Exploratory Data Analysis (EDA) in R

David Gerard

2019-02-08

Learning Objectives

- Strategies for EDA
- Chapter 7 of RDS

General Strategies

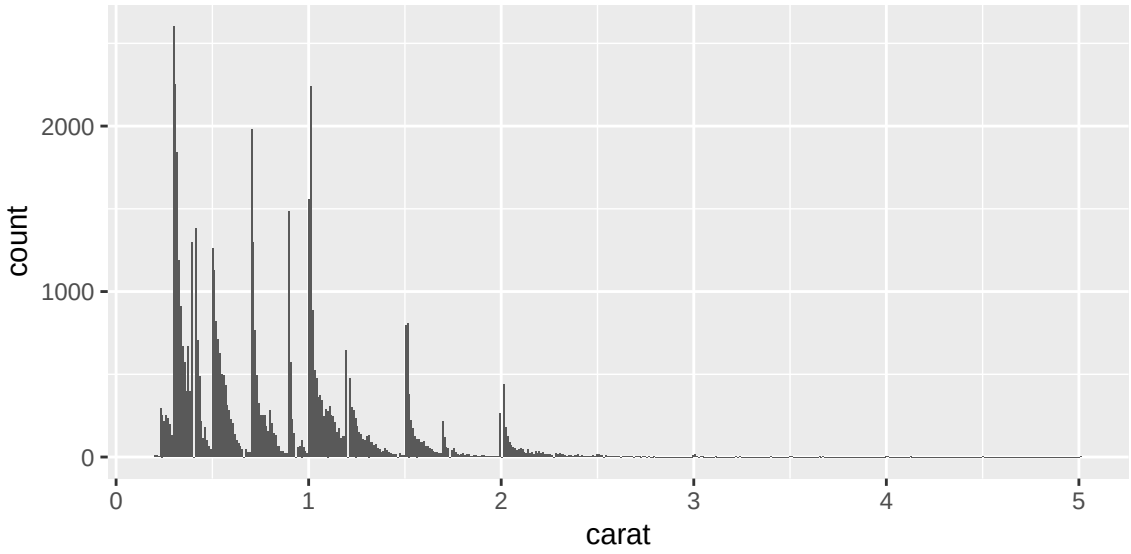
- Plot the distribution of every variable.
- Plot the bivariate distribution of every pair of variables (to find which variables are associated).
- Color code by variables to try and see if relationships can be explained.
- Calculate lots of summary statistics.
- Look at missingness.
- Look at outliers.
- EDA is about **curiosity**. Ask *many* questions, use *many* plots, investigate *many* aspects of your data. This will let you hone in on the few *interesting* questions you want to pursue deeper.

```
library(tidyverse)
data("diamonds")
```

Distribution of Every Variable:

- Quantitative: Use a histogram.
  - Look for modality. Indicates multiple groups of units. What can explain the modes? Can any of the other variables explain the modes?
  - Are certain values more likely than other values?
  - Look for skew.
  - Does it make sense to think of these data on the log scale? For example, do you expect to see about the same number of units in the two adjacent bins of (i) [1,2) and [2,3) or (ii) [1, 10) and [10, 100)? You should log in case (ii) and not log in case (i).
  - `geom_histogram()`
  - Mean, median, standard deviation, five number summary.

```
ggplot(data = diamonds, mapping = aes(x = carat)) +
  geom_histogram(bins = 500)
```




```
fivenum(diamonds$carat)
```

```
## [1] 0.20 0.40 0.70 1.04 5.01
```

```
mean(diamonds$carat)
```

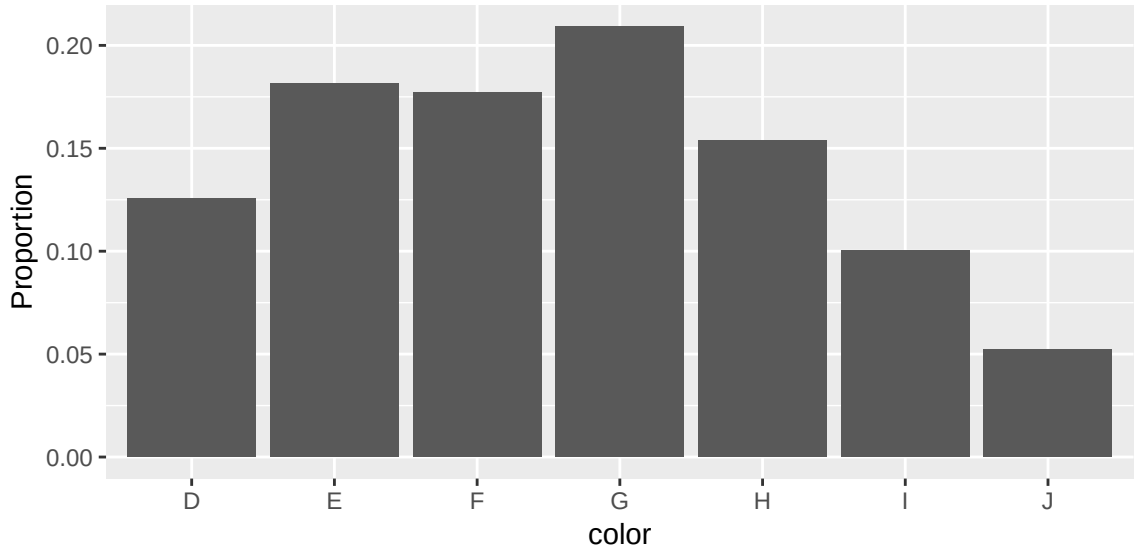
```
## [1] 0.7979
```

```
sd(diamonds$carat)
```

```
## [1] 0.474
```

- Categorical: Use a bar chart. Or just a table of *proportions* (`table()` then `prop.table()`).
  - Absolute counts are sometimes interesting, but usually you want to look at the proportion of observations in each category.
  - Is there a natural ordering of the categories (bad, medium, good)?
  - Why are some categories more represented than others?
  - `geom_bar()`, `geom_col()`
  - Proportion of observations within each group.

```
ggplot(diamonds, aes(x = color, y = ..)) +
  geom_bar(aes(y = ..count.. / sum(..count..))) +
  ylab("Proportion")
```




```
table(diamonds$color)
```

```
##
##     D     E     F     G     H     I     J
##  6775  9797  9542 11292  8304  5422  2808
```

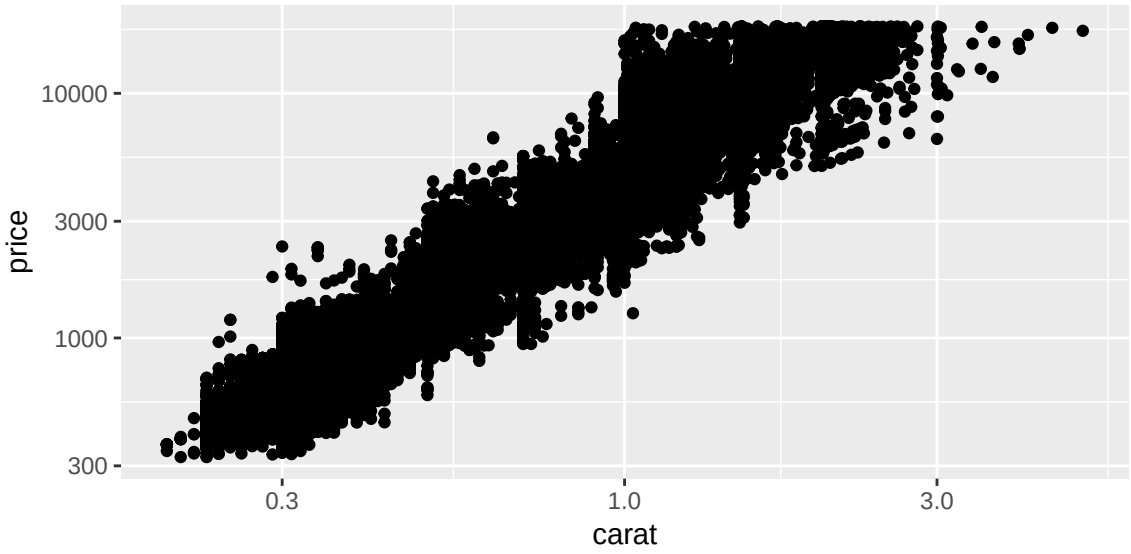
```
prop.table(table(diamonds$color))
```

```
##
##       D       E       F       G       H       I       J
## 0.12560 0.18163 0.17690 0.20934 0.15395 0.10052 0.05206
```

Bivariate Distribution of Every Pair of Variables

- Quantitative vs Quantitative: Use a scatterplot
  - Is the relationship linear? Quadratic? Exponential?
  - Logging is useful tool to make some associations linear. If the relationship is (i) monotonic and (ii) curved, then try logging the x-variable *if the x-variable is all positive*. If it is also (iii) more variable at larger y-values, then try logging the y-variable *instead* of the x-variable *if the y-variable is all positive*. Try logging both if you still see curvature *if both variables are all positive*.
  - Ask if an observed association can be explained by another variable?
  - Correlation coefficient (only appropriate if association is linear).
  - Kendall's tau (always appropriate).

```
ggplot(diamonds, aes(x = carat, y = price)) +
  geom_point() +
  scale_y_log10() +
  scale_x_log10()
```



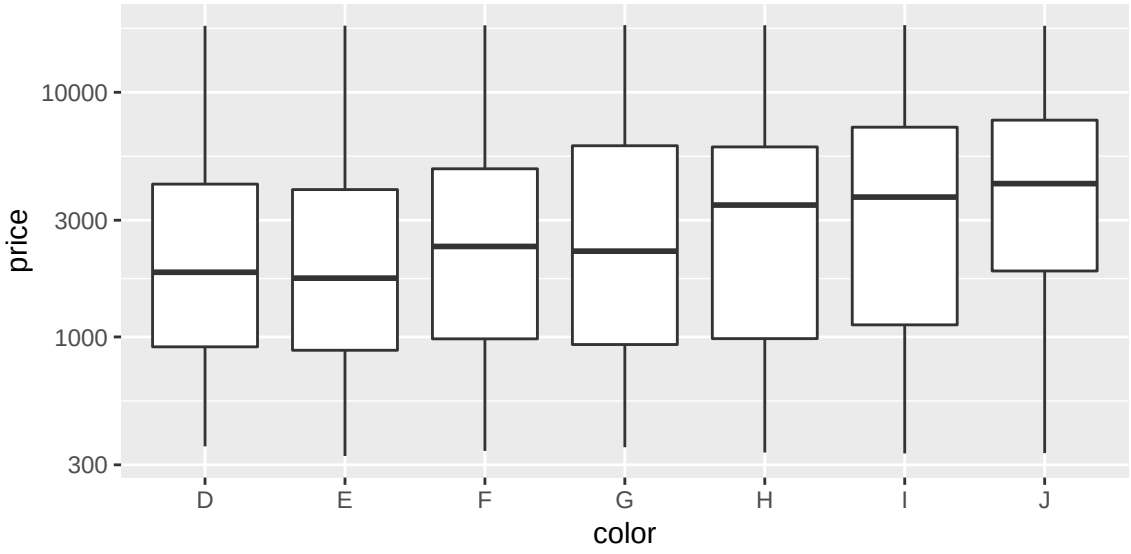

```
cor(diamonds$carat, diamonds$price)
```

```
## [1] 0.9216
```

```
## cor(diamonds$carat, diamonds$price, method = "kendall")
```

- Categorical vs Quantitative: Use a boxplot
  - For which levels of the categorical variable is the quantitative variable higher or lower?
  - For which levels is the quantitative variable more spread out?
  - Aggregated means, medians, standard deviations, quantiles

```
ggplot(diamonds, aes(x = color, y = price)) +
  geom_boxplot() +
  scale_y_log10()
```

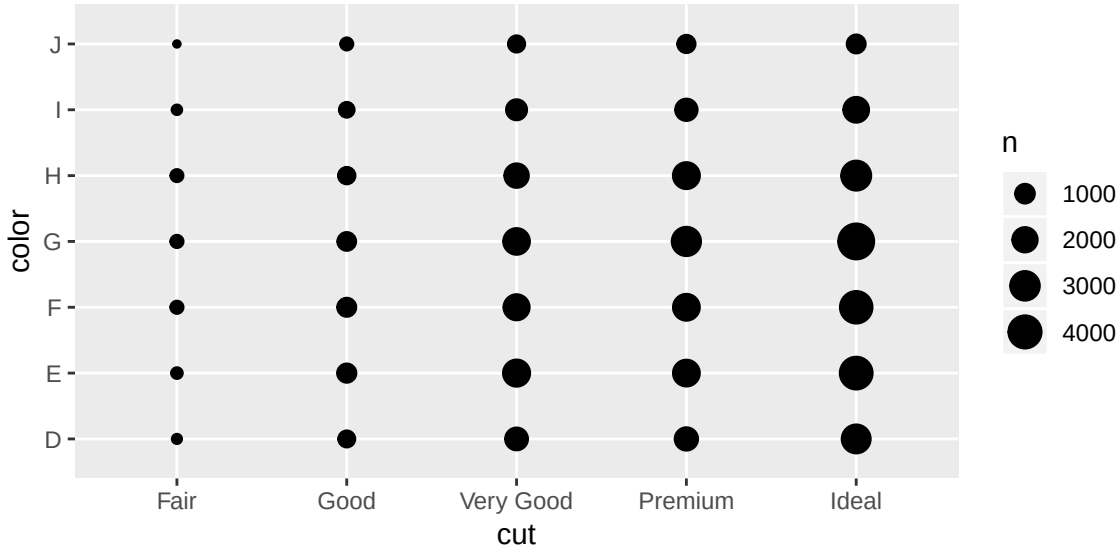


```r
diamonds %>%
  mutate(logprice = log(price)) %>%
  group_by(color) %>%
  summarize(mean   = mean(logprice),
            sd     = sd(logprice),
            median = median(logprice),
            Q1     = quantile(logprice, 0.25),
            Q3     = quantile(logprice, 0.75))
```

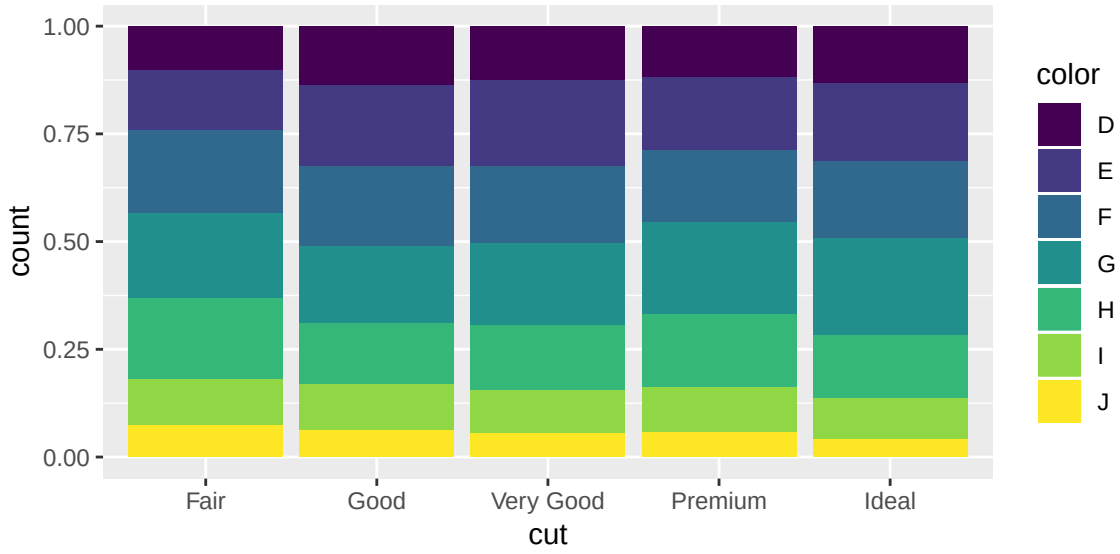
```
## # A tibble: 7 x 6
##   color  mean    sd median    Q1    Q3
##   <ord> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 D      7.62 0.926   7.52  6.81  8.35
## 2 E      7.58 0.925   7.46  6.78  8.29
## 3 F      7.76 0.968   7.76  6.89  8.49
## 4 G      7.79 1.03    7.72  6.84  8.71
## 5 H      7.92 1.06    8.15  6.89  8.70
## 6 I      8.02 1.11    8.22  7.02  8.88
## 7 J      8.15 1.04    8.35  7.53  8.95
```

- Categorical vs Categorical: Use a mosaic plot or a count plot
  - For which pairs of values of the categorical variables are there the most number of units?
  - Does the conditional distribution of a categorical variable change at different levels of the other categorical variable?
  - `prop.table()`

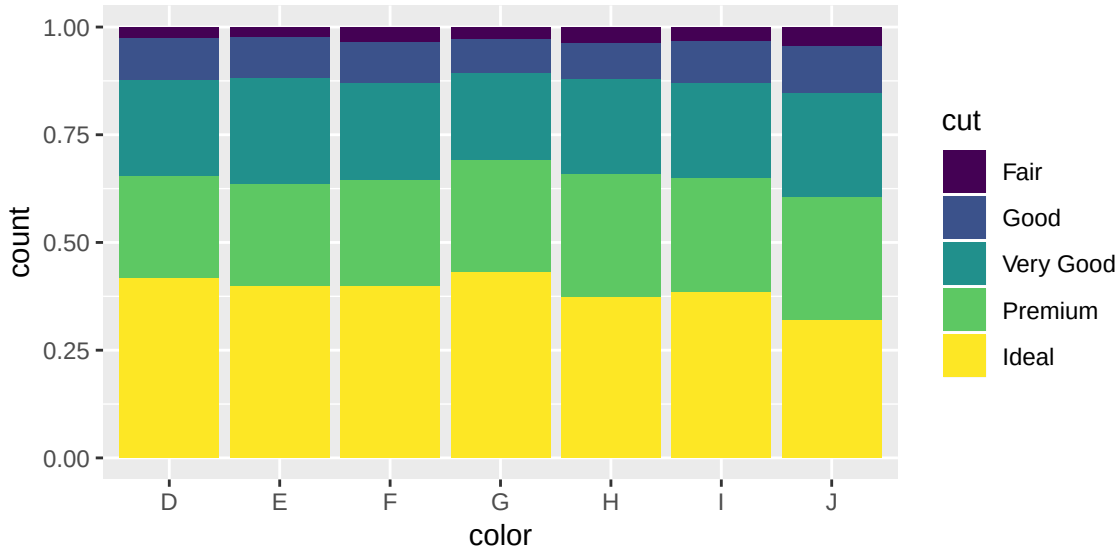
```r
## Only gives you the bivariate distribution
ggplot(diamonds, aes(x = cut, y = color)) +
  geom_count()
```




```r
## Gives you the conditional distributions of color given cut
ggplot(diamonds, aes(x = cut, fill = color)) +
  geom_bar(position = "fill")
```




```
## Gives you the conditional distributions of cut given color
ggplot(diamonds, aes(x = color, fill = cut)) +
  geom_bar(position = "fill")
```




```
## Bivariate Distribution
prop.table(table(diamonds$color, diamonds$cut))
```

```
##
##         Fair     Good Very Good  Premium    Ideal
##   D 0.003022 0.012273  0.028050 0.029718 0.052540
##   E 0.004153 0.017297  0.044494 0.043326 0.072358
##   F 0.005784 0.016852  0.040119 0.043215 0.070931
##   G 0.005821 0.016148  0.042621 0.054208 0.090545
##   H 0.005617 0.013014  0.033815 0.043752 0.057749
##   I 0.003244 0.009677  0.022321 0.026474 0.038802
##   J 0.002206 0.005692  0.012570 0.014980 0.016611
```

```r
##  Conditional distributions of column variable conditional on row variable
prop.table(table(diamonds$color, diamonds$cut), margin = 1)
```

```
##
##        Fair    Good Very Good Premium   Ideal
##   D 0.02406 0.09771   0.22332 0.23661 0.41830
##   E 0.02286 0.09523   0.24497 0.23854 0.39839
##   F 0.03270 0.09526   0.22679 0.24429 0.40096
##   G 0.02781 0.07713   0.20360 0.25894 0.43252
##   H 0.03649 0.08454   0.21965 0.28420 0.37512
##   I 0.03228 0.09627   0.22206 0.26337 0.38602
##   J 0.04238 0.10933   0.24145 0.28775 0.31909
```

```r
## Conditional distributions of row variable conditional on column variable
prop.table(table(diamonds$color, diamonds$cut), margin = 2)
```

```
##
##        Fair    Good Very Good Premium   Ideal
##   D 0.10124 0.13494   0.12523 0.11624 0.13150
##   E 0.13913 0.19018   0.19864 0.16946 0.18111
##   F 0.19379 0.18528   0.17911 0.16902 0.17753
##   G 0.19503 0.17754   0.19028 0.21202 0.22663
##   H 0.18820 0.14309   0.15097 0.17113 0.14454
##   I 0.10870 0.10640   0.09965 0.10355 0.09712
##   J 0.07391 0.06258   0.05612 0.05859 0.04158
```