Diamonds Dataset Analysis

Immanuel David

Introduction

The diamonds dataset, available in R through the 'ggplot2' package, contains data on over 50,000 diamonds, including features such as carat, cut, color, clarity, and price. This dataset is widely used to understand the characteristics that influence diamond pricing.

Objective

- Identify key factors influencing diamond prices.
- Visualize the relationships between these features.
- Uncover any trends or patterns that could inform pricing strategies

Dataset Overview

This following code will load the 'diamonds' dataset through 'ggplot2' package.

```
library(ggplot2)
data(diamonds)
summary(diamonds)
```

```
##      carat               cut        color        clarity          depth
##  Min.   :0.2000   Fair      : 1610   D: 6775   SI1    :13065   Min.   :43.00
##  1st Qu.:0.4000   Good      : 4906   E: 9797   VS2    :12258   1st Qu.:61.00
##  Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194   Median :61.80
##  Mean   :0.7979   Premium   :13791   G:11292   VS1    : 8171   Mean   :61.75
##  3rd Qu.:1.0400   Ideal     :21551   H: 8304   VVS2   : 5066   3rd Qu.:62.50
##  Max.   :5.0100                     I: 5422   VVS1   : 3655   Max.   :79.00
##                                     J: 2808   (Other): 2531
##      table           price             x                y
##  Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720
##  Median :57.00   Median : 2401   Median : 5.700   Median : 5.710
##  Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735
##  3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540
##  Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900
##
##        z
##  Min.   : 0.000
##  1st Qu.: 2.910
##  Median : 3.530
```

```
##  Mean   : 3.539
##  3rd Qu.: 4.040
##  Max.   :31.800
##
```

Data Exploration

Check the basic structure of the dataset.

```
str(diamonds)
```

```
## tibble [53,940 x 10] (S3: tbl_df/tbl/data.frame)
##  $ carat  : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
##  $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
##  $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
##  $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
##  $ depth  : num [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
##  $ table  : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...
##  $ price  : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...
##  $ x      : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
##  $ y      : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
##  $ z      : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

A quick summary for the dataset.

```
summary(diamonds)
```

```
##      carat               cut        color        clarity          depth
##  Min.   :0.2000   Fair      : 1610   D: 6775   SI1    :13065   Min.   :43.00
##  1st Qu.:0.4000   Good      : 4906   E: 9797   VS2    :12258   1st Qu.:61.00
##  Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194   Median :61.80
##  Mean   :0.7979   Premium   :13791   G:11292   VS1    : 8171   Mean   :61.75
##  3rd Qu.:1.0400   Ideal     :21551   H: 8304   VVS2   : 5066   3rd Qu.:62.50
##  Max.   :5.0100                     I: 5422   VVS1   : 3655   Max.   :79.00
##                                     J: 2808   (Other): 2531
##      table           price             x                y
##  Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720
##  Median :57.00   Median : 2401   Median : 5.700   Median : 5.710
##  Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735
##  3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540
##  Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900
##
##        z
##  Min.   : 0.000
##  1st Qu.: 2.910
##  Median : 3.530
##  Mean   : 3.539
##  3rd Qu.: 4.040
##  Max.   :31.800
##
```

Data Cleaning

Now let's check for any missing value.

```
colSums(is.na(diamonds))
```

```
##   carat     cut   color clarity   depth   table   price       x       y       z
##       0       0       0       0       0       0       0       0       0       0
```

As there's no missing values, let's check for any duplicate entries.

```
any(duplicated(diamonds))
```

```
## [1] TRUE
```

There seems to be some duplicate data. So let's filters them out.

```
diamonds <- diamonds[!duplicated(diamonds), ]
```

Now let's check for any incorrect or inconsistent values

```
unique(diamonds$cut)
```

```
## [1] Ideal     Premium   Good      Very Good Fair
## Levels: Fair < Good < Very Good < Premium < Ideal
```

```
unique(diamonds$color)
```

```
## [1] E I J H F G D
## Levels: D < E < F < G < H < I < J
```

```
unique(diamonds$clarity)
```

```
## [1] SI2  SI1  VS1  VS2  VVS2 VVS1 I1   IF
## Levels: I1 < SI2 < SI1 < VS2 < VS1 < VVS2 < VVS1 < IF
```

```
unique(diamonds$carat)
```

```
##   [1] 0.23 0.21 0.29 0.31 0.24 0.26 0.22 0.30 0.20 0.32 0.33 0.25 0.35 0.42 0.28
##  [16] 0.38 0.70 0.86 0.71 0.78 0.96 0.73 0.80 0.75 0.74 0.81 0.59 0.90 0.91 0.61
##  [31] 0.77 0.63 0.76 0.64 0.72 0.79 0.58 1.17 0.60 0.83 0.54 0.98 0.52 1.01 0.53
##  [46] 0.84 0.51 1.05 0.55 0.87 1.00 0.57 0.82 1.04 0.93 1.20 0.99 0.34 0.43 0.36
##  [61] 0.95 0.89 1.02 0.97 0.56 0.85 0.92 1.27 0.66 1.12 0.68 1.03 0.62 1.22 1.08
##  [76] 0.88 0.50 1.19 0.39 0.65 1.24 1.50 0.27 0.41 1.13 1.06 0.69 0.40 1.14 0.94
##  [91] 1.29 1.52 1.16 1.21 1.23 1.09 0.67 1.11 1.10 1.18 1.15 1.25 1.07 1.28 1.51
## [106] 0.37 1.31 1.26 1.39 1.44 1.35 1.30 1.32 1.41 1.36 1.45 1.34 1.58 1.54 1.38
## [121] 1.33 1.74 1.64 1.47 1.40 1.55 1.95 2.00 1.37 1.83 1.62 1.57 1.69 2.06 1.72
## [136] 1.66 2.14 1.49 1.46 2.15 1.96 2.22 1.70 1.53 1.85 2.01 2.27 1.68 1.56 1.81
## [151] 1.65 1.82 2.03 1.73 1.59 1.42 1.43 2.08 1.48 1.60 2.49 1.71 2.02 2.07 3.00
```

```
## [166] 2.21 2.10 1.91 2.25 2.17 2.32 2.72 1.61 2.23 2.11 2.05 1.63 2.30 2.31 1.75
## [181] 2.04 2.12 1.77 2.50 1.80 1.67 1.84 2.20 3.01 1.88 2.33 2.68 2.34 1.90 2.16
## [196] 2.74 1.78 1.76 2.28 1.79 1.94 2.43 1.86 3.11 1.87 2.09 1.89 2.52 2.19 2.18
## [211] 2.77 2.63 3.05 2.46 3.02 2.38 2.24 2.26 2.36 1.99 2.29 3.65 2.45 2.40 2.54
## [226] 3.24 2.13 2.58 3.22 3.50 2.48 1.98 2.44 2.75 1.93 2.41 2.61 2.35 2.51 2.70
## [241] 2.55 1.97 2.53 2.37 2.47 2.80 4.01 2.56 3.04 1.92 2.39 3.40 4.00 3.67 2.42
## [256] 2.66 2.65 2.59 2.60 2.57 2.71 4.13 2.64 5.01 4.50 2.67 3.51 0.44 0.45 0.47
## [271] 0.46 0.48 0.49
```

Data Analysis

Now let's identify which factor affects the price of a diamond.

```
cor(diamonds[, c("carat", "depth", "table", "price")])
```

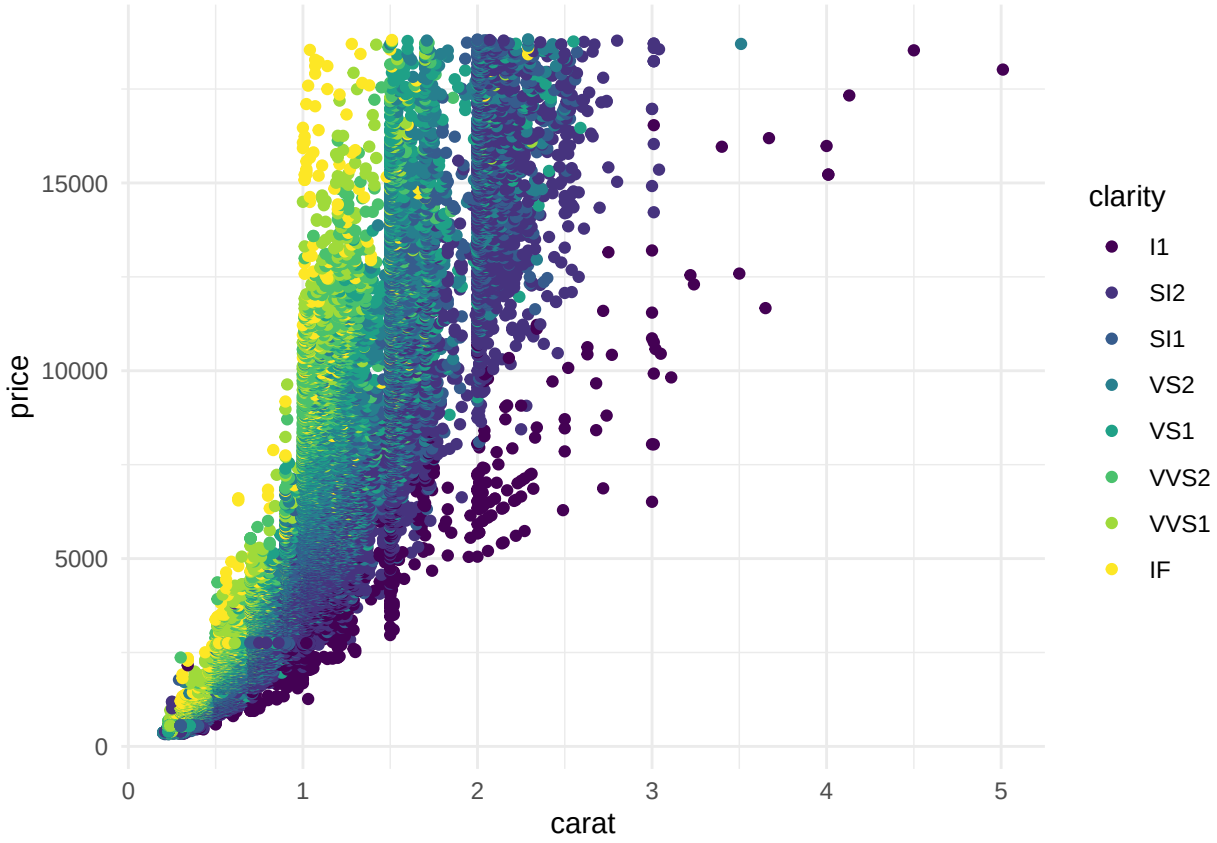
```
##            carat       depth      table       price
## carat 1.00000000  0.02786089  0.1810911  0.92154832
## depth 0.02786089  1.00000000 -0.2976691 -0.01104752
## table 0.18109111 -0.29766912  1.0000000  0.12656609
## price 0.92154832 -0.01104752  0.1265661  1.00000000
```

As this reveals that *carat* and *price* has a strong correlation (~0.92), *depth* and *price* are nearly uncorrelated (~-0.01) and *table* has a small positive correlation with *price* (~0.12). Now let's visualize the findings to see things more clearly!

Data Visualization

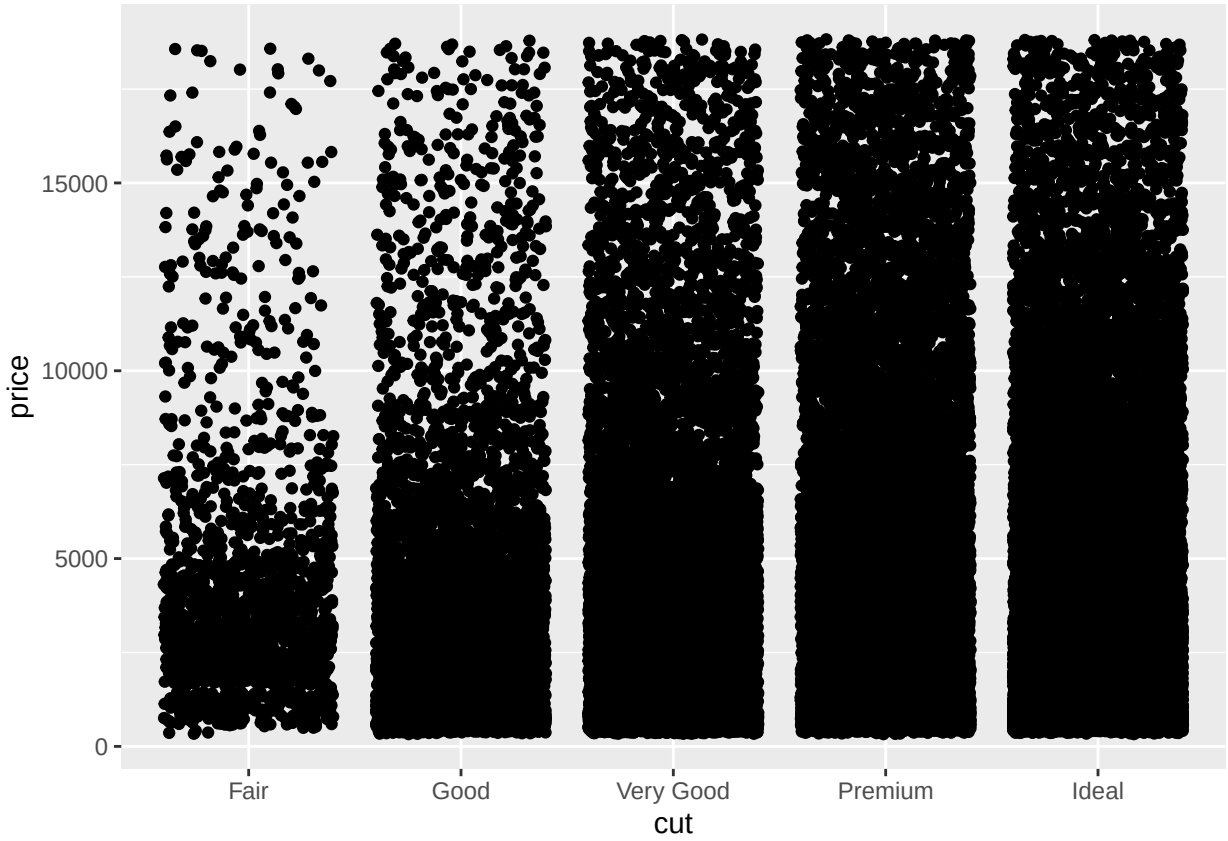
Let's compare price by carat.

```
library(ggplot2)
ggplot(diamonds, aes(x=carat, y=price, color=clarity))+
  geom_point()+
  theme_minimal()
```

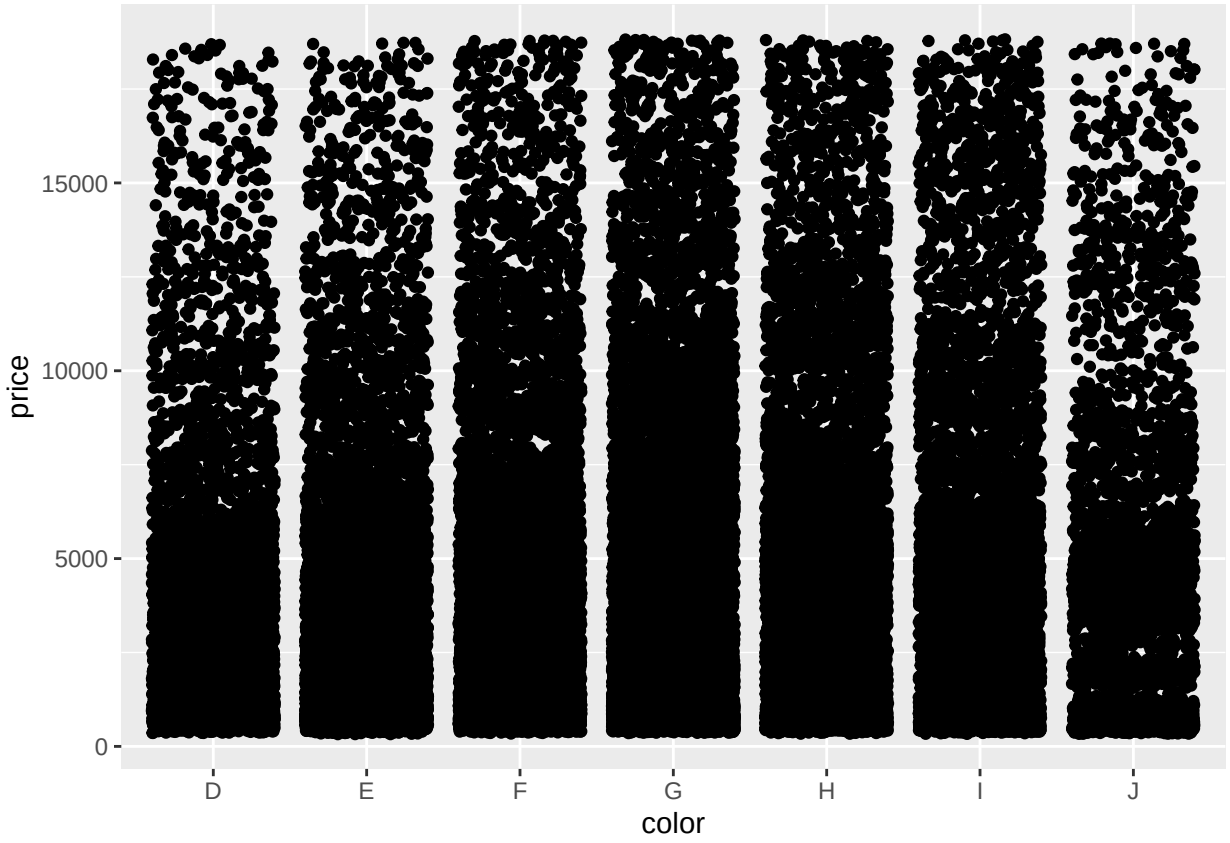


This helps us visualize how carat influence the pricing of diamonds. Let's also visualize cut, color and clarity with price.

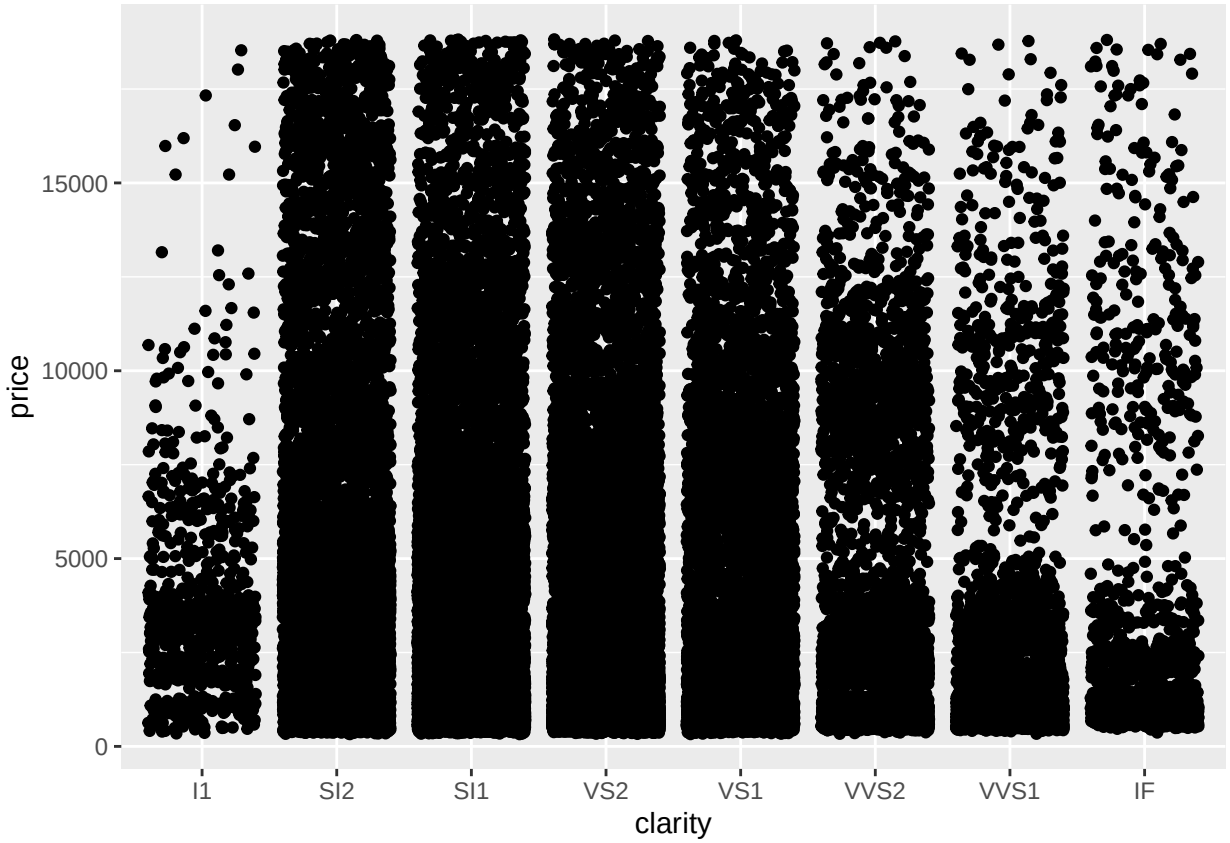
```
ggplot(diamonds, aes(x=cut, y=price))+
  geom_jitter()
```




```
ggplot(diamonds, aes(x=color, y=price))+
  geom_jitter()
```




```
ggplot(diamonds, aes(x=clarity, y=price))+
  geom_jitter()
```



This clearly shows the value of cut, color and clarity doesn't influence the pricing of diamonds.

Conclusion

Based on the analysis, carat emerges as the significant factor that influences the pricing of a diamond. Understanding the weight of carat in pricing can be valuable for both customers and sellers in making informed decisions.