Lecture 1
Data Basics, Scatterplots, Correlation Coefficient
STAT 432, Spring 2021

Data Tables

- Statisticians usually prepare data as tables, where the columns are the variables and the rows are the individual cases or **observations**.
- A **variable** can be thought of as a characteristic of an observation.

Data Tables

Here is an example of a data set in R called `mtcars` that was originally from the 1974 *Motor Trend* magazine. The columns are variables on automobile design and performance (e.g., mileage, weight, horsepower). The rows are the different automobile models.

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1

> dim(mtcars)
[1] 32 11
```

The `head()` command is used to preview the first several rows of the data table, and the `dim()` command gives the dimensions (32 car models and 11 variables).

Variable Types

- **Numerical variables** take on numerical values and that are usually measurements or counts. It makes sense to take the sum or mean of a numerical variable.
  - For example, mileage (`mpg`) and weight (`wt`) are numerical variables.
- **Categorical variables** take on values that fall into distinct categories.
  - For example, transmission type (`am`) is a categorical variable since each car model either has an automatic (coded as 0) or manual transmission (coded as 1).

Scatterplots

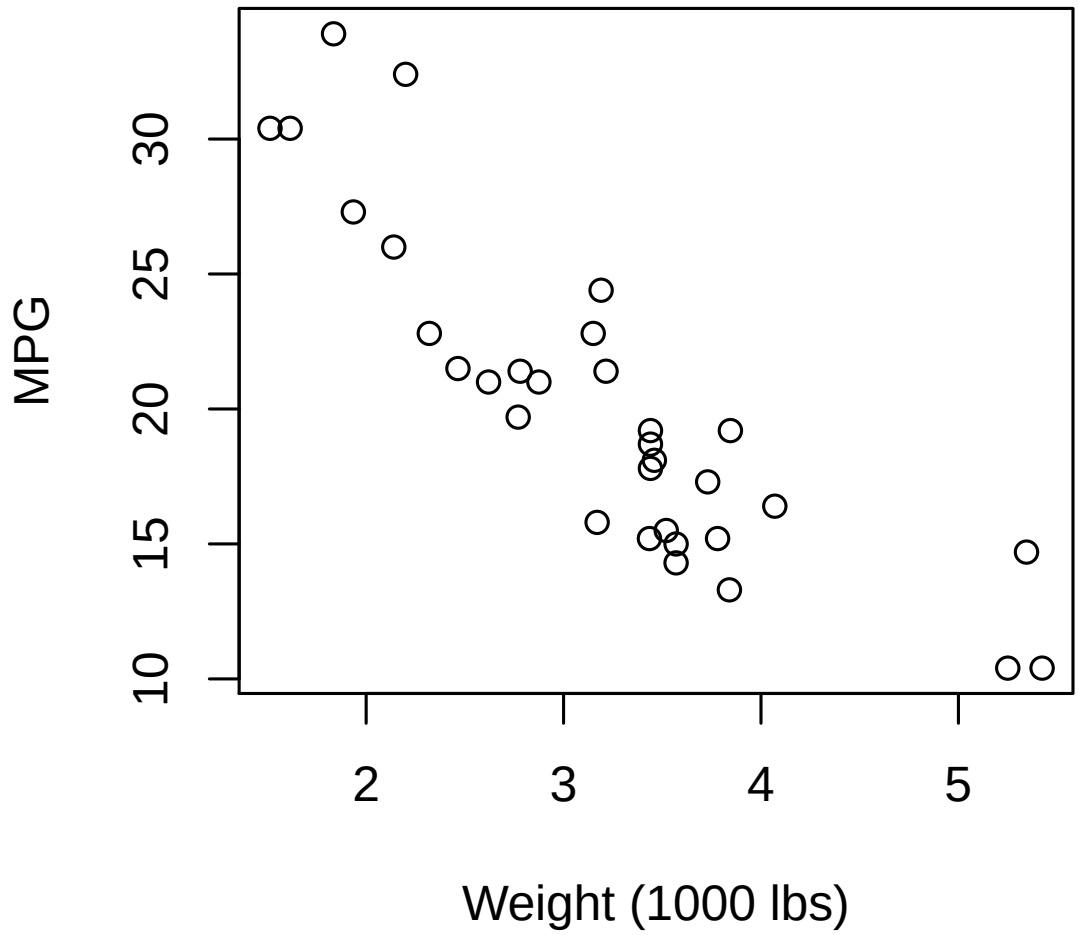
- A scatterplot a graphical display used to study the relationship between two variables X and Y.
- Data displayed on a scatter plot are collected in pairs:

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

  where n denotes the total number of cases or pairs.
- Y is commonly called the **response variable** and X the **explanatory variable**.
- A scatter plot will provide insight into how two variables are related.

Example

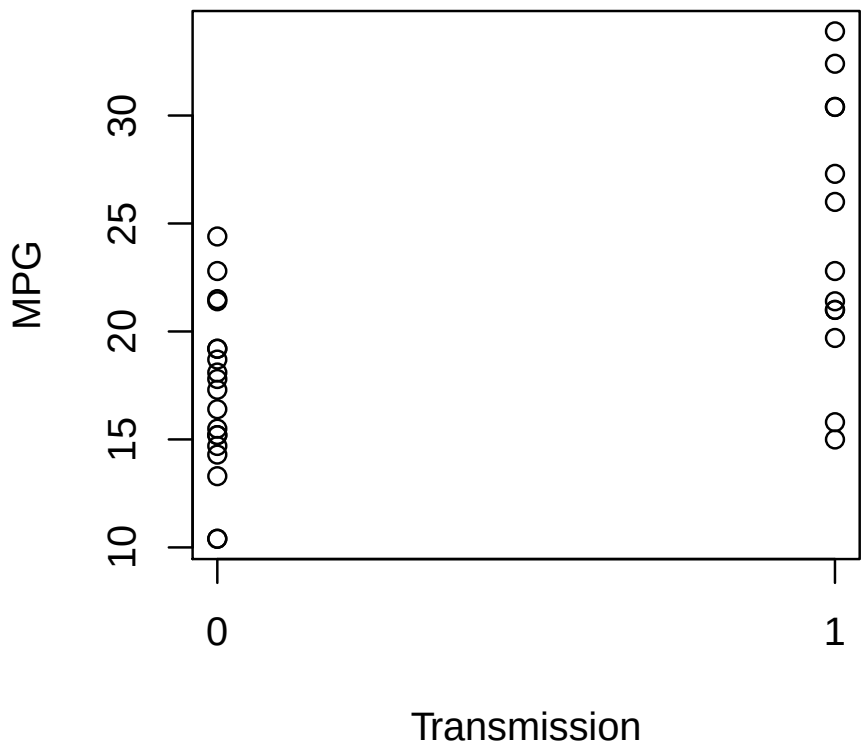
```
> plot(mtcars$wt, mtcars$mpg,
    xlab='Weight (1000 lbs)', ylab='MPG')
```



Example

We can use a scatter plot when X is categorical. Here transmission type (X) is coded as automatic $= 0$ and manual $= 1$.

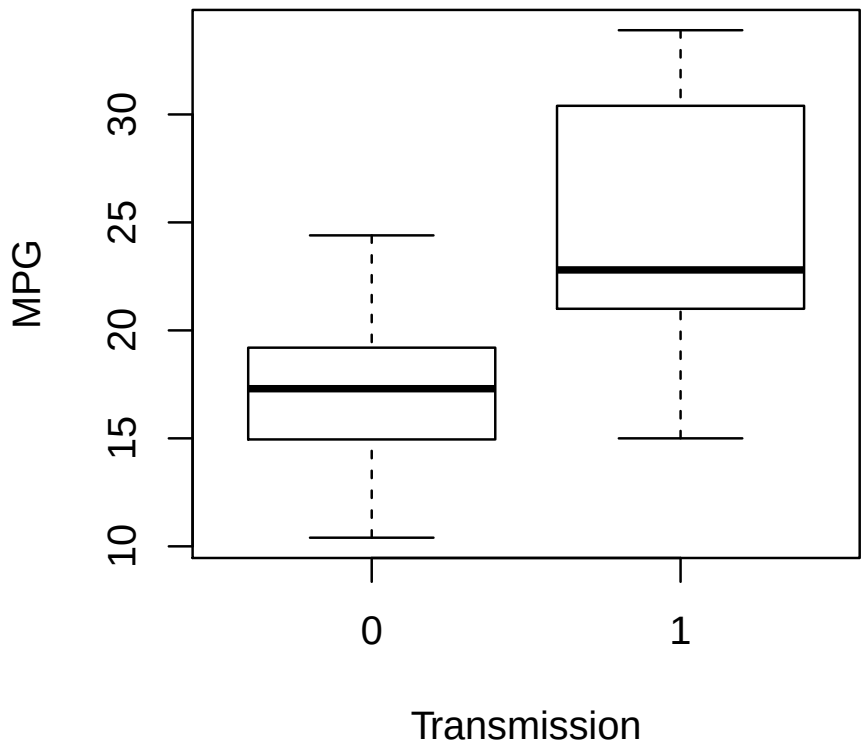
```
> plot(mtcars$am, mtcars$mpg,
    xlab='Transmission', ylab='MPG')
```



Example

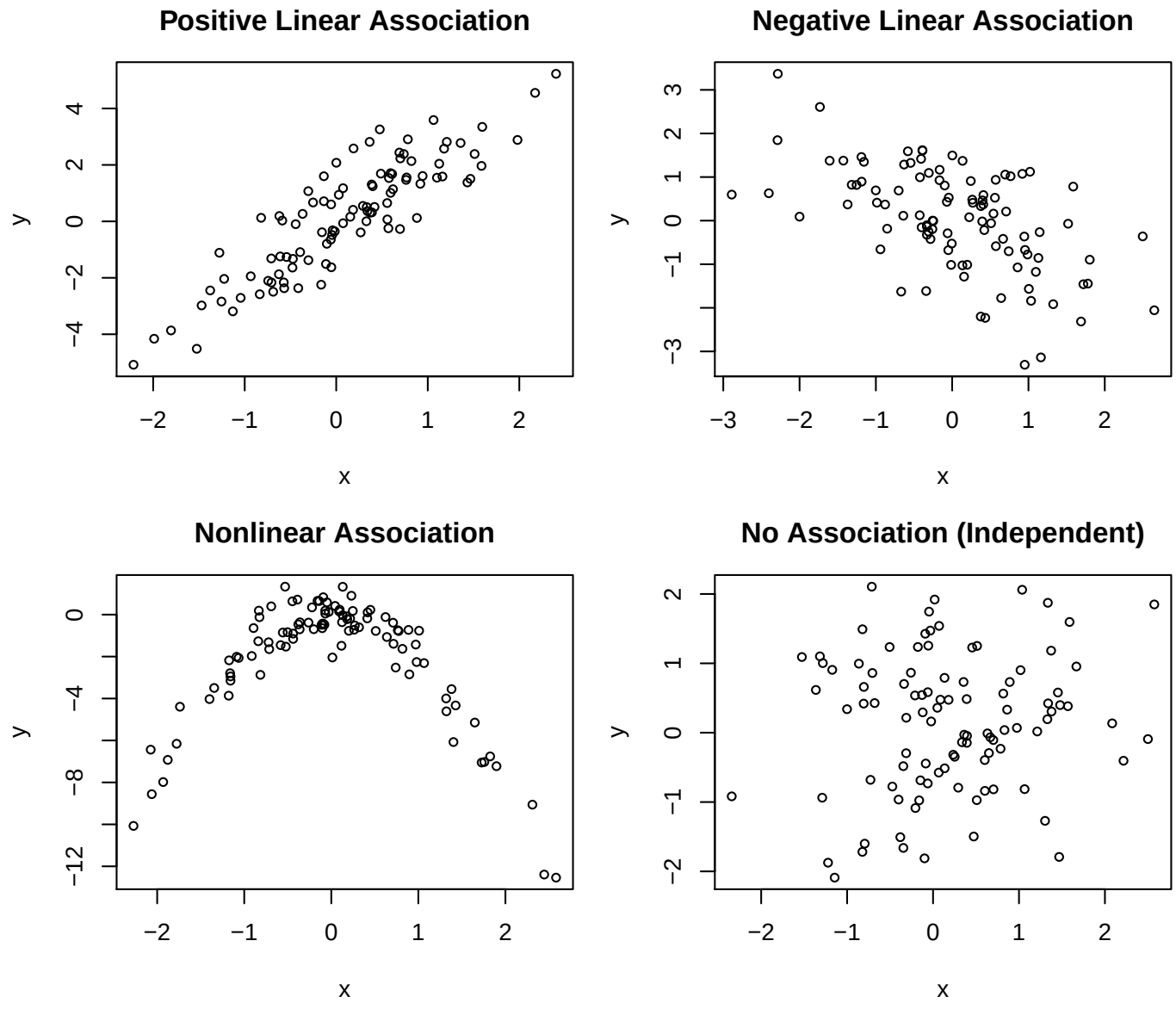
Boxplots can also be used as a more informative display when X is categorical.

```
> boxplot(mtcars$mpg~mtcars$am,
    xlab='Transmission', ylab='MPG')
```



Types of Relationships between Variables

- Two variables are said to be **associated** if the scatterplot shows a discernible pattern or trend.
- An association is **positive** if Y increases as X increases.
- An association is **negative** if Y decreases as X increases.
- An association is **linear** if the scatterplot between X and Y has a linear trend; otherwise, the association is called **nonlinear**.


Positive Linear Association
Negative Linear Association
Nonlinear Association
No Association (Independent)
x
y

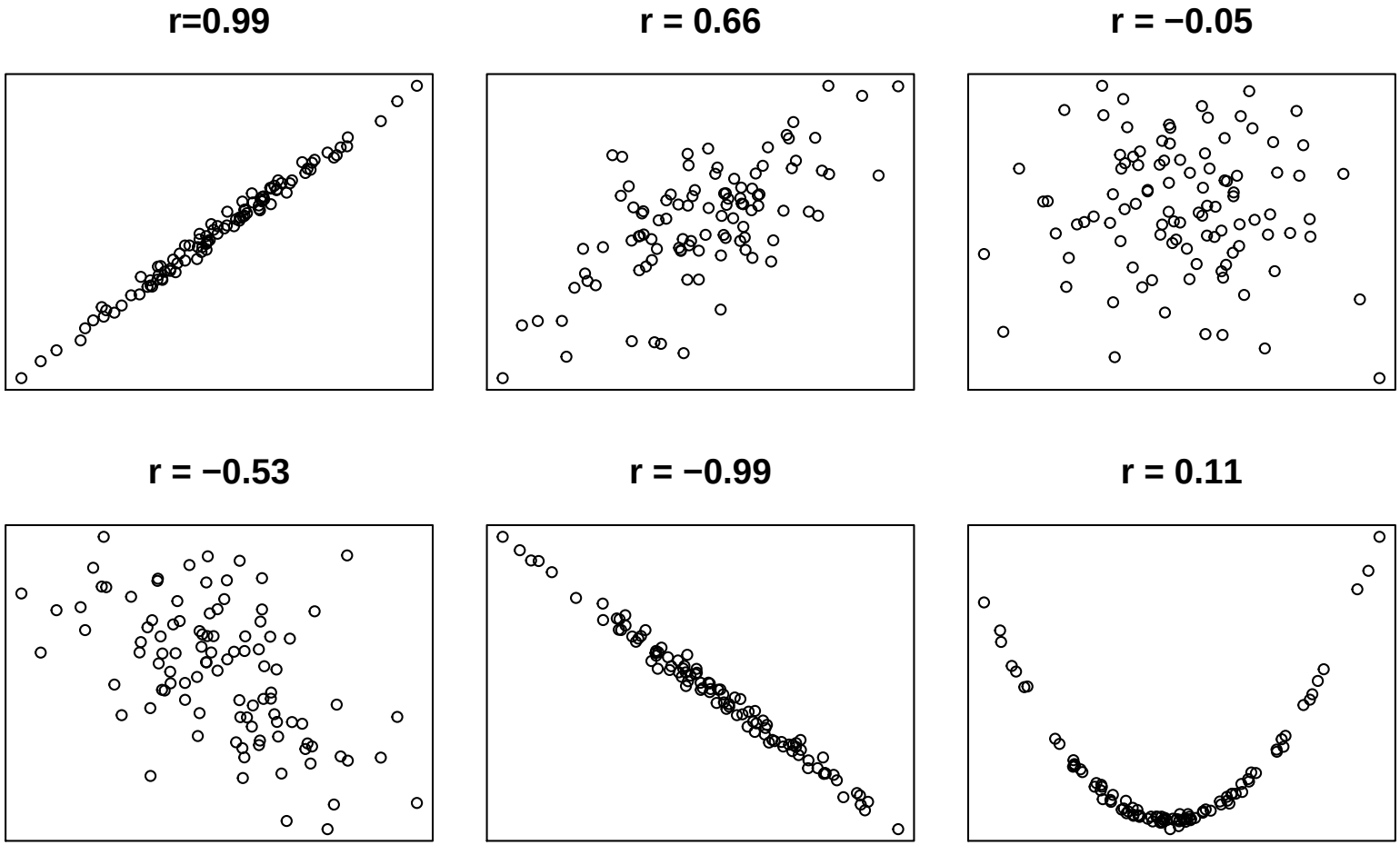
Correlation Coefficient

The **correlation coefficient**, denoted by r, is a number between -1 and 1 that describes the strength of the linear association between two variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- $\bar{x}$ and $\bar{y}$ are the sample means
- s_x and s_y are the sample standard deviations

Correlation Coefficient



Correlation Coefficient

- $r \approx 1$ when there is a strong positive linear association between the variables.
- $r \approx -1$ when there is a strong negative linear association between the variables.
- $r \approx 0$ when there is no association between the variables (i.e., independent).
- The correlation coefficient is only useful for evaluating the linear association between two variables. It is not a useful measure for nonlinear relationships.

Exercise

Match each correlation to the corresponding scatterplot.

(a) $r = -0.63$

(b) $r = 0.85$

(c) $r = 0.19$

(d) $r = 0.95$