

# Correlation and Regression

Exploratory Data Analysis (EDA) helps us **understand** our dataset:

- What variables exist
- Their distributions
- Trends and patterns
- Outliers
- Missing values

However, **EDA alone cannot answer deeper analytical questions**, such as:

- *Does one variable predict another?* For example, Do higher ad spends lead to higher sales? Can we quantify how much sales change if ad spend change by 100K?
- *How strong is the relationship between two variables?*

To answer these questions, we need **Correlation** and **Regression**.

---

## Why Correlation: Correlation tells us **how strongly two numerical variables move together**.

Examples: - Does sales increase when advertising increases? - Are discounts associated with lower profit? - Do online visits correlate with conversions?

Correlation coefficient is a **number between -1 and 1**:

- **1** → Perfect positive relationship
- **0** → No relationship
- **-1** → Perfect negative relationship

Please note that, **Correlation does not imply causation**, but it tells us whether a predictive relationship may exist.

---

# Why Regression: While correlation tells us “**there is a relationship**”, regression tells us:

- Can we predict one variable from another?
- Is the relationship statistically significant?
- How well does our prediction perform?

In business analytics, regression helps answer practical questions like: - *“If we increase advertising by \$1000, how much will sales increase?”* - *“How does price affect demand?”* - *“Can we forecast future sales?”*

---

## Create a small dataset for analysis

- Amount spent on advertising ( `Ad_Spend` )
- Units sold ( `Units_Sold` )

```
library(tidyverse)

# Create a simple toy dataset
sales_data <- tibble(
  Ad_Spend    = c(1000, 1500, 2000, 2500, 3000, 3500, 4000),
  Units_Sold  = c(50,   60,   65,   70,   80,   85,   90)
)

sales_data
```

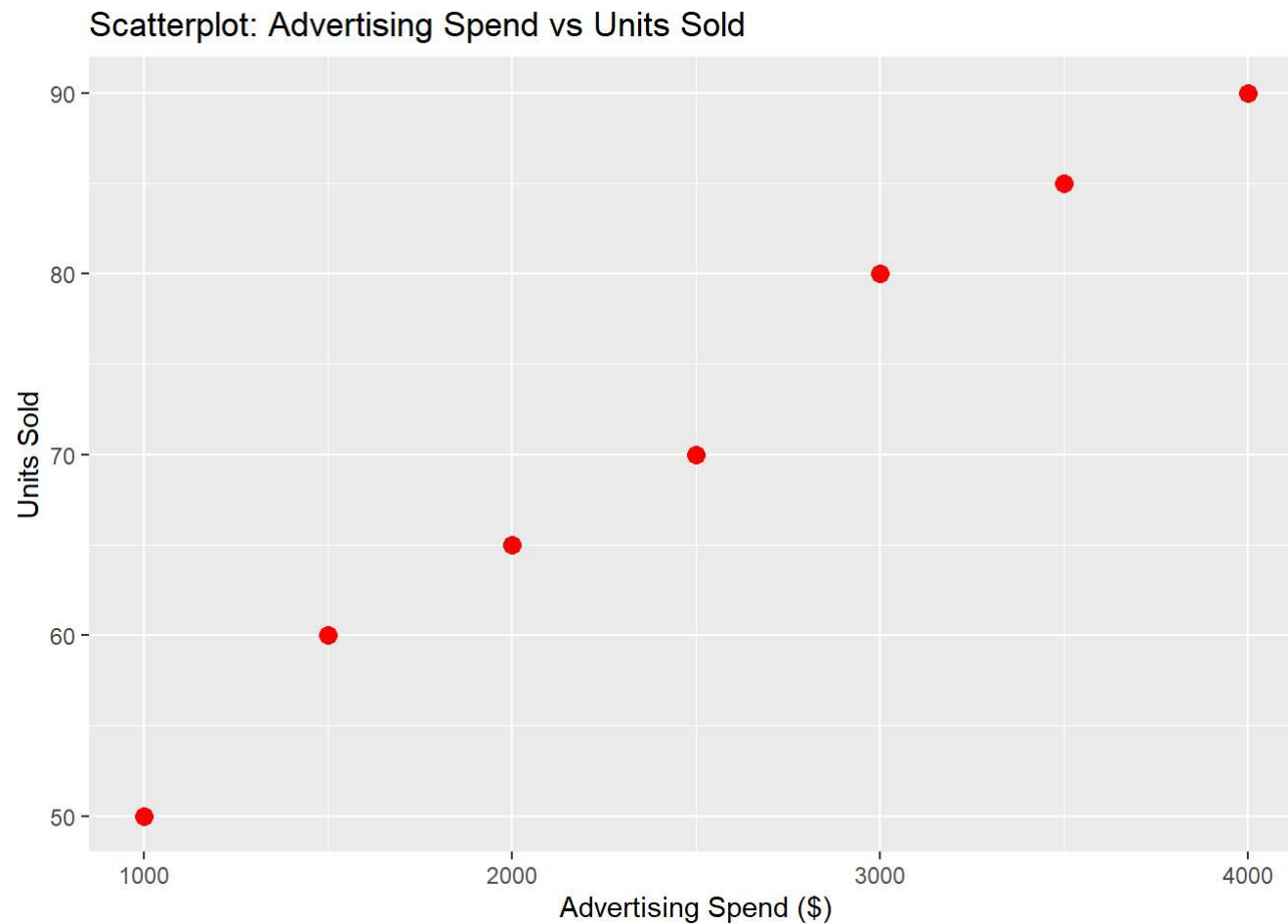
```
## # A tibble: 7 × 2
##   Ad_Spend Units_Sold
##   <dbl>     <dbl>
## 1    1000         50
## 2    1500         60
## 3    2000         65
## 4    2500         70
## 5    3000         80
## 6    3500         85
## 7    4000         90
```

## Quick EDA

```
# Summary of variables
summary(sales_data)
```

```
##      Ad_Spend      Units_Sold
## Min.   :1000   Min.   :50.00
## 1st Qu.:1750   1st Qu.:62.50
## Median :2500   Median :70.00
## Mean   :2500   Mean    :71.43
## 3rd Qu.:3250   3rd Qu.:82.50
## Max.   :4000   Max.    :90.00
```

```
# Scatterplot: good visual check for linear trend
sales_data |>
  ggplot(aes(x = Ad_Spend, y = Units_Sold)) +
  geom_point(size = 3, color = "red") +
  labs(title = "Scatterplot: Advertising Spend vs Units Sold",
       x = "Advertising Spend ($)",
       y = "Units Sold")
```



**Observation:** You should see a **clear upward trend** — more advertising → more sales.

## Correlation Analysis

```
correlation_value <- cor(sales_data$Ad_Spend, sales_data$Units_Sold)
correlation_value
```

```
## [1] 0.9945662
```

## How to interpret

- If your output is around **0.90–0.99**, this means **very strong positive correlation**.
- Higher ad spend is strongly associated with higher sales.
- This strong relationship suggests trying regression can be meaningful.
- In general a correlation coefficient closer to (0.75–0.99) or (–0.75–0.99) indicates strong relationship.
- The sign of the coefficient indicates whether the variable are directly or inversly proportional to each other.
- Closer to 0 (either +ve or -ve) numbers indicate very very weak or no relationship at all.

---

## Simple Linear Regression

We try to model the relationship between Ad\_Spend and Sales as a straightline (its like fitting a straightline through the data points). It is absolutely not necessary to model the relationship as linear as it may not be true at all. But, it is the simplest model possible.

$\text{Units\_Sold} = b_0 + b_1 \text{ Ad\_Spend}$

Now let's code in R to build our predictive model.

```
model <- lm(Units_Sold ~ Ad_Spend, data = sales_data)
summary(model)
```

```
##
## Call:
## lm(formula = Units_Sold ~ Ad_Spend, data = sales_data)
##
## Residuals:
##      1      2      3      4      5      6      7
## -1.6071  1.7857  0.1786 -1.4286  1.9643  0.3571 -1.2500
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.839e+01  1.666e+00   23.05 2.86e-06 ***
## Ad_Spend     1.321e-02  6.186e-04   21.36 4.17e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.637 on 5 degrees of freedom
## Multiple R-squared:  0.9892, Adjusted R-squared:  0.987
## F-statistic: 456.3 on 1 and 5 DF,  p-value: 4.168e-06
```

# Understanding the Regression Output

The `summary(model)` output includes several key elements.

## 1. Coefficients

Example interpretation:

- **Intercept ( $\beta_0$ )** =What is the predicted mean value of dependent variable when our Independent variable is zero. For our example, it is around 35 That means, If advertising is zero, the model predicts about 35 units sold on average.
- **Slope ( $\beta_1$ )** =What is the change in dependent variable in average, if independent variable changes by one unit. For our example, it is around 0.015. That means, For every additional **\$1 spent**, units sold increase by **0.015**. For every additional **\$1000 spent**, sales increase by:

[ 0.015 ]

## 2. p-values

If p-value < 0.05:

- The relationship is statistically significant.
- Advertising spend **does have a real effect** on sales.

---

## 3. R-squared

- Shows how much percentage of the variation in sales is explained by ad spend.
- A value like **0.95** means:
  - 95% of sales changes can be explained by advertising.
  - Very strong model.

---

## Step 5: Regression Line Visualization

```
sales_data |>
  ggplot(aes(x = Ad_Spend, y = Units_Sold)) +
  geom_point(size = 3, color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "darkred") +
  labs(title = "Regression Line: Advertising Spend Predicting Sales",
       x = "Advertising Spend ($)",
       y = "Units Sold")
```

