Homework_data_viz

KNP

2024-07-25

Homework diamonds graph

Prepare data and library

```
#library
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
```

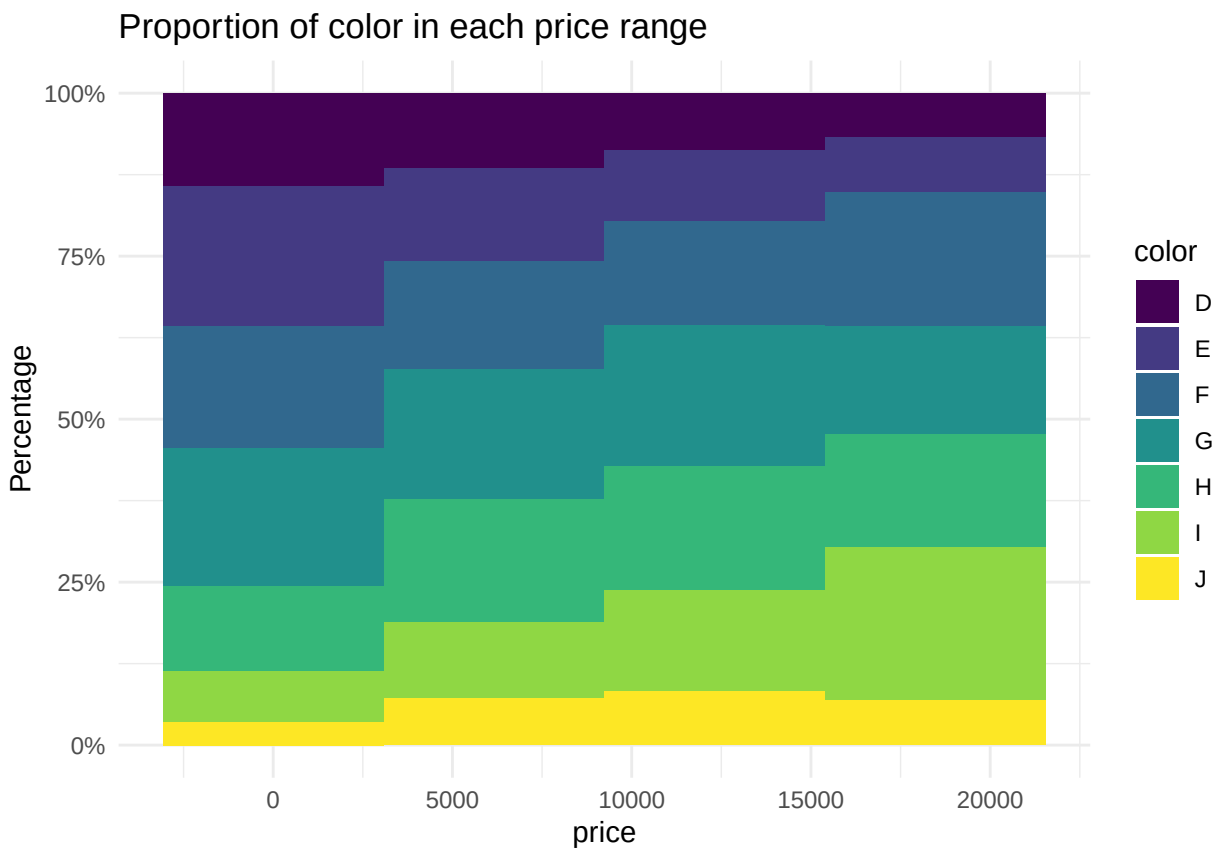
```
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
#Homework diamonds graph
set.seed(12)
sample_dia = diamonds%>%
  sample_frac(0.2)
```

Plot1 - quality of diamonds

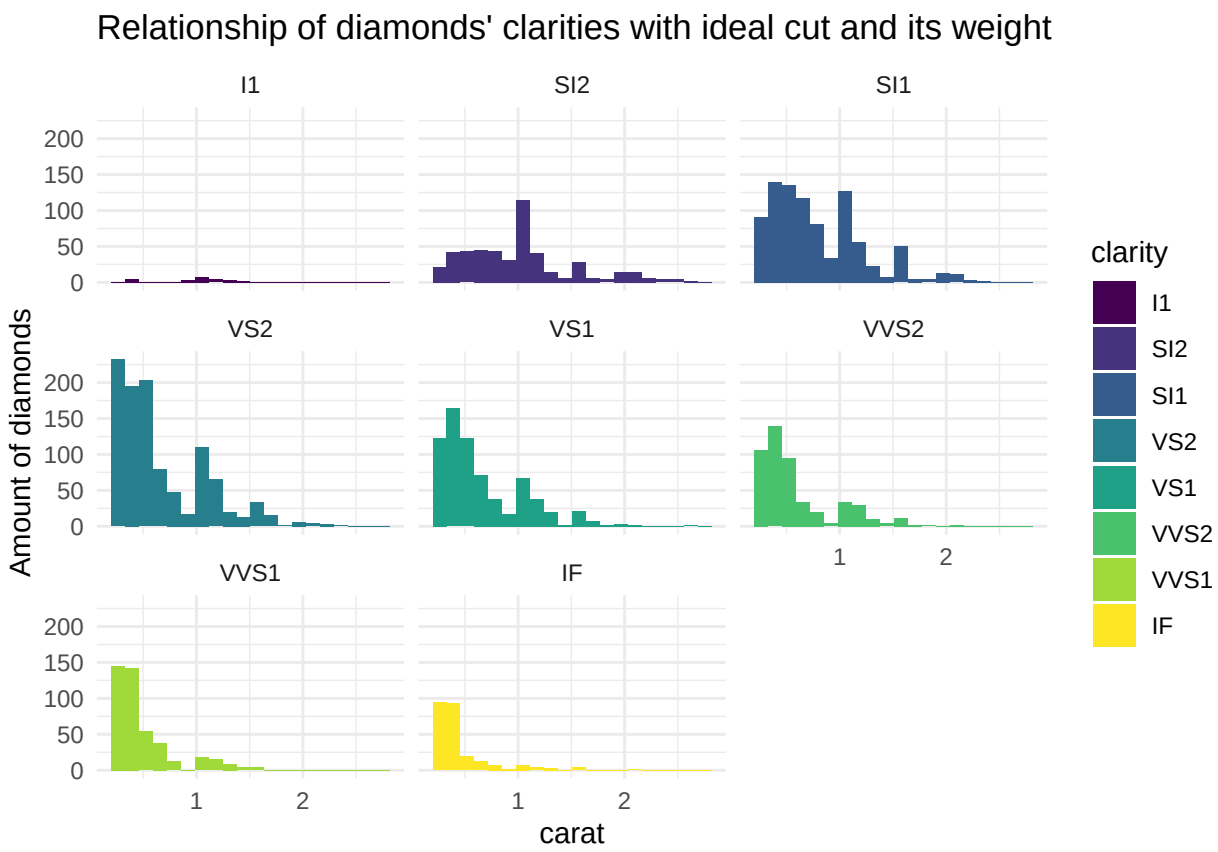
```
ggplot(sample_dia,aes(price,fill=color))+
  geom_histogram(bins = 4, position = "fill",  )+
  theme_minimal()+
  scale_y_continuous(labels = percent_format()) +
  labs(title = "Proportion of color in each price range",
       y = "Percentage")
```



color I has the biggest proportion in range 15000+ price

Plot2 - carat and clarity in ideal cut

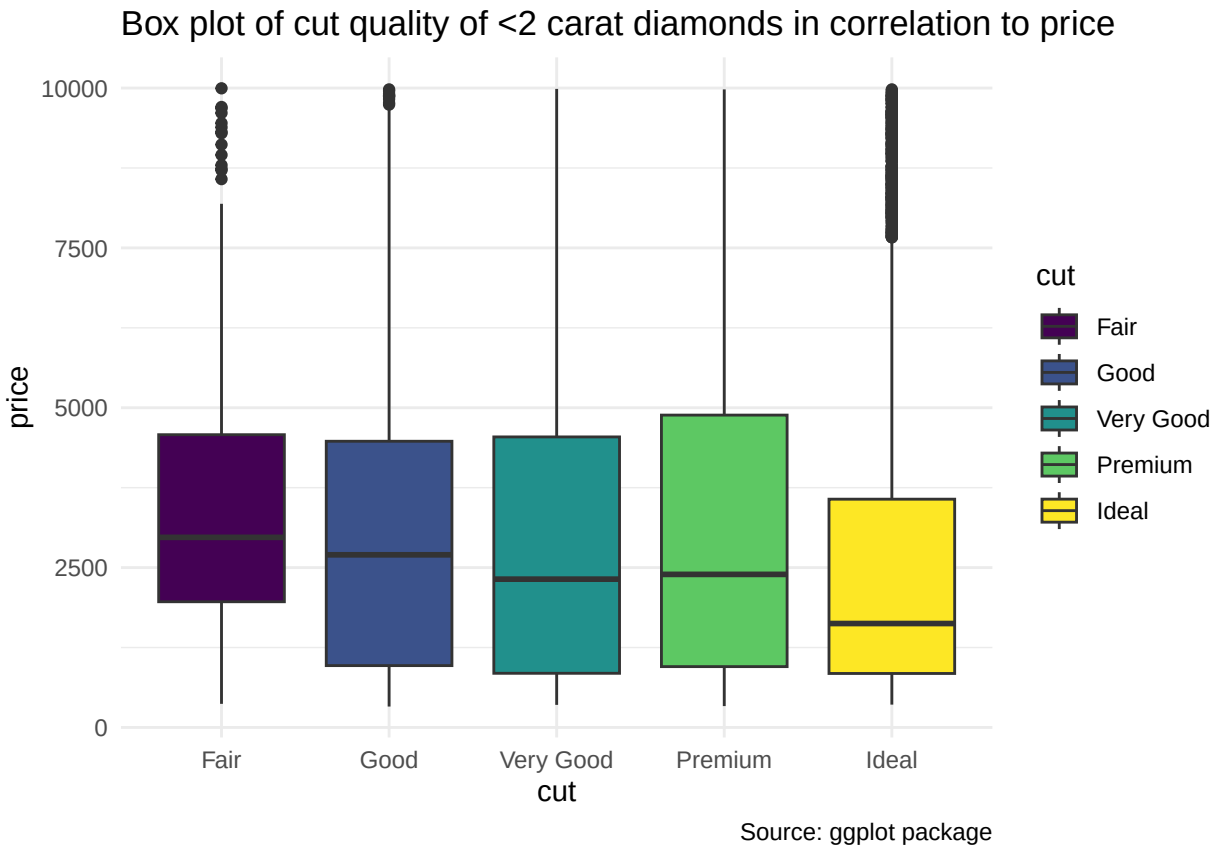
```
ggplot(sample_dia%>%filter(cut %in% "Ideal"),aes(carat,fill = clarity))+
  geom_histogram(bins = 20)+
  theme_minimal()+
  facet_wrap(~clarity)+
  labs(title = "Relationship of diamonds' clarities with ideal cut and its weight",
       y = "Amount of diamonds")
```



Only a few of I1 clarity diamonds are in the ideal cut. Most of the clarity which has ideal cut are less than 1 carat with SI2 having the largest no. of ideal cut diaonds that weigh more than 2 carats.

Plot 3 - box plot cut and price

```
ggplot(sample_dia%>% filter(carat <= 2, price < 10000), aes(cut,price,fill = cut))+
  geom_boxplot()+
  theme_minimal()+
  labs(title = "Box plot of cut quality of <2 carat diamonds in correlation to price",
       caption = "Source: ggplot package")
```

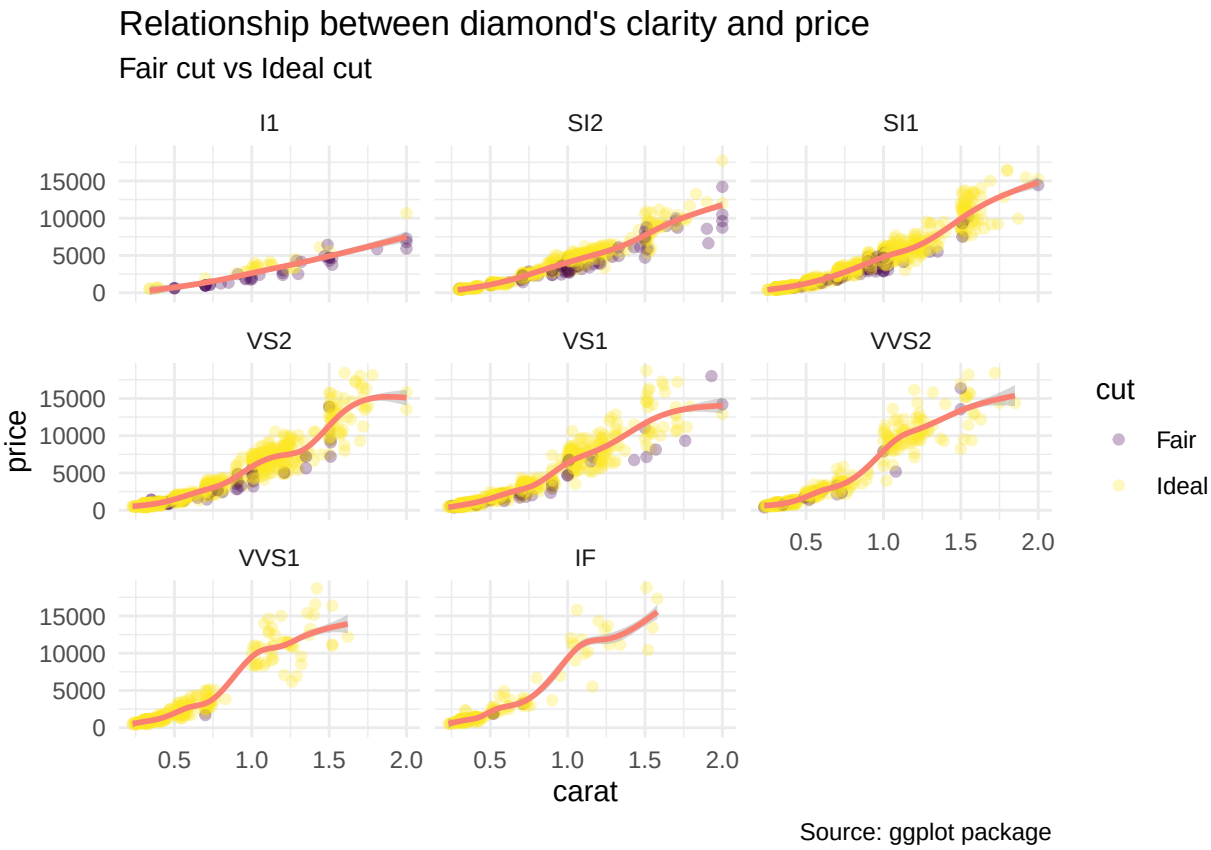


There are way more outliers in ideal cut compared to fair cut. It can be inferred that other factors (like clarity and carat) may influence the price as well not only the cut quality.

Plot4 - carat and cut and price

```
ggplot(sample_dia%>%filter(cut %in% c("Ideal","Fair"),carat <= 2),aes(carat,price,col = cut))+
  geom_point(alpha=0.3)+
  geom_smooth(col = "salmon")+
  theme_minimal()+
  facet_wrap(~clarity)+
  labs(title ="Relationship between diamond's clarity and price",
       subtitle = "Fair cut vs Ideal cut",
       caption = "Source: ggplot package")
```

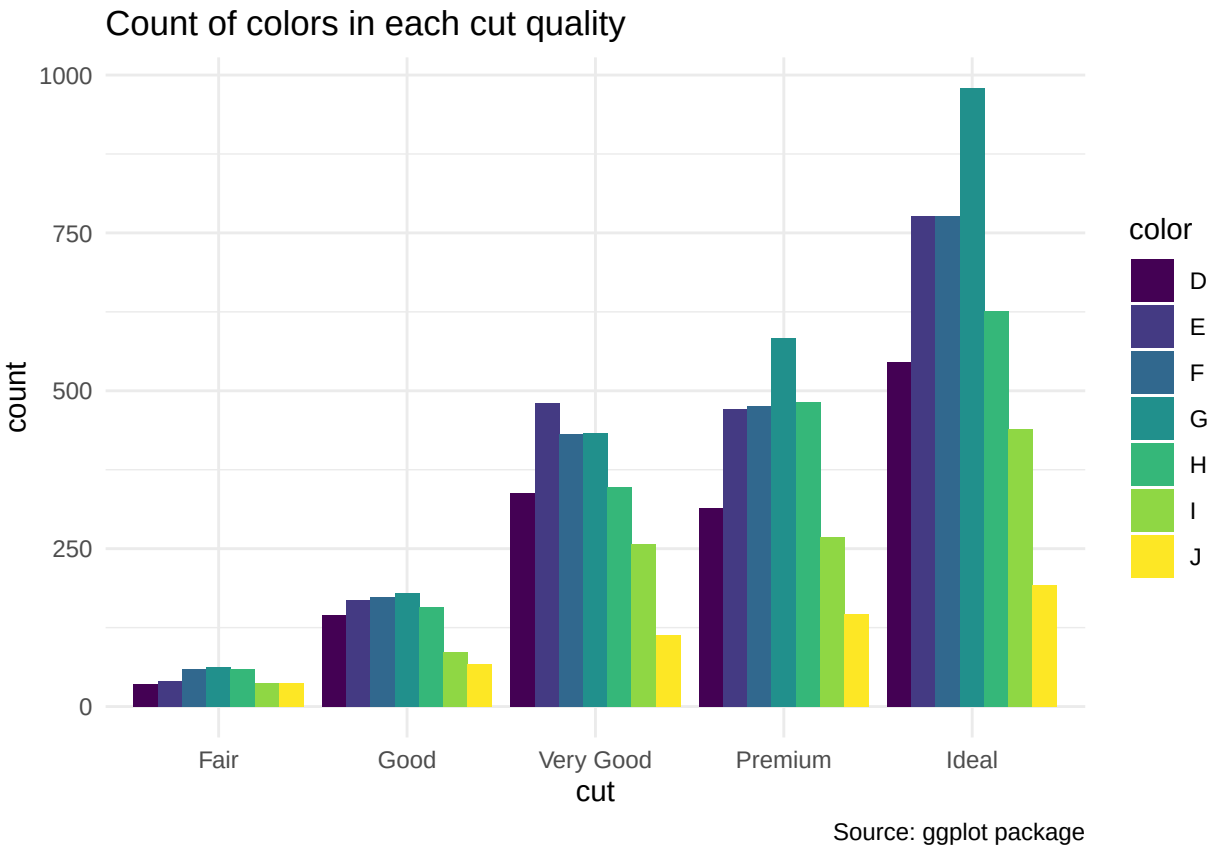
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Clarity played the most role in increasing the price of diamonds. Comparing the lower clarities which has mostly fair cus and higher clairty(like IF) which are all ideal cuts, the trend showed that the price rise exponentially as size increased in higher clarity unlike I1 which only has a steady rise in price.

Plot5 - color and cut

```
ggplot(sample_dia, aes(cut, fill = color))+
  geom_bar(position = "dodge")+
  theme_minimal()+
  labs(title = "Count of colors in each cut quality",
       caption= "Source: ggplot package" )
```



There are more diamonds in each cut categorities as the cut quality increase (fair has the least amount of diamonds while ideal has the most amount). Color J seems to be the rarest in every cut categories while color H diamonds are the most common in ideal cut category.