Diamonds Report with Data Visualization

Tanyamon Sirikanjanakul

2023-03-03

Table of Contents

[Sprint 05 - Data Visualization Homework 1](#_Toc130467455)

[1. Introduction 2](#_Toc130467456)

[2. Loading Libraries and Data 2](#_Toc130467457)

[3. Understanding the dataset 3](#_Toc130467458)

[4. Data Visualization 5](#_Toc130467459)

## Sprint 05 - Data Visualization Homework

: Explore dataframe diamonds and create 5 visualizations



Diamond\_ring

The dataset diamonds is a built-in dataset in ggplot2 library which containing the prices and other attributes of almost 54,000 round cut diamonds.

The variable in diamonds*(53,940 rows x 10 columns)* are as follows:

* **carat** : weight of the diamond (0.2–5.01)
* **cut** : quality of the cut (Fair, Good, Very Good, Premium, Ideal)
* **color** : diamond color, from D (best) to J (worst)
* **clarity** : a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
* **price** : price in US dollars ($326–$18,823)
* **x** : length in mm (0–10.74)
* **y** : width in mm (0–58.9)
* **z** : depth in mm (0–31.8)
* **depth** : total depth percentage = z / mean(x, y) = 2 \* z / (x + y) (43–79)
* **table** : width of top of diamond relative to widest point (43–95)

### 1. Introduction

**Objective** Let’s say I’m looking for an engagement ring (for myself). I have done some research in these diamond attributes and come across some questions:

* What is the distribution of price in round cut diamonds ?
* I heard that in diamonds, table should be fairly large. Therefore, I want to compare the tablevalues between cut grade
* As diamond is grading by its clarity, let’s explore this variable by comparing each clarity group
* Color and carat also play a major role in diamond price. What is the correlation between these two variables with respect to clarity?
* For buying engagement ring on a suitable budget, What is the average price of diamonds by color and clarity?

### 2. Loading Libraries and Data

# Import Library  
library(tidyverse)  
library(dplyr)  
library(patchwork)  
library(scales)

# loading dataset  
data(diamonds)

### 3. Understanding the dataset

Review dataset

head(diamonds)

## # A tibble: 6 × 10  
## carat cut color clarity depth table price x y z  
## <dbl> <ord> <ord> <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>  
## 1 0.23 Ideal E SI2 61.5 55 326 3.95 3.98 2.43  
## 2 0.21 Premium E SI1 59.8 61 326 3.89 3.84 2.31  
## 3 0.23 Good E VS1 56.9 65 327 4.05 4.07 2.31  
## 4 0.29 Premium I VS2 62.4 58 334 4.2 4.23 2.63  
## 5 0.31 Good J SI2 63.3 58 335 4.34 4.35 2.75  
## 6 0.24 Very Good J VVS2 62.8 57 336 3.94 3.96 2.48

Review data structure

glimpse(diamonds)

## Rows: 53,940  
## Columns: 10  
## $ carat <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, 0.23, 0.…  
## $ cut <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very Good, Ver…  
## $ color <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, J, J, I,…  
## $ clarity <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, SI1, VS1, …  
## $ depth <dbl> 61.5, 59.8, 56.9, 62.4, 63.3, 62.8, 62.3, 61.9, 65.1, 59.4, 64…  
## $ table <dbl> 55, 61, 65, 58, 58, 57, 57, 55, 61, 61, 55, 56, 61, 54, 62, 58…  
## $ price <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339, 340, 34…  
## $ x <dbl> 3.95, 3.89, 4.05, 4.20, 4.34, 3.94, 3.95, 4.07, 3.87, 4.00, 4.…  
## $ y <dbl> 3.98, 3.84, 4.07, 4.23, 4.35, 3.96, 3.98, 4.11, 3.78, 4.05, 4.…  
## $ z <dbl> 2.43, 2.31, 2.31, 2.63, 2.75, 2.48, 2.47, 2.53, 2.49, 2.39, 2.…

#### The Diamonds Grading System

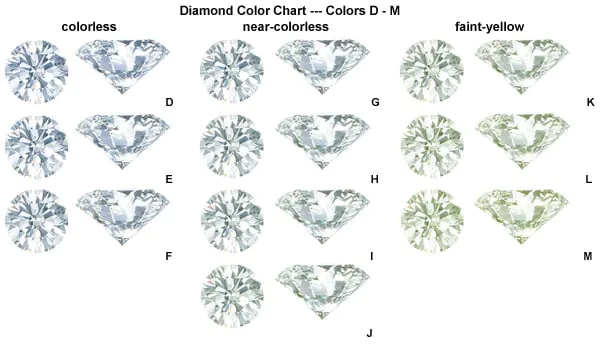
To gain a better understanding in this dataset, here’s some study that I did on diamond grading system and its value (in this case – *price*). The grading system is based on four attributes: carat, color, clarity, and cut, known as **The Four Cs of Diamonds.**

**Carat**

Diamond weight is measured in carats. One carat (or 1 ct, when abbreviated) equals 0.2 gram. The carat weight is sometimes directly related to price and the prices go up as diamond carat weight increase.

**Color**

Diamonds are graded on a color scale from D-Z. According to [GIA color grading scale](https://www.gemsociety.org/article/diamond-color/), a diamond graded D-F is considered to be “colorless” and diamond graded G-J are considered to be “near colorless”.



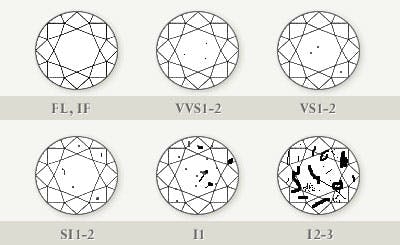
GIA color chart

**Clarity**

[Diamonds clarity](https://www.gemsociety.org/article/diamond-clarity/) are graded on a scale from “Internal Flawless (IF)” to “Included (I3)”. Most diamonds have inclusions (defined as anything that will interfere with the free passage of light) but the less inclusions it has, the more valuable it is.

Clarity grading are as follows:

* IF: Internally Flawless
* VVS1, VVS2: Very, Very Slightly Included
* VS1, VS2: Very Slightly Included
* SI1,SI2: Slightly Included
* I1: Included



Diamond Clarity Chart

**Cut**

[Diamond cut quality](https://www.gemsociety.org/article/diamond-cuts/) is what gives a diamond its sparkle. Therefore, Cut is considered to be the most important determining the diamond’s overall beauty. The overall quality of the cut is determined by the anatomy of diamond such as *table, total depth percentage, etc.* as well as the quality of the polish.

For round brilliant diamonds, cut grading are as follows:

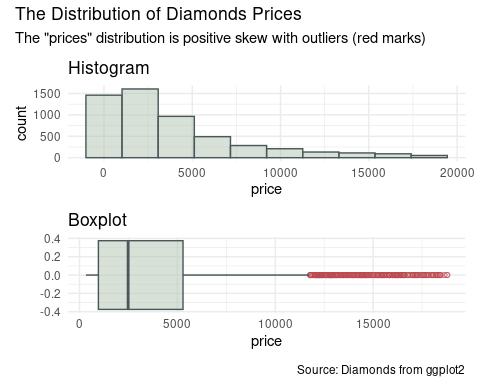
* Ideal or Excellence
* Premium
* Very Good
* Good
* Fair
* Poor (does not include in this dataset)

### 4. Data Visualization

#### Chart 01 - Diamonds prices distribution

To simply explore the distribution of diamonds price, the histogram along with boxplot is plotted as shown below:

# Using sampling method  
set.seed(42)   
p1 <- ggplot(diamonds %>% sample\_n(5400),  
 aes(price)) +  
 geom\_histogram(bins=10, fill = "#b0c4b1", alpha = 0.5, color = "#4a5759" ) +  
 theme\_minimal() +  
 ggtitle("Histogram")  
  
p2 <- ggplot(diamonds %>% sample\_n(5400),  
 aes(price)) +  
 geom\_boxplot(fill = "#b0c4b1", alpha = 0.5, color = "#4a5759",  
 outlier.colour = "#bc4b51", outlier.shape = 1) +  
 theme\_minimal() +  
 ggtitle("Boxplot")  
  
patchwork <- p1 / p2  
patchwork +   
 plot\_annotation(  
 title = "The Distribution of Diamonds Prices",  
 subtitle = 'The "prices" distribution is positive skew with outliers (red marks)',  
 caption = 'Source: Diamonds from ggplot2'  
)

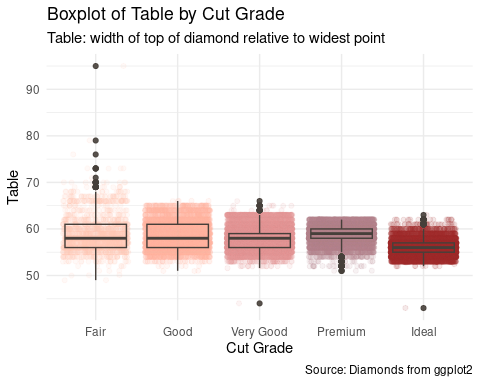


* The diamonds “prices” distribution is positive skew, and there are outliers (red marks) as shown in the boxplot.

#### Chart 02 - Boxplot of Table by Cut Grade

As table is one of the attribute that affect the cut quality. I want to compare the distribution of table between cut grade side-by-side. The boxplot of table is created as shown below:

diamonds %>%  
 ggplot(aes( cut ,table,fill=cut, col=cut)) +  
 #geom\_point(position = "jitter") +  
 geom\_jitter(alpha=0.1, show.legend = FALSE) +  
 geom\_boxplot(alpha=0.9, fill= NA, color = "#463f3a", show.legend = FALSE)+  
 theme\_minimal() +  
 scale\_color\_manual(values = c(  
 "#ffcdb2", "#ffb4a2", "#e5989b", "#b5838d", "#9e2a2b")) +  
 labs(title = "Boxplot of Table by Cut Grade",  
 subtitle = "Table: width of top of diamond relative to widest point",  
 x = "Cut Grade",   
 y = "Table",  
 caption = "Source: Diamonds from ggplot2")



* From this chart, it can be seen that only diamonds with Good-cut has no outlier, and the range of table is much more narrower in Ideal-cut compared to the Fair-cut (worst).
* The median of Ideal-cut table is also lower than other cut grade. Therefore, there is a certain suitable range for diamond table.

#### Chart 03 - Percentage of diamonds by Clarity Segments

As in this dataset, clarity is scaling from IF to I1, and the flaw is usually noticeable in SI1 to I1. Therefore, instead of showing all scaling percentage in pie chart, the clarity is grouped into 4 segments as follows:

* Internal Flawless (IF)
* Very, Very Slightly Included (VVS1-2)
* Very Slightly Included (VS1-2)
* All other (SI1 to I1)

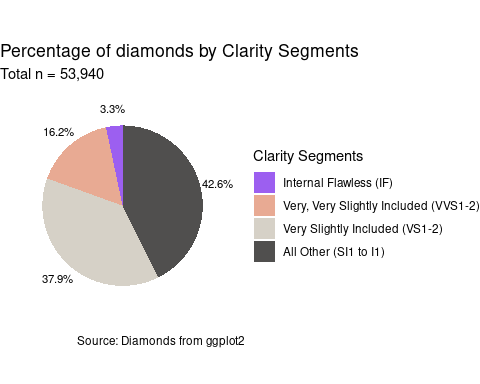
In this case, I can focus more on comparing the clean and unnoticeable grade.

## create new segment level  
# data transformation  
dia\_cl <- diamonds %>%  
 mutate(segment="")  
  
dia\_cl$segment[dia\_cl$clarity == "IF"] <- "Internal Flawless (IF)"  
dia\_cl$segment[dia\_cl$clarity %in% c("VVS1","VVS2")] <- "Very, Very Slightly Included (VVS1-2)"  
dia\_cl$segment[dia\_cl$clarity %in% c("VS1","VS2")] <- "Very Slightly Included (VS1-2)"  
dia\_cl$segment[dia\_cl$clarity %in% c("SI1","SI2", "I1")] <- "All Other (SI1 to I1)"  
  
dia\_cl$segment <- factor(dia\_cl$segment, levels = c("Internal Flawless (IF)", "Very, Very Slightly Included (VVS1-2)","Very Slightly Included (VS1-2)","All Other (SI1 to I1)"))  
  
dia\_cl <- dia\_cl %>%  
 group\_by(segment) %>%   
 count() %>%  
 ungroup() %>%  
 mutate(pc =`n`/sum(`n`)) %>%  
 arrange(pc ) %>%  
 mutate(labels = scales::percent(pc, accuracy = 0.1))  
dia\_cl

## # A tibble: 4 × 4  
## segment n pc labels  
## <fct> <int> <dbl> <chr>   
## 1 Internal Flawless (IF) 1790 0.0332 3.3%   
## 2 Very, Very Slightly Included (VVS1-2) 8721 0.162 16.2%   
## 3 Very Slightly Included (VS1-2) 20429 0.379 37.9%   
## 4 All Other (SI1 to I1) 23000 0.426 42.6%

Create Pie Chart

# build pie chart  
ggplot(dia\_cl, aes(x="", y = pc, fill= segment)) +  
 geom\_col(alpha=0.8) +  
 geom\_text(aes(x = 1.65, label = labels),  
 position = position\_stack(vjust = 0.5),  
 size = 3)+  
 coord\_polar(theta = "y") +  
 theme\_void() + #empty theme   
 scale\_fill\_manual(values=c("#8338ec","#e29578","#ccc5b9","#252422")) +  
 guides(fill = guide\_legend(title = "Clarity Segments")) +  
 labs(title = "Percentage of diamonds by Clarity Segments",  
 subtitle = "Total n = 53,940",  
 caption = "Source: Diamonds from ggplot2")



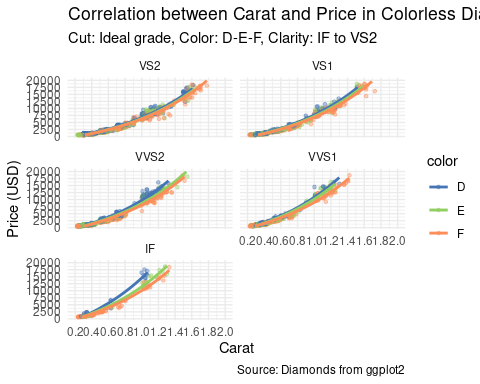
* It can be seen that the smallest fraction in this case is IF or Internal Flawless diamond with 3.3%.

#### Chart 04 - Distribution of Price by Color

As diamond prices depend on various variables. I want to explore the correlation between carat and price in colorless diamonds (D, E, and F), focusing solely on clean & unnoticeable flaw clarity grades (IF to VS2).

Scatter plots with trendlines are used to illustrate these correlations.

## Data Transformation  
colorless\_dia <- diamonds %>%   
 filter(color %in% c("D","E","F") & cut == "Ideal"  
 & clarity %in% c("IF","VVS1","VVS2","VS1","VS2"))  
  
# summary(colorless\_dia)  
  
  
## Scatterplot  
ggplot(colorless\_dia, aes(x=carat, y=price, col = color) ) +  
 geom\_point(size=1, position = "jitter", alpha=0.4) +  
 geom\_smooth(method="lm", aes(color=color),  
 formula = (y ~ exp(x)),  
 se = FALSE) +  
 scale\_x\_continuous(limits =c(0.20,2.00) , n.breaks = 10) +  
 scale\_y\_continuous(limits = c(500,20000), n.breaks = 10) +  
 theme\_minimal() +  
 scale\_color\_manual( values = c("#4575b4", "#91cf60", "#fc8d59")) +  
 facet\_wrap(~ clarity, ncol = 2) +  
 labs(title = "Correlation between Carat and Price in Colorless Diamonds",x = "Carat", y = "Price (USD)",  
 subtitle = "Cut: Ideal grade, Color: D-E-F, Clarity: IF to VS2 ",  
 caption = "Source: Diamonds from ggplot2")

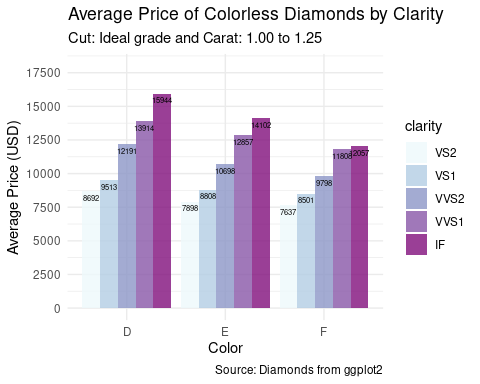


* According to the chart above, there’s a positive correlation between carat and price in all colorless diamonds
* In terms of clarity grade, the price jump between D&E-color was found in IF-grade as the price of D-color diamonds increase exponentially

#### Chart 05 - Average Price of Colorless Diamonds by Clarity

Assume the engagement ring that I’m looking for is 1.00-1.25 carat diamond with Ideal-cut. Before making decision, I want to know the average price of ideal cut diamond based on color and clarity.

colorless\_dia %>%   
 filter(carat >= 1 & carat <= 1.25) %>%  
 group\_by(color,clarity) %>%  
 summarise(avg\_price = round(mean(price))) %>%  
 ggplot( aes(x=color, y=avg\_price, fill=clarity)) +  
 geom\_bar(alpha = 0.8,stat="identity", position=position\_dodge())+  
 geom\_text(aes(label=avg\_price), vjust=1.6, color="black",  
 position = position\_dodge(0.9), size=2)+  
 scale\_fill\_brewer(palette="BuPu", direction = 1) +  
 theme\_minimal() +  
 scale\_y\_continuous(limits = c(0,18000), n.breaks = 8) +  
 labs(title = "Average Price of Colorless Diamonds by Clarity",  
 subtitle = "Cut: Ideal grade and Carat: 1.00 to 1.25",  
 x = "Color",   
 y = "Average Price (USD)",  
 caption = "Source: Diamonds from ggplot2")



* Refer to the chart above, it can be seen that the average price of diamond increase rapidly with the clarity and color grade
* This could be helpful information if I wish to trade off some aspects to lower the price.