Introduction

Grinnell College

September 6, 2024

Summarizing Data

Data collection has made remarkable progress in the last decades, giving us a greater quantity of data than most could ever dream of

In it's raw form, it becomes nearly incomprehensible to comprehend

Typically, we present either *numerical or visual summaries* of our data to facilitate interpretation

blah

Datasets

We'll be using a few datasets to help illustrate our points today

1. Tips data
2. 2019 College Survey
3. `mtcars` 1979 Motor Trend magazine

Univariate Visual Summaries

Much in the same way that all people have a height and weight, or how all combustion engines have a displacement and a number of cylinders, we can expect data and variables to have certain attributes, depending on their type

One attribute that is relevant for *all* types of data is its **distribution**.

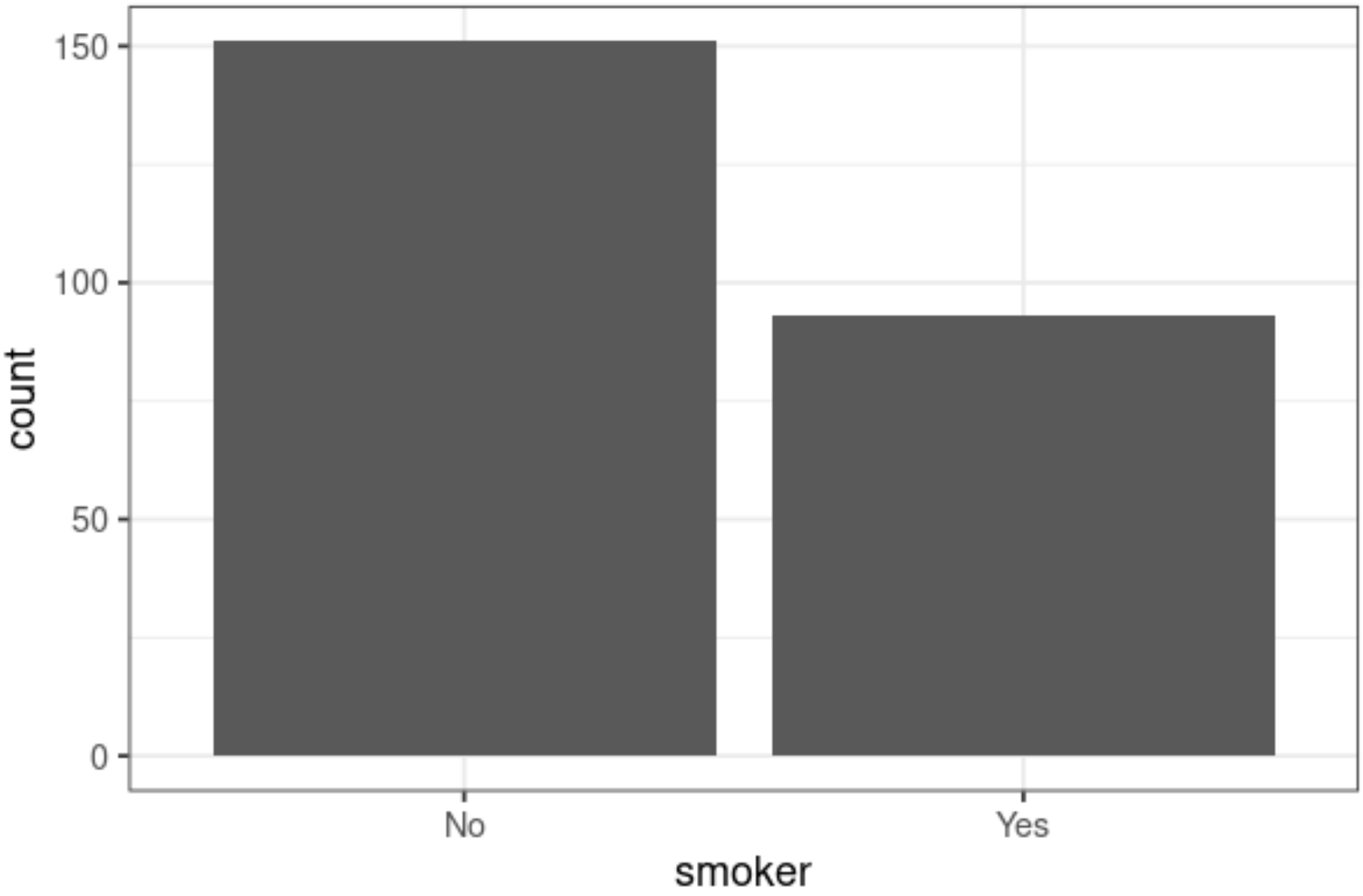
A distribution tells us two things:

1. What values does are data take?
2. How frequently do those values appear?

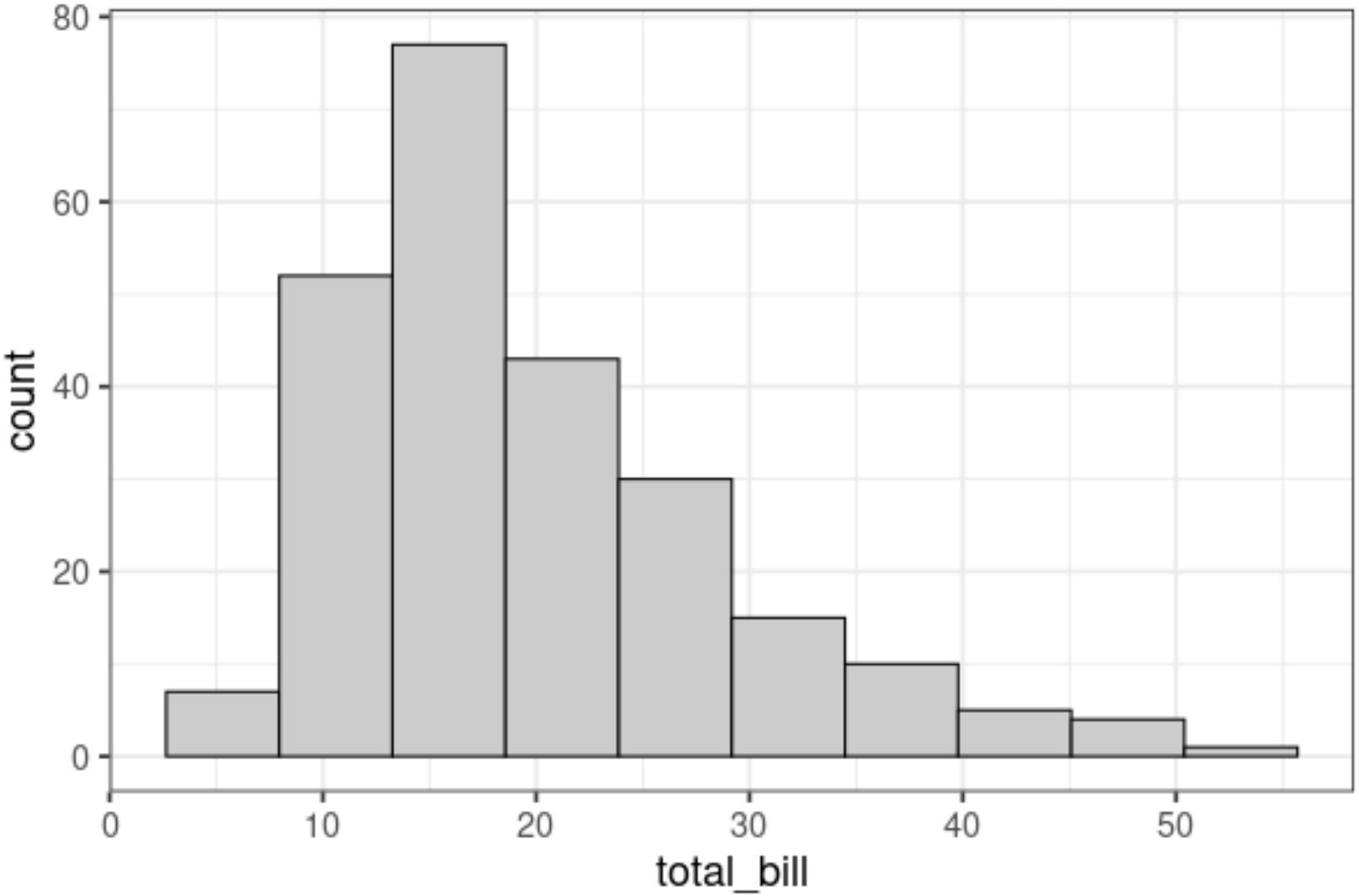
Outline

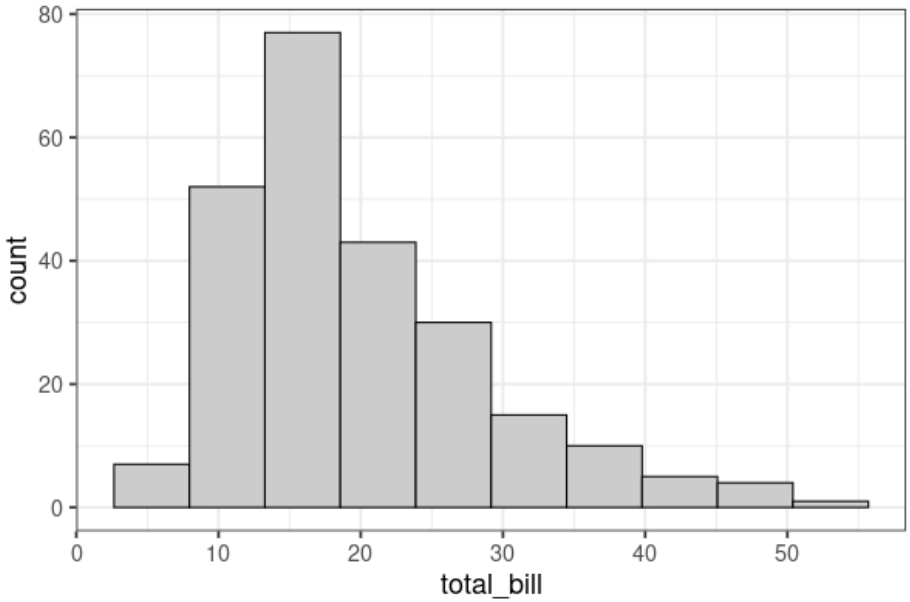
1. Univariate plots
   - Useful for showing distribution of data
   - Often provides insight into other relevant characteristics, according to data type
2. Bivariate plots
   - Most useful for demonstrating relationship between two variables
   - Often framed as indepenent/dependent variables (prediction)
3. Multivariate plots
   - Visualize relationship between three or more variables
   - Help identify trends that exist across category values

Bar plots (Univariate)

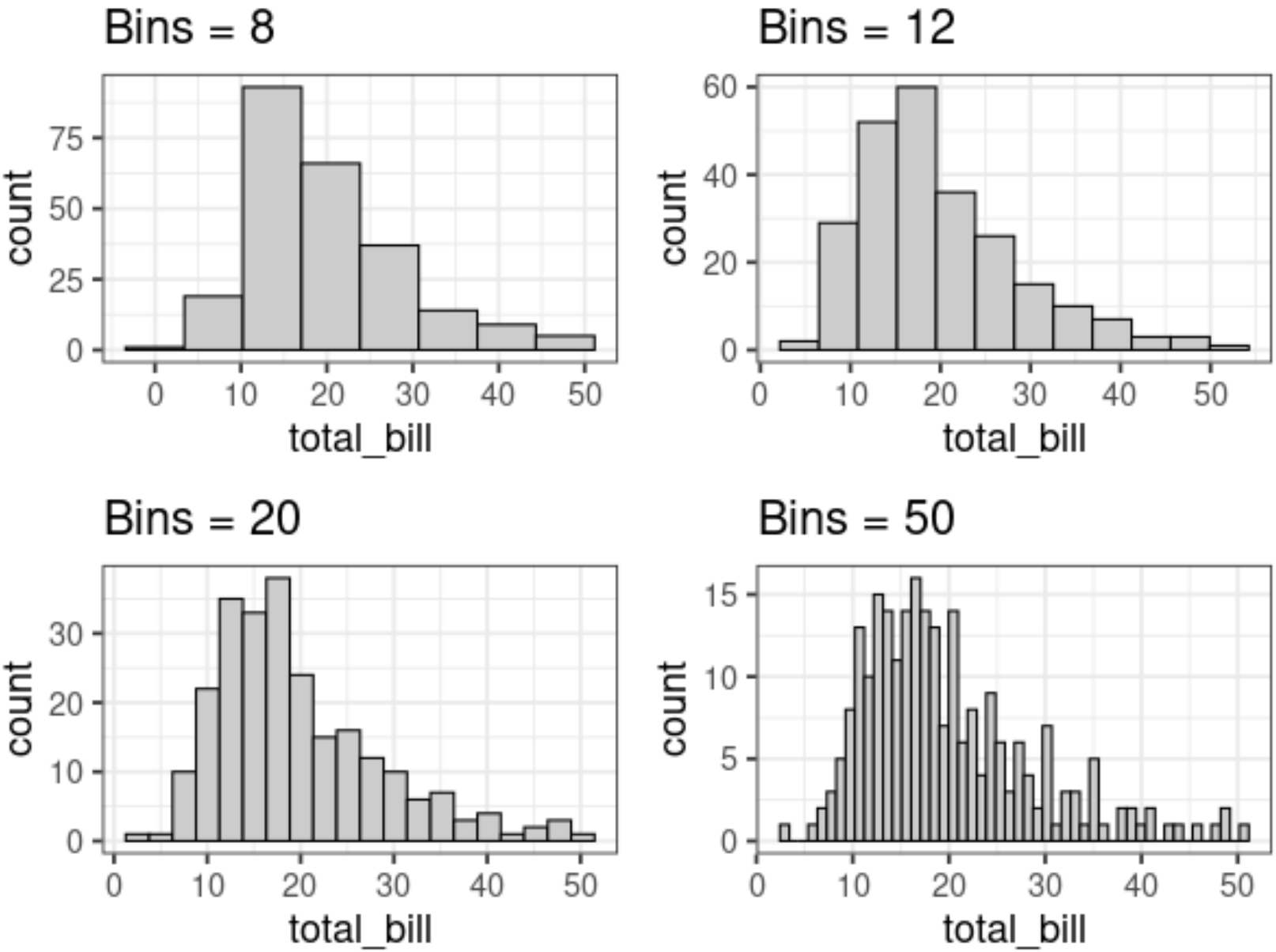


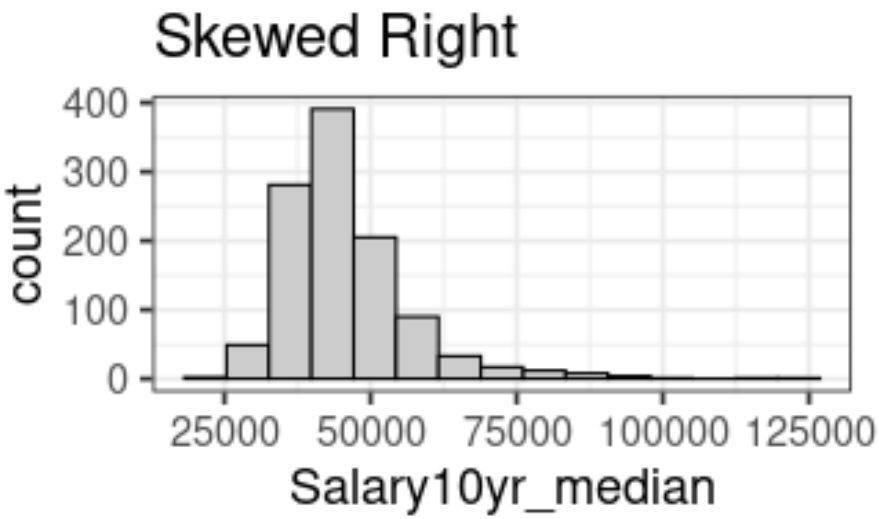
Histogram (Univariate)

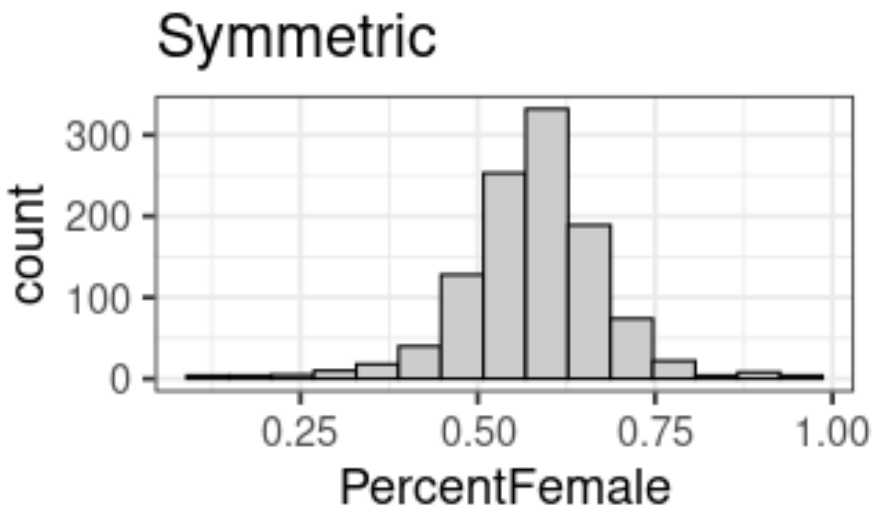


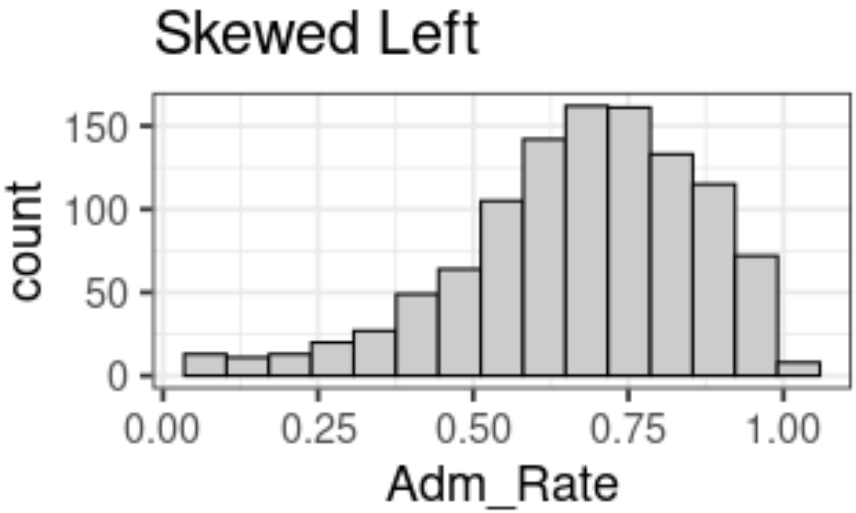


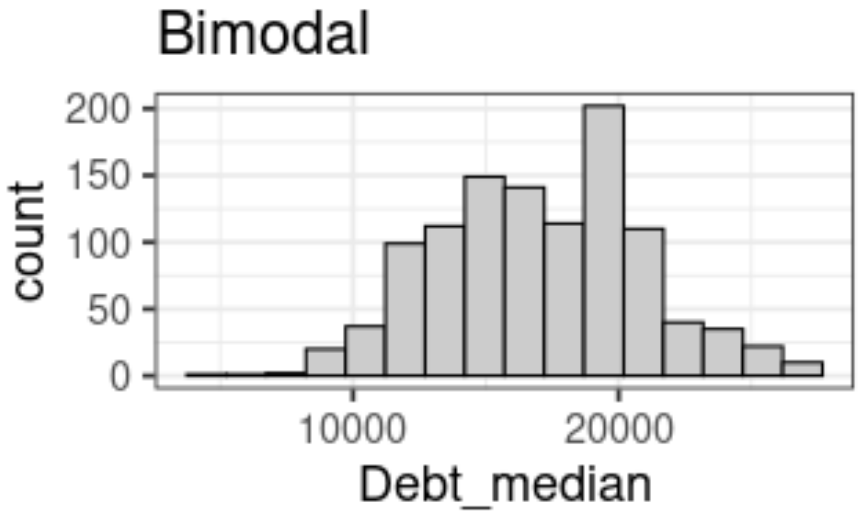
- Number of bins
- Distribution properties
  - Center
  - Dispersion
  - Shape


Bins = 8
count
total_bill
Bins = 12
count
total_bill
Bins = 20
count
total_bill
Bins = 50
count
total_bill

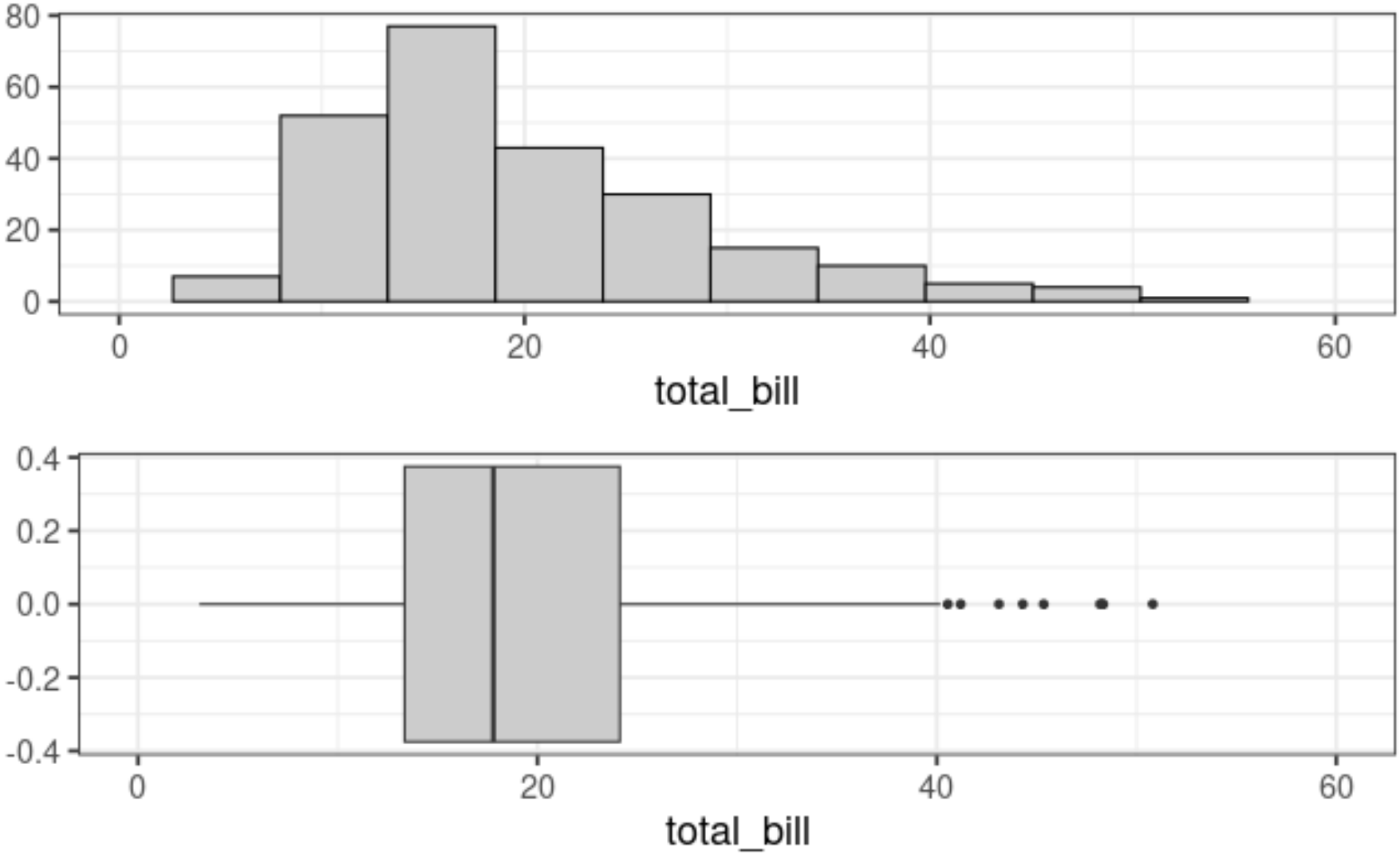

Skewed Right
count
400
300
200
100
0
25000
50000
75000
100000
125000
Salary10yr_median


Symmetric
count
300
200
100
0
0.25
0.50
0.75
1.00
PercentFemale


Skewed Left
count
150
100
50
0
0.00
0.25
0.50
0.75
1.00
Adm_Rate


Bimodal
count
200
150
100
50
0
10000
20000
Debt_median

Box plots (Univariate)



Bivariate Plots – Association

Bivariate plots are used to illustrate the *relationship* between two variables. Typically, we are intereted in knowing if two variables are **associated**

We say that two variables are **associated** if knowing the value of one variable gives us information about the other

1. Are the "centers" similar between groups?
2. Does an change in one tend to suggest a systematic change in another?
3. Are the proportions of outcomes similar between categories?

Variables that are not associated are said to be **independent**

Bivariate Plots – Explanatory and Response

Often when discussing the relationship between two variables, we will designate one variable as an **explanatory/independent variable** and the other as a **response/dependent variable**

This language can be a bit misleading as it is *not* intended to suggest a causal relationship between the two, i.e., if X is our explanatory variable and y is our response, we do not mean to suggest X *causes* y

Rather, we think:

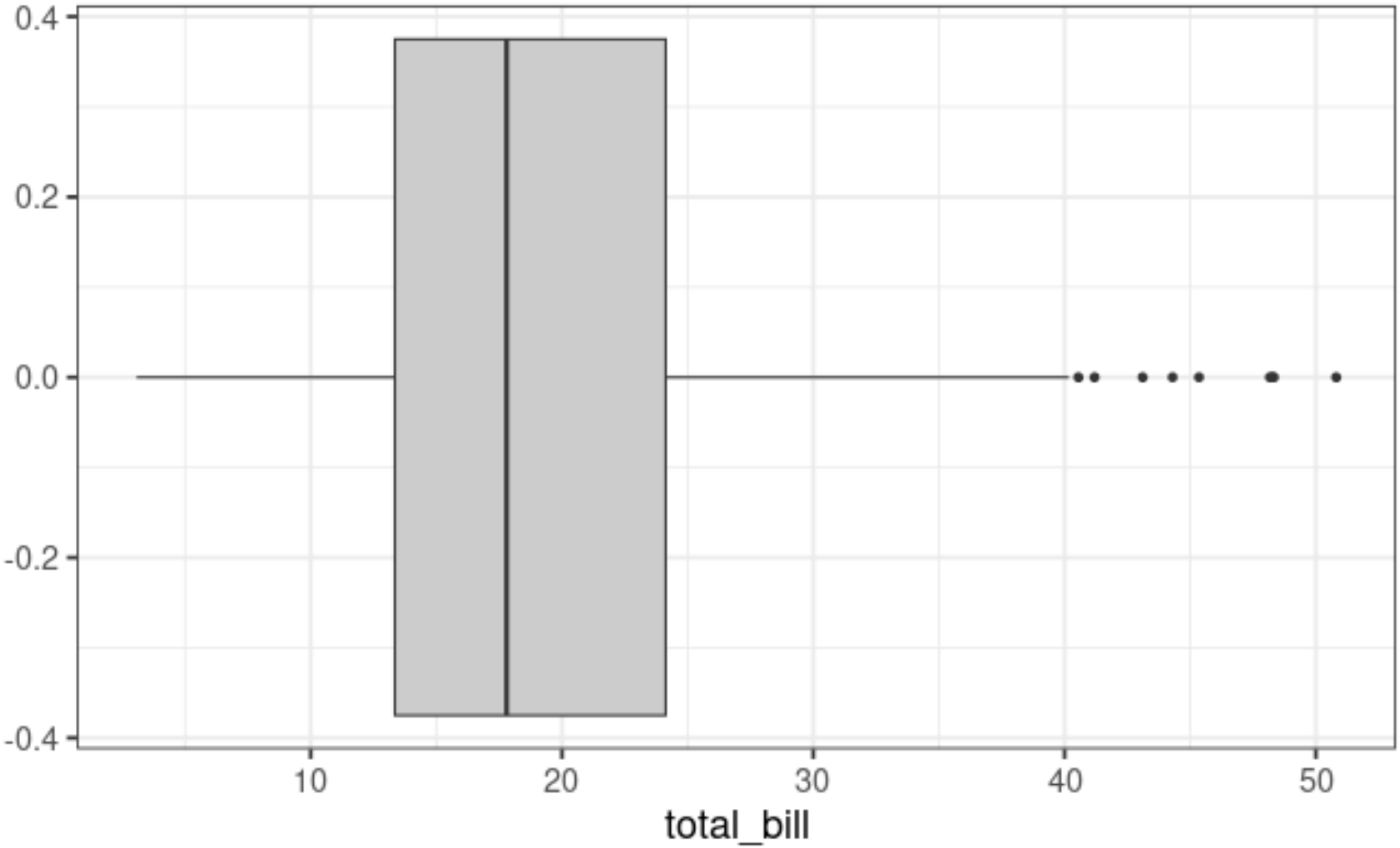
"Given this value of X, we can predict this value for y"

Bivariate Summaries

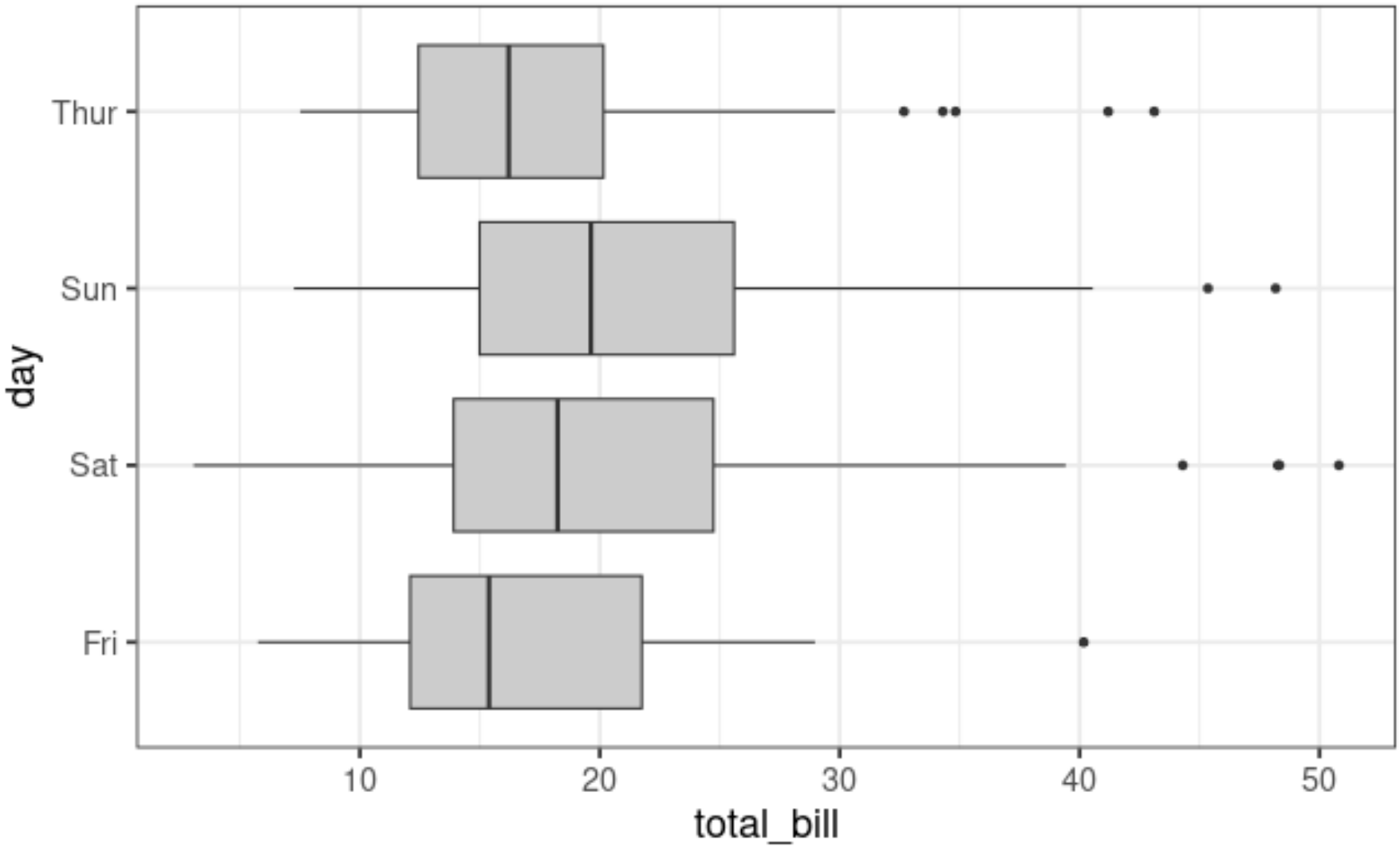
By the very nature of it, we have the following arrangements of variables and their associated plots:

1. Categorical and quantitative
   - Boxplots
2. Quantiative and quantitative
   - Scatterplots
3. Categorical and categorical
   - Stacked bar charts
   - Clustered bar charts
   - Conditional/proportional bar charts

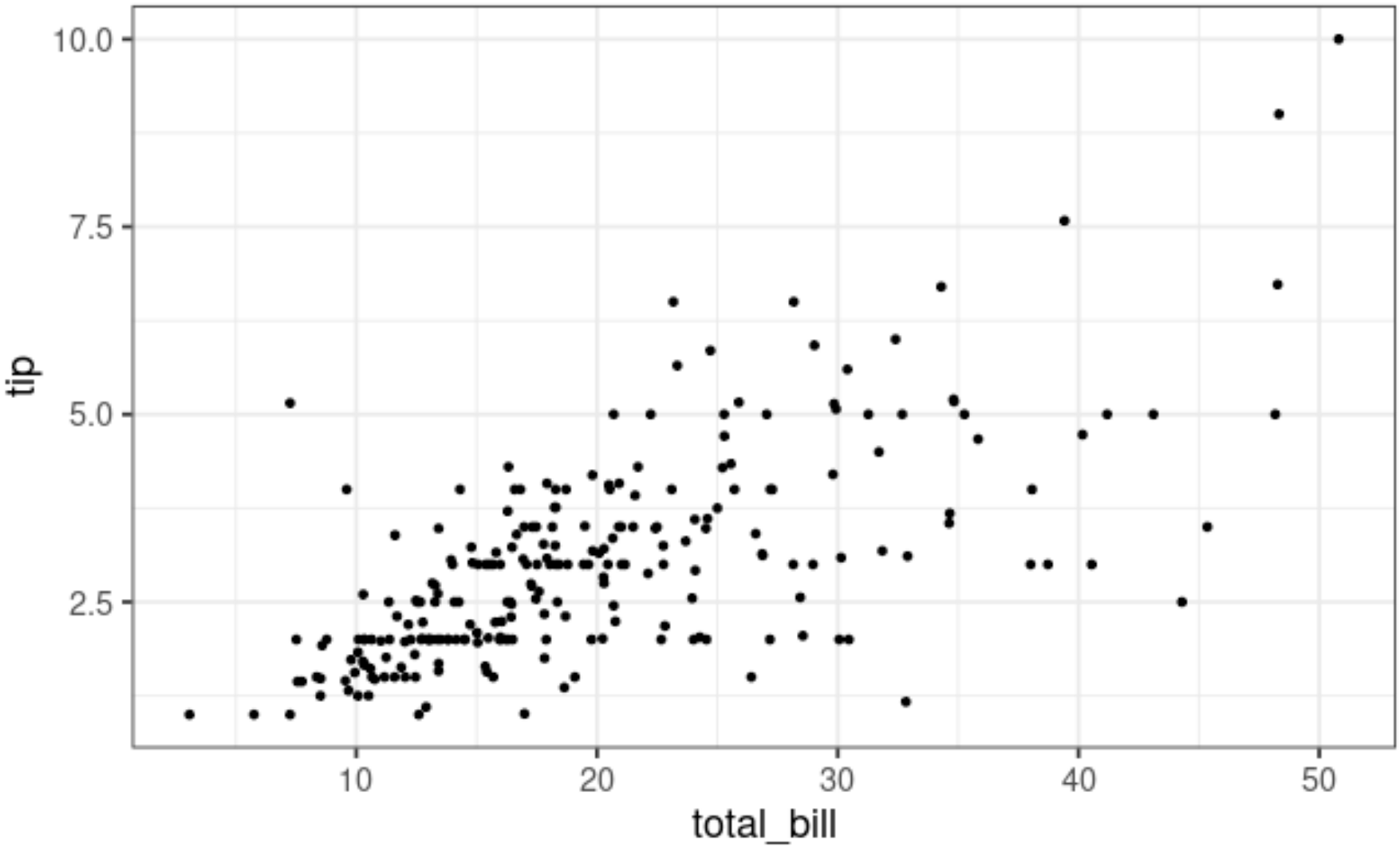
Categorical and Quantitative (Bivariate)

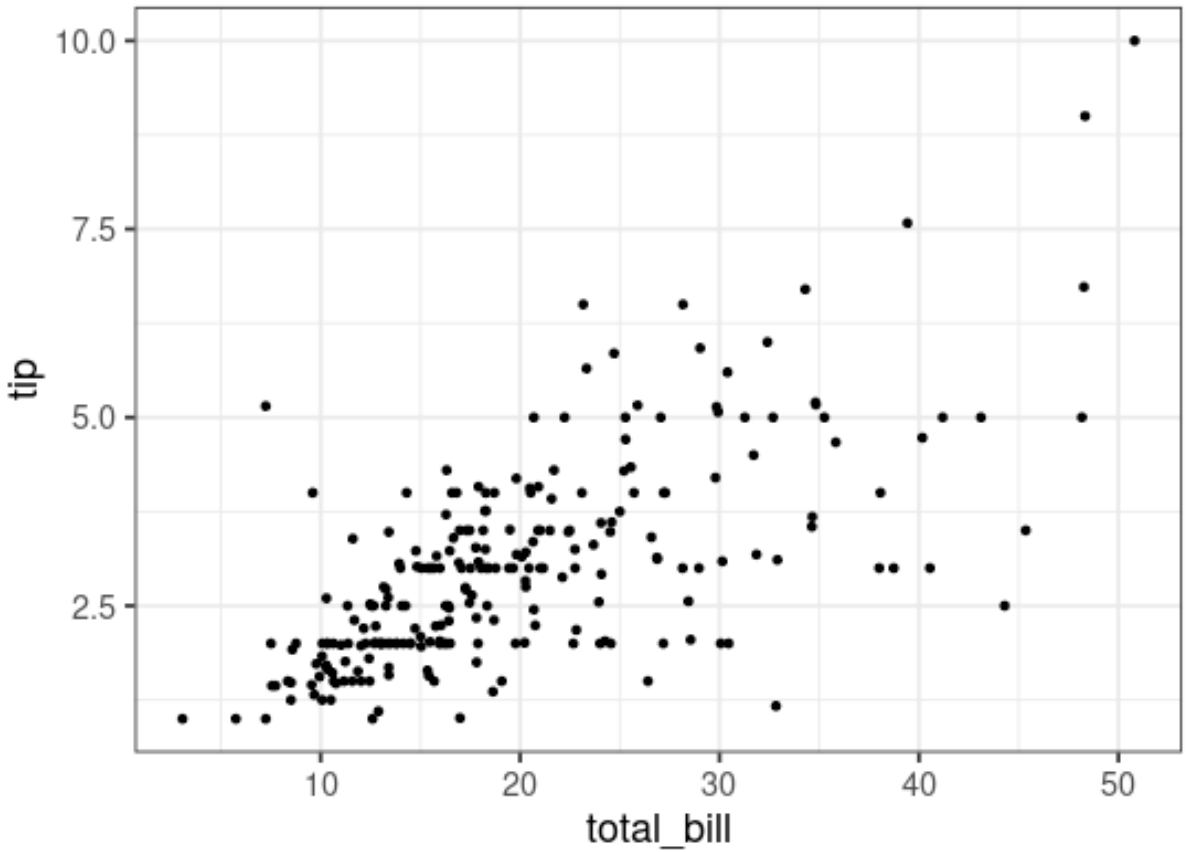


Categorical and Quantitative (Bivariate)



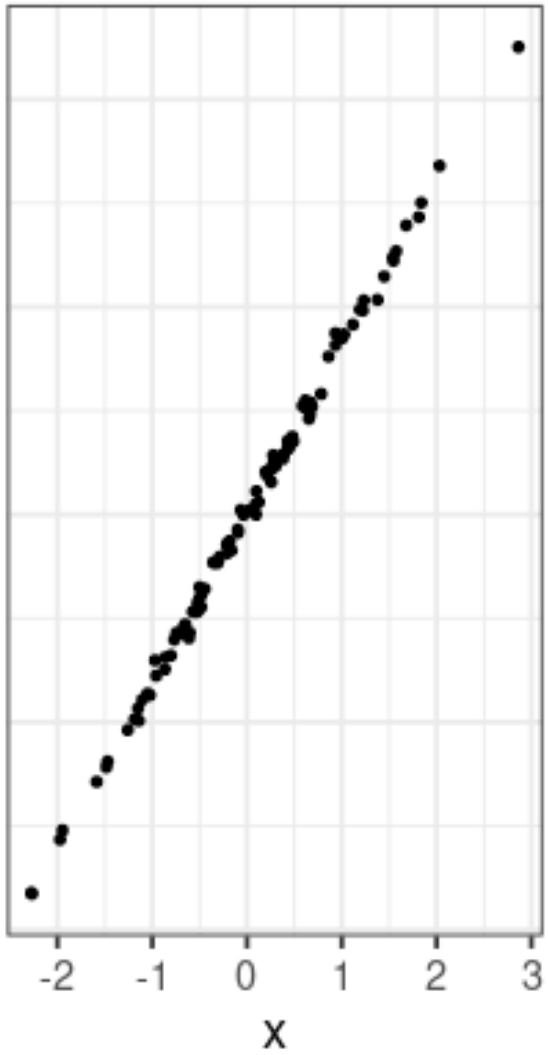
Quantitative and Quantitative (Bivariate)

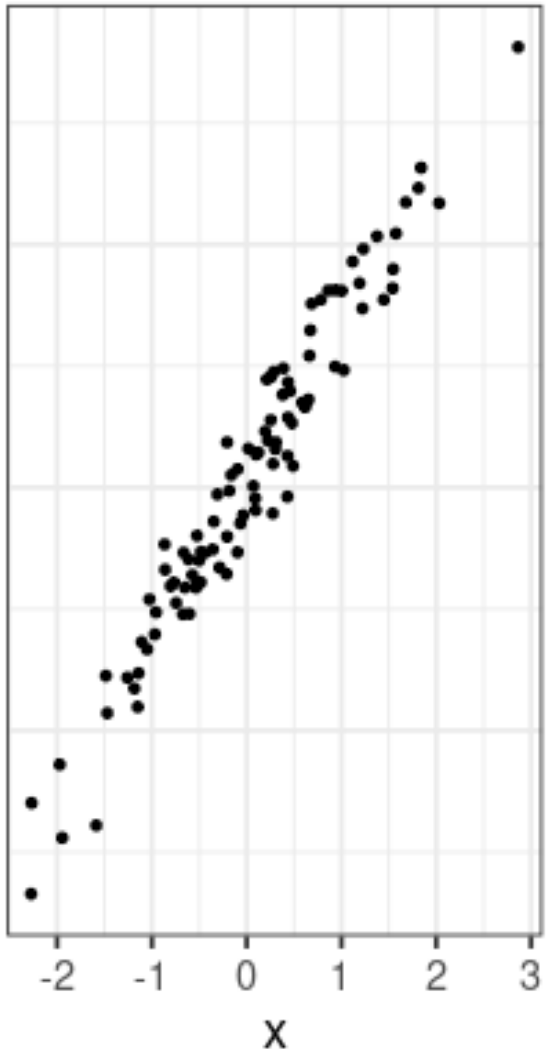


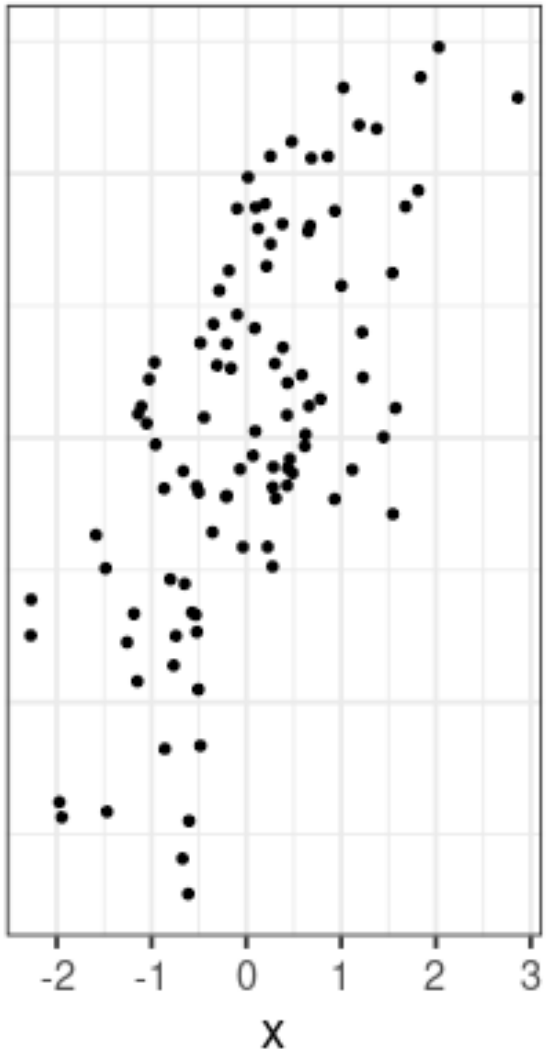


- Strength
- Form
- Direction

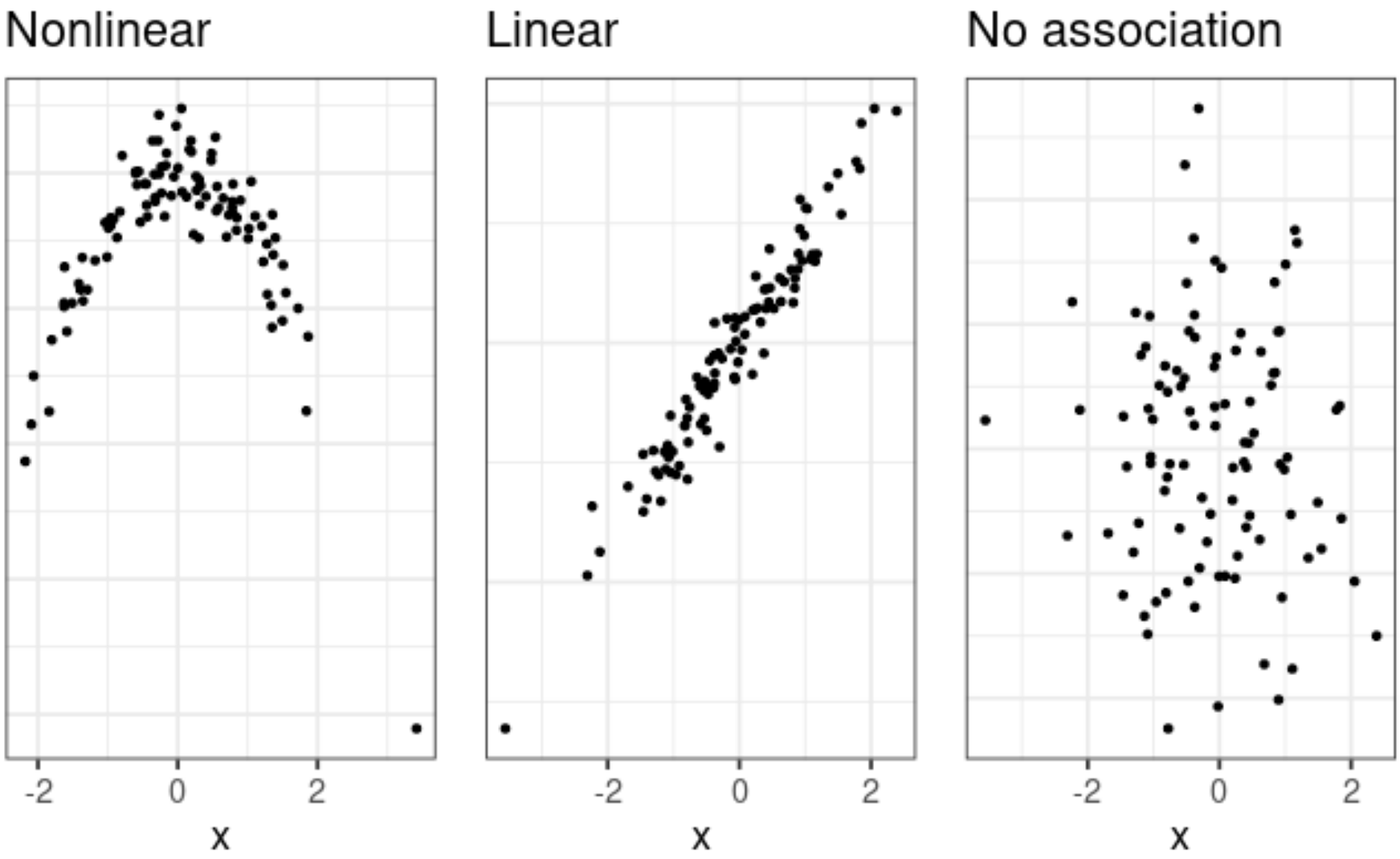
Strength



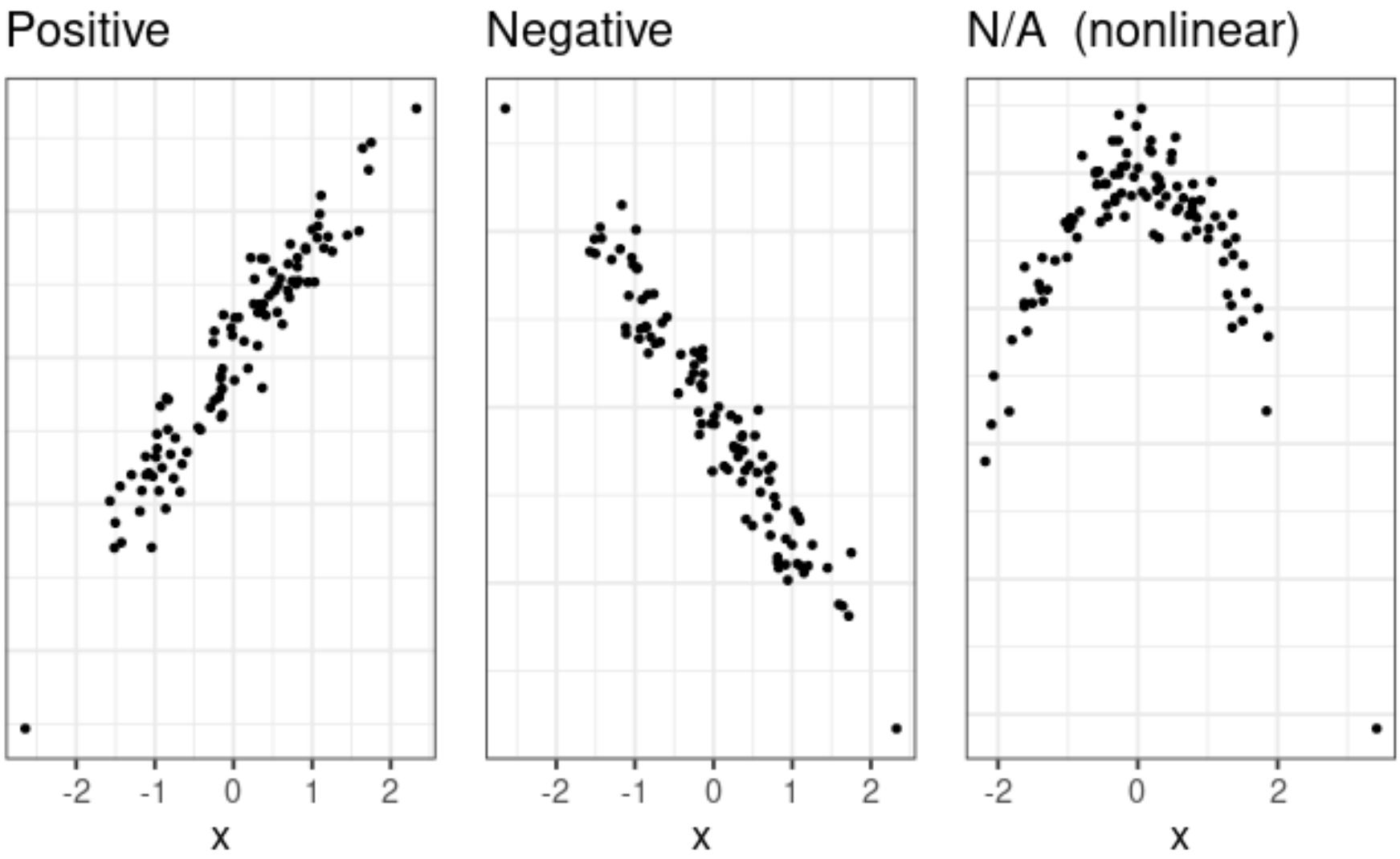




Form



Direction



Bar Plots

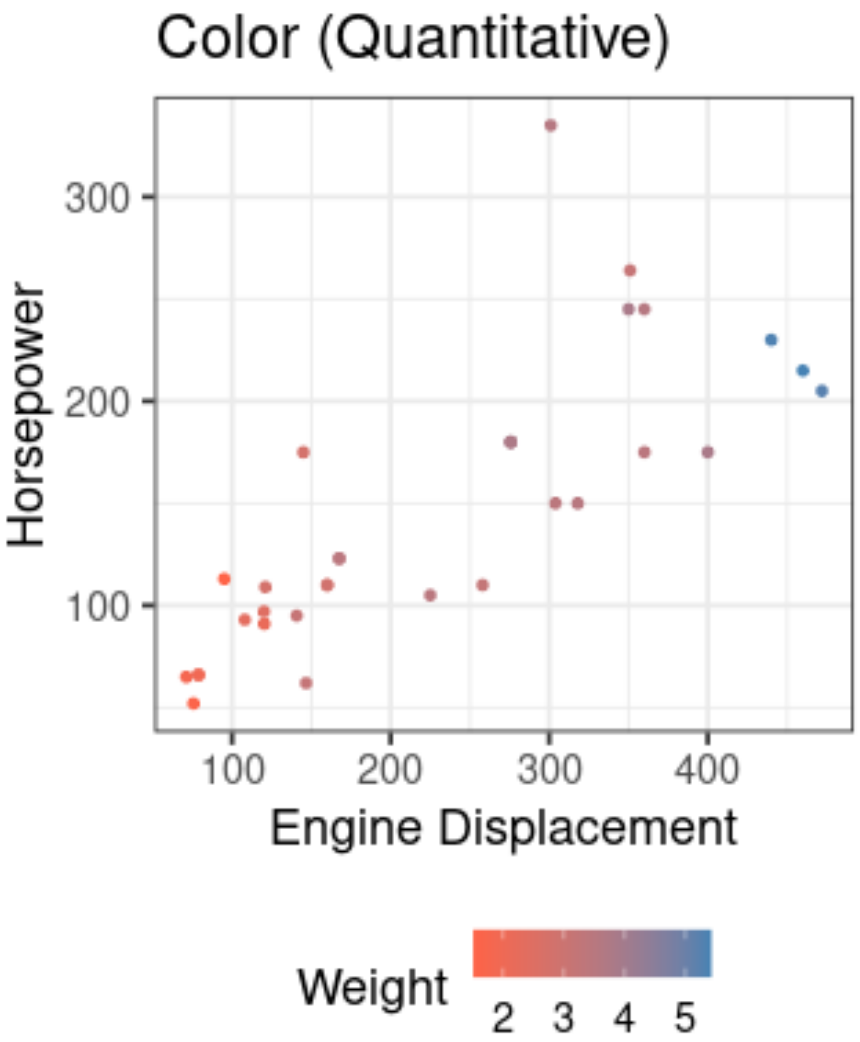
We will do this next week

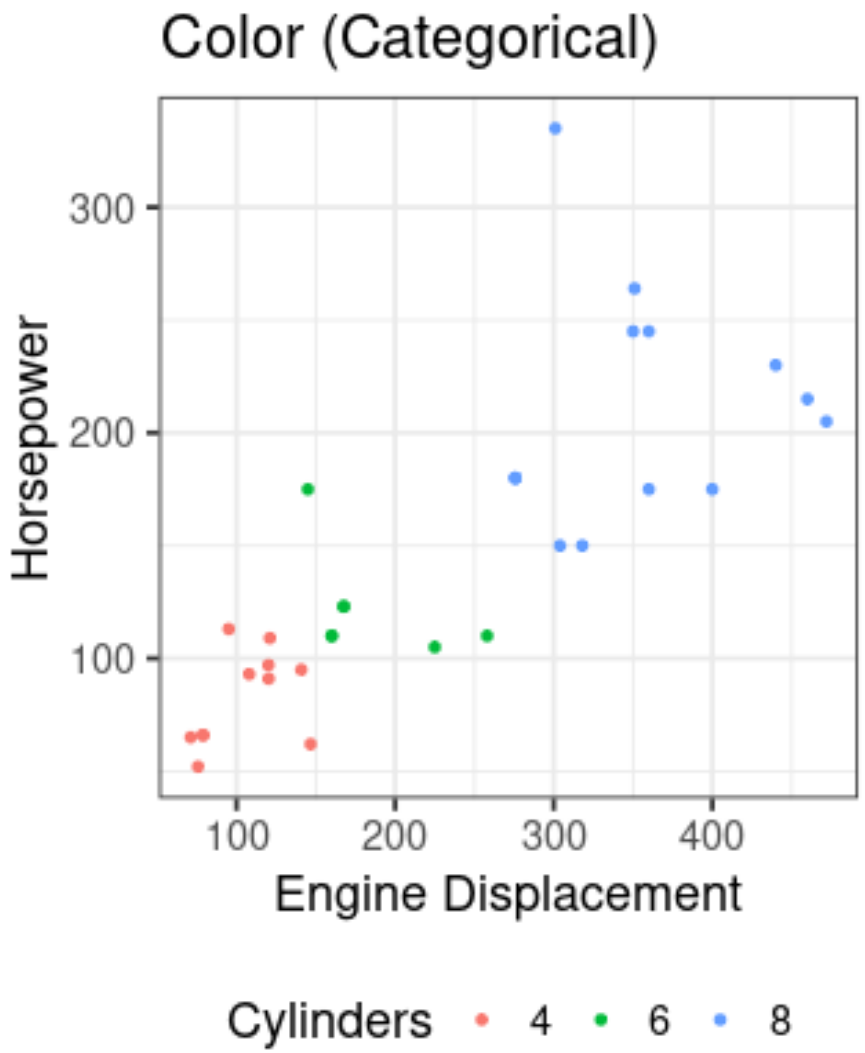
Multivariate plots

Multivariate plots are those illustrating the relationships between three or more variables

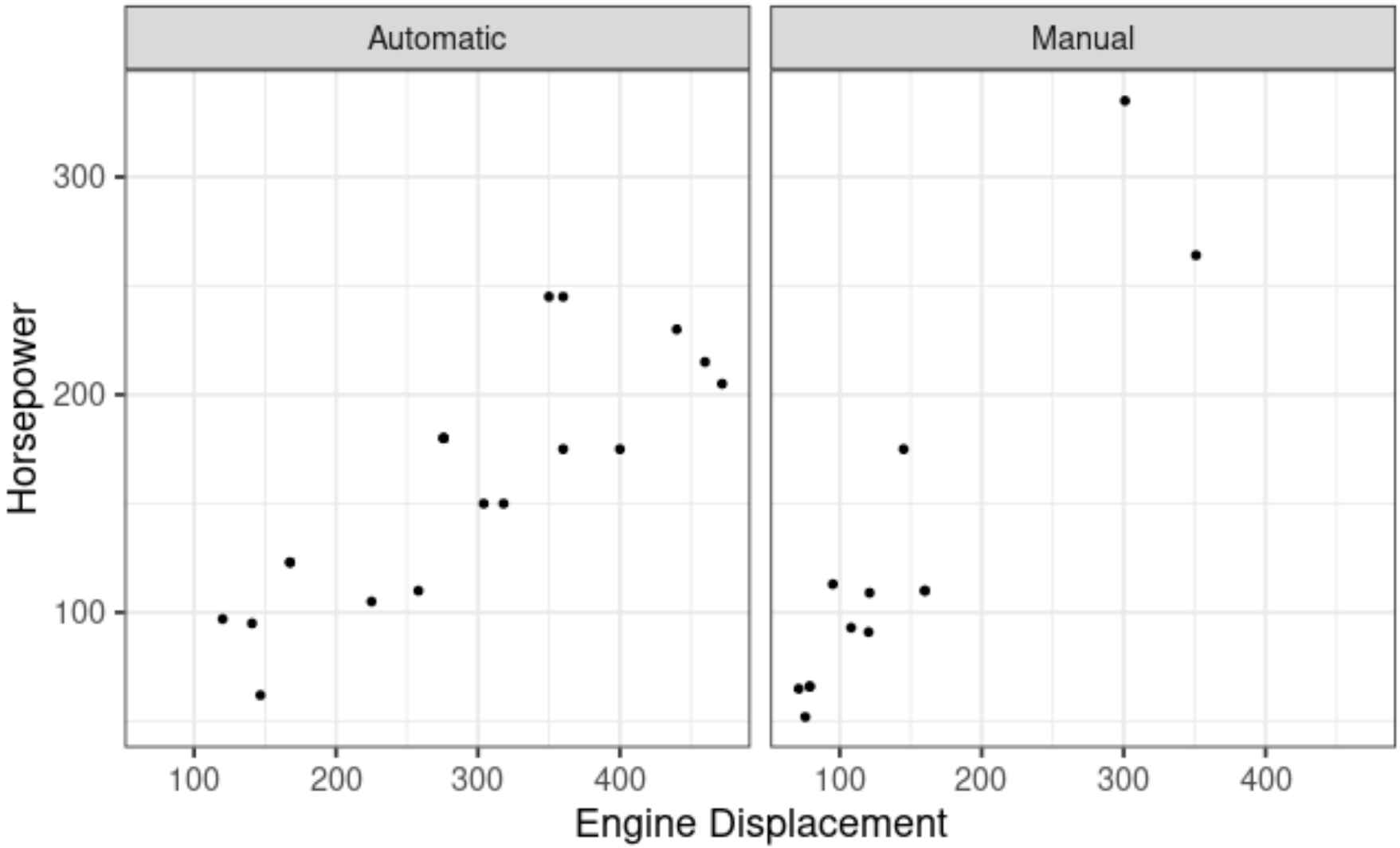
In the langauge of `ggplot2`, this equates to creating additional mappings from our dataset through the use of either *aesthetics* or *faceting*

Color aesthetic

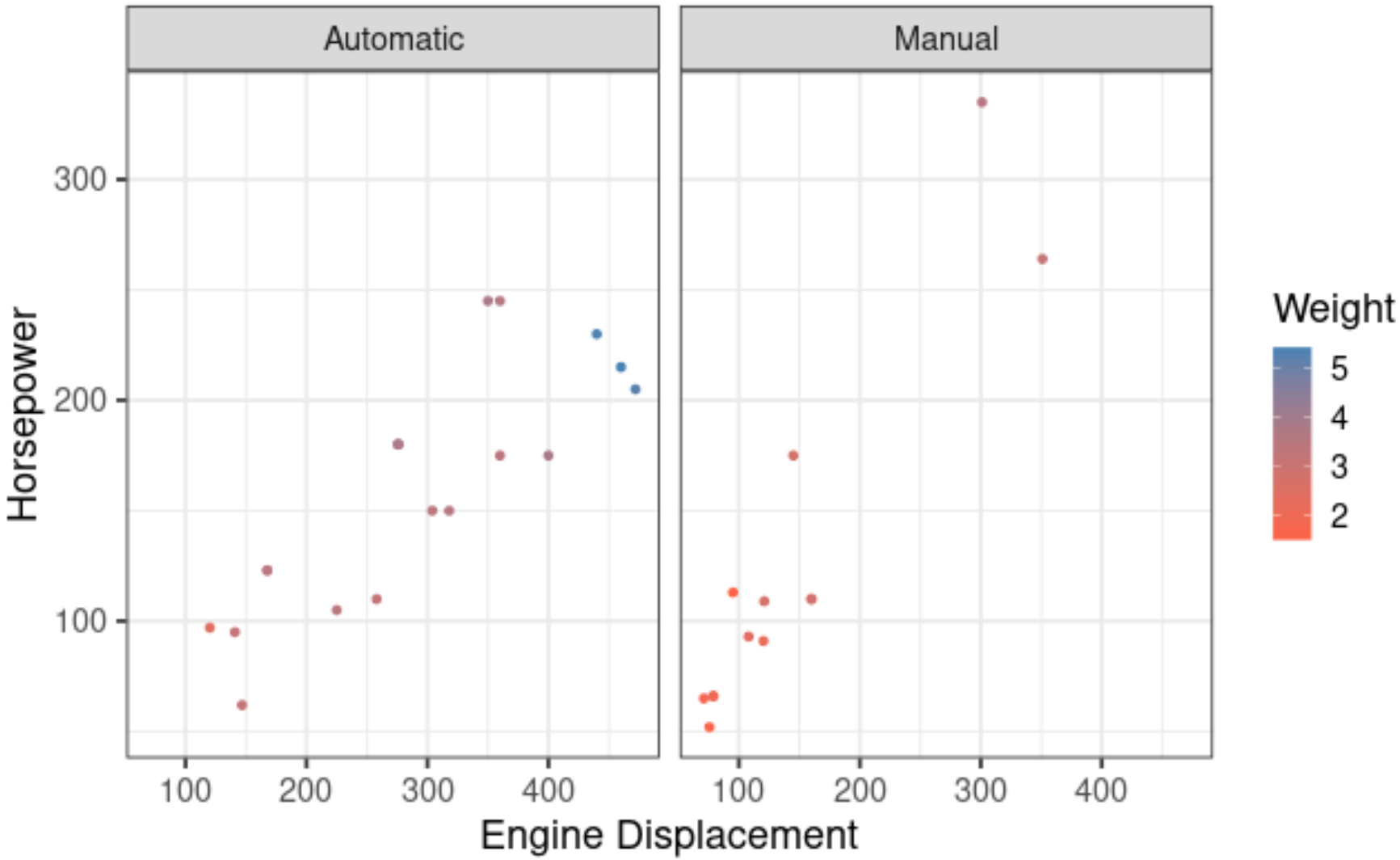




Faceting


Automatic
Manual
Horsepower
300
200
100
100
200
300
400
100
200
300
400
Engine Displacement

Multivariate Plots



Summary

We need to be able to correctly identify which types of plots are associated with which types of variables

Plots are exceptionally important in helping familiarize ourselves with new datasets as they can quickly and clearly summarize large quantities of information 4mm
Questions we can ask ourselves:

- What type of variable(s) am I working with?
- What information about these variables do I want to know? Distribution? Relationship? Other?
- What plots best convey this information to myself and others?