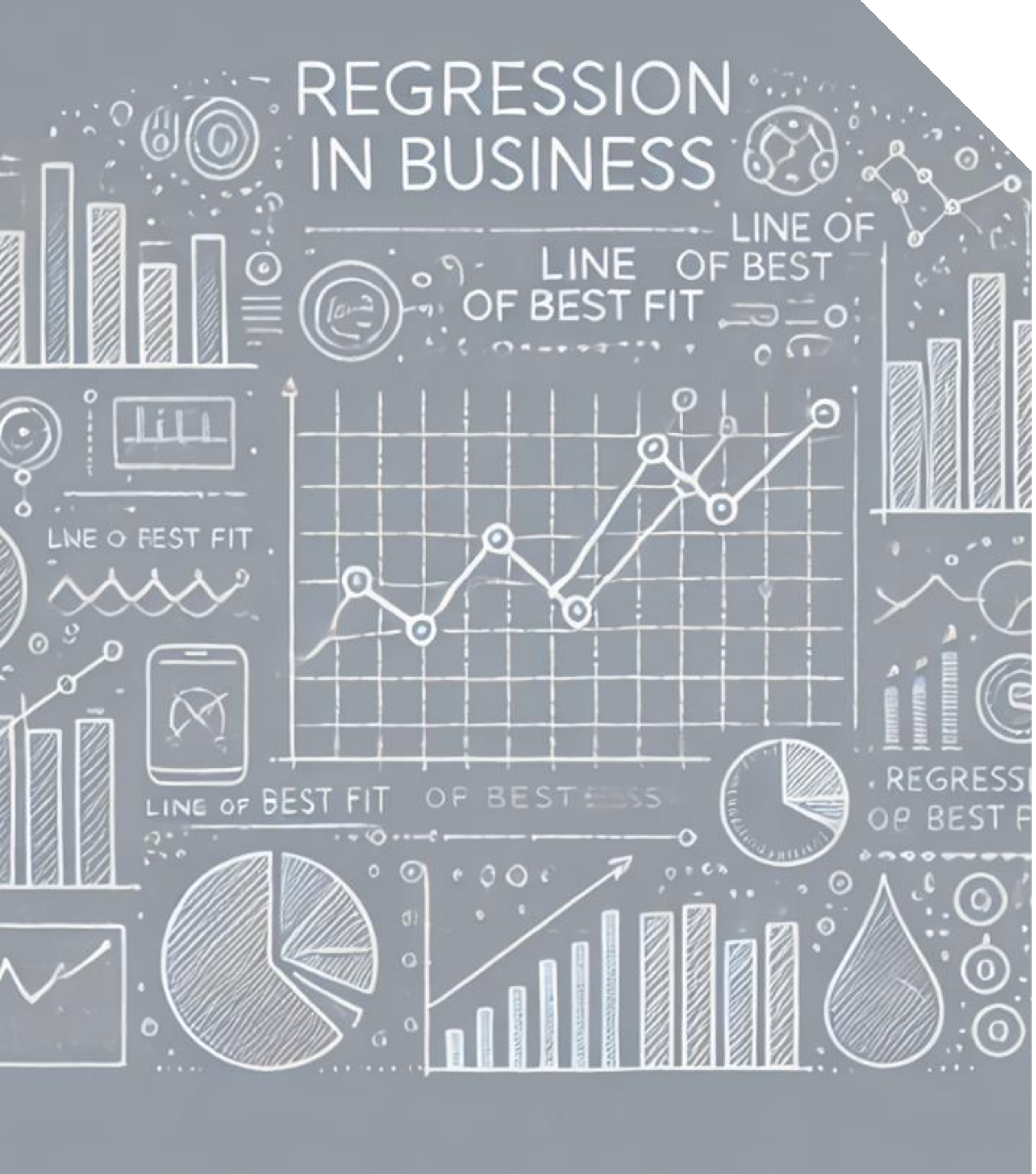
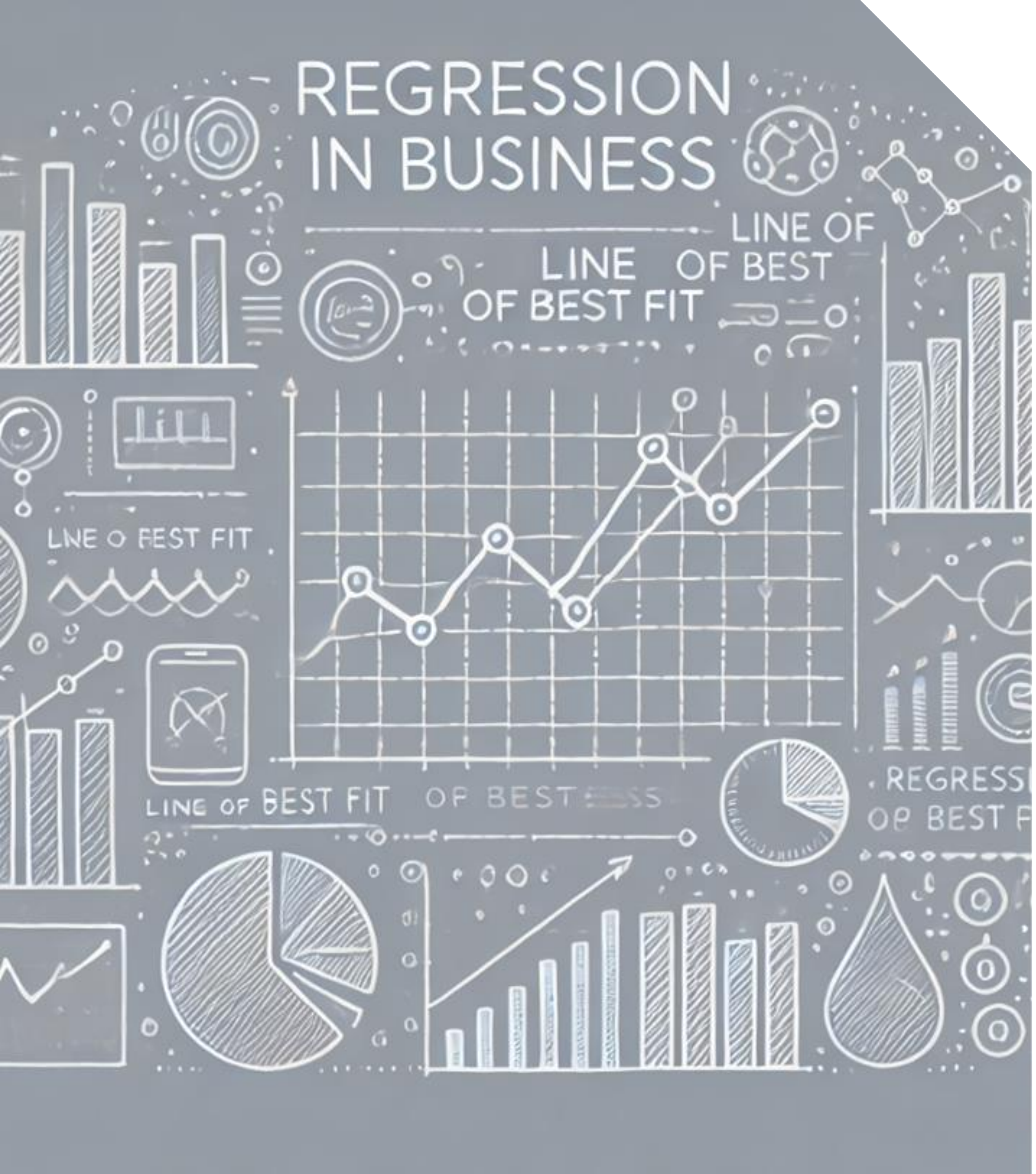


Business use cases and Regression

(Understanding Regression Analysis and it’s interpretation in Business)

Anant Prakash Awasthi

Lead Data Scientist, Optum Global Solutions



Disclaimer

All the points and views in this discussion are my own and no where represents view or policies of my organization.

Agenda

- Greetings & Session Logistics
- Introduction to AI
- Introduction & Types of Machine Learning Methods
- Introduction & History of Regression
- Machine Learning Methods for Regression Problems
- Regression Problem - Use Cases
- A quick case study of a regression problem
- Developing an ML Solution using Linear Regression
- Case Study

Session Logistics

- Session Microsite
- Programming Environment
  - On Machine Setup
    - Visual Studio Code
    - Anaconda
  - Cloud Setup
    - Google Colab
    - Kaggle Notebooks
    - AWS Sagemaker Notebooks
  - Support Software
    - Git
    - GitHub CLI
    - Sublime Editor
- Accounts
  - Google Account | Kaggle Account | GitHub Account

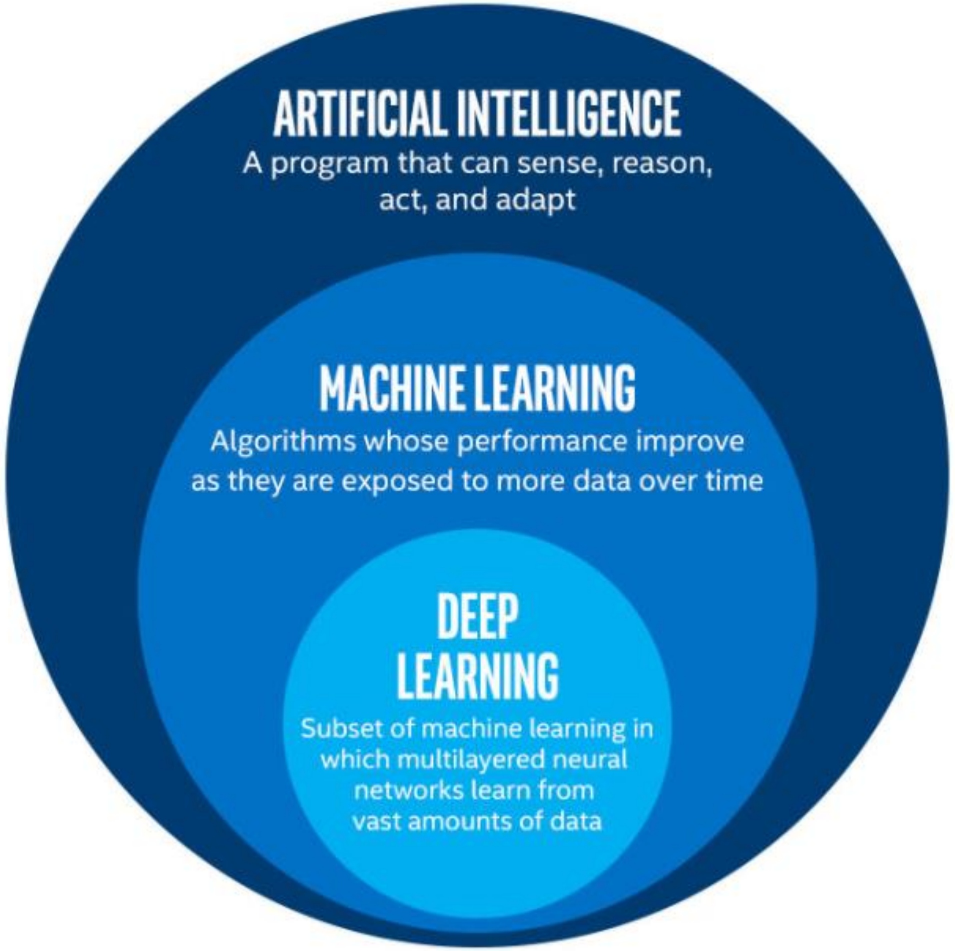
Introduction to AI

Artificial Intelligence (AI) refers to the ability of machines or computers to perform tasks that typically require human intelligence.

This includes things like recognizing patterns, solving problems, learning from experience, and making decisions.

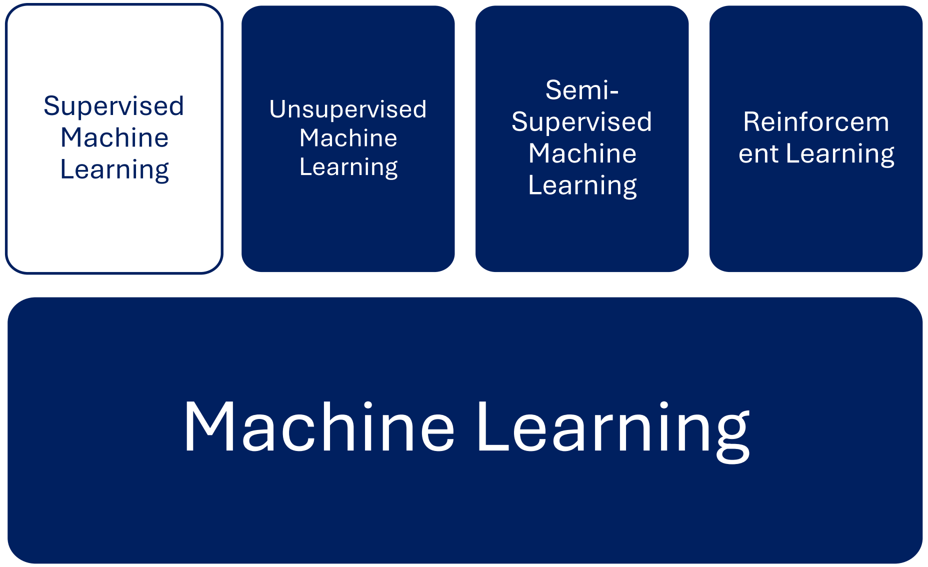
AI enables machines to understand and respond to their environment in ways that mimic human thinking, often helping businesses and individuals complete tasks more efficiently and accurately.

Introduction to Machine Learning Methods


ARTIFICIAL INTELLIGENCE
A program that can sense, reason, act, and adapt
MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time
DEEP LEARNING
Subset of machine learning in which multilayered neural networks learn from vast amounts of data

Source: https://miro.medium.com/v2/resize:fit:631/0Q3PICBlib-932hhH.png

Types of Machine Learning



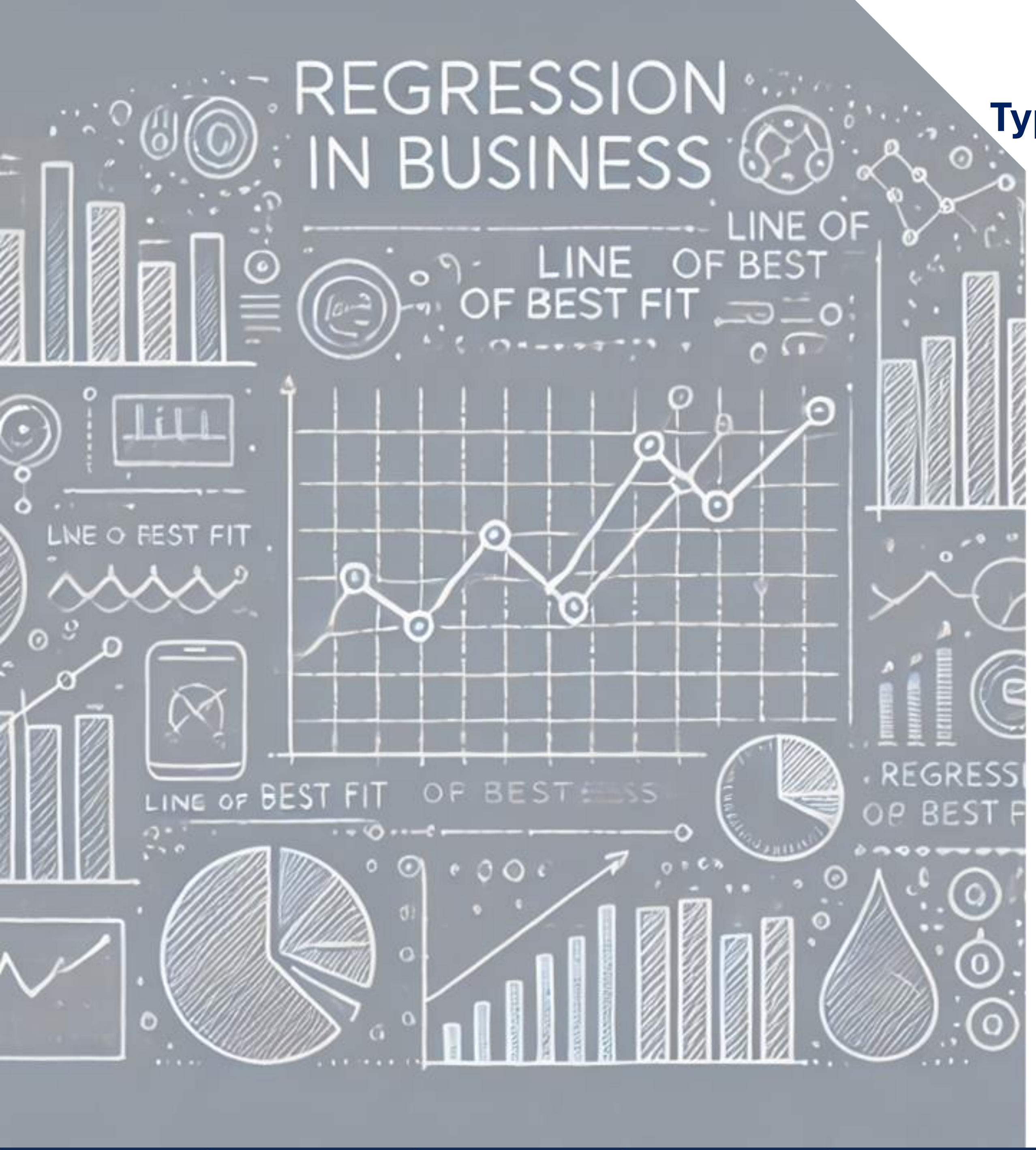
Supervised Machine Learning

Unsupervised Machine Learning

Semi-Supervised Machine Learning

Reinforcement Learning

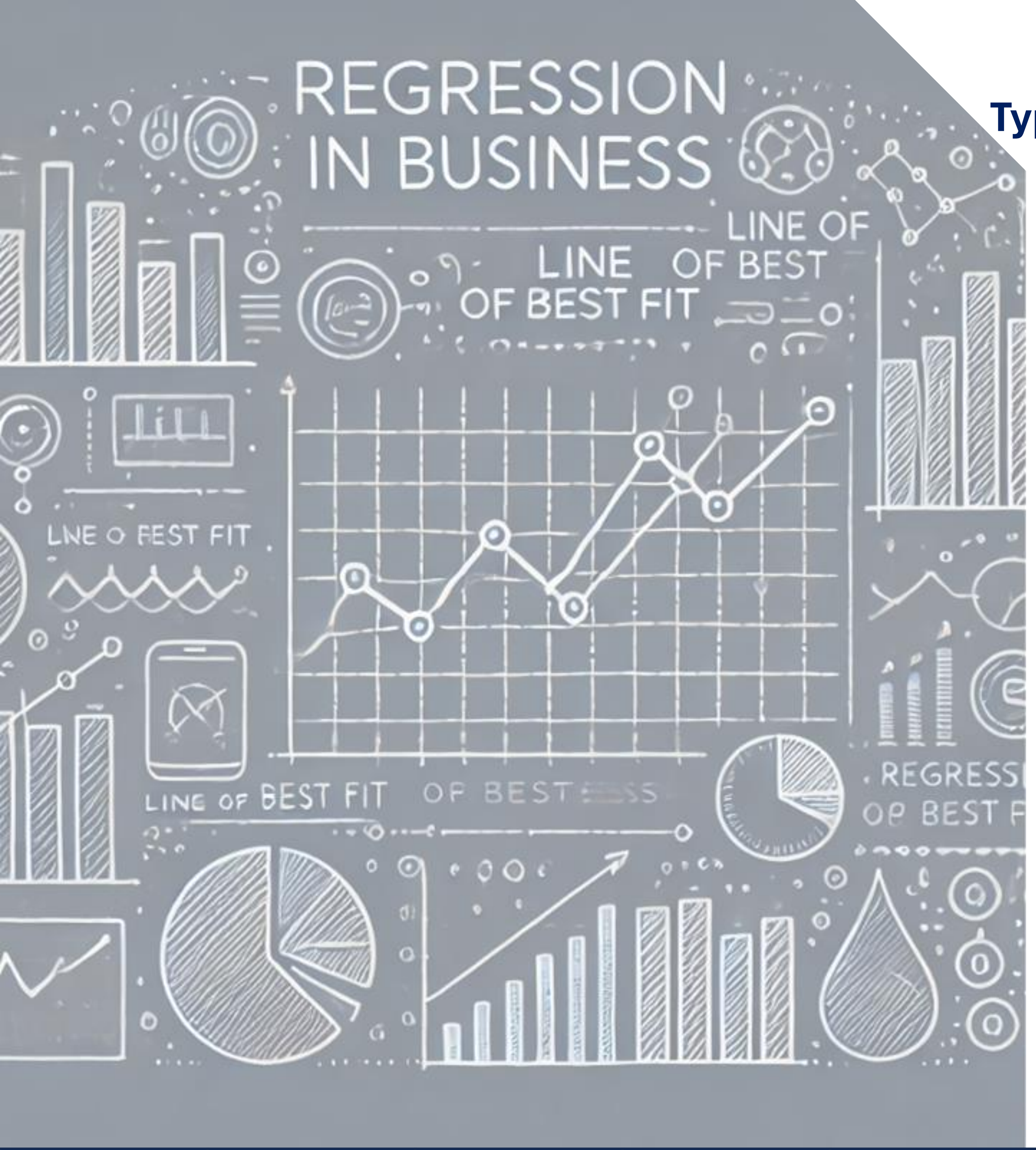
Machine Learning



Types of Problems in Supervised machine Learning

In supervised machine learning, there are two main types of problems:

- Regression Problems
  - What it is: Regression is used when the goal is to predict a continuous outcome or numerical value.
  - Example: Predicting house prices based on features like size, location, and number of rooms.
  - Business Use Cases: Sales forecasting, stock price prediction, demand prediction

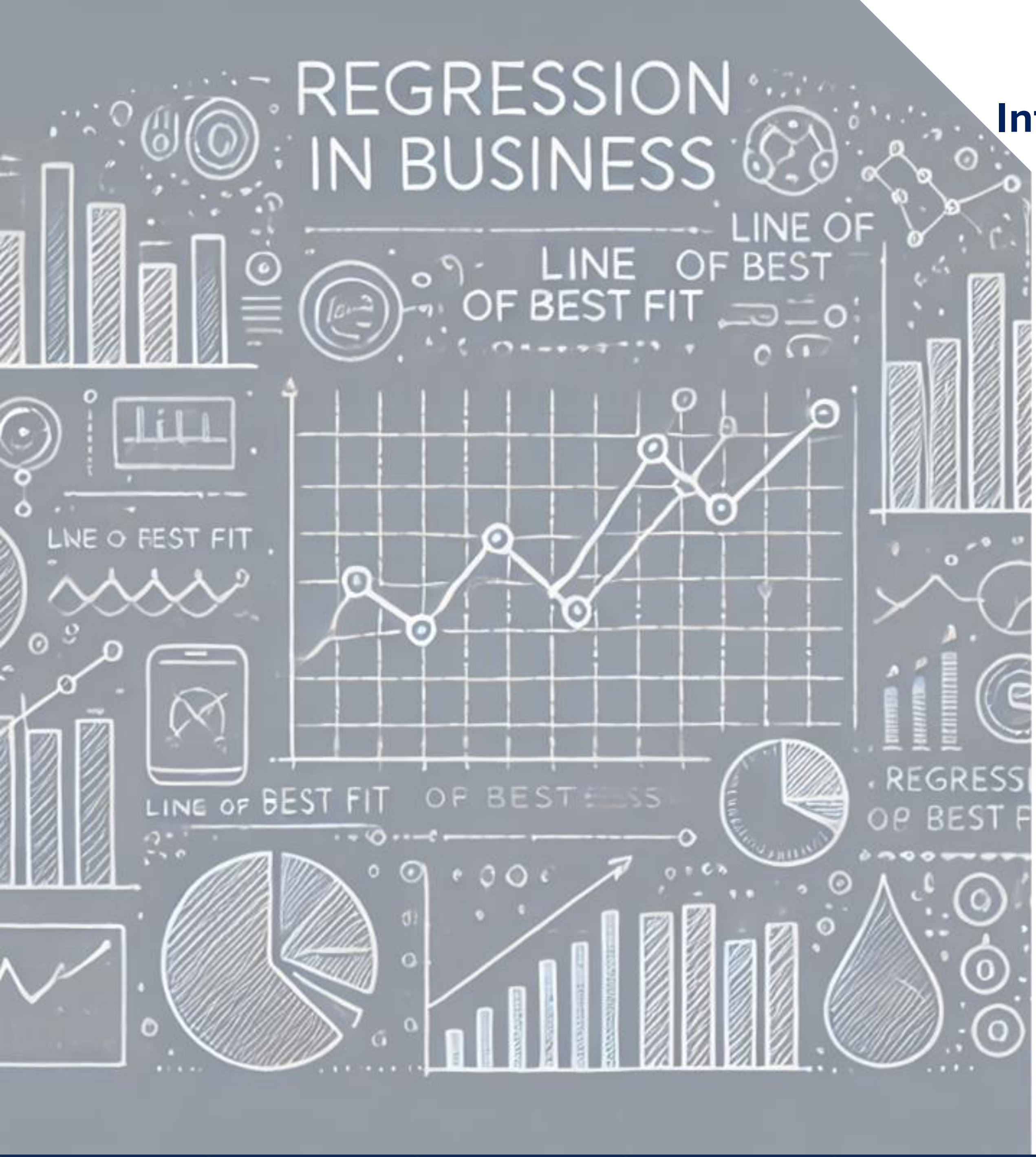


Types of Problems in Supervised machine Learning

In supervised machine learning, there are two main types of problems:

- Classification Problems
  - What it is: Classification is used when the goal is to predict discrete categories or classes.
  - Example: Identifying whether an email is spam or not (binary classification), or categorizing a customer as high, medium, or low value (multi-class classification).
  - Business Use Cases: Customer segmentation, fraud detection, sentiment analysis

NO MATH
ON SATURDAYS
SATURDAY

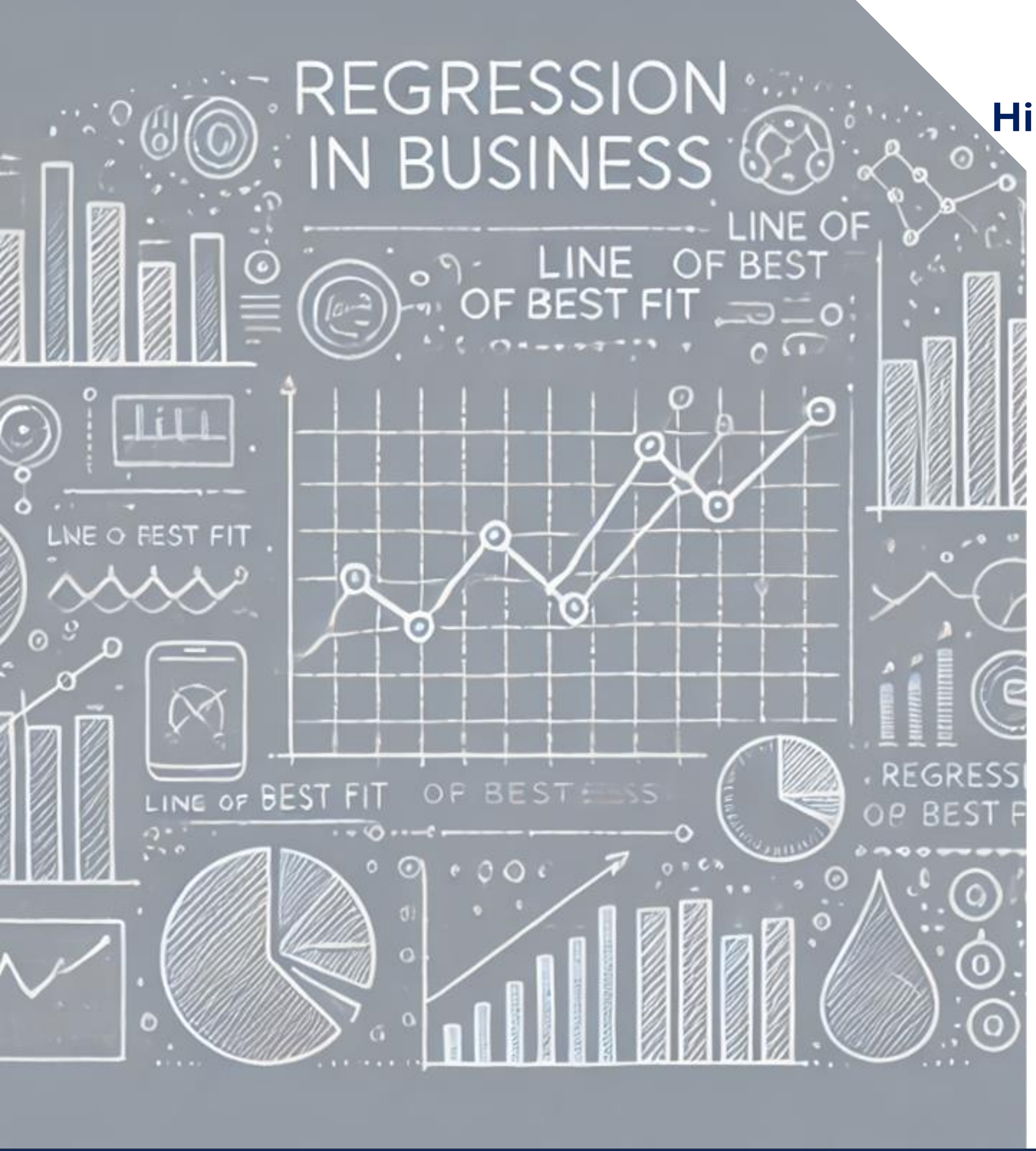


Introduction to Regression

Regression is a statistical method used to understand the relationship between a dependent variable (the outcome we are interested in) and one or more independent variables (factors that may influence the outcome).

It helps in predicting future outcomes based on past data. For example, businesses often use regression to predict sales based on factors like marketing spend or seasonality, enabling them to make informed decisions.

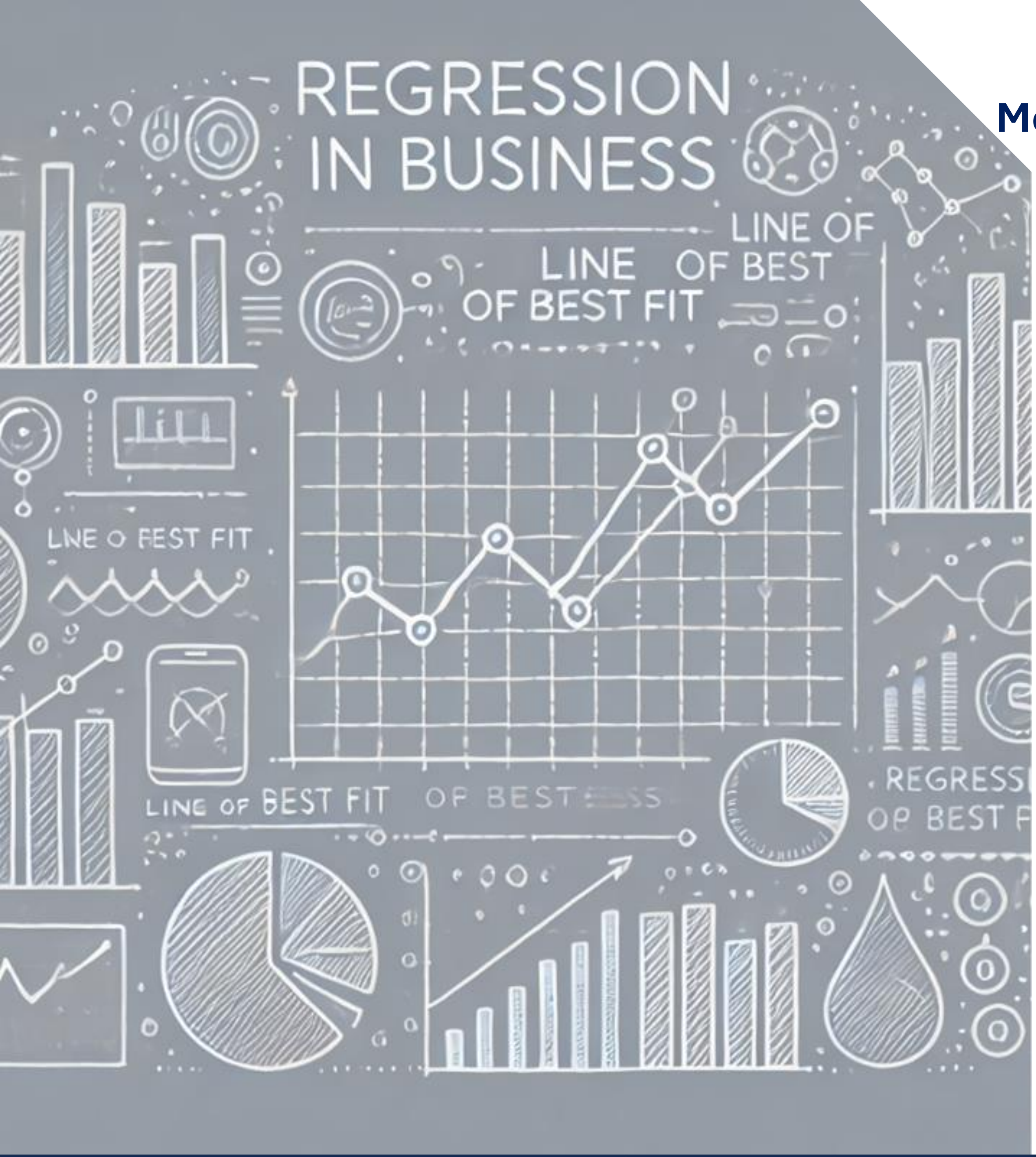
Dependent Variables are Called Y and Independent Variables are called X in Machine learning.



History of Regression

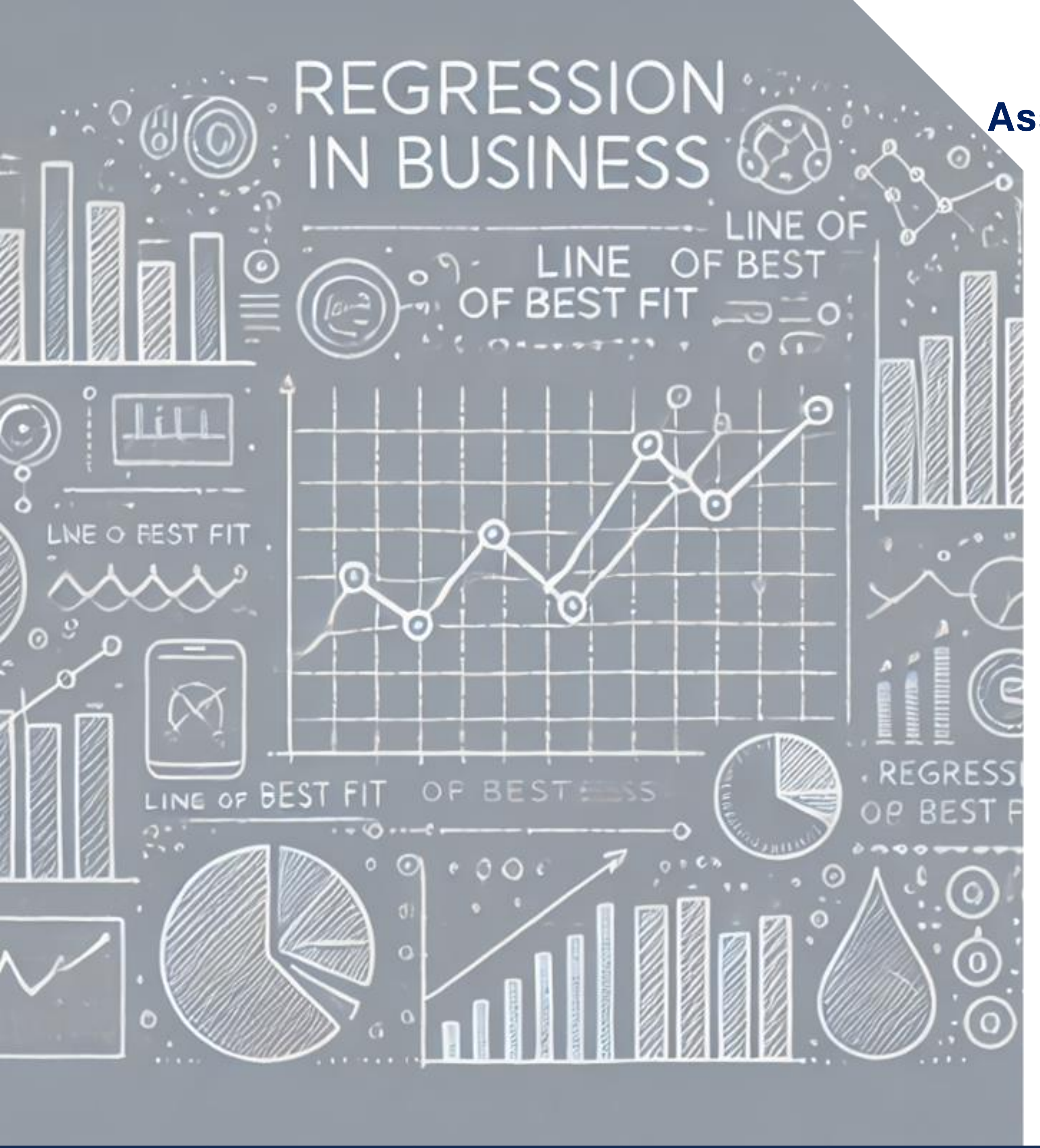
The history of regression dates back to the late 19th century, when British statistician Sir Francis Galton introduced the concept while studying the relationship between the heights of parents and their children. He coined the term "regression to the mean" after observing that extreme traits (such as very tall parents) tend to produce offspring closer to the average height.

Galton's work laid the foundation for the development of modern regression analysis. His work was further advanced by Karl Pearson, who formalized the method with mathematical principles. Regression analysis has since become a fundamental tool in statistics, widely used across disciplines such as economics, biology, and business to predict outcomes and model relationships between variables.



Methods to solve Regression Problems in ML

- Linear Regression
- Ridge Regression
- Lasso Regression
- Polynomial Regression
- Decision Tree Regression
- Random Forest Regression
- Support Vector Regression (SVR)
- K-Nearest Neighbors (KNN) Regression
- Gradient Boosting Regression
- XGBoost Regression
- ElasticNet Regression



Assumptions of Linear Regression

Linearity:

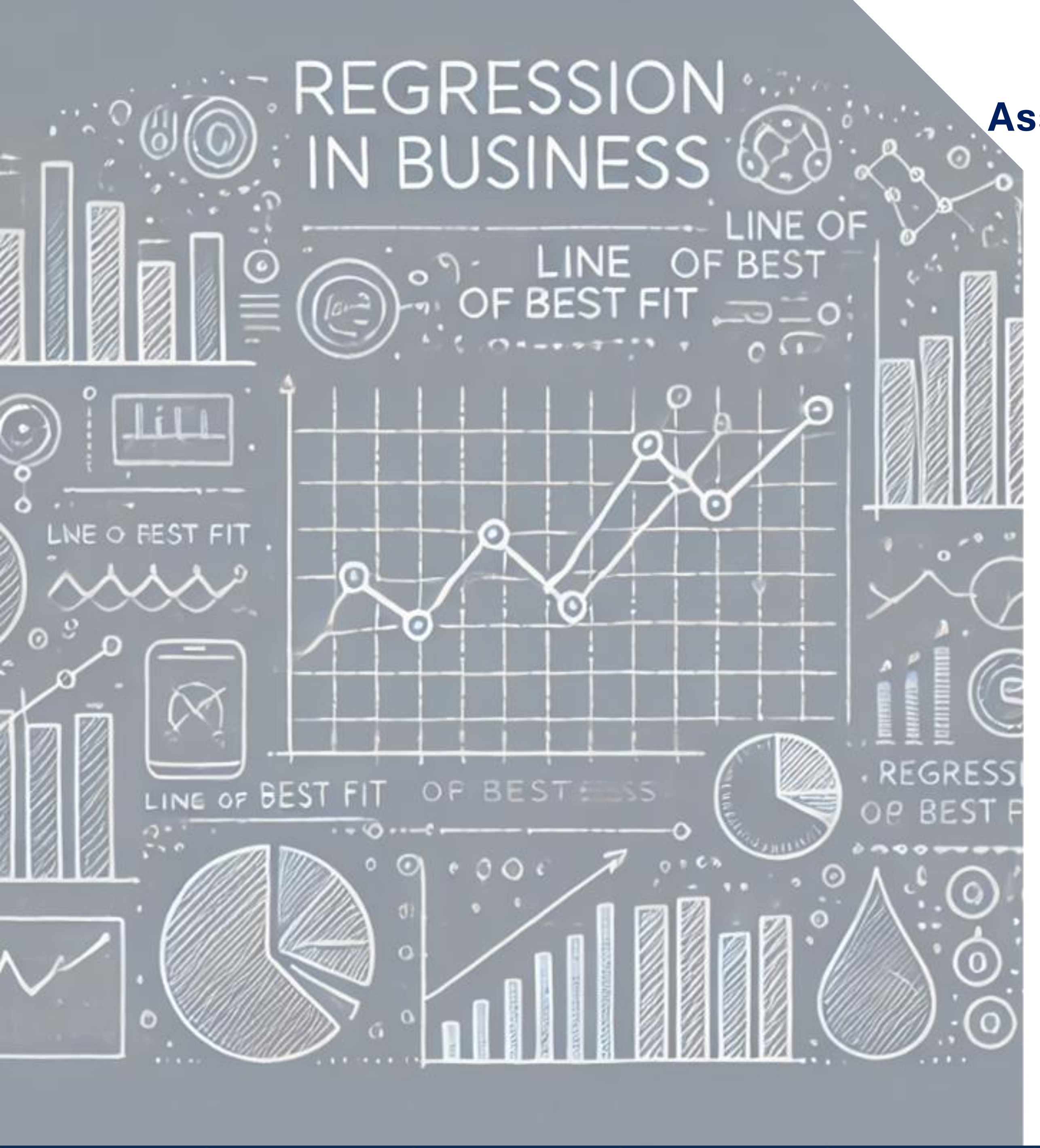
Assumption: The relationship between the dependent variable and the independent variables is linear.

Validation: Plot the residuals (errors) versus the fitted values. If the plot shows no clear pattern and the residuals are randomly scattered, this supports linearity. Additionally, you can use a scatterplot of the independent variables against the dependent variable to visually inspect linear relationships.

Independence of Errors:

Assumption: The residuals (errors) are independent of each other.

Validation: For time series data, use the Durbin-Watson test to check for autocorrelation. For other types of data, you can plot residuals in the order of their occurrence (or sort by some other criterion) and check for patterns.



Assumptions of Linear Regression

Homoscedasticity:

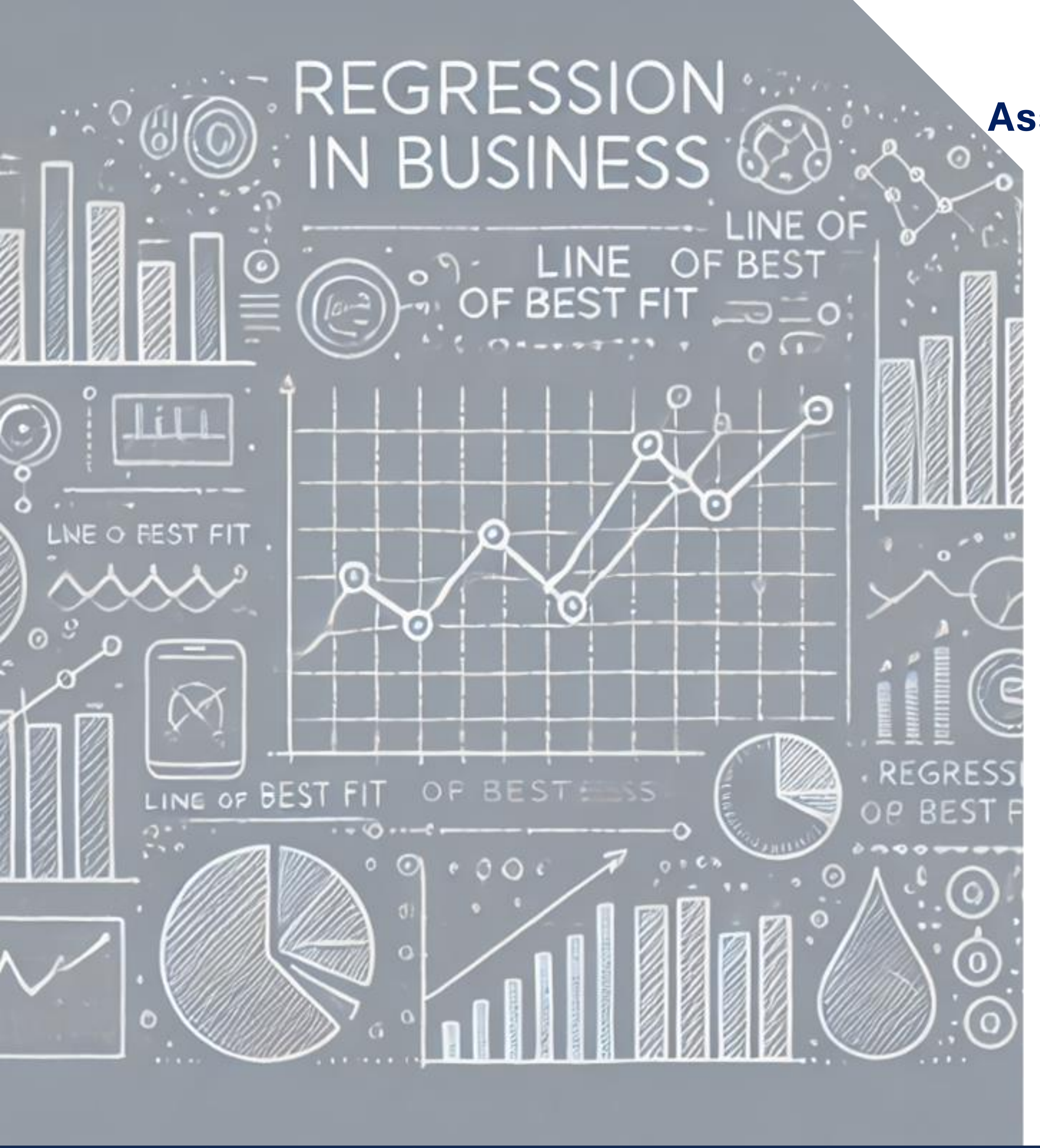
Assumption: The residuals have constant variance at all levels of the independent variables.

Validation: Plot the residuals versus the fitted values. If the spread of the residuals increases or decreases systematically with the fitted values, this may indicate heteroscedasticity. You can also use statistical tests like Breusch-Pagan or White's test to formally assess homoscedasticity.

Normality of Errors:

Assumption: The residuals are normally distributed.

Validation: Examine a Q-Q (quantile-quantile) plot of the residuals, which should ideally follow a straight line if the residuals are normally distributed. Alternatively, use the Shapiro-Wilk test or Kolmogorov-Smirnov test to statistically test for normality.



Assumptions of Linear Regression

No Multicollinearity:

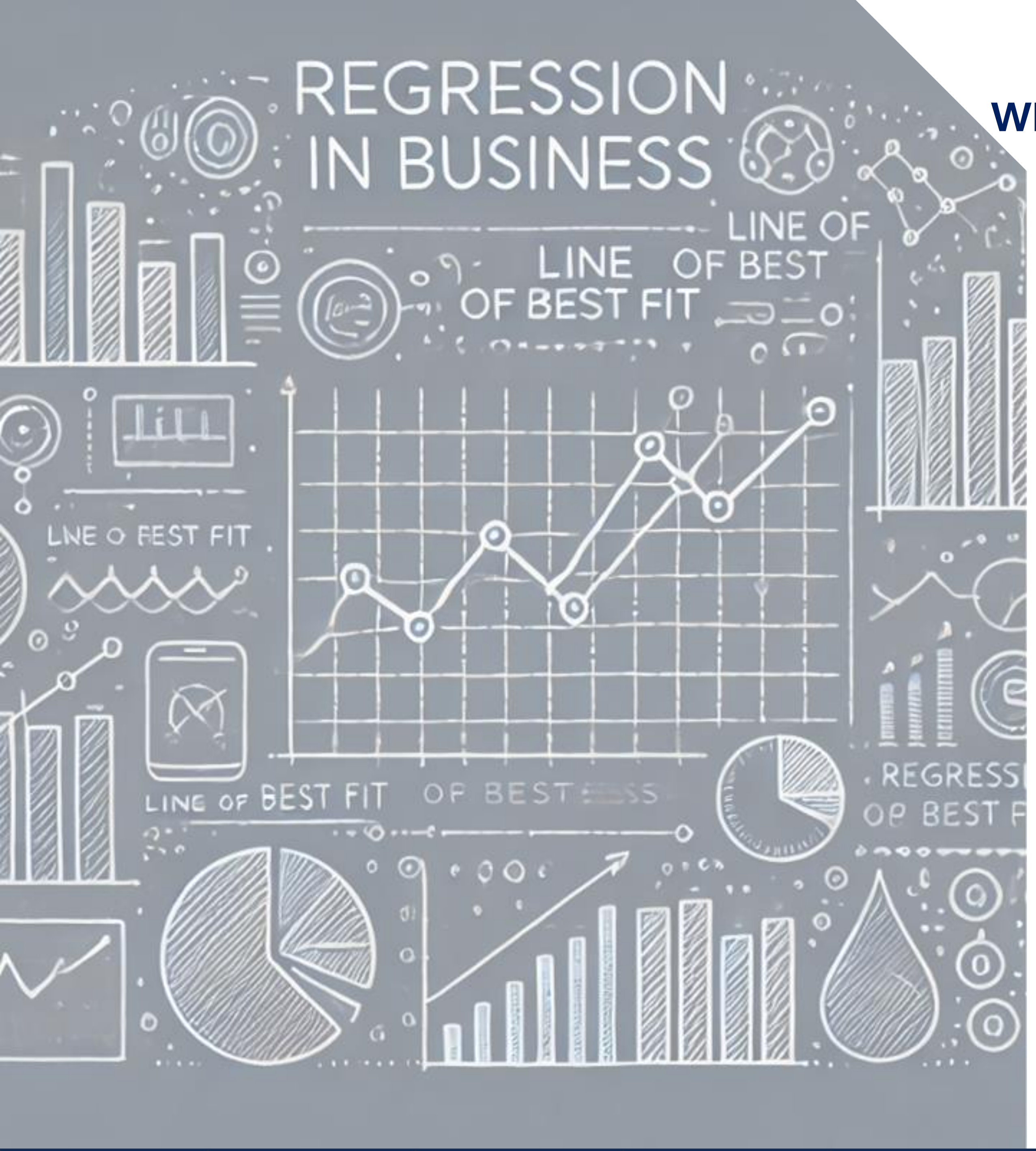
Assumption: The independent variables are not highly correlated with each other.

Validation: Calculate the Variance Inflation Factor (VIF) for each independent variable. A VIF value greater than 10 may indicate problematic multicollinearity. Additionally, check the correlation matrix of the independent variables for high correlations.

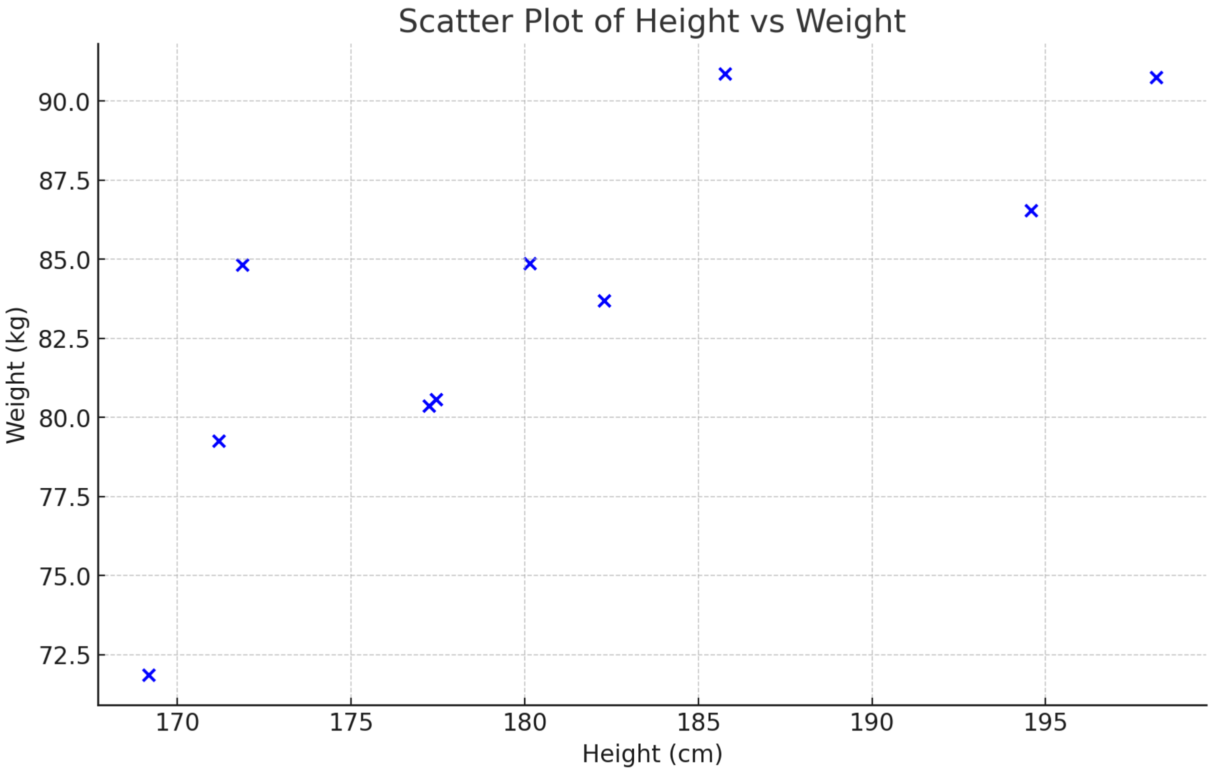
No Endogeneity:

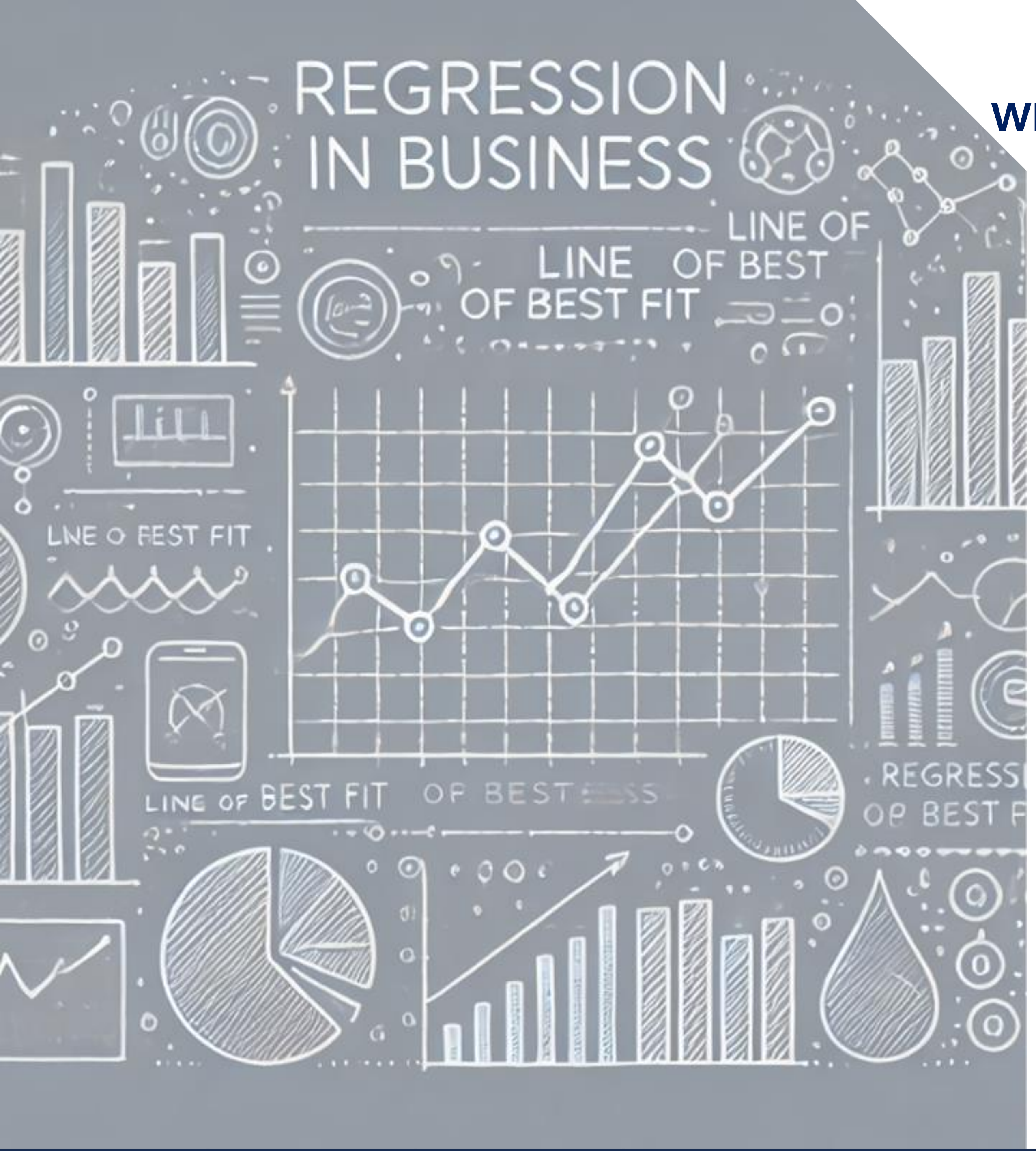
Assumption: The independent variables are not correlated with the error term.

Validation: This can be more complex to test directly, but you can check if there’s a need for instrumental variables or if there are omitted variables that might be influencing the results.

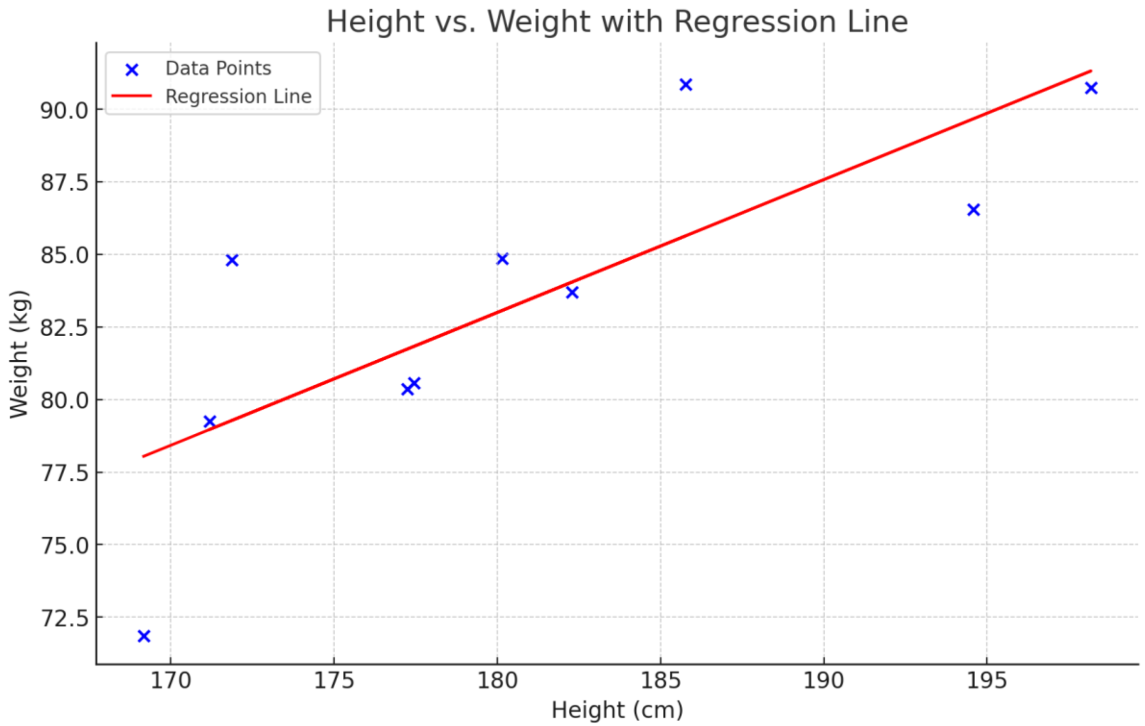


What is Linear regression?





What is Linear regression?



Weight (kg) = 0.46*Height (cm) + 0.64

(Y = β*x + Intercept + Some Error)

Use Cases of Regression

1. Sales Forecasting: Predict future sales based on factors like past sales data, advertising spend, and market conditions.

2. Price Prediction: Estimate the price of commodities, housing, or stocks based on factors like size, location, and market trends.

3. Customer Lifetime Value (CLV): Predict a customer's future spending and lifetime value based on past purchasing behavior.

4. Demand Forecasting: Estimate product demand by analyzing past sales, seasonality, and economic factors.

5. Risk Assessment in Finance: Predict credit scores and the likelihood of loan defaults based on income, age, and other factors.

Use Cases of Regression

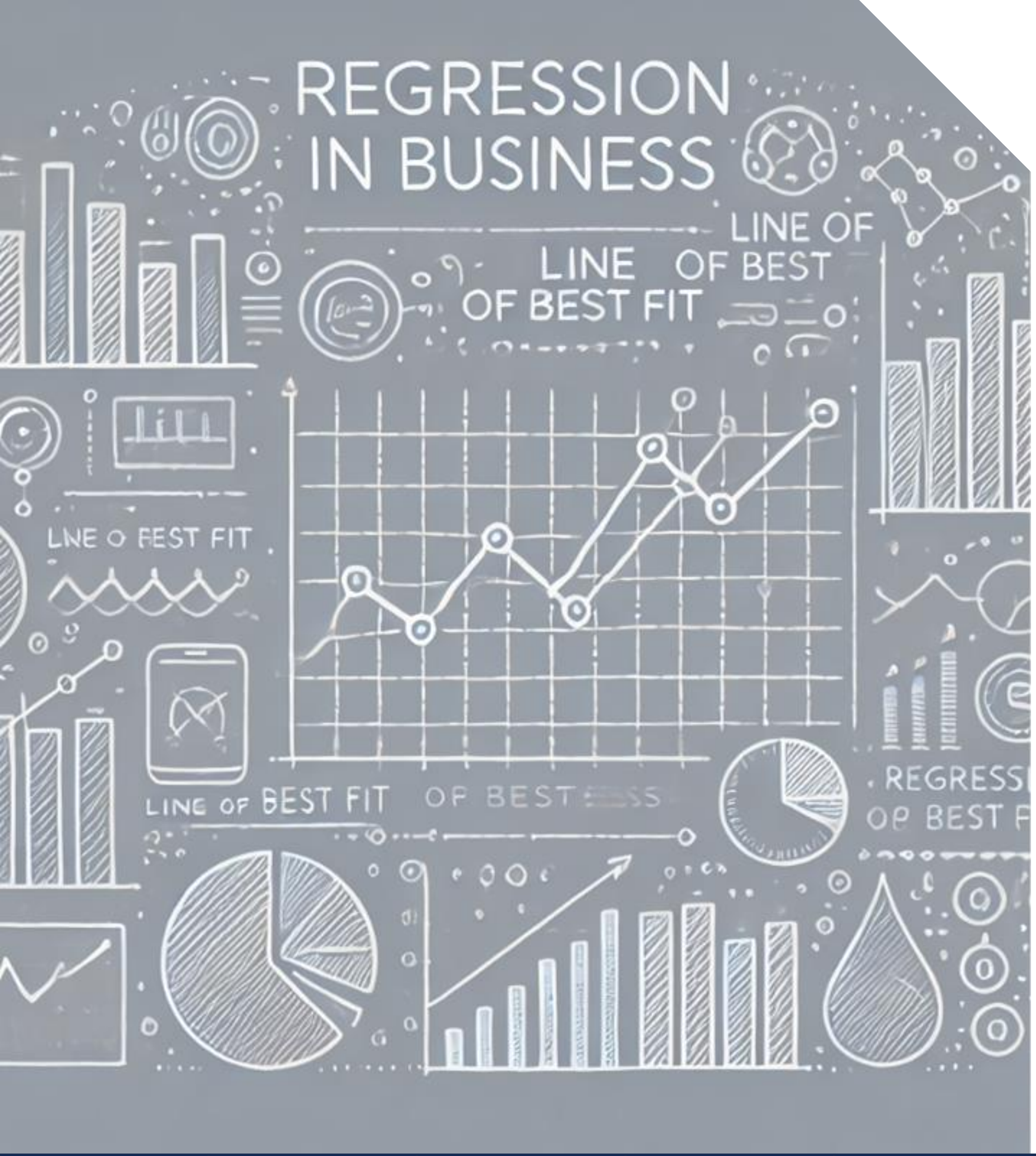
6. Health Care Cost Prediction: Estimate future medical expenses based on patient health data, age, and historical costs.

7. Employee Salary Prediction: Determine salary trends by analyzing experience, education, and other job-related variables.

8. Energy Consumption Forecasting: Predict future energy usage based on factors such as weather conditions and historical consumption patterns.

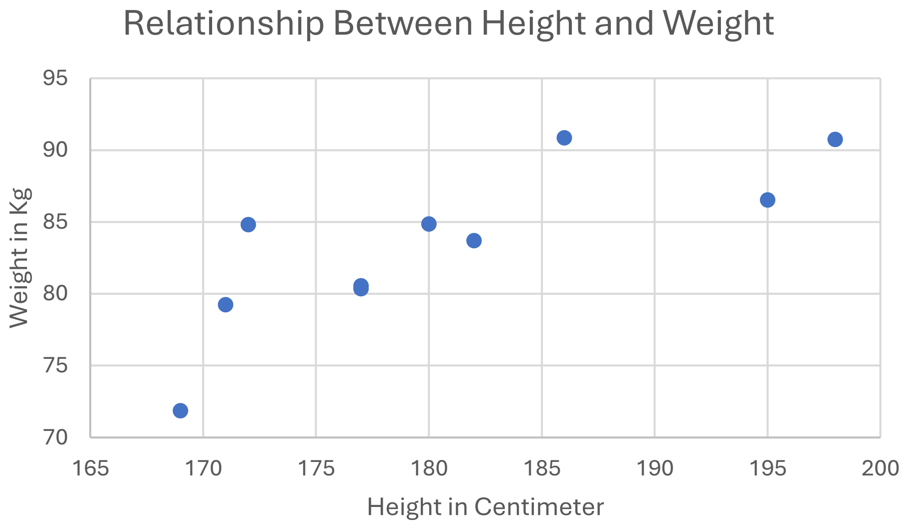
9. Advertising Impact: Analyze the relationship between marketing spend and customer acquisition or revenue growth.

10. Manufacturing Process Optimization: Predict the yield of a production process based on input factors such as temperature, pressure, and time.



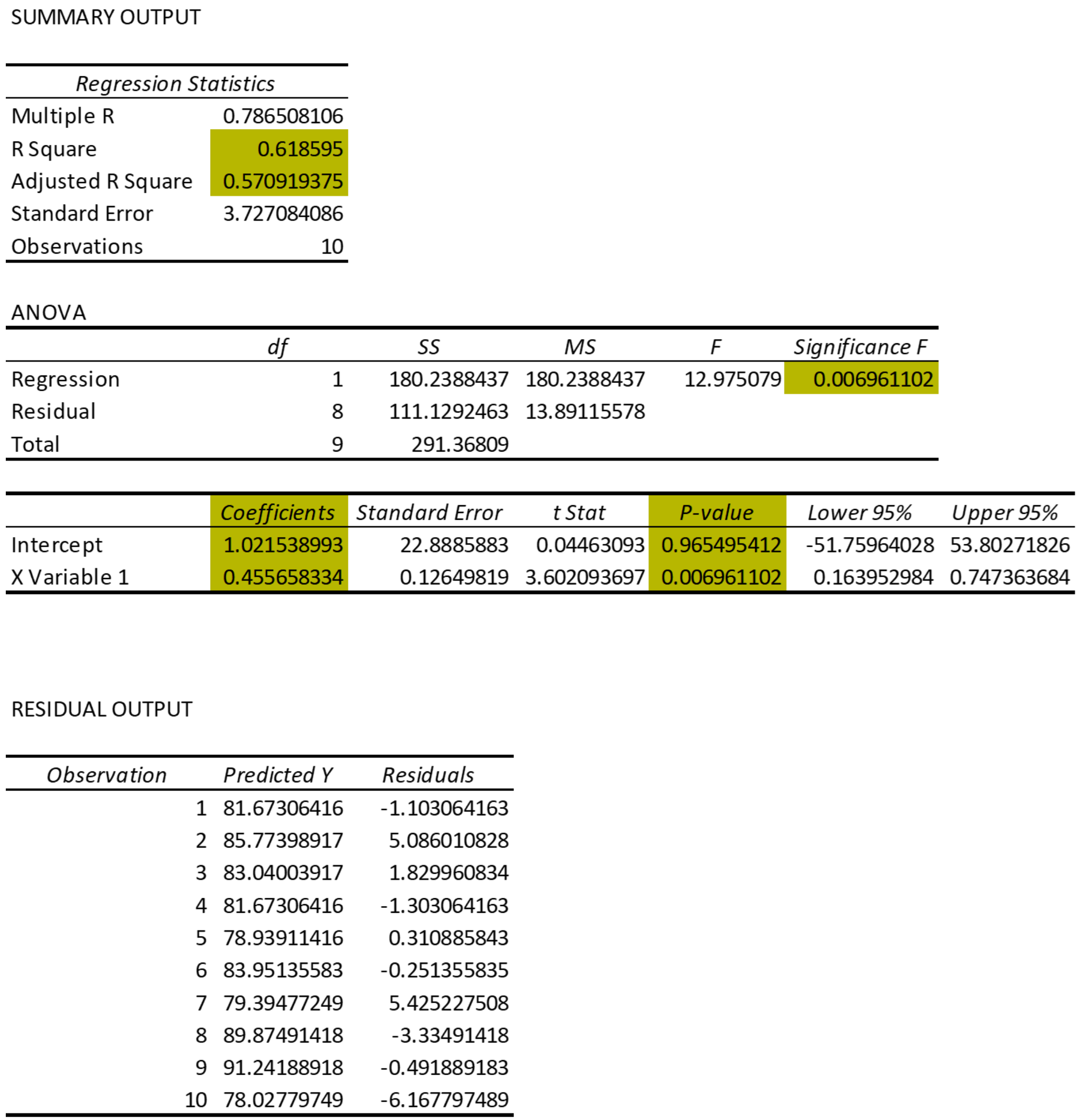
Fit your first Regression Model

Height (cm)	Weight (kg)
177	80.57
186	90.86
180	84.87
177	80.37
171	79.25
182	83.7
172	84.82
195	86.54
198	90.75
169	71.86


Relationship Between Height and Weight

Let’s go to Excel

How Regression Looks like in Excel



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.786508106
R Square	0.618595
Adjusted R Square	0.570919375
Standard Error	3.727084086
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	180.2388437	180.2388437	12.975079	0.006961102
Residual	8	111.1292463	13.89115578		
Total	9	291.36809			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.021538993	22.8885883	0.04463093	0.965495412	-51.75964028	53.80271826
X Variable 1	0.455658334	0.12649819	3.602093697	0.006961102	0.163952984	0.747363684

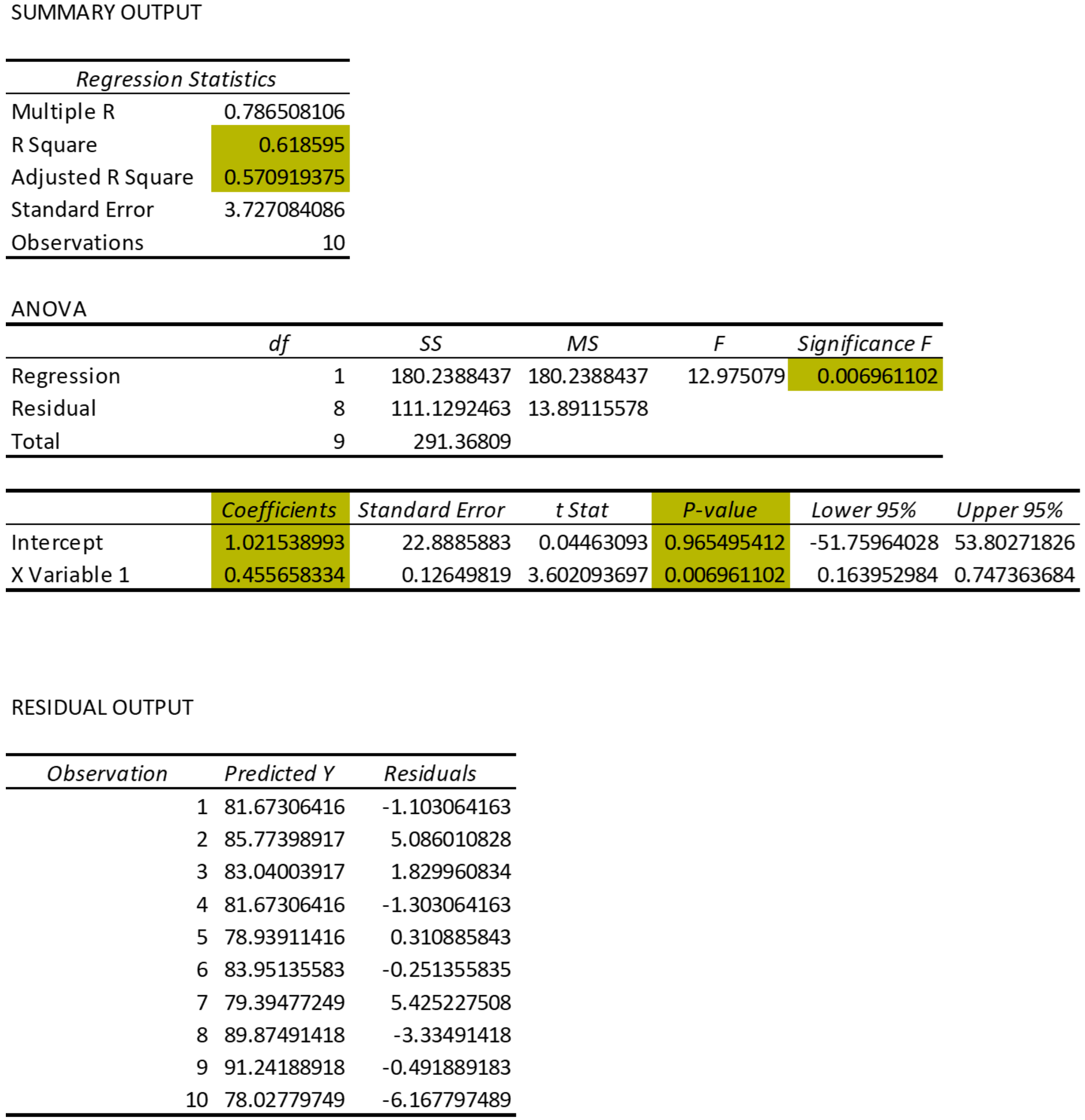
RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*
1	81.67306416	-1.103064163
2	85.77398917	5.086010828
3	83.04003917	1.829960834
4	81.67306416	-1.303064163
5	78.93911416	0.310885843
6	83.95135583	-0.251355835
7	79.39477249	5.425227508
8	89.87491418	-3.33491418
9	91.24188918	-0.491889183
10	78.02779749	-6.167797489

R-Squared (R^2)

- R^2 = 0.618595 (Adjusted: 0.570919375)
- Explanation: R-squared indicates the proportion of variance in the dependent variable (the variable you're trying to predict) that is explained by the independent variable(s).An R^2 value of 0.618 means that approximately 61.9% of the variation in the dependent variable is explained by the model.
- The adjusted R^2 is slightly lower (0.571), which accounts for the number of predictors and adjusts for their significance. It is a better indicator when you have multiple predictors.

How Regression Looks like in Excel



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.786508106
R Square	0.618595
Adjusted R Square	0.570919375
Standard Error	3.727084086
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	180.2388437	180.2388437	12.975079	0.006961102
Residual	8	111.1292463	13.89115578		
Total	9	291.36809			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.021538993	22.8885883	0.04463093	0.965495412	-51.75964028	53.80271826
X Variable 1	0.455658334	0.12649819	3.602093697	0.006961102	0.163952984	0.747363684

RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*
1	81.67306416	-1.103064163
2	85.77398917	5.086010828
3	83.04003917	1.829960834
4	81.67306416	-1.303064163
5	78.93911416	0.310885843
6	83.95135583	-0.251355835
7	79.39477249	5.425227508
8	89.87491418	-3.33491418
9	91.24188918	-0.491889183
10	78.02779749	-6.167797489

Coefficients

Intercept (1.021538993): This is the constant term in the regression equation, representing the predicted value when all other independent variables are zero. In this case, when the independent variables are 0, the dependent variable is expected to be around 1.02.

Slope of X(0.455658334): This is the coefficient for the independent variable. It indicates that for every one-unit increase in the independent variable, the dependent variable increases by 0.456.

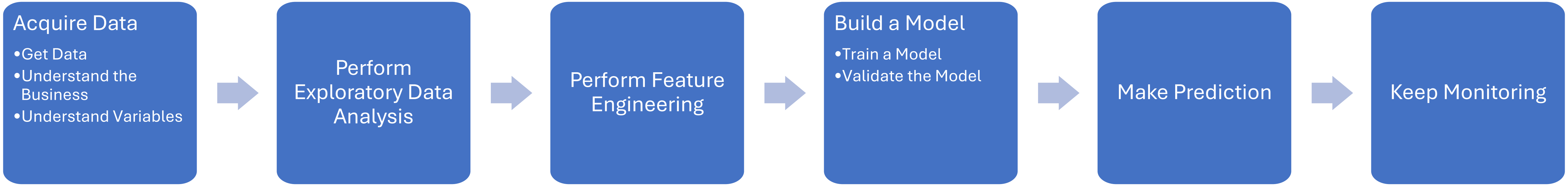
Regression equation

Weight =1.0215+0.4557×Height

Lets Make Predictions

Height (cm)	Weight (kg)
177	
182	
172	
169	

How to Build a model?



Acquire Data

- Get Data
- Understand the Business
- Understand Variables

Perform Exploratory Data Analysis

Perform Feature Engineering

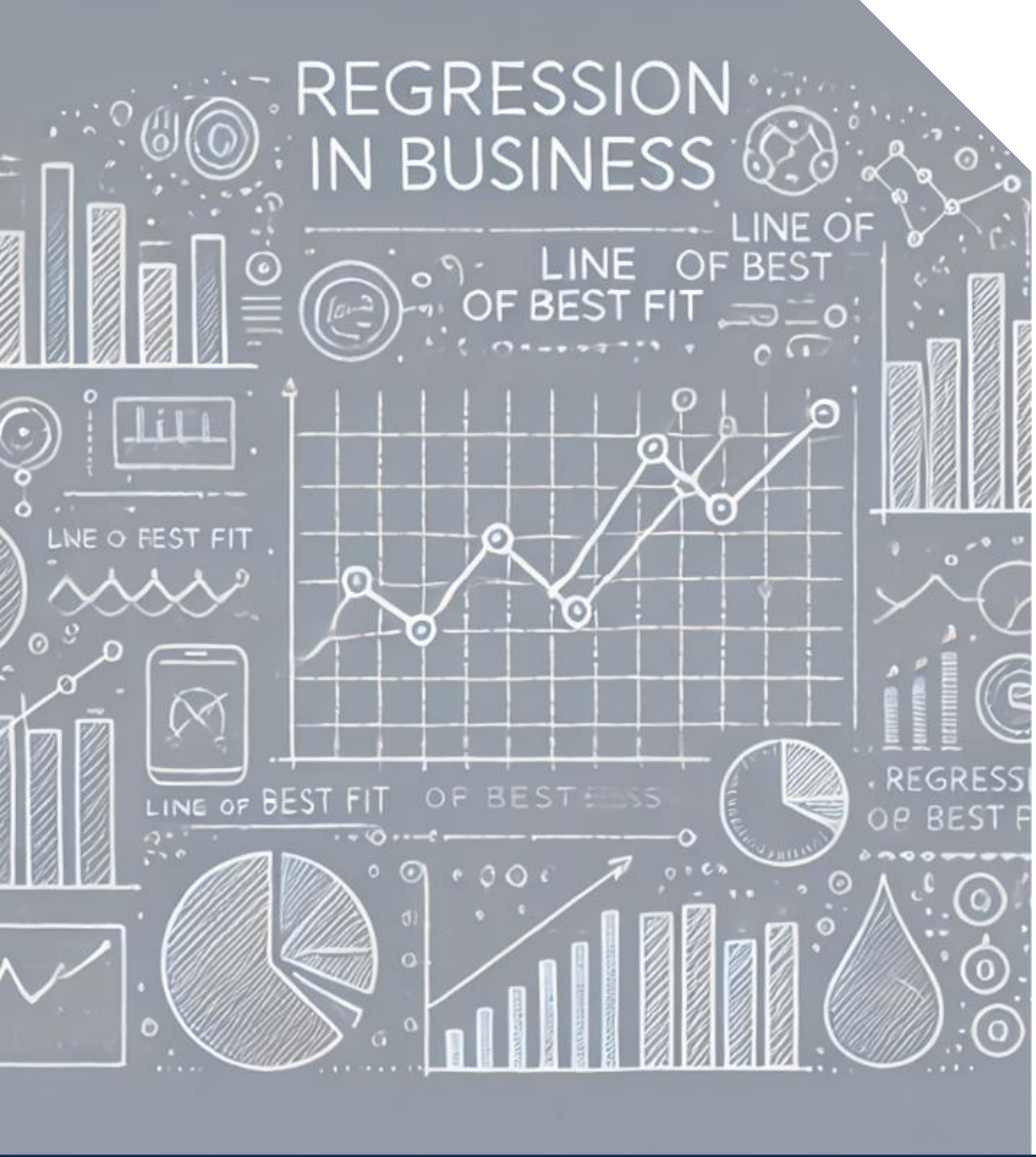
Build a Model

- Train a Model
- Validate the Model

Make Prediction

Keep Monitoring

Let's brew some coffee together



Have a Question?

anant.awasthi@outlook.com

Phone: 8884692929