042 -1.1101163035  5.9652716094  0.2853573062
## z      0.0399048498 -0.3226799256  1.0868425469  3.8779408496
##
## Proportion of trace:
##    LD1    LD2    LD3    LD4
## 0.7989 0.1670 0.0332 0.0009
```

The first thing we notice is that the prior probabilities for each class are very different. We see small amounts of variation in group means for different variables across different classes

but there is nothing to striking. As the data are on different scales it is harder to interpret the group mean differences. The variables with strongest weight in LD1 are carat, x, y and table while the variables with the most weight in LD2 are x, y, carat and depth. LD1 retains almost 80% of the variation in the data while LD2 retains only 17%.

```
confusionMatrix(RCMcut)
```

```
## Confusion Matrix and Statistics
##
##            actual
## RPcut       Fair Good Very Good Premium Ideal
##   Fair       193  109        27       4     2
##   Good        15  137        95      25     0
##   Very Good   16  223       611     369   101
##   Premium     61  262       691    1746   182
##   Ideal       44  262       992     584  4109
##
## Overall Statistics
##
##                Accuracy : 0.6258
##                  95% CI : (0.6166, 0.6349)
##     No Information Rate : 0.4046
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4484
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: Fair Class: Good Class: Very Good Class: Premium
## Sensitivity              0.58663     0.13797          0.25290         0.6400
## Specificity              0.98652     0.98632          0.91604         0.8529
## Pos Pred Value           0.57612     0.50368          0.46288         0.5935
## Neg Pred Value           0.98708     0.91915          0.81080         0.8760
## Prevalence               0.03029     0.09144          0.22247         0.2512
## Detection Rate           0.01777     0.01262          0.05626         0.1608
## Detection Prevalence     0.03085     0.02505          0.12155         0.2709
## Balanced Accuracy        0.78657     0.56214          0.58447         0.7465
##                      Class: Ideal
## Sensitivity                0.9351
## Specificity                0.7089
## Pos Pred Value             0.6859
## Neg Pred Value             0.9415
## Prevalence                 0.4046
```

```
## Detection Rate              0.3784
## Detection Prevalence        0.5517
## Balanced Accuracy           0.8220
```

The confusion matrix clearly shows a large amount of errors for each individual class. We again get the overall accuracy measurement of 0.6258 with a 95% confidence interval of (0.6166,0.6349). In terms of the accuracy of classification for individual classes Ideal, Fair and Premium had higher accuracy ratings than Good or Very Good. It was important to check individual class accuracy as the prior probabilities of an observation belonging to each class were very different.

We conclude that the data is not linearly separable by using the levels of categorical variables as classes. It's possible performance may be better by using the 280 interactions as known classes but we couldn't see much real world use for such a complex classifier.

8.1 Why prediction without the categorical variables is not working well

8.1.1 Create modified version of dataset only containing carat values between 1 and 1.1

```
library(dplyr)
diamonds_car2 <- diamonds[,c(1,4,7)]
diamonds_car2[1:4,]
```

```
##   carat clarity price
## 1  0.23     SI2   326
## 2  0.21     SI1   326
## 3  0.23     VS1   327
## 4  0.29     VS2   334
```

```
diamonds_car19_21 <- filter(diamonds_car2, between(carat, 1, 1.1))
diamonds_car19_21[1:4,]
```

```
##   carat clarity price
## 1  1.01      I1  2781
## 2  1.01      I1  2788
## 3  1.01     SI2  2788
## 4  1.05     SI2  2789
```

```
# scatterplot of carat vs price, but also colouring the observations
# by 'cut'
ggplot(data=diamonds_car19_21, aes(x=carat, y=price, color=clarity))+
  geom_point()+
  guides(colour=guide_legend(reverse = T))+
  labs(title = "'Carat' vs 'price', coloured by 'clarity'")
```

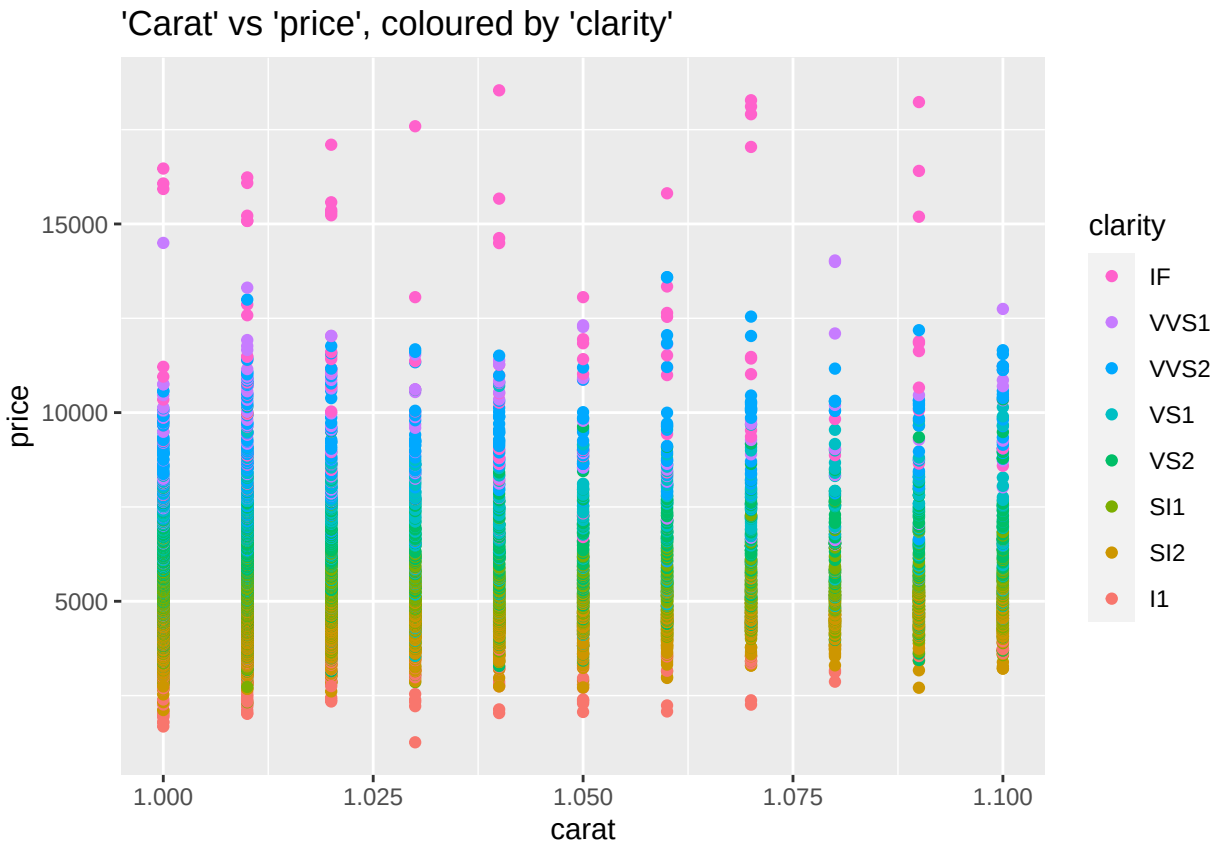


Figure 16: Carat (limited values) vs Price coloured by Clarity

Figure 16 shows a the scatterplot of 'price' vs 'carat' vs 'clarity', but just for values of 'carat' between 1 and 1.1, which helps us see the pattern more clearly. Firstly, note the vertical spread of the data points; most of the values begin at around 2500 USD and generally have extreme upper values larger than 15,000 USD, so a spread of approximately 10,000 USD. There are clear colour bands running horizontally, with the lowest quality 'clarity' diamonds near the bottom of the price range, while the highest values are near the top.

So, using 'carat' alone to predict these prices clearly is not enough; given the spread of price for any given value of carat, the categorical variables play an important in determining where in the range any given diamond will be. This was the reason our regression models that excluded the categorical variables did not perform nearly as well as those that included them.

8.2 Discriminant analysis using the reduced dataset (price, carat and clarity)

To address the issue above, we attempted another discriminant analysis using the reduced dataset above, that is, for 'carat' values between 1.0 and 1.1. The hope was that the relatively clear bands seen in figure 16 would enable the LDA to this time, as there is less overlap between levels in the reduced dataset than there is in the original.

So, we repeat the Linear Discriminant Analysis process using the reduced dataset.

```
# split dataset into 'train' and 'test'
set.seed(1234567890, kind = "Mersenne-Twister")
ind <- sample(c("Tr","Te"),
              nrow(diamonds_car19_21),replace = TRUE,prob = c(0.6,0.4))

# display first few elements to check
ind[1:15]
```

```
##  [1] "Te" "Tr" "Te" "Te" "Te" "Te" "Tr" "Te" "Te" "Te" "Te" "Tr" "Tr" "Te" "Te"
```

```
# check that the proprortions have come out as expected
nrow(diamonds_car19_21)
```

```
## [1] 7568
```

```
table(ind)
```

```
## ind
##   Te   Tr
## 3008 4560
```

```
prop.table(table(ind))
```

```
## ind
##       Te       Tr
## 0.397463 0.602537
```

The output above confirms that the random allocation of values to either the training or testing datasets produced the desired result of approximately 60% in the training set and 40% in the testing set.

8.2.1 Create training and test subsets

```
Train <- diamonds_car19_21[ind=="Tr",]
Test <- diamonds_car19_21[ind=="Te",]
```

8.2.2 Create LDA and display summary

```
library(MASS)
LDA.diam <- lda(clarity~carat+price,
           data=Train)
```

8.2.3 Boxplots of the LDA means

```
boxplot(t(LDA.diam$means), main=
          "Distribution of LDA means for each level of 'clarity'",
```

```
        xlab="Levels of 'clarity' (lowest on the left, highest on the right)",
        ylab="Count")
```

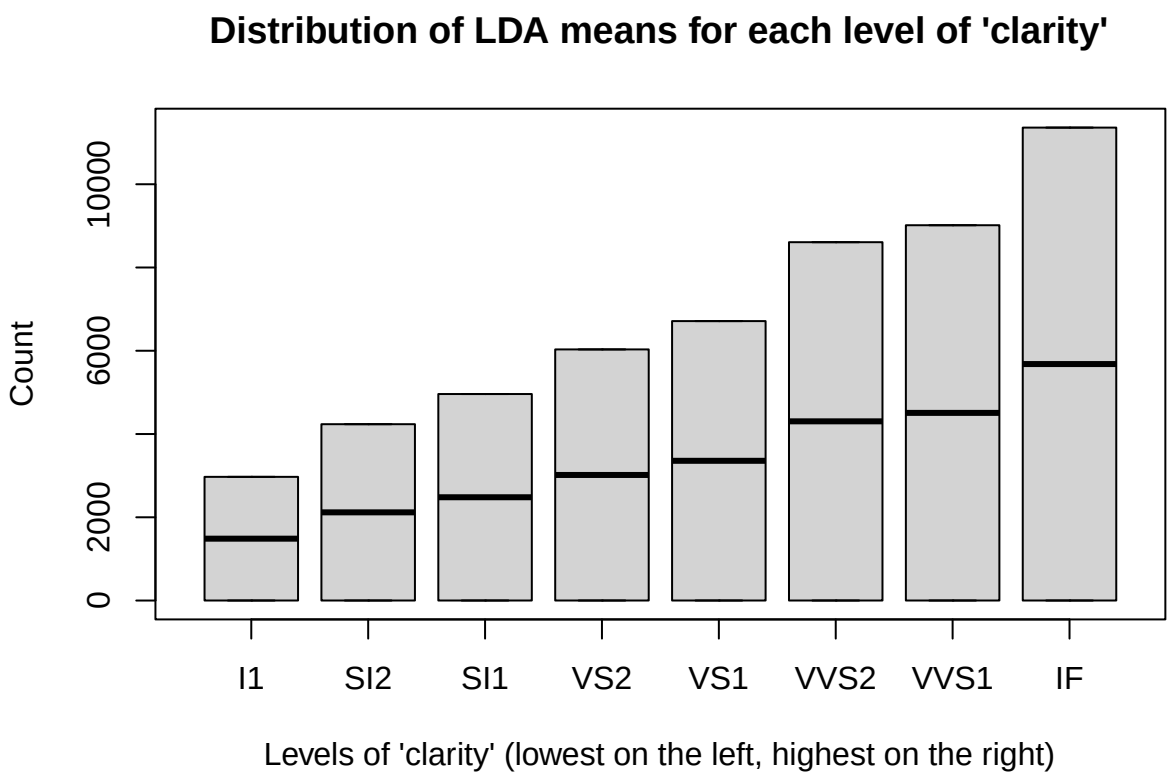


Figure 17: Distribution of LDA means

Figure 17 shows the distribution of the LDA means. It is obvious that there is an increasing number of diamonds of higher 'clarity' levels.

8.2.4 Create Prediction object

```
Pred <- predict(LDA.diam)
```

8.2.5 Biplot

```
library(klaR)
library(ggord)
library(devtools)
ggord(LDA.diam,Train$clarity)
```

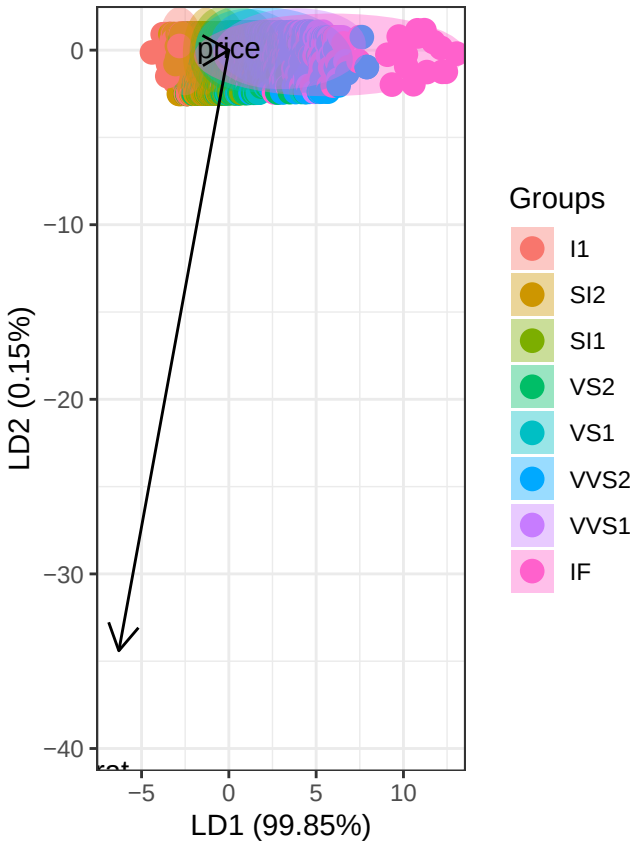


Figure 18: Biplot of 'carat', 'price' against 'clarity'

Figure 18 shows the biplot for 'carat' versus 'price' against the different levels of 'clarity'. There is still a large amount of overlap visible near the top of the plot (the different coloured dots), meaning that separation of the levels is not going to be easy.

8.2.6 Partition plot

```
partimat(clarity~carat+price, data = Train, method="lda")
```

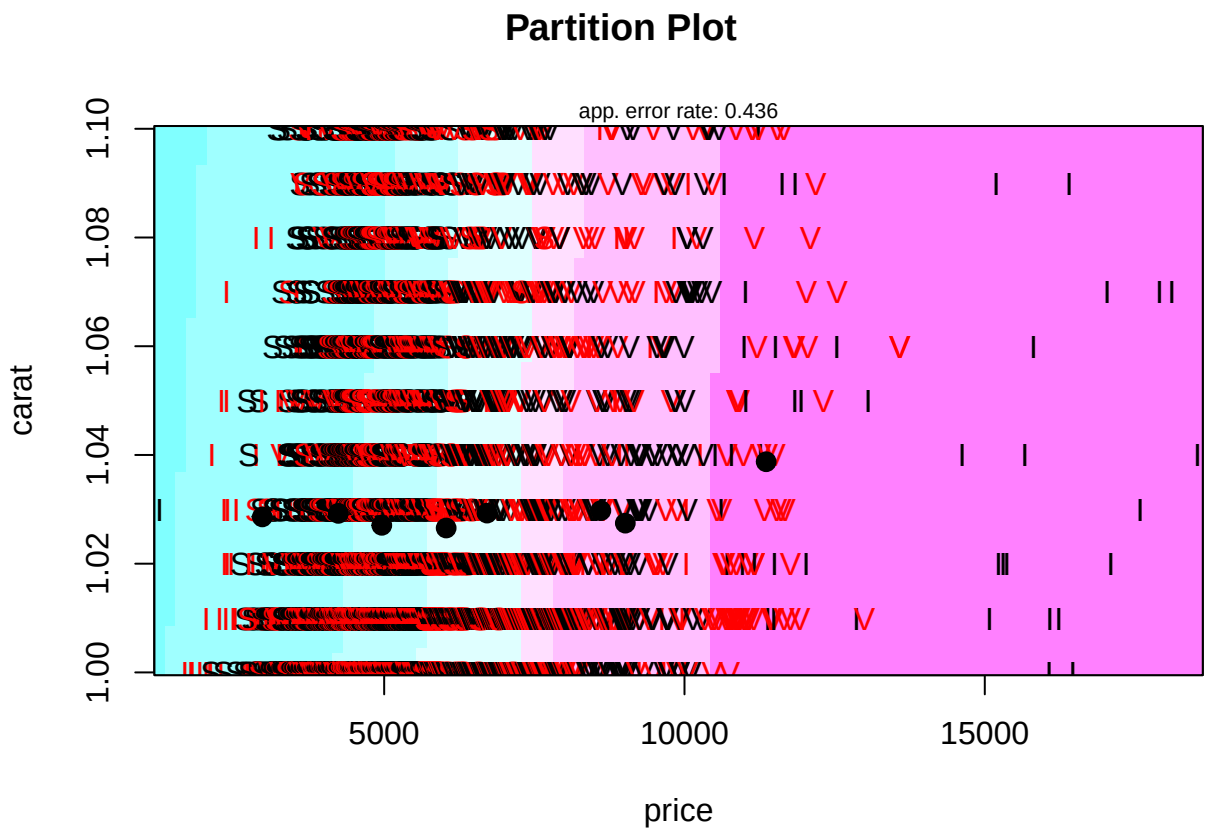


Figure 19: Partition plot for levels of 'clairty' as predicted by 'carat' and 'price'

Figure 19 shows the partition plot for the discriminant analysis (DA) of the different levels of the 'clarity' variable as predicted by 'carat' and 'price'. As a reminder, the levels are: I1 (lowest clarity), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (highest clarity). We can see from the number of red font values that there are many mis-categorised observations, suggesting that the LDA has not worked particularly well.

```
RealisticPredicted <- predict(LDA.diam,Test)$class
RCM <- table(RealisticPredicted, Actual=Test$clarity)
RCM
```

```
##                    Actual
## RealisticPredicted  I1 SI2 SI1 VS2 VS1 VVS2 VVS1  IF
##                I1    0   0   0   0   0    0    0   0
##                SI2  65 545 185  40  24    4    0   0
##                SI1   0 260 578 158  57    5    4   0
##                VS2   0   3  76 366 155   30    3   1
##                VS1   0   0   1  45  81   28    9   1
##                VVS2  0   1   0  12  43  118   31  21
##                VVS1  0   0   0   0   0    0    0   0
##                IF    0   0   0   0   1   14   17  26
```

The confusion matrix above shows that there were many mis-categorisations. There is a pattern evident. The leading diagonal shows the correct predictions. There are many incorrect predictions for values immediately above and below the leading diagonal, meaning that the DA partitions struggled to separate adjacent categories.

```
prop_succ <- sum(diag(RCM))/sum(RCM)
prop_succ
```

```
## [1] 0.5698138
```

```
per_succ <- signif(prop_succ*100,4)
per_succ
```

```
## [1] 56.98
```

The output above shows that the percentage of successful predictions was 56.98%, indicating that this LDA is a mediocre predictor.

8.2.7 Summary of LDA

The relatively poor performance of LDA in this instance is perhaps not surprising given the amount of overlap we saw in figure 16 earlier (the scatterplot of 'price' vs 'carat' vs 'clarity' for 'carat' values between 1 and 1.1); there simply isn't enough geometrical separation between the different levels of 'clarity' for the algorithm to accurately categorise the boundary cases.

9 Cluster Analysis

As LDA did not yield useful results in terms of classifying the diamonds data we also decided to conduct Cluster Analysis to investigate if the diamonds data set could be divided into clusters. We used a K-means algorithm to achieve this. As we were initially unsure of how many clusters we should attempt to divide the data into we will first examine the sum of the squares of the intra-class disatances for different numbers of clusters. We will again use the smaller sample to aid in the interpretation of visual displays.

9.1 Optimal cluster number

```
fviz_nbclust((scale(smallersample[,-c((2:4),11)])), kmeans, method = "wss") +
    geom_vline(xintercept = 4, linetype = 2)+
  labs(subtitle = "Elbow method")
```

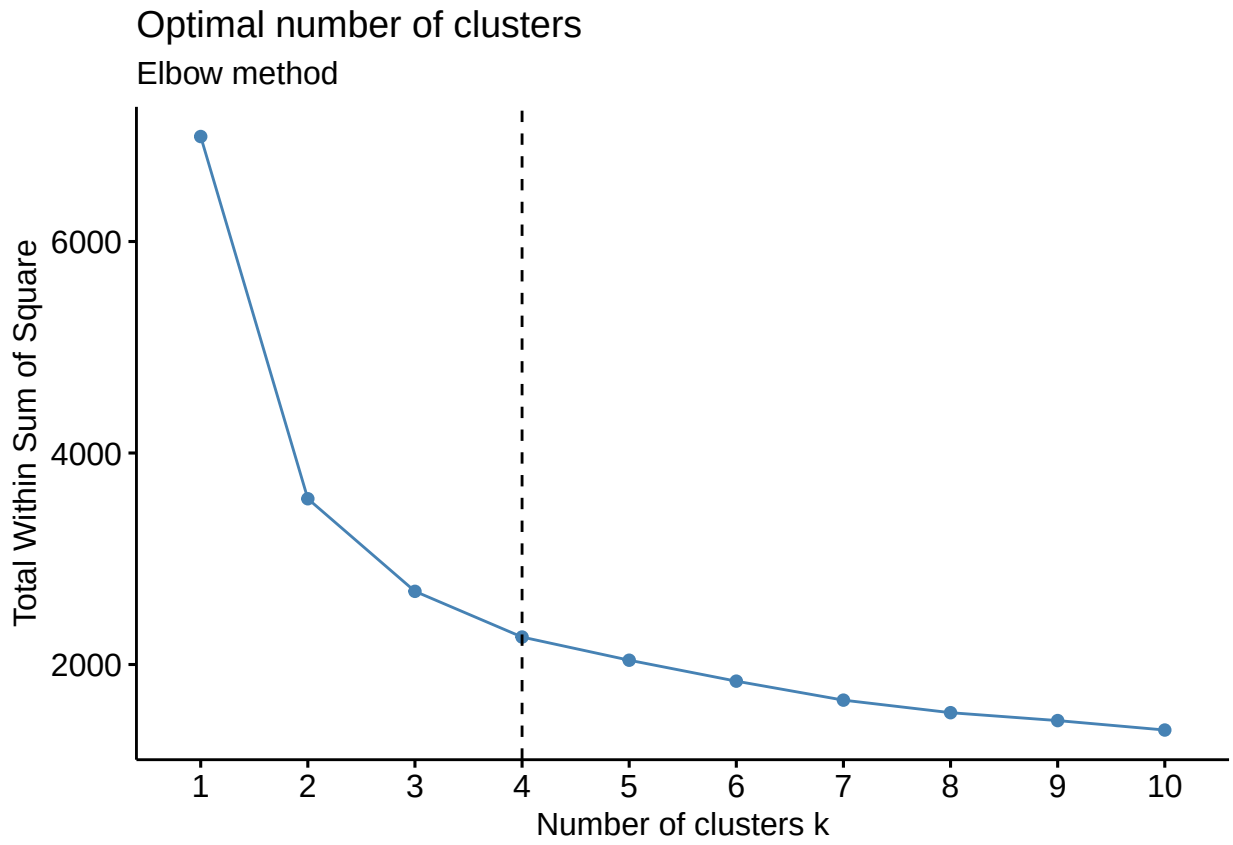


Figure 20: Optimal number of clusters

Figure 20 shows that when the elbow method is used, we see that 4 clusters is the optimum number.

9.2 kmeans function

```
clusters <- kmeans(scale(smallersample[,-c((2:4),11)]),
                   algorithm = "MacQueen", centers = 4,
                   iter.max = 100, nstart = 50)

# print clusters means
clusters$centers
```

```
##        carat      depth       table       price          x          y
## 1  0.3436636  0.4625156 -0.19171931  0.17367610  0.4846585  0.4876979
## 2 -0.8302592  0.0195466 -0.31324680 -0.71288391 -0.9107135 -0.9109156
## 3  1.9509740  0.1578009  0.05304506  2.09674493  1.7073677  1.7102105
## 4  0.2424997 -1.1838989  1.40775657  0.06611287  0.4472637  0.4388989
##            z
## 1  0.5477885
```

```
## 2 -0.9007307
## 3  1.7128313
## 4  0.2761387
```

```
# print first 20 values
clusters$cluster[1:20]
```

```
## 44678  2911  2255 27164 48502 47496 47160 11609 35207 33369 29772 33632 47294
##     4     4     1     3     2     2     1     1     2     2     2     2     2
## 39481 14569 25563 45445 23592 13383  1549
##     2     1     3     2     3     2     1
```

We create four clusters with sizes of 127, 455, 137 and 281 respectively. We see clear differences in the mean of each numeric variable across the four clusters. The within cluster sum of squares indicates that the clusters are not very compact. We shall now visualize the clusters with a scatterplot using the first two principal components of our dataset as the two dimensions. The first two principal components account for over 86% of the variance so this should be an appropriate visualization.

9.3 Cluster plot

```
fviz_cluster(clusters, data = (scale(smallersample[,-c((2:4),11)])),
             geom = "point",
             ellipse.type = "norm",
             ggtheme = theme_bw()
             )
```

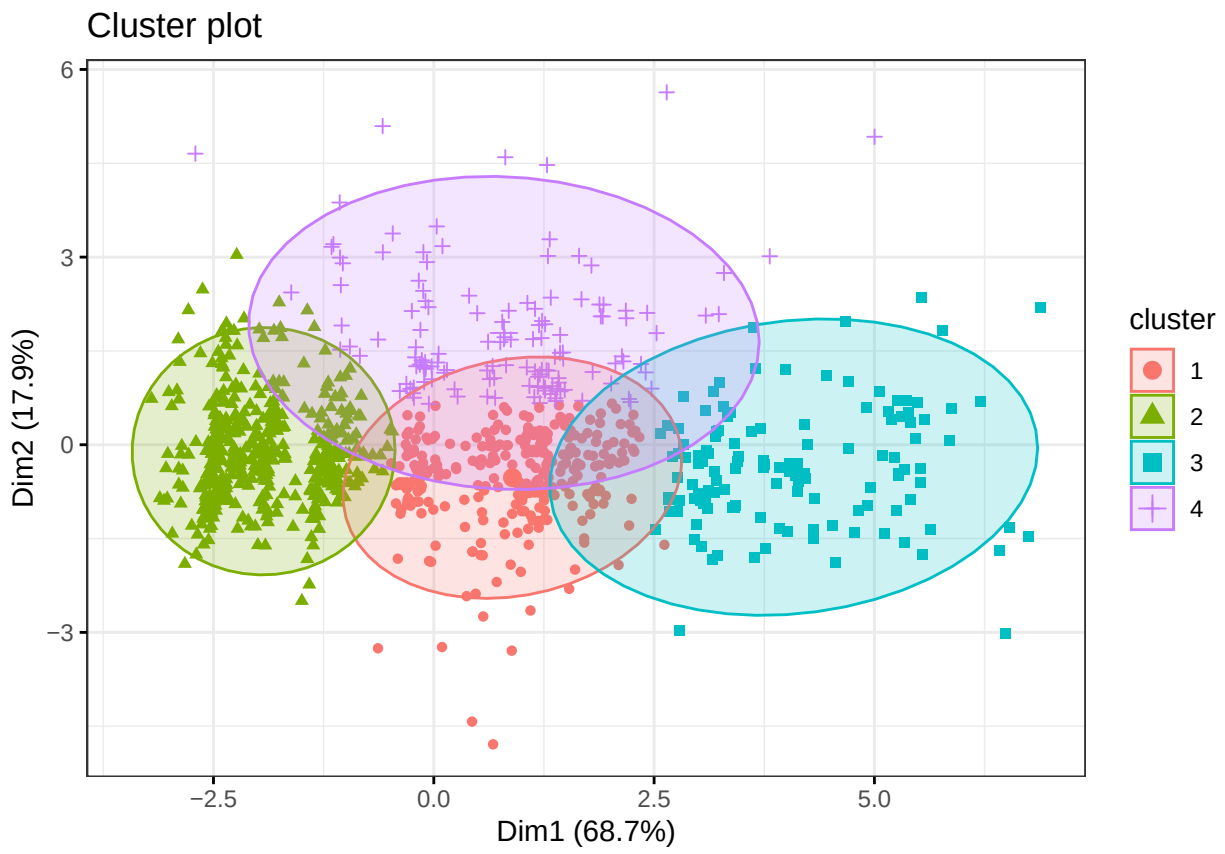


Figure 21: Cluster plot

In the cluster plot (figure 21) we see significant overlap between the four clusters as well as many observations that do not seem to fit well in any of the clusters. It seems that the diamond dataset does not easily split into distinct clusters.

It is worth mentioning that 'diamonds' already has categories and levels within those categories, and that the clustering above does not take the pre-existing categories into account, but attempts to find naturally occurring classes within the data.

9.4 Silhouette plot

```
library(cluster)
sil <- silhouette(clusters$cluster, dist((scale(smallersample[,-c((2:4),11)]))))
fviz_silhouette(sil)
```

```
##   cluster size ave.sil.width
## 1       1  283          0.29
## 2       2  455          0.51
## 3       3  127          0.34
## 4       4  135          0.21
```

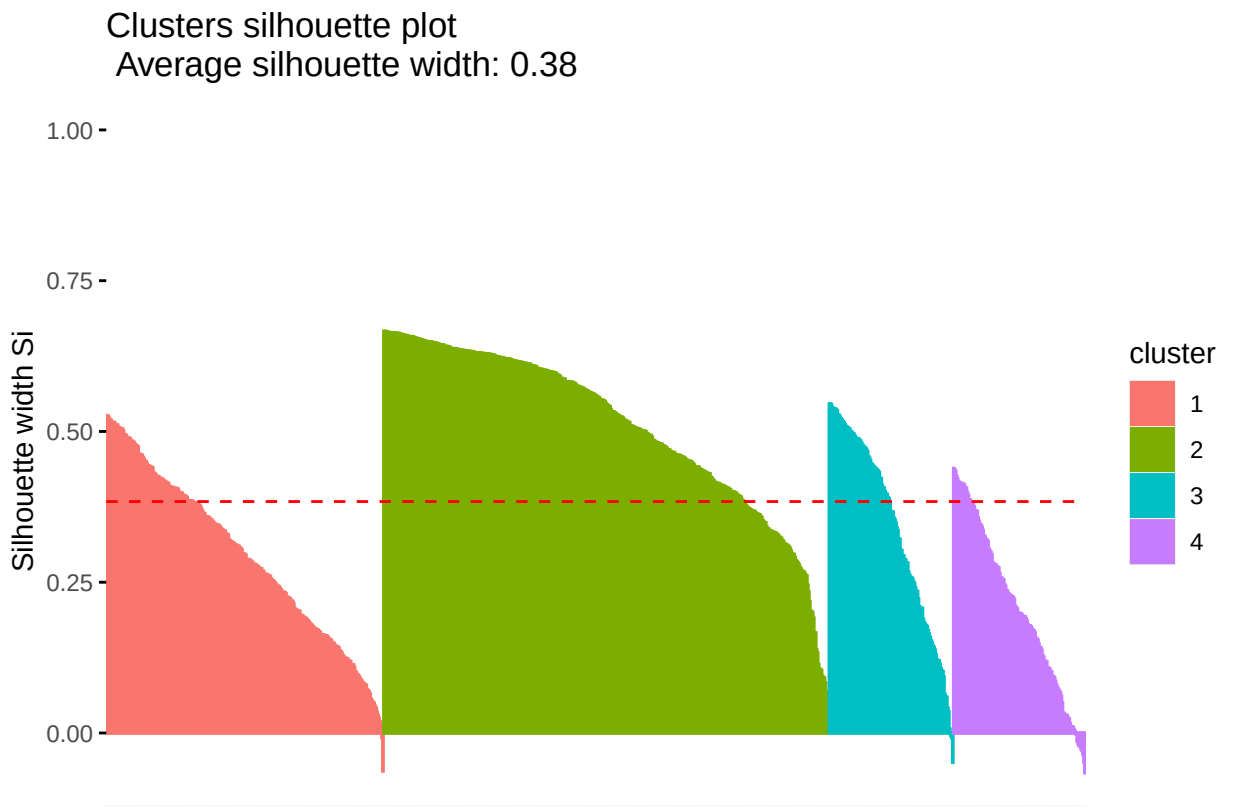


Figure 22: Clusters silhouette plot

The silhouette plot (figure 22) helps us to visualize how similar points each observation is to the cluster that it is assigned too. We see very few negative values indicating that very few observations have been assigned to the wrong cluster. However we do have a large proportion of values with low positive values indicating that these observations were close to the borderline between two clusters. We see that the average silhouette width is 0.38.

10 Conclusion

During this investigation we applied a number of the new analytical methods that we learned in STAT394. As this project was principally a learning task we attempted them all, even though an experienced statistician may have concluded beforehand that they were inappropriate or of dubious value. We mostly performed these techniques on a subset of the data that only included the numerical variables. Not all of the analytical techniques worked as we had hoped, and when this occurred we investigated why.

The primary aims of our multivariate statistical investigation on the diamonds data set were as follows

- To find the best way of predict diamond price using the other variables in the data set

- To attempt to classify observations, in particular the different levels of the three categorical variables
- To explore the new statistical techniques of PCA, LDA, FA and CA in relation to our data set

We had the most success with our first aim; predicting diamond price. This was first achieved using the multiple regression model with all variables except for 'y' and 'z', in other words the model; price = carat + cut + color + clarity + depth + table + x. The best model was obtained using stepwise AIC and BIC tests. The diagnostics for the multiple regression model showed that all the assumptions were violated, meaning the model might not produce accurate predictions. Secondly, a principal components analysis using the numerical variables was successful in reducing the dimensionality of the dataset while still retaining a large proportion of the variance. We were able to use the first two principal components in a model using principal components regression and this performed almost as well as the "best" model identified through stepwise multiple regression despite being much more parsimonious. We did however note that a model including all principal components was more effective at predicting price.

An examination of the key principal components inspired us to conduct a factor analysis with the hypothesis that the factors "price + dimension" and "light conductance" might be sufficient to explain variation in the dataset. While we saw evidence of our hypothesized structure in the factor loadings two factors was not found to be sufficient to capture all the variation in the data.

We attempted to address our second aim through the techniques of LDA and CA. A LDA was undertaken using each categorical variable as a classifier. This failed to reliably discriminate observations. As using known classes in the dataset was unsuccessful we attempted a cluster analysis. The cluster analysis identified four clusters as the optimal number to describe the data however there was significant overlap between these. We conclude that the diamonds dataset presents a difficult classification problem and that the reason for this is the high number of possible interactions between levels of our three categorical variables.

11 Appendices

11.1 Code for Producing KS Tests

```
ks.test(diamonds$carat, "pnorm", mean=mean(diamonds$carat), sd=sd(diamonds$carat))
```

```
## Warning in ks.test.default(diamonds$carat, "pnorm", mean =
## mean(diamonds$carat), : ties should not be present for the Kolmogorov-Smirnov
## test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
```

```
## data:  diamonds$carat
## D = 0.12274, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(diamonds$depth, "pnorm", mean=mean(diamonds$depth), sd=sd(diamonds$depth))
```

```
## Warning in ks.test.default(diamonds$depth, "pnorm", mean =
## mean(diamonds$depth), : ties should not be present for the Kolmogorov-Smirnov
## test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  diamonds$depth
## D = 0.075871, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(diamonds$table, "pnorm", mean=mean(diamonds$table), sd=sd(diamonds$table))
```

```
## Warning in ks.test.default(diamonds$table, "pnorm", mean =
## mean(diamonds$table), : ties should not be present for the Kolmogorov-Smirnov
## test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  diamonds$table
## D = 0.13225, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(diamonds$price, "pnorm", mean=mean(diamonds$price), sd=sd(diamonds$price))
```

```
## Warning in ks.test.default(diamonds$price, "pnorm", mean =
## mean(diamonds$price), : ties should not be present for the Kolmogorov-Smirnov
## test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  diamonds$price
## D = 0.18467, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(diamonds$x, "pnorm", mean=mean(diamonds$x), sd=sd(diamonds$x))
```

```
## Warning in ks.test.default(diamonds$x, "pnorm", mean = mean(diamonds$x), : ties
## should not be present for the Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
```

```
##
## data:  diamonds$x
## D = 0.093545, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(diamonds$y, "pnorm", mean=mean(diamonds$y), sd=sd(diamonds$y))
```

```
## Warning in ks.test.default(diamonds$y, "pnorm", mean = mean(diamonds$y), : ties
## should not be present for the Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  diamonds$y
## D = 0.088528, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(diamonds$z, "pnorm", mean=mean(diamonds$z), sd=sd(diamonds$z))
```

```
## Warning in ks.test.default(diamonds$z, "pnorm", mean = mean(diamonds$z), : ties
## should not be present for the Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  diamonds$z
## D = 0.089273, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

11.2 Code for producing ANOVA/Levene's Test/Kruskal Wallis/Tukey of Price by Cut

```
cutanova <- aov(price ~ cut, data = diamonds)
summary(cutanova)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## cut             4 1.104e+10 2.760e+09   175.7 <2e-16 ***
## Residuals   53935 8.474e+11 1.571e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(price ~ cut, data= diamonds)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     4   123.6 < 2.2e-16 ***
##       53935
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
kruskal.test(price ~ cut, data = diamonds)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  price by cut
## Kruskal-Wallis chi-squared = 978.62, df = 4, p-value < 2.2e-16
```

```
TukeyHSD(cutanova, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = price ~ cut, data = diamonds)
##
## $cut
##                           diff         lwr        upr     p adj
## Good-Fair           -429.89331  -740.44880  -119.3378 0.0014980
## Very Good-Fair      -376.99787  -663.86215   -90.1336 0.0031094
## Premium-Fair         225.49994   -59.26664   510.2665 0.1950425
## Ideal-Fair          -901.21579 -1180.57139  -621.8602 0.0000000
## Very Good-Good        52.89544  -130.15186   235.9427 0.9341158
## Premium-Good         655.39325   475.65120   835.1353 0.0000000
## Ideal-Good          -471.32248  -642.36268  -300.2823 0.0000000
## Premium-Very Good    602.49781   467.76249   737.2331 0.0000000
## Ideal-Very Good     -524.21792  -647.10467  -401.3312 0.0000000
## Ideal-Premium      -1126.71573 -1244.62267 -1008.8088 0.0000000
```

11.3 Code for producing ANOVA/Levene's Test/Kruskal Wallis/Tukey of Price by Clarity

```
clarityanova <- aov(price ~ clarity, data = diamonds)
summary(clarityanova)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## clarity         7 2.331e+10 3.330e+09     215 <2e-16 ***
## Residuals   53932 8.352e+11 1.549e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(price ~ clarity, data= diamonds)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
```

```
##          Df F value    Pr(>F)
## group     7  77.809 < 2.2e-16 ***
##       53932
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
kruskal.test(price ~ clarity, data = diamonds)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  price by clarity
## Kruskal-Wallis chi-squared = 2718.2, df = 7, p-value < 2.2e-16
```

```
TukeyHSD(clarityanova, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = price ~ clarity, data = diamonds)
##
## $clarity
##                    diff          lwr         upr     p adj
## SI2-I1     1138.8599147   683.395891  1594.32394 0.0000000
## SI1-I1       71.8324571  -378.570901   522.23582 0.9997320
## VS2-I1        0.8207037  -450.377702   452.01911 1.0000000
## VS1-I1      -84.7132999  -542.298929   372.87233 0.9992819
## VVS2-I1    -640.4316203 -1109.531923  -171.33132 0.0009165
## VVS1-I1   -1401.0540535 -1881.569711  -920.53840 0.0000000
## IF-I1     -1059.3295848 -1580.334655  -538.32451 0.0000000
## SI1-SI2   -1067.0274575 -1229.386830  -904.66808 0.0000000
## VS2-SI2   -1138.0392109 -1302.591274  -973.48715 0.0000000
## VS1-SI2   -1223.5732146 -1404.907129 -1042.23930 0.0000000
## VVS2-SI2  -1779.2915349 -1987.983831 -1570.59924 0.0000000
## VVS1-SI2  -2539.9139681 -2773.136347 -2306.69159 0.0000000
## IF-SI2    -2198.1894995 -2506.318797 -1890.06020 0.0000000
## VS2-SI1     -71.0117534  -220.988718    78.96521 0.8410824
## VS1-SI1    -156.5457571  -324.764949    11.67343 0.0899007
## VVS2-SI1   -712.2640774  -909.667681  -514.86047 0.0000000
## VVS1-SI1  -1472.8865106 -1696.064436 -1249.70859 0.0000000
## IF-SI1    -1131.1620420 -1431.760399  -830.56369 0.0000000
## VS1-VS2     -85.5340037  -255.870471    84.80246 0.7958312
## VVS2-VS2   -641.2523240  -840.463263  -442.04138 0.0000000
## VVS1-VS2  -1401.8747572 -1626.652874 -1177.09664 0.0000000
## IF-VS2    -1060.1502885 -1361.938605  -758.36197 0.0000000
## VVS2-VS1   -555.7183203  -769.001243  -342.43540 0.0000000
```

```
## VVS1-VS1  -1316.3407535 -1553.679770 -1079.00174 0.0000000
## IF-VS1     -974.6162849 -1285.873083  -663.35949 0.0000000
## VVS1-VVS2  -760.6224332 -1019.466585  -501.77828 0.0000000
## IF-VVS2    -418.8979645  -746.848084   -90.94785 0.0027364
## IF-VVS1     341.7244687    -2.356168   685.80510 0.0531204
```

11.4 Code for producing ANOVA/Levene's Test/Kruskal Wallis/Tukey of Price by Color

```
coloranova <- aov(price ~ color, data = diamonds)
summary(coloranova)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## color           6 2.685e+10 4.475e+09   290.2 <2e-16 ***
## Residuals   53933 8.316e+11 1.542e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(price ~ color, data= diamonds)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     6  219.12 < 2.2e-16 ***
##       53933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
kruskal.test(price ~ color, data = diamonds)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  price by color
## Kruskal-Wallis chi-squared = 1335.6, df = 6, p-value < 2.2e-16
```

```
TukeyHSD(coloranova, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = price ~ color, data = diamonds)
##
## $color
##            diff         lwr         upr     p adj
## I-J  -231.94307  -501.11666    37.23053 0.1449244
## H-J  -837.14882 -1089.88345  -584.41420 0.0000000
```

```
## G-J -1324.68235 -1568.82093 -1080.54376 0.0000000
## F-J -1598.93162 -1847.48867 -1350.37458 0.0000000
## E-J -2247.06554 -2494.88601 -1999.24508 0.0000000
## D-J -2153.86392 -2413.70535 -1894.02250 0.0000000
## H-I  -605.20576  -807.34909  -403.06243 0.0000000
## G-I -1092.73928 -1284.02646  -901.45210 0.0000000
## F-I -1366.98856 -1563.88381 -1170.09331 0.0000000
## E-I -2015.12248 -2211.08707 -1819.15789 0.0000000
## D-I -1921.92086 -2132.88224 -1710.95948 0.0000000
## G-H  -487.53352  -654.89884  -320.16821 0.0000000
## F-H  -761.78280  -935.53004  -588.03556 0.0000000
## E-H -1409.91672 -1582.60860 -1237.22484 0.0000000
## D-H -1316.71510 -1506.25419 -1127.17600 0.0000000
## F-G  -274.24927  -435.23673  -113.26182 0.0000106
## E-G  -922.38320 -1082.23107  -762.53532 0.0000000
## D-G  -829.18158 -1007.09708  -651.26607 0.0000000
## E-F  -648.13392  -814.65208  -481.61576 0.0000000
## D-F  -554.93230  -738.86403  -371.00057 0.0000000
## D-E    93.20162   -89.73351   276.13675 0.7437450
```

Bibiography

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2020. *RMarkdown: Dynamic Documents for r.* https://rmarkdown.rstudio.com.

"Diamonds Dataset, Kaggle.com." 2016. https://www.kaggle.com/datasets/shivam2503/diamonds.

"Github.com." 2022. https://github.com/arinicholas/STAT394-Group-Project.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

RStudio Team. 2022. *RStudio: Integrated Development Environment for r.* Boston, MA: RStudio, PBC. http://www.rstudio.com/.