Data 6 Summer 2023

# Visualizations

Week 3 Tutoring Worksheet

July 2023

## 1. Visualization Fundamentals

Visualization is an incredibly important tool in a data scientist's toolkit, enabling you to better understand the data you're working with and to share insights with others. However, not all visualizations are created equal — what **visualization type** works best for your data depends heavily on the **type(s) of data** you're working with and what you're trying to show with your visualization.

a. Describe what is meant by ‘encoding‘ in the context of visualization.

b. For each of the following variables, determine its variable type.

i. Phone number

ii. Occupation (e.g. accountant, construction worker, etc.)

iii. Day of the week

iv. Income

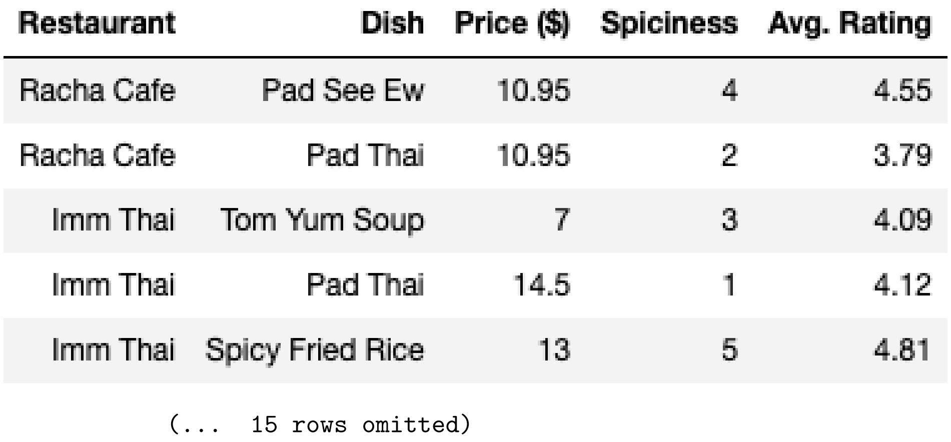
v. Number of books owned

c. Match the variable type(s) to the most appropriate visualization type.

| | |
|---|---|
| 1 Numerical Discrete Variable | |
| 1 Numerical Continuous Variable | Bar Chart |
| 1 Categorical Variable (Ordinal/Nominal) | Histogram |
| 1 Categorical Variable, 1 Numerical Variable | Line Plot |
| 2 Numerical Variables | Scatter Plot |
| 2 Numerical Variables (one of which is time) | |

## 2. Charts, Graphs and Plots Galore

thai_restaurants table:



| Restaurant | Dish | Price ($) | Spiciness | Avg. Rating |
|---|---|---|---|---|
| Racha Cafe | Pad See Ew | 10.95 | 4 | 4.55 |
| Racha Cafe | Pad Thai | 10.95 | 2 | 3.79 |
| Imm Thai | Tom Yum Soup | 7 | 3 | 4.09 |
| Imm Thai | Pad Thai | 14.5 | 1 | 4.12 |
| Imm Thai | Spicy Fried Rice | 13 | 5 | 4.81 |

```
(... 15 rows omitted)
```

Using the table thai_restaurants above, write code to create the following visualizations.

a. A histogram showing the distribution of prices across all dishes in the thai_restaurants table.

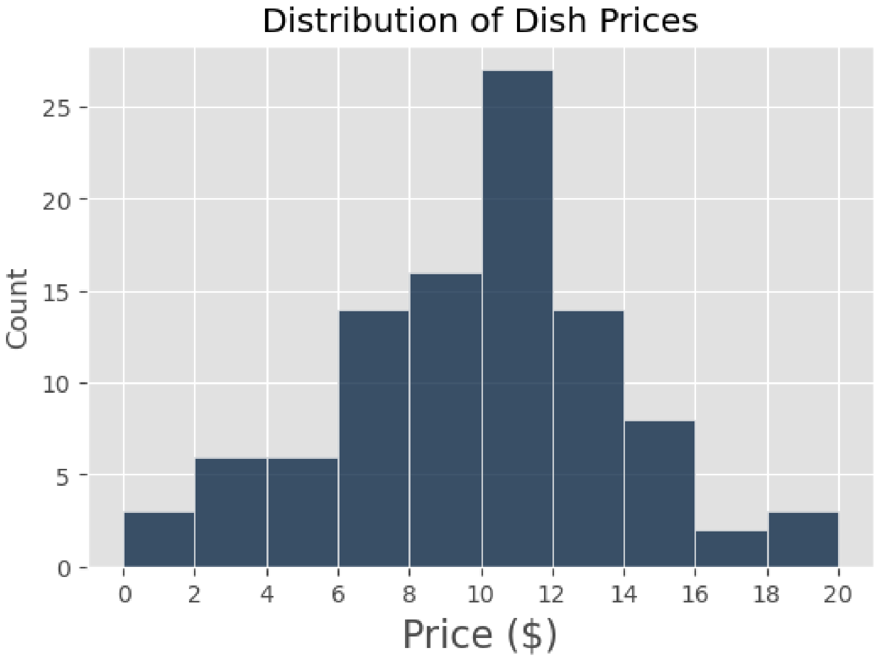
b. A histogram showing the price distribution of dishes, grouped by restaurant.

c. A bar chart showing the spiciness level for each dish across all restaurants.

d. A scatter plot showing the relationship between price and average rating, with different colors for each unique dish.

## 3. Interpreting Histograms

Histograms allow us to visualize the distribution of a single numerical variable by grouping numerical values into **bins** and encoding the **frequency** (count) of each bin as its height. While it is perfectly acceptable to combine histogram bins, you cannot split a bin as you don't know the distribution of values within a single bin.



Using the histogram above, answer the following questions.

a. What is the most common range of prices for dishes?

b. Approximately how many dishes cost $14 or more?

c. True or False: Most dishes cost at least $10.

d. True or False: More dishes cost between $5 and $6 than between $4 and $5.