Exercise2_HaoranXu

Haoran Xu

2025-01-24

1. Why is EDA an important part of a modeling exercise?

EDA allows us get a brief glimpse of the dataset to understand its structure, characteristics, and quality before conducting data analysis. It also helps us to possibly spot patterns, trends, and correlations within and between variables so that we can detect dataset defects in advance and choose the right model.

2. How many aesthetic elements of visualization can you think of? List them and explain how they could be mapped to variables in different scales.

x-axis, y-axis

Quantitative variables: mapped to precise positions Categorical variables: mapped to discrete positions
e.g.,

```
ggplot(aes(x = mode, y = pollutants))
```

color

Quantitative variables: use distinct and discrete colors Categorical variables: use gradients or sequential colors
e.g.,

```
geom_bar(color = "black",
           fill = "white")
```

size

Adjust the sizes of points, lines, or areas.
e.g.,

```
geom_segment(size = 1)
```

shape

Adjust the shapes of points.
e.g.,

```
geom_sf(data = data,
          size = 2,
          shape = 1)
```

line type / thickness

Adjust the shapes or width of lines.

transparency (alpha)

Higher values are more opaque, lower values more transparent.

labels

Texts for axises or legends.
e.g.,

```
labs(y="Mode",
     x = expression("Sidewalk density (km/km"^2*")"),
     # Add a label for the fill
     fill = "Mode")
```

faceting

Create multiple subplots.
e.g.,

```
facet_wrap(~ available.Walk,
           labeller = label_both)
```

fill

Quantitative variables: use distinct and discrete colors Categorical variables: use gradients or sequential colors
e.g.,

```
scale_fill_brewer(palette = "Dark2")
```

3. The data set explored in this chapter was in wide format: each row was a single respondent. However, discrete choice data is often found in long format: each row is a choice situation, and each individual now appears in the table as many times as they faced a choice situation. Would this change in any way how you approach data analysis?

I think this would change someway about how to conduct data analysis. Wide format data have rows corresponding to a single respondent, which is better for individual-level summary statistics. While long format data have rows corresponding to a single choice situation, which is better for discrete choice models.

Long format datasets have advantages in choice-specific analysis and are needed for models like Multinomial Logit Models, Mixed Logit Models, and Nested Logit Models. They are also good at visualizing choice-specific data analysis.

```
library(mlogit)
data("Car")
```

4. How many variables are there in this data set and of which type (i.e., categorical/quantitative)?

5. Choose four relevant categorical variables from this data set and describe them using univariate, bivariate, and multivariate techniques. Discuss your results.

6. What ideas about individuals' choices regarding car do you develop from this EDA exercise?