Homework Data Viz Batch 10

Chanakarn Chuklin

2024-07-20

Create ggplot2 charts to explore the Diamonds dataset

* Load Library

```r
## install.packages("tidyverse")
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
print("Loaded library for data visualization")
```

```
## [1] "Loaded library for data visualization"
```

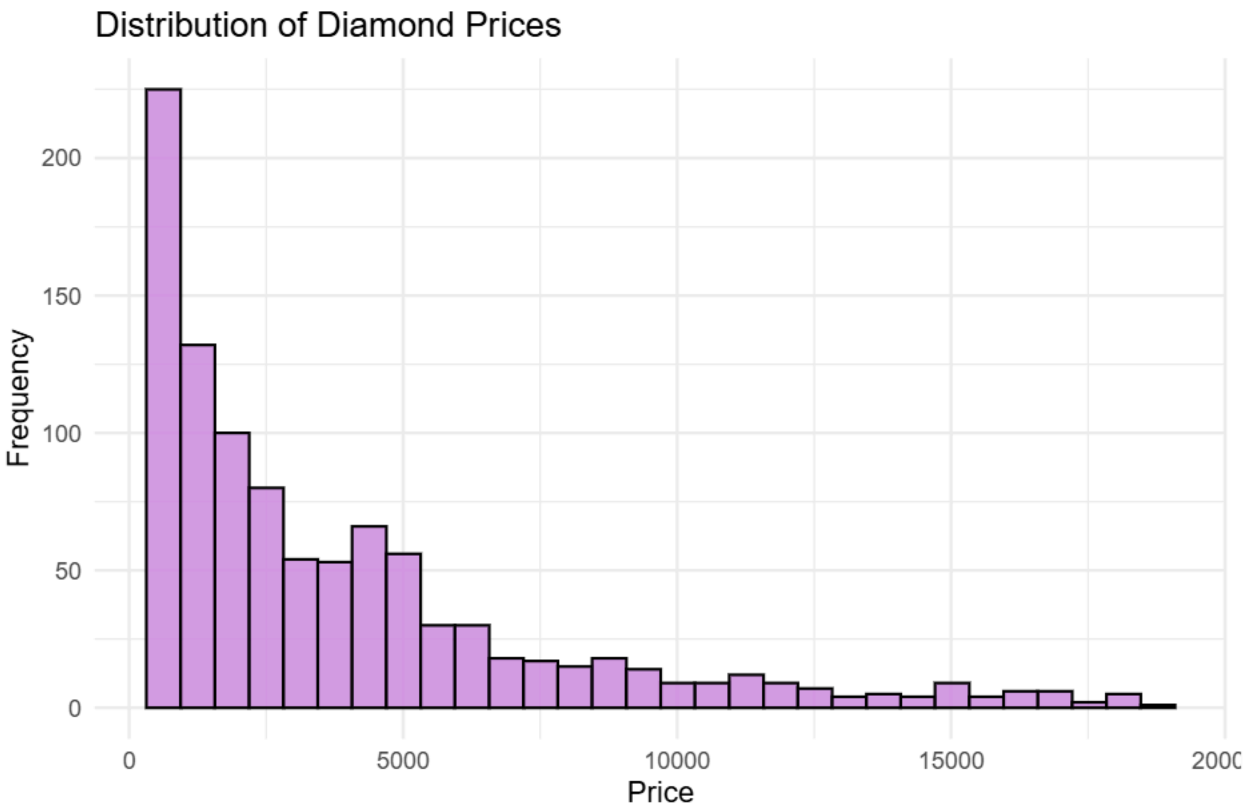
** Prepare Data

```r
set.seed(42)
small_df <- diamonds %>%
  sample_n(1000)
```

1. Histogram of Diamond Prices

```r
ggplot(small_df, aes(x = price)) +
  geom_histogram(fill = "#CC83E0", alpha = 0.8, color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Diamond Prices",
       caption = "Source: ggplot package",
       x = "Price",
       y = "Frequency")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Source: ggplot package

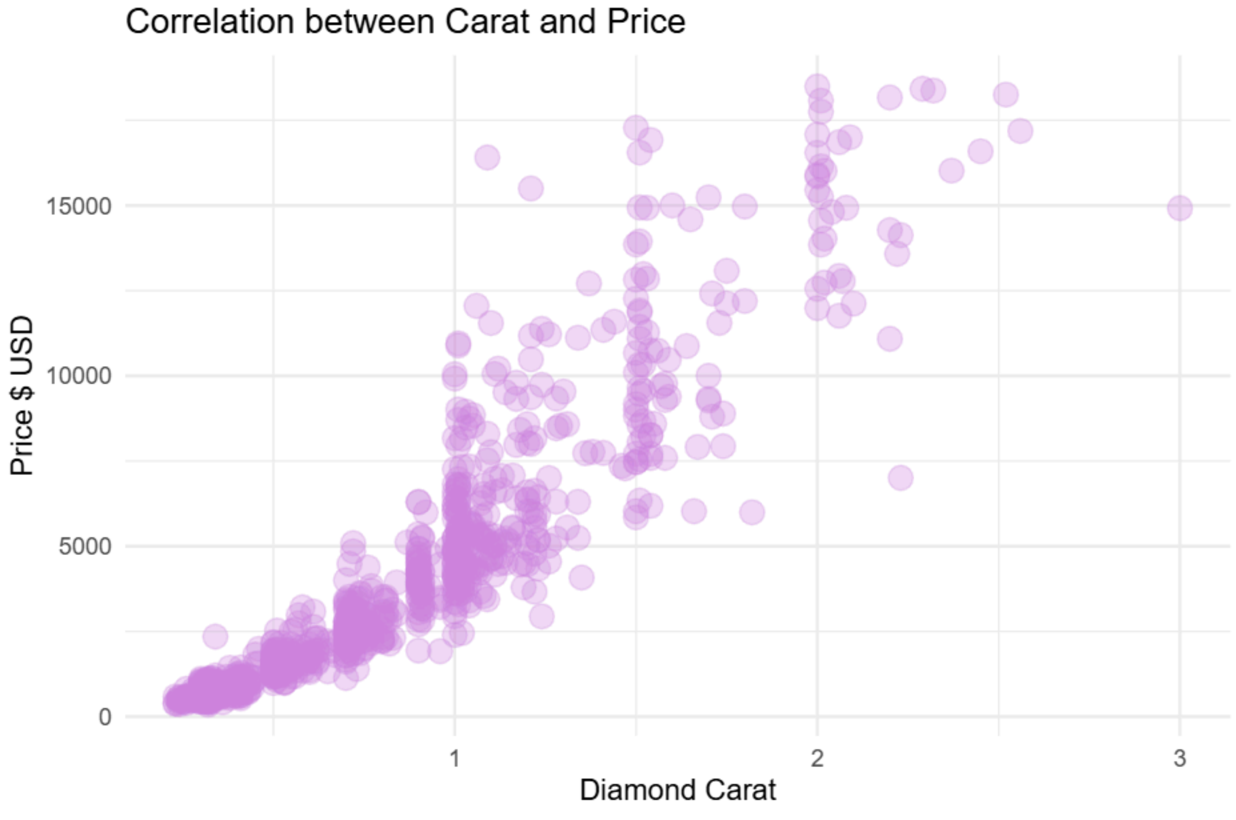
```r
# Calculate mean and median
summary_stats <- small_df %>%
  summarize(mean_price = mean(price),
            median_price = median(price))

# Approximate mode using the most frequent price
mode_price <- small_df %>%
  count(price) %>%
  arrange(desc(n)) %>%
  slice(1) %>%
  pull(price)

# Combine results into a single data frame
final_result <- summary_stats %>%
  tibble(mode_price = mode_price)
print(final_result)
```

```
## # A tibble: 1 x 3
##   mean_price median_price mode_price
##        <dbl>        <dbl>      <int>
## 1      3882.         2500        984
```

Analysis: Diamond prices exhibit a heavily right-skewed distribution, indicating a majority of lower-priced diamonds and a smaller proportion of very expensive ones. This skewness is typical of price distributions and suggests that a few high-value diamonds significantly influence the overall market value. The wide price range highlights the significant variation in diamond pricing, which is determined by factors like carat, cut, color, and clarity.

2. Scatter Plot of Carat vs. Price

```r
ggplot(small_df, aes(x = carat, y = price)) +
  geom_point(size = 4, col = "#CC83E0", alpha = 0.3) +
  theme_minimal() +
  labs(title = "Correlation between Carat and Price",
       caption = "Source: ggplot package",
       x = "Diamond Carat",
       y = "Price $ USD")
```

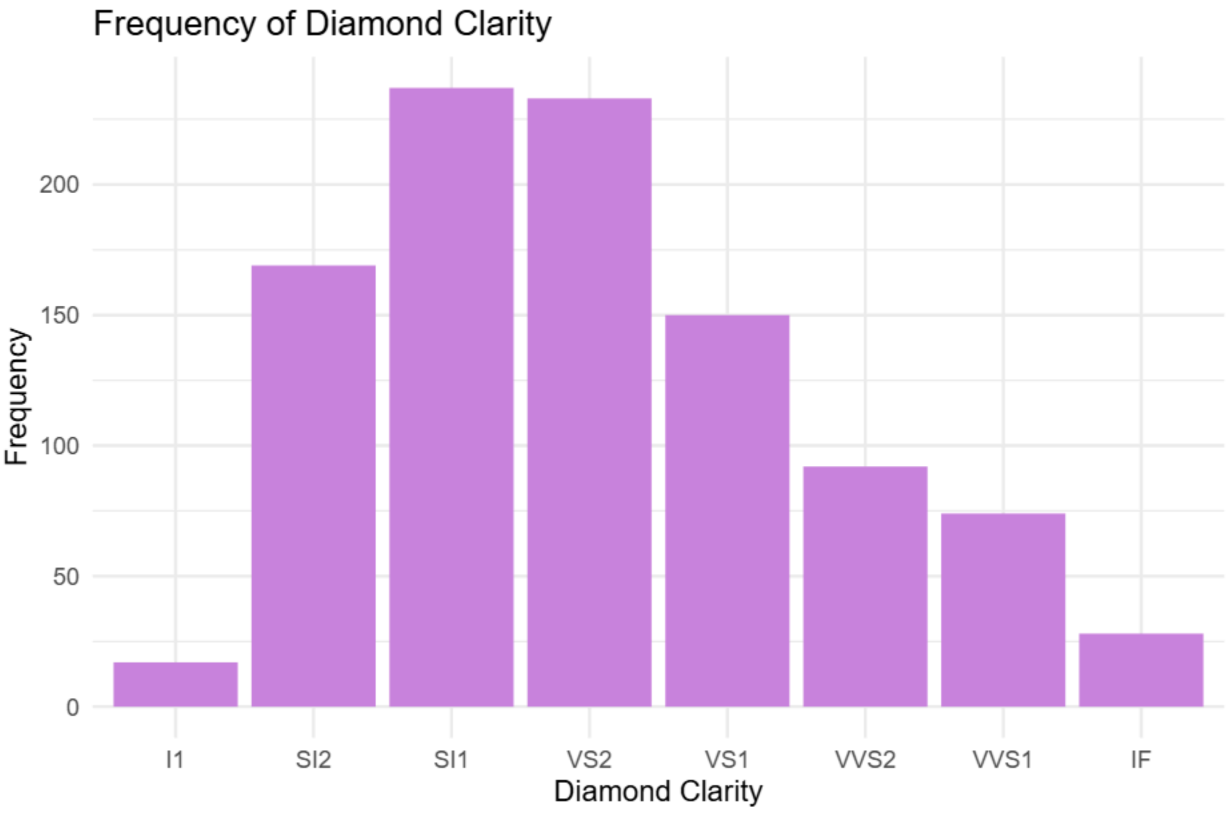


Source: ggplot package

Analysis: Diamond price exhibits a strong positive correlation with carat weight, meaning larger diamonds generally command higher prices. However, significant scatter in the data indicates that other factors such as cut, color, and clarity also influence pricing. The presence of outliers suggests that exceptional quality or rarity can significantly impact a diamond's value beyond its carat weight.

3. Bar Chart of Diamond Clarity Frequency

```r
ggplot(small_df, aes(x = clarity)) +
  geom_bar(fill = "#CC83E0") +
  theme_minimal() +
  labs(title = "Frequency of Diamond Clarity",
       caption = "Source: ggplot package",
       x = "Diamond Clarity",
       y = "Frequency")
```



Source: ggplot package

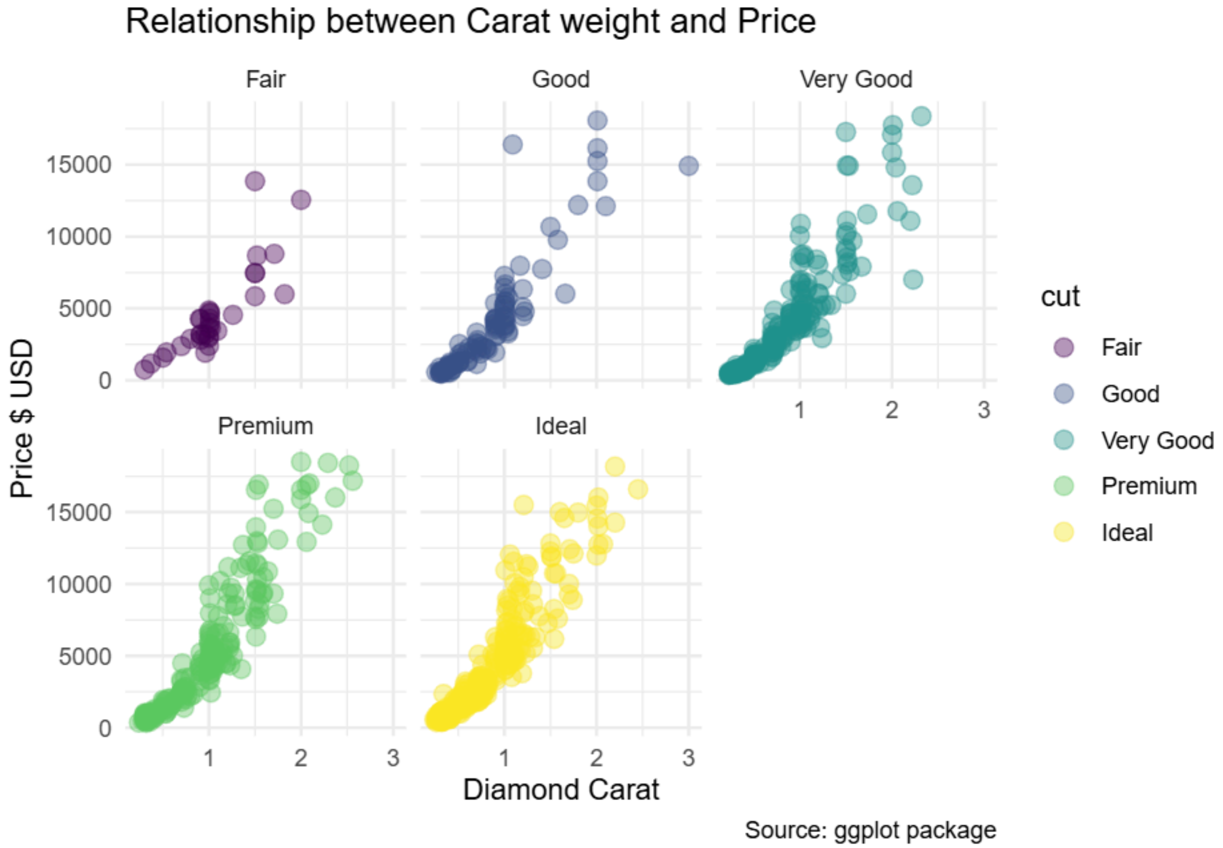
```
small_df %>%
  count(clarity) %>%
  arrange(desc(n))
```

```
## # A tibble: 8 x 2
##   clarity     n
##   <ord>   <int>
## 1 SI1       237
## 2 VS2       233
## 3 SI2       169
## 4 VS1       150
## 5 VVS2       92
## 6 VVS1       74
## 7 IF         28
## 8 I1         17
```

Analysis: Clarity grades SI1 and VS2 are the most prevalent within the dataset, likely reflecting higher consumer demand and more affordable price points. While higher clarity grades such as IF and VVS1 are typically associated with premium prices due to fewer inclusions, their lower frequency in the dataset may indicate a lower demand stemming from potentially prohibitive price points for the average consumer. Despite fewer inclusions, many consumers prioritize more accessible, mid-range clarity grades, which often exhibit comparable aesthetic qualities.

4. Scatter Plot of Carat vs. Price, Colored by Cut

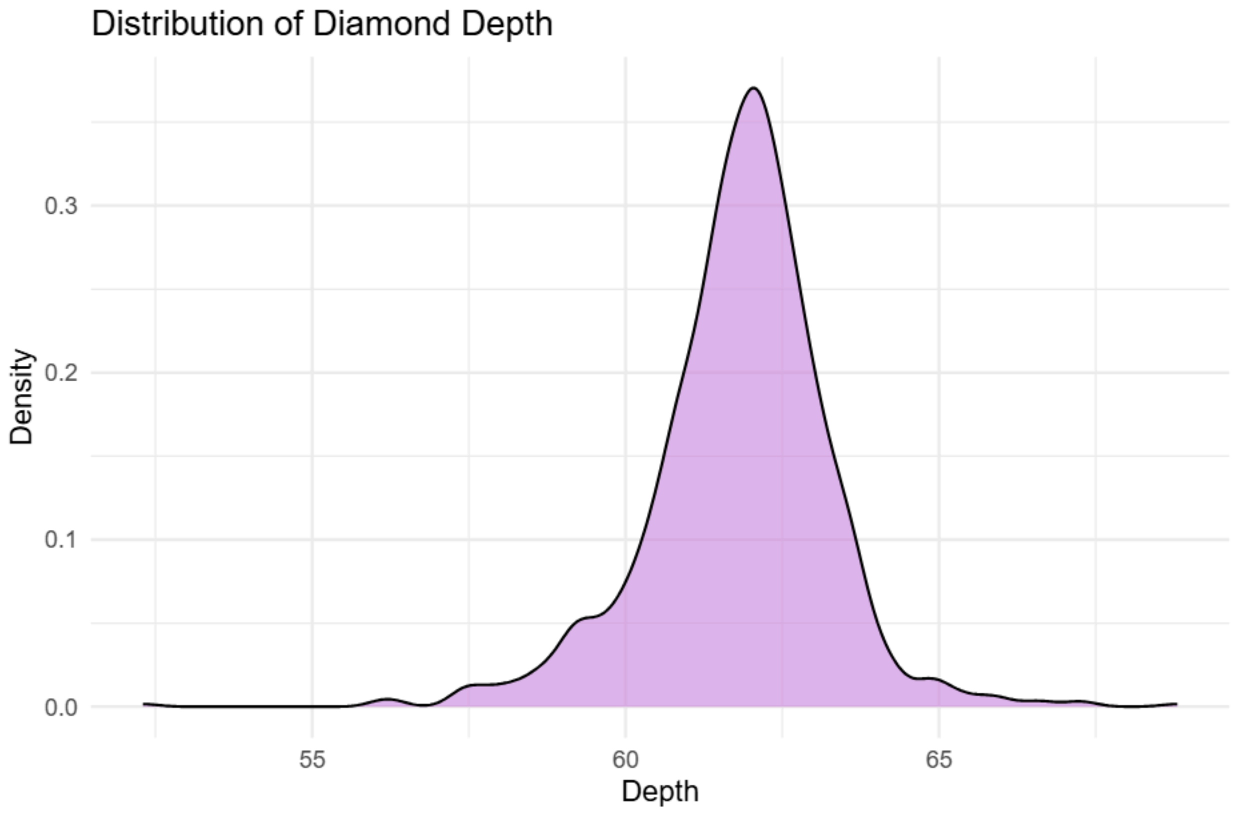
```
ggplot(small_df, aes(x = carat, y = price, color = cut)) +
  geom_point(size = 3, alpha = 0.4) +
  theme_minimal() +
  facet_wrap(~cut) +
   labs(title = "Relationship between Carat weight and Price",
       caption = "Source: ggplot package",
       x= "Diamond Carat",
       y = "Price $ USD")
```



Analysis: Diamond price exhibits a strong positive correlation with carat weight across all cut qualities. However, `Ideal and Premium` cuts consistently command higher prices, particularly for larger diamonds. While carat weight is a significant factor, `Good, Very Good, and Fair` cuts generally have lower prices, especially for larger stones. This highlights the interplay of various factors, including cut quality, color, clarity, and specific cutting characteristics, in determining a diamond's final price.

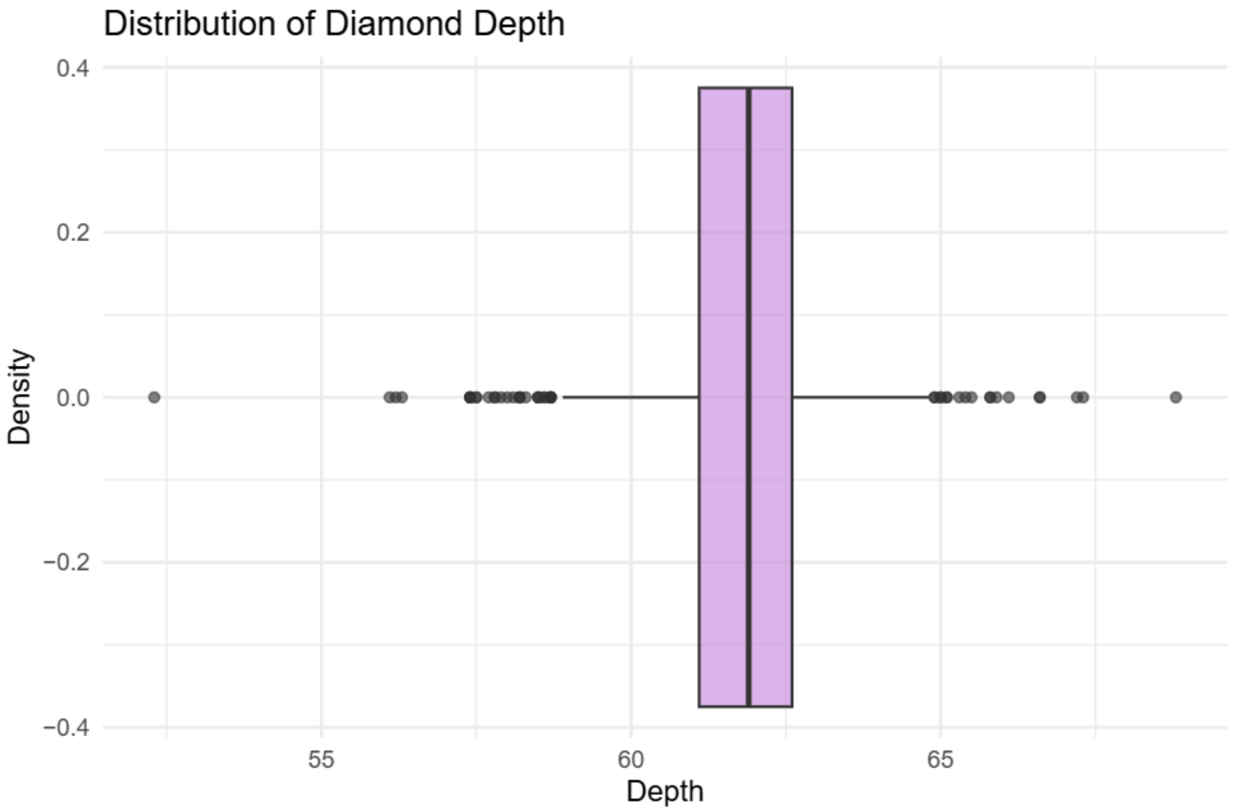
5. Density Plot of Diamond Depth

```
ggplot(small_df, aes(x = depth)) +
  geom_density(fill = "#CC83E0", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Distribution of Diamond Depth",
       caption = "Source: ggplot package",
       x = "Depth",
       y = "Density")
```



Source: ggplot package

```
ggplot(small_df, aes(x = depth)) +
  geom_boxplot(fill = "#CC83E0", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Distribution of Diamond Depth",
       caption = "Source: ggplot package",
       x = "Depth",
       y = "Density")
```



Source: ggplot package

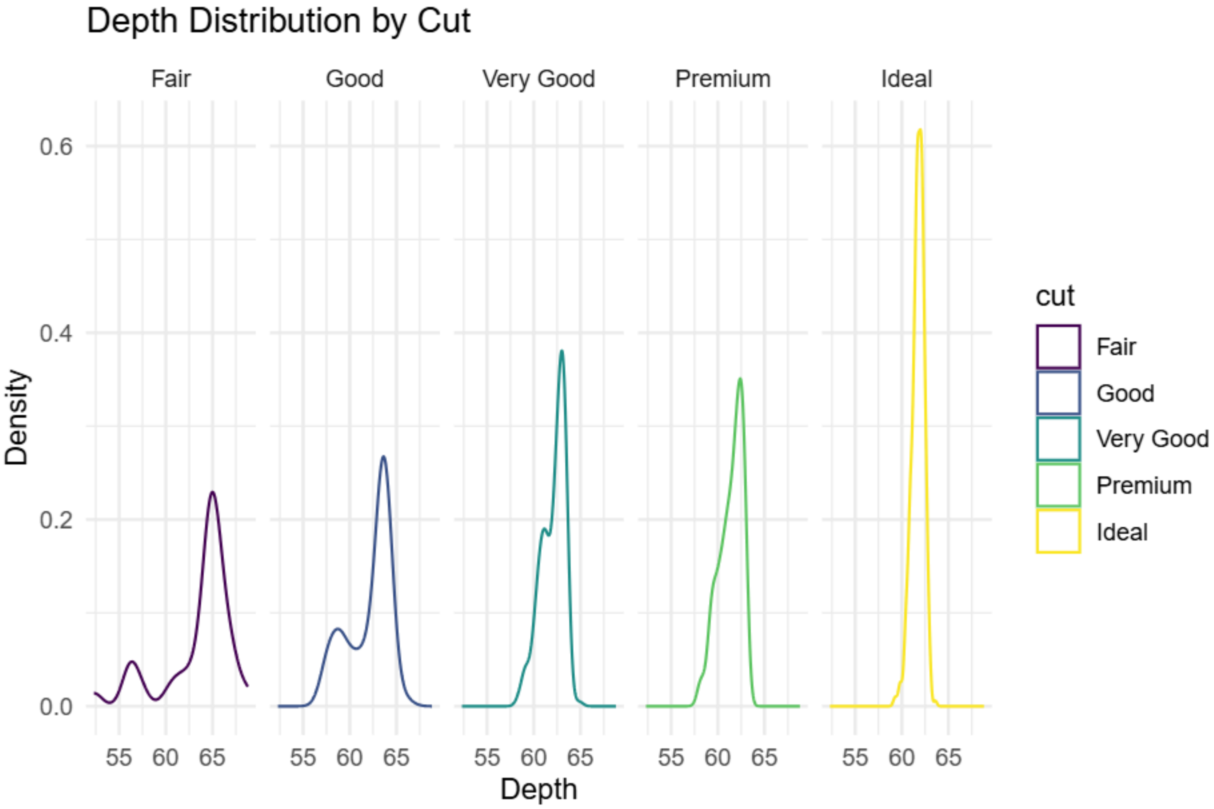
```
sum_depth <- small_df %>%
  summarize(
    min_depth = min(depth),
    q1_depth = quantile(depth, probs = 0.25),
    avg_depth = mean(depth),
    median_depth = quantile(depth, probs = 0.5),
    q3_depth = quantile(depth, probs = 0.75),
    max_depth = max(depth)
  )

print(sum_depth)
```

```
## # A tibble: 1 x 6
##   min_depth q1_depth avg_depth median_depth q3_depth max_depth
##       <dbl>    <dbl>     <dbl>        <dbl>    <dbl>     <dbl>
## 1      52.3     61.1      61.8         61.9     62.6      68.8
```

Analysis: Most diamonds in this dataset have a depth around 62%, which is considered the best for making them sparkle the most. This is likely because diamond cutters aim for this depth to maximize their brilliance and fire. However, there's some variation in the depths of these diamonds. This could be because of differences in their shapes, how they were cut, and the original shape of the rough diamond. Interestingly, there's a smaller group of diamonds with a depth around 55%, which might suggest they were cut differently or for a different purpose.

6. Depth Distribution by Cut

```r
ggplot(small_df, aes(x = depth, color = cut)) +
  geom_density(alpha = 0.8) +
  facet_grid(~cut) +
  theme_minimal() +
  labs(title = "Depth Distribution by Cut",
       caption = "Source: ggplot package",
       x = "Depth",
       y = "Density")
```



Source: ggplot package

Analysis: Most diamonds, regardless of cut quality, have a depth around 60-62%, indicating a common cutting practice. However, `Ideal and Premium` cuts show more consistent depths, suggesting greater attention to symmetry. In contrast, `Very Good, Good, and Fair` cuts exhibit a wider range of depths, potentially reflecting more flexibility in their cutting process. This flexibility might allow for adjustments to preserve weight or clarity, even if it slightly impacts brilliance.

Summary

This document explores various characteristics of `diamonds` using the `ggplot2` package and a sample of 1000 diamonds from the diamonds dataset.

1. Library and Data Preparation:

- We loaded the `tidyverse` library, which includes `ggplot2` for creating visualizations.
- A sample of 1,000 diamonds was selected from the full dataset using `sample_n`.

2. Price Distribution:

- Diamond prices are right-skewed, with many lower-priced diamonds and a few very expensive ones. Carat weight is a significant factor influencing price, but other factors like cut, color, and clarity also play a role.

3. Carat vs. Price:

- There is a positive correlation between carat weight and price. Larger diamonds tend to be more expensive. However, there is also scatter in the data, indicating the influence of other factors.

4. Diamond Clarity Frequency:

- `SI1` and `VS2` are the most frequent clarity grades, while `IF` and `I1` are less common. Higher clarity grades generally correspond to higher prices.

5. Carat vs. Price by Cut:

- The relationship between carat weight and price varies across cut qualities. `Ideal and Premium` cuts command higher prices, especially for larger diamonds.

6. Diamond Depth Distribution:

- Diamond depth shows a peak around 62%, with some variation. `Ideal and Premium` cuts have a narrower depth distribution compared to lower quality cuts.

Overall, this document provides a foundational understanding of the `diamond dataset's` characteristics using `ggplot2` visualizations. It offers valuable insights into the relationships between various diamond attributes and their impact on price and overall value.