Homework Data Viz Batch 10

Kornkamol H

2024-09-10

Data Visualization using Diamonds Dataset

```
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
```

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

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

Prepare sample data

```
set.seed(42)
small_df <- diamonds %>%
  sample_n(2000) %>%
  arrange(carat)

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  0.2  Premium   D     VS2      62.3    60   367  3.73  3.68  2.31
## 2  0.21 Premium   E     SI2      61.9    56   394  3.84  3.82  2.37
## 3  0.23 Very Good E     VVS1     62.4    54   583  3.95  3.98  2.47
## 4  0.23 Very Good F     VS2      61.6    59   402  3.96  4     2.45
## 5  0.23 Very Good F     VS1      62.1    58   373  3.91  3.95  2.44
## 6  0.23 Ideal     G     SI1      61.2    56   375  3.97  4     2.44
```

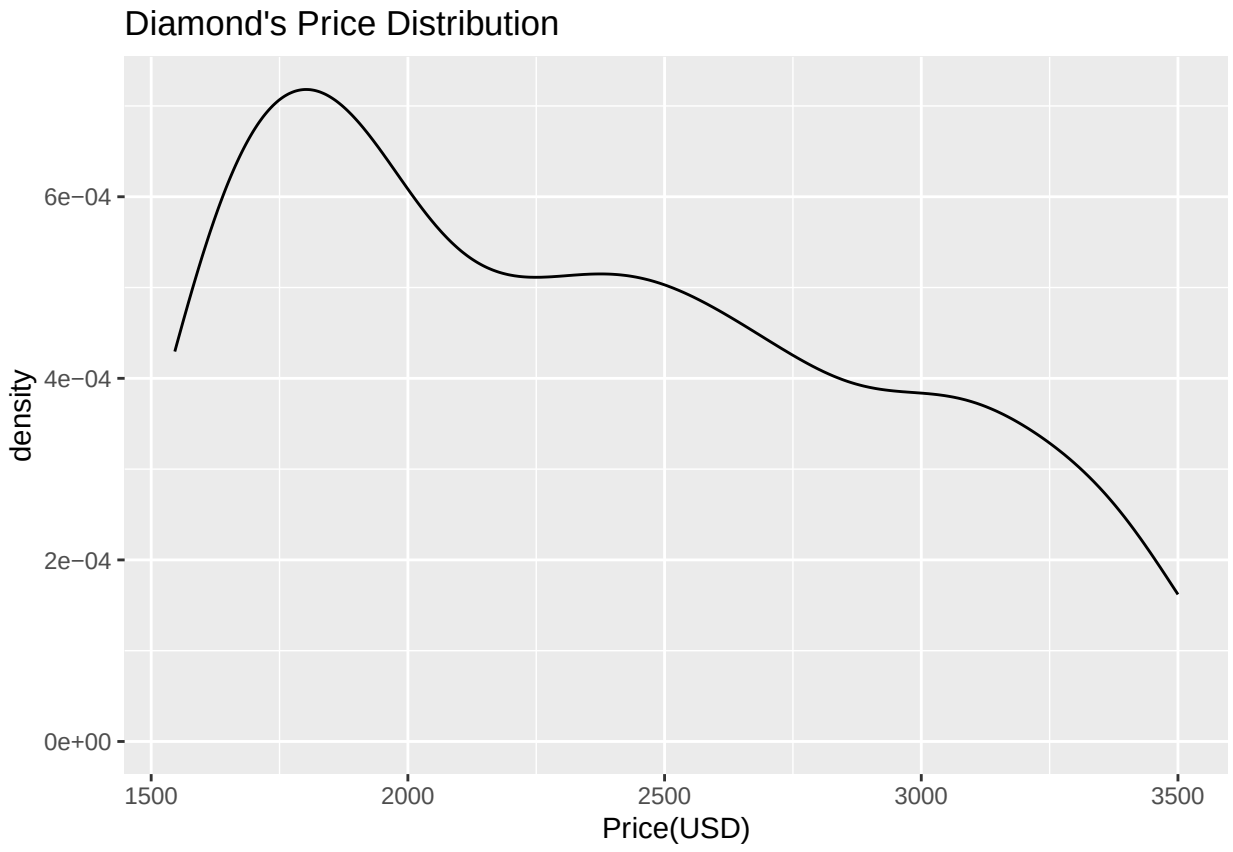
5 Chart and Analysis

1. A histogram of the price of diamonds

```
cat(paste("min:",min(diamonds$price)),"\nmax:",max(diamonds$price),
    "\nmedian:",median(diamonds$price))
```

```
## min: 326
## max: 18823
## median: 2401
```

```
ggplot(data = small_df %>%
         filter(between(price,1500,3500)),
       mapping = aes(x = price)) +
  geom_density() +
  labs(title = "Diamond's Price Distribution",
       x = "Price(USD)")
```



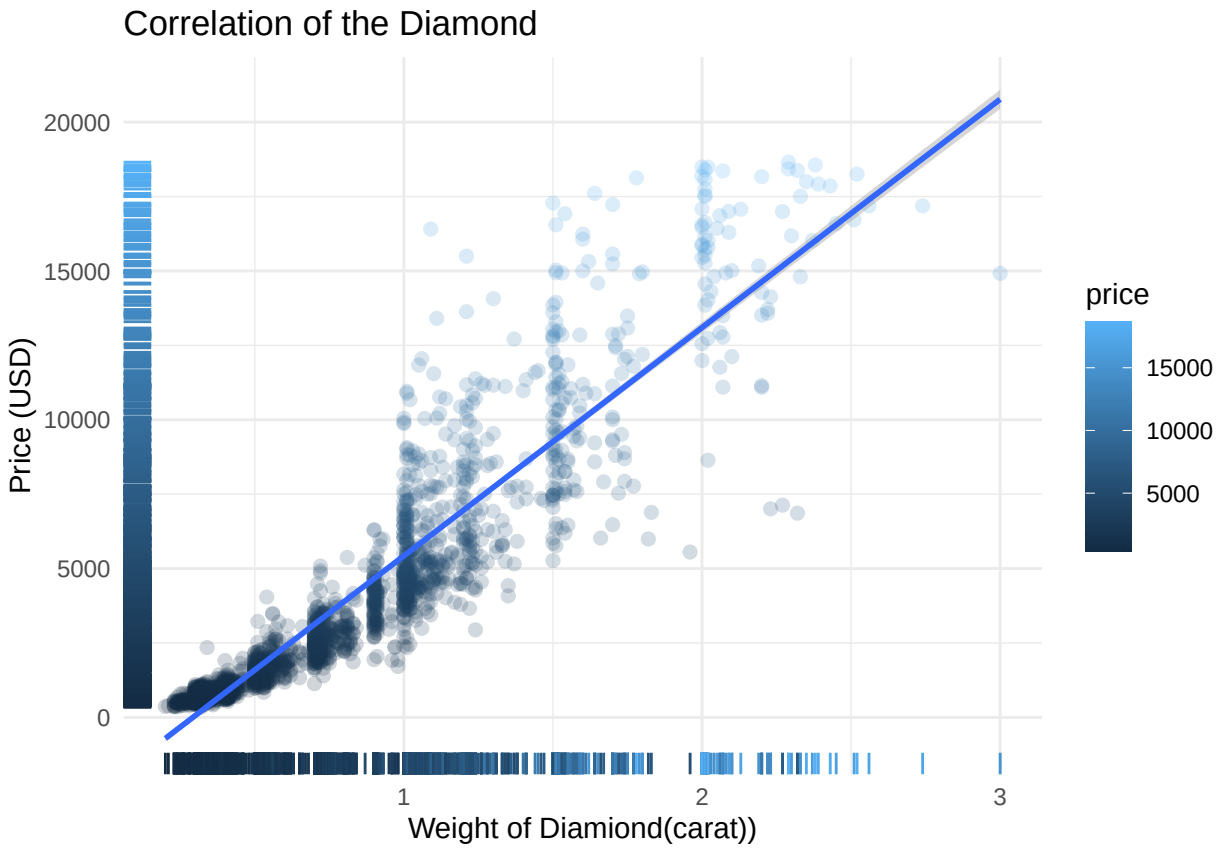
2. Correlation between two variables

```
ggplot(data = small_df,
       mapping = aes(carat, price, col=price)) +
  geom_point(size=2, alpha=0.2) +
```

```
  geom_smooth(method = "lm") +
  geom_rug() +
  theme_minimal() +
  labs(title = "Correlation of the Diamond",
       x = "Weight of Diamiond(carat))",
       y = "Price (USD)")
```

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

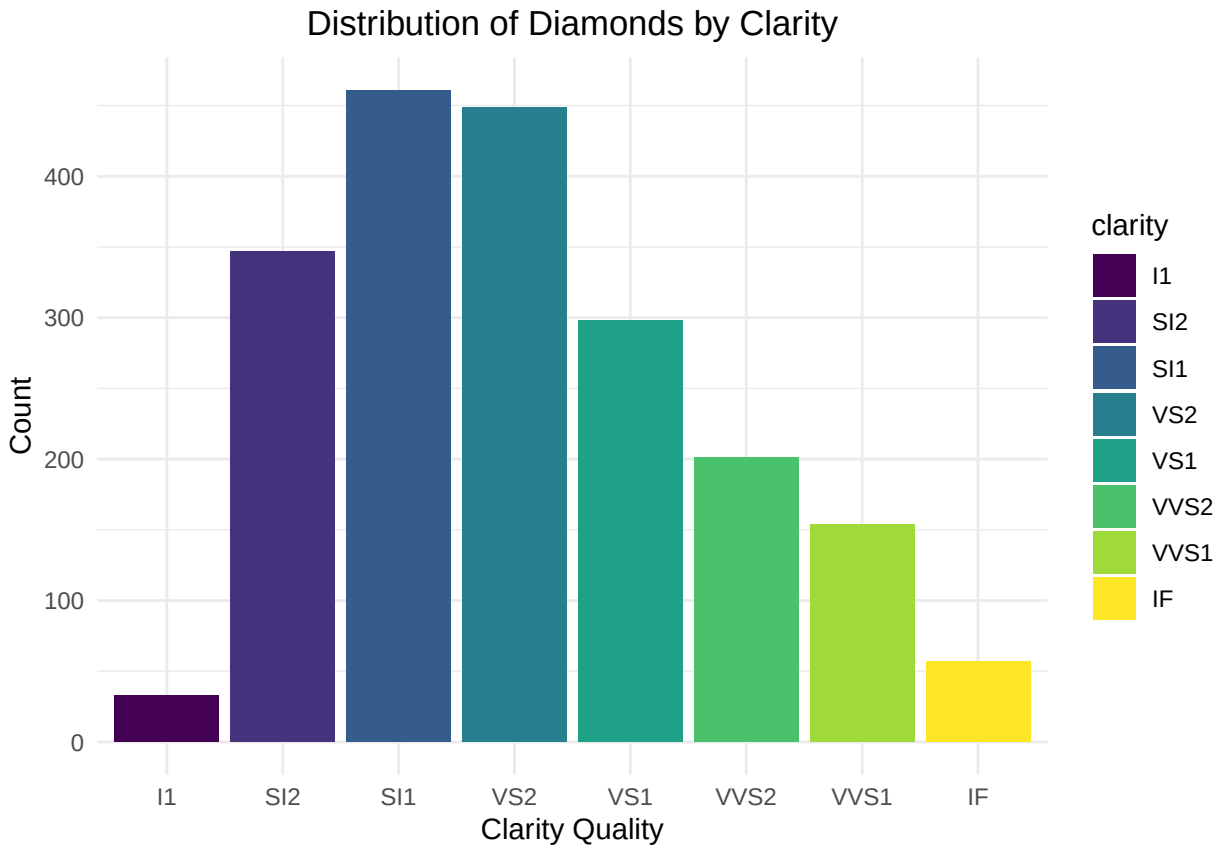
```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



3. The clarity of a diamond

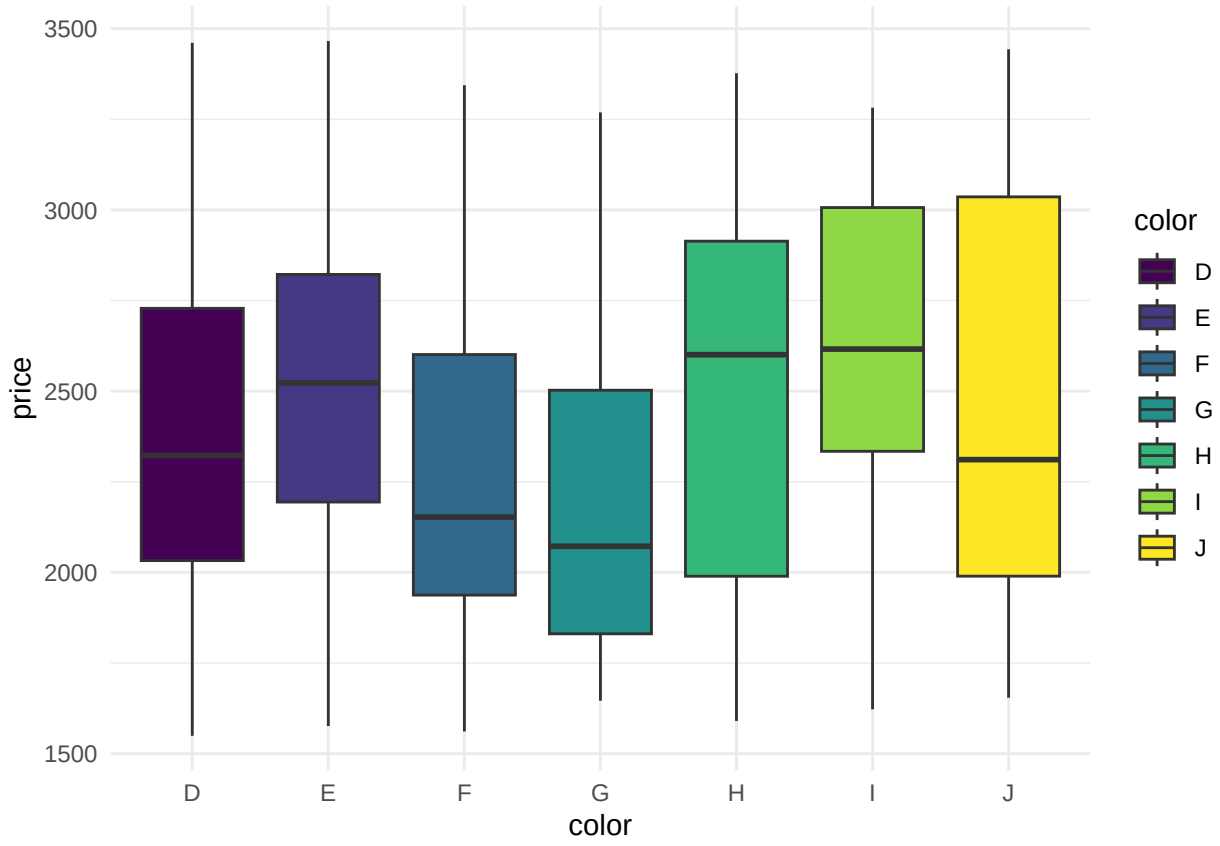
```
ggplot(data = small_df,
       mapping = aes(clarity, fill = clarity)) +
  geom_bar()+
  theme_minimal()+
  labs(title = "Distribution of Diamonds by Clarity",
```

```
       x = "Clarity Quality",
       y = "Count")+
  theme(plot.title = element_text(hjust = 0.5))
```



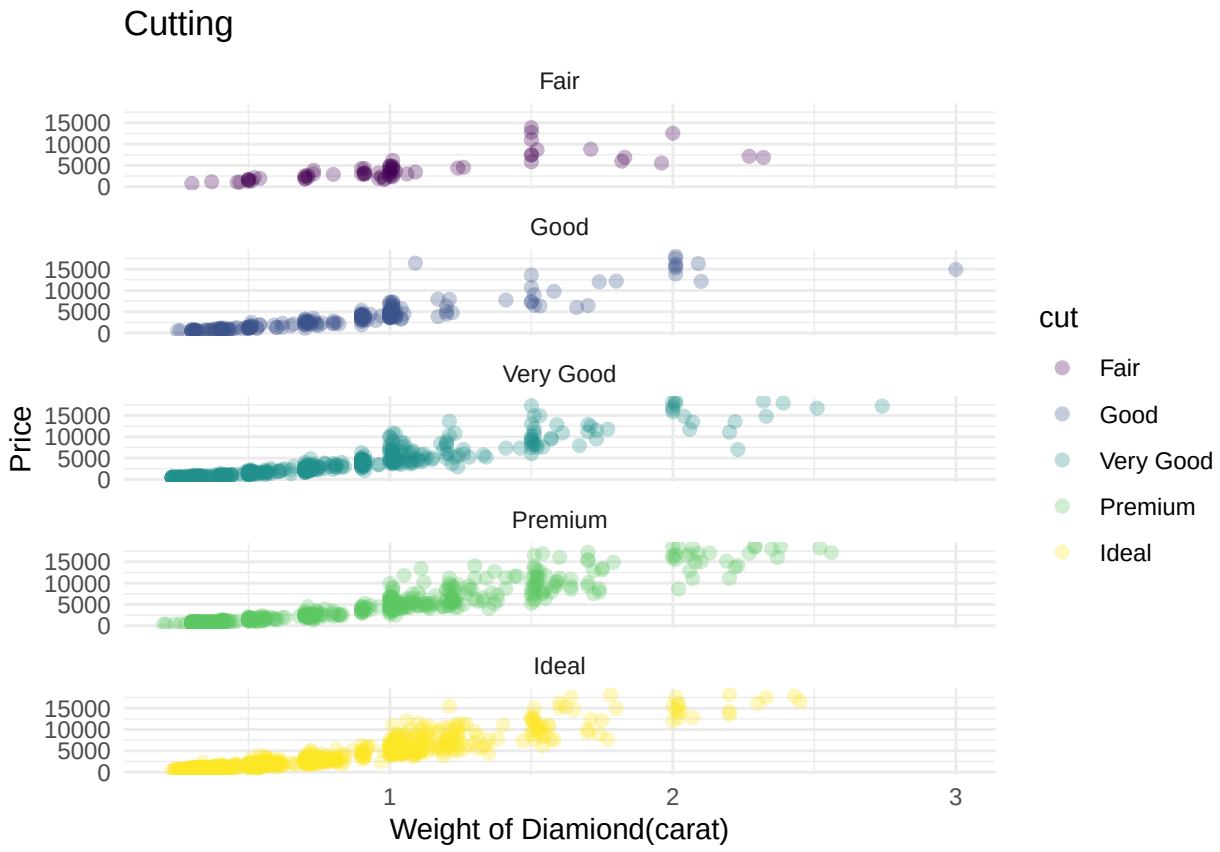
4.Distribution of Diamond Prices by Color

```
small_df %>%
  filter(cut %in% c("Fair", "Good", "Premium"),
         between(price,1500,3500)) %>%
ggplot(aes(color, price, fill=color))+
  geom_boxplot() +
  theme_minimal()
```



5. Scatter Plot between two variables

```
ggplot(data = small_df,
       aes(carat, price, col=cut))+
  geom_point(alpha= 0.3, size = 2) +
  facet_wrap(~ cut ,ncol=1) +
  labs(title = "Cutting",
       x = "Weight of Diamiond(carat)",
       y = "Price") +
  theme_minimal()
```


Cutting
Fair
Good
Very Good
Premium
Ideal
15000
10000
5000
0
Price
1
2
3
Weight of Diamiond(carat)
cut
Fair
Good
Very Good
Premium
Ideal