

2.exploration (exercises)

Joao Lopes

2024-04-29

Contents

1. VARIATION	3
1.1. TYPICAL VALUES	3
1.2. UNUSUAL VALUES	3
2. MISSING VALUES	4
2.1. EXPLICIT MISSING VALUES	4
2.2. IMPLICIT MISSING VALUES	4
2.3. FACTORS AND EMPTY GROUPS	4
3. COVARIATION	5
3.1. CATEGORICAL AND NUMERICAL VARIABLES	5
3.2. TWO CATEGORICAL VARIABLES	5
3.3. TWO NUMERICAL VARIABLES	5
4. PATTERNS AND MODELS	7

1. VARIATION

[from <https://r4ds.hadley.nz/eda#variation>]

[see <https://jrnold.github.io/r4ds-exercise-solutions/exploratory-data-analysis.html>]

[see <https://mine-cetinkaya-rundel.github.io/r4ds-solutions/EDA.html>]

1.1. TYPICAL VALUES

- a) Explore the distribution of each of the x, y, and z variables in diamonds. What do you learn? Think about a diamond and how you might decide which dimension is the length, width, and depth.
- b) Explore the distribution of price. Do you discover anything unusual or surprising? (Hint: Carefully think about the binwidth and make sure you try a wide range of values.)
- c) How many diamonds are 0.99 carat? How many are 1 carat? What do you think is the cause of the difference?
- d) Compare and contrast `coord_cartesian()` vs `xlim()` or `ylim()` when zooming in on a histogram. What happens if you leave `binwidth` unset? What happens if you try and zoom so only half a bar shows?

1.2. UNUSUAL VALUES

[No exercises]

2. MISSING VALUES

[from <https://r4ds.hadley.nz/missing-values>]

[see <https://jrnold.github.io/r4ds-exercise-solutions/exploratory-data-analysis.html>]

[see <https://mine-cetinkaya-rundel.github.io/r4ds-solutions/EDA.html>]

2.1. EXPLICIT MISSING VALUES

a) What happens to missing values in a histogram? What happens to missing values in a bar chart? Why is there a difference?

b) What does `na.rm = TRUE` do in `mean()` and `sum()`?

2.2. IMPLICIT MISSING VALUES

a) Can you find any relationship between the carrier and the rows that appear to be missing from planes?

2.3. FACTORS AND EMPTY GROUPS

[no exercises]

3. COVARIATION

[from <https://r4ds.hadley.nz/eda#covariation>]

[see <https://jrnold.github.io/r4ds-exercise-solutions/exploratory-data-analysis.html>]

[see <https://mine-cetinkaya-rundel.github.io/r4ds-solutions/EDA.html>]

3.1. CATEGORICAL AND NUMERICAL VARIABLES

- a) Use what you've learned to improve the visualization of the departure times of cancelled vs. non-cancelled flights.
- b) What variable in the diamonds dataset is most important for predicting the price of a diamond? How is that variable correlated with cut? Why does the combination of those two relationships lead to lower quality diamonds being more expensive?
- c) One problem with boxplots is that they were developed in an era of much smaller datasets and tend to display a prohibitively large number of “outlying values”. One approach to remedy this problem is the letter value plot. Install the `lvplot` package, and try using `geom_lv()` to display the distribution of price vs. cut. What do you learn? How do you interpret the plots?
- d) Create a visualization of diamond prices vs. a categorical variable from the diamonds dataset using `geom_violin()`, then a faceted `geom_histogram()`, then a colored `geom_freqpoly()`, and then a colored `geom_density()`. Compare and contrast the four plots. What are the pros and cons of each method of visualizing the distribution of a numerical variable based on the levels of a categorical variable?
- e) If you have a small dataset, it's sometimes useful to use `geom_jitter()` to avoid overplotting to more easily see the relationship between a continuous and categorical variable. The `ggbeeswarm` package provides a number of methods similar to `geom_jitter()`. List them and briefly describe what each one does.

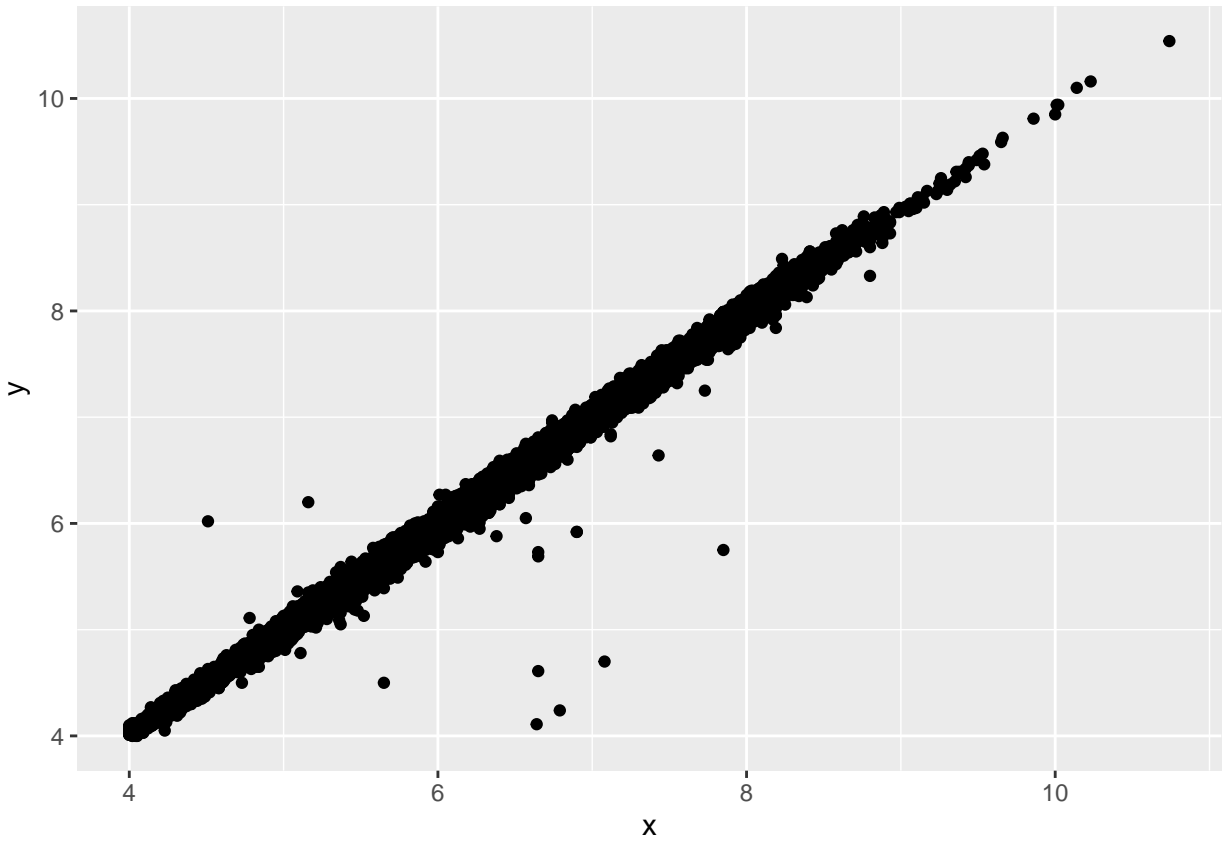
3.2. TWO CATEGORICAL VARIABLES

- a) How could you rescale the count dataset above to more clearly show the distribution of cut within color, or color within cut?
- b) What different data insights do you get with a segmented bar chart if color is mapped to the x aesthetic and cut is mapped to the fill aesthetic? Calculate the counts that fall into each of the segments.
- c) Use `geom_tile()` together with `dplyr` to explore how average flight delays vary by destination and month of year. What makes the plot difficult to read? How could you improve it?

3.3. TWO NUMERICAL VARIABLES

- a) Instead of summarizing the conditional distribution with a box plot, you could use a frequency polygon. What do you need to consider when using `cut_width()` vs `cut_number()`? How does that impact a visualization of the 2d distribution of carat and price?
- b) Visualize the distribution of carat, partitioned by price.
- c) How does the price distribution of very large diamonds compare to small diamonds. Is it as you expect, or does it surprise you?
- d) Combine two of the techniques you've learned to visualize the combined distribution of cut, carat, and price.
- e) Two dimensional plots reveal outliers that are not visible in one dimensional plots. For example, some points in the plot below have an unusual combination of x and y values, which makes the points outliers even though their x and y values appear normal when examined separately. Why is a scatterplot a better display than a binned plot for this case?

```
diamonds |>  
  filter(between(x, 4, 11), between(y, 4, 11)) |>  
  ggplot(aes(x = x, y = y)) +  
  geom_point()
```



4. PATTERNS AND MODELS

[no exercises]