Correlation and Linear Regression

Why?

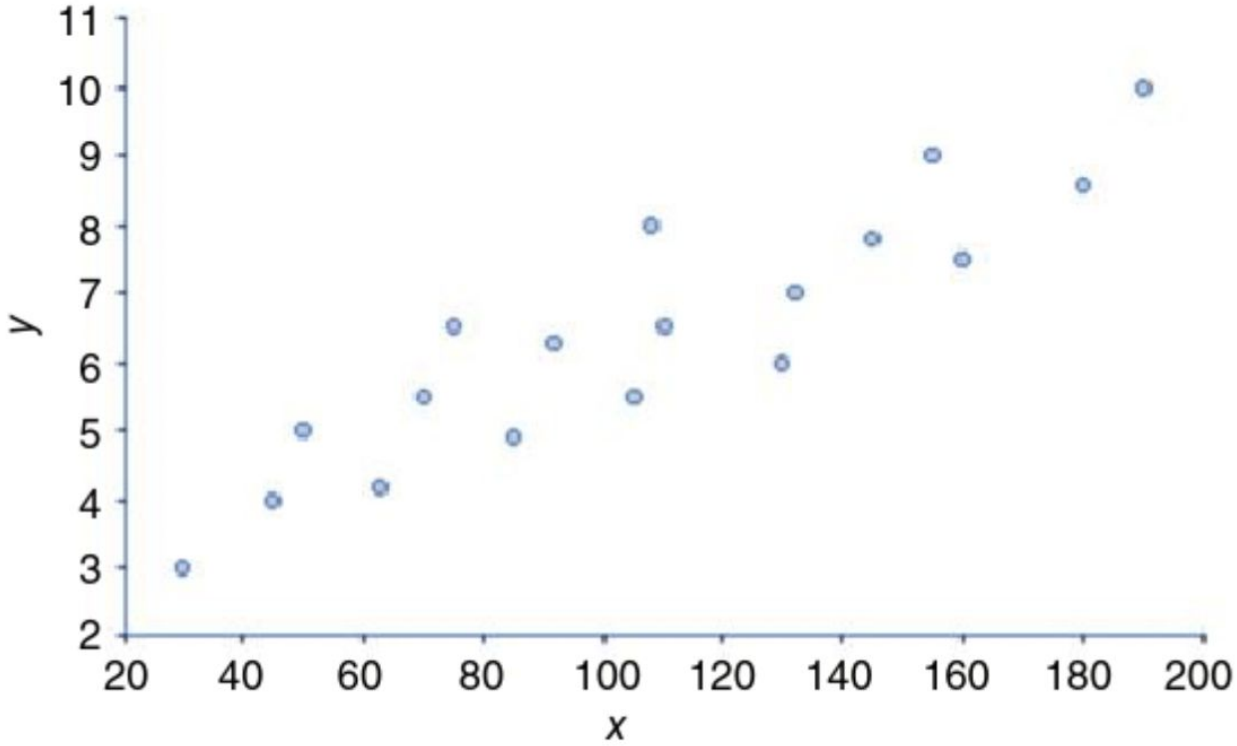
Comparing two variables, and finding the relationship between them useful for data storytelling and EDA

Typical visualizations associated with two variables

Scatter plot!

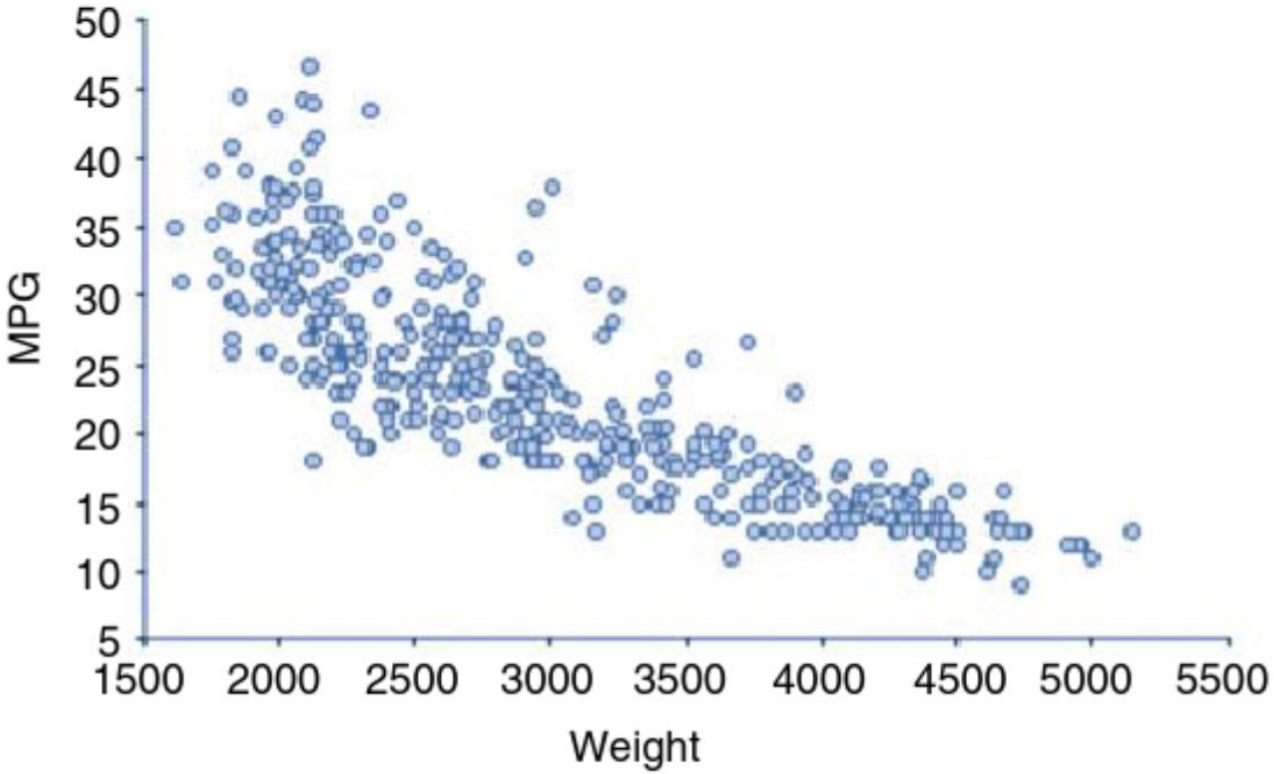
easy to see relationship between x and y

This example is positive relationship with positive slope



Typical visualizations associated with two variables

negative, non-linear relationship



Recall:

formula for a (linear) line:

$y = mx + b$

b is y intercept

m is slope $\Delta y / \Delta x$

if b and m are known, then any y can be calculated given any x

$m = r * s_y/s_x$

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

Example Data

TABLE 4.1 Table of Data with Values for the x and y Variables

x	y
92	6.3
145	7.8
30	3.0
70	5.5
75	6.5
105	5.5
110	6.5
108	8.0
45	4.0
50	5.0
160	7.5
155	9.0
180	8.6
190	10.0
63	4.2
85	4.9
130	6
132	7

What is r?

What is the slope?

What is the y-intercept?

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
92	6.3	−14.94	−0.11	1.64
145	7.8	38.06	1.39	52.90
30	3	−76.94	−3.41	262.37
70	5.5	−36.94	−0.91	33.62
75	6.5	−31.94	0.09	−2.87
105	5.5	−1.94	−0.91	1.77
110	6.5	3.06	0.09	0.28
108	8	1.06	1.59	1.69
45	4	−61.94	−2.41	149.28
50	5	−56.94	−1.41	80.04
160	7.5	53.06	1.09	58.07
155	9	48.06	2.59	124.68
180	8.6	73.06	2.19	160.00
190	10	83.06	3.59	298.19
63	4.2	−43.94	−2.21	97.11
85	4.9	−21.94	−1.51	33.13
130	6	23.06	−0.41	−9.45
132	7	25.06	0.59	14.79
$\bar{x} = 106.94$	$\bar{y} = 6.41$			*Sum = 1,357.06*
$s_x = 47.28$	$s_y = 1.86$			

Aside - is r significant?

If abs(r) > abs(r_{crit}), then r is significant

t it t_{crit} from t distribution with df = n-2

$$r_{crit} = \frac{t}{\sqrt{t^2+n-2}}$$

Matplotlib

import matplotlib.pyplot as plt

? plt.scatter

```
Signature:
 plt.scatter(
    x: 'float | ArrayLike',
    y: 'float | ArrayLike',
    s: 'float | ArrayLike | None' = None,
    c: 'ArrayLike | Sequence[ColorType] | ColorType | None' = None,
    marker: 'MarkerType | None' = None,
    cmap: 'str | Colormap | None' = None,
    norm: 'str | Normalize | None' = None,
    vmin: 'float | None' = None,
    vmax: 'float | None' = None,
    alpha: 'float | None' = None,
    linewidths: 'float | Sequence[float] | None' = None,
    *,
    edgecolors: "Literal['face', 'none'] | ColorType | Sequence[ColorType] | None" = None,
    plotnonfinite: 'bool' = False,
    data=None,
    **kwargs,
) -> 'PathCollection'
```

Scatter plots and logical indexing!

live coding