Data wrangling

Duong Vu

November 23, 2017

Library

Loading all the libraries:

```
library(tidyverse)
```

Loading data

```
summary(diamonds)
```

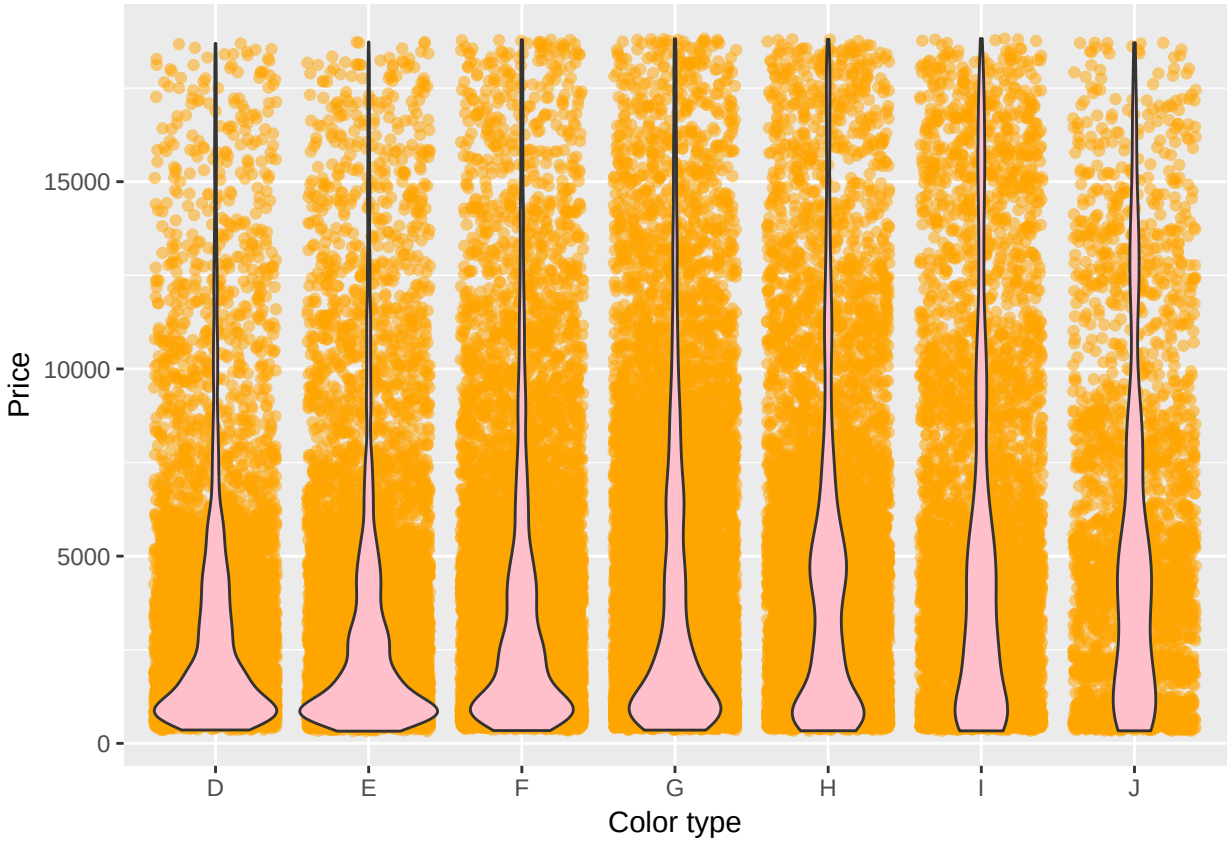
```
##      carat               cut        color        clarity     
##  Min.   :0.2000   Fair      : 1610   D: 6775   SI1    :13065  
##  1st Qu.:0.4000   Good      : 4906   E: 9797   VS2    :12258  
##  Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194  
##  Mean   :0.7979   Premium   :13791   G:11292   VS1    : 8171  
##  3rd Qu.:1.0400   Ideal     :21551   H: 8304   VVS2   : 5066  
##  Max.   :5.0100                     I: 5422   VVS1   : 3655  
##                                     J: 2808   (Other): 2531  
##      depth           table           price             x         
##  Min.   :43.00   Min.   :43.00   Min.   :  326   Min.   : 0.000  
##  1st Qu.:61.00   1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710  
##  Median :61.80   Median :57.00   Median : 2401   Median : 5.700  
##  Mean   :61.75   Mean   :57.46   Mean   : 3933   Mean   : 5.731  
##  3rd Qu.:62.50   3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540  
##  Max.   :79.00   Max.   :95.00   Max.   :18823   Max.   :10.740  
##                                                                  
##        y                z        
##  Min.   : 0.000   Min.   : 0.000  
##  1st Qu.: 4.720   1st Qu.: 2.910  
##  Median : 5.710   Median : 3.530  
##  Mean   : 5.735   Mean   : 3.539  
##  3rd Qu.: 6.540   3rd Qu.: 4.040  
##  Max.   :58.900   Max.   :31.800  
##
```

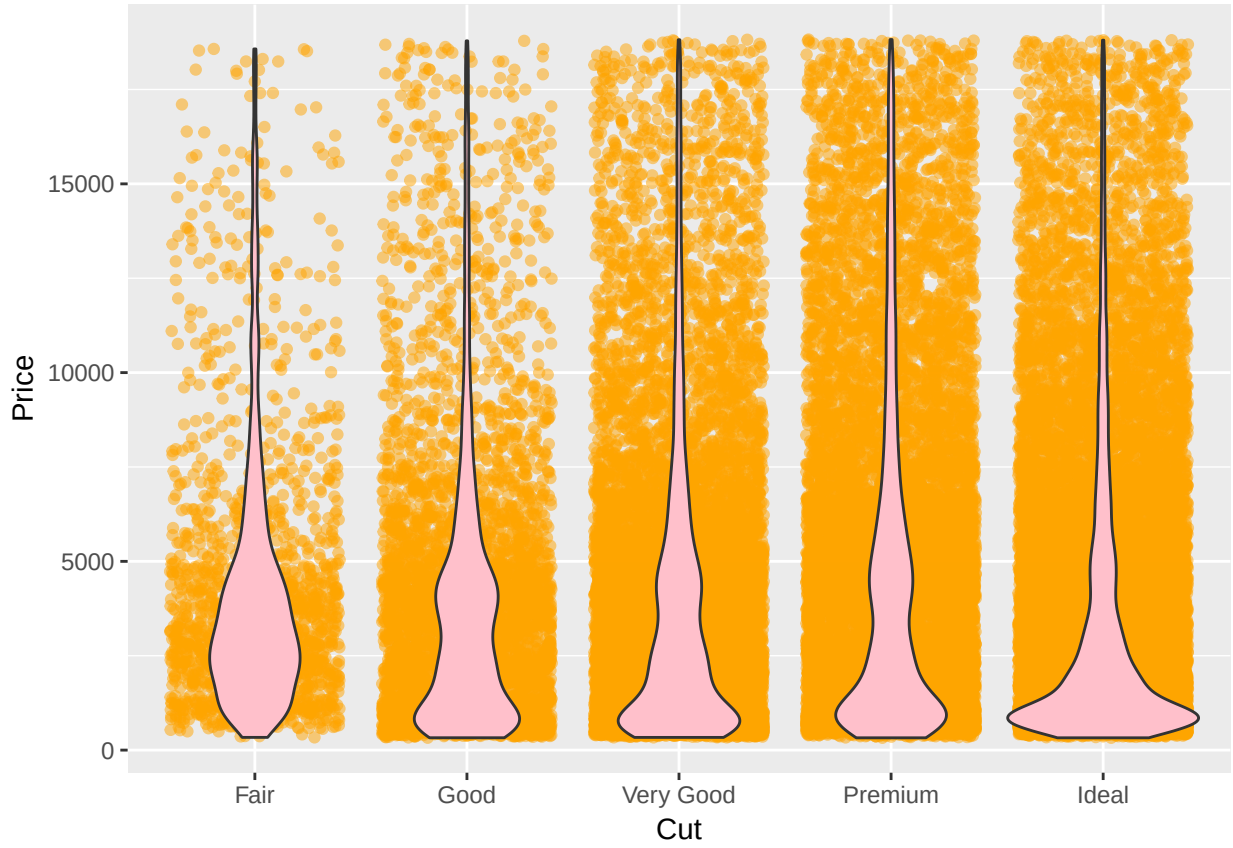
```
head(diamonds)
```

```
## # 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  0.23     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  4.20  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
```

Including Plots

Looking at the dataset and the relationship between Price and color and cut.

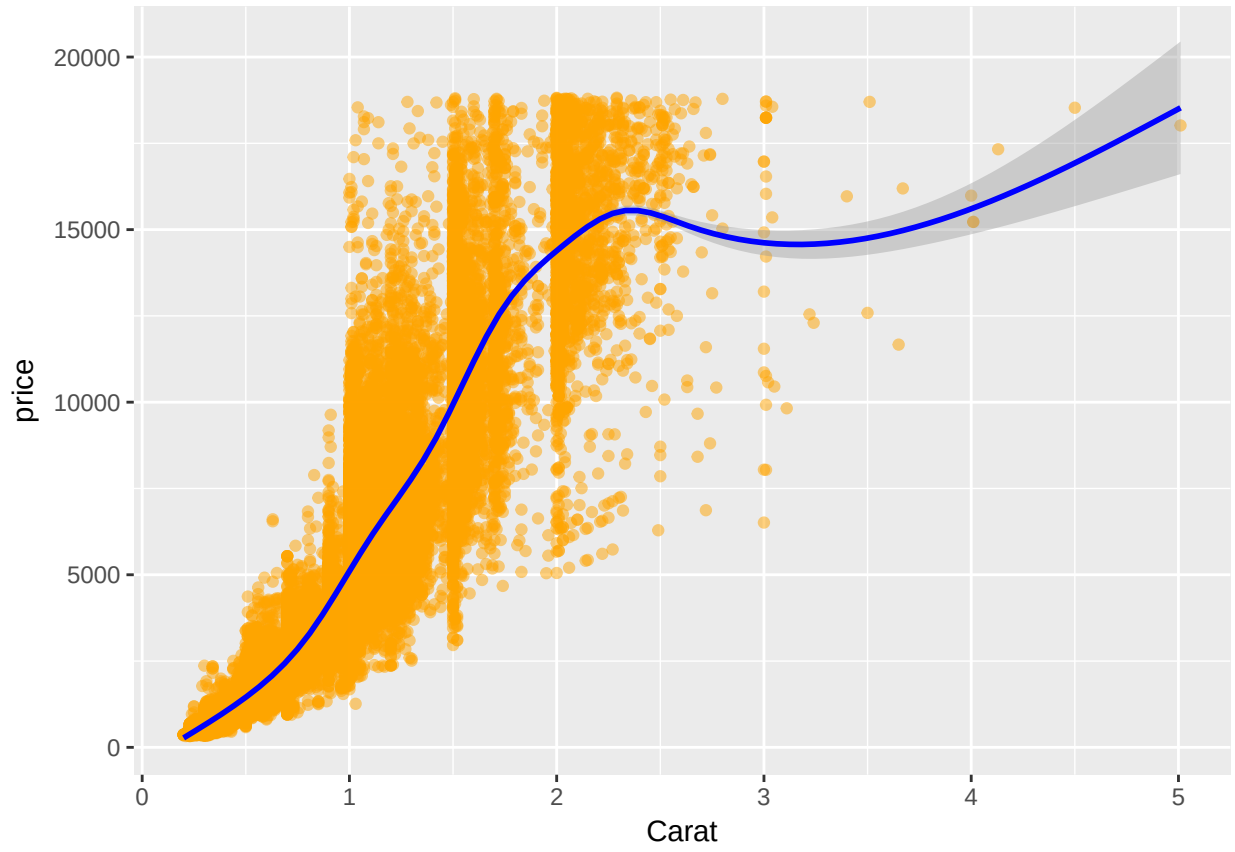




How about the carat?

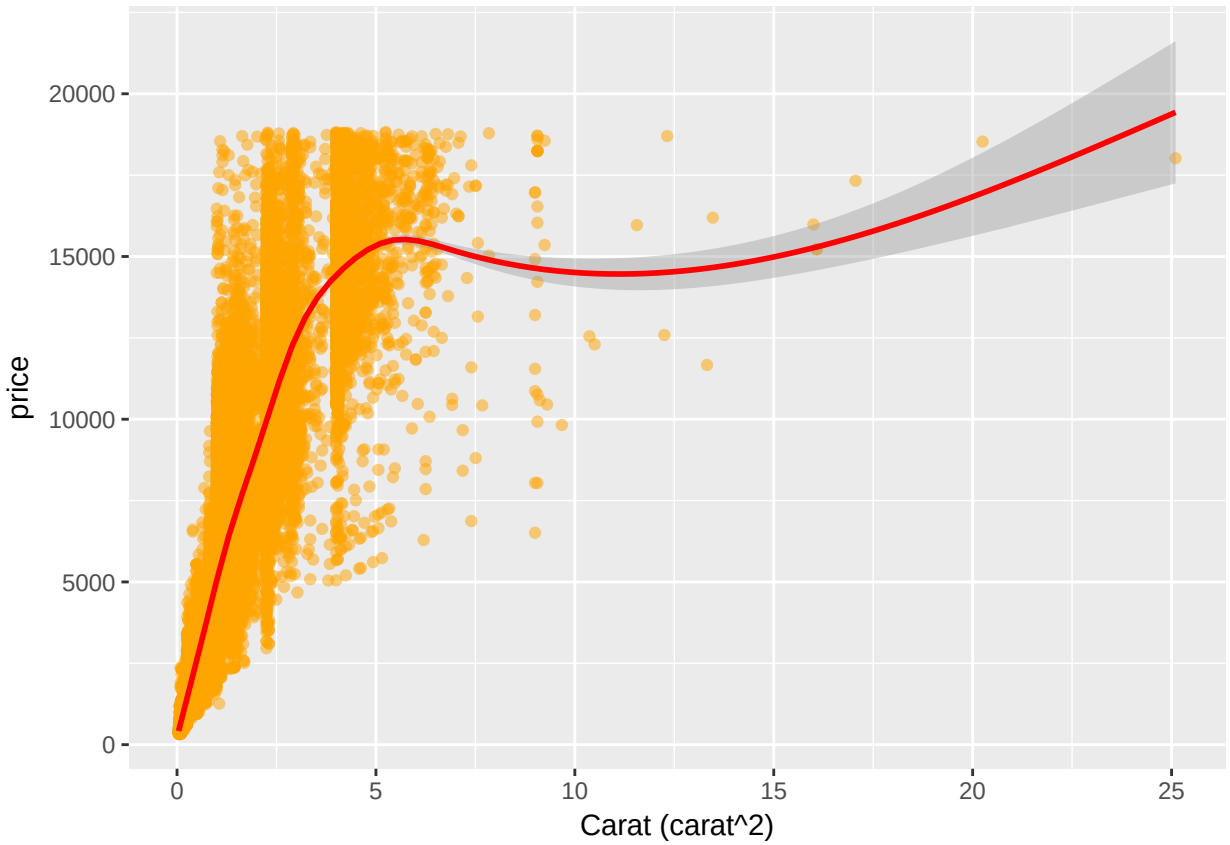
```
ggplot(diamonds, aes(carat, price)) +
  geom_point(alpha = 0.5, color = "orange") +
  geom_smooth(color = "blue") +
  labs(x = "Carat")
```

```
## `geom_smooth()` using method = 'gam'
```




```
ggplot(diamonds, aes(carat^2, price)) +
  geom_point(alpha = 0.5, color = "orange") +
  geom_smooth(color = "red") +
  labs(x = "Carat (carat^2)")
```

```
## `geom_smooth()` using method = 'gam'
```


price
20000
15000
10000
5000
0
0
5
10
15
20
25
Carat (carat^2)