CAAP Statistics - Lec05

R Session2

Jul 12, 2022

Review

- Numerical Data
  - Graphical summary
    - Scatterplot
    - Histogram
    - Boxplot
  - Numerical summary
    - Mean and Variance
- Categorical Data
  - Graphical Summary
    - Contingency tables and Bar plot
    - Mosaic plot(If time allows)

Data

At a first glance, does there appear to be a relationship between vaccine and infection?

		outcome		
		infection	no infection	Total
treatment	vaccine	5	9	14
	placebo	6	0	6
	Total	11	9	20

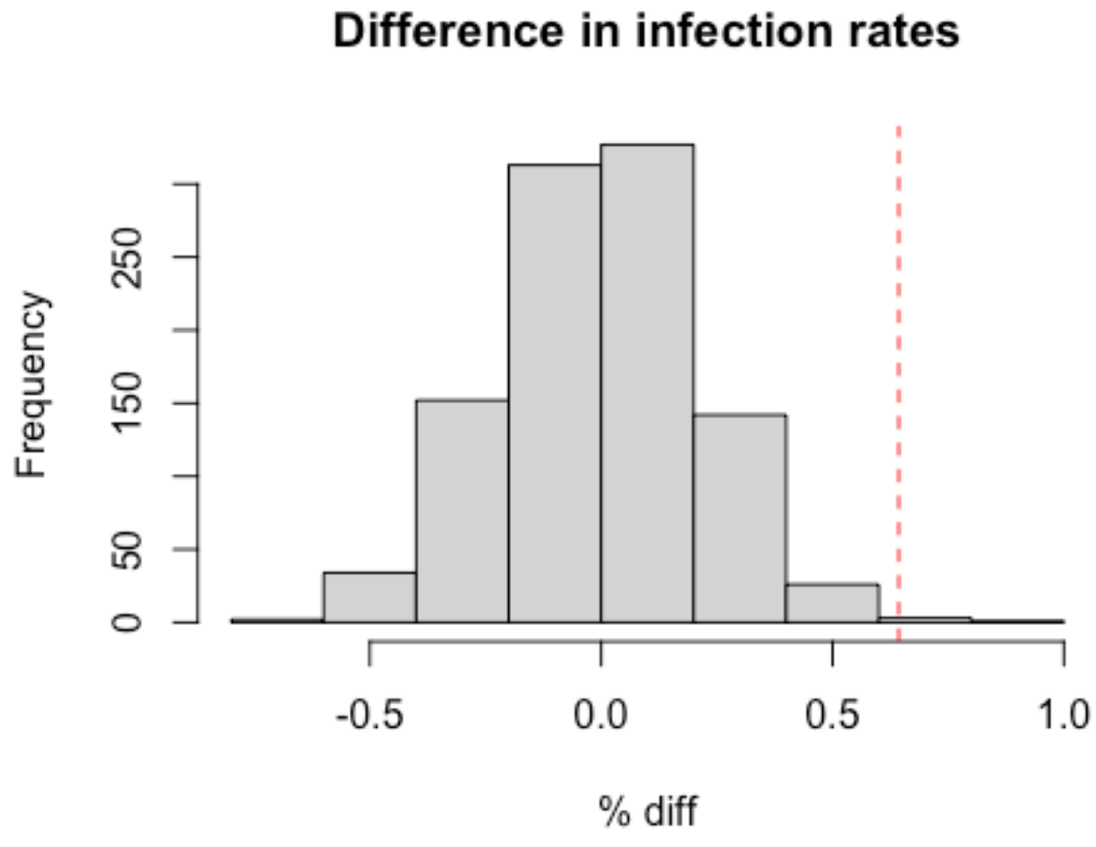
Figure 2.29: Summary results for the malaria vaccine experiment.

% of treatment group got infected: 5 / 14 = 0.357

% of control group got infected: 6 / 6 = 1.000

Simulations Using Software

In reality, we use software to generate the simulations. The histogram below shows the distribution of simulated differences in promotion rates based on 1000 simulations.



Practice

Do the results of the simulation you just ran provide convincing evidence that the vaccine is effective, i.e. dependence between the vaccination and infection rate?

A. No, the data do not provide convincing evidence for the alternative hypothesis, therefore we can't reject the null hypothesis of independence between the vaccination and infection rate. The observed difference between the two proportions was due to chance.

B. Yes, the data provide convincing evidence for the alternative hypothesis that the vaccine is effective against the malaria. The observed difference between the two proportions was due to a real effect of vaccination.

Practice

Do the results of the simulation you just ran provide convincing evidence that the vaccine is effective, i.e. dependence between the vaccination and infection rate?

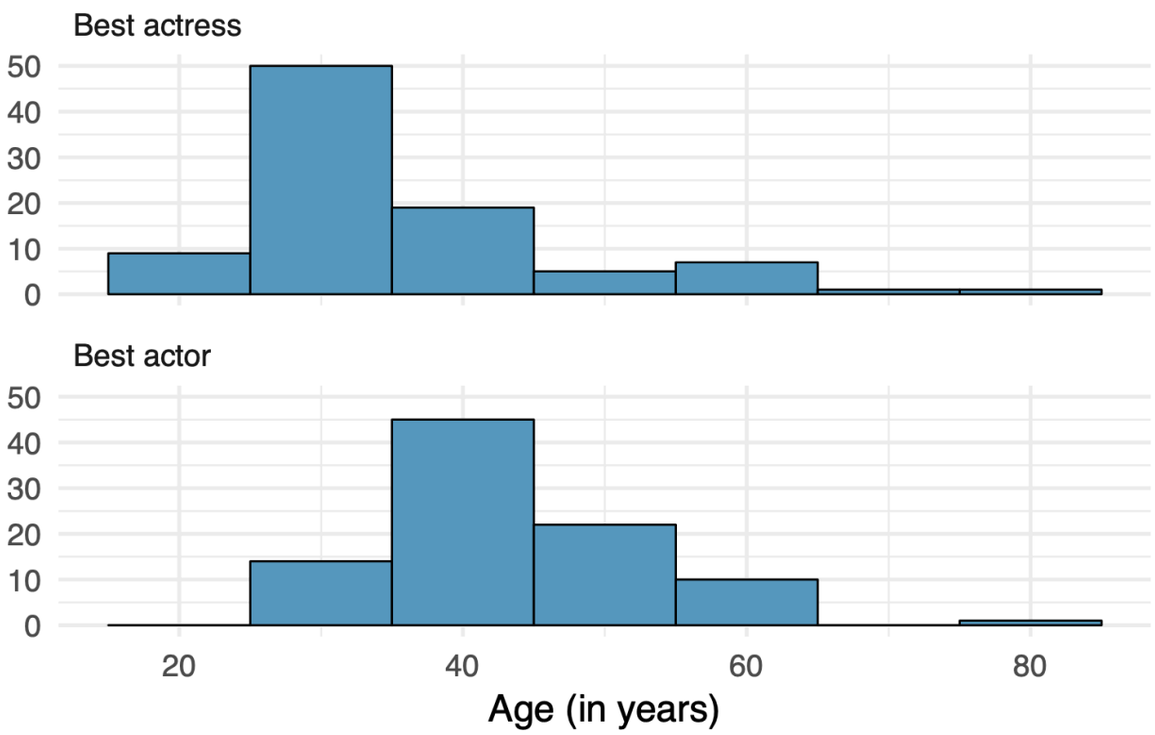
A. No, the data do not provide convincing evidence for the alternative hypothesis, therefore we can't reject the null hypothesis of independence between the vaccination and infection rate. The observed difference between the two proportions was due to chance.

B. Yes, the data provide convincing evidence for the alternative hypothesis that the vaccine is effective against the malaria. The observed difference between the two proportions was due to a real effect of vaccination.

Let’s discuss!

Oscar Winners

The first Oscar awards for best actor and best actress were given out in 1929. The histograms below show the age distribution for all of the best actor and best actress winners from 1929 to 2018. Summary statistics for these distributions are also provided. Compare the distributions of ages of best actor and actress winners.

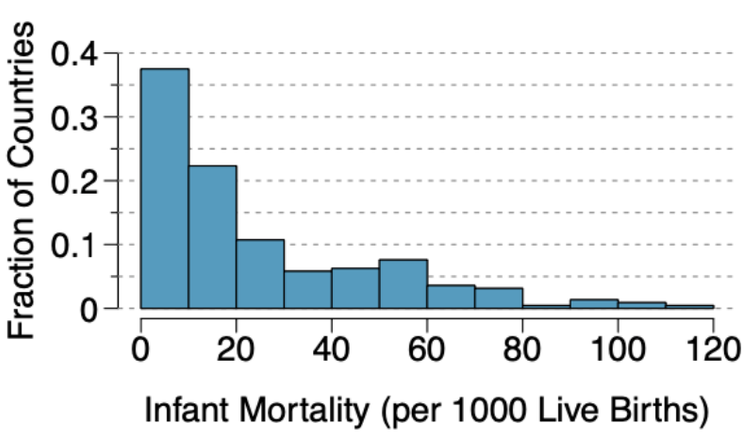


	Best Actress
Mean	36.2
SD	11.9
n	92

	Best Actor
Mean	43.8
SD	8.83
n	92

Infant mortality

The infant mortality rate is defined as the number of infant deaths per 1,000 live births. This rate is often used as an indicator of the level of health in a country. The relative frequency histogram below shows the distribution of estimated infant death rates for 224 countries for which such data were available in 2014.

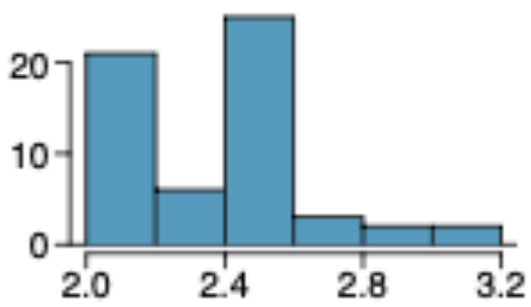


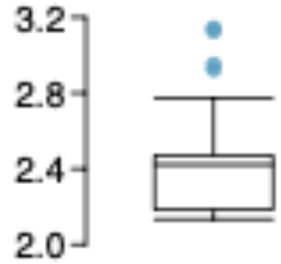
(a) What can you observe from the above histogram?(skewness, mean, median..)

(b) Would you expect the mean of this data set to be smaller or larger than median? Explain your reasoning.

Marathon winners

The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1970 and 1999.

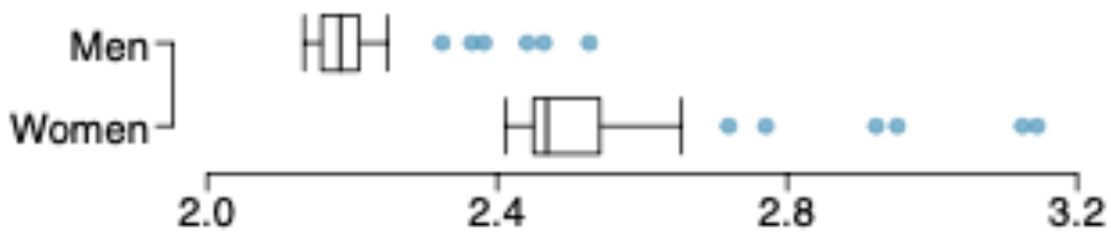




(a) What features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?

(b) What may be the reason for the bimodal distribution? Explain.

(c) Compare the distribution of marathon times for men and women based on the box plot shown below.



Learning Objectives

- Introduction to RMarkdown
- Data manipulation using R
- Playing with bin width of histogram
- Boxplot using R
- Mosaic plot using R

Load packages

```
# install.packages("lattice")
library(tidyverse)
library(openintro)
library(ggplot2)
```

Let's see the actual data

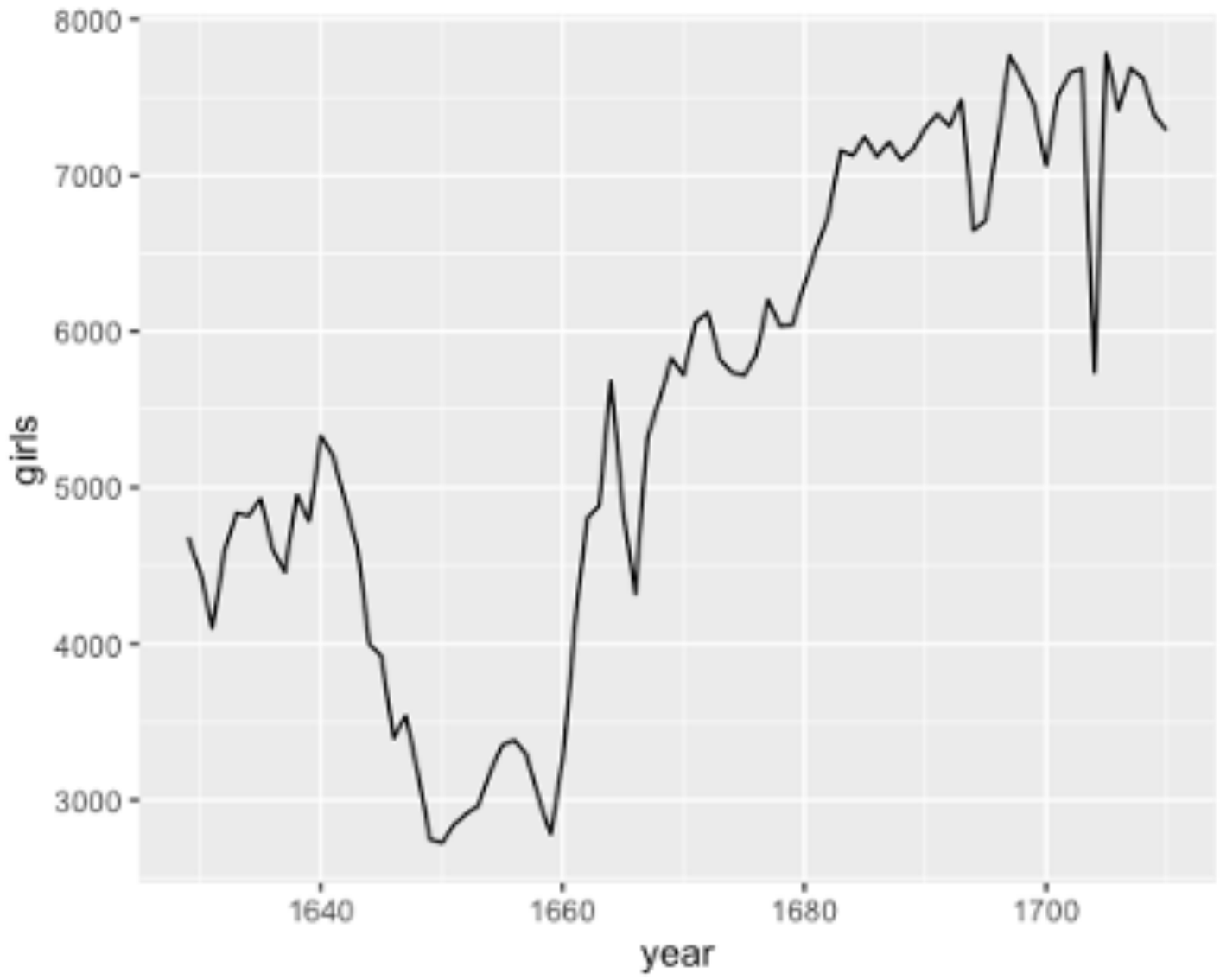
```
arbuthnot # from openintro package
data_web = read.csv("https://www.openintro.org/book/
statdata/arbuthnot.csv") # from web
# getwd() # check for the current working directory
# data = read.csv("arbuthnot.csv") # read from the
working directory
```

How does the data look like?

```
glimpse(arbuthnot)
## Rows: 82
## Columns: 3
## $ year  <int> 1629, 1630, 1631, 1632, 1633, 1634,
1635, 1636, 1637, 1638, 1639…
## $ boys  <int> 5218, 4858, 4422, 4994, 5158, 5035,
5106, 4917, 4703, 5359, 5366…
## $ girls <int> 4683, 4457, 4102, 4590, 4839, 4820,
4928, 4605, 4457, 4952, 4784…
```

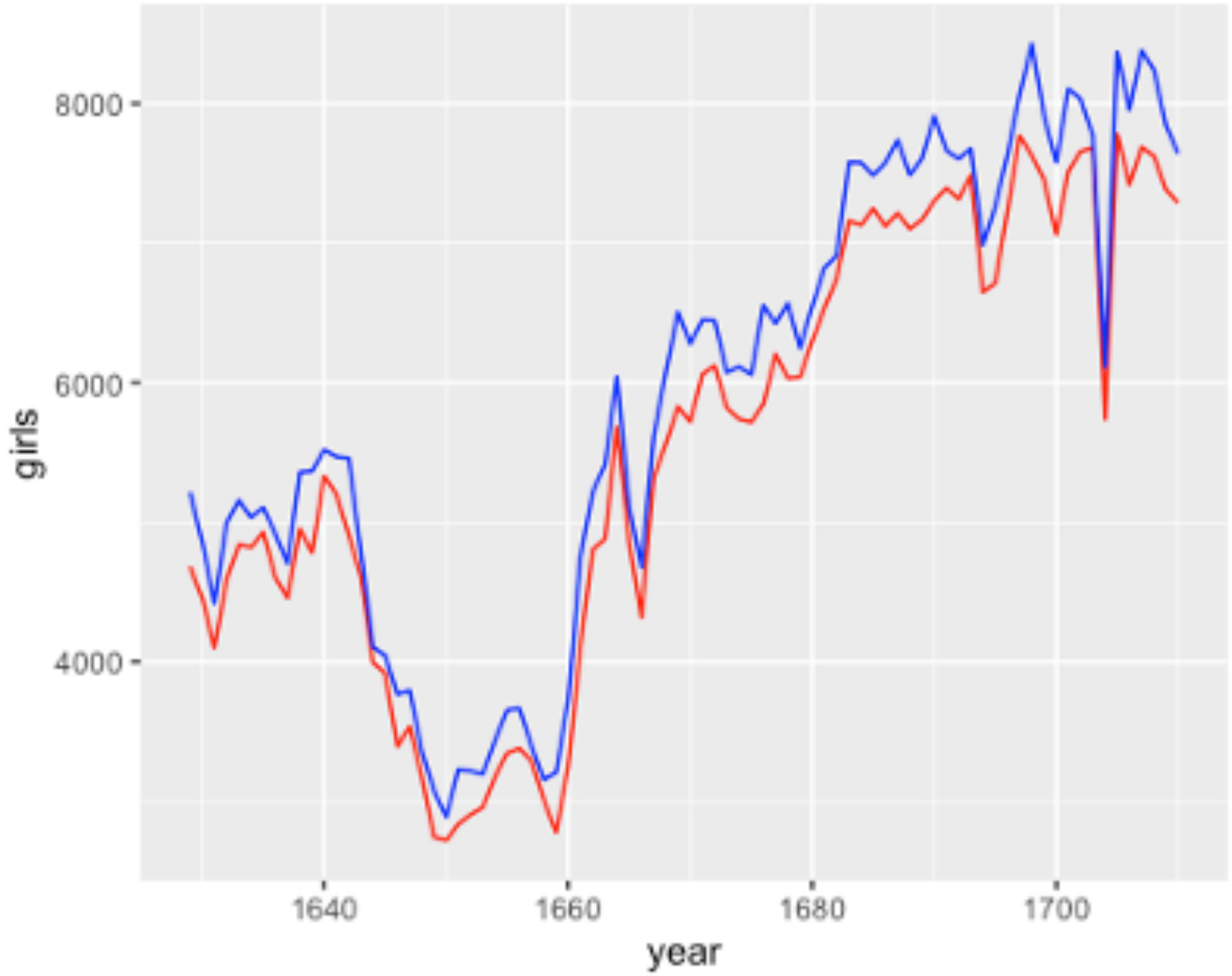
Visualize the Data

```
ggplot(data = arbuthnot,
aes(x=year, y = girls))+
  geom_line()
```



Visualize the Data

```
ggplot(data = arbuthnot)+
  geom_line(aes(x=year, y =
girls),colour = "red")+
  geom_line(aes(x=year, y =
boys),colour="blue")
```

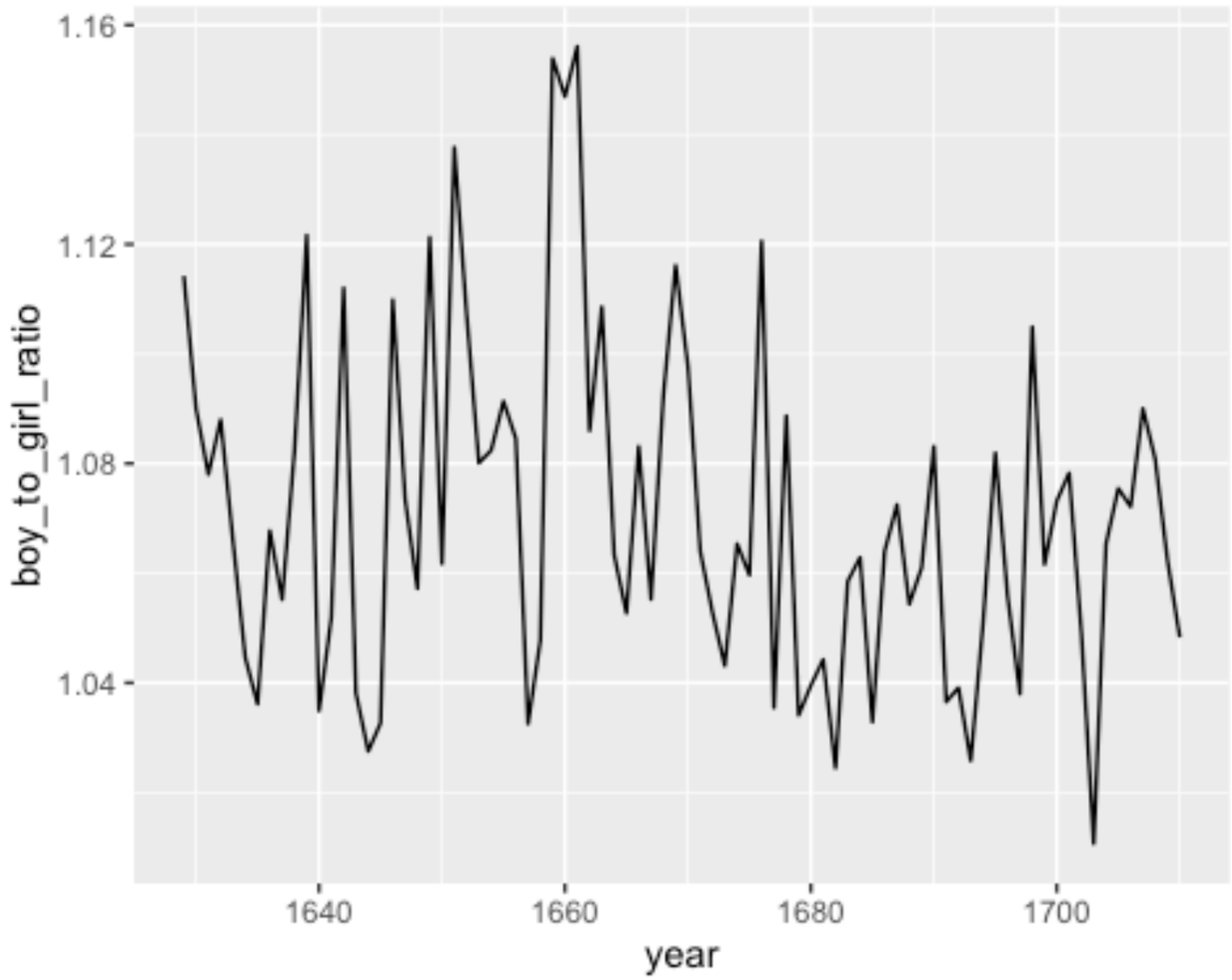


Manipulate the data matrix

```
arbuthnot = arbuthnot %>%
  mutate(boy_to_girl_ratio = boys / girls)
head(arbuthnot)
## # A tibble: 6 × 4
##    year  boys girls boy_to_girl_ratio
##   <int> <int> <int>              <dbl>
## 1  1629  5218  4683               1.11
## 2  1630  4858  4457               1.09
## 3  1631  4422  4102               1.08
## 4  1632  4994  4590               1.09
## 5  1633  5158  4839               1.07
## 6  1634  5035  4820               1.04
```

Visualize the Data over time

```
ggplot(arbuthnot, aes(x=year,y
= boy_to_girl_ratio))+
  geom_line()
```



On Your Own: Try the same analysis on `present` dataset and compare with `arbuthnot`

```
data(present)
glimpse(present)
## Rows: 63
```

- ```
  ## Columns: 3
  ```
- ```
  ## $ year  <dbl> 1940, 1941, 1942, 1943, 1944, 1945,
  1946, 1947, 1948, 1949, 1950…
  ```
- ```
  ## $ boys  <dbl> 1211684, 1289734, 1444365, 1508959,
  1435301, 1404587, 1691220, 1…
  ```
- ```
  ## $ girls <dbl> 1148715, 1223693, 1364631, 1427901,
  1359499, 1330869, 1597452, 1…
  ```

Exploring Numerical Data

Numerical Dataset: diamonds

The `diamonds` dataset consists of prices and quality information from about 54,000 diamonds, and is included in the ggplot2 package.

```
data(diamonds)
```

You can see the information about the data using `?diamonds` command

- `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 `D` (best) to `J` (worst)
- `clarity`: a measurement of how clear the diamond is (`I1` (worst), `SI2`, `SI1`, `VS2`, `VS1`, `VVS2`, `VVS1`, `IF` (best))

Overview of the dataset

```
str(diamonds)
## tibble [53,940 × 10] (S3: tbl_df/tbl/data.frame)
##  $ carat  : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
##  $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
##  $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
##  $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
##  $ depth  : num [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
##  $ table  : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...
##  $ price  : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...
##  $ x      : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
##  $ y      : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
##  $ z      : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

Numerical Summary - Five Number

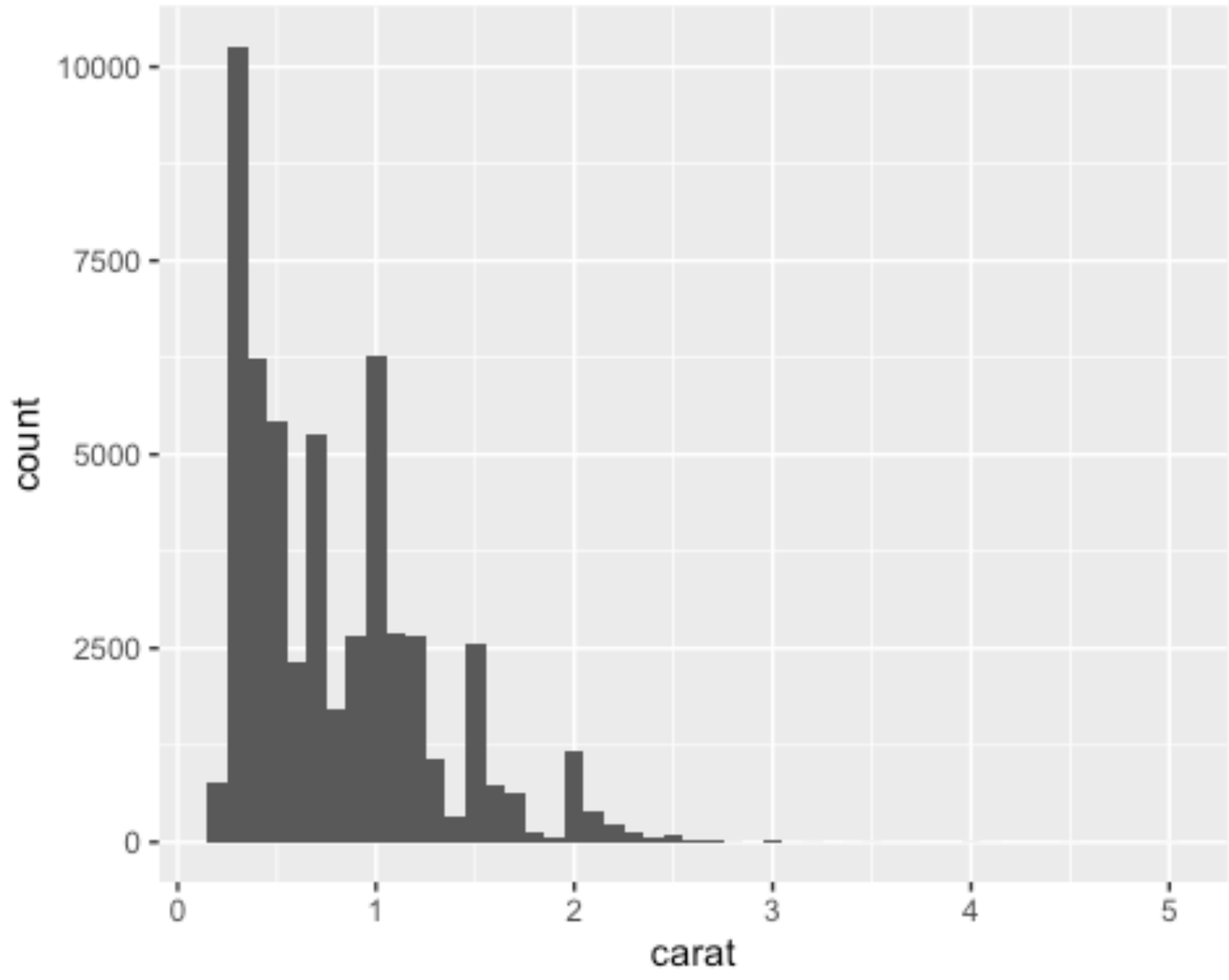
```
mean(diamonds$price)
## [1] 3932.8
sd(diamonds$price)
## [1] 3989.44
median(diamonds$price)
## [1] 2401
fivenum(diamonds$price)
## [1]   326.0   950.0  2401.0  5324.5 18823.0
```

Numerical Summary - Aggregation using pipeline via tidyverse

```
diamonds %>%
  group_by(cut)%>% # categorical variable
  summarise(mean = mean(price), median = median(price), sd =
sd(price))
## # A tibble: 5 × 4
##   cut         mean median    sd
##   <ord>      <dbl>  <dbl> <dbl>
## 1 Fair       4359.  3282  3560.
## 2 Good       3929.  3050. 3682.
## 3 Very Good 3982.  2648  3936.
## 4 Premium    4584.  3185  4349.
## 5 Ideal      3458.  1810  3808.
```

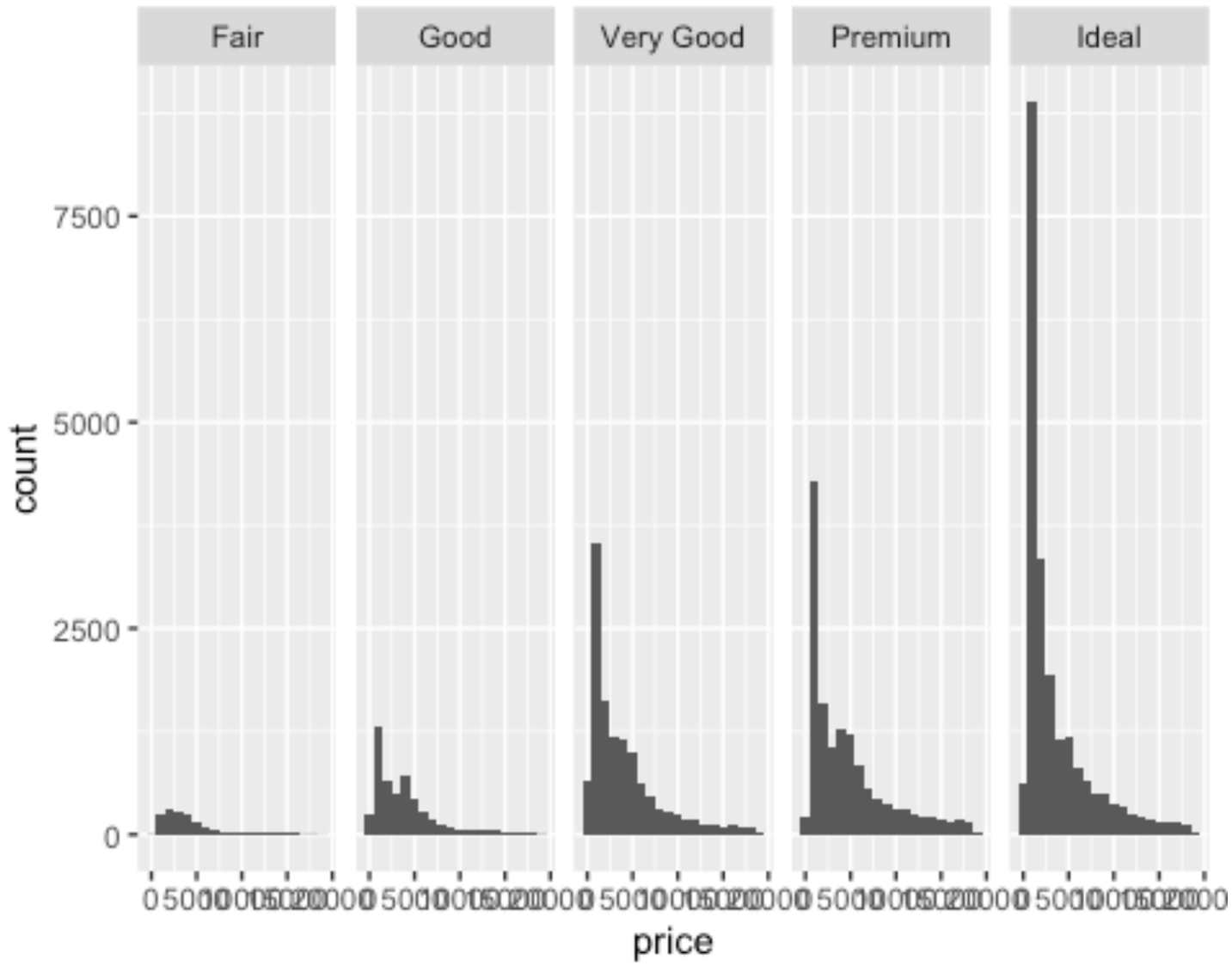
Graphical Summary -Histogram

```
ggplot(diamonds, aes(x =
carat))+
  geom_histogram(binwidth =
0.1)
```



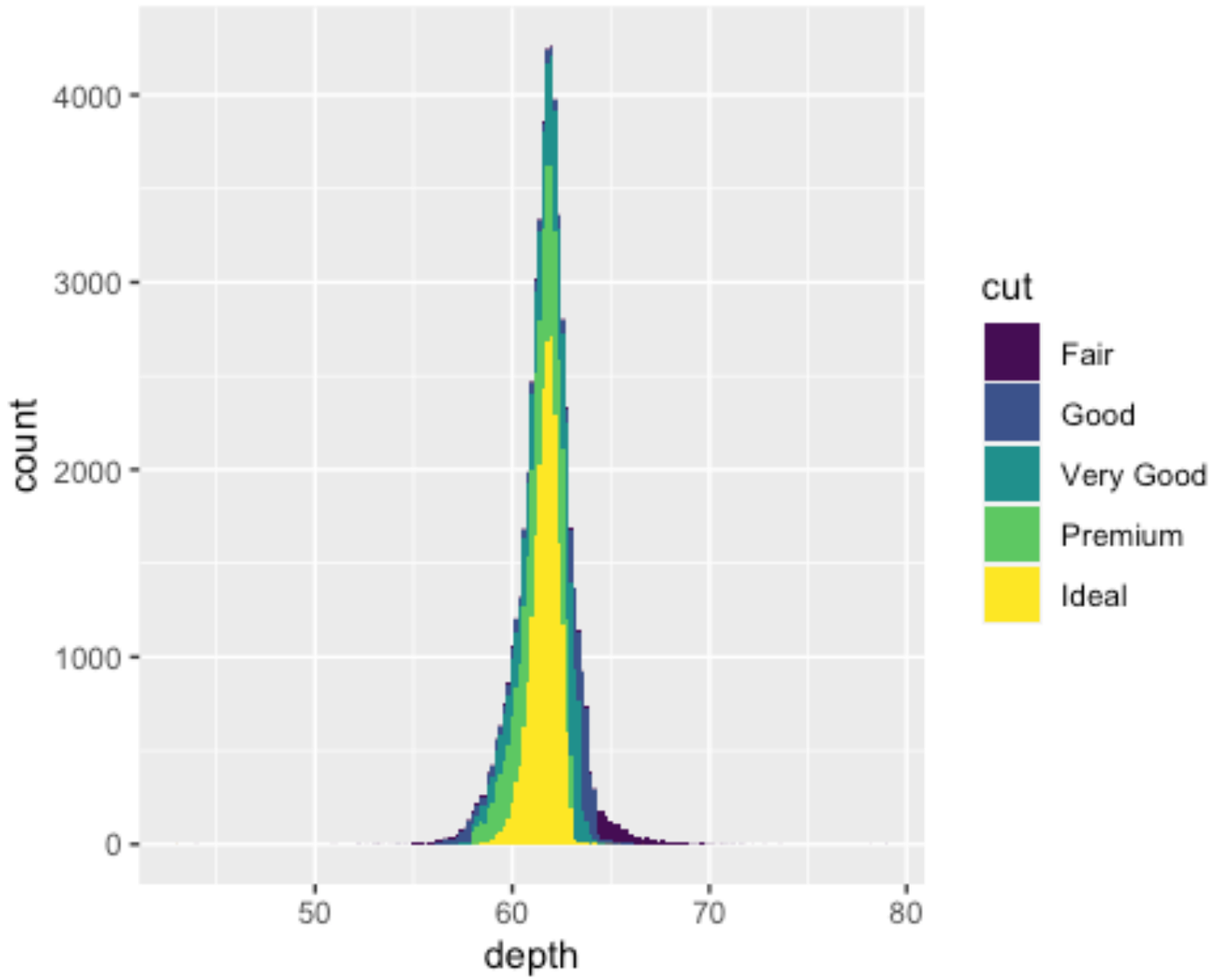
Graphical Summary -Histogram

```
ggplot(diamonds, aes(x=price))
+
  facet_grid(~cut)+
  geom_histogram(binwidth =
1000)
```



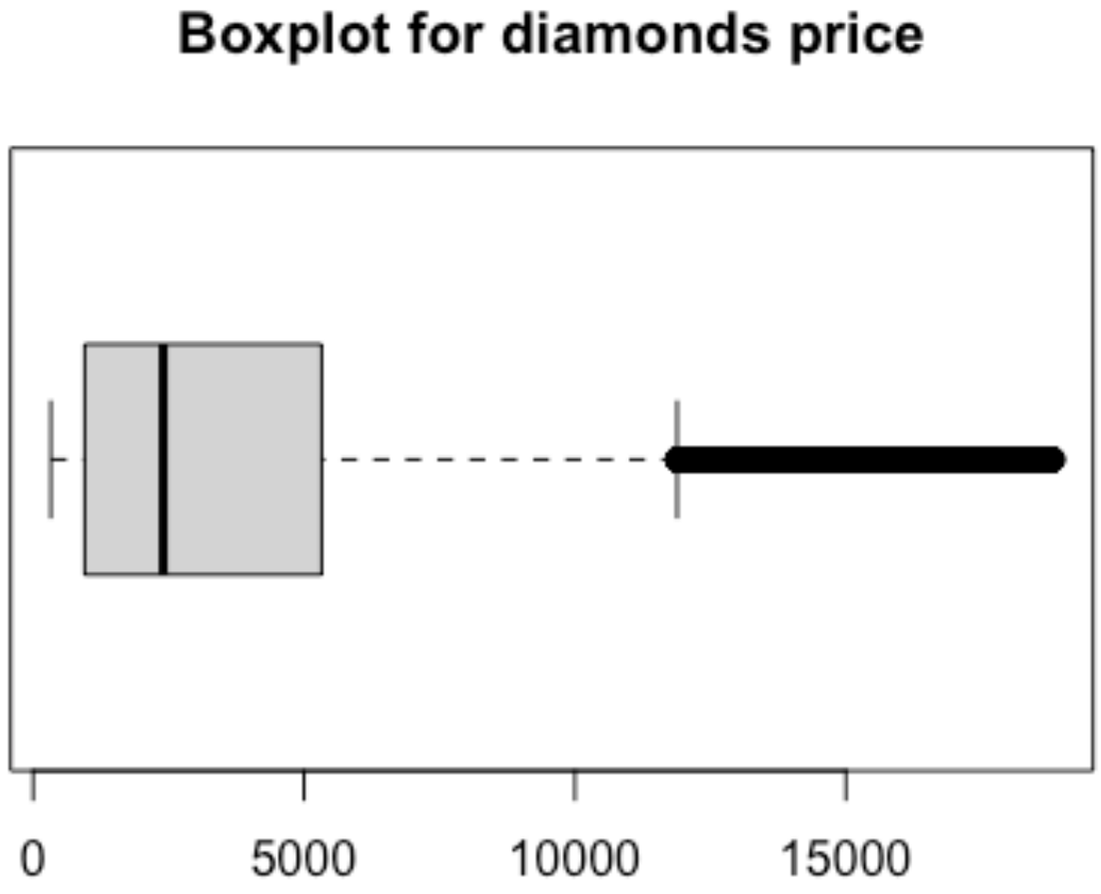
Graphical Summary -Histogram

```
ggplot(data = diamonds,
aes(x=depth, fill = cut))+
  geom_histogram(binwidth =
0.2)
```



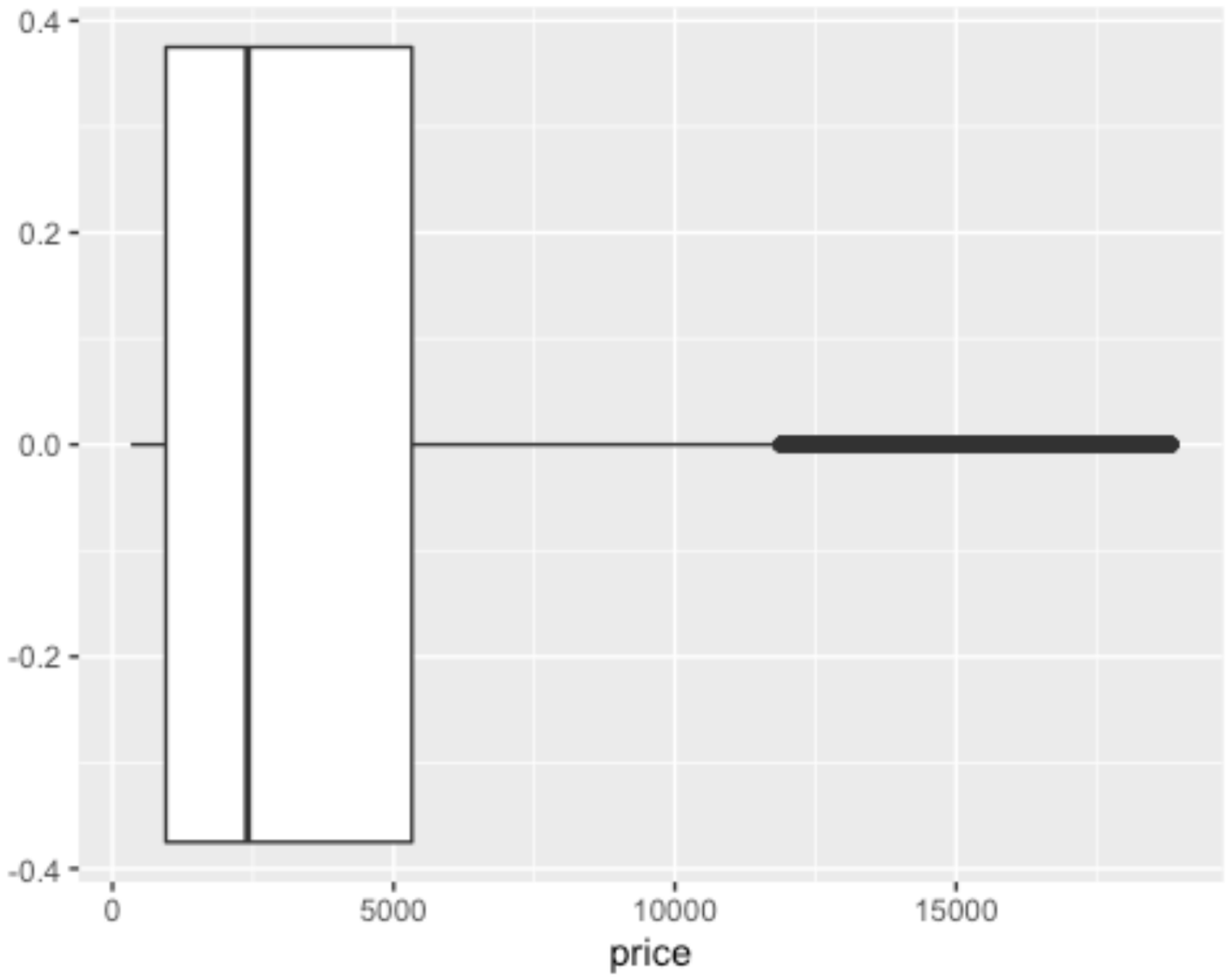
Graphical Summary -Boxplot

```
boxplot(diamonds$price,
horizontal= T, main = "Boxplot
for diamonds price")
```



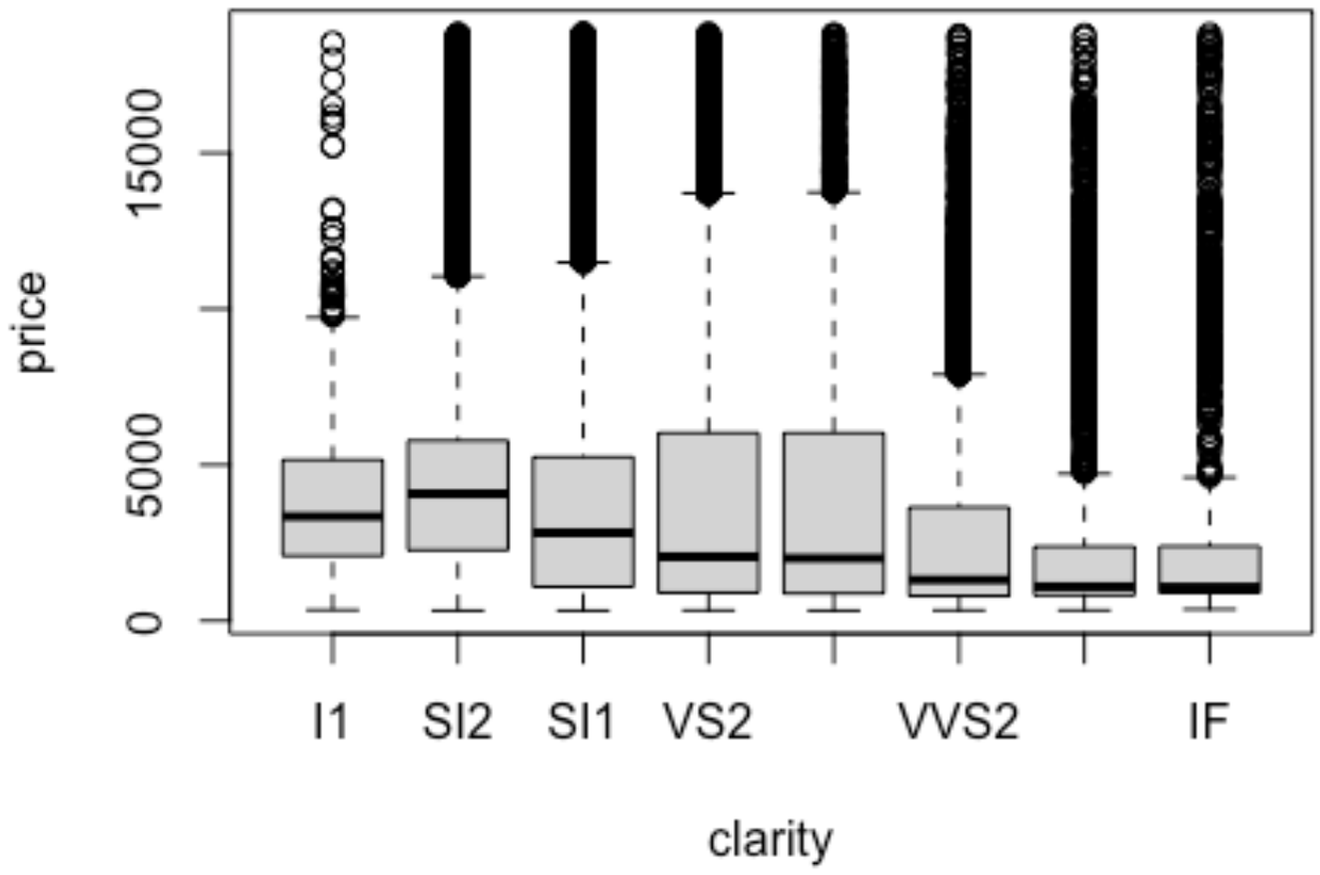
Graphical Summary - Boxplot using ggplot2

```
ggplot(diamonds, aes(x=price))
+
  geom_boxplot()
```



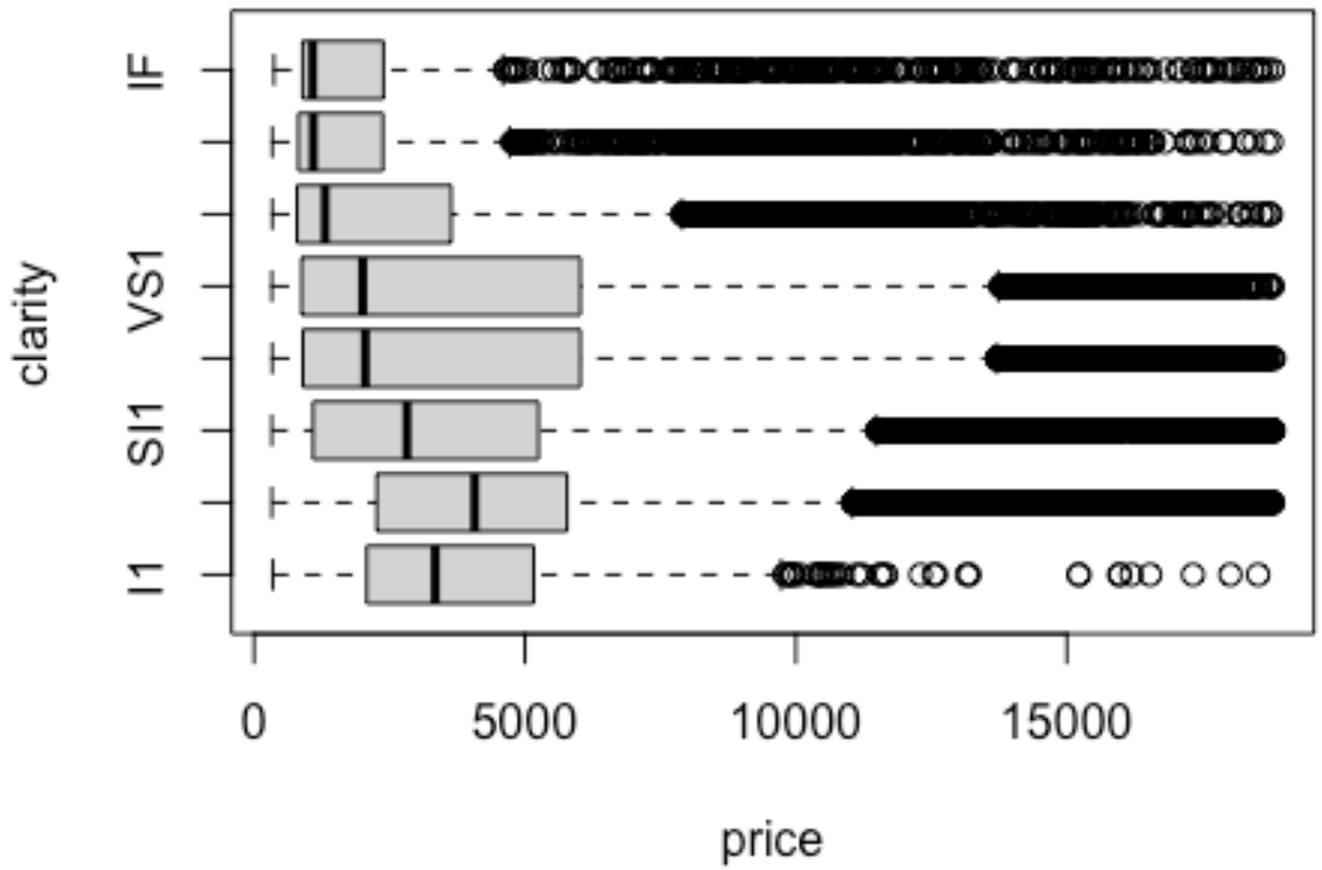
Graphical Summary - Boxplot to see the relationship among groups

```
boxplot(price~clarity, xlab =
"clarity",ylab="price", data =
diamonds)
```



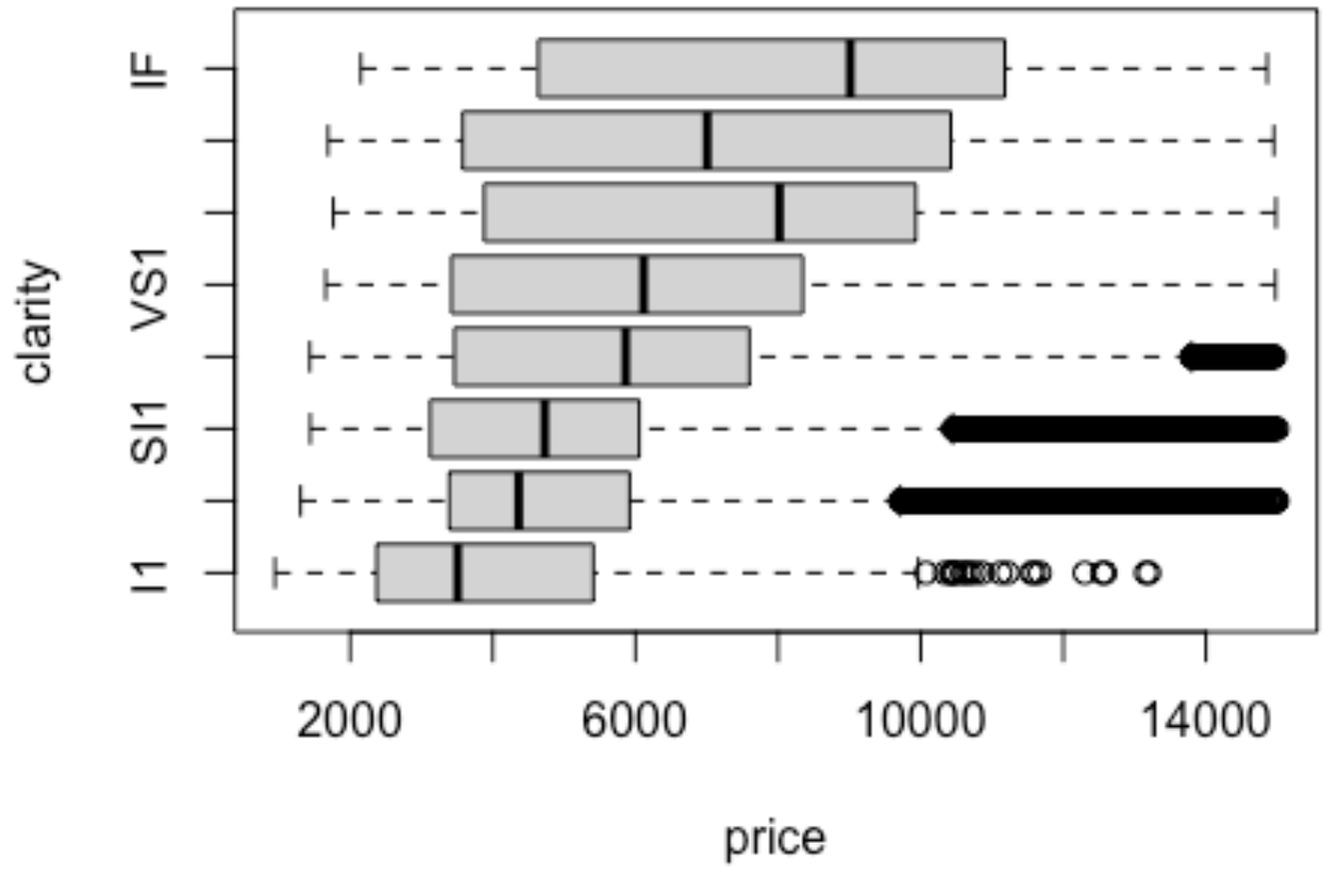
Graphical Summary - Boxplot to see the relationship among groups

```
boxplot(price~clarity, xlab =
"price",ylab="clarity",
horizontal = TRUE,
        data = diamonds)
```



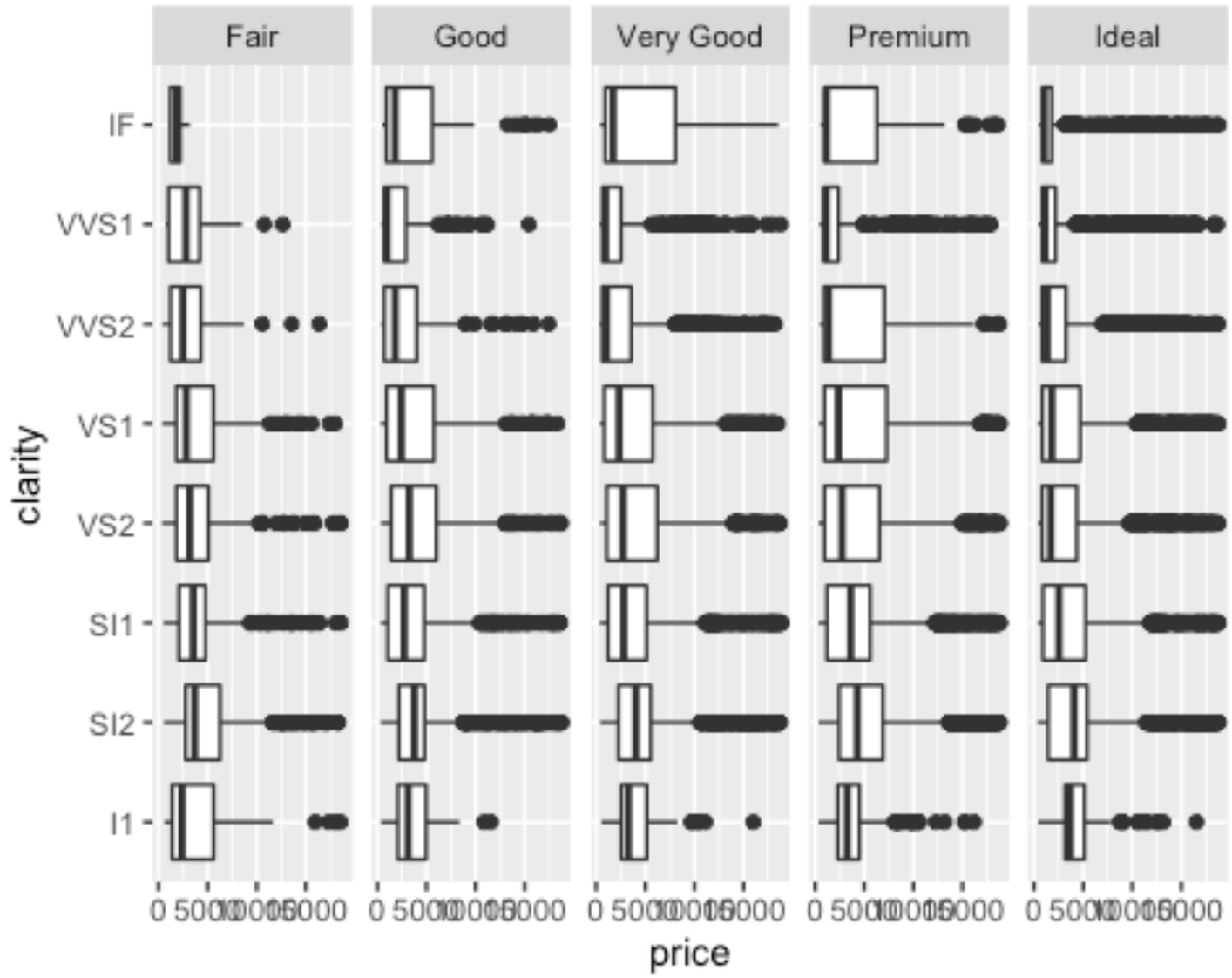
Graphical Summary - Boxplot to see the relationship among groups

```
boxplot(price~clarity, xlab =
"price",ylab="clarity",
horizontal = TRUE,
        data =
subset(diamonds, price<15000 &
carat >= 0.7))
```



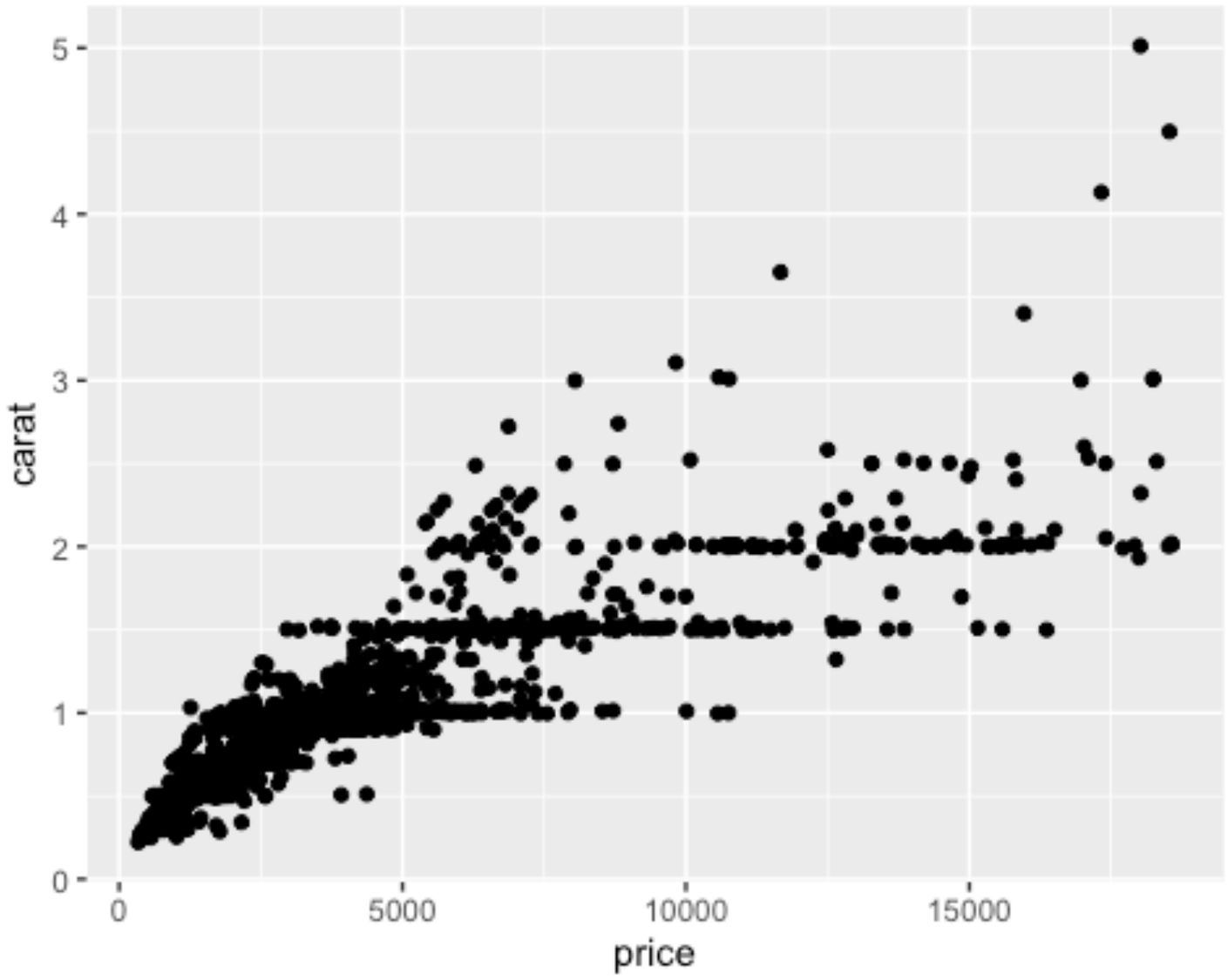
Graphical Summary - Boxplot to see the relationship among groups

```
ggplot(diamonds, aes(x= price,
y = clarity))+
  facet_grid(~cut)+
  geom_boxplot()
```



Graphical Summary - Scatterplot

```
diamonds %>%
  filter(cut == "Fair") %>%
  ggplot(aes(x = price, y =
carat)) +
  geom_point(position =
"jitter")
```



Exploring Categorical Data

Exploring Categorical data - Frequency Table

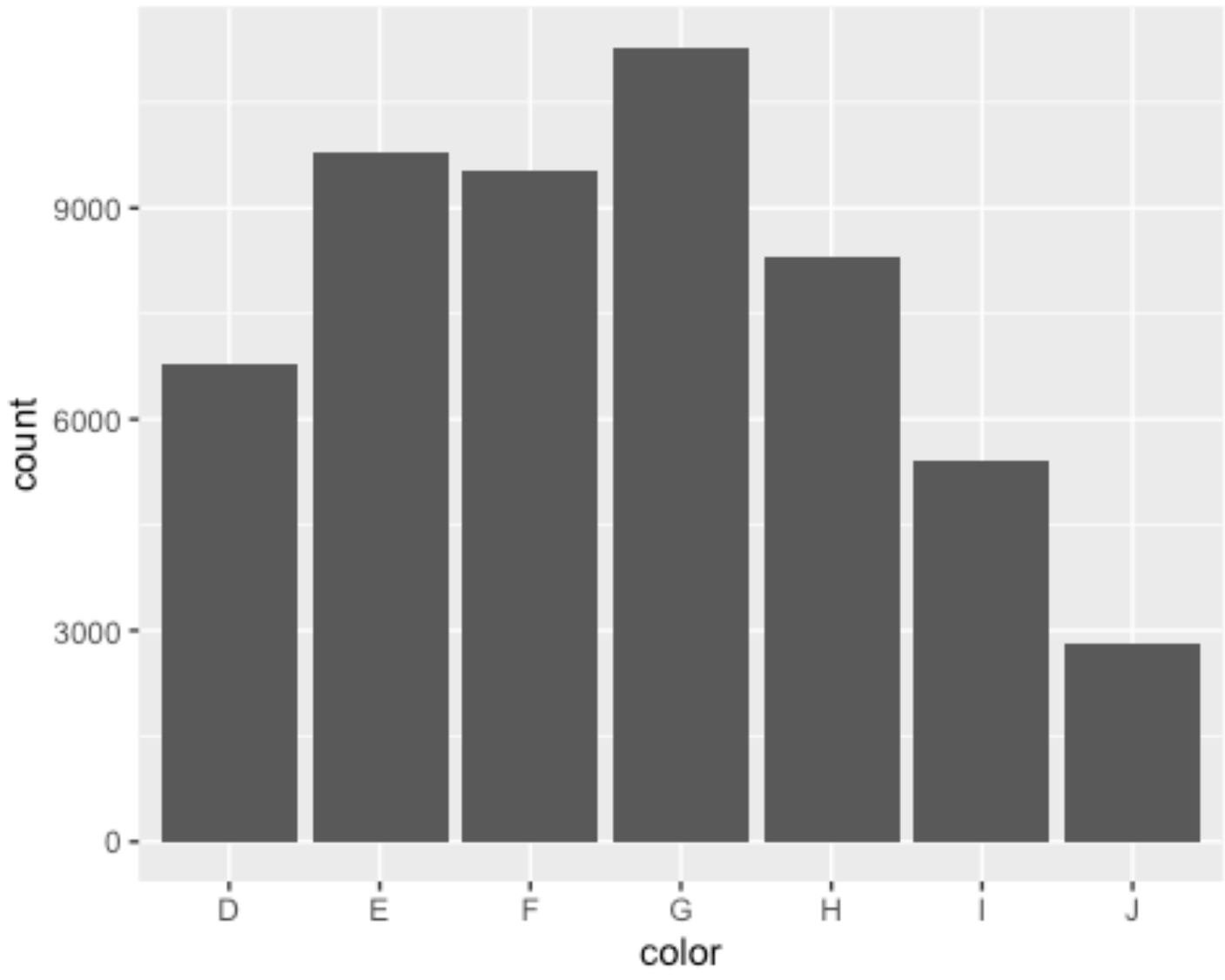
```
diamonds %>%
  group_by(cut) %>%
  summarise(counts = n(), proportions = n()/nrow(diamonds))
## # A tibble: 5 × 3
##   cut       counts proportions
##   <ord>      <int>       <dbl>
## 1 Fair        1610      0.0298
## 2 Good        4906      0.0910
## 3 Very Good  12082      0.224
## 4 Premium    13791      0.256
## 5 Ideal      21551      0.400
```

Categorical data - Contingency Table

```
xtabs(~cut+clarity, data = diamonds) %>%
  addmargins()
##             clarity
## cut            I1   SI2   SI1   VS2   VS1  VVS2  VVS1    IF   Sum
##   Fair        210   466   408   261   170    69    17     9  1610
##   Good         96  1081  1560   978   648   286   186    71  4906
##   Very Good    84  2100  3240  2591  1775  1235   789   268 12082
##   Premium     205  2949  3575  3357  1989   870   616   230 13791
##   Ideal       146  2598  4282  5071  3589  2606  2047  1212 21551
##   Sum         741  9194 13065 12258  8171  5066  3655  1790 53940
```

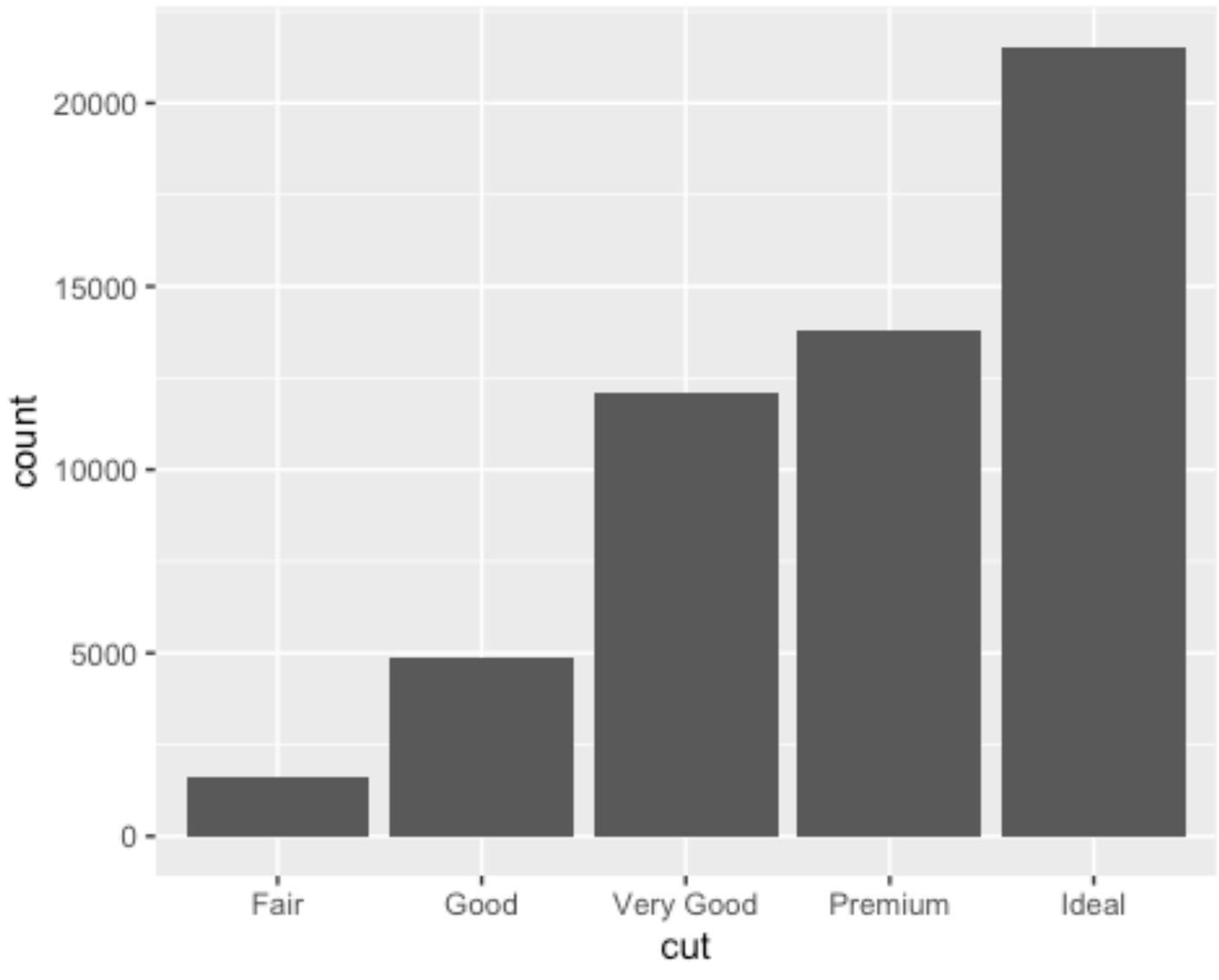
Graphical Summary - Barplot

```
ggplot(data = diamonds, aes(x
= color))+
  geom_bar()
```



Graphical Summary - Barplot

```
ggplot(data = diamonds, aes(x
= cut))+
  geom_bar()
```

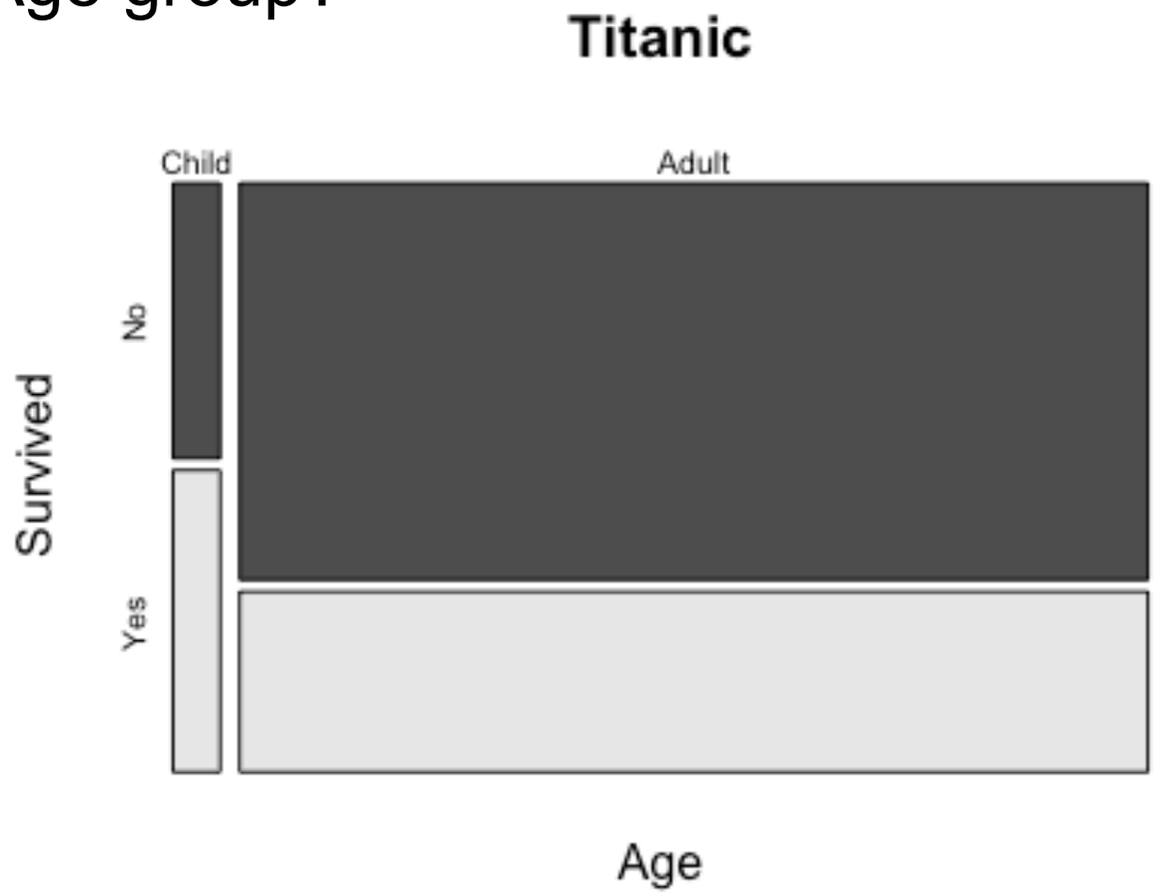


Mosaic Plot: `Titanic` data

Does the survival rates differ by Age group?

```
xtabs(~Age+Survived, data =
Titanic)
##         Survived
## Age      No Yes
##   Child  8   8
##   Adult  8   8
```

- `mosaicplot(~ Age + Survived, data = Titanic, color = TRUE)`



Review is important!

- Please review R code and play around — feel free to bring any questions you might want to solve by observing the dataset to office hour!
- I highly recommend you to follow the readings from the course archive
- Before Thursday's lecture, practice **Sec 3.1.1 Introductory examples**
- Office hour from **7pm** via Zoom.
  - **This week, make sure to participate at least one office hour!!**
  - 5 out of 50 participation pts