BST5920 Data Visualization - Homework 3

Miao Cai miao.cai@slu.edu

2018-10-23

The question is described below:

- Find a univariate (i.e., one variable) set of data and summarize its distribution. Use three or four of the methods for visualizing its distribution. Consider things like: histogram, dot plot, box plot, violin plot, kernel density estimate, Box-Cox transformation, etc. Write a brief summary of the methods you used and the conclusions you have drawn.
- Be sure to put your name at the top, save it as a PDF file.
- Due Monday 10/29 at 11:59 pm.

I use the diamonds data from the R package "ggplot2". This dataset contains the prices and other attributes of over 54,000 diamonds. The description of this data is shown below.

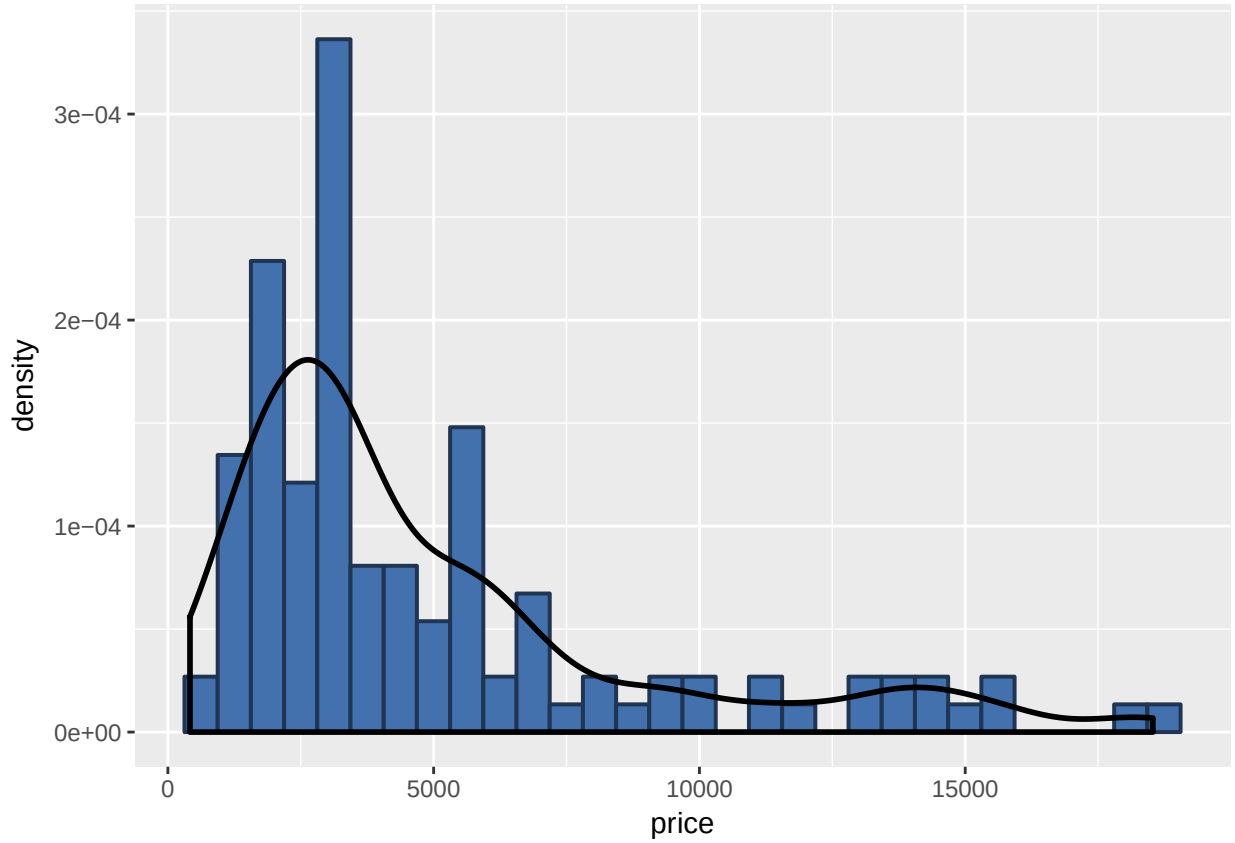
- **price**: price in US dollars ($326–$18,823)
- **carat**: weight of the diamond (0.2–5.01)
- **cut** quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- **color**: diamond colour, from J (worst) to D (best)
- **clarity**: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
- **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)

To reduce sample size and make cleaner descriptive plots, I focus on the diamonds with "Fair" quality of the cut (cut = "Fair") and worst diamond color (color = "J"). This leads to a sample of 119 diamonds. I use histogram and kernel density plots, violin and boxplots, and log transformations to explore the price distribution of these diamonds.

```
## # 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 0.230 Ideal     E     SI2      61.5   55.   326  3.95  3.98  2.43
## 2 0.210 Premium   E     SI1      59.8   61.   326  3.89  3.84  2.31
## 3 0.230 Good      E     VS1      56.9   65.   327  4.05  4.07  2.31
## 4 0.290 Premium   I     VS2      62.4   58.   334  4.20  4.23  2.63
## 5 0.310 Good      J     SI2      63.3   58.   335  4.34  4.35  2.75
## 6 0.240 Very Good J     VVS2     62.8   57.   336  3.94  3.96  2.48
```

1 Histogram and kernel density plot

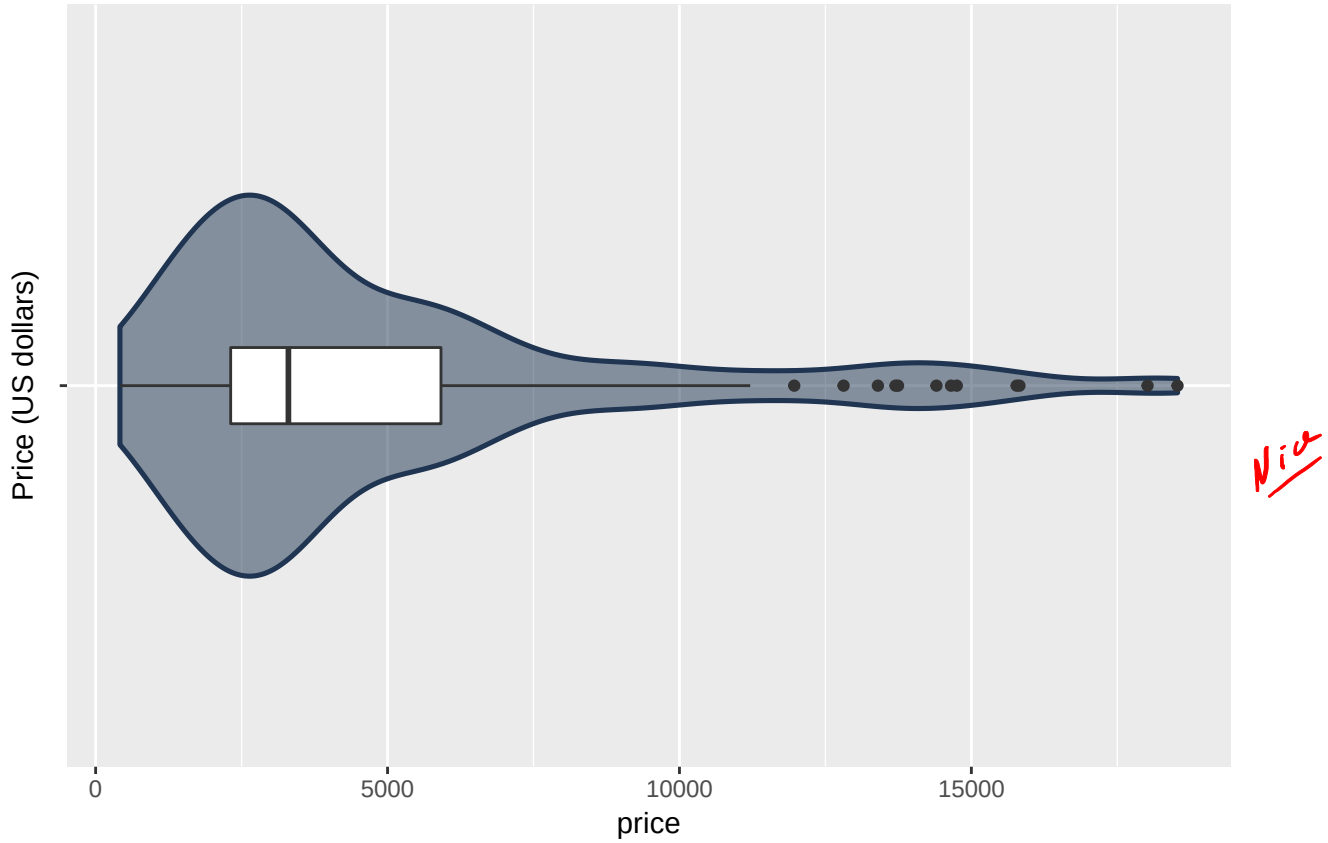
```
diamond_sub %>%
  ggplot(aes(x = price)) +
  geom_histogram(aes(y = ..density..),
    color = "#1F3552",
    fill = "#4271AE",
    size = 0.7) +
  geom_density(size = 1)
```



The histogram and kernel density plots show that price distribution is right skewed. The price of these diamonds are as high as over 15,000 dollars and as low as nearly 0 dollars. The highest density (mode) is at around 3,000 dollars.

2 Violin plot and boxplot

```
diamond_sub %>%
  ggplot(aes(x = "", y = price)) +
  geom_violin(
    width = 0.5,
    fill = "#1F3552",
    alpha = 0.5,
    color = "#1F3552",
    size = 0.9) +
  geom_boxplot(width = 0.1) +
  xlab("Price (US dollars)") +
  coord_flip()
```



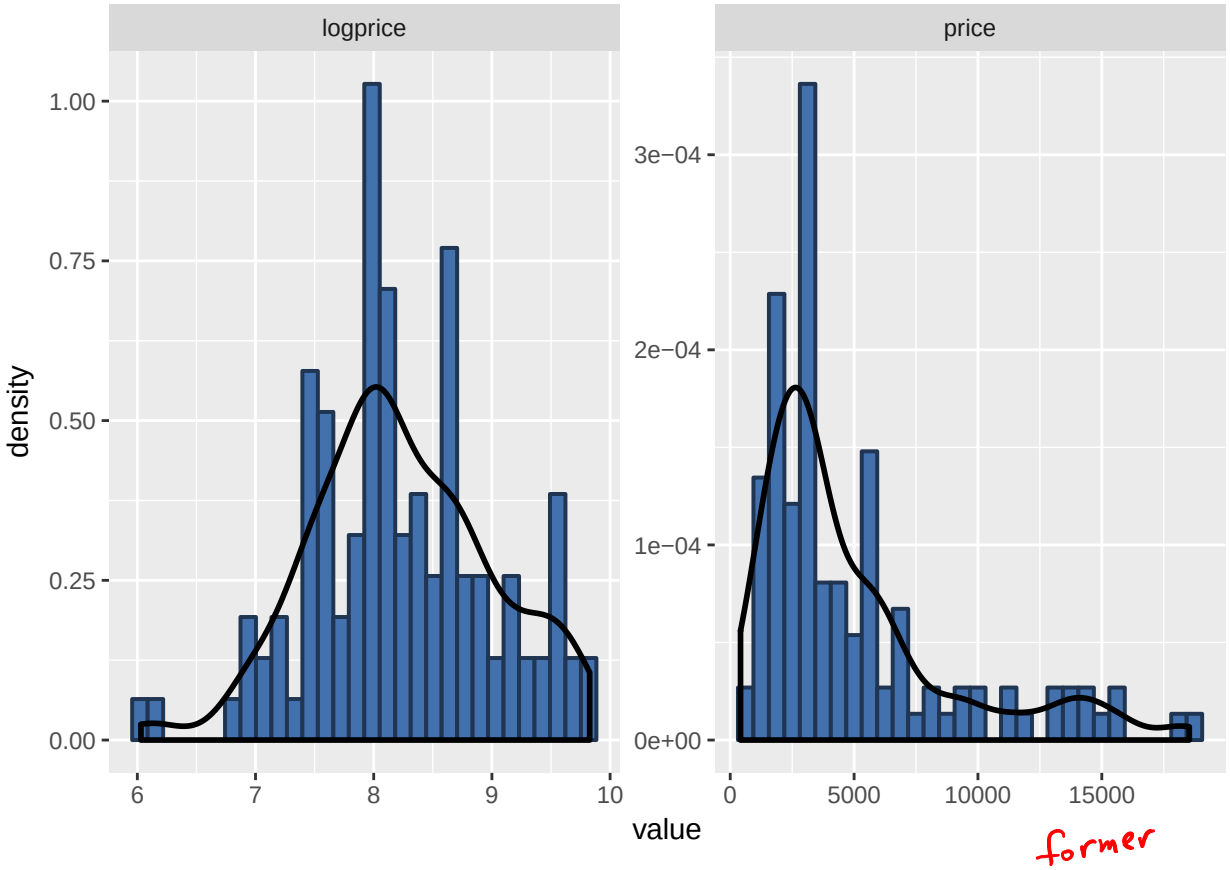
Violin plot and boxplot also demonstrates the price distributions of these diamonds. It also shows that the prices are right skewed. The highest density is at around 2,500 dollars from the violin plot. The 25% quartile, median and 75% quartiles are approximately 2,500, 3,000, and 6,000 dollars. Diamonds with prices higher than about 11,500 are defined as outliers in the boxplot.

3 Log transformation of price

```
require(tidyr)

subtran = diamond_sub %>%
  mutate(id = as.character(1:length(cut)), logprice = log(price)) %>%
  select(price, logprice) %>%
  gather(key = type, value = value)

subtran %>%
  ggplot(aes(x = value)) +
  geom_histogram(aes(y = ..density..),,
    color = "#1F3552",
    fill = "#4271AE",
    size = 0.7) +
  geom_density(size = 1) +
  facet_wrap(~type, scales = "free")
```



A comparison of log transformed price (left) and original price (right) shows that the latter one has a distribution more similar to the normal distribution.