Exploratory Analysis Of Diamonds Dataset

Cody Collie-Szach

2022-06-24

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(modelr)
```

Introduction

In this report I will be conducting an Exploratory Data Analysis. Specifically, data visualization and transformation to explore the `dimonds` data set.

1. What type of variation occurs within my variables?
2. What type of covariation occurs within my variables?

```
glimpse(diamonds)
```

```
## Rows: 53,940
## Columns: 10
## $ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, 0.23, 0.~
## $ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very Good, Ver~
## $ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, J, J, I,~
## $ clarity <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, SI1, VS1, ~
## $ depth   <dbl> 61.5, 59.8, 56.9, 62.4, 63.3, 62.8, 62.3, 61.9, 65.1, 59.4, 64~
## $ table   <dbl> 55, 61, 65, 58, 58, 57, 57, 55, 61, 61, 55, 56, 61, 54, 62, 58~
## $ price   <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339, 340, 34~
## $ x       <dbl> 3.95, 3.89, 4.05, 4.20, 4.34, 3.94, 3.95, 4.07, 3.87, 4.00, 4.~
## $ y       <dbl> 3.98, 3.84, 4.07, 4.23, 4.35, 3.96, 3.98, 4.11, 3.78, 4.05, 4.~
## $ z       <dbl> 2.43, 2.31, 2.31, 2.63, 2.75, 2.48, 2.47, 2.53, 2.49, 2.39, 2.~
```

Variation

- Categorical Variables

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

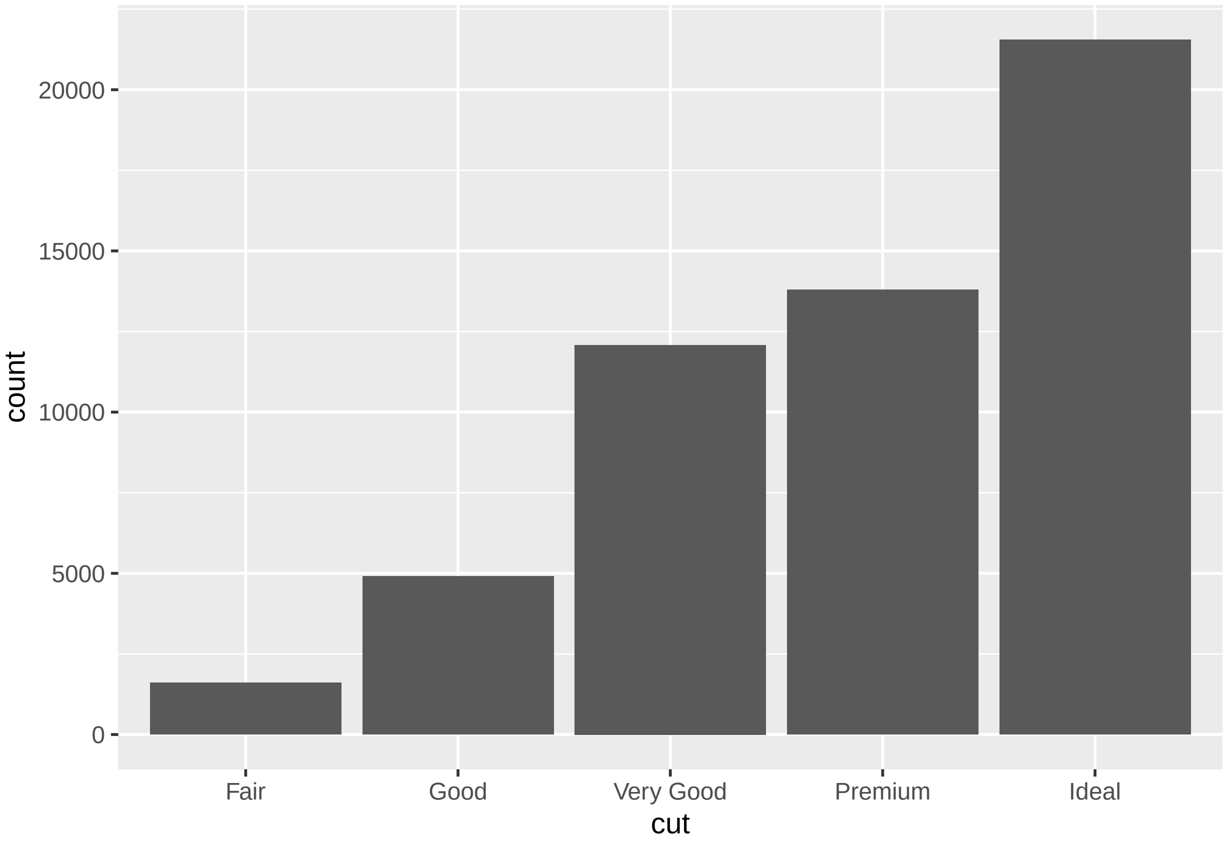


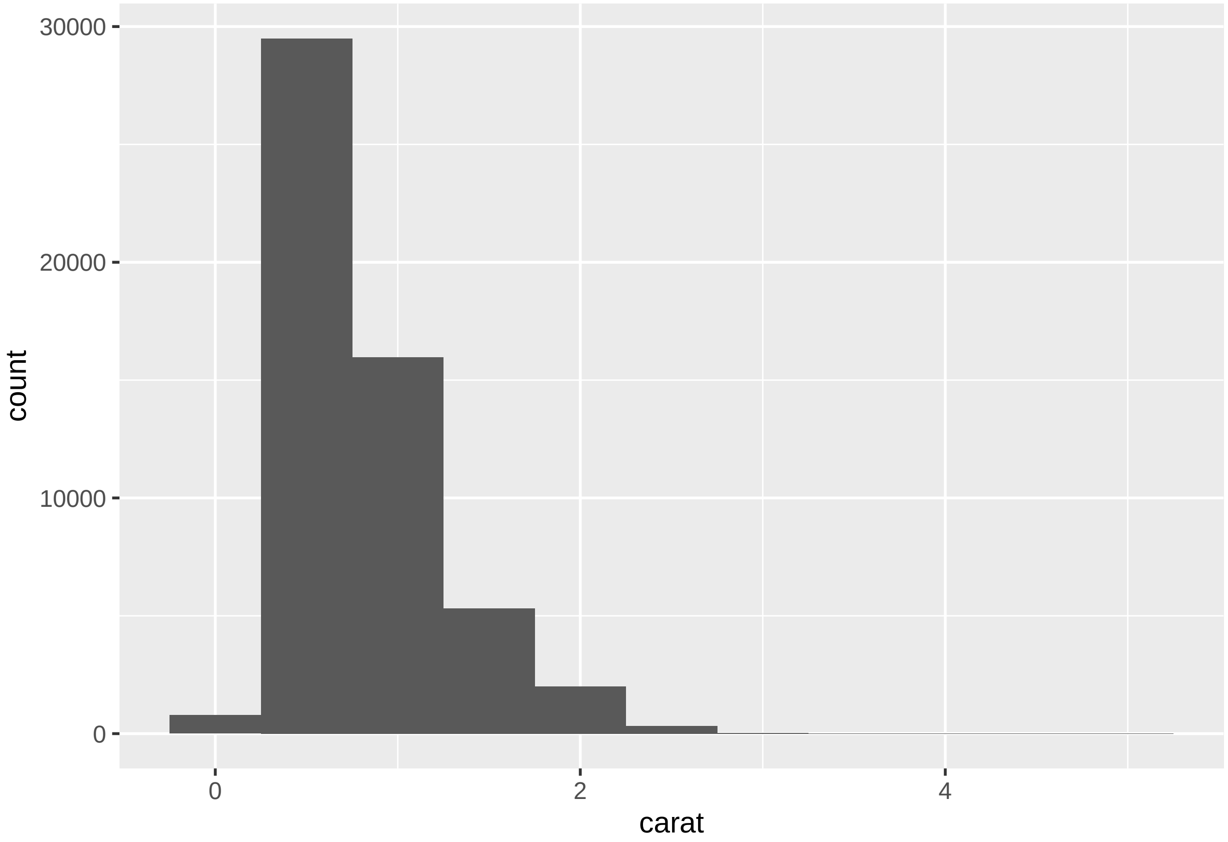
Figure 1: Bar Chart Looking at Diamond Cut

```
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 Variables

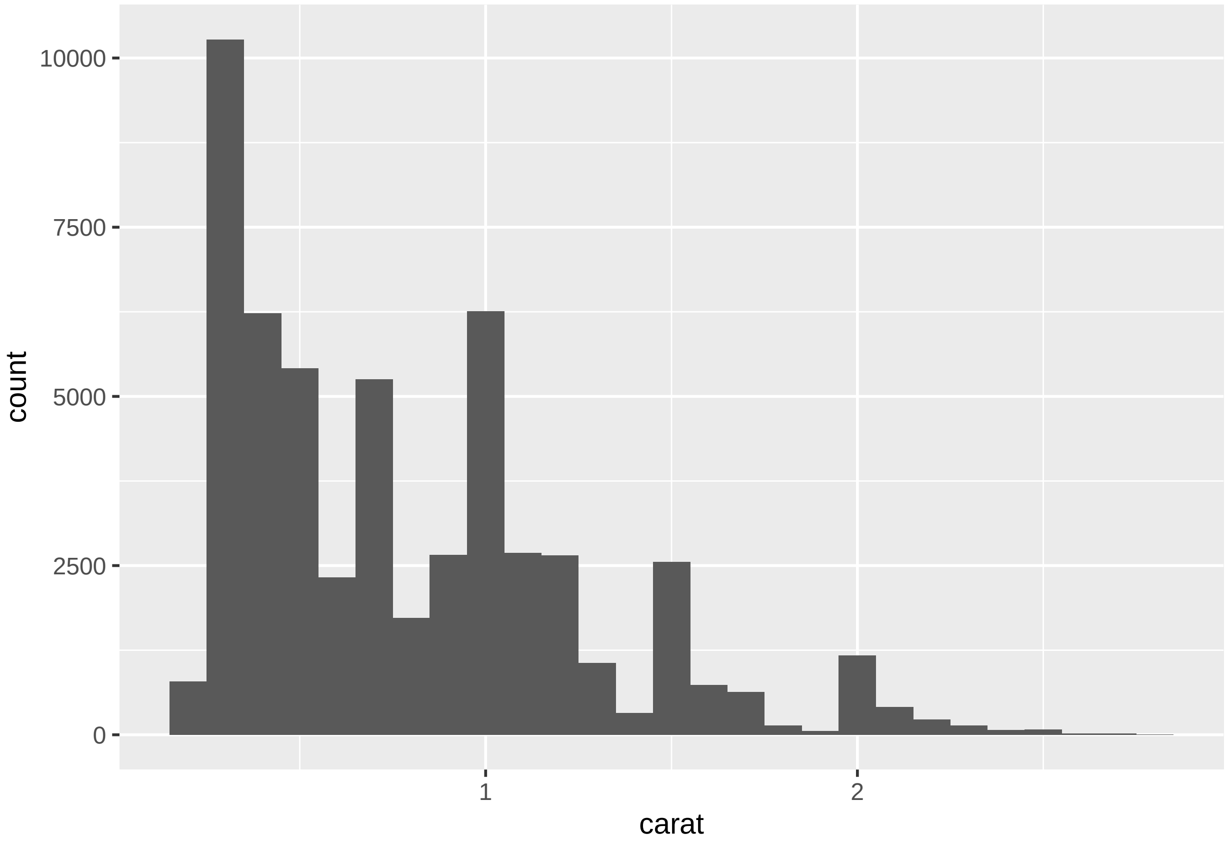
```
ggplot(diamonds) + geom_histogram(aes(x = carat), binwidth = 0.5)
```



Data Cleaning

```
small <- diamonds %>% filter(carat < 3)
```

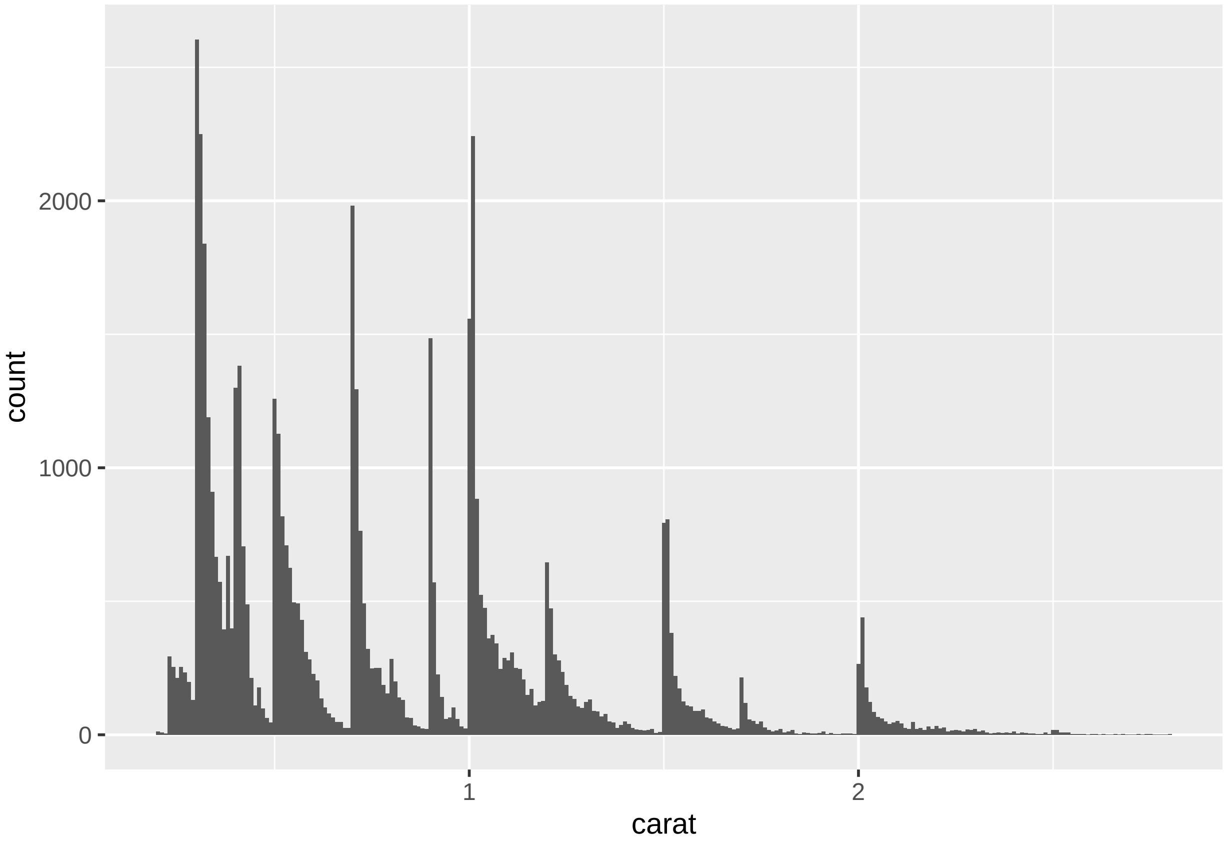
```
ggplot(small, aes(x = carat)) + geom_histogram(binwidth = 0.1)
```



Typical Values

1. Which values are the most common? Why?
2. Which values are rare? Why? Does that match our expectaions?
3. Can you see any unusual patterns? What might explain them?

```
ggplot(small, aes(x = carat)) + geom_histogram(binwidth = 0.01)
```

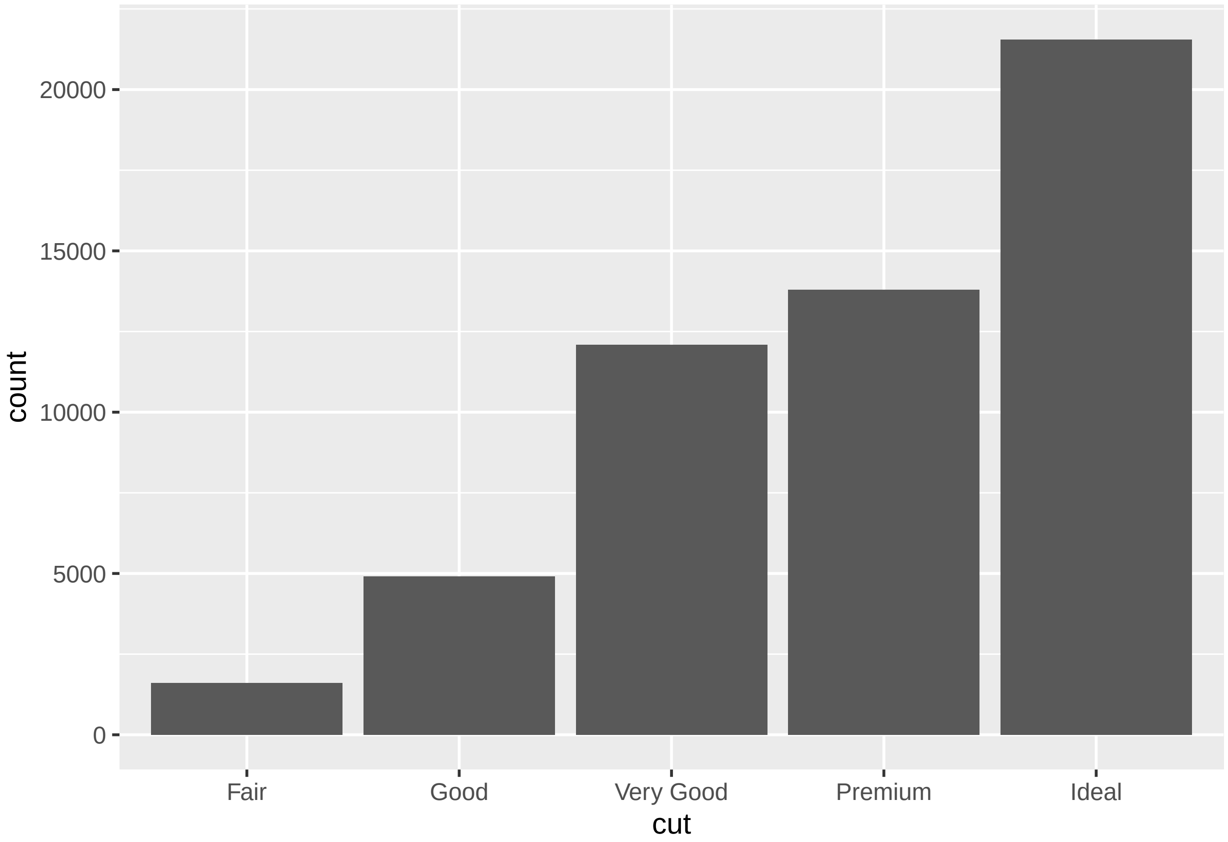


Questions that need to be answered - Why are more diamonds at whole carats? - Why are there more diamonds that are more to the right of the common carats?

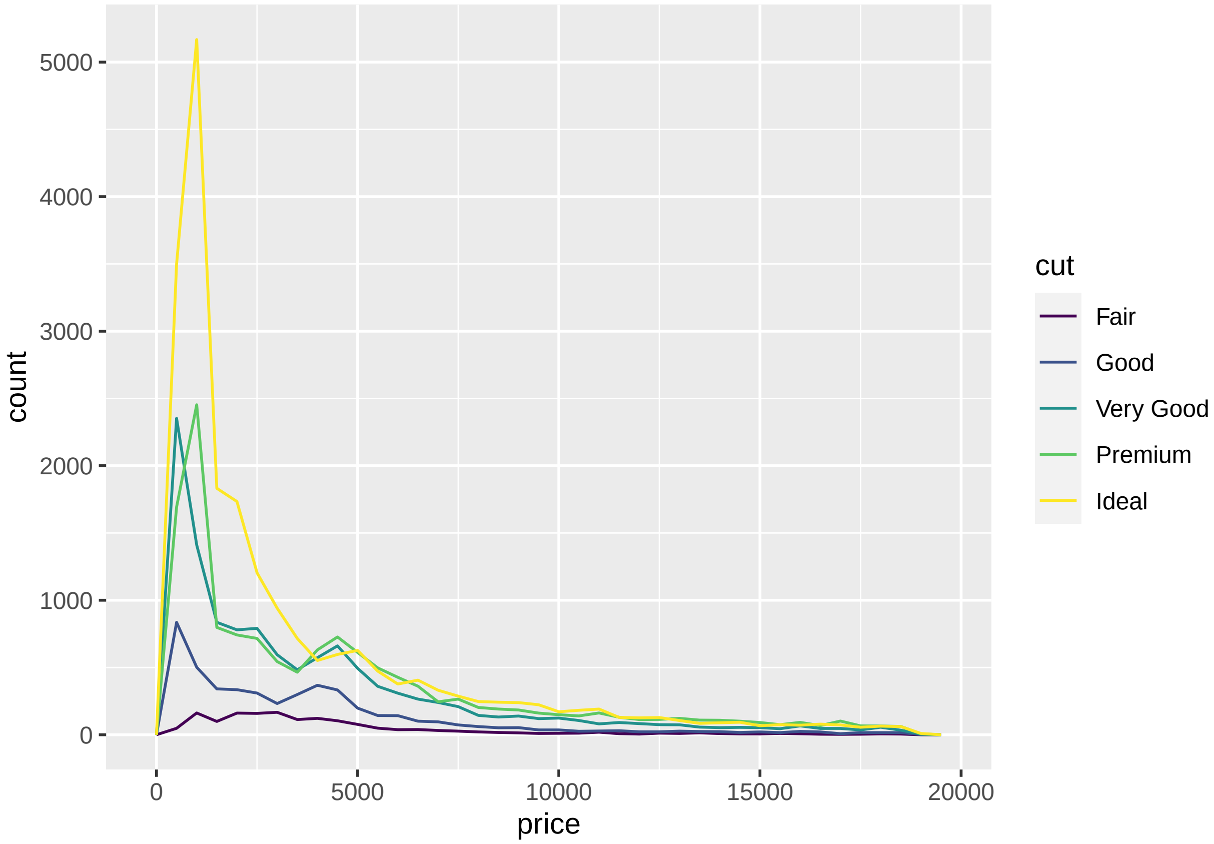
Covariation

Categorical vs Continuous

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

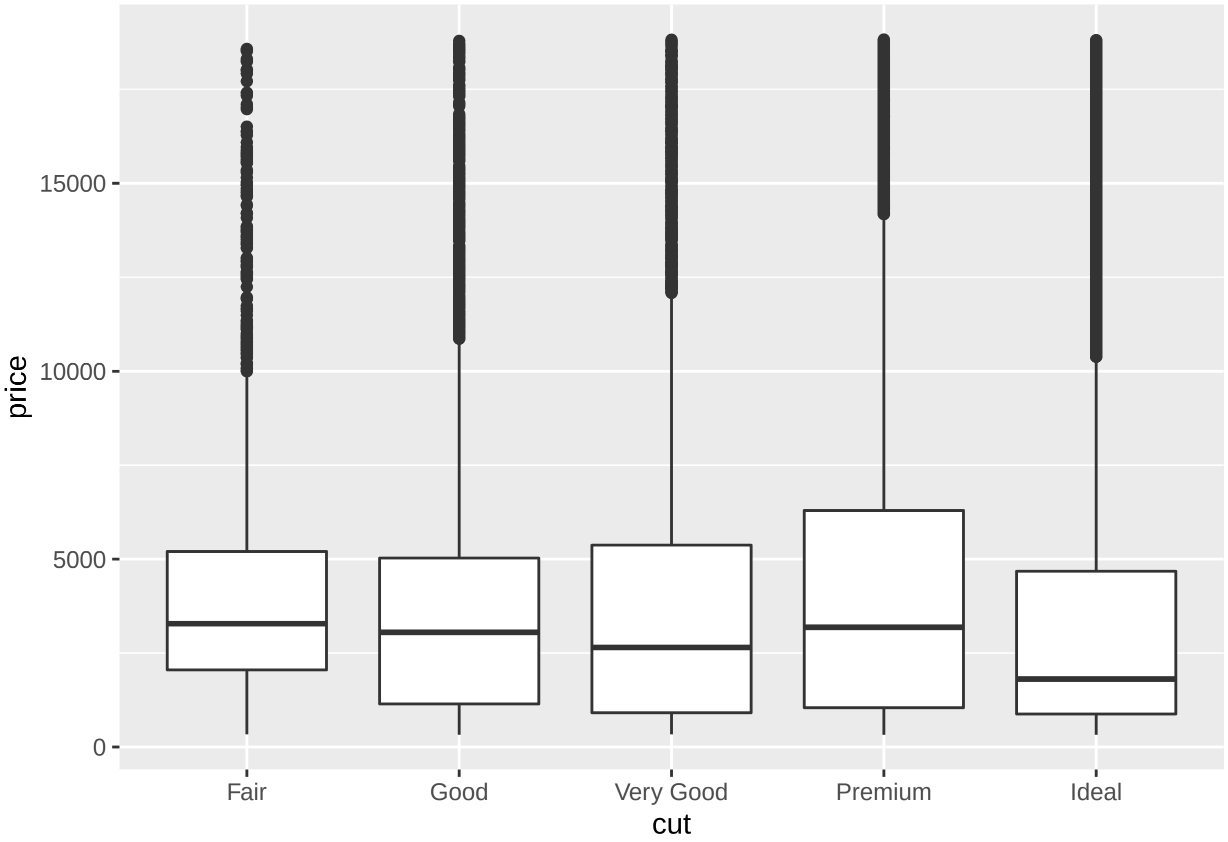


```
ggplot(diamonds, aes(x = price)) + geom_freqpoly(aes(color = cut), binwidth = 500)
```



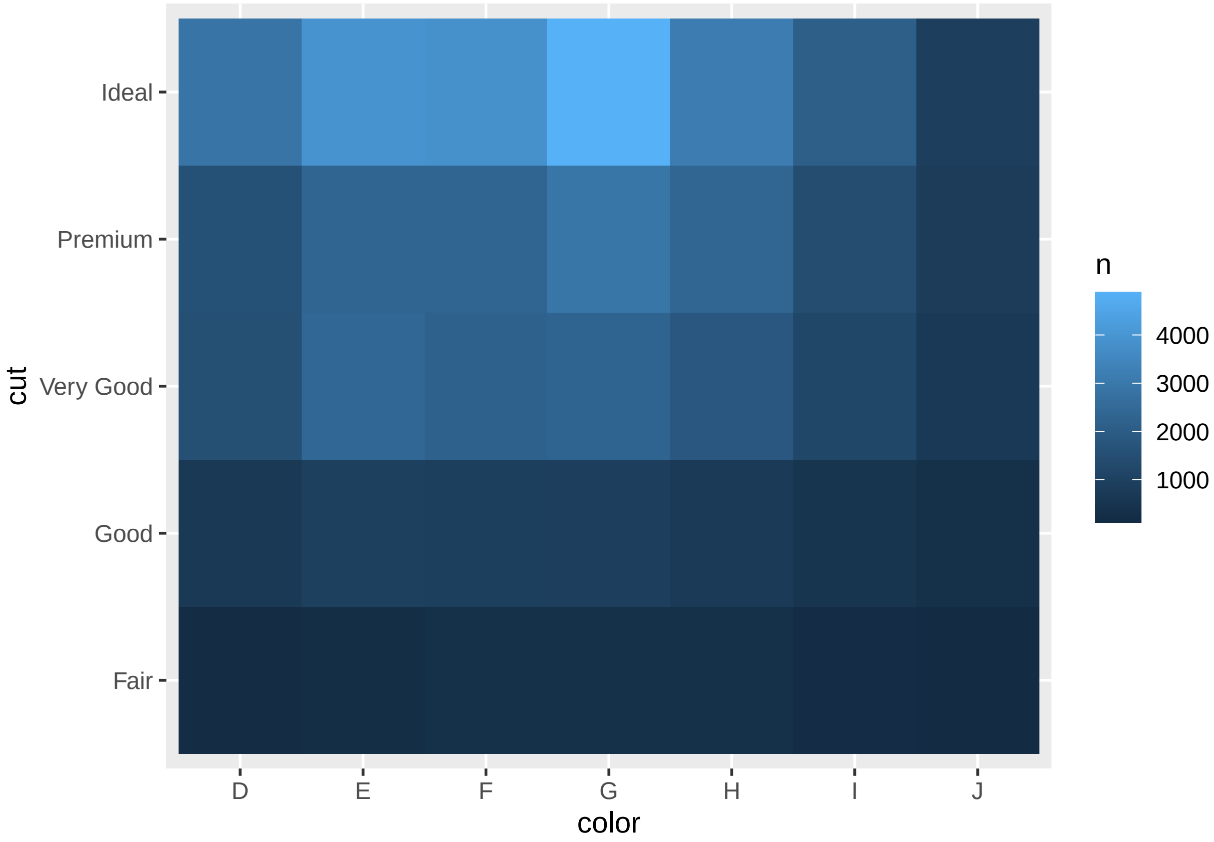
-Boxplot

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



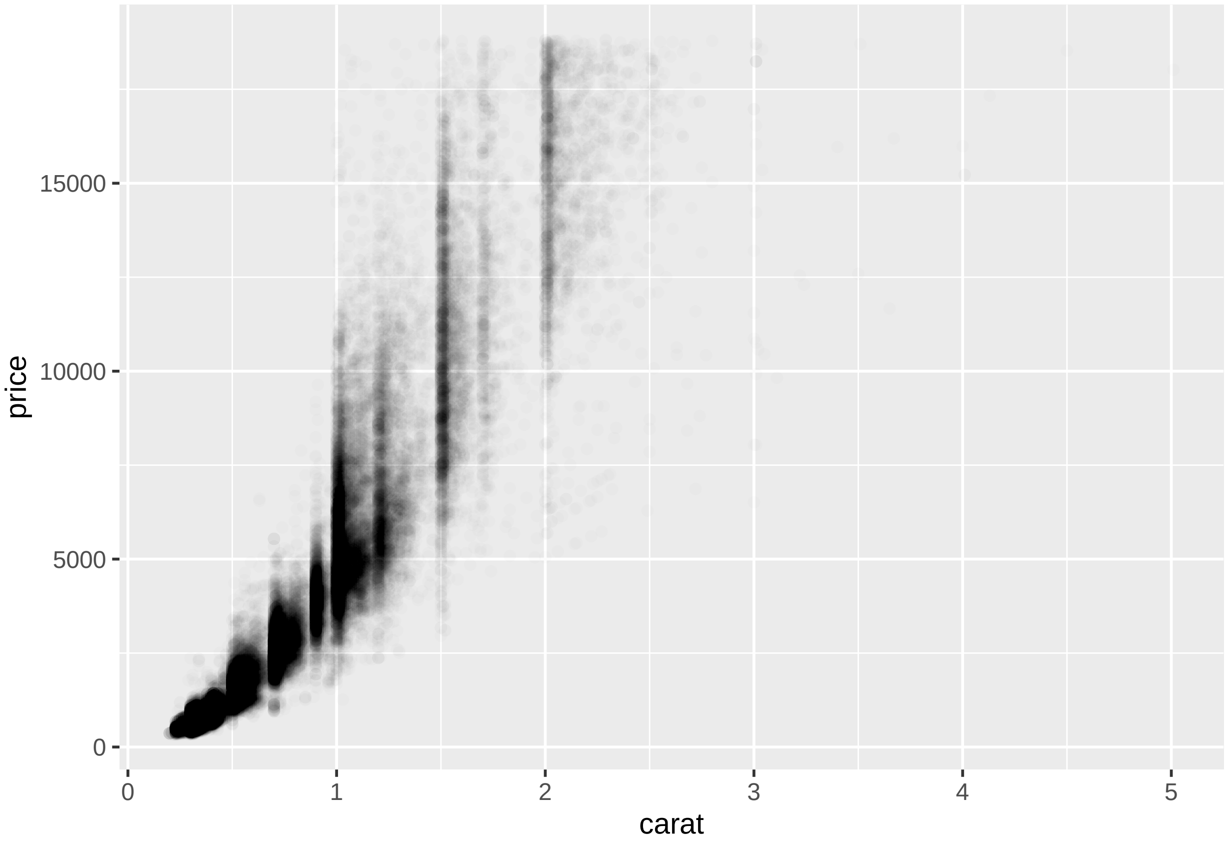
Categorical vs Categorical

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

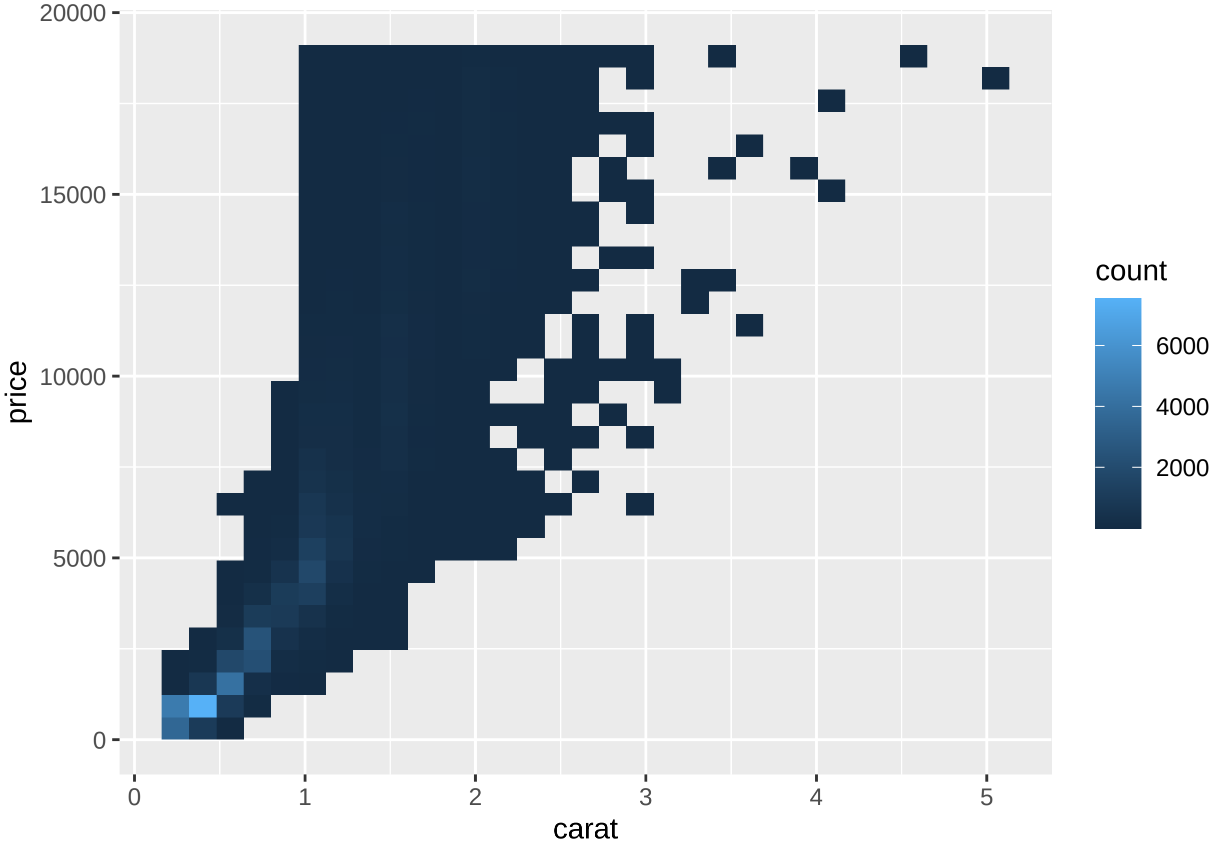


Continuous vs Continuous

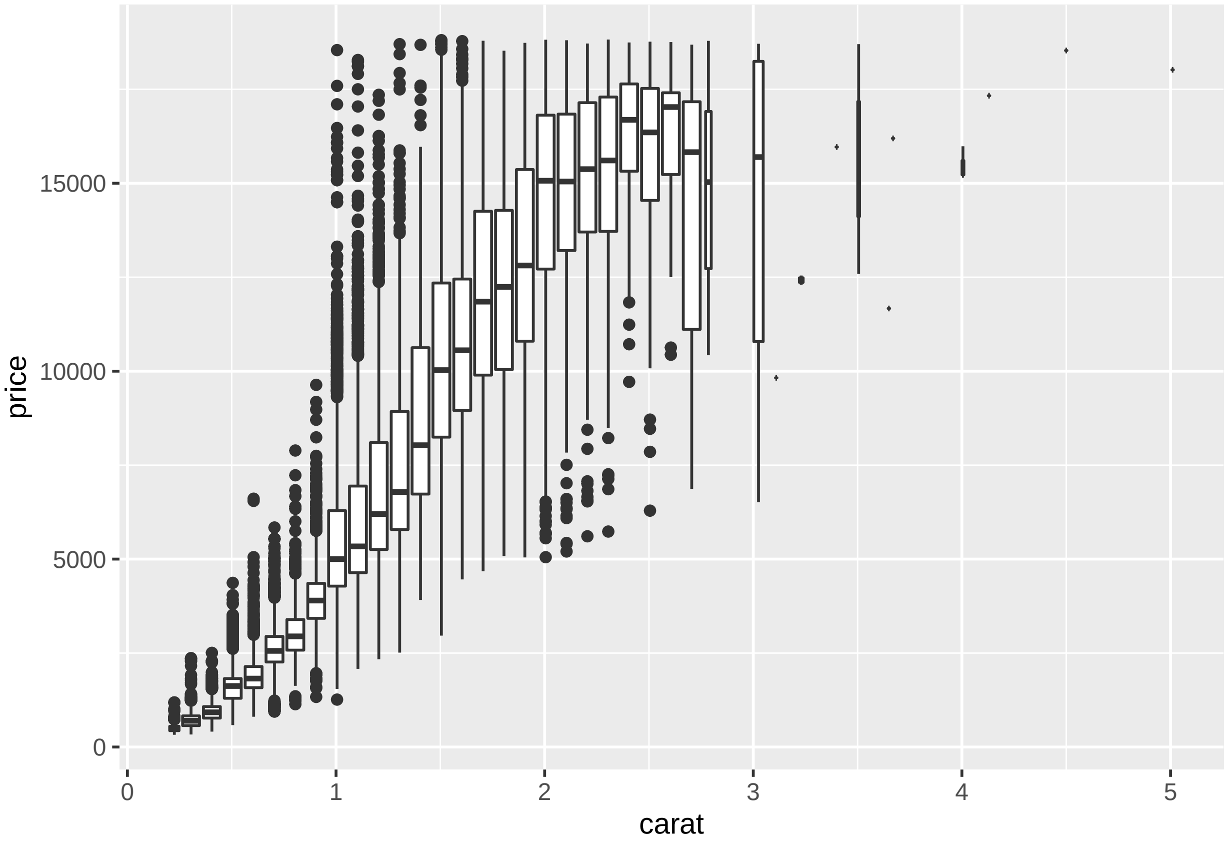
```
ggplot(diamonds) + geom_point(aes(x = carat, y = price), alpha=1/100)
```



```
ggplot(diamonds) + geom_bin_2d(aes(x=carat, y=price))
```



```
ggplot(diamonds, aes(x=carat, y=price)) + geom_boxplot(aes(group=cut_width(carat, 0.1)))
```



Pattern and models

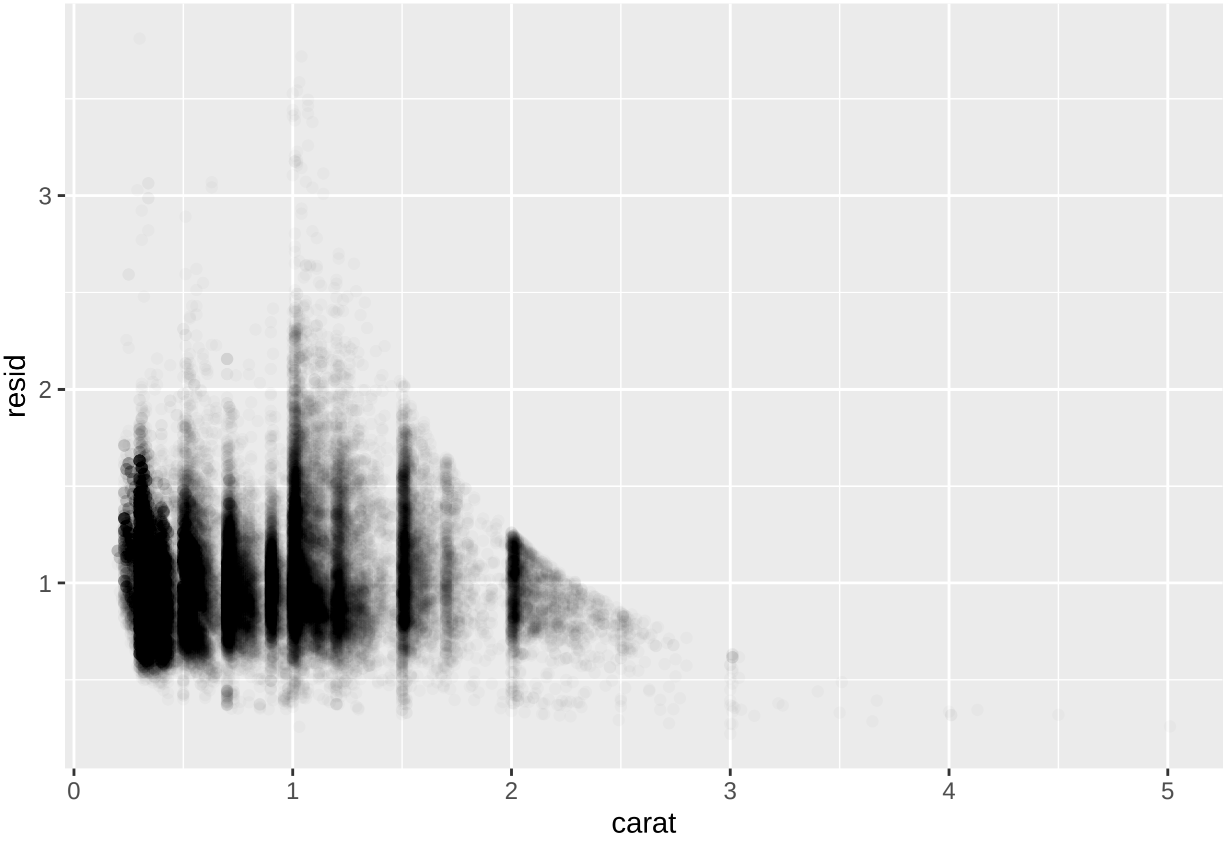
```
mod <- lm(log(price) ~ log(carat), data=diamonds)
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = log(price) ~ log(carat), data = diamonds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.50833 -0.16951 -0.00591  0.16637  1.33793
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.448661   0.001365  6190.9   <2e-16 ***
## log(carat)  1.675817   0.001934   866.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2627 on 53938 degrees of freedom
## Multiple R-squared:  0.933,  Adjusted R-squared:  0.933
## F-statistic: 7.51e+05 on 1 and 53938 DF,  p-value: < 2.2e-16
```

```
diamonds2 <- diamonds %>% add_residuals(mod) %>% mutate(resid=exp(resid))
```

```
ggplot(diamonds2) + geom_point(aes(x=carat, y=resid), alpha=1/50)
```


resid
carat