**INFERENTIAL STATISTICAL TESTS**

**Hypothesis Testing (Comparison Tests)**

* **t-Test (Independent & Paired)** – Comparing means of two groups (A/B testing, before-after analysis).
* **ANOVA (One-Way & Two-Way)** – Comparing means of multiple groups (customer segmentation, experimental design).
* **Z-Test** – Comparing sample mean to population mean (large sample size).

**1. t-Test (Independent & Paired) – Detailed Explanation & Real-World Example**

**What is a t-Test?**

A **t-test** is a statistical test used to compare the means of two groups to determine if the difference between them is statistically significant. It helps answer questions like:

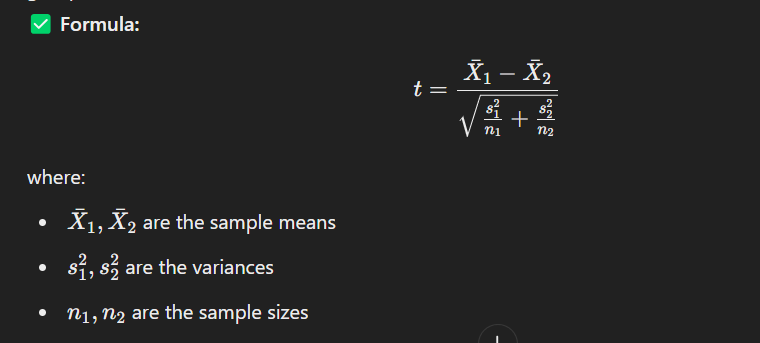
* Is there a significant difference in customer spending before and after a marketing campaign?
* Do two different website designs lead to different conversion rates?

The t-test assumes that the data follows a **normal distribution** and that variances are equal in the two groups (for independent samples). If these assumptions are violated, non-parametric tests (like the Mann-Whitney U test) are better alternatives.

**Types of t-Tests**

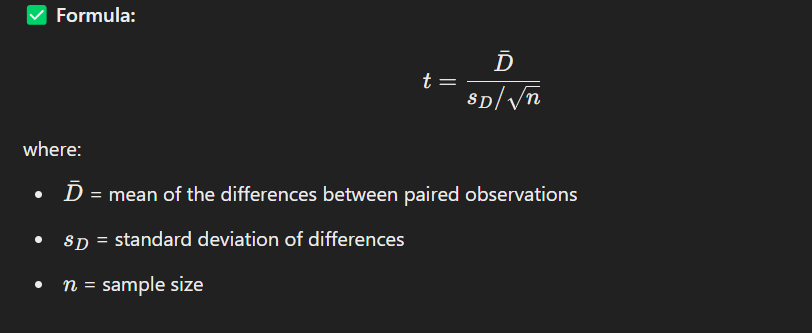
**1. Independent Samples t-Test (Two-Sample t-Test)**

✅ **Use case:** Compare the means of two independent groups.  
✅ **Assumption:** The two samples are unrelated (e.g., different customer segments, different product groups).



**2. Paired Samples t-Test (Dependent t-Test)**

✅ **Use case:** Compare means before and after an intervention (e.g., impact of a training program).  
✅ **Assumption:** The two samples are related (e.g., same individuals before and after an event).



* Use an **Independent t-Test** for comparing two different groups (e.g., A/B testing).
* Use a **Paired t-Test** for comparing before-and-after situations (e.g., impact analysis).
* In real-world business, these tests help in decision-making, A/B testing, and assessing marketing impact.

**(A). Independent t-Test (A/B Testing for Website Conversion)**

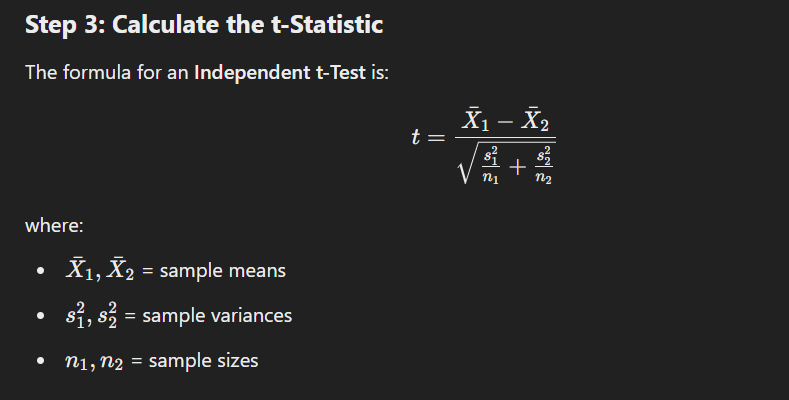
**Scenario:** A company wants to test if a new website design (Version B) improves conversion rates compared to the old design (Version A).

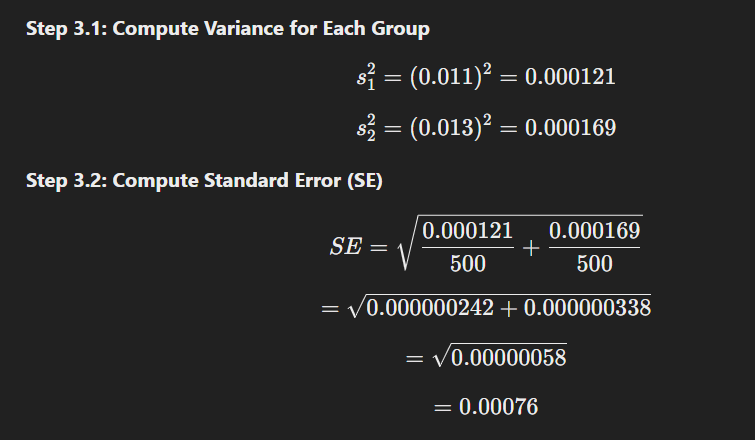
**Step 1: Define Hypotheses**

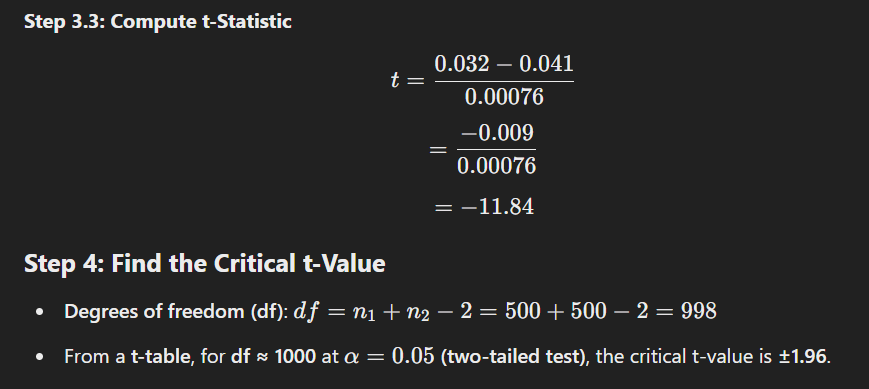
* Null Hypothesis (H0​): There is no significant difference in conversion rates between Version A and Version B.
* Alternative Hypothesis (H1​): Version B has a significantly higher conversion rate.

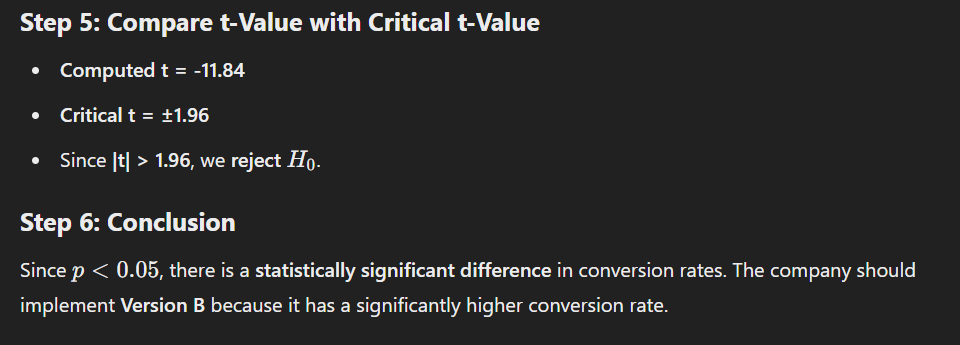
**Step 2: Given Data**

| **Version** | **Sample Size (n)** | **Mean Conversion Rate (x bar)** | **Standard Deviation (s)** |
| --- | --- | --- | --- |
| A (Old) | 500 | 3.2% (0.032) | 1.1% (0.011) |
| B (New) | 500 | 4.1% (0.041) | 1.3% (0.013) |









**(B). Paired t-Test (Sales Before vs. After Marketing Campaign)**

**Scenario:**

A company wants to see if a **marketing campaign** significantly increases customer sales.

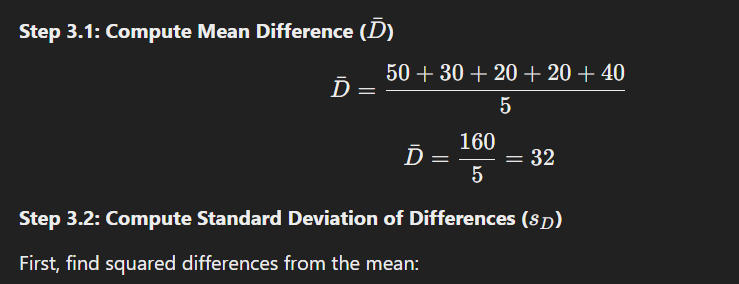
**Step 1: Define Hypotheses**

* **Null Hypothesis (H0)**: No significant difference in customer sales before and after the campaign.
* **Alternative Hypothesis (H1​)**: The campaign increased customer sales.

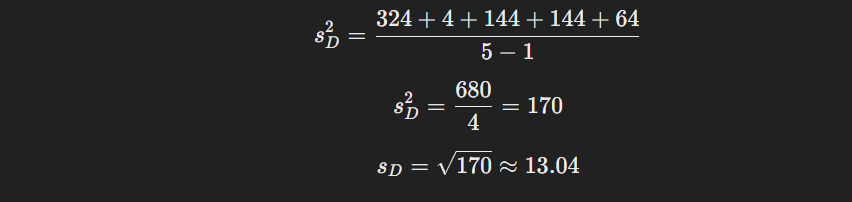
**Step 2: Given Data**

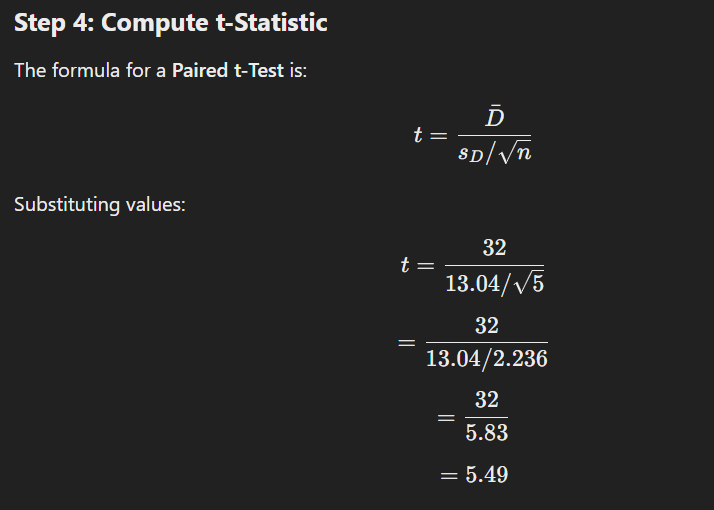
| **Customer** | **Sales Before ($X\_1$)** | **Sales After ($X\_2$)** | **Difference ($D = X\_2 - X\_1$)** |
| --- | --- | --- | --- |
| 1 | 200 | 250 | 50 |
| 2 | 150 | 180 | 30 |
| 3 | 300 | 320 | 20 |
| 4 | 180 | 200 | 20 |
| 5 | 250 | 290 | 40 |

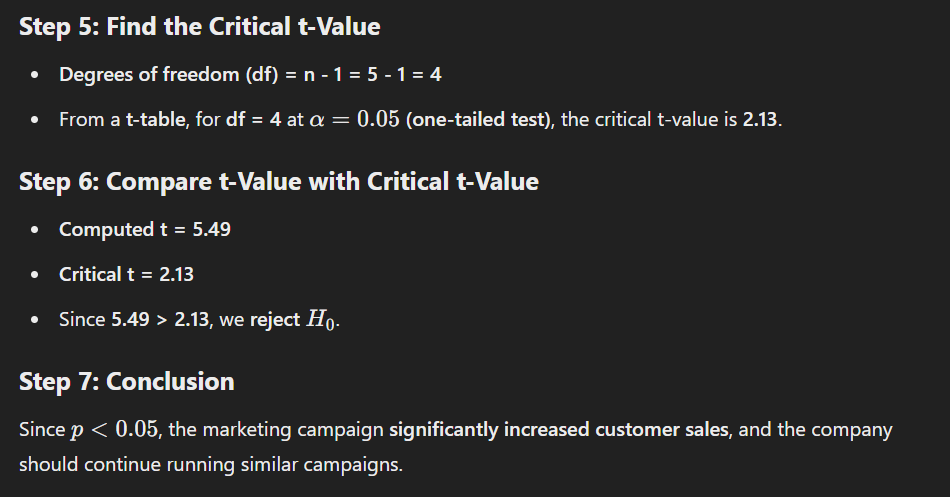
**Step 3: Compute Mean and Standard Deviation of Differences**

****







****

**2. Z-Test: Detailed Explanation & Real-World Industrial Example**

**What is a Z-Test?**

A Z-Test is a statistical test used to compare a sample mean to a population mean when the sample size is large (n≥30). It determines whether the observed difference is statistically significant or due to random chance.

**When to Use a Z-Test?**

✅ Population standard deviation (σ) is known  
✅ Sample size is large (n≥30)  
✅ Data is normally distributed (or follows the Central Limit Theorem)

If the sample size is small (n<30) and population standard deviation is unknown, use a t-Test instead.

**Types of Z-Tests**

1. **One-Sample Z-Test:** Compares the sample mean to a known population mean.
2. **Two-Sample Z-Test:** Compares the means of two independent samples.

**(A). One-Sample Z-Test: Manual Calculation & Example**

**Industry Example:** Average Salary of Data Analysts

A company claims that the average salary of a data analyst in New York is $85,000 per year. A researcher collects a random sample of 50 data analysts and finds the average salary is $82,000, with a known population standard deviation of $12,000.

We want to test if the average salary in the sample is significantly different from the claimed $85,000.

Step-by-Step Calculation of One-Sample Z-Test

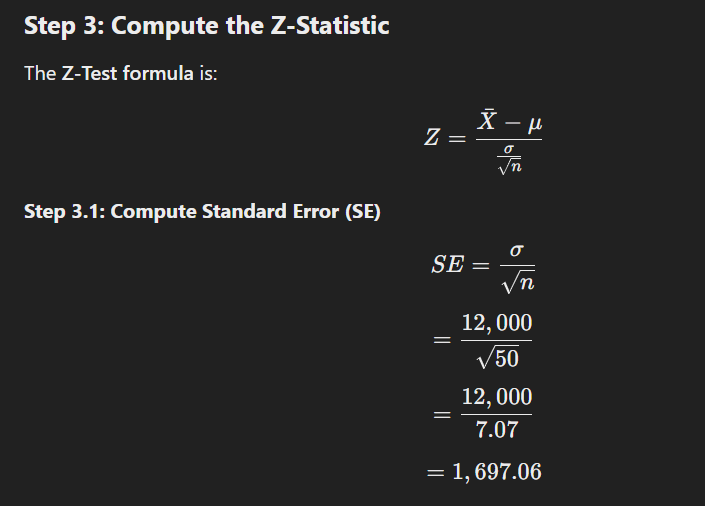
**Step 1: Define Hypotheses**

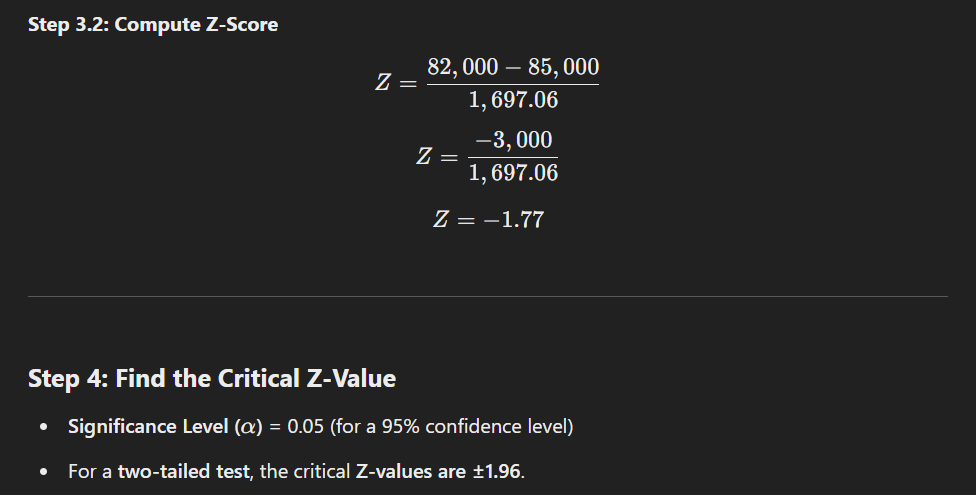
* Null Hypothesis (H0): The population mean is $85,000 (μ=85,000)
* Alternative Hypothesis (H1​): The population mean is not $85,000 (μ≠85,000)

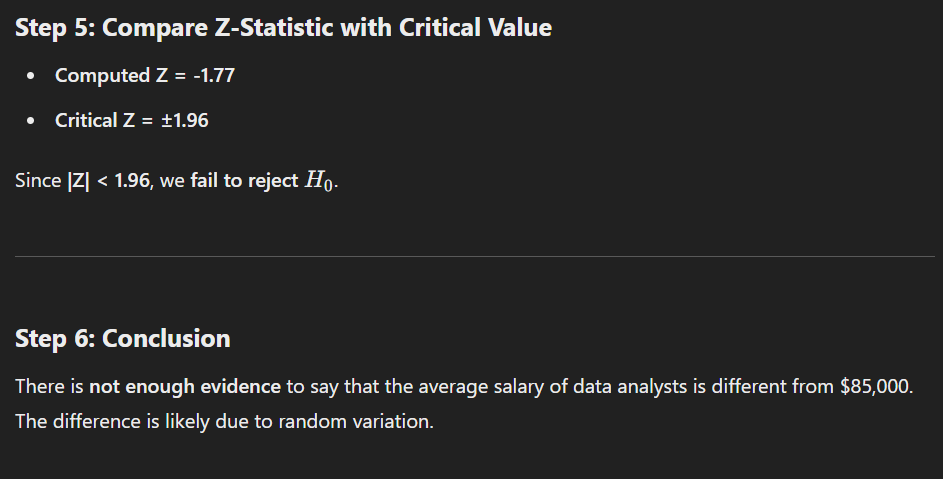
This is a two-tailed test since we are checking for any difference (higher or lower).

**Step 2: Given Data**

* Population Mean (μ) = 85,000
* Sample Mean (X bar) = 82,000
* Population Standard Deviation (σ) = 12,000
* Sample Size (n) = 50



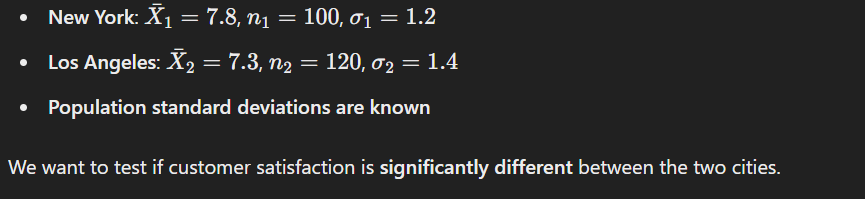




**(B). Two-Sample Z-Test: Manual Calculation & Example**

**Industry Example:** Customer Satisfaction in Two Cities

A retail company wants to compare customer satisfaction scores between New York and Los Angeles. They collect random samples from both cities.

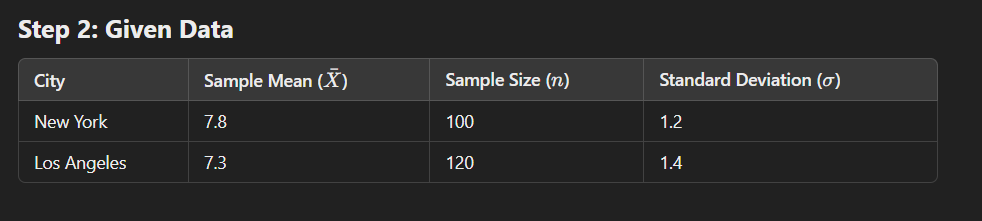
****

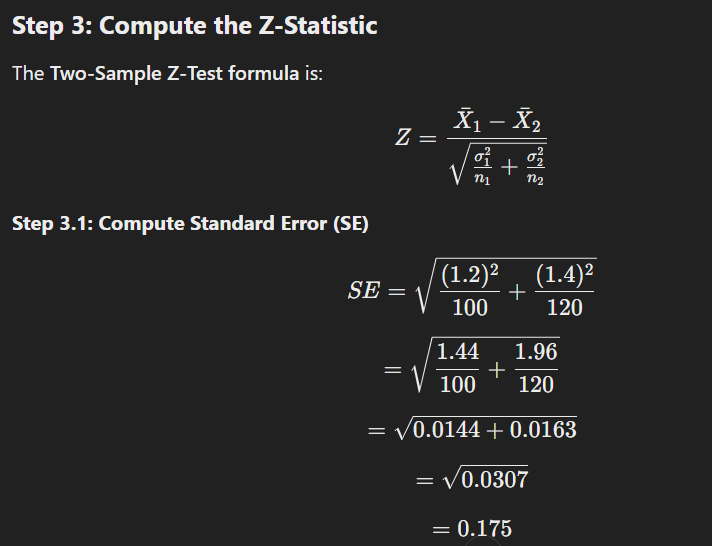
Step-by-Step Calculation of Two-Sample Z-Test

**Step 1: Define Hypotheses**

* Null Hypothesis (H0​): There is no difference in customer satisfaction (μ1=μ2)
* Alternative Hypothesis (H1​): Customer satisfaction is different (μ1≠μ2​)

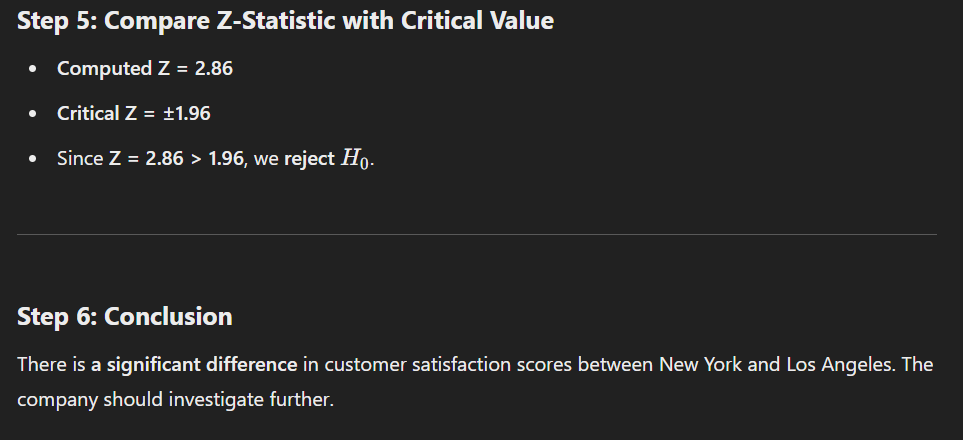
This is a two-tailed test.

****

****

**A screenshot of a computer

Description automatically generated**

****

**A screenshot of a computer

Description automatically generated**

**3. ANOVA (Analysis of Variance) – Detailed Explanation & Industrial Example**

**What is ANOVA?**

ANOVA (Analysis of Variance) is a hypothesis testing technique used to compare the means of three or more groups to determine if they are significantly different.

It extends the t-Test, which compares only two groups, to multiple groups.

**Types of ANOVA**

1. **One-Way ANOVA** – Compares the means of multiple groups based on one independent variable.
   * Example: Comparing customer satisfaction across three different store locations.
2. **Two-Way ANOVA** – Compares means based on two independent variables.
   * Example: Analysing customer satisfaction across three store locations and two different shopping seasons (Summer vs. Winter).

**(A). One-Way ANOVA:** Industrial Example & Step-by-Step Manual Calculation

**Industry Example:** Customer Satisfaction Across Three Store Locations

A retail company wants to check whether customer satisfaction scores differ across three store locations (New York, Los Angeles, and Chicago).

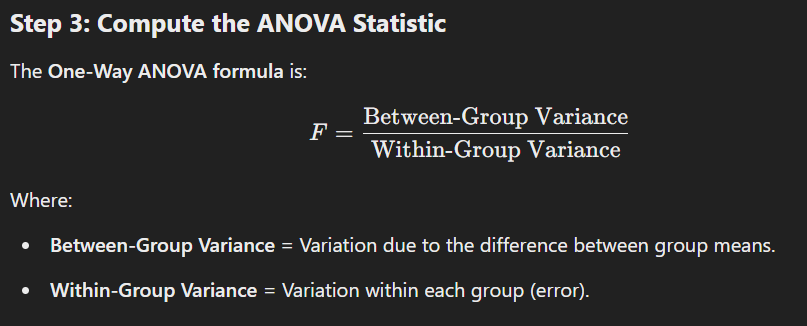
**Step 1: Define Hypotheses**

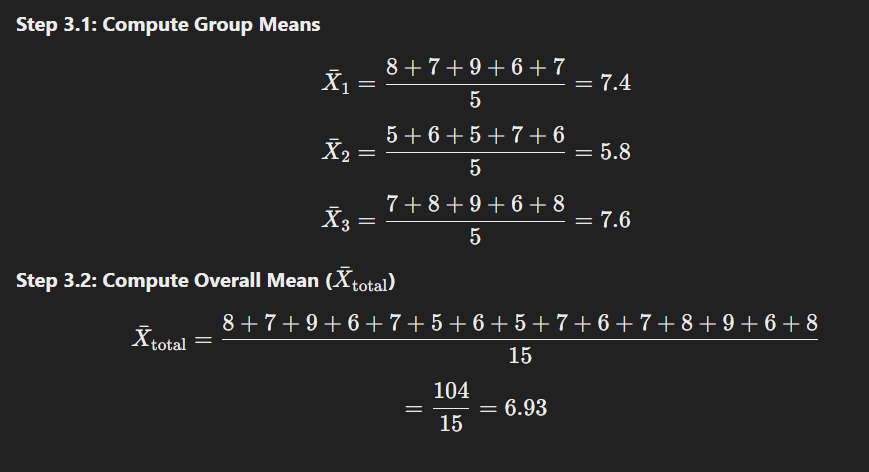
* Null Hypothesis (H0​): All three store locations have the same average customer satisfaction score.
* Alternative Hypothesis (H1​): At least one store has a different customer satisfaction score.

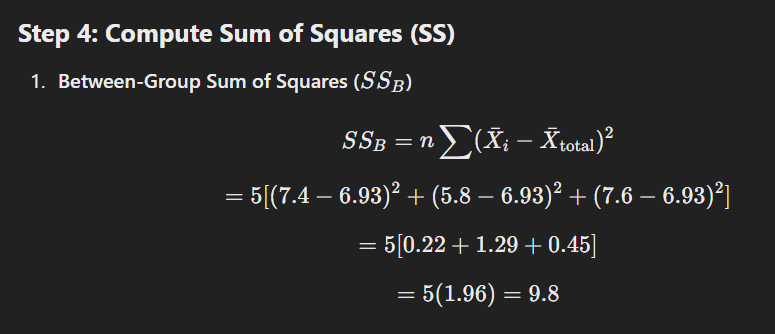
This is a one-way ANOVA test because we are testing only one independent variable (Store Location).

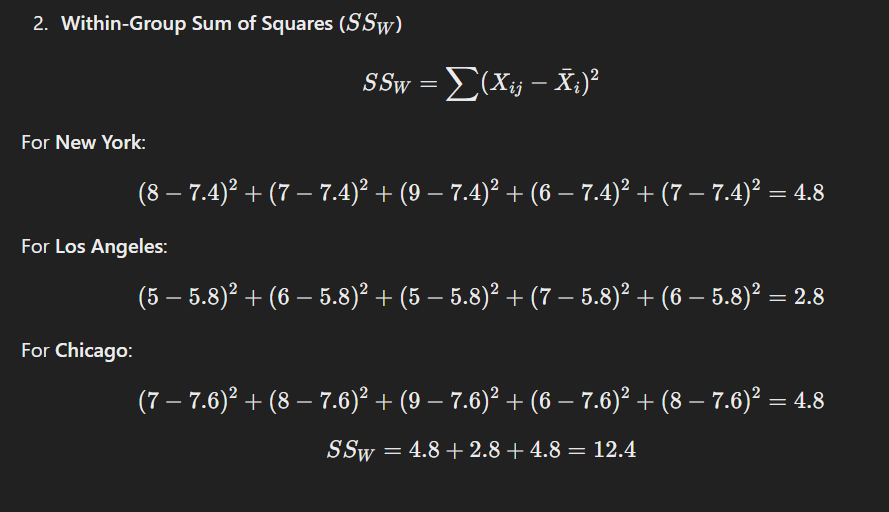
**A screenshot of a computer

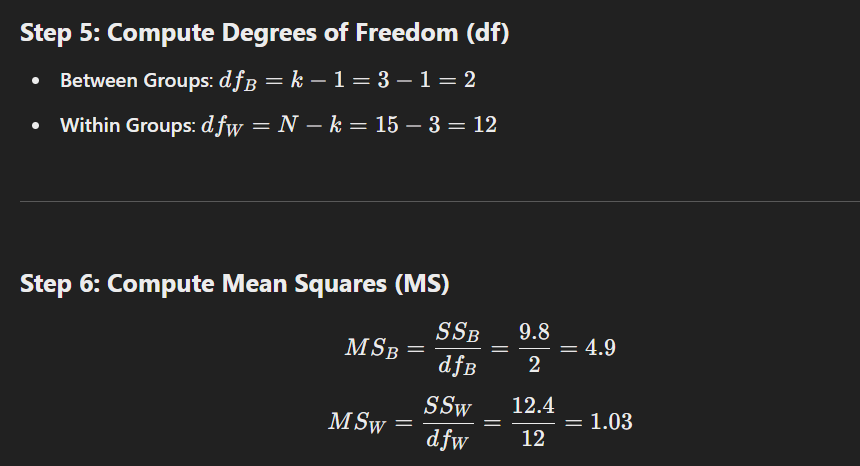
Description automatically generated**

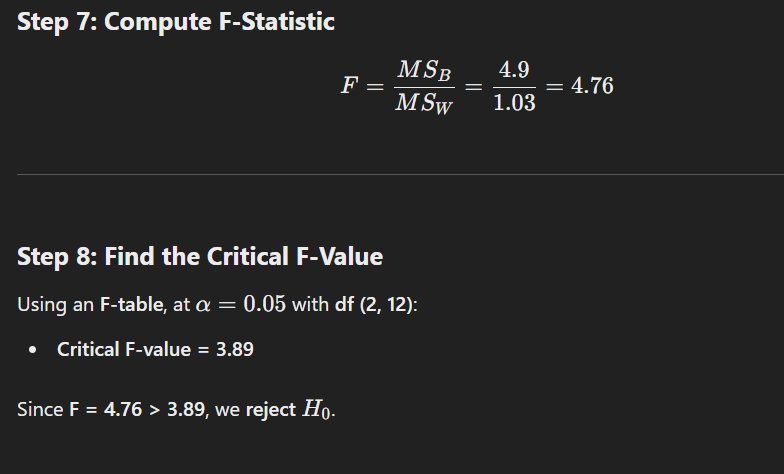
****

****

****

****

****

****

**Step 9: Conclusion**

At least one store has a significantly different customer satisfaction score. The company should investigate which store has the highest or lowest satisfaction.

**(B). Two-Way ANOVA: Practical Example with Step-by-Step Calculation**

**Scenario: Customer Satisfaction Across Store Locations and Shopping Seasons**

A retail company wants to analyse how **customer satisfaction (out of 10)** is affected by two factors:

1. **Store Location** (New York, Los Angeles, Chicago)
2. **Shopping Season** (Summer, Winter)

This is a **Two-Way ANOVA** because we have **two independent variables** (Store Location & Shopping Season).

**Step 1: Define Hypotheses**

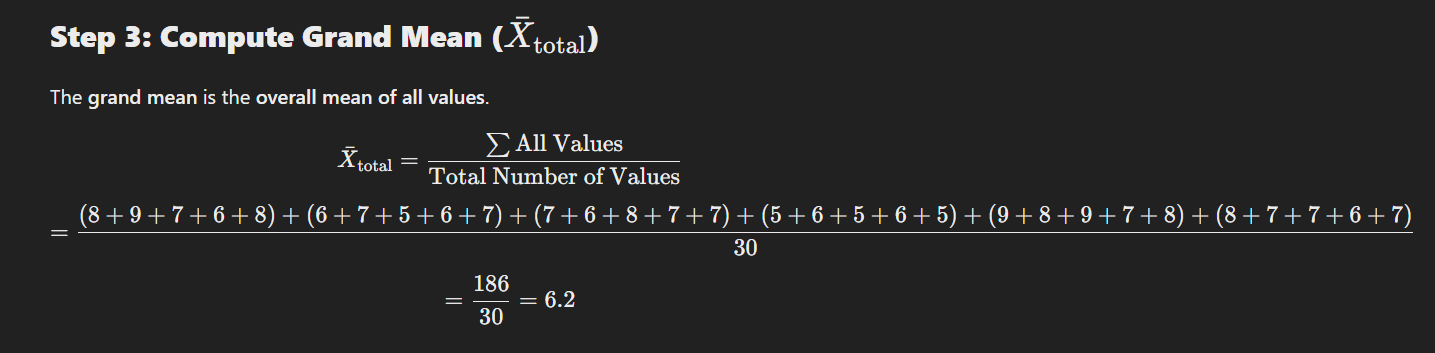
* **Null Hypotheses (H0)**:
  1. **No significant difference in satisfaction across store locations**.
  2. **No significant difference in satisfaction between seasons**.
  3. **No interaction effect** between store location and season.
* **Alternative Hypotheses (H1​)**:
  1. **At least one store has a different satisfaction score**.
  2. **At least one season has a different satisfaction score**.
  3. **There is an interaction effect** (i.e., location and season together influence satisfaction).

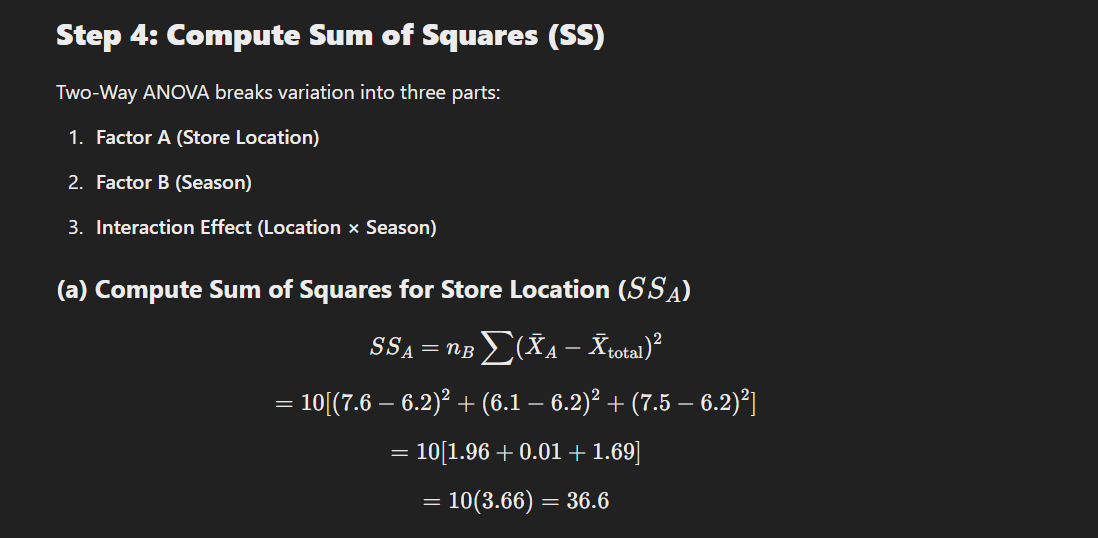
**Step 2: Given Data**

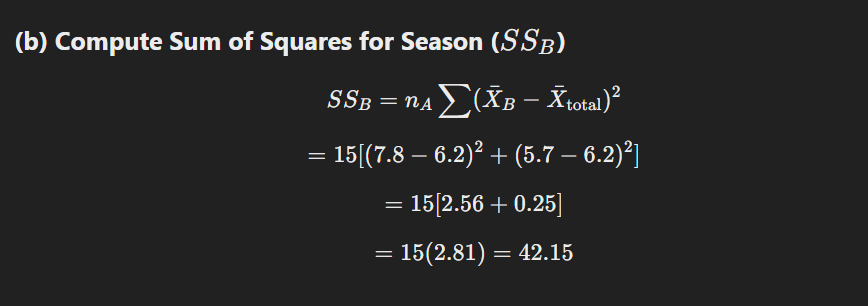
The company surveys **customers from each store during both Summer and Winter**.  
The table below shows **customer satisfaction scores (out of 10):**

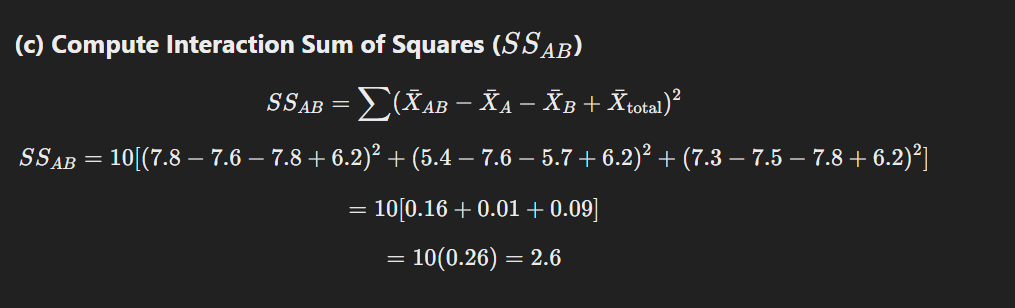
| **Store Location** | **Summer Ratings** | **Winter Ratings** |
| --- | --- | --- |
| **New York** | 8, 9, 7, 6, 8 | 6, 7, 5, 6, 7 |
| **Los Angeles** | 7, 6, 8, 7, 7 | 5, 6, 5, 6, 5 |
| **Chicago** | 9, 8, 9, 7, 8 | 8, 7, 7, 6, 7 |

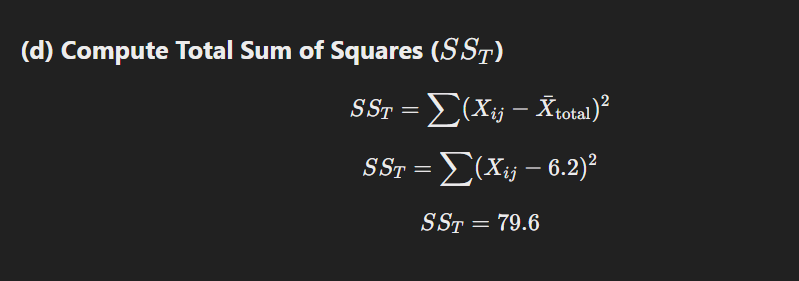
Each store has **two groups** (Summer & Winter), with **5 samples per group**.

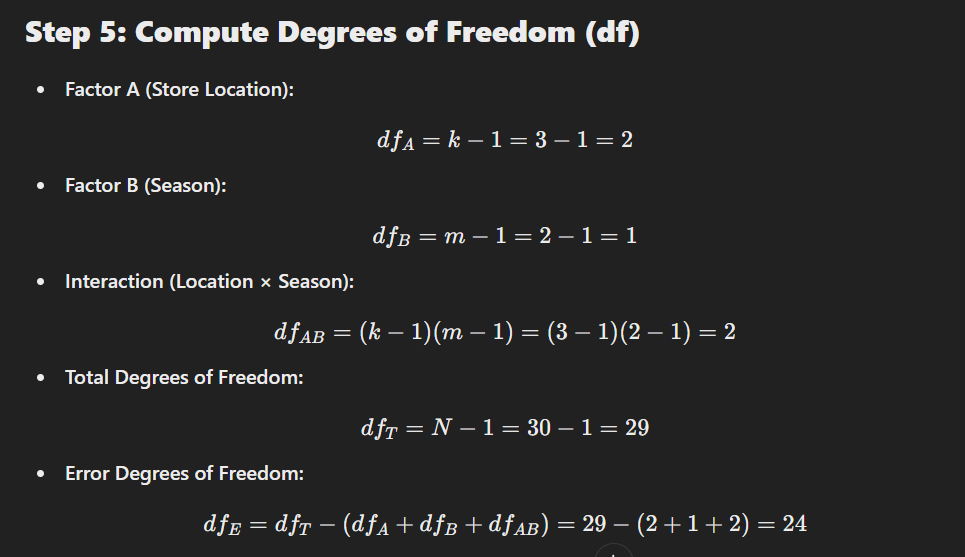


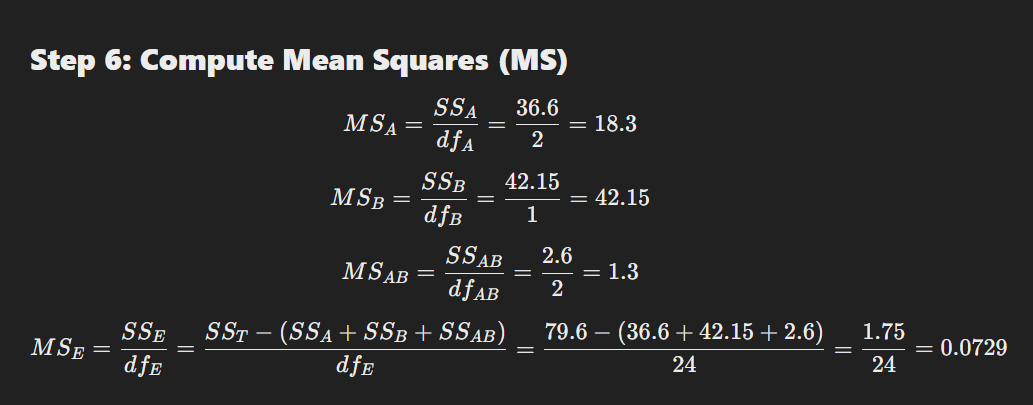


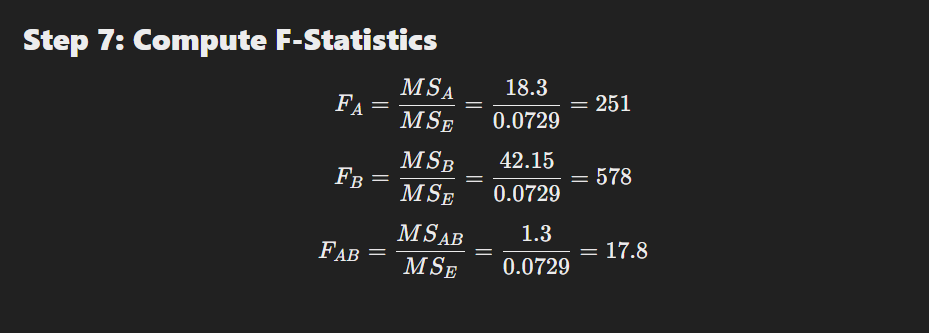


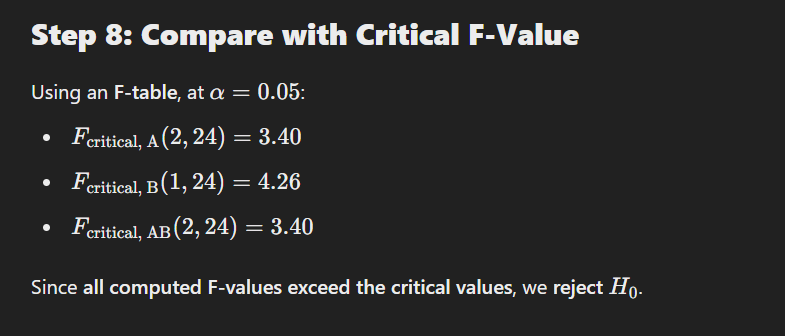


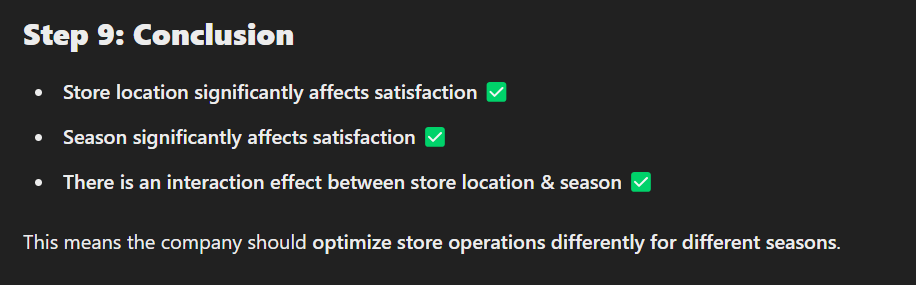












**2. Relationship & Correlation Tests**

* **Pearson Correlation –** Linear relationship between two continuous variables.
* **Spearman’s Rank Correlation –** Non-linear relationship between ordinal or skewed continuous data.
* **Chi-Square Test** – Association between categorical variables (customer preferences, survey analysis).

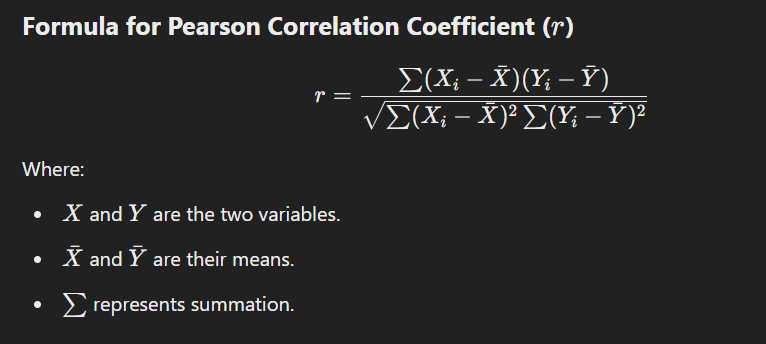
**Pearson Correlation: Detailed Explanation & Industrial Example**

**What is Pearson Correlation?**

Pearson correlation (r) measures the strength and direction of the linear relationship between two continuous variables.

It ranges from -1 to +1:

* r=+1: Perfect positive correlation (as one variable increases, the other increases).
* r=−1: Perfect negative correlation (as one variable increases, the other decreases).
* r=0: No correlation (no linear relationship).



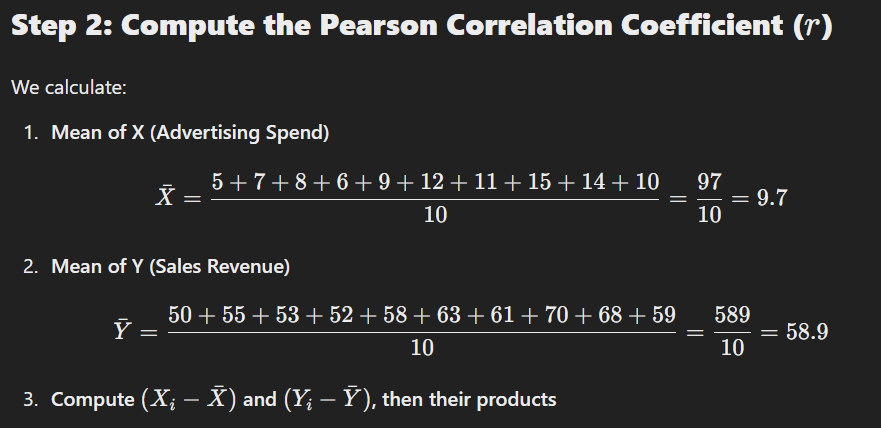
**Industrial Example:** Pearson Correlation in Retail Sales & Advertising Spend

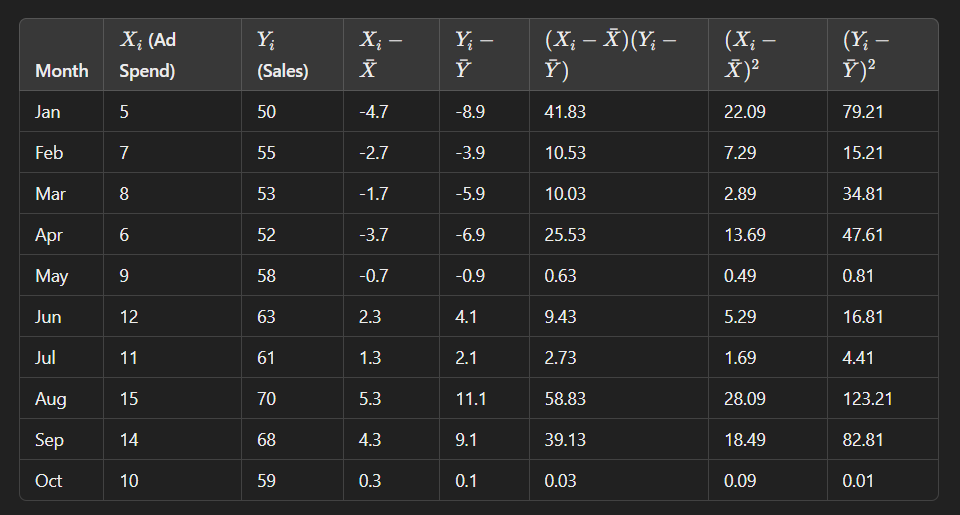
A retail company wants to understand if there is a linear relationship between advertising spend and sales revenue.

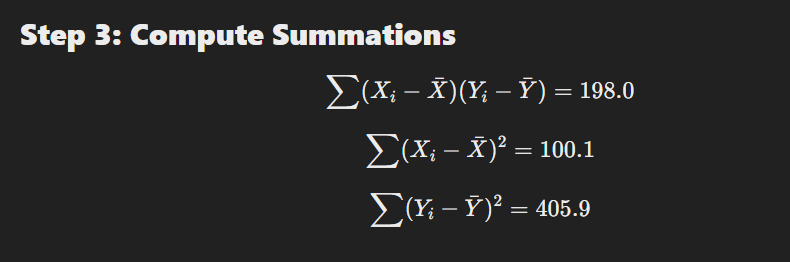
**Step 1: Collect Real-Time Data**

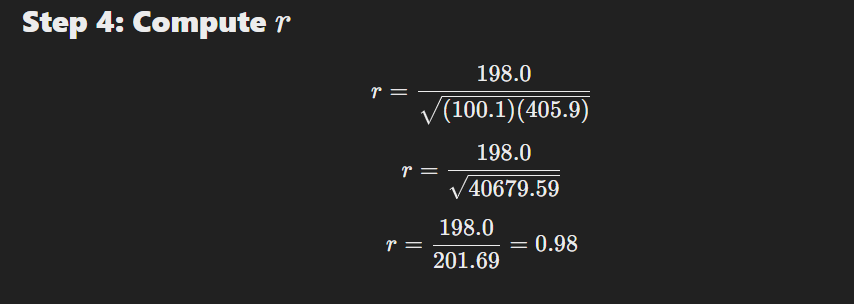
The company collects data over **10 months**:

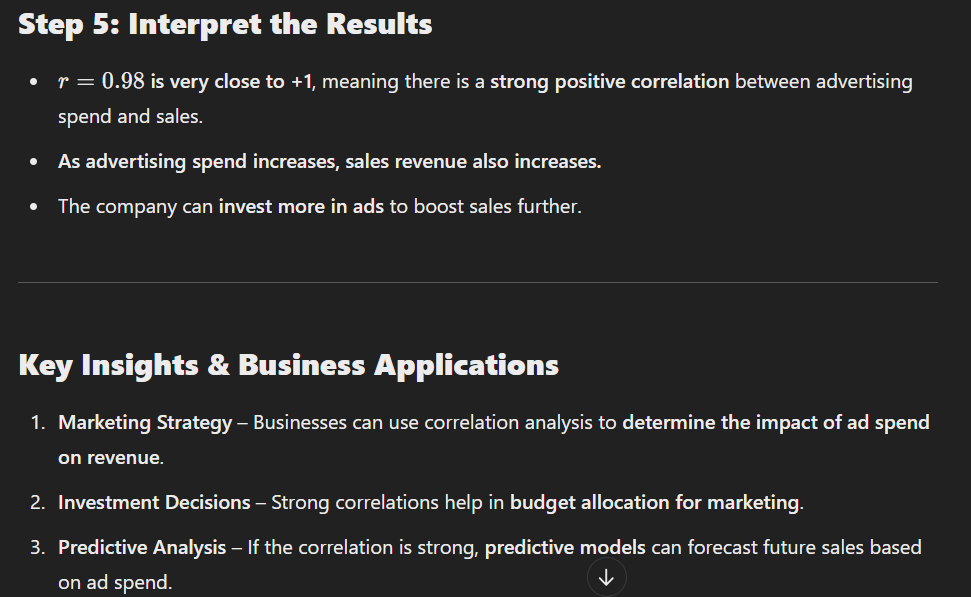












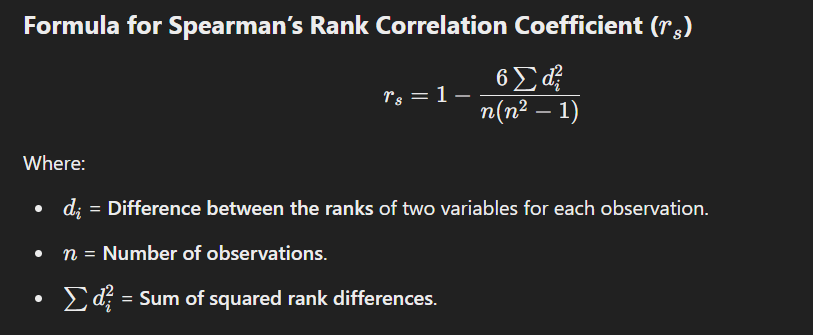
**Spearman’s Rank Correlation: Explanation & Industrial Example**

**What is Spearman’s Rank Correlation?**

Spearman’s Rank Correlation (rs​) measures the **strength and direction of the monotonic relationship** between **two ordinal or non-normally distributed continuous variables**.

Unlike Pearson correlation, which measures **linear relationships**, Spearman correlation captures **monotonic relationships**, meaning:

* **If one variable increases, the other tends to increase or decrease consistently** (but not necessarily at a constant rate).



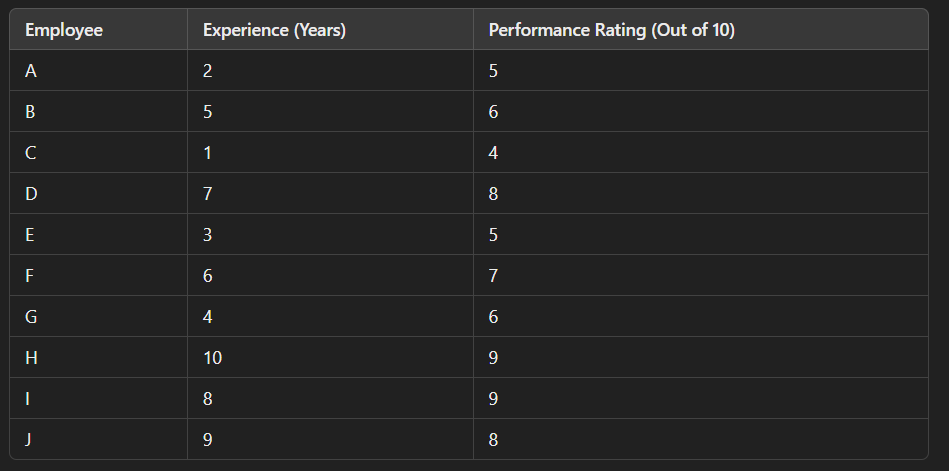
**Industrial Example: Employee Experience & Job Performance**

**Business Case:**

A company wants to analyse whether **employee experience (years of service)** is correlated with **job performance ratings**.

**Step 1: Collect Real-Time Data**

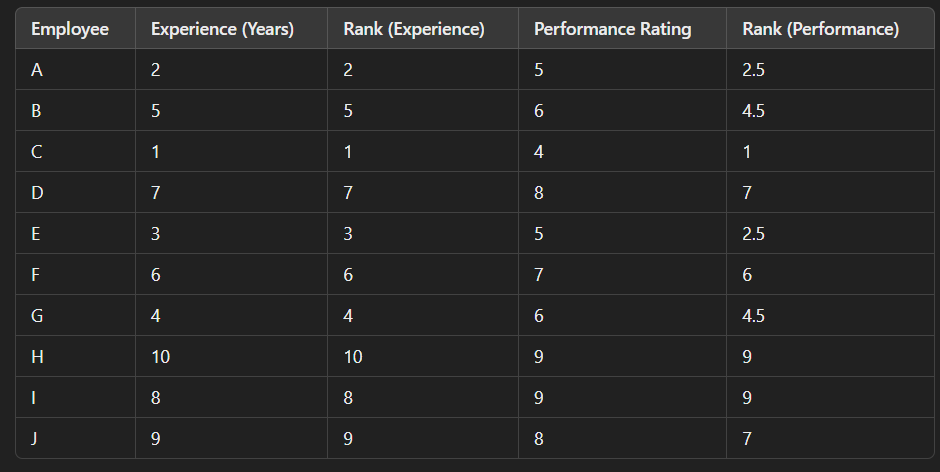
The HR department collects data from **10 employees**.



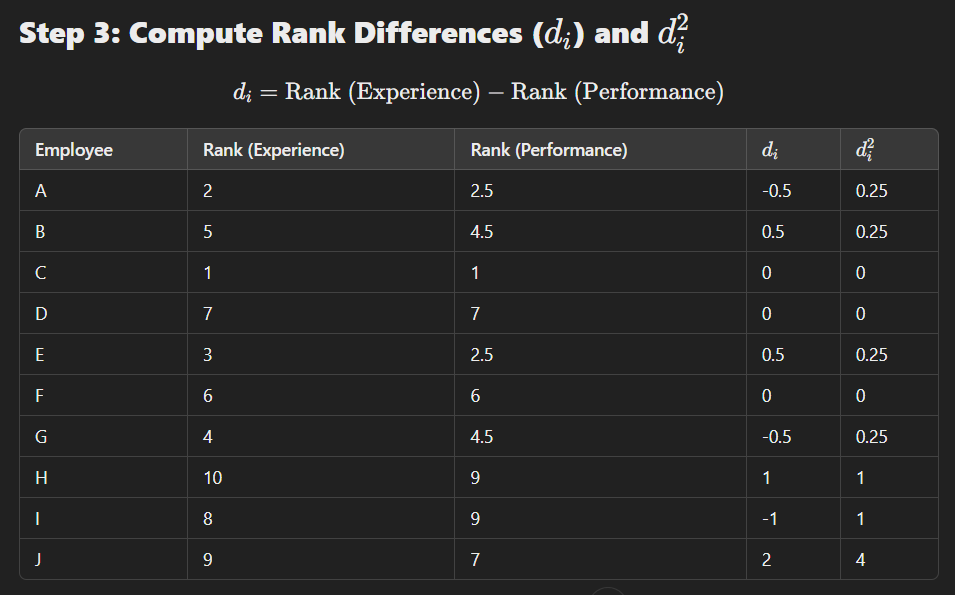
Since **experience is continuous**, but not normally distributed (skewed), and **performance rating is ordinal**, Spearman correlation is ideal.

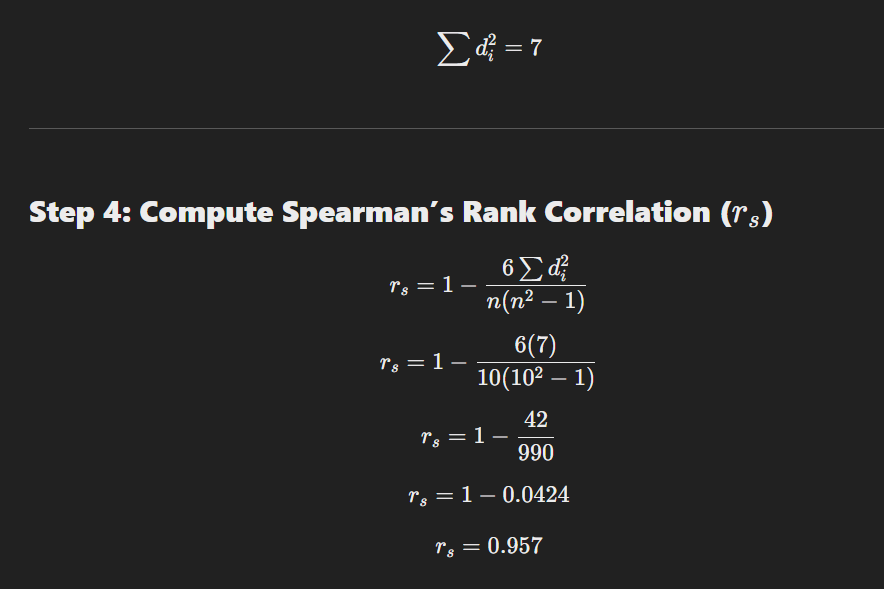
**Step 2: Rank the Data**

We replace raw values with their **ranks**.



If values **tie**, assign them the **average rank** (e.g., both ratings of **5 get rank 2.5**).





**Key Business Insights**

1. **HR Policy** – Spearman’s rank can help HR understand **what truly drives employee satisfaction**.
2. **Non-Linear Relationships** – Unlike Pearson, it detects **non-linear effects** (e.g., salary beyond a certain point has diminishing returns).
3. **Market Research** – Used in **customer preference analysis**, where **brand perception vs. actual sales might not be linear**.

**Chi-Square Test: Explanation & Industrial Example**

**What is the Chi-Square Test?**

The **Chi-Square test** (χ2) is a **statistical test** used to **determine the association between two categorical variables** in a dataset. It compares observed frequencies with expected frequencies under the assumption of independence.

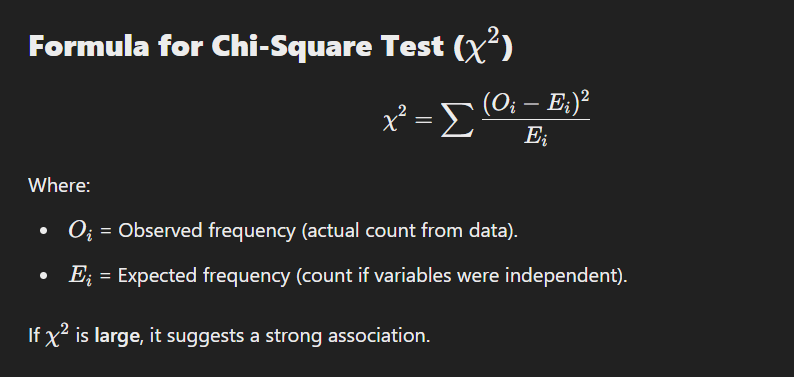
**When to Use the Chi-Square Test?**

* When both variables are **categorical** (e.g., Gender vs. Product Preference).
* To check if there is a **statistically significant relationship** between them.
* Common in **market research, healthcare, and business analytics**.

**Types of Chi-Square Tests**

1. **Chi-Square Goodness of Fit** – Tests whether a sample follows a known distribution.
2. **Chi-Square Test of Independence** – Tests whether two categorical variables are related.

We will focus on **Chi-Square Test of Independence**.

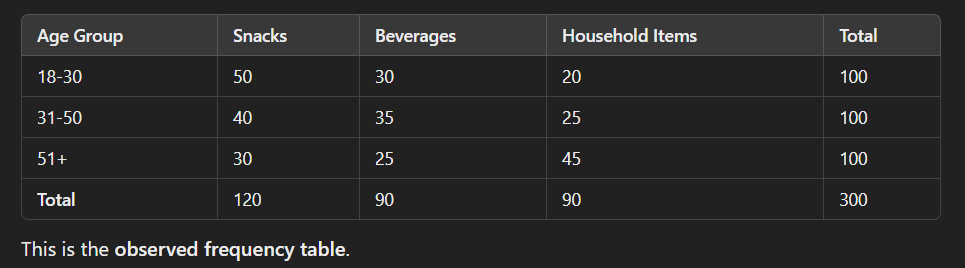


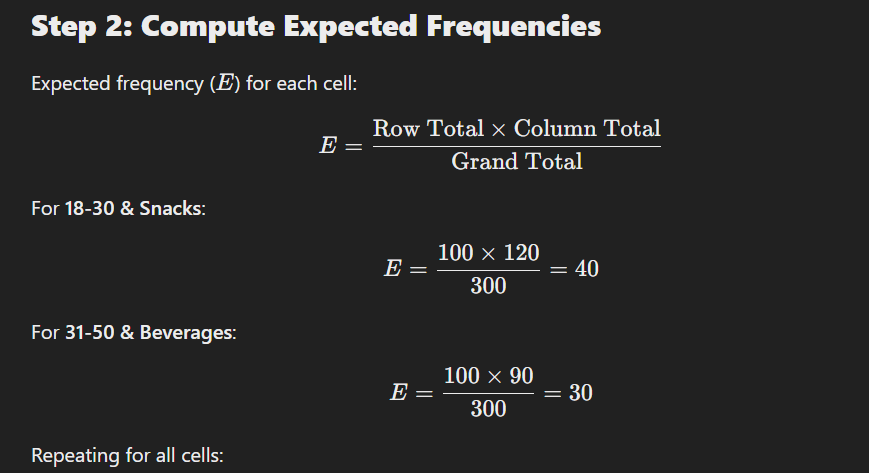
**Industrial Example: Chi-Square Test in Customer Preferences**

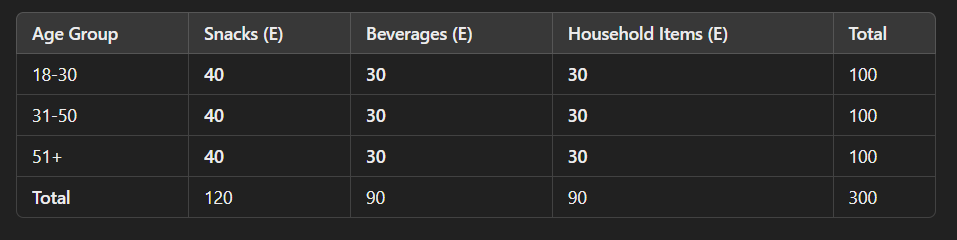
A **supermarket** wants to analyse if there is a **relationship between Age Group and Preferred Product Category**.

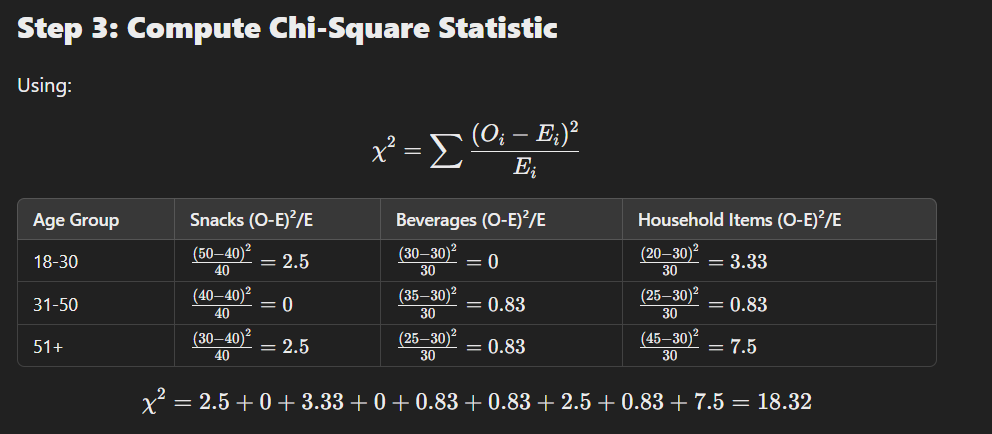
**Step 1: Collect Data (Observed Frequency)**

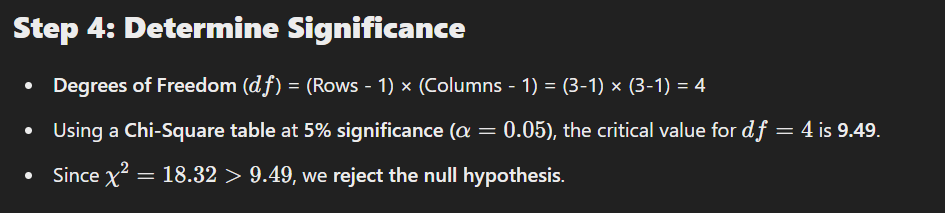
The store surveys **300 customers** about their preferred product category:











**Step 5: Conclusion & Business Insights**

* There is a **significant relationship** between **Age Group and Product Preference**.
* The supermarket can now **target age-specific promotions**:
  + **Younger customers (18-30)** prefer **snacks**, so they should offer **discounts on chips and chocolates**.
  + **Older customers (51+)** prefer **household items**, so they should **increase inventory for cleaning supplies**.
* This helps in **better inventory management and targeted marketing**.

**Regression Analysis (Predictive Analytics)**

**What is Regression Analysis?**

Regression analysis is a **predictive modelling technique** that estimates the relationship between **independent variables (predictors)** and a **dependent variable (outcome)**.

It is widely used for:  
✅ **Sales Forecasting**  
✅ **Trend Analysis**  
✅ **Risk Assessment**  
✅ **Customer Demand Prediction**

**Types of Linear Regression**

1. **Simple Linear Regression** – One independent variable.

Y= β0 + β1X + ϵ

Example: Predicting sales based on advertising spend.

1. **Multiple Linear Regression** – Two or more independent variables.

Y= β0 + β1X1 + β2X2 +...+ βnXn + ϵ

Example: Predicting house prices based on size, location, and number of bedrooms.

**1. Dependent Variable (Y)**

* Y is the **outcome** or **target variable** we are trying to predict.
* It is a **continuous variable** (e.g., sales revenue, customer spending, temperature, etc.).
* Example: **Predicting monthly sales ($1000s) based on advertising spend** → **Y=Sales**.

**2. Intercept (β0)**

* β0 is the **baseline value** of Y when X=0.
* It represents the value of the dependent variable **when there is no influence from the independent variable**.
* Example: If β0=5, then when **ad spend is zero**, the sales will still be **$5000** due to other factors.

**3. Slope Coefficient (β1​)**

* β1 represents the **rate of change** of Y for every **1-unit increase** in X.
* It shows how strongly the independent variable affects the dependent variable.
* Example: If β1= 0.8, it means **every $1000 increase in ad spend increases sales by $800**.

**4. Independent Variable (X)**

* X is the **predictor variable** or **independent variable**.
* It is the **input variable** we use to predict Y.
* Example: **Advertising Spend** is an independent variable used to predict **Sales**.

**5. Error Term (ϵ)**

* ϵ\epsilonϵ represents the **random error** in the model.
* It accounts for all **unexplained factors** that affect Y but are **not included in the model**.
* Example: Sales may be affected by **economic conditions, competitor promotions, weather, etc.**, which are **not captured** by ad spend alone.

**Final Interpretation**

Sales = 5 + 0.8 × (Ad Spend) + ϵ

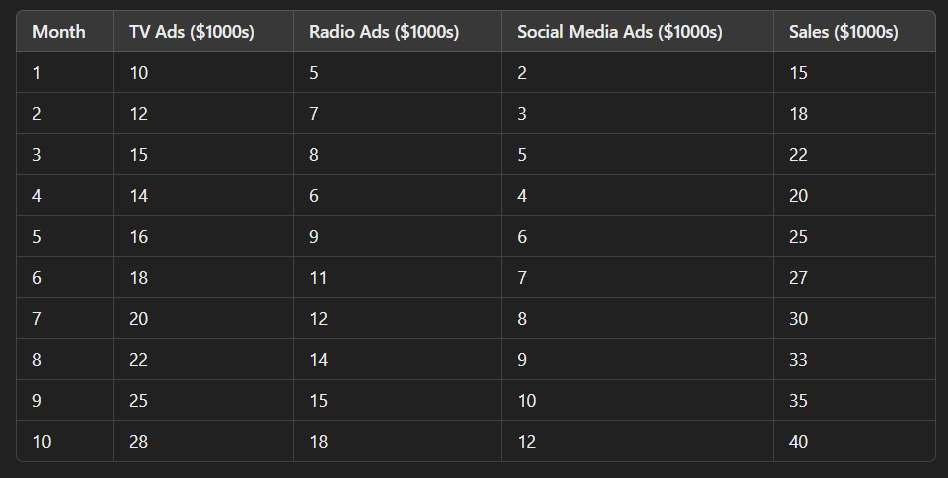
* When **Ad Spend = 0**, the baseline sales are **$5000**.
* For every **$1000 increase in Ad Spend**, sales increase by **$800**.
* The actual sales will **fluctuate** due to external factors (ϵ\epsilonϵ).

**Industrial Example: Sales Forecasting with Multiple Regression**

A company wants to **predict monthly sales ($1000s) based on advertising spend (TV, Radio, social media)**.

**Step 1: Collect Data**

We collect **10 months of data** on sales and ad spending.



**Step 2: Define the Regression Model**

Sales = β0 + β1(TV Ads) + β2(Radio Ads) + β3(Social Media Ads) + ϵ

Where:

* **Sales** = Target variable (Dependent variable).
* **TV Ads, Radio Ads, Social Media Ads** = Independent variables (Predictors).
* **β0​ (Intercept)** = Base sales when no ads are run.
* **β1, β2, β3​** = Regression coefficients showing impact per $1000 spent.

**Step 3: Compute Regression Coefficients**

We calculate the regression coefficients using **Least Squares Method** or software like **Python, Excel, or R**.

Let's assume we obtain:

**Sales = 5 + 0.8(TV Ads) + 0.5(Radio Ads) + 0.7(Social Media Ads)**

**Step 4: Make Predictions**

Now, if the company plans to spend $20,000 on TV, $10,000 on Radio, and $6,000 on social media, the predicted sales:

Sales = 5 + (0.8×20) + (0.5×10) + (0.7×6)

🔹 Predicted Monthly Sales = $30,200

**Step 5: Interpret the Results**

1. TV Ads have the biggest impact (β1=0.8), meaning every $1000 increase in TV ads increases sales by $800.
2. Social Media Ads (β3=0.7) are also effective, so investing more in digital ads is recommended.
3. Radio Ads have a lower impact (β2=0.5), meaning they contribute less to sales.

**Business Insights & Applications**

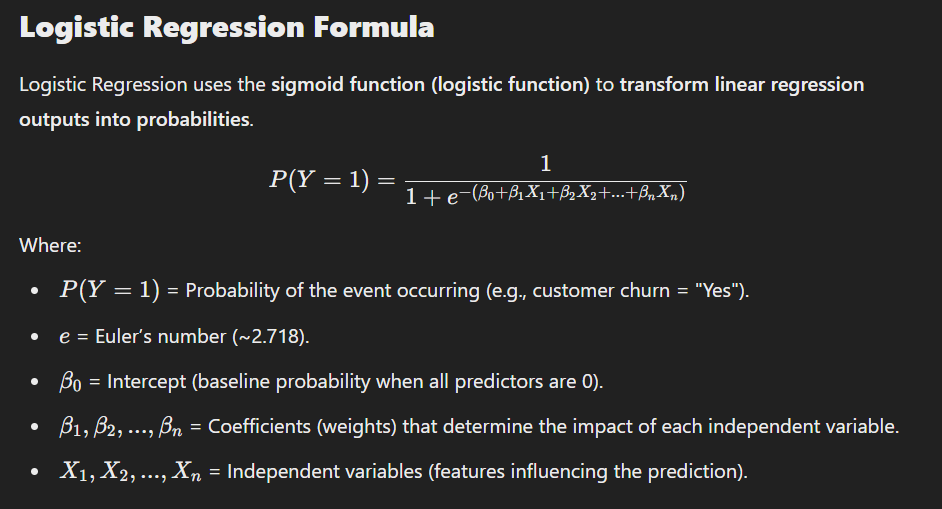
✔ **Budget Allocation** – Shift marketing budget towards the most effective channel (**TV & social media**).  
✔ **Forecasting Revenue** – Helps companies predict future revenue based on planned ad spend.  
✔ **Product Launch Strategy** – Companies can estimate potential sales before launching a new product.

**Logistic Regression: Predictive Analytics for Categorical Outcomes**

**What is Logistic Regression?**

Logistic Regression is a statistical method used for **predicting categorical outcomes**, especially **binary classification problems (Yes/No, 0/1, Success/Failure)**. Unlike **Linear Regression**, which predicts continuous values, Logistic Regression predicts **probabilities**.

It is widely used in:  
✅ **Customer Churn Prediction** (Will a customer leave or stay?)  
✅ **Fraud Detection** (Is a transaction fraudulent or not?)  
✅ **Disease Diagnosis** (Does a patient have a disease or not?)

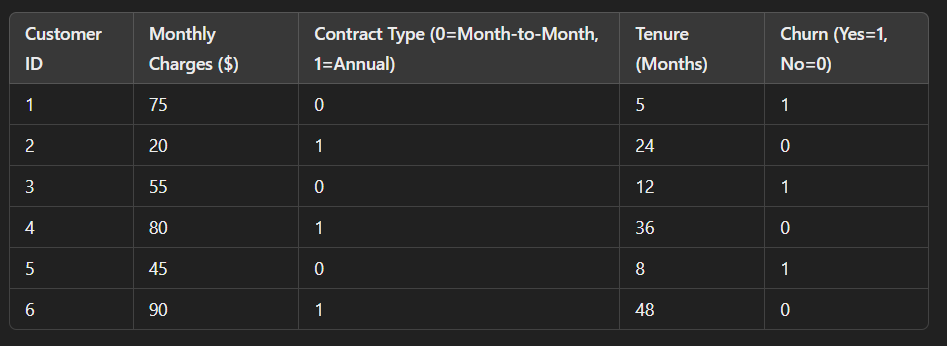


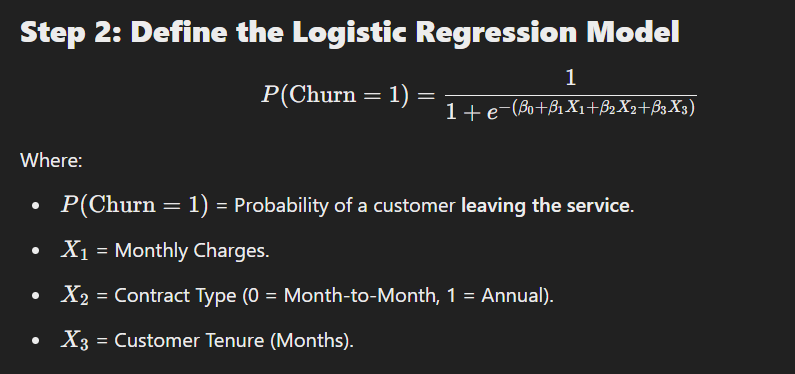
**Industrial Example:** Customer Churn Prediction

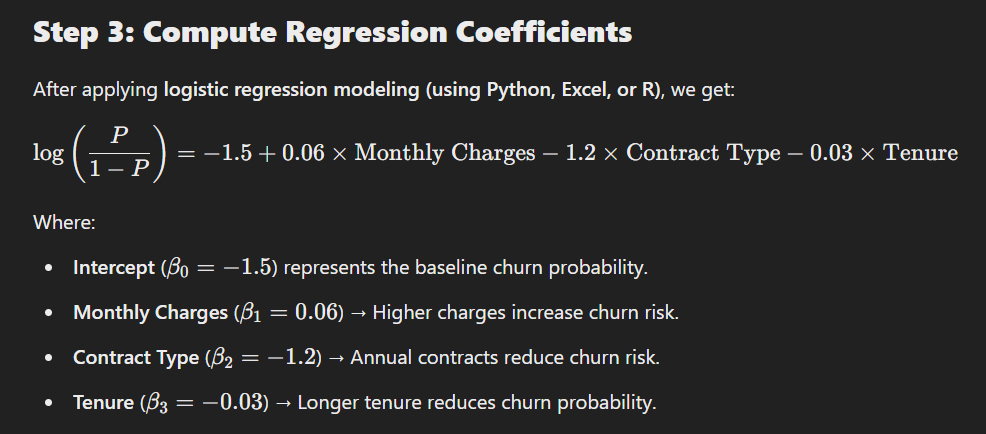
A telecom company wants to predict whether a customer will churn (leave the service) or not based on customer data.

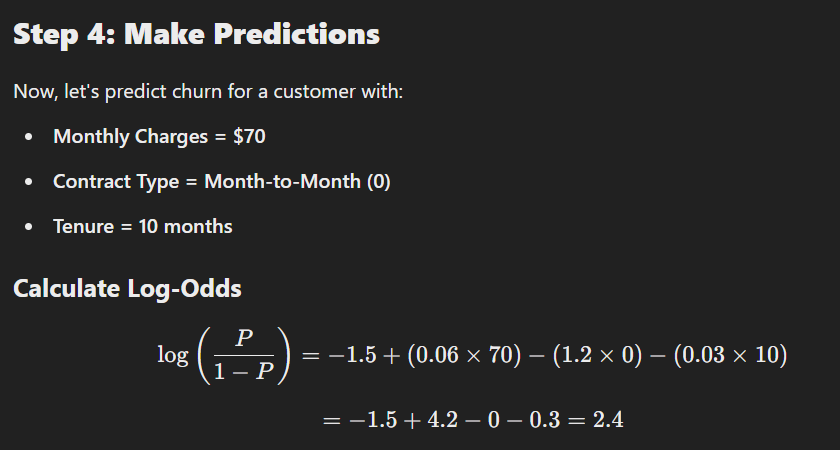
**Step 1: Collect Data**

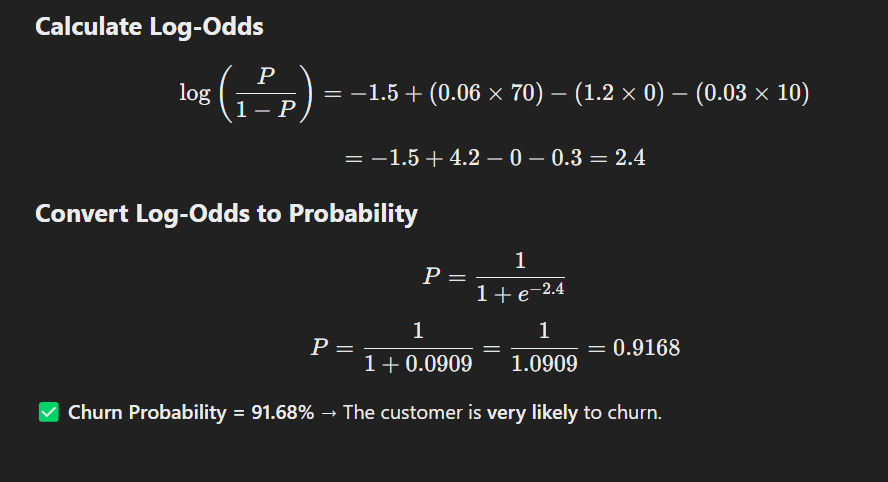
The company collects historical customer data, including Monthly Charges, Contract Type, and Customer Tenure.











**Step 5: Business Insights**

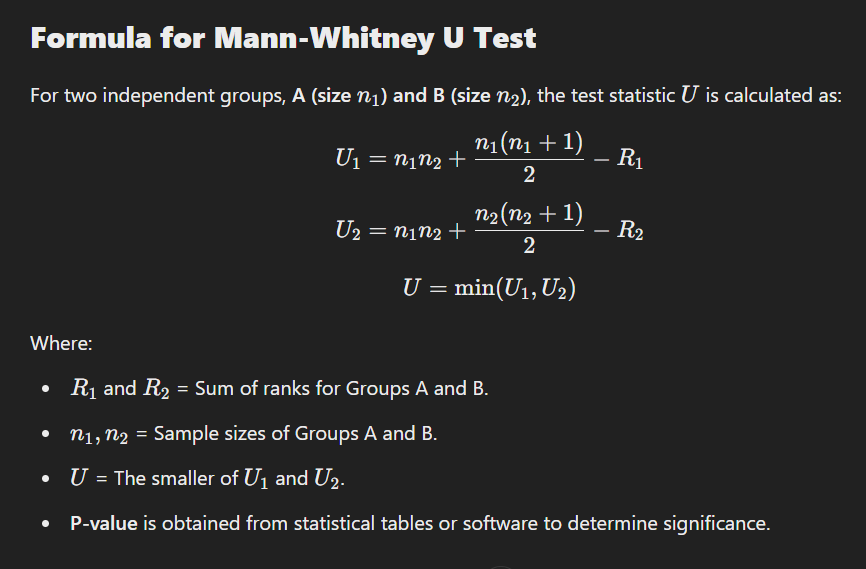
📌 **High Monthly Charges** → Encourage customers to switch to annual plans by offering discounts.  
📌 **Month-to-Month Customers** → Target them with **loyalty programs** to prevent churn.  
📌 **Short Tenure Customers** → Engage them with **onboarding support & discounts** to improve retention.

**Mann-Whitney U Test: A Non-Parametric Alternative to t-Test**

**What is the Mann-Whitney U Test?**

The **Mann-Whitney U test** (also called the **Wilcoxon Rank-Sum test**) is a **non-parametric test** used to compare two independent groups when:  
✔ Data is **not normally distributed** (violates the normality assumption).  
✔ Data is **ordinal, skewed, or contains outliers**.  
✔ Sample sizes are small.

🔹 Unlike the **t-test**, which compares means, the **Mann-Whitney U test compares medians** and determines whether one group tends to have **higher or lower** values than the other.



**Industrial Example: Employee Productivity Analysis**

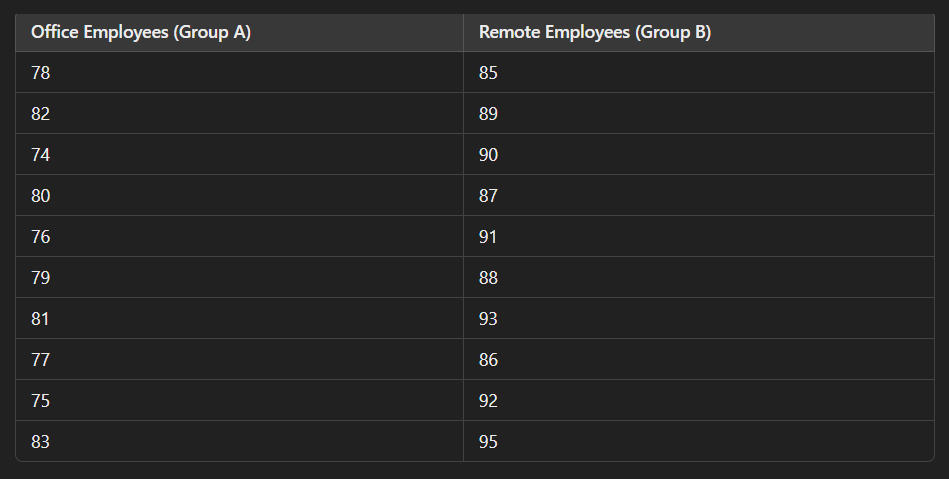
**Scenario**

A company wants to determine if **remote employees** have significantly different productivity compared to **office employees**.

✔ **Group A**: Employees working **from office**.  
✔ **Group B**: Employees working **remotely**.  
✔ **Productivity Score**: Measured on a **100-point scale**.  
✔ Data is **not normally distributed** (skewed), so we use **Mann-Whitney U test** instead of a t-test.

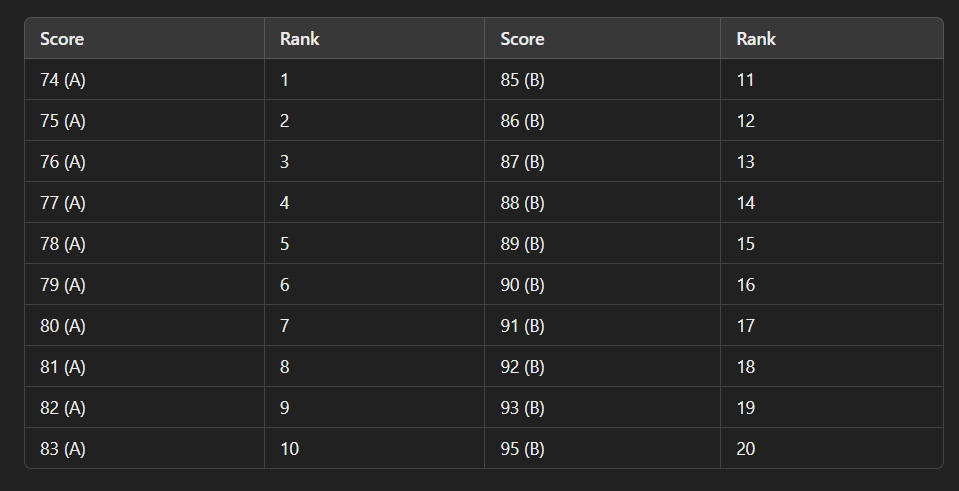
**Step 1: Collect Data**

We collect **productivity scores** from **10 employees** in each group.



**Step 2: Rank the Data**

We rank all **20 scores** from lowest to highest.

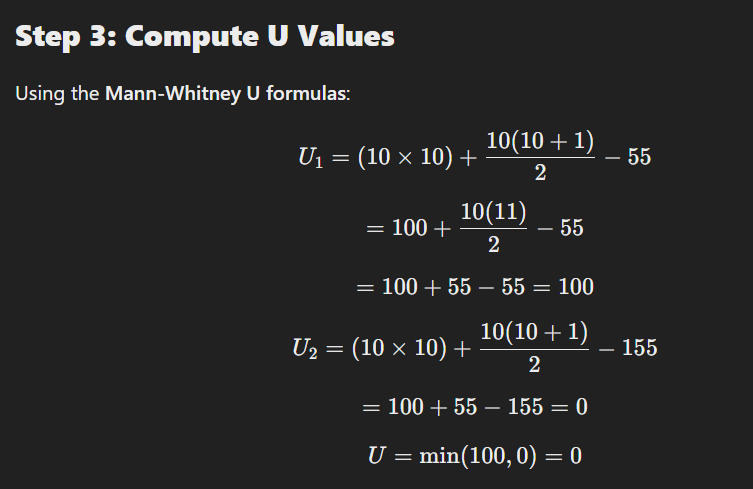


* **Sum of Ranks (Office employees)**:

R1=1+2+3+4+5+6+7+8+9+10=55

* **Sum of Ranks (Remote employees)**:

R2=11+12+13+14+15+16+17+18+19+20=155



**Step 4: Determine Statistical Significance**

We compare the **calculated U-value** with the **critical U-value** from statistical tables at **α=0.05 (95% confidence level)**.

For **n1=10, n2=10 the critical U-value from the table is 23**.

Since **U = 0 < 23**, we **reject the null hypothesis**.

**Step 5: Conclusion & Business Insights**

📌 **Remote employees have significantly higher productivity than office employees.**  
📌 The company should **consider hybrid work models** to improve productivity.  
📌 If there are **other factors** (distractions, work environment), further research should be conducted.

**Key Takeaways**

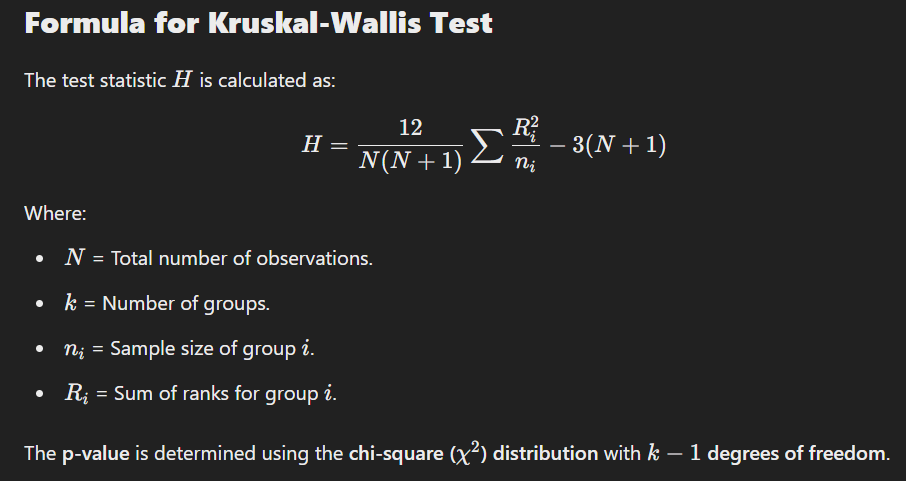
✅ **Mann-Whitney U Test** is useful for comparing **non-normal or skewed data**.  
✅ It is an **alternative to the t-test** when assumptions of **normality** and **equal variance** are violated.  
✅ Widely used in **HR analytics, medicine, customer behaviour studies, and operational performance analysis**.

**Kruskal-Wallis Test: A Non-Parametric Alternative to ANOVA**

**What is the Kruskal-Wallis Test?**

The **Kruskal-Wallis test** is a **non-parametric alternative to One-Way ANOVA** used to compare **three or more independent groups** when:  
✔ The **data is not normally distributed** (violates normality assumption).  
✔ The **groups have different sample sizes**.  
✔ The **data is ordinal or skewed**.

🔹 Unlike **ANOVA**, which compares **means**, the Kruskal-Wallis test **compares medians and rank distributions** to determine if at least one group differs significantly.



**Industrial Example: Customer Satisfaction Analysis**

**Scenario**

A company wants to analyse if **customer satisfaction** levels **differ across three different service centres**.

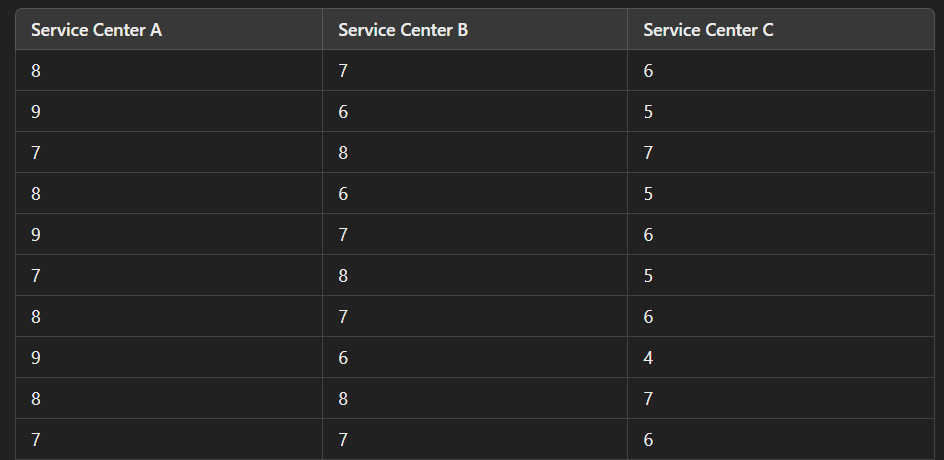
✔ **Groups**:

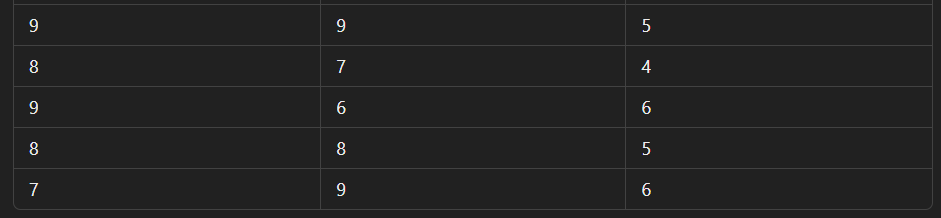
* **Service Centre A**
* **Service Centre B**
* **Service Centre C**

✔ **Customer Satisfaction Scores** (on a scale of 1-10).  
✔ Data is **not normally distributed** (skewed), so we use the **Kruskal-Wallis test** instead of ANOVA.

**Step 1: Collect Data**

Customer satisfaction scores from 15 customers at each center:



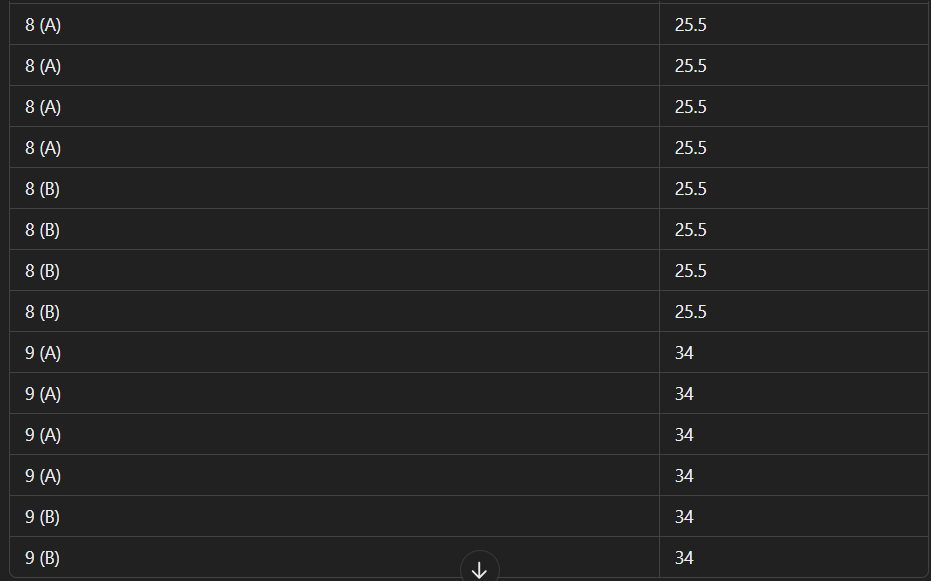


**Step 2: Rank the Data**

We rank all **45 scores** from lowest to highest.







* **Sum of Ranks (A)**:

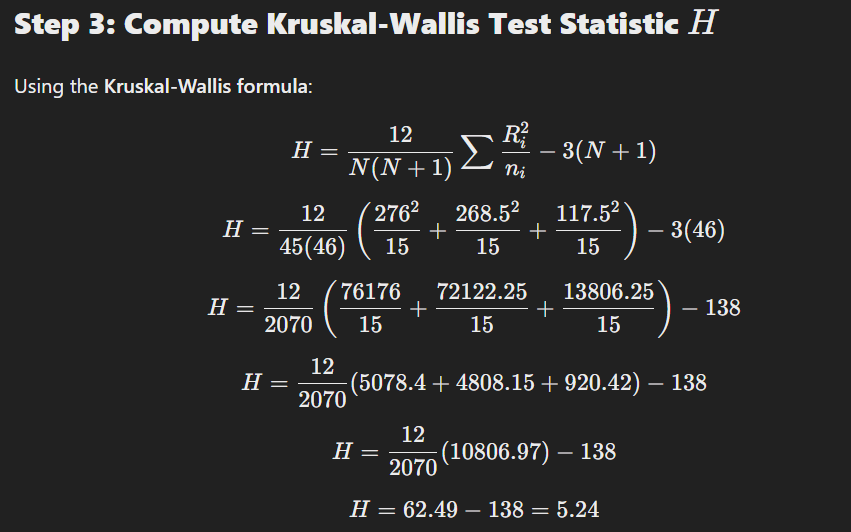
RA=25.5+25.5+25.5+25.5+34+34+34+34+17+17=276

* **Sum of Ranks (B)**:

RB=17+17+17+17+17+17+17+17+25.5+25.5+25.5+25.5+34+34=268.5

* **Sum of Ranks (C)**:

RC=1.5+1.5+3.5+3.5+3.5+3.5+3.5+10+10+10+10+10+10+10+17=117.5



**Step 4: Determine Statistical Significance**

We compare the **calculated H=5.24** with the **critical chi-square value** at α=0.05 for **df = k−1 = 3−1 = 2**

📌 **Critical χ2 value from table** = **5.99**.

Since **H = 5.24 < 5.99,** we **fail to reject the null hypothesis**.

**Step 5: Conclusion & Business Insights**

📌 **No significant difference** in customer satisfaction among the three service centres.  
📌 The company does **not** need to prioritize changes in a specific centre.  
📌 **Further investigation** may be needed into **individual service factors**.

**Key Takeaways**

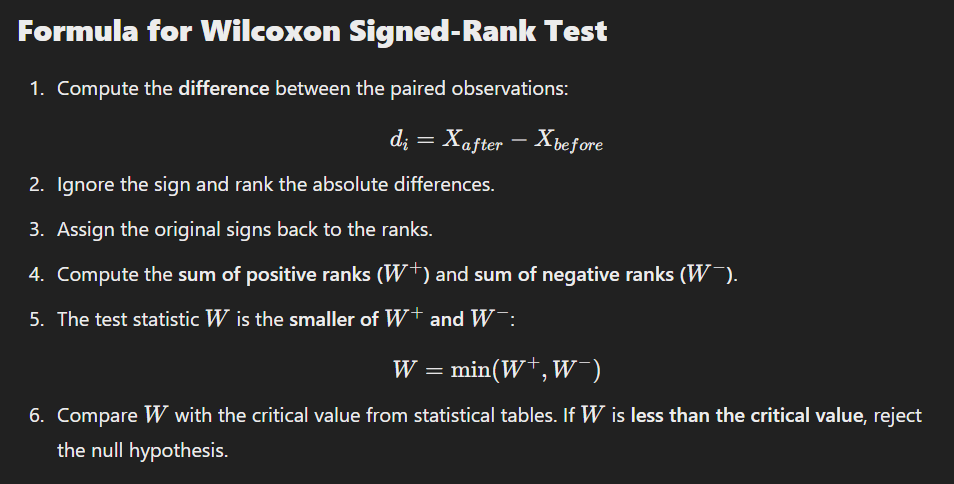
✅ **Kruskal-Wallis Test** is a powerful alternative when **ANOVA assumptions are violated**.  
✅ It is widely used in **customer satisfaction analysis, HR evaluations, healthcare studies, and market research**.  
✅ Works best when **data is ordinal, skewed, or has outliers**.

**Wilcoxon Signed-Rank Test: A Non-Parametric Alternative to Paired t-Test**

**What is the Wilcoxon Signed-Rank Test?**

The Wilcoxon Signed-Rank Test is a non-parametric alternative to the paired t-test used when:  
✔ The data is not normally distributed (violates normality assumption).  
✔ The same group is measured before and after a treatment (paired samples).  
✔ The goal is to check whether the median difference between paired observations is significantly different from zero.

🔹 Unlike the paired t-test, which compares means, the Wilcoxon Signed-Rank Test compares ranks of differences.



**Industrial Example: Productivity Improvement After Training**

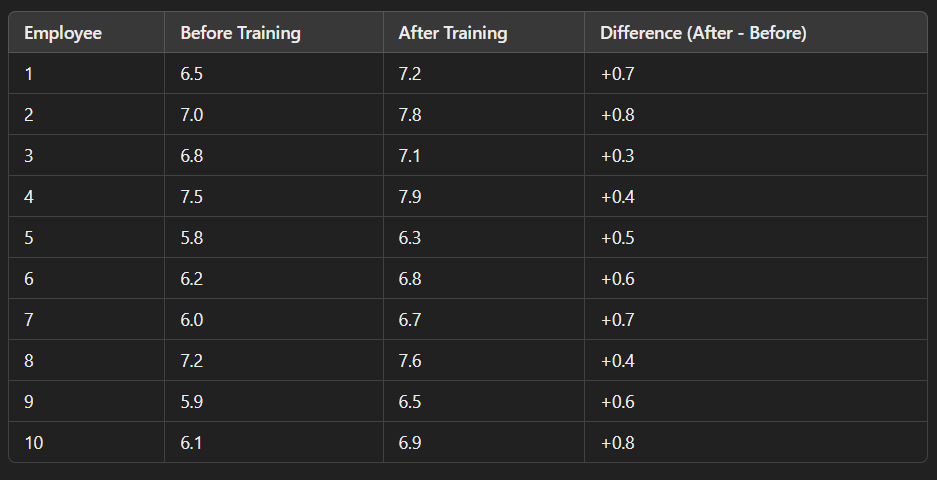
**Scenario**

A company wants to test if a **new employee training program** improves **employee productivity**.

✔ **10 employees** are selected for training.  
✔ Productivity (hours worked effectively per day) is recorded **before** and **after** training.  
✔ Data is **not normally distributed** (so we use Wilcoxon instead of paired t-test).

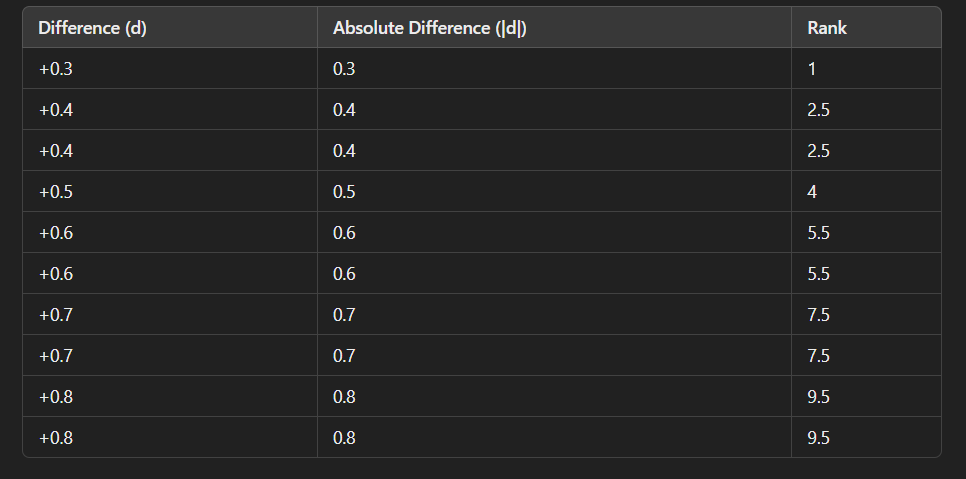
**Step 1: Collect Data**

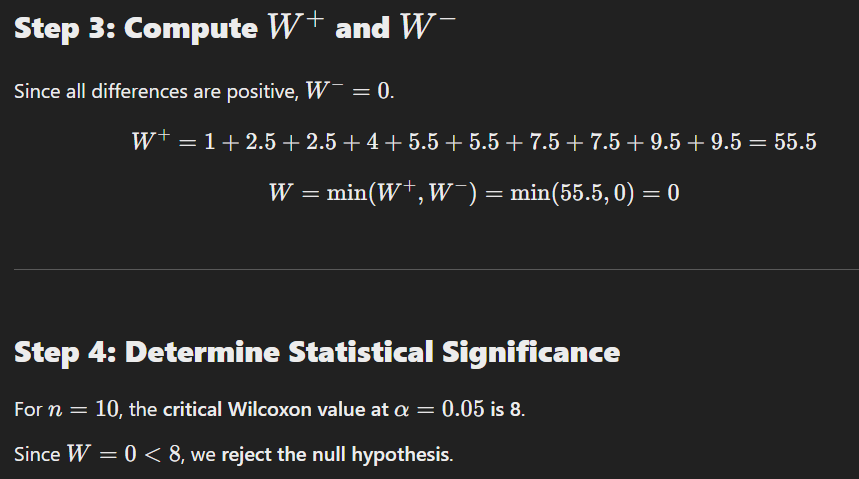
Employee productivity before and after training:



**Step 2: Rank Absolute Differences**

We rank the absolute differences from smallest to largest.





**Step 5: Conclusion & Business Insights**

📌 The **training program significantly improves employee productivity**.  
📌 The company should **continue the training program** for all employees.  
📌 **Further research** can explore **which aspects** of the training contributed most to the improvement.

**Key Takeaways**

✅ **Wilcoxon Signed-Rank Test** is useful for **paired samples** when **data is non-normal**.  
✅ It is widely used in **HR analytics, healthcare research, product testing, and finance**.  
✅ Works best for **small sample sizes** with **skewed data or outliers**.

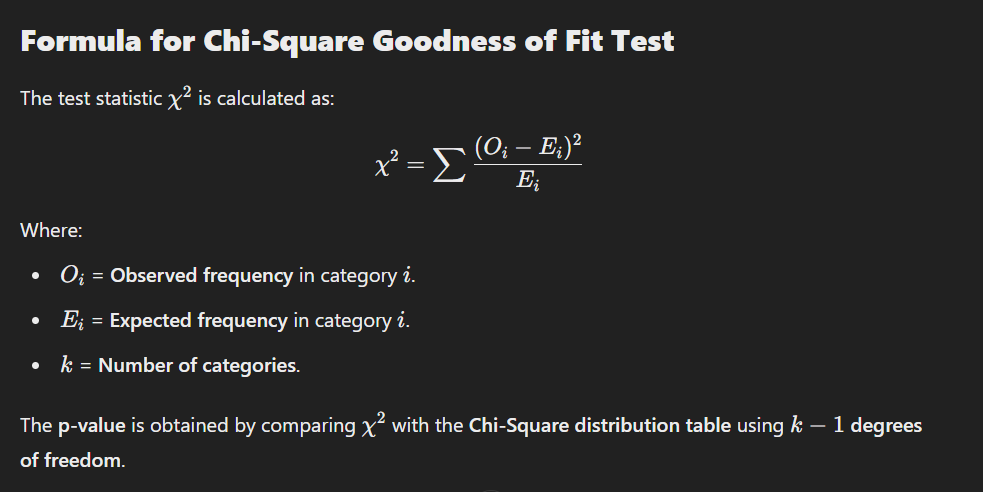
**Chi-Square Goodness of Fit Test: Testing If Observed Frequencies Match Expected Frequencies**

**What is the Chi-Square Goodness of Fit Test?**

The **Chi-Square Goodness of Fit test** is a **non-parametric test** used to determine **if an observed distribution fits an expected distribution**.

**When to Use This Test?**

✔ The data consists of **categorical variables**.  
✔ You want to check **if an observed frequency distribution differs from an expected distribution**.  
✔ The sample size is **large enough** (expected frequency in each category should be **≥ 5**).



**Industrial Example: Customer Purchase Behavior in a Retail Store**

**Scenario**

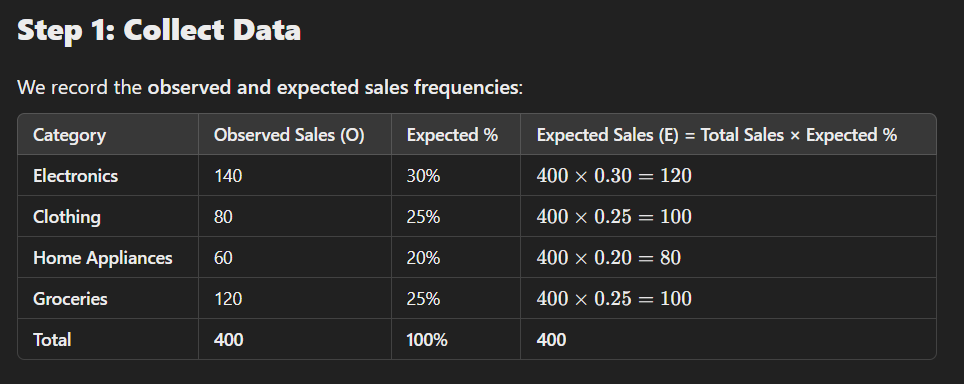
A retail store sells products in **four categories**:

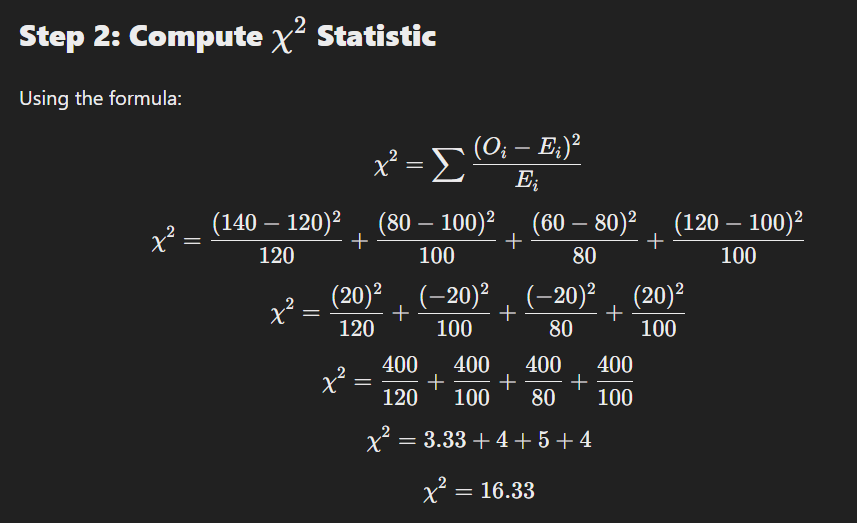
1. **Electronics**
2. **Clothing**
3. **Home Appliances**
4. **Groceries**

The store **expects** sales to be **proportional** to the following breakdown, based on past data:

* **Electronics**: 30%
* **Clothing**: 25%
* **Home Appliances**: 20%
* **Groceries**: 25%

However, after a **marketing campaign**, the store collects **actual sales data** from **400 customers** and wants to check if the **new purchase pattern follows the expected distribution**.





**Step 3: Determine Statistical Significance**

We compare the **calculated χ2=16.33** with the **critical Chi-Square value** from a statistical table.

**Degrees of Freedom**

df = k – 1 = 4−1 = 3

From the **Chi-Square table**, the **critical value at α=0.05 or df = 3** is **7.815**.

**Compare**

Since **16.33 > 7.815**, we **reject the null hypothesis**.

**Step 4: Conclusion & Business Insights**

📌 The **new purchase pattern is significantly different** from the expected sales distribution.  
📌 The **marketing campaign changed customer purchasing behaviour**.  
📌 The store should analyse **why Electronics sales increased** and **why Clothing & Home Appliances underperformed**.

**Key Takeaways**

✅ **Chi-Square Goodness of Fit Test** is used to check if **observed frequencies match expected frequencies**.  
✅ It is widely used in **retail, finance, healthcare, marketing, and manufacturing**.  
✅ Works best for **categorical data with large sample sizes**.

**5. Proportion & Frequency Tests**

* **Chi-Square Goodness of Fit** – Testing if observed frequencies match expected frequencies.
* **Z-Test for Proportions** – Comparing proportions of two groups (conversion rate analysis)

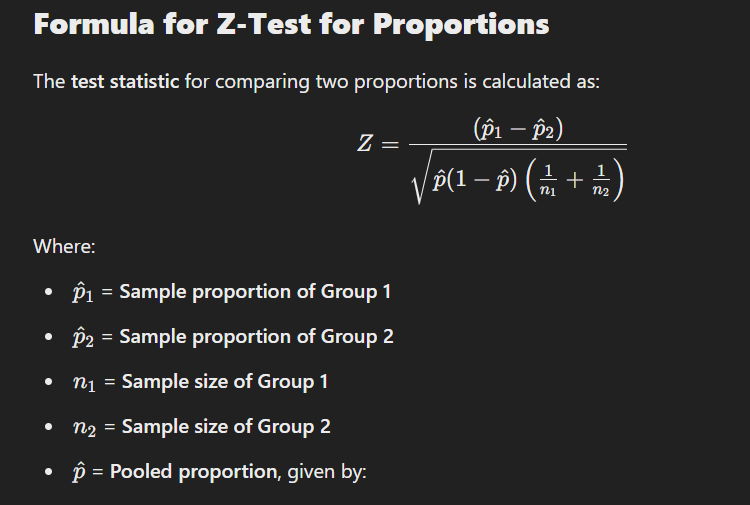
**Z-Test for Proportions: Comparing Proportions of Two Groups**

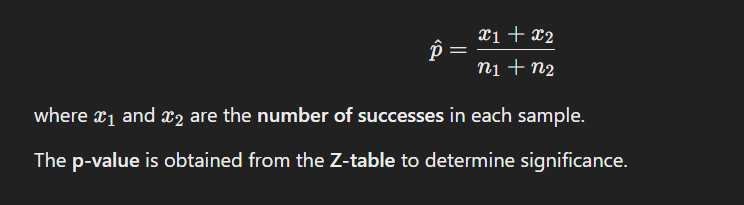
**What is the Z-Test for Proportions?**

A **Z-Test for Proportions** is a **statistical test** used to compare the **proportions of two independent groups** and determine if the difference between them is statistically significant.

**When to Use the Z-Test for Proportions?**

✔ The data is **categorical** (e.g., success/failure, converted/not converted).  
✔ You are comparing **two independent proportions**.  
✔ The sample size is **large enough** (np ≥ 5 and nq ≥ 5 ).





**Industrial Example: A/B Testing in Digital Marketing**

**Scenario**

A company runs an **A/B test** to determine if a **new landing page design** improves the **conversion rate** compared to the **existing design**.

✔ **Group 1 (Old Design)**: 1000 visitors, **120 converted**  
✔ **Group 2 (New Design)**: 1200 visitors, **180 converted**

The company wants to **test if the new design leads to a significantly higher conversion rate**.

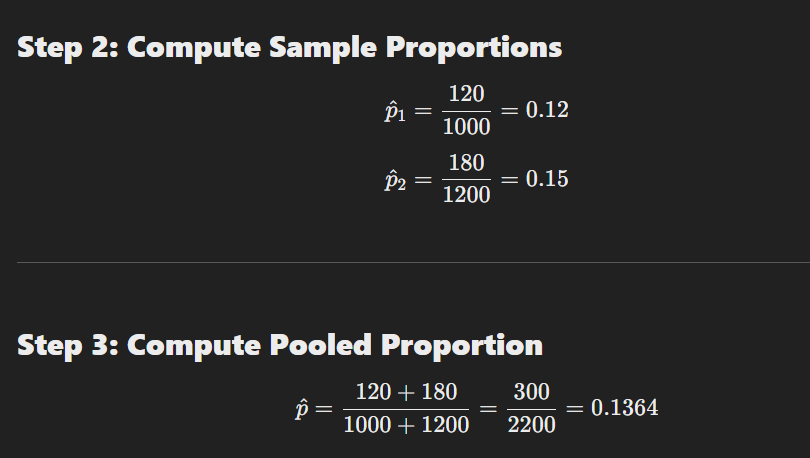
**Step 1: Define Hypotheses**

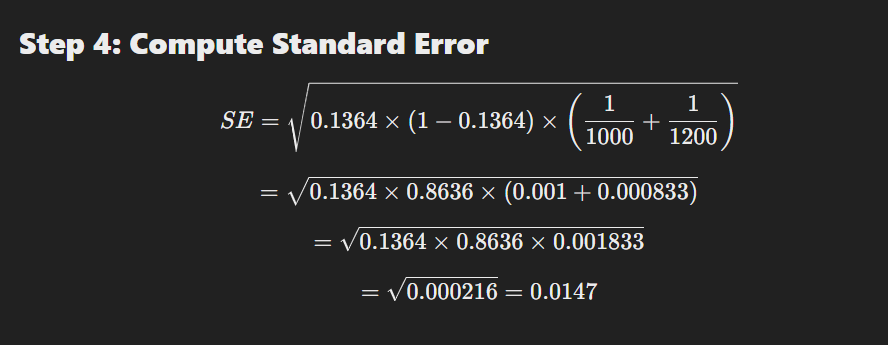
* **Null Hypothesis H0​**: The conversion rates of the two groups are **equal**.

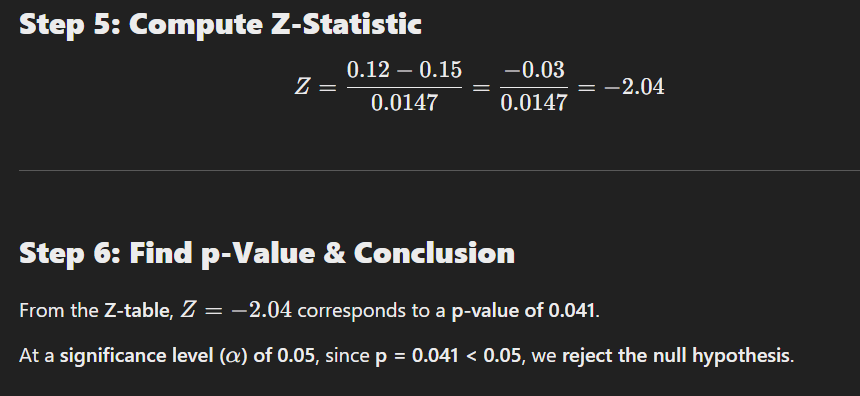
H0: p1=p2​

* **Alternative Hypothesis H1​**: The conversion rates are **different**.

Ha: p1≠p2 ​







**Step 7: Business Insights**

📌 **The new landing page design significantly improves conversion rates**.  
📌 The company should **replace the old design with the new one**.  
📌 Further tests can analyse **why the new design performs better** (e.g., better UX, improved CTA).

**Key Takeaways**

✅ **Z-Test for Proportions** is useful for **A/B testing, marketing analysis, customer surveys, and medical studies**.  
✅ Helps determine if a **change (e.g., new product, policy, design) leads to a significant impact**.  
✅ Works best for **large sample sizes** with **categorical data**.

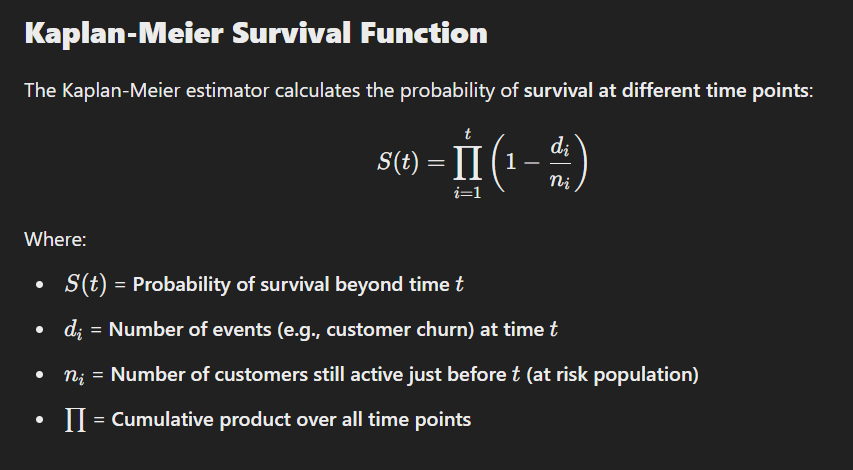
**Kaplan-Meier Curve: Analyzing Survival Data (Customer Retention)**

**What is the Kaplan-Meier Curve?**

The **Kaplan-Meier (KM) curve** is a **non-parametric statistical method** used to estimate the **survival function** over time. It helps analyse **time-to-event data**, such as:  
✔ Customer retention (how long customers stay subscribed)  
✔ Machine failure rates (when a machine breaks down)  
✔ Patient survival rates (how long patients survive after treatment)

**When to Use the Kaplan-Meier Curve?**

✅ When analysing **the time until an event happens** (e.g., customer churn, machine failure).  
✅ When **not all subjects experience the event** (some customers might still be active at the end of the study).  
✅ When survival data is **censored** (some subjects are lost to follow-up before the event occurs).



**Industrial Example: Customer Retention in a Subscription Service**

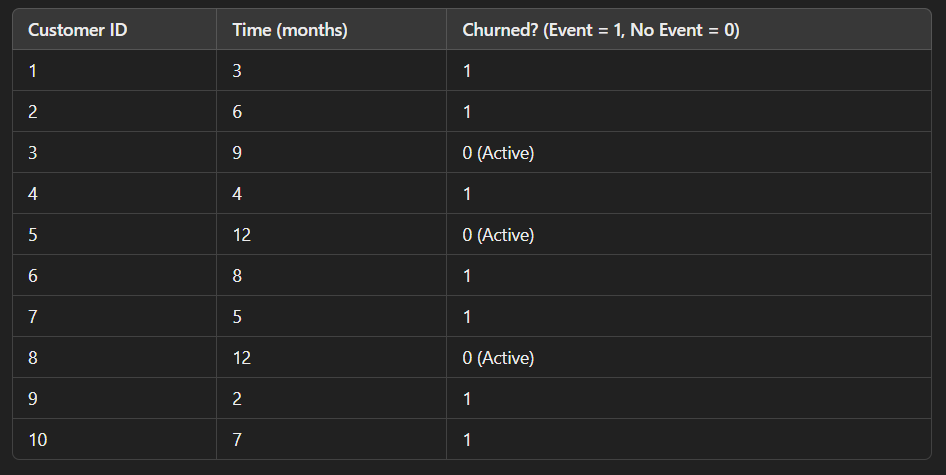
**Scenario**

A streaming service wants to analyse **how long customers stay subscribed** after signing up.

📌 **Data Collected:**

* Customers are **monitored for 12 months**.
* Some customers **churn (cancel subscription)**, while others **remain active** at the end.
* Some customers are **lost to follow-up** (e.g., stopped responding but didn’t officially cancel).

**Step 1: Collect Data**

****

**Step 2: Calculate Survival Probability**

Using the Kaplan-Meier estimator, we compute the cumulative survival rate at each time event.



**Step 3: Plot the Kaplan-Meier Curve**

A **Kaplan-Meier survival curve** is plotted with:

* **X-axis** = Time (months)
* **Y-axis** = Survival probability S(t)

📉 The curve **drops at each churn event**, showing how retention declines over time.

**Step 4: Business Insights**

📌 **50% of customers churn by the 6th month.**  
📌 Retention stabilizes after **9 months**, suggesting a core loyal user base.  
📌 The company should **introduce retention strategies (discounts, new features) by Month 3–5 to prevent early churn**.

**Key Takeaways**

✅ **Kaplan-Meier Curve** is a powerful tool for analysing **customer retention, machine reliability, and patient survival**.  
✅ It accounts for **censored data**, making it more accurate for real-world survival analysis.  
✅ Helps businesses **predict when customers will churn** and **implement strategies to improve retention**.

**Cox Proportional-Hazards Model:** Understanding Factors Affecting Survival Time

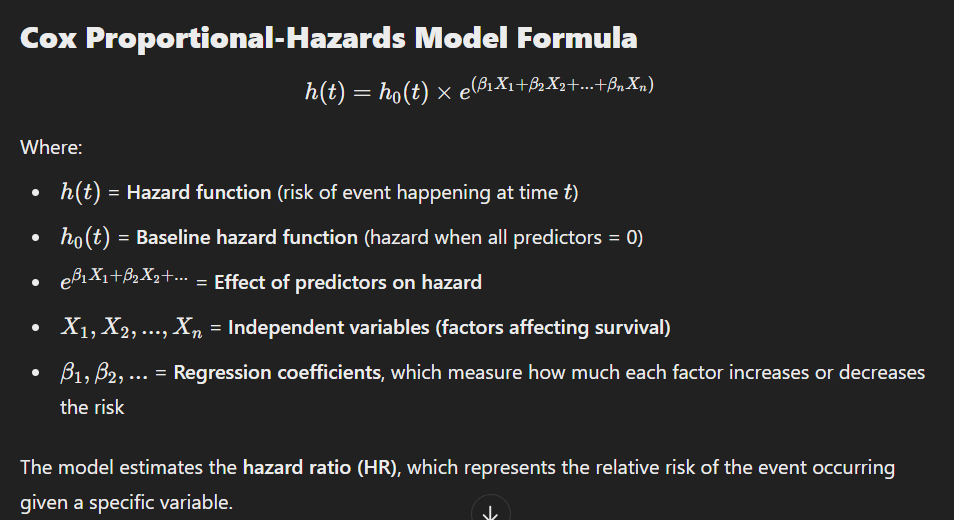
**What is the Cox Proportional-Hazards Model?**

The Cox Proportional-Hazards (Cox PH) model is a semi-parametric survival analysis method used to examine the relationship between multiple predictor variables (features) and the time until an event occurs (survival time).

Unlike the Kaplan-Meier curve, which only describes survival probabilities over time, the Cox PH model can analyse the impact of different factors (e.g., customer demographics, machine usage, patient characteristics) on survival.

**When to Use the Cox Proportional-Hazards Model?**

✅ When analysing time-to-event data (e.g., customer churn, machine failure, patient survival).  
✅ When you want to quantify the impact of multiple factors on survival time.  
✅ When the hazard ratio is assumed to be proportional over time (meaning the effect of a variable remains constant).



**Industrial Example: Customer Churn Prediction for a SaaS Company**

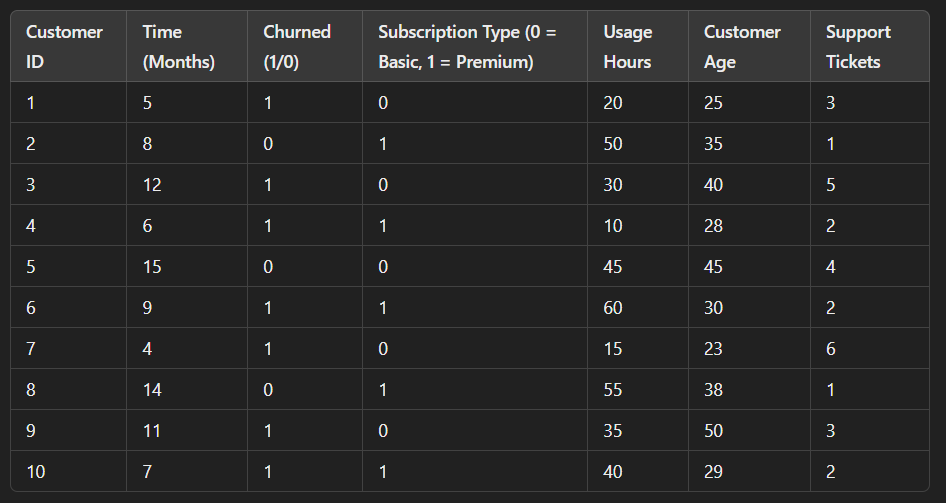
**Scenario**

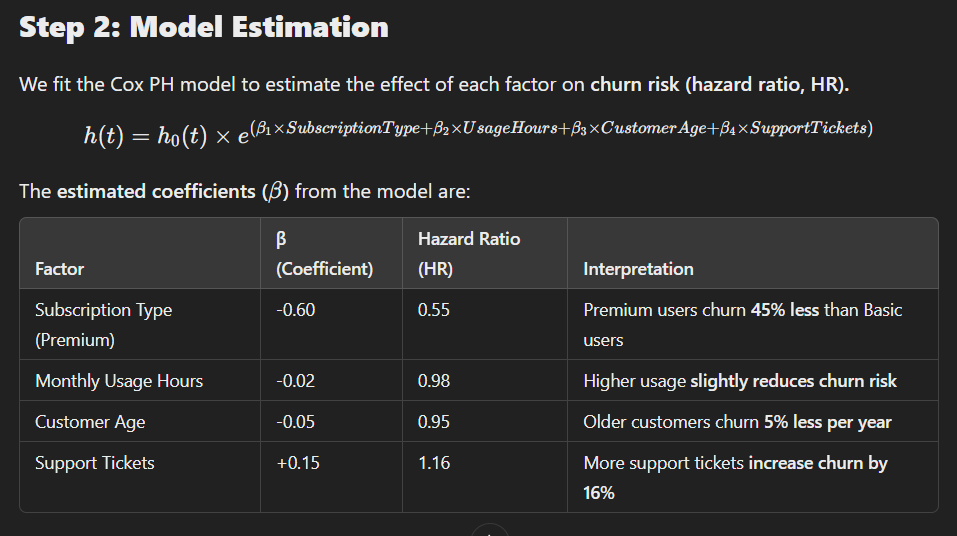
A SaaS (Software as a Service) company wants to understand **which factors influence customer churn** and **when customers are most likely to leave** after signing up.

📌 **Dataset Includes:**

* **Time (months)** before the customer churned or was last observed.
* **Churn (1 = churned, 0 = still active at the end of the study).**
* **Independent variables affecting churn:**
  + **Subscription Type** (Basic, Premium)
  + **Monthly Usage Hours**
  + **Customer Age**
  + **Support Ticket Frequency** (number of times a customer contacted support)

**Step 1: Data Collection**

****

****

**Step 3: Interpret Results**

✔ **Subscription Type (Premium)** → Reduces churn risk by 45% compared to Basic users.  
✔ **Higher Usage Hours** → Customers who use the service more are less likely to churn.  
✔ **Older Customers** → More loyal compared to younger customers.  
✔ **Support Tickets** → More issues lead to higher churn, suggesting customer dissatisfaction.

**Step 4: Business Insights & Recommendations**

📌 Offer incentives for Basic users (discounts, exclusive features) to encourage retention.  
📌 Encourage higher usage through personalized recommendations, reminders, and engagement campaigns.  
📌 Improve customer support to resolve issues faster and reduce churn.  
📌 Focus retention efforts on younger users who tend to leave earlier.

**Key Takeaways**

✅ Cox Proportional-Hazards Model is useful for analysing time-to-event data with multiple predictors.  
✅ It estimates the hazard ratio (HR), which shows how each factor affects survival (or churn, failure, etc.).  
✅ Helps businesses identify key drivers of churn and develop targeted retention strategies.