Exploratory Analysis

2022-10-12

Mujahid Akber Ali

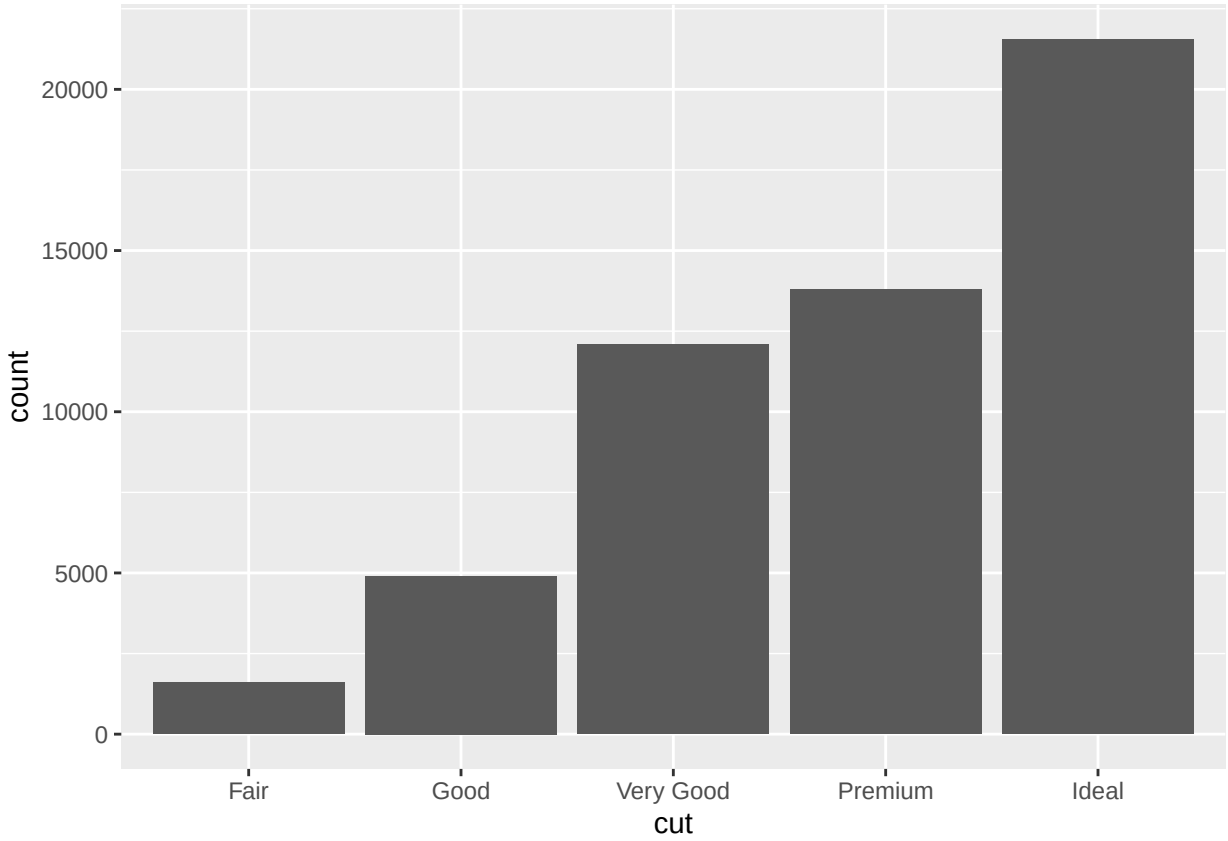
Importing Library

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Visualizing distribution

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



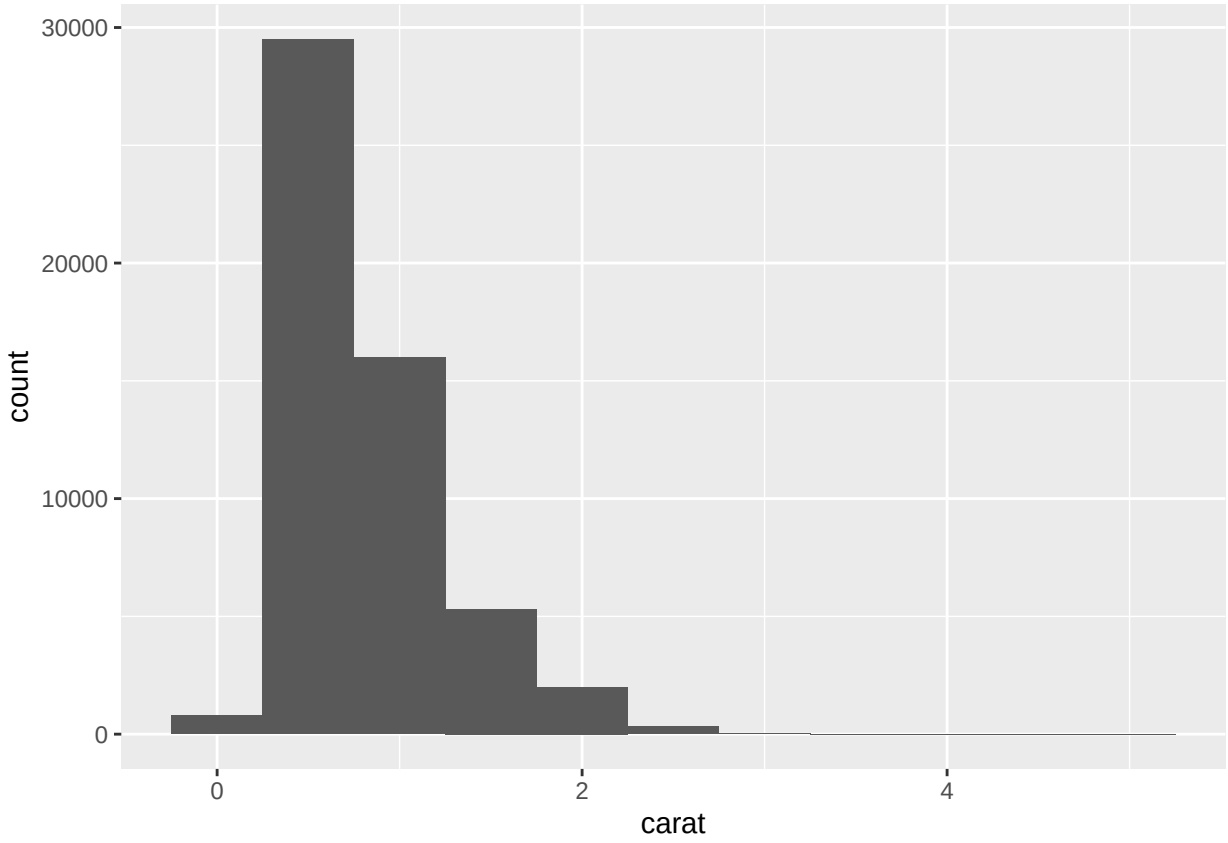
Counting diamond Cuts

```
library(tidyverse)
diamonds %>% count(cut)
```

```
## # A tibble: 5 x 2
##   cut           n
##   <ord>     <int>
## 1 Fair       1610
## 2 Good       4906
## 3 Very Good 12082
## 4 Premium   13791
## 5 Ideal     21551
```

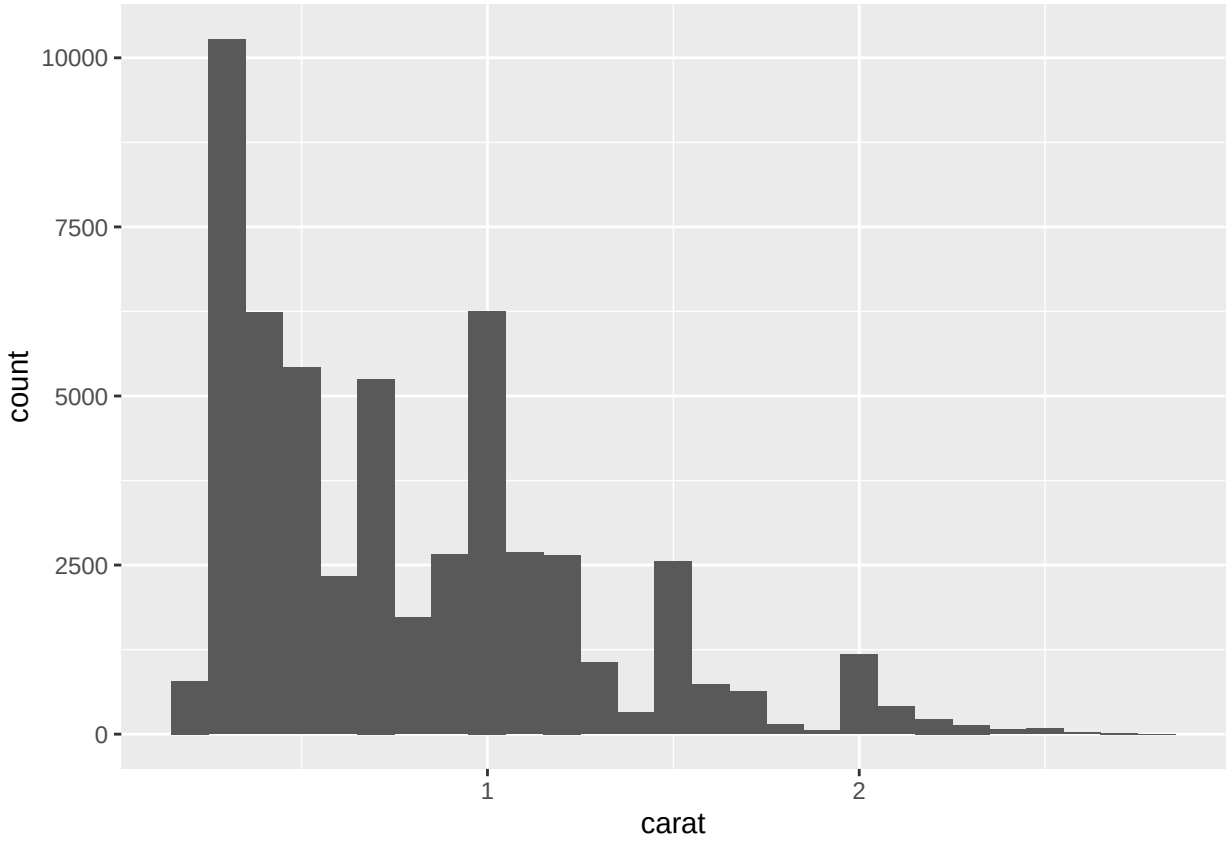
Continuous chart

```
ggplot(data = diamonds) + geom_histogram(mapping = aes(x = carat), binwidth = 0.5)
```



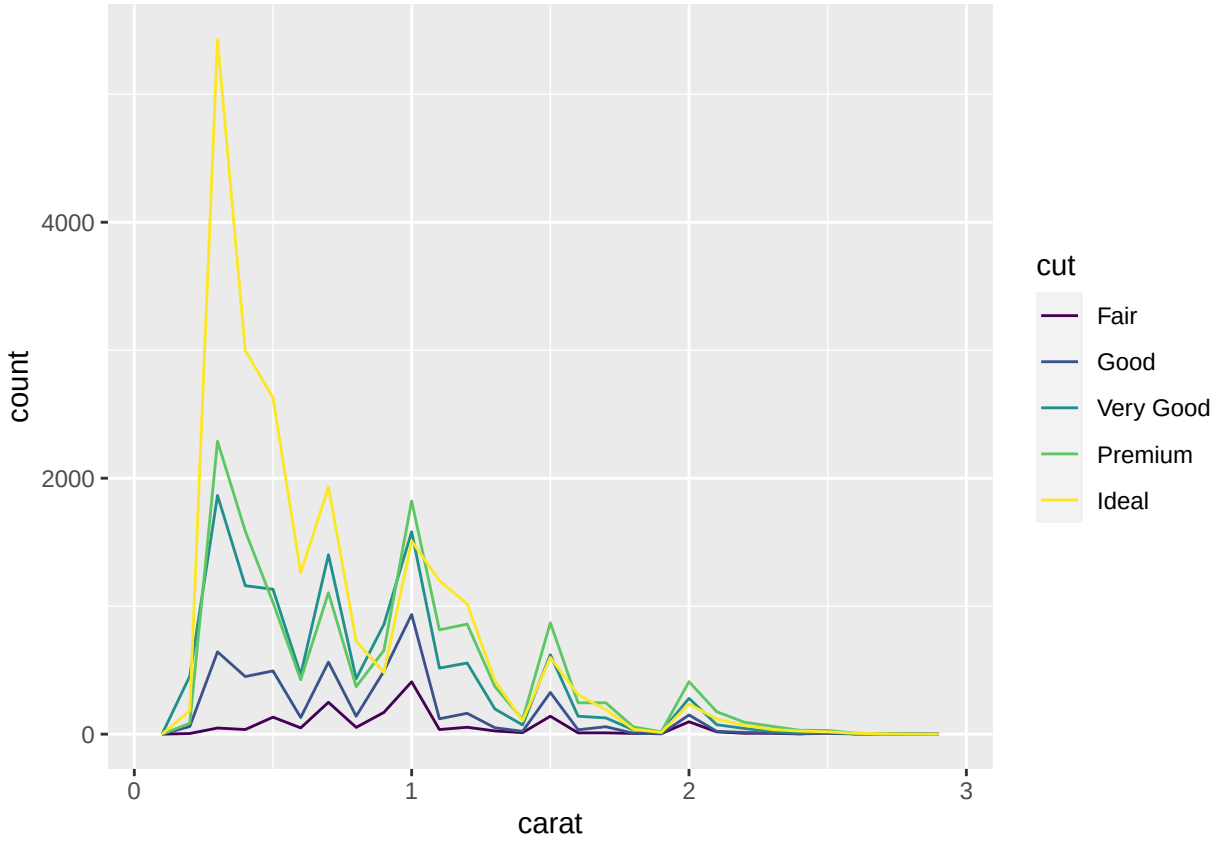
Filter carat Less than 3 and Plotting

```
smaller <- diamonds %>% filter(carat < 3)
ggplot(data = smaller, mapping = aes(x = carat)) + geom_histogram(binwidth = 0.1)
```



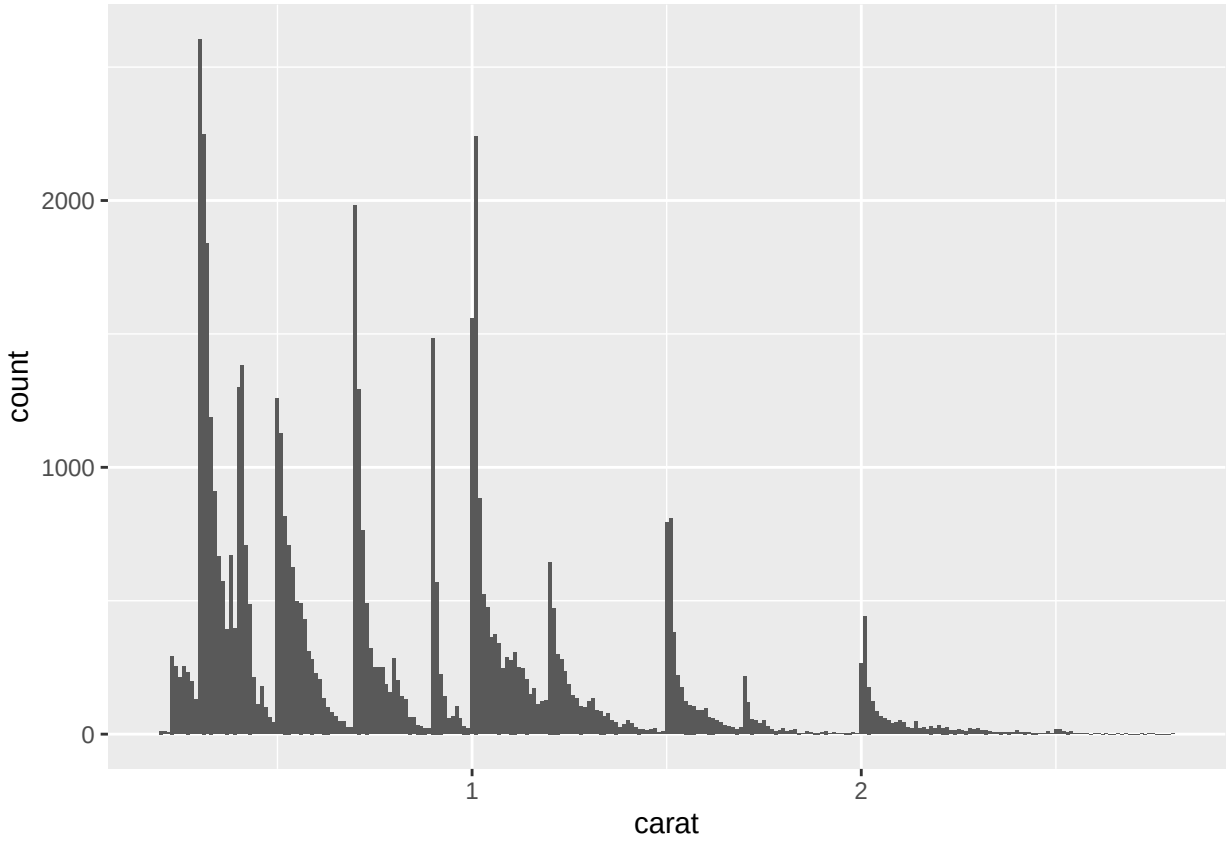
Frquency Plot with Legends

```
ggplot(data = smaller, mapping = aes(x = carat, colour = cut)) + geom_freqpoly(binwidth = 0.1)
```



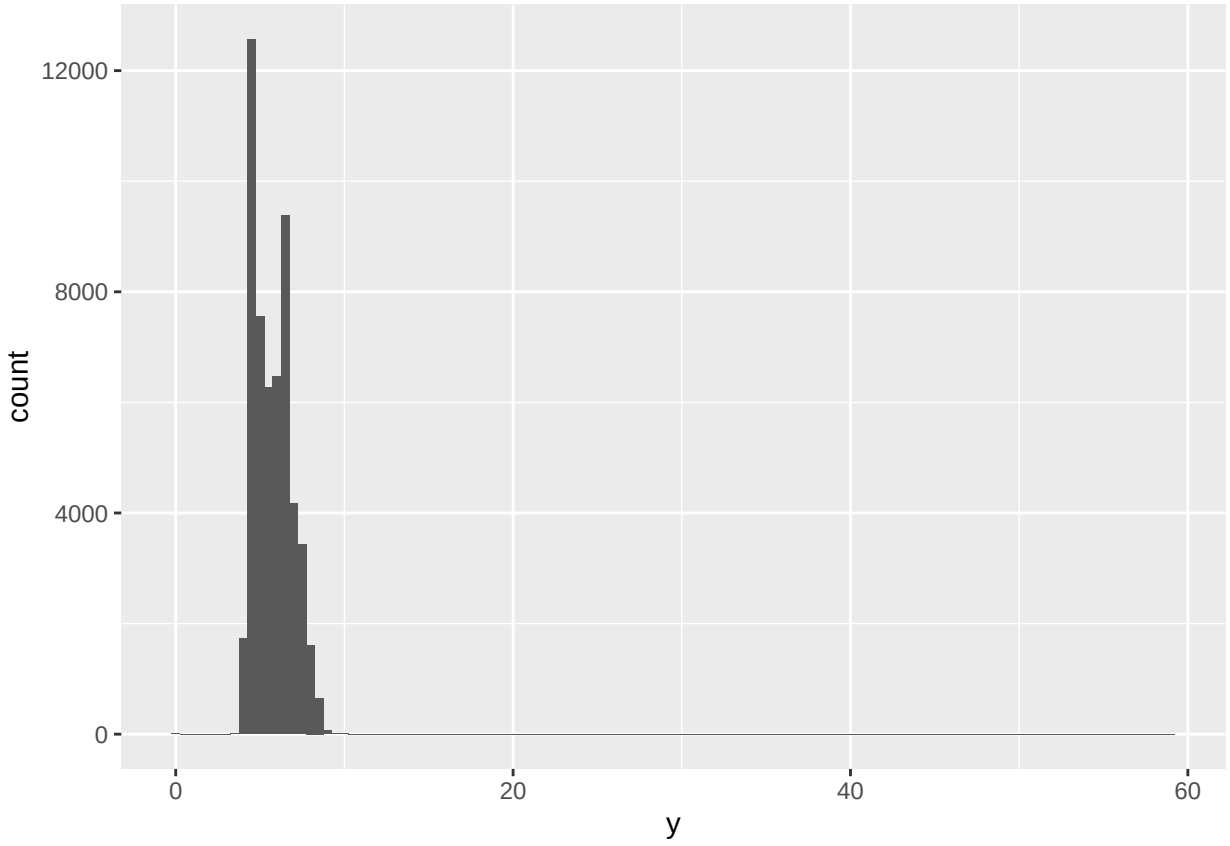
Histogram

```
ggplot(data = smaller, mapping = aes(x = carat)) + geom_histogram(binwidth = 0.01)
```



Plotting Outliers

```
ggplot(diamonds) + geom_histogram(mapping = aes(x = y), binwidth = 0.5)
```



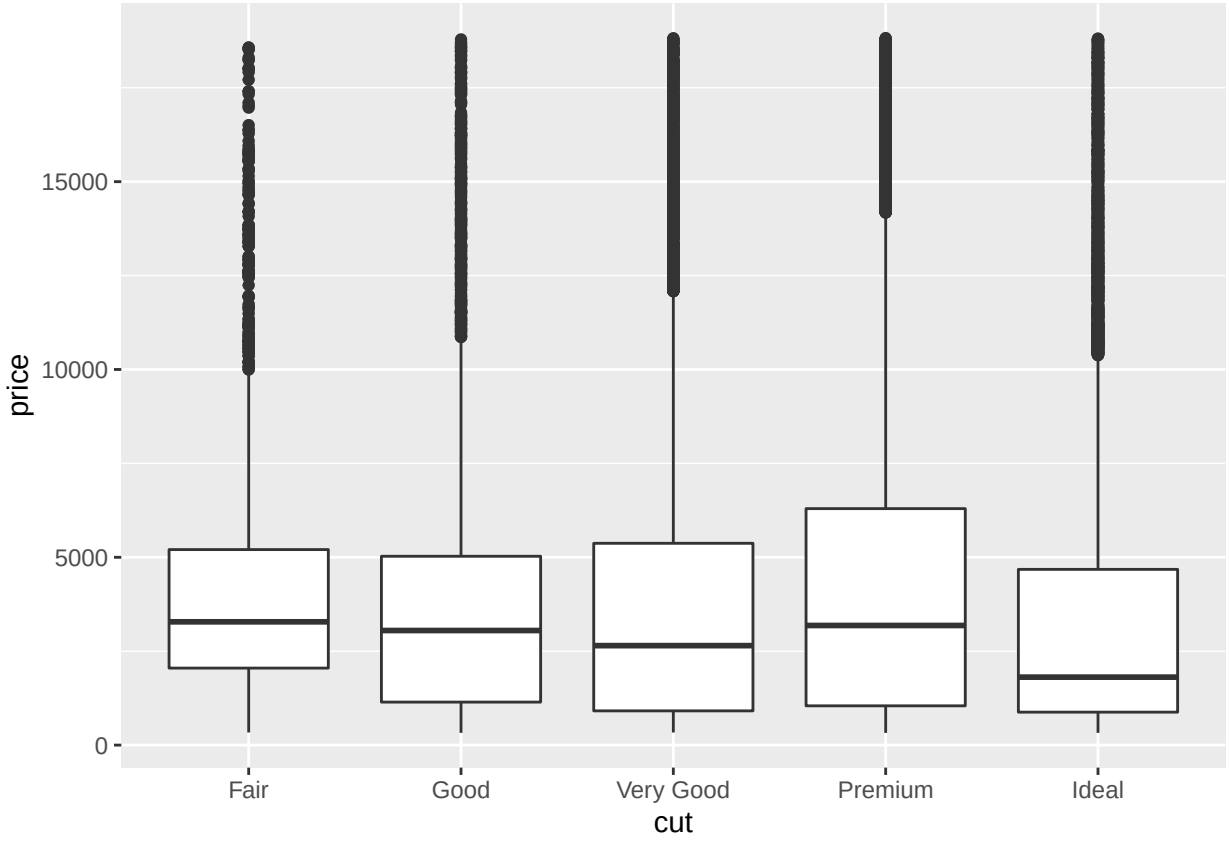
Filtering

```
diamonds %>% filter(between(y, 3, 20))
```

```
## # A tibble: 53,931 x 10
##    carat cut       color clarity depth table price     x     y     z
##    <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
##  2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
##  3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
##  4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
##  5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
##  6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
##  7  0.24 Very Good I     VVS1     62.3    57   336  3.95  3.98  2.47
##  8  0.26 Very Good H     SI1      61.9    55   337  4.07  4.11  2.53
##  9  0.22 Fair      E     VS2      65.1    61   337  3.87  3.78  2.49
## 10  0.23 Very Good H     VS1      59.4    61   338  4     4.05  2.39
## # ... with 53,921 more rows
```

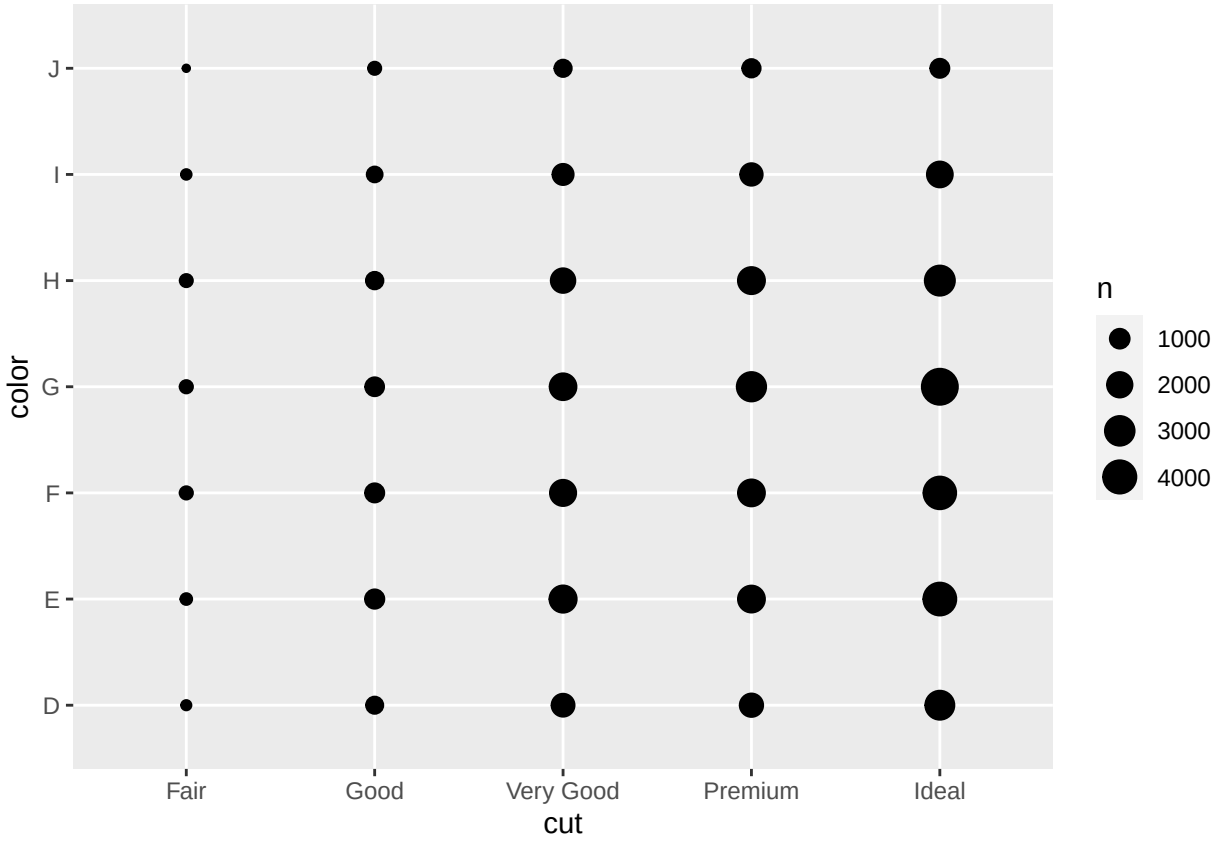
Box Plot

```
ggplot(data = diamonds, mapping = aes(x = cut, y = price)) + geom_boxplot()
```



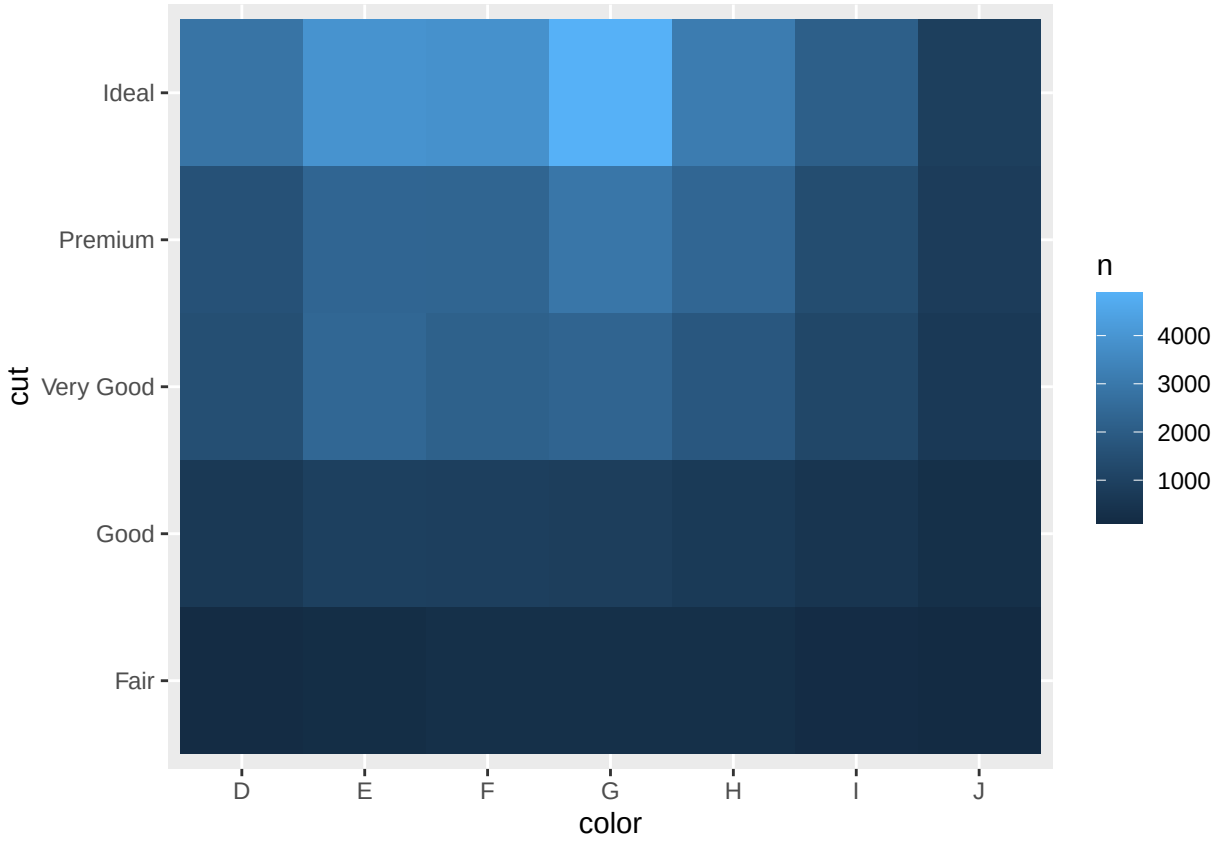
Geomatrical counting

```
ggplot(data = diamonds) + geom_count(mapping = aes(x = cut, y = color))
```



Color Mapping

```
diamonds %>% count(color, cut) %>% ggplot(mapping = aes(x = color, y = cut)) + geom_tile(mapping = aes(
```


Ideal
Premium
Very Good
Good
Fair
cut
D
E
F
G
H
I
J
color
n
4000
3000
2000
1000