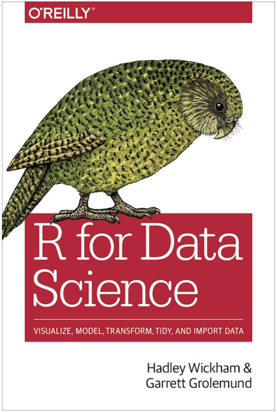
Socio-Informatics 348

Data Visualisation
Exploratory Data Analysis

Dr Lisa Martin

Department of Information Science
Stellenbosch University

Today's Reading



R for Data Science, Chapter 10

EDA: A Creative, Iterative Process

Exploratory data analysis is not a rigid protocol. It is an iterative cycle:

1. Generate questions about your data.
2. Seek answers via visualisation, transformation, and modeling.
3. Use new insights to ask better or new questions.

Even when your research questions are predefined, EDA is invaluable for checking data quality and guiding cleaning steps.

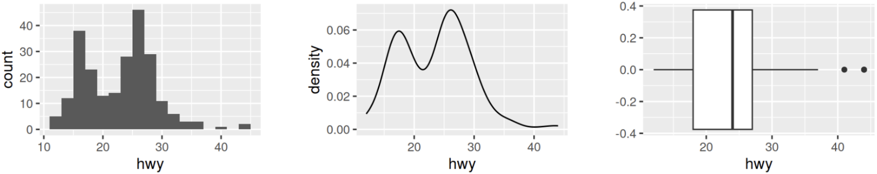
Key Questions in EDA

You should ask many questions during EDA, but there are two core types of questions that guide exploration:

1. **Variation:** What type of variation occurs within my variables?
2. **Covariation:** What type of covariation occurs between my variables?

1. Variation

- Variation is the tendency of the values of a variable to change from measurement to measurement.
- Captures how a variable's values spread, cluster, or distribute across observations.
- Visual tools: histograms, density plots, boxplots.



R for Data Science, Chapter 9

1. Variation

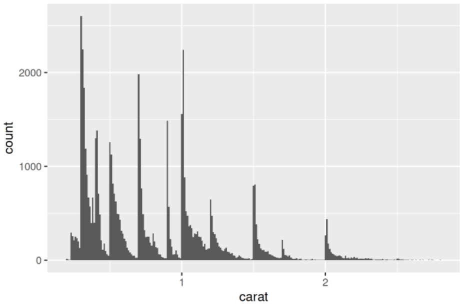
Typical Values:

- Which values are the most common? Why?
  - These are the tallest bars or the densest areas in your visualisations.
- Which values are rare? Why? Is this expected?
  - These are the shorter bars or the less dense areas in your visualisations.
- Any unusual patterns? What might explain them?

1. Variation

```
smaller <- diamonds |>
  filter(carat < 3)

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



- Why are there more diamonds at whole carats and common fractions of carats?
- Why are there more diamonds slightly to the right of each peak than there are slightly to the left of each peak?

1. Variation

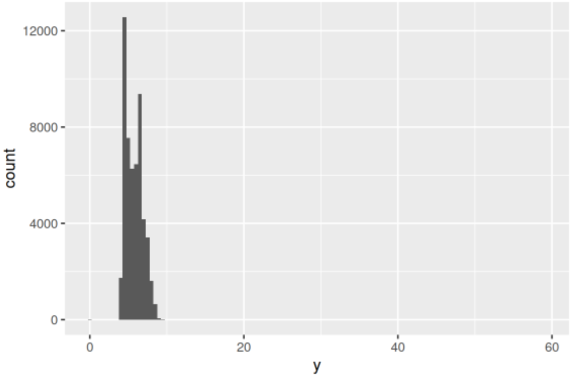
Unusual Values:

- These are values that deviate noticeably from the norm.
  - Data entry errors – not meaningful
  - Genuine extremes – meaningful
  - Something else?
- Regardless, important to identify and understand them.
- Visual tools: boxplots, histograms, scatterplots.
  - Boxplots use a point to indicate unusual values/outliers.
  - In a historogram, unusual values are often visible as isolated bars.
  - In scatterplots, unusual values may appear as points far from the main cluster.

1. Variation

Unusual Values:
With large data sets, unusual values can be hard to spot.

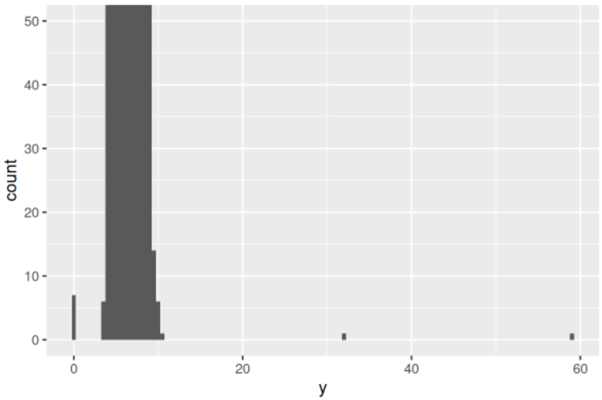
```
ggplot(diamonds, aes(x = y)) +
  geom_histogram(binwidth = 0.5)
```



1. Variation

Unusual Values:

```
ggplot(diamonds, aes(x = y)) +
  geom_histogram(binwidth = 0.5) +
  coord_cartesian(ylim = c(0, 50))
```



1. Variation

Unusual Values:

```
unusual <- diamonds |>
  filter(y < 3 | y > 20) |>
  select(price, x, y, z) |>
  arrange(y)
unusual
#> # A tibble: 9 × 4
#>   price     x     y     z
#>   <int> <dbl> <dbl> <dbl>
#> 1  5139  0      0    0
#> 2  6381  0      0    0
#> 3 12800  0      0    0
#> 4 15686  0      0    0
#> 5 18034  0      0    0
#> 6  2130  0      0    0
#> 7  2130  0      0    0
#> 8  2075  5.15  31.8  5.12
#> 9 12210  8.09  58.9  8.06
```

2. Unusual Values

Handling Unusual Values:

- Decide whether to keep, remove, or transform unusual values based on their context.
- Drop:
  Simple, but you lose potentially valuable information in other variables.

```
diamonds2 <- diamonds |>
  filter(between(y, 3, 20))
```

- Replace with NA: Keeps the remaining data intact

```
diamonds2 <- diamonds |>
  mutate(y = if_else(y < 3 | y > 20, NA, y))
```

- Good practice:
  - Document your decisions and the rationale behind them.
  - If kept, run analysis with and without them included.

3. Covariation

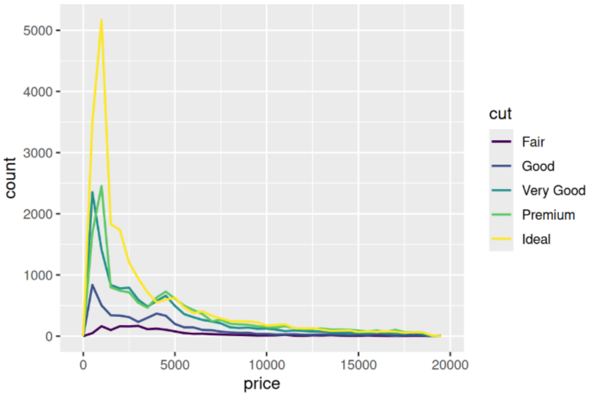
Covariation is the tendency for the values of two or more variables to vary together in a related way.

- The best way to spot covariation is to visualise the relationship between two or more variables.
- Critical for spotting relationships and informing further modeling.

3. Covariation

Example 1: Categorical and Numerical Variables

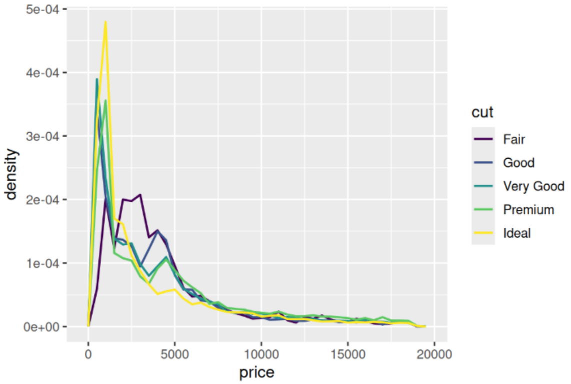
```
ggplot(diamonds, aes(x = price)) +
  geom_freqpoly(aes(color = cut), binwidth = 500, linewidth = 0.75)
```



3. Covariation

Example 1: Categorical and Numerical Variables

```
ggplot(diamonds, aes(x = price, y = after_stat(density))) +
  geom_freqpoly(aes(color = cut), binwidth = 500, linewidth = 0.75)
```

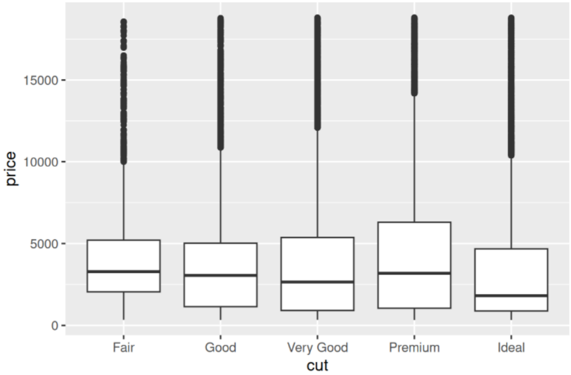


- Appears that ‘Fair’ has a higher average price?

3. Covariation

Example 1: Categorical and Numerical Variables

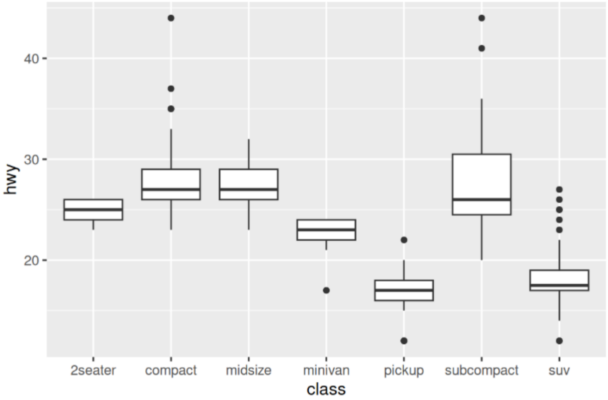
```
ggplot(diamonds, aes(x = cut, y = price)) +
  geom_boxplot()
```



- Boxplot confirms our suspicion. Needs further exploration.

Example 2: Categorical and Numerical Variables

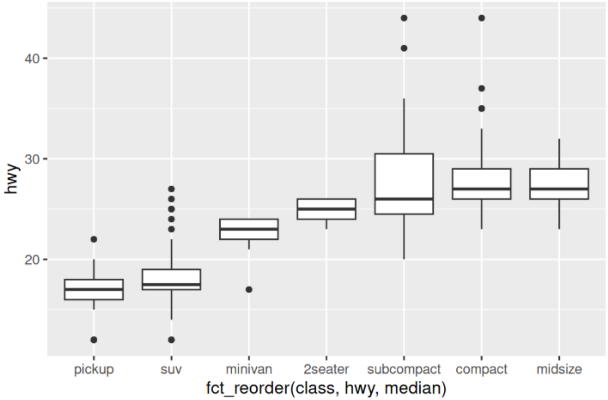
```
ggplot(mpg, aes(x = class, y = hwy)) +
  geom_boxplot()
```



3. Covariation

Example 2: Categorical and Numerical Variables

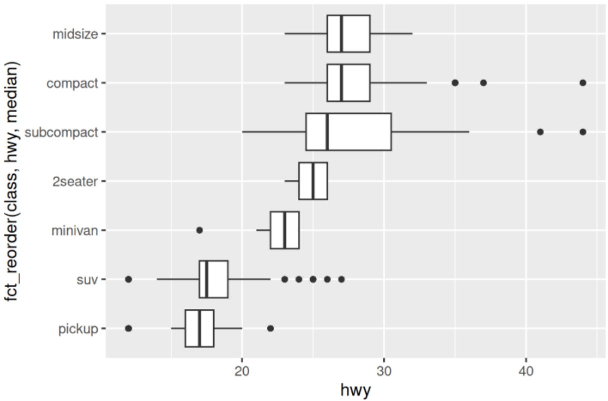
```
ggplot(mpg, aes(x = fct_reorder(class, hwy, median), y = hwy)) +
  geom_boxplot()
```



3. Covariation

Example 2: Categorical and Numerical Variables

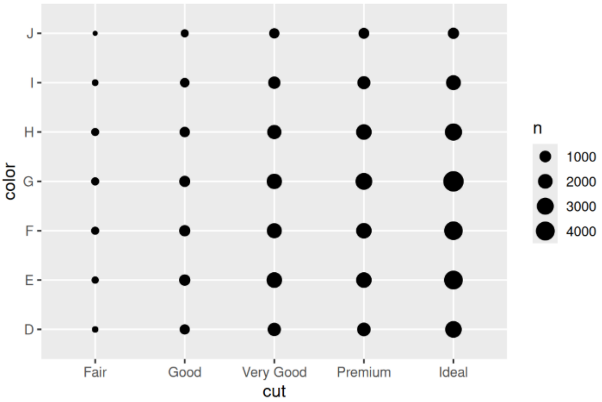
```r
ggplot(mpg, aes(x = hwy, y = fct_reorder(class, hwy, median))) +
  geom_boxplot()
```



3. Covariation

Example: Two Categorical Variables

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



3. Covariation

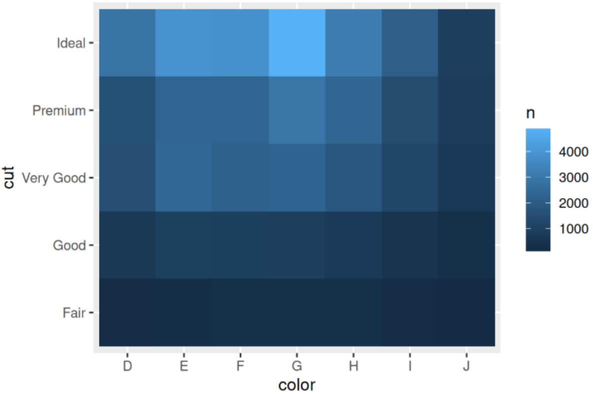
Example: Two Categorical Variables

```
diamonds |>
  count(color, cut)
#> # A tibble: 35 × 3
#>   color cut           n
#>   <ord> <ord>     <int>
#> 1 D     Fair        163
#> 2 D     Good        662
#> 3 D     Very Good  1513
#> 4 D     Premium    1603
#> 5 D     Ideal      2834
#> 6 E     Fair        224
#> # ℹ 29 more rows
```

3. Covariation

Example: Two Categorical Variables

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

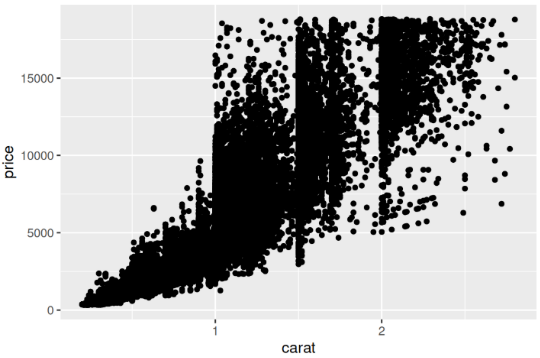


3. Covariation

Example: Two Numerical Variables

You are already familiar with the scatterplot, which is a common way to visualize the relationship between two numerical variables.

```
ggplot(smaller, aes(x = carat, y = price)) +
  geom_point()
```

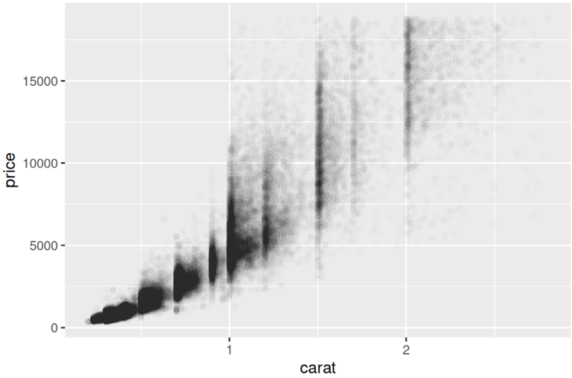


3. Covariation

Example: Two Numerical Variables

Combat overlap with transparrency, but this is still a problem with very large datasets.

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



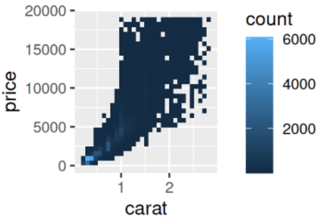
3. Covariation

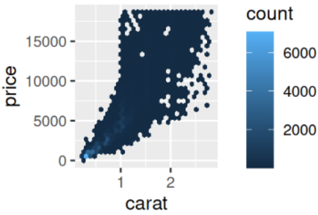
Example: Two Numerical Variables

Use bins to reduce the number of points plotted.

```r
ggplot(smaller, aes(x = carat, y = price)) +
  geom_bin2d()

# install.packages("hexbin")
ggplot(smaller, aes(x = carat, y = price)) +
  geom_hex()
```



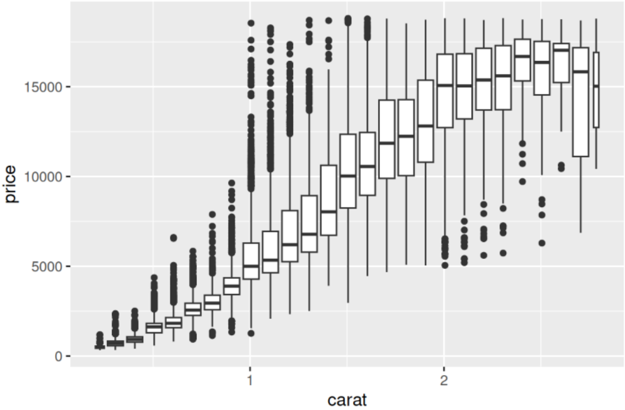


3. Covariation

Example: Two Numerical Variables

Another option: Bin one of the variables to create a categorical variable.

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



4. Patterns and Models

If a systematic relationship exists between two variables it will appear as a pattern in the data.

- Could this pattern be due to coincidence (i.e. random chance)?
- How can you describe the relationship implied by the pattern?
- How strong is the relationship implied by the pattern?
- What other variables might affect the relationship?
- Does the relationship change if you look at individual subgroups of the data?

4. Patterns and Models

- Patterns in the data can often be summarised by simple models.
- Modeling in EDA is exploratory—used to reveal trends, not confirm hypotheses.