ANOVA – Analysis of Variance

ANOVA, or Analysis of Variance, in data science refers to a statistical method used to analyze the differences among group means in a sample.

Types of ANOVA:

- **One-way ANOVA**: Compares the means of three or more independent groups to determine if there are statistically significant differences among them.
- **Two-way ANOVA**: Extends the one-way ANOVA by assessing the influence of two categorical independent variables (factors) simultaneously.
- **ANOVA with repeated measures**: Analyzes data where measurements are taken from the same subjects at different points in time or under different conditions.

Application in Data Science:

ANOVA is widely used in data science for various purposes, including:

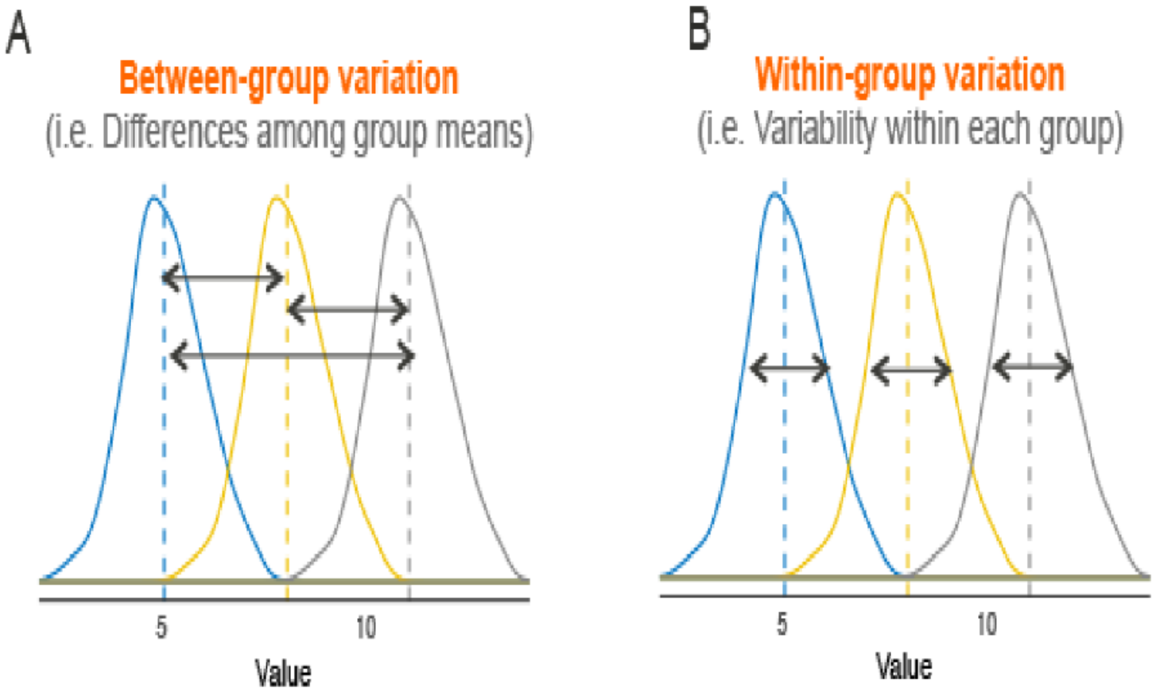
- **Comparing multiple groups**: For example, comparing the effectiveness of different drugs in a clinical trial.
- **Variable selection**: Determining which variables (factors) have a statistically significant impact on the outcome variable.
- **Interaction effects**: Assessing whether the effect of one variable depends on the level of another variable.

Assumptions of one-way ANOVA

The assumptions of ANOVA include:

- **Independence:** The observations within each group must be independent of each other. This means that the value of one observation should not be related to the values of another observation in the same group.
- **Normality:** The dependent variable should be normally distributed within each group. This means that the distribution of the values should be symmetrical and bell-shaped.
- **Homogeneity of Variance:** The variance of the dependent variable should be equal across all levels of the independent variable. This means that the spread of the values should be the same for each group.
- **Random sampling:** The observations in each group should be randomly selected from the population.

ONE WAY ANOVA:



Two-way ANOVA

A **two-way ANOVA** ("analysis of variance") is used to determine whether or not there is a statistically significant difference between the means of three or more independent groups that have been split on two variables (sometimes called "factors").

When to Use a Two-Way ANOVA

You should use a two-way ANOVA when you'd like to know how two factors affect a response variable and whether or not there is an interaction effect between the two factors on the response variable.

For example, suppose a botanist wants to explore how sunlight exposure and watering frequency affect plant growth. She plants 40 seeds and lets them grow for two months under different conditions for sunlight exposure and watering frequency. After two months, she records the height of each plant.

In this case, we have the following variables:

- **Response variable:** plant growth
- **Factors:** sunlight exposure, watering frequency

And we would like to answer the following questions:

- Does sunlight exposure affect plant growth?
- Does watering frequency affect plant growth?
- Is there an interaction effect between sunlight exposure and watering frequency? (e.g. the effect that sunlight

exposure has on the plants is dependent on watering frequency)

We would use a two-way ANOVA for this analysis because we have **two** factors. If instead we wanted to know how only watering frequency affected plant growth, we would use a one-way ANOVA since we would only be working with one factor.

Two-Way ANOVA Assumptions

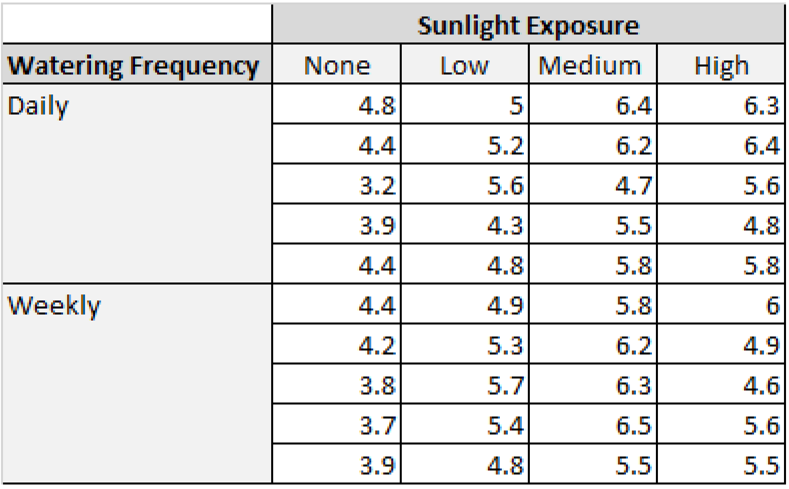
For the results of a two-way ANOVA to be valid, the following assumptions should be met:

1. Normality – The response variable is approximately normally distributed for each group.

2. Equal Variances – The variances for each group should be roughly equal.

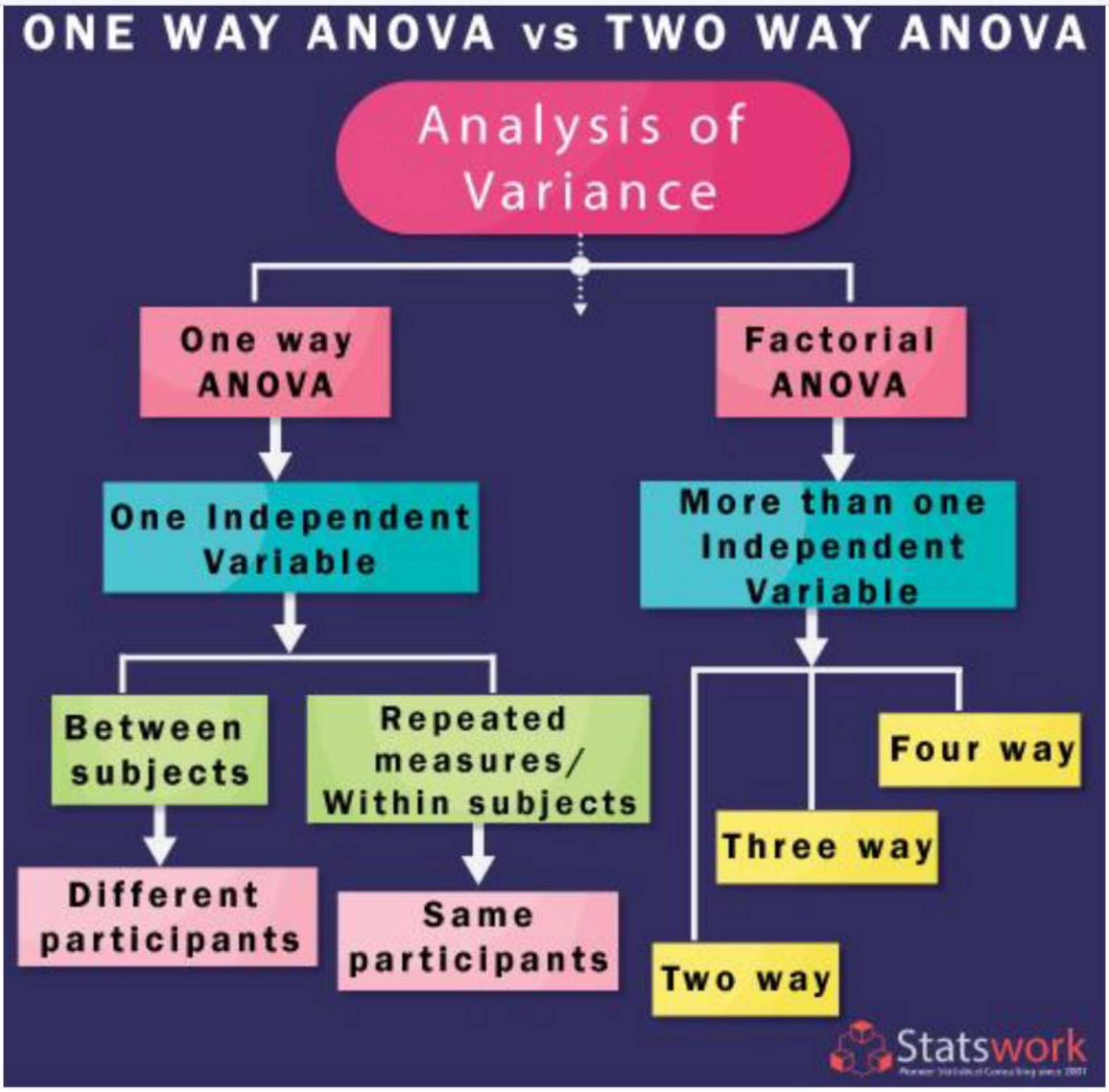
3. Independence – The observations in each group are independent of each other and the observations within groups were obtained by a random sample.

A botanist wants to know whether or not plant growth is influenced by sunlight exposure and watering frequency. She plants 40 seeds and lets them grow for two months under different conditions for sunlight exposure and watering frequency. After two months, she records the height of each plant. The results are shown below:



	Sunlight Exposure			
Watering Frequency	None	Low	Medium	High
Daily	4.8	5	6.4	6.3
	4.4	5.2	6.2	6.4
	3.2	5.6	4.7	5.6
	3.9	4.3	5.5	4.8
	4.4	4.8	5.8	5.8
Weekly	4.4	4.9	5.8	6
	4.2	5.3	6.2	4.9
	3.8	5.7	6.3	4.6
	3.7	5.4	6.5	5.6
	3.9	4.8	5.5	5.5

In the table above, we see that there were five plants grown under each combination of conditions.



Conclusion:

In summary, ANOVA is a powerful statistical tool in data science that helps analyze the differences among group means and determine whether these differences are statistically significant, providing valuable insights into the relationships between variables in a dataset.