Homework_Data_Viz

oOFourthOo

2024-07-21

Load library for data visualization

```
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   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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
```

```
library(tinytex)
```

Prepare Data

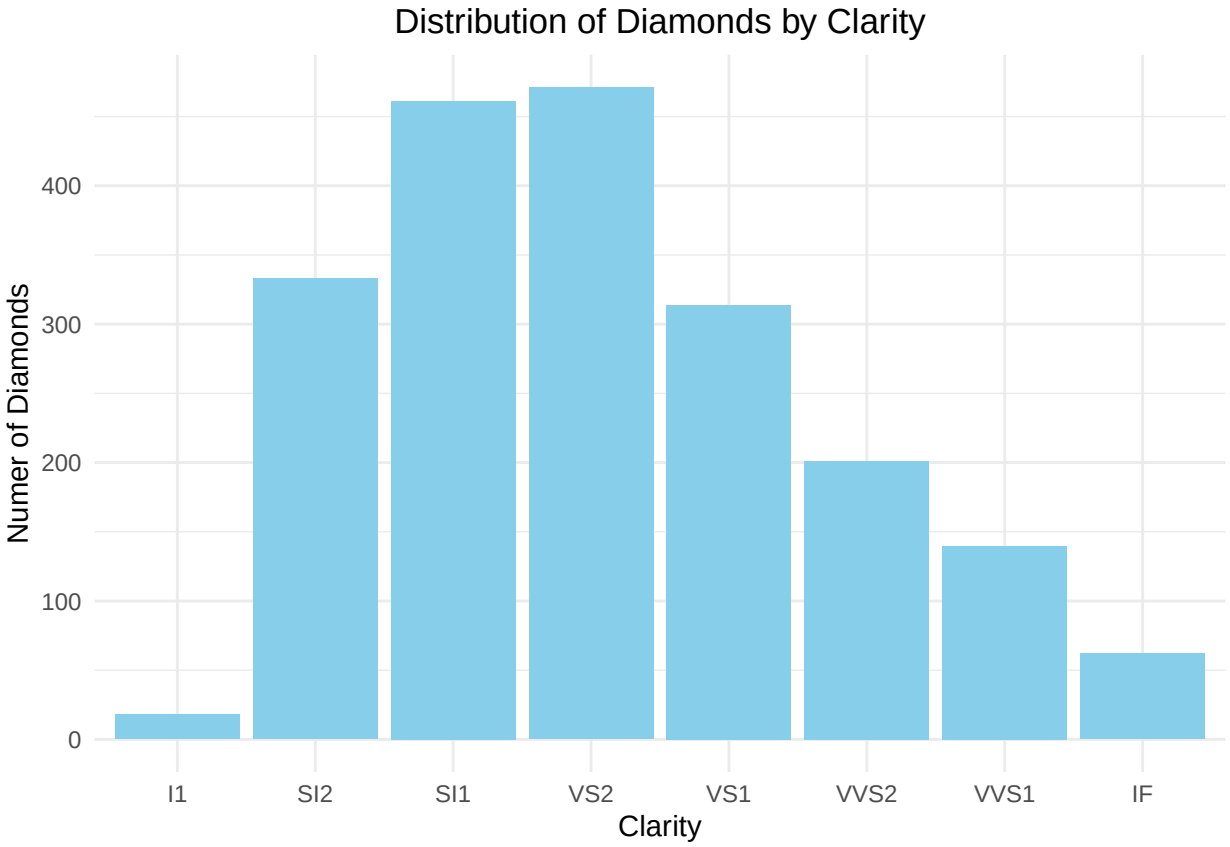
```
small_df <- diamonds %>%
  sample_n(2000)

head(small_df)
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.32 Ideal     D     SI1      62.1    54   715  4.4   4.45  2.75
## 2  1.18 Ideal     H     SI1      61.4    55  6013  6.84  6.78  4.18
## 3  0.8  Good      F     SI2      63.9    57  2189  5.86  5.84  3.74
## 4  2.01 Very Good I     SI2      61.8    58 13323  8.01  8.05  4.96
## 5  0.38 Ideal     F     VS2      61.2    57   889  4.67  4.64  2.85
## 6  0.32 Ideal     D     VS2      61.1    56   972  4.41  4.39  2.69
```

1.Barchart Diamonds Clarity

```
ggplot(small_df, aes(clarity)) +
  geom_bar(fill = "skyblue") +
  theme_minimal() +
  labs(title = "Distribution of Diamonds by Clarity",
       x = "Clarity",
       y = "Numer of Diamonds") +
  theme(plot.title = element_text(hjust = 0.5))
```



- From the bar chart, clarity levels SI1 and VS2 have the highest number of diamonds. This is probably the most easily found or most sought after level on the market.

2.Histogram Diamonds Price

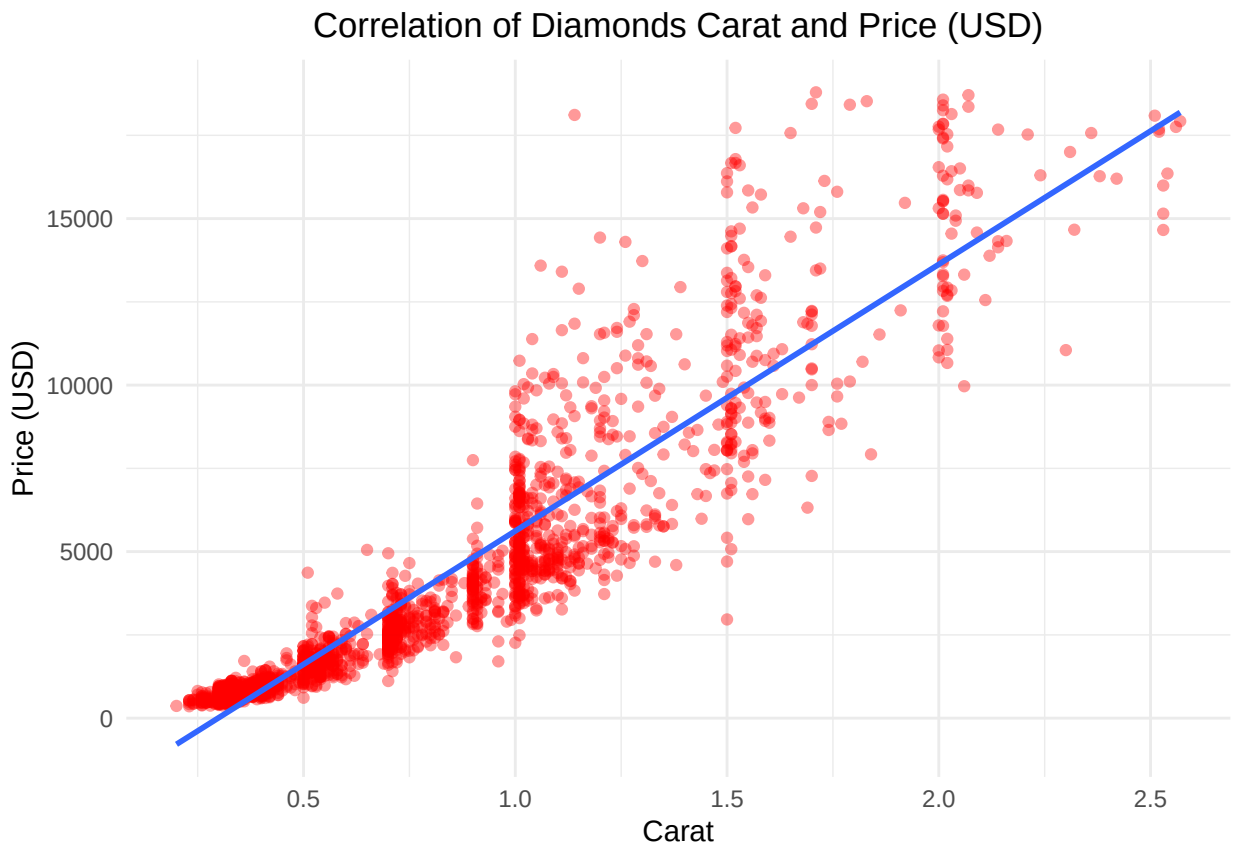
```
ggplot(small_df, aes(price)) +
  geom_histogram(fill = "salmon",bins = 20, color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Diamond Prices (USD)",
       x = "Price (USD)",
       y = "Numer of Diamonds") +
  theme(plot.title = element_text(hjust = 0.5))
```

- From the Histogram chart, it can be seen that the price of most diamonds is less than $2000.

3.Scatter Plot Carat and Price

```
ggplot(small_df, aes(carat, price)) +
  geom_point(color = "red", alpha = 0.4) +
    geom_smooth(method = "lm", se = FALSE) +
  theme_minimal() +
  labs(title = "Correlation of Diamonds Carat and Price (USD)",
       x = "Carat",
       y = "Price (USD)") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



- From the scatter chart, when the carat weight increases, the price of diamonds increases and Data becomes more distributed as carat increases.

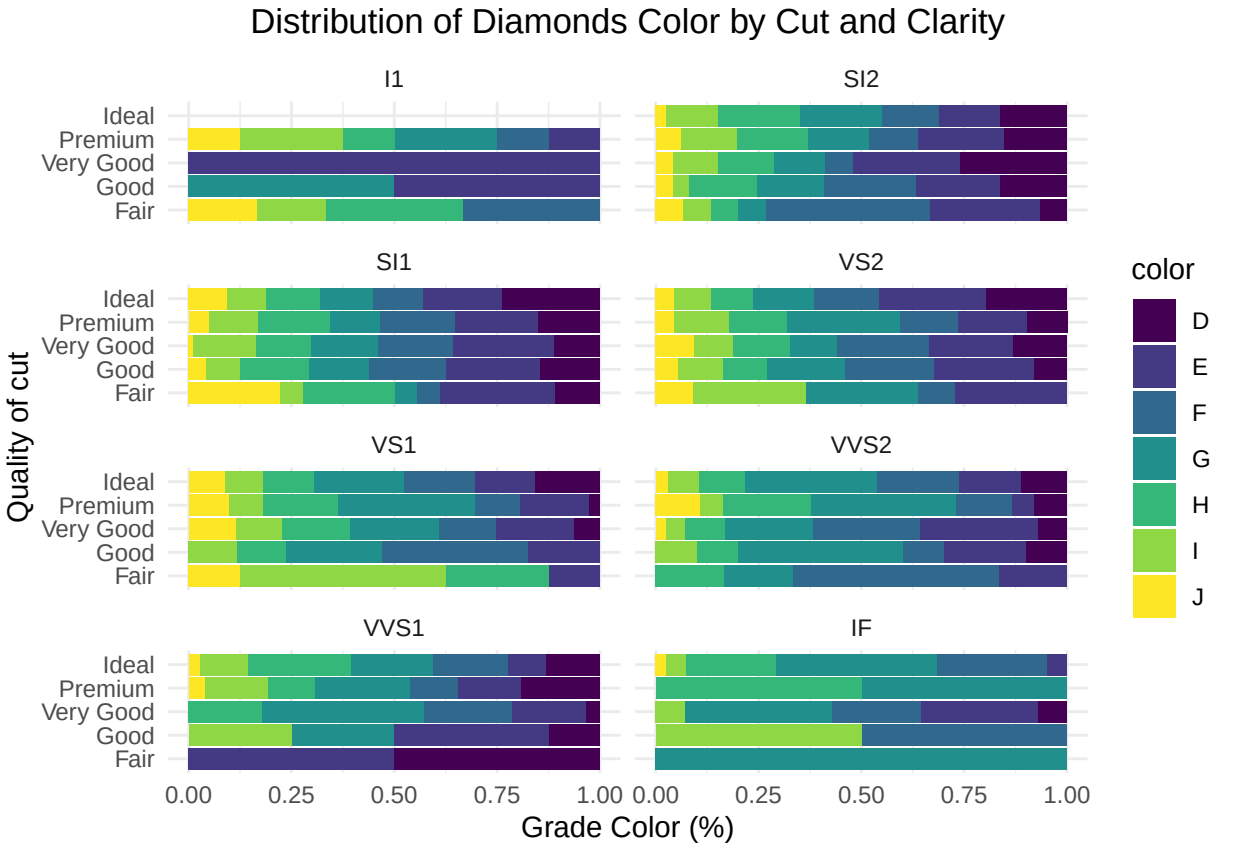
4.Count Plot Cut and Color

```
ggplot(small_df, aes(cut, color, colour  = color)) +
  geom_count() +
  theme_minimal() +
  labs(title = "Distribution of Diamonds Cut Quality by Color Grade",
       x = "Quality of Cut",
       y = "Quality of color") +
  theme(plot.title = element_text(hjust = 0.5))
```

- From the count chart, it can be seen that a high quality cut has an effect on having more diamonds, and G and H colored diamonds are found in large numbers in every cut, while J color diamonds tend to have a low number of diamonds in every cut.

5. Multiple Stacked Bar Plot Color by Cut and Clarity

```
ggplot(small_df, aes(cut, fill = color)) +
  geom_bar(position = "fill") +
  facet_wrap(~ clarity, ncol = 2) +
  coord_flip() +
  theme_minimal() +
  labs(title = "Distribution of Diamonds Color by Cut and Clarity",
       x = "Quality of cut",
       y = "Grade Color (%)") +
  theme(plot.title = element_text(hjust = 0.5))
```



- From the scatter plot, G and H colored diamonds are most commonly found in all cuts and clarity, while D colored diamonds are found in a smaller proportion compared to other colors in all cuts and clarity.