Homework 5
of
STAT 3355 Data Analysis for Statisticians & Actuaries

Due: 11:30 am

March 30 (Wednesday), 2022

Let's work on the dataset `diamonds` in the package `ggplot2`. You can use the following code to load the data. Use necessary code to read the description of the dataset, which contains 53940 samples and 10 variables.

```
# Install the package if you never did
install.packages("ggplot2")

# Load the pacakge
library(ggplot2)

# Load the mpg dataset
data("diamonds")
```

Problem 1 ($1 \times 5 = 5$ points)

Use `ggplot2` to visualize the data. You need to paste the resulting plots and related code in order to get the full points. For each `ggplot2` plot:

- make it complete/readable, in other words, it should include axis label(s), title, and legend if necessary;

- write 1–2 sentence about what the chart tells you about the data.

(a) Choose a bin number or a binwidth (Hint: See page 11 of lecture_04c.pdf), explain why, and create a histogram of `carat`

(b) Make a scatter plot of $y =$ `price` against $x =$ `carat` and set the color to `clarity`

(c) Make a scatter plot of $y =$ `price` against $x =$ `carat` and add a smooth line to each group of points defined by `clarity`

(d) Make a scatter plot of $y =$ `price` against $x =$ `carat` and facet it by `clarity`

(e) Show `carat` vs `cut`, make a point, a jitter, a box plot and a violin plot, respectively. Which one is the best for visualization?
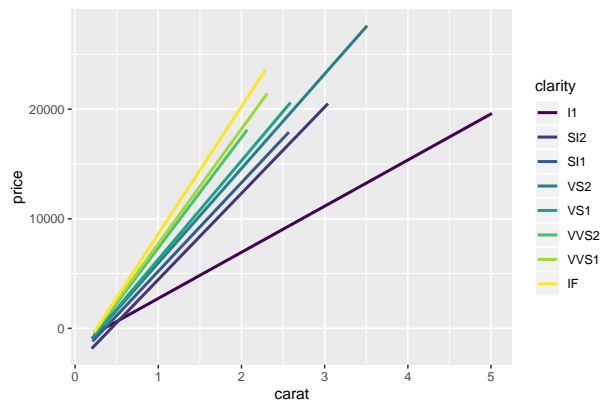
Problem 2 ($1 \times 5 = 5$ **points**)

Use `ggplot2` to recreate the following plots with title. You need to paste the new plots and related code in order to get full points.
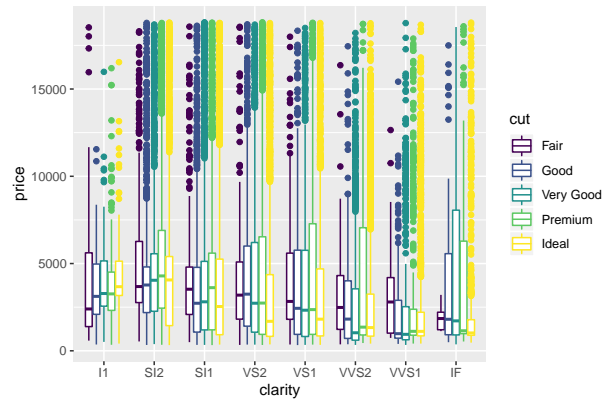
(a) Recreate the following two plots, add a short title, and comment on the merits of each one compared to the other
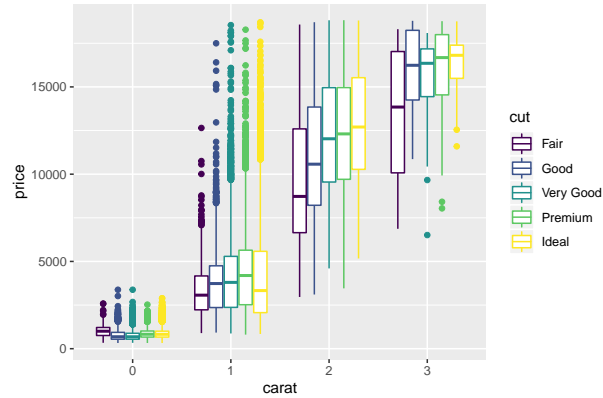

(b) Recreate the following plot and add a short title

(c) Recreate the following plot and add a short title

(d) Recreate the following plot and add a short title

(e) Recreate the following plot and add a short title (Hint: Choose binwidth $= 0.1$)