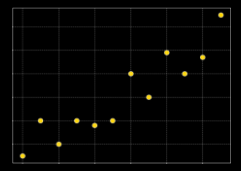
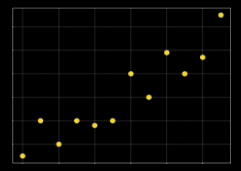


Scatter Plot

We will explore one of the most important visualization tools in statistics — the scatter plot. We will learn **what** it is, **why** it's useful, **how** to interpret it with real-life examples and some **common pitfalls** to avoid when using scatterplots



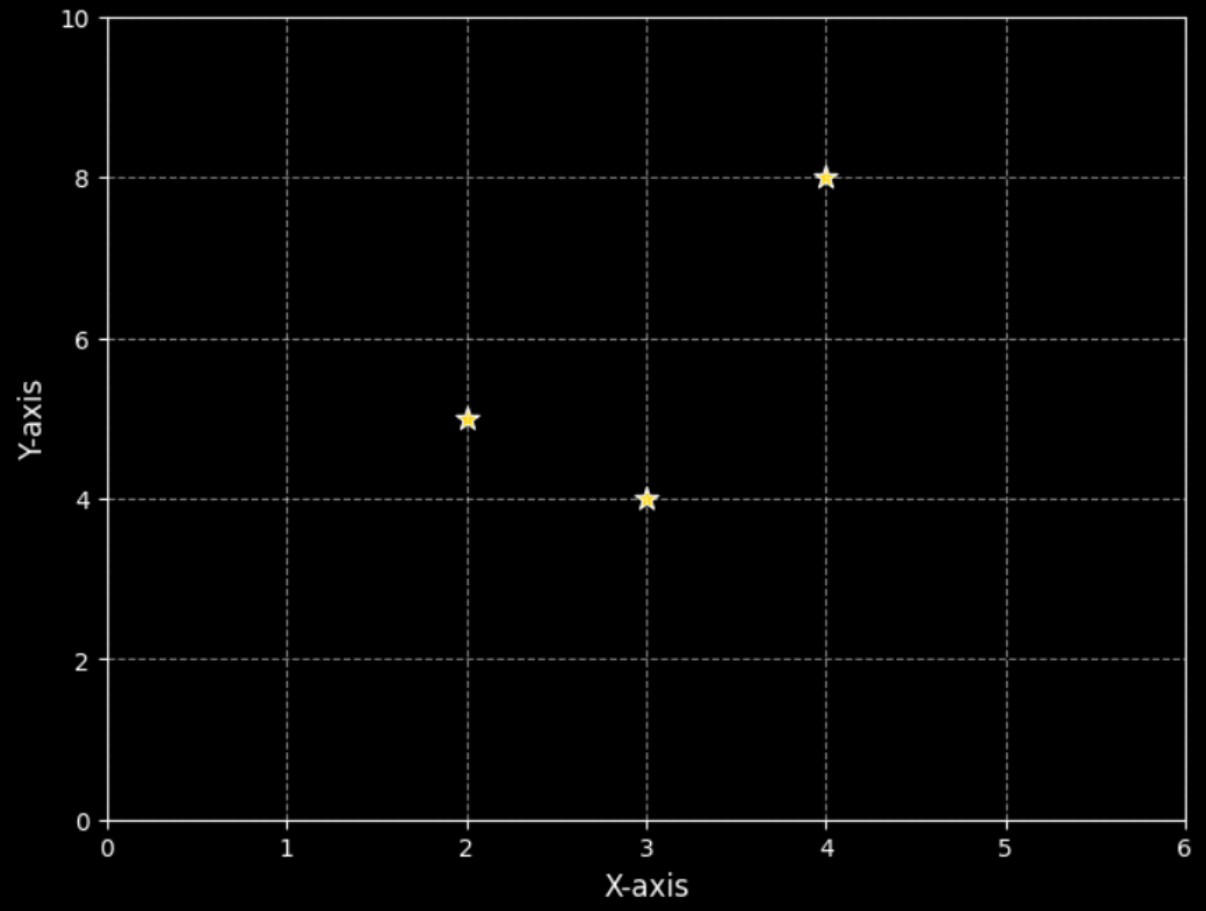
Scatter Plot

A scatter plot is a type of graph that shows the relationship between two quantitative or numerical variables using dots.

X-axis → **independent** variable
Y-axis → **dependent** variable (AKA response)
Each dot is a data point (x, y).

Example:
(2,5)
(4,8)
(3,4)

Easy to interpret and often the **first step** in data analysis.

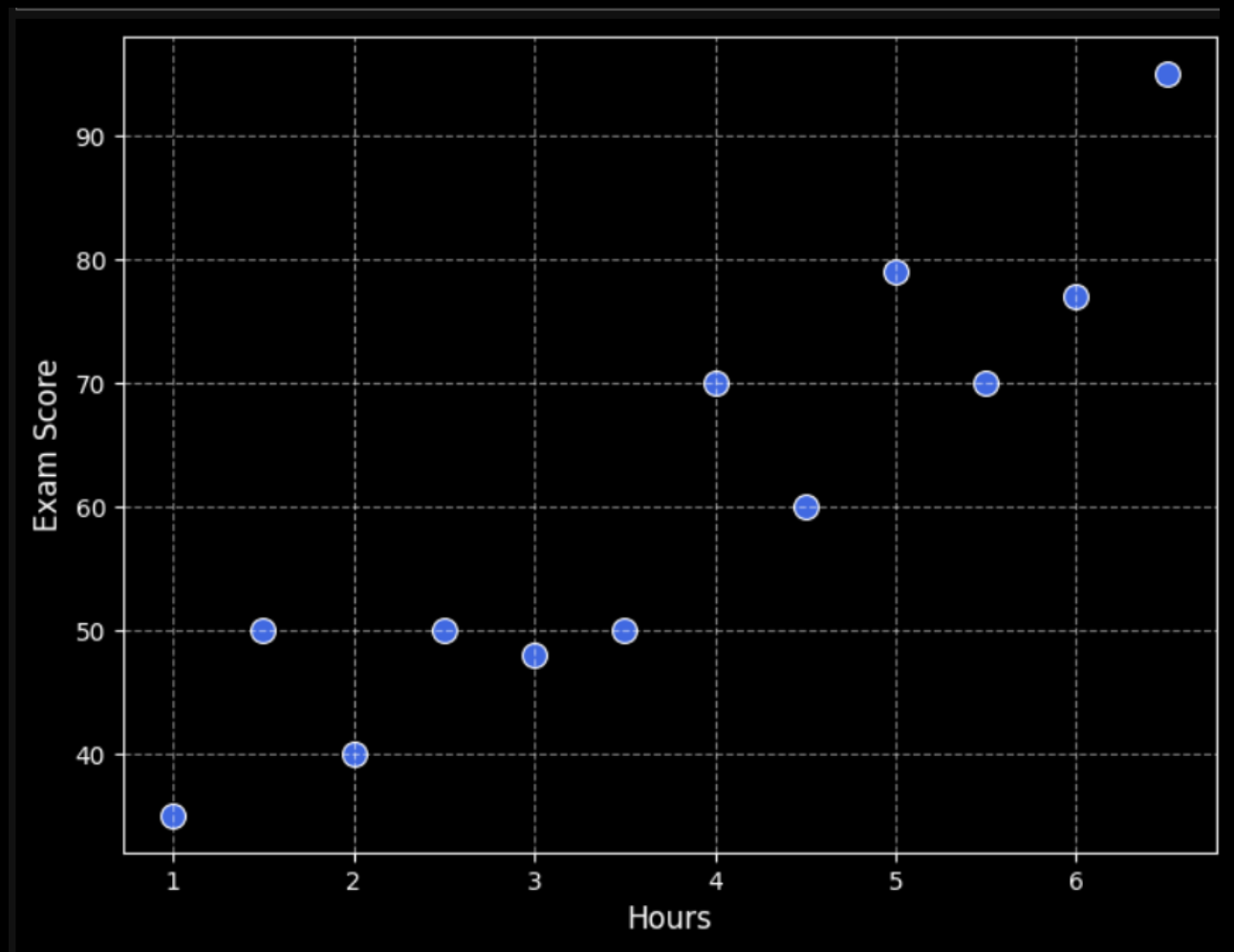


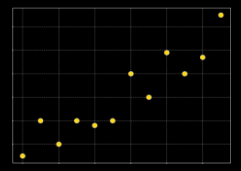
Scatter Plot

Example1: Positive linear relationship

Suppose we collect data on students — their hours of study (X) and exam scores (Y). Each point shows how much a student studied and what score they achieved.

Hours Studied	Scores
1	35
2	40
3	48
4	70
5	79
6	77
2.5	50
3.5	50
4.5	60
5.5	70
1.5	50
6.5	95



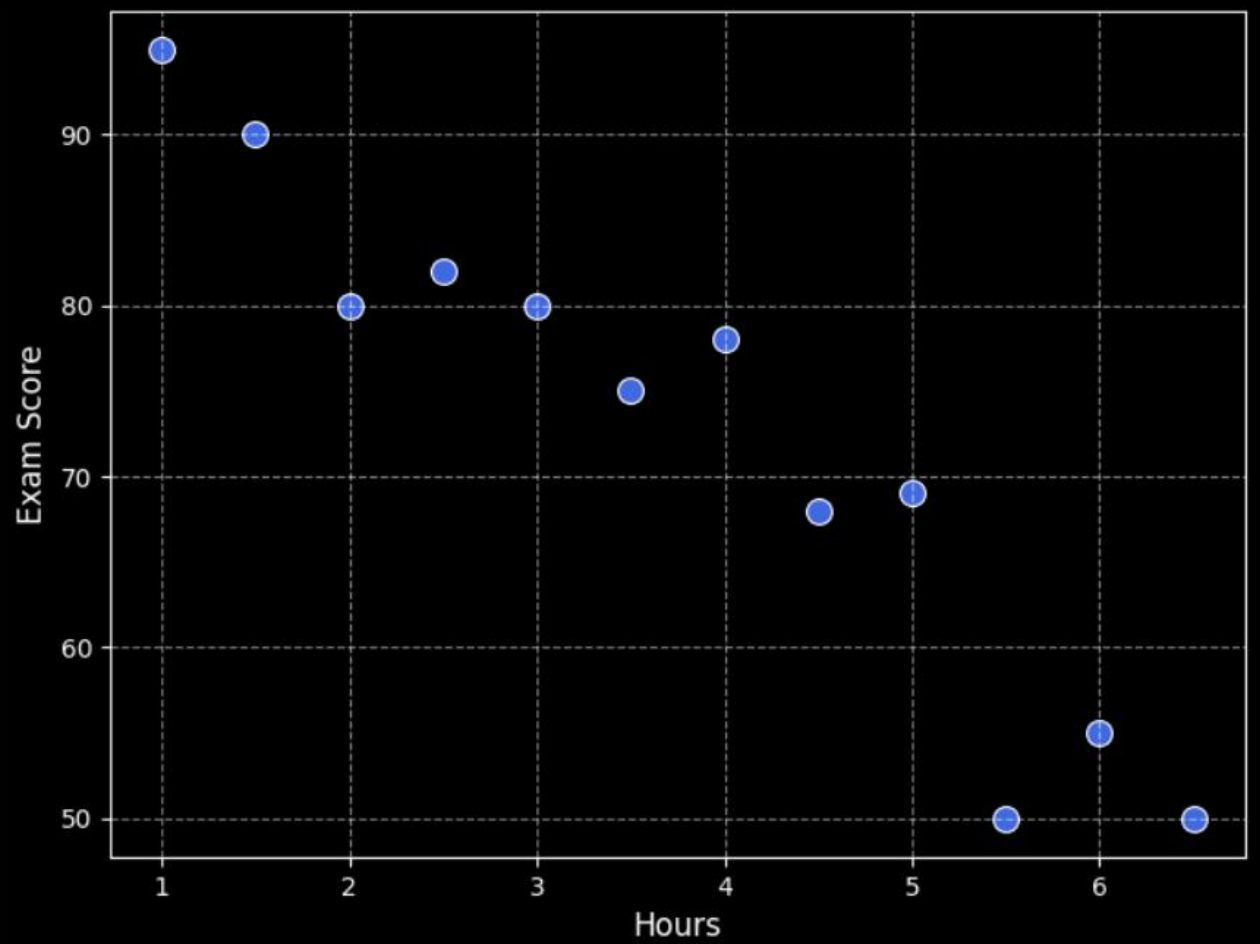


Scatter Plot

Example2: Negative linear relationship

Suppose we collect data on students — their hours of video game playtime (X) and exam scores (Y). Each point shows how much a student played video games and what score they achieved.

Hours Gaming	Scores
1	95
2	80
3	80
4	78
5	69
6	55
2.5	82
3.5	75
4.5	68
5.5	50
1.5	90
6.5	50



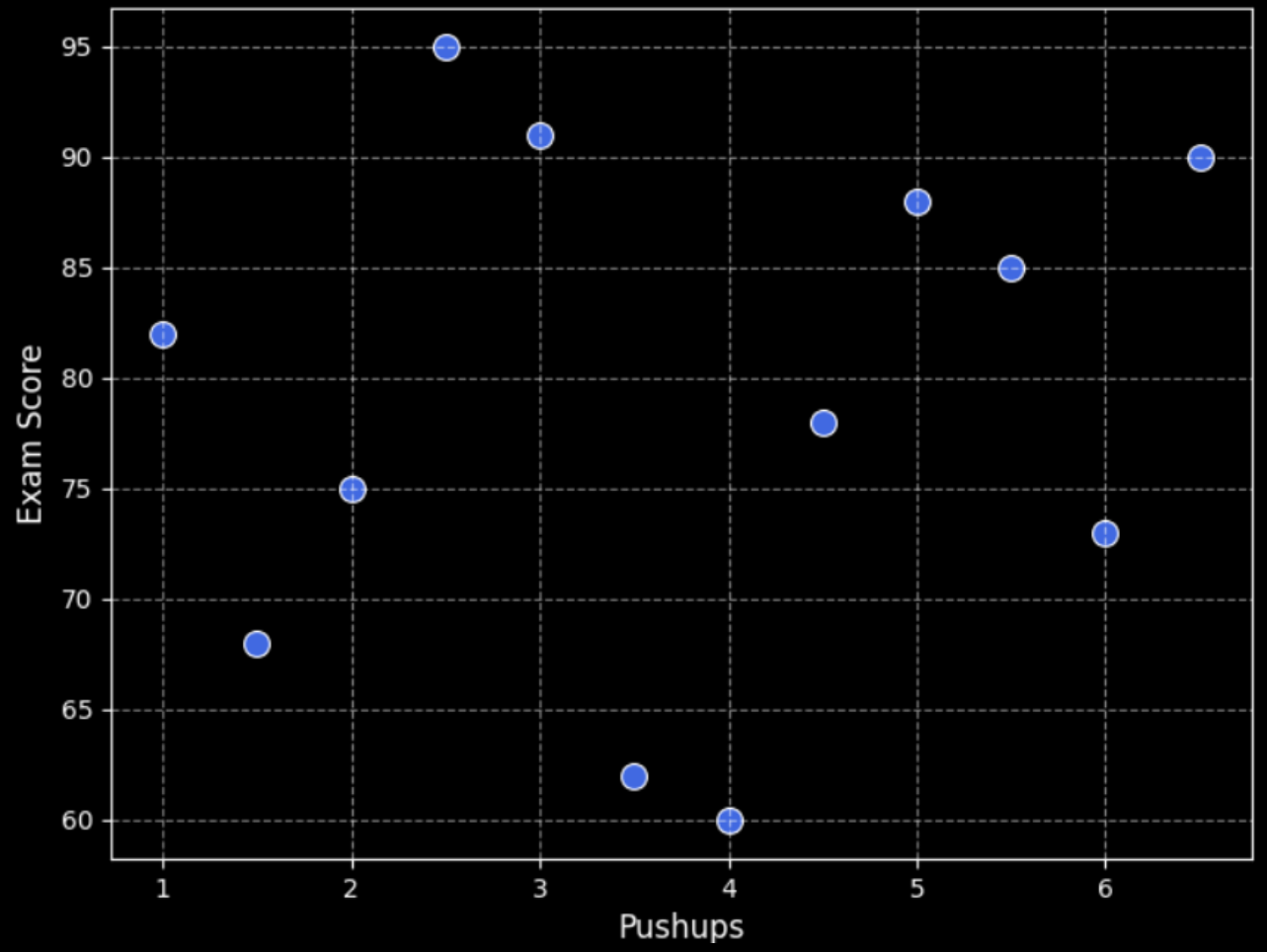
Scatter Plot

Example3: No linear relationship

Suppose we collect data on students — the number of pushups (X) and exam scores (Y).

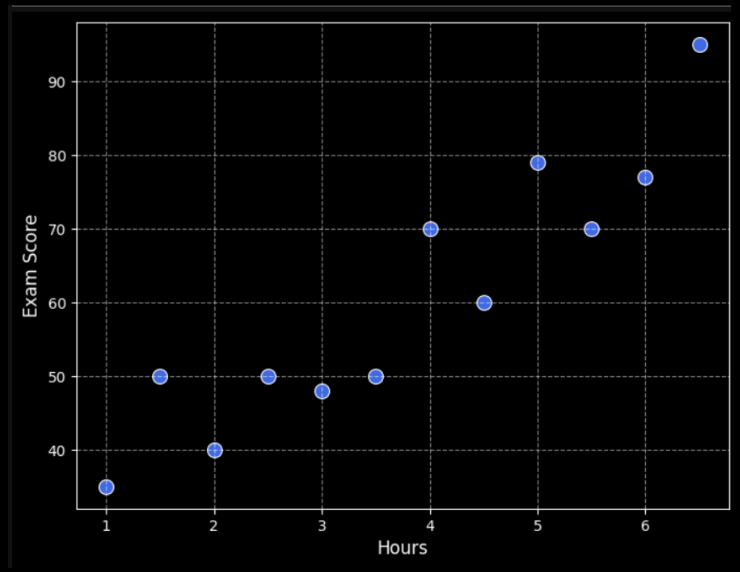
Each point shows how many pushups a student can do and what score they achieved.

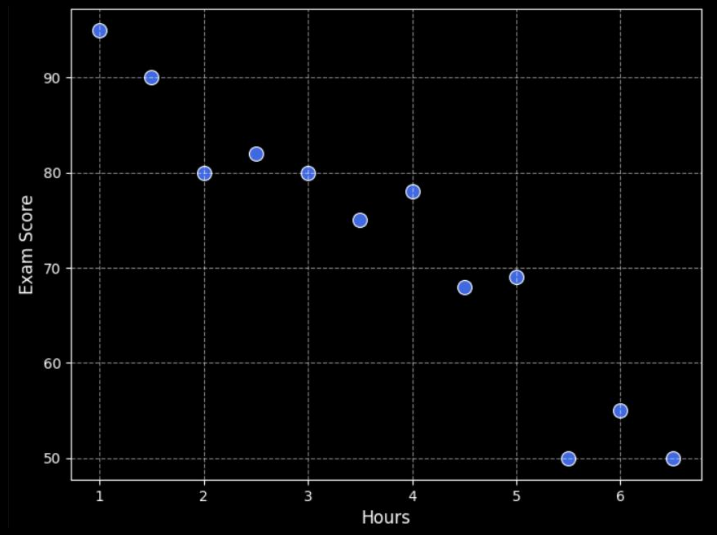
Push-Ups	Scores
10	82
15	75
20	91
25	60
30	88
35	73
40	95
45	62
50	78
55	85
60	68
65	90

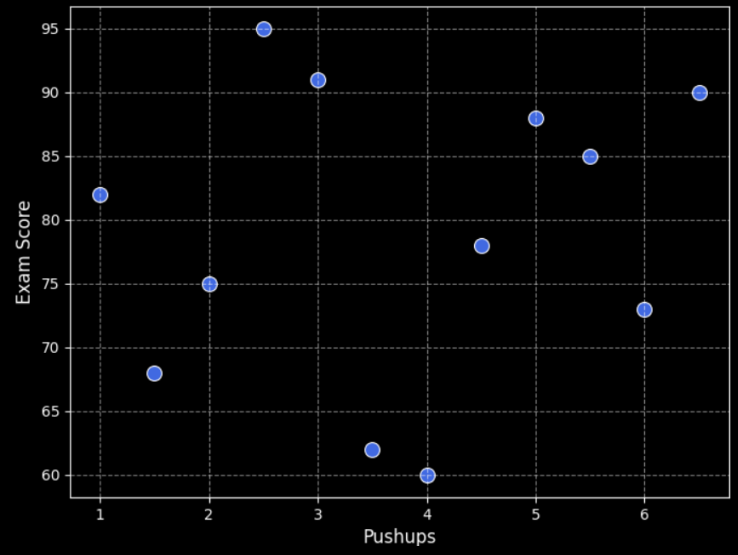


Why Scatter Plots Are Useful: Identify Relationships

- Detect *correlation* between variables.
- Direction of relationship:
  - Positive correlation: both increase
  - Negative correlation: one increases, the other decreases
  - No correlation: scattered, no clear trend





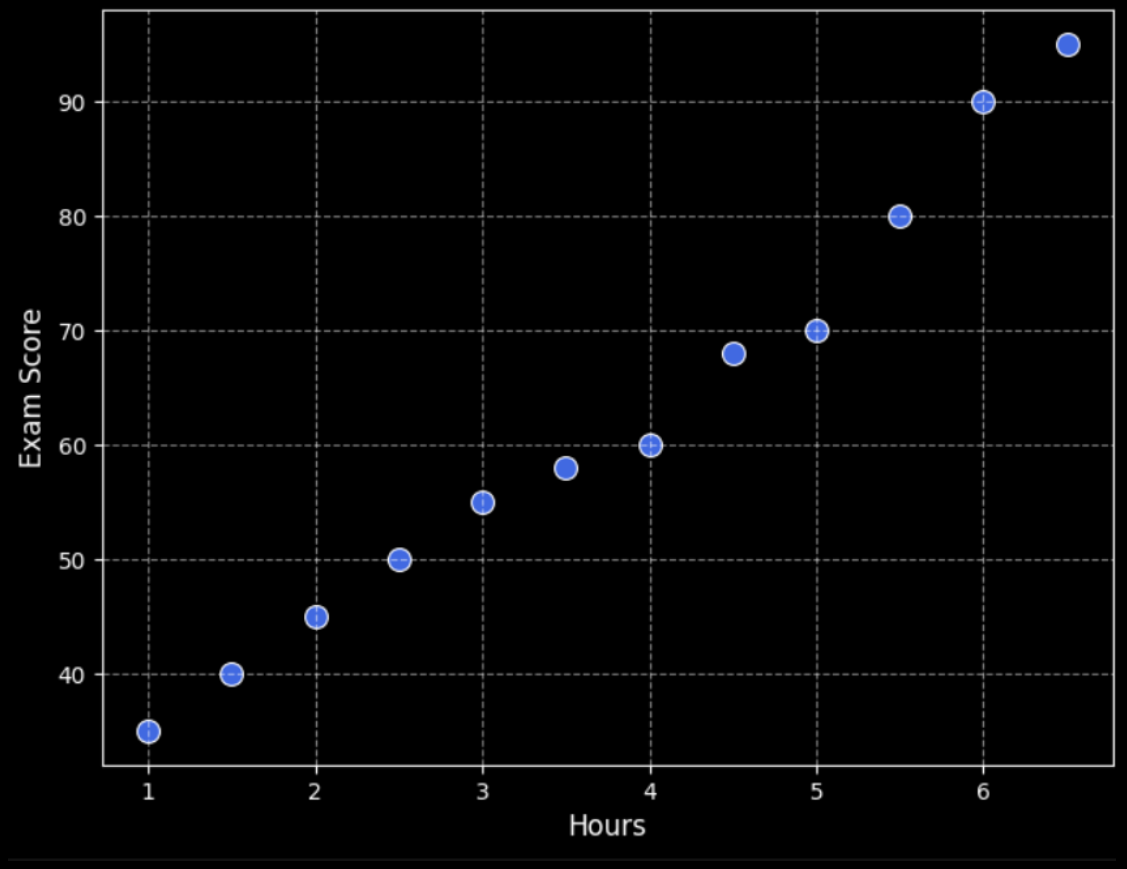


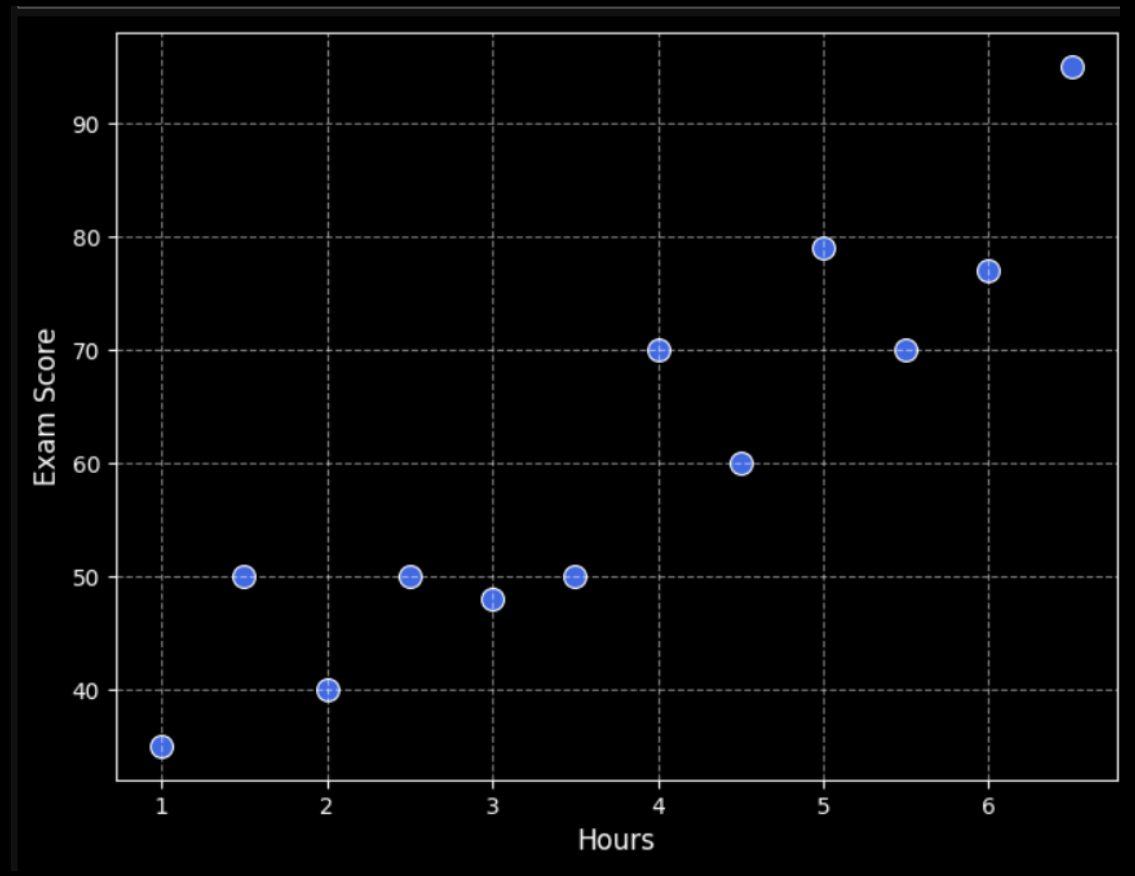
Why Scatter Plots Are Useful: Identify Strength of Relationships

- **Interpret strength:**

If strong correlation then points are close to a straight line.

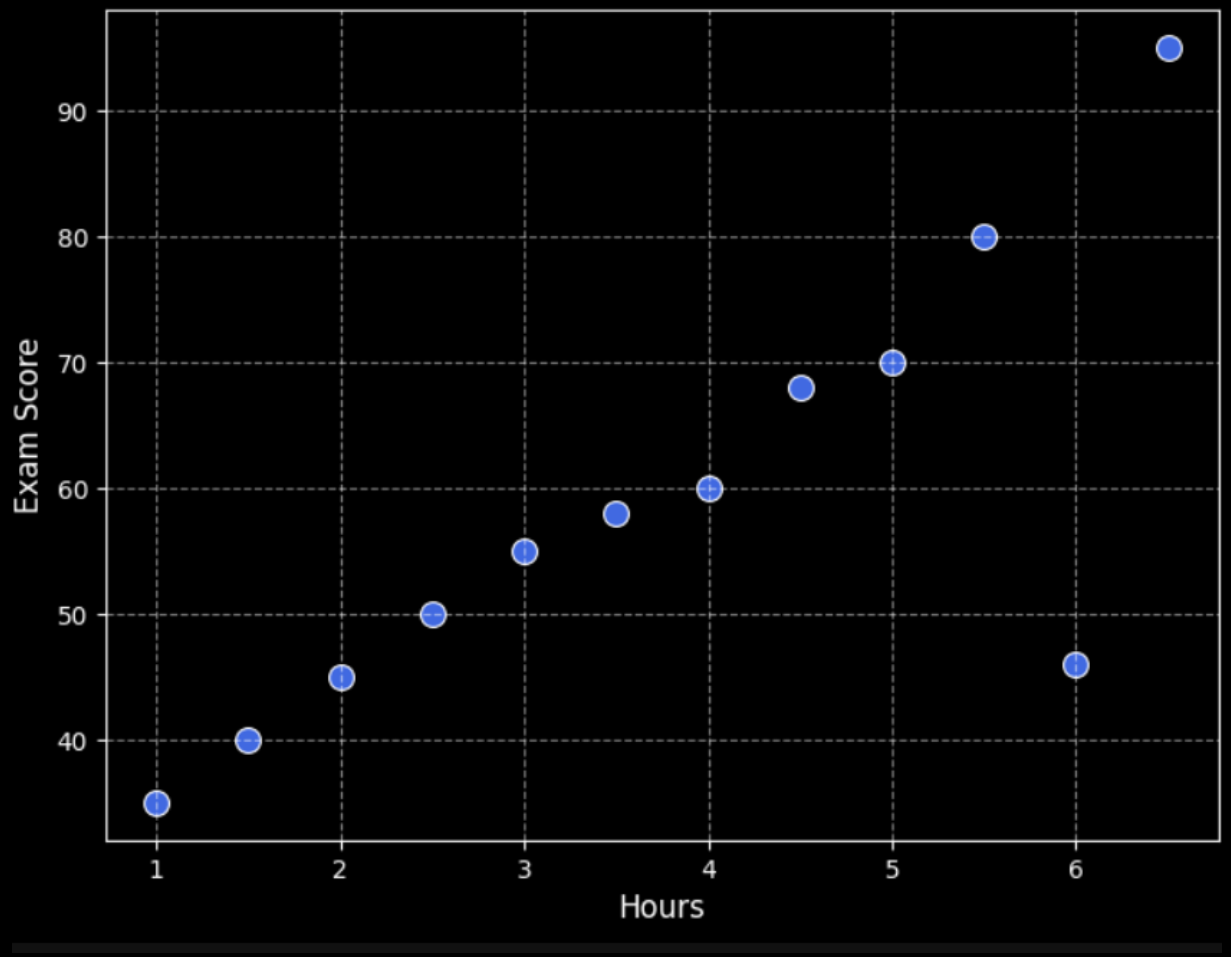
If weak correlation then points are loosely spread.





Why Scatter Plots Are Useful: Spot Outliers

- Points that don’t fit the pattern indicate unusual observations.



Why Scatter Plots Are Useful: Visualize Clusters

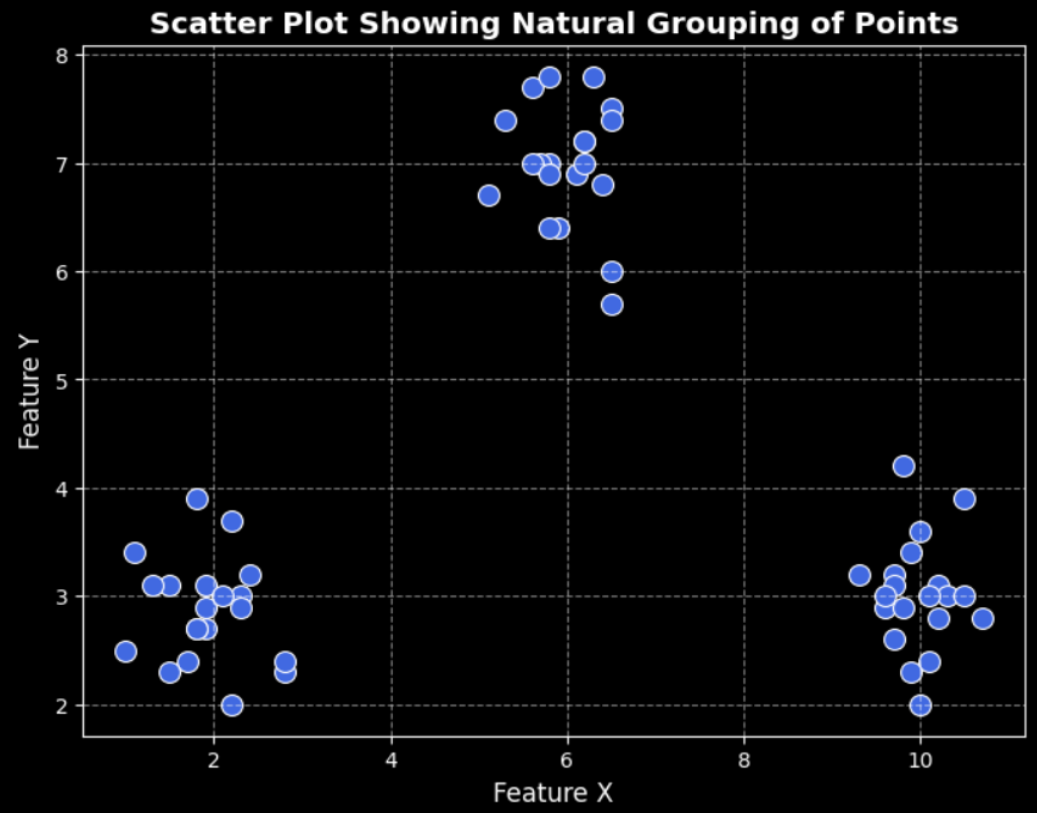
- You can detect natural grouping of points.

X (distance from home) = [2.2, 1.9, 2.3, 2.8, 1.9, 1.9, 2.8, 2.4, 1.8, 2.3, 1.8, 1.8, 2.1, 1.0, 1.1, 1.7, 1.5, 2.2, 1.5, 1.3, 6.4, 6.1, 5.9, 5.8, 5.3, 5.6, 5.8, 6.5, 6.2, 5.1, 6.2, 5.8, 5.7, 6.3, 6.5, 6.5, 5.6, 5.8, 6.2, 6.5, 9.9, 10.2, 10.7, 9.7, 9.6, 9.7, 10.5, 10.2, 9.7, 10.3, 10.0, 10.5, 9.6, 9.8, 9.8, 9.3, 10.1, 10.1, 10.0, 9.9]

Y (money spent on food) = [3.7, 2.9, 3.0, 2.3, 2.7, 3.1, 2.4, 3.2, 2.7, 2.9, 2.7, 3.9, 3.0, 2.5, 3.4, 2.4, 3.1, 2.0, 2.3, 3.1, 6.8, 6.9, 6.4, 6.4, 7.4, 7.7, 7.0, 7.5, 7.2, 6.7, 7.2, 7.8, 7.0, 7.8, 5.7, 7.4, 7.0, 6.9, 7.0, 6.0, 2.3, 2.8, 2.8, 2.6, 2.9, 3.2, 3.9, 3.1, 3.1, 3.0, 2.0, 3.0, 3.0, 4.2, 2.9, 3.2, 3.0, 2.4, 3.6, 3.4]

It is hard to detect clusters by looking at numbers,

but easy to see them when they are plotted.



Common Mistakes and Tips

1. Using it for categorical data:

Scatter plots only make sense for columns that have numerical data. For example, below, the feature Department although has numeric data, but it is categorical in nature with following mapping:

1 -> HR

2 -> Sales

3 -> Finance

Name	Department	Years_of_Experience	Salary
A	1	2	32000
B	2	5	55000
C	2	7	72000
D	3	3	40000
E	1	6	50000
F	3	8	65000
G	2	4	43000
H	2	9	80000
I	1	1	30000
J	3	10	85000

Common Mistakes and Tips

2. Misinterpreting correlation as causation:

“Just because two things move together doesn’t mean one causes the other.”

A study finds that **ice cream sales** and **electricity demand** both increase during certain months of the year. We may misinterpret this and claim increased ice cream consumption led to higher electricity demand.

Actual reason could be a lurking variable. The real cause is **seasonal temperature** — warm weather causes both:

- More ice cream sales (because people want cold desserts)
- Electricity demand grows (due to more AC, etc.)

Real-World Applications

1. **Education:** Hours studied vs. marks scored.
2. **Health:** Calorie intake vs. weight gain.
3. **Business:** Advertising spend vs. sales revenue.
4. **Economics:** Inflation vs. unemployment (Phillips Curve).
5. **Data Science:** Feature relationships before modeling — to check if variables are related.

Etc…

Hfdslfds

fksdljfsD|]

fkdslfa