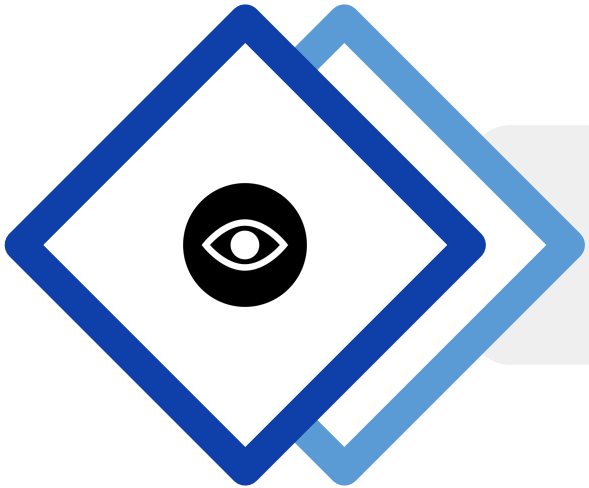
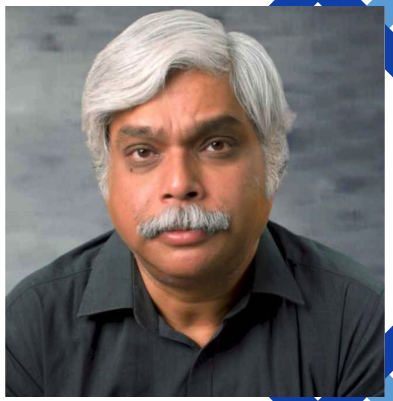


Linear Regression

Meet Your Speaker



Dr. Abhinanda Sarkar
Academic Director at Great Learning

- Alumnus - Indian Statistical Institute, Stanford University
- Faculty - MIT, Indian Institute of Management, Indian Institute of Science
- Experienced in applying probabilistic models, statistical analysis and machine learning to diverse areas
- Certified Master Black Belt in Lean Six Sigma and Design for Six Sigma in GE

Learning Objectives

By the end of this session, you should be able to:

- Relate correlation and simple linear regression in the context of understanding linear relationships.
- Explore simple linear regression models to capture the linear relationship between a pair of attributes.
- Build multiple linear regression to model relationships between two or more input attributes and the output, to predict business outcomes.
- Evaluate linear regression models and identify the levers to improve their performance.
- Discover the applications of linear regression to solve a variety of business problems.

Agenda

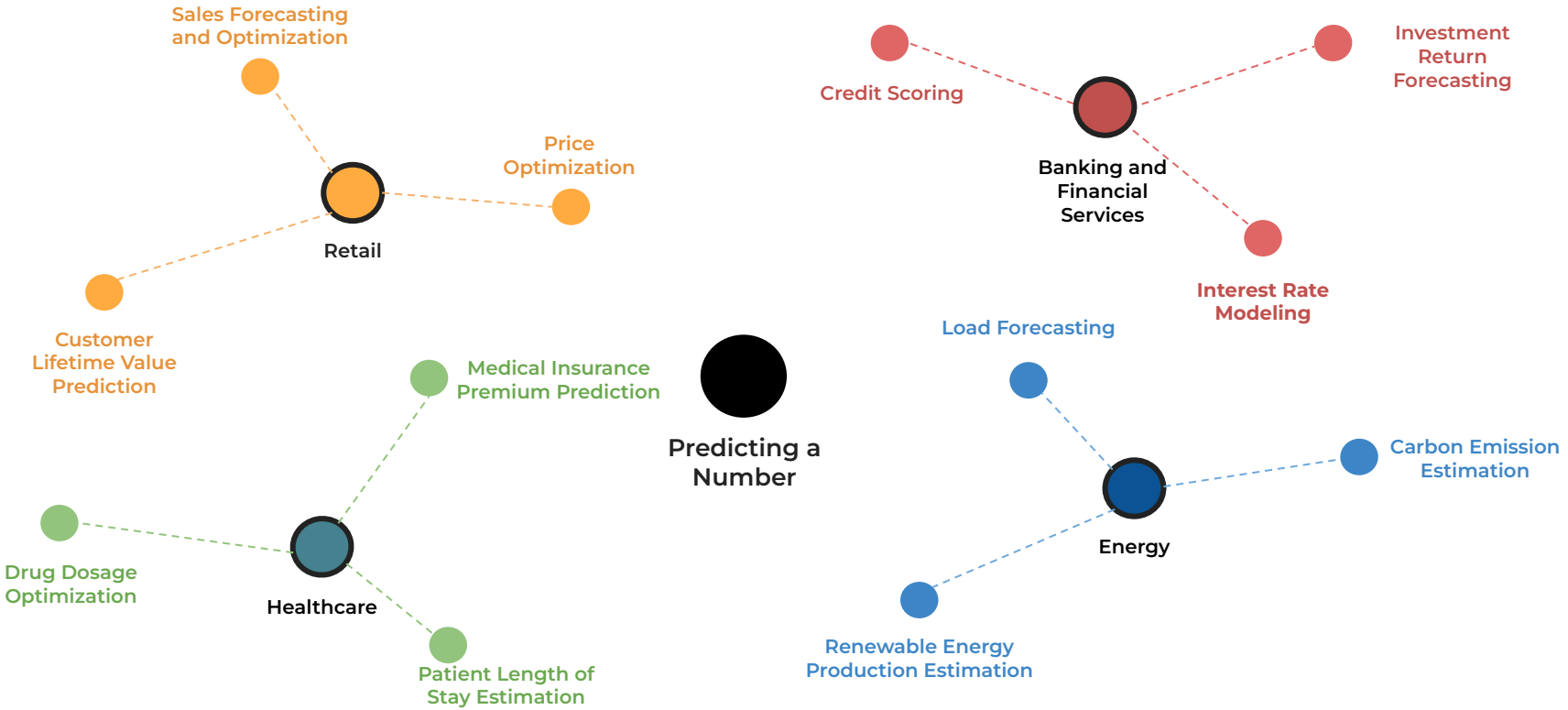
In this session, we'll discuss:

- Business Problem and Solution Space
- Correlation and Linear Relationships
- Simple Linear Regression
- Multiple Linear Regression
- Categorical Variables in Regression
- Evaluation Metrics for Regression

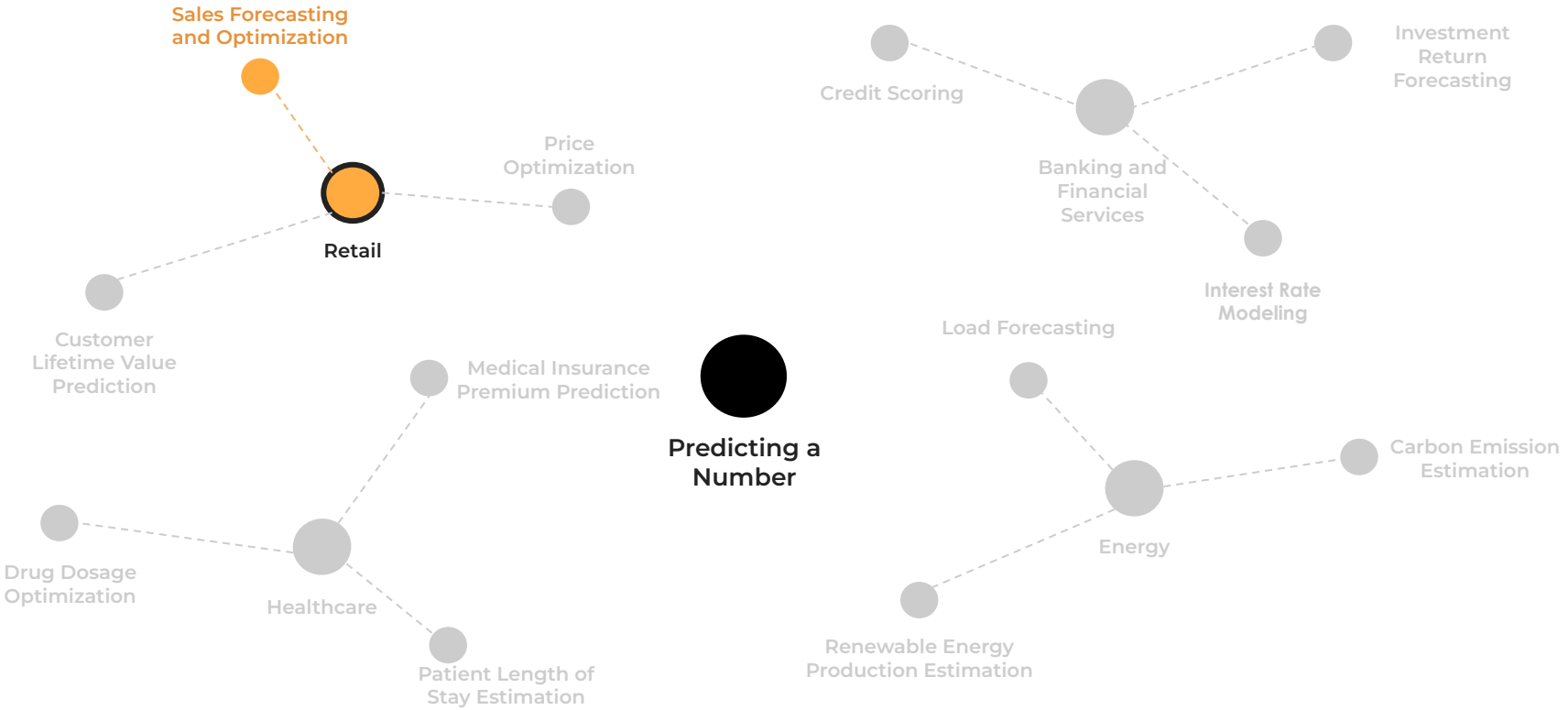
Common Business Questions

- How can we forecast sales based on historical sales data and marketing expenditure?
- How do we determine medical insurance premiums for customers based on attributes like blood pressure, blood sugar level, and smoking habits?
- How do we determine the credit card limit to be assigned to customers based on their past spending behavior, demographic information, etc?
- How can we predict future power load requirements to ensure reliable grid operation and prevent outages?

Problem Space

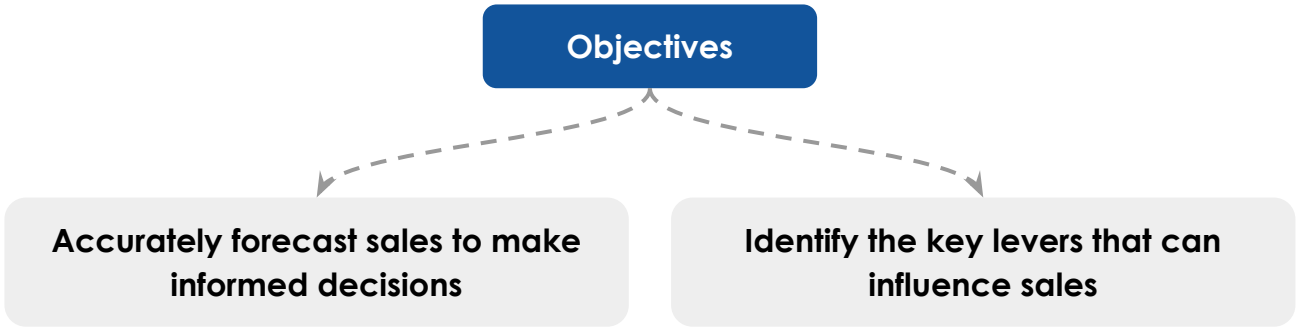


Problem Space

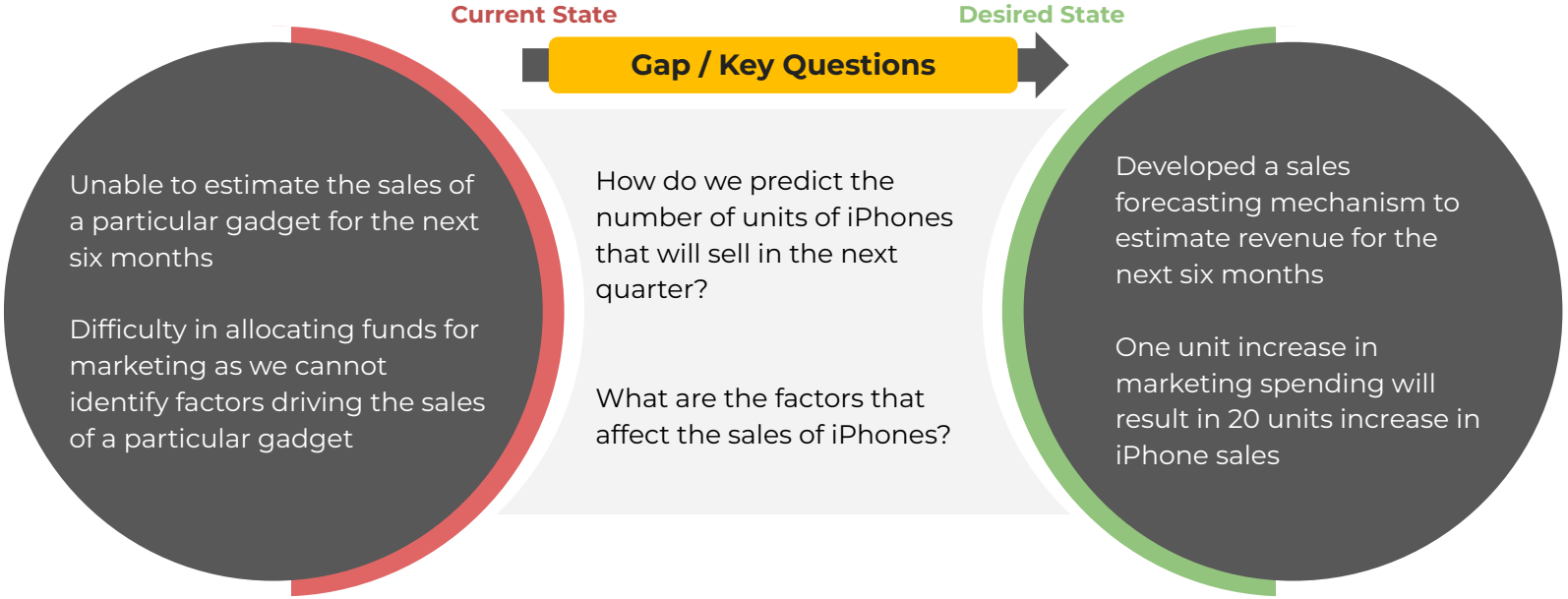


Problem Statement

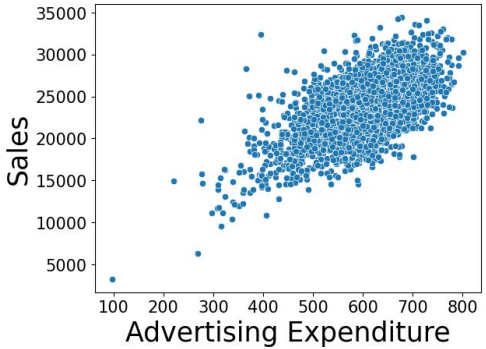
- Consider an online retailer of mobiles and tablets
- Crucial to stay ahead of market trends and consumer preferences to maximize sales
- Need to effectively manage inventory and marketing efforts to attract and retain customers

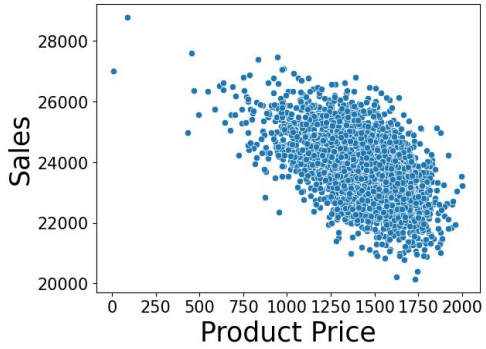


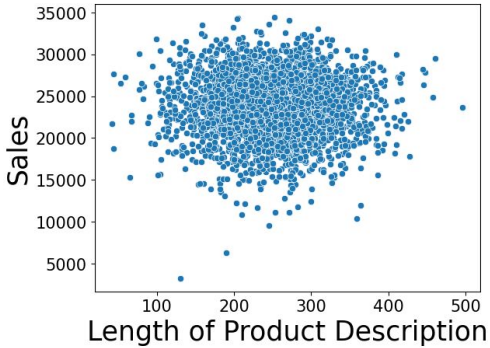
Sales Forecasting and Optimization



Visualizing Relationships







What happens to Sales as Advertising Expenditure increases?

Advertising Expenditure ↑ ✕ **Sales** ↑

Positive relationship

What happens to Sales as Product Price increases?

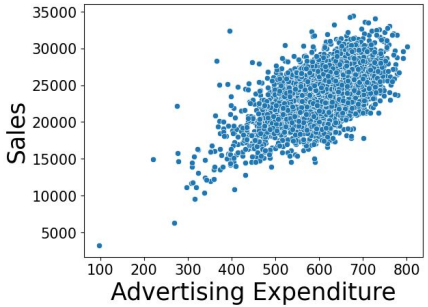
Product Price ↑ ✕ **Sales** ↓

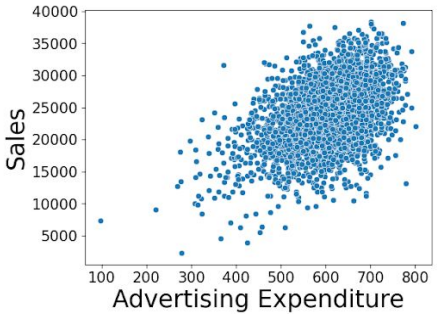
Negative relationship

No relationship

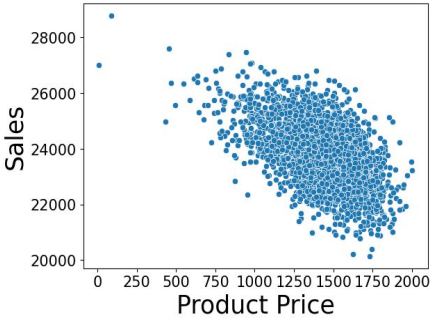
Visualizing Relationships

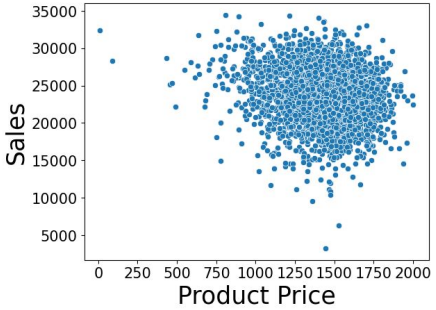
(I)





(II)





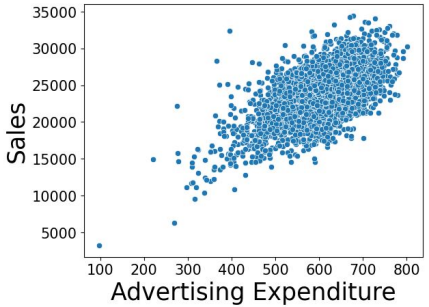
(I) In both the cases, we observe a **positive relationship** between sales and advertising expenditure

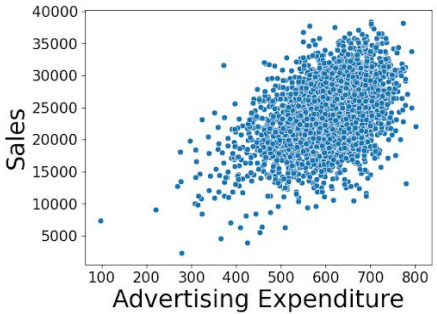
What is the **difference**?

(II) In both the cases, we observe a **negative relationship** between the sales and product price

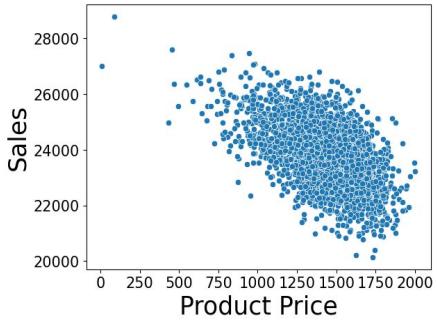
Visualizing Relationships

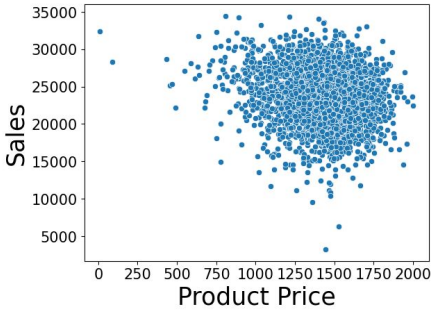
(I)





(II)





Strong

Weak

The cases on the left - in both (I) and (II) - exhibit a **stronger relationship (positive** or **negative)** than the ones on the right

Correlation

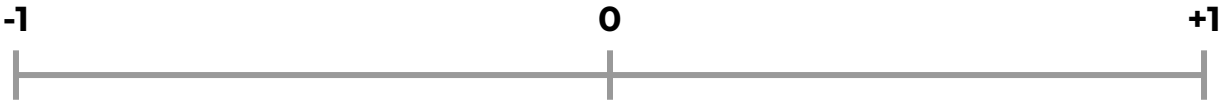
- We have seen how to **visually identify relationships** between a pair of variables from two aspects - **direction** and **strength**
- But we need a **quantitative measure** of the relationship

Correlation is a **statistical measure** that describes the **strength and direction** of a **relationship** between two variables.

- Indicates the **degree** to which two variables tend to **change together**
- Quantifies both the **direction** and **strength** of the relationship

Correlation

Correlation typically **ranges between -1 and 1.**



Perfect negative correlation	No correlation	Perfect positive correlation
One variable decreases as the other increases	Variables are independent of each other	Both variables increase together

Pearson's Correlation Coefficient

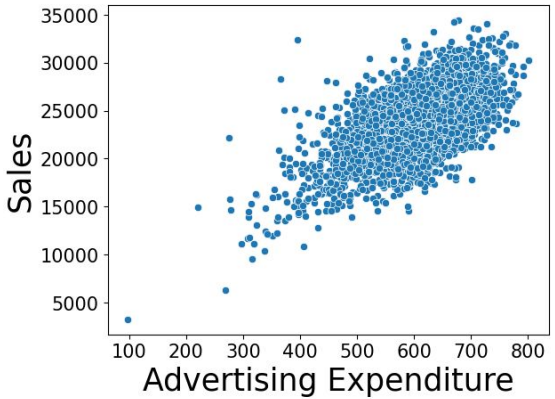
- One of the most commonly used measures of correlation.

A statistical measure that quantifies the **strength** and **direction** of the **linear relationship** between **two continuous variables.**

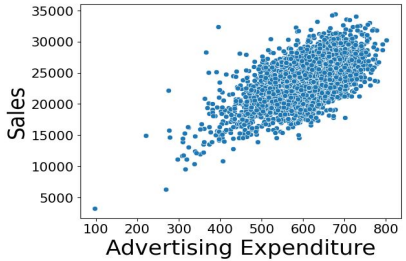
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

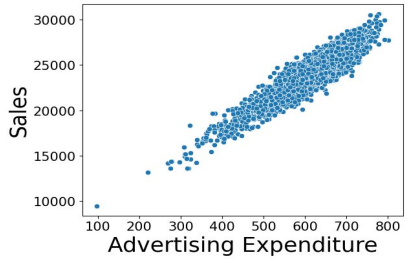
Correlation vs. Causation

- We observed advertising expenditure exhibits a strong positive correlation with sales.
- As advertising expenditure increased, sales increased.
- Does it mean advertising expenditure **causes** an increase in sales?
- **Not necessarily true!**
- There might be **other factors** at play.

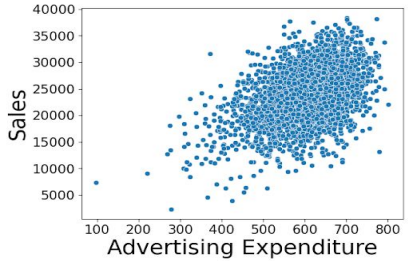


Correlation vs. Causation





Economic Zone 1



Economic Zone 2

- Let's split the data with respect to **another factor** - **economic zone.**
- Economic Zone 1 has a **booming economy** - sales will be higher here even if we don't spend as much on marketing.
- Economic Zone 2 has a **stagnant economy** - sales might have been higher due to data collected in a festive season.

Correlation vs. Causation

Correlation ≠ Causation

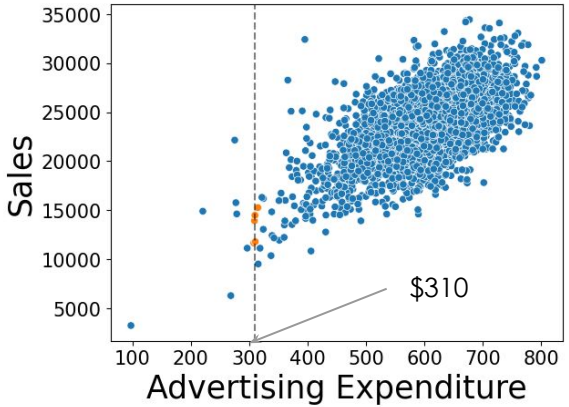
Variable 1 and **Variable 2** are **highly correlated**

≠

Variable 1 causes a change in **Variable 2**

The Need for Regression

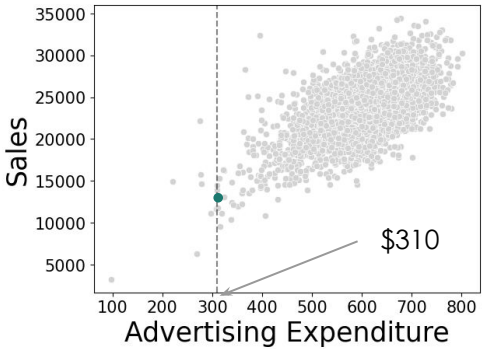
- We observed advertising expenditure exhibits a strong positive correlation with sales.
- Let's say we now decide to spend $310 for the marketing campaign of the latest iPhone.
- **How much sales should we expect?**
- **We don't know!**
- **Historically**, we've had **different sales** for **similar marketing spending.**



Correlation measures the strength and direction of the **relationship**, but **doesn't** provide a way to **predict** the output given an input.

The Need for Regression

- It is important for us to be able to determine the output (sales).
- It is also important to identify the lever(s) that drive the output (sales).
- Hence, the need for a mathematical model.



Input value
($310 marketing spend)

→

Mathematical model

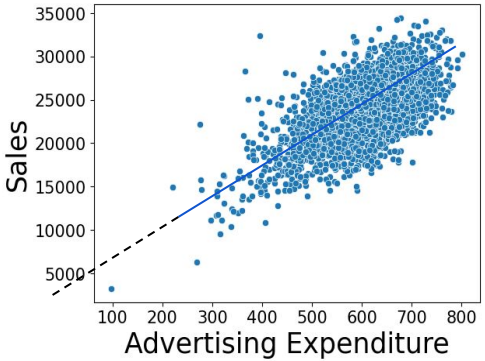
→

Output value
($13040.50 sales)

Simple Linear Regression

- The **simplest** mathematical model is **linear** - a **straight line.**

Linear Regression is a **statistical model** which **estimates** the **linear relationship** between a **response** and one or more **explanatory variables.**



- Simple Linear Regression - **one explanatory** and **one response variable.**

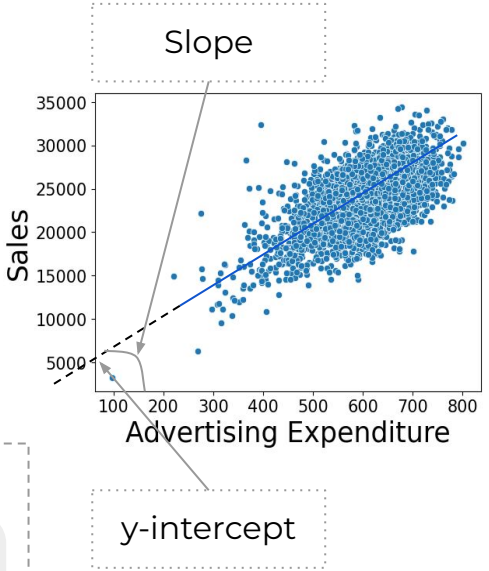
- Assumes that there is a linear relationship between the explanatory (independent) variable and the response (dependent) variable.

advertising expenditure → **sales**

Simple Linear Regression

- The equation of line is represented by:

sales ← $y = b_0 + b_1 x$ → **advertising expenditure**



y-intercept (b_0)

The average value of the response when the explanatory variable is zero

Slope (b_1)

The average change in the response for one-unit increase in the explanatory variable

Coefficients of Simple Linear Regression

Coefficient Interpretation

- Consider the following model for our context:

sales = 1.01 + 2.45 * advertising expenditure

- For a **unit increase** in advertising expenditure, the sales will increase by **2.45 units.**

This **interpretation** is **valid ONLY IF** the **assumptions** of linear regression hold **true.**

Coefficient Interpretation

- Consider the following model for our context:

sales = 1.01 + 2.45 * advertising expenditure

- If we have zero marketing expenditure:

sales = 1.01 + 2.45 * 0 = 1.01

- Makes **business sense** — we can have **organic sales.**

What if the business context changes?

Coefficient Interpretation

- Consider the case of predicting the price of a house using the following model:

house price = 291.07 + 105.45 * square footage

- For a **unit increase** in square footage, the price of the house increases by **105.45 units.**

This **interpretation** is **valid ONLY IF** the **assumptions** of linear regression hold **true.**

Coefficient Interpretation

- Consider the case of predicting the price of a house using the following model:

house price = 291.07 + 105.45 * square footage

- In the case of zero square footage:

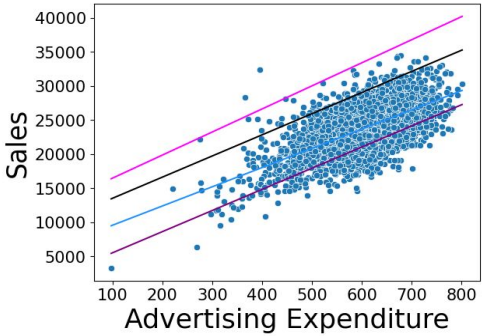
house price = 291.07 + 105.45 * 0 = 291.07

- **Doesn't make business sense!**

y-intercept doesn't always make business sense.

Best-Fit Line

- We observed one line that described the relationship between sales and advertising expenditure.
- But we can draw multiple lines!

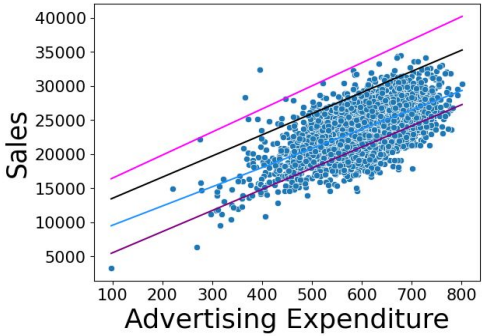


Which line do we choose?

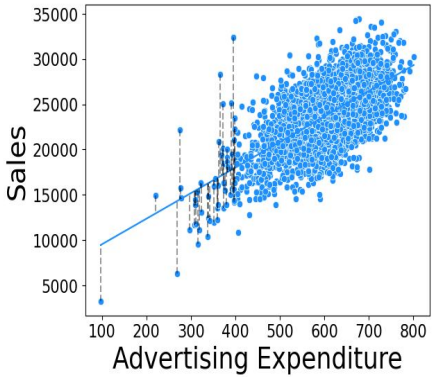
Best-Fit Line

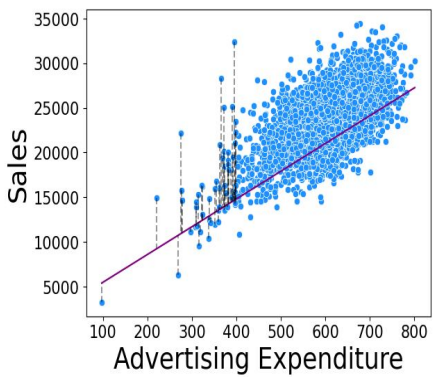
- We first need to understand the **difference** between these **lines.**
- We have actual data points (actual sales) and predicted data points (model's predicted sales).

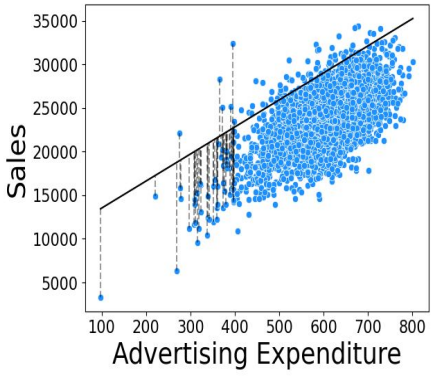
Prediction Error = Actual Value - Predicted Value

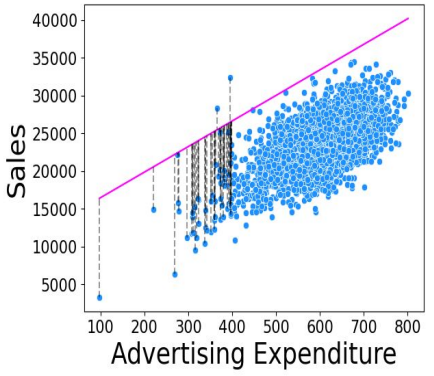


Best-Fit Line



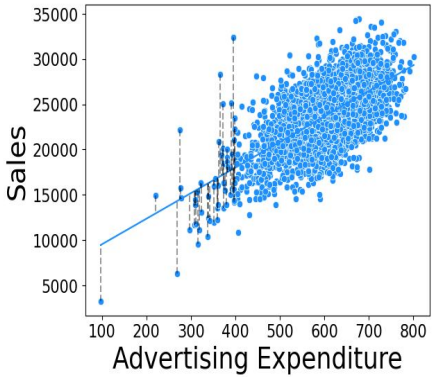


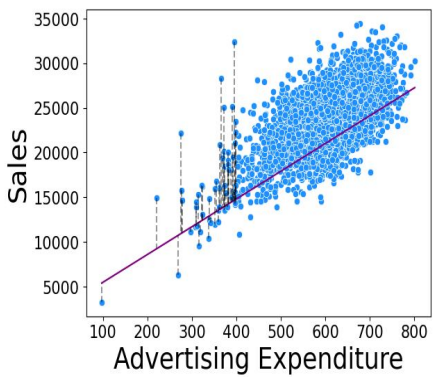


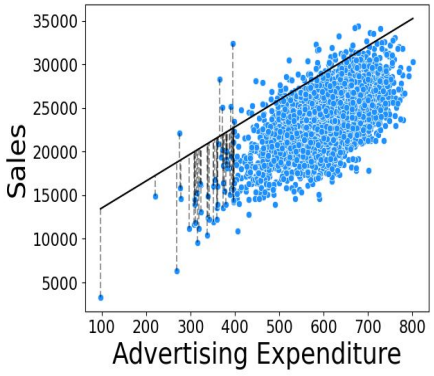


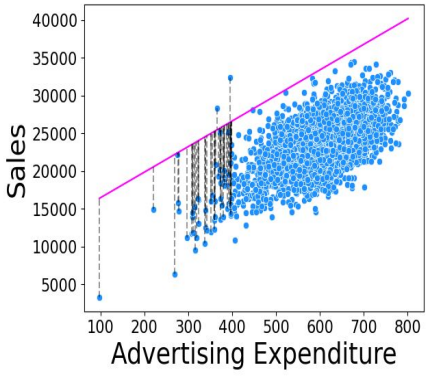
- There are multiple data points to consider.
- Take the aggregate of the errors across the data points.

Best-Fit Line









- The **line** with the **least aggregate error** across all data points is the one **we want.**

This is called the **best-fit line.**

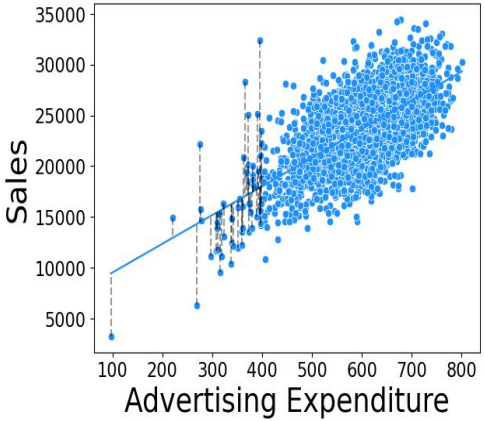
Best-Fit Line Computation

- How to find the error?

$$Error = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

Actual Value — y_i

Predicted Value — $\hat{y}_i$



- **Difference** between actual and predicted values can be **positive** or **negative**
- Direct addition will give a false picture of low overall error

Best-Fit Line Computation

- Take **absolute values**

$$Error = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$Error = \frac{1}{n}\sum_{i=1}^{n} |y_i - (b_0 + b_1 x_i)|$$

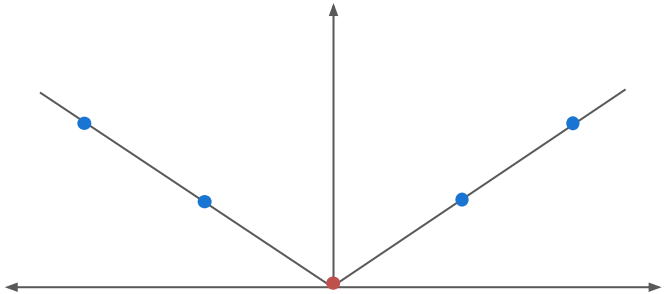
How to minimize the error?

Best-Fit Line Computation

- Need to find the **values** of the **coefficients** (b_0 and b_1) that yield the **minimum error**
- Use **differentiation**

Differentiate the **error** with respect to the **coefficients** (b_0 and b_1)

- Differentiating absolute values is **mathematically inconvenient**



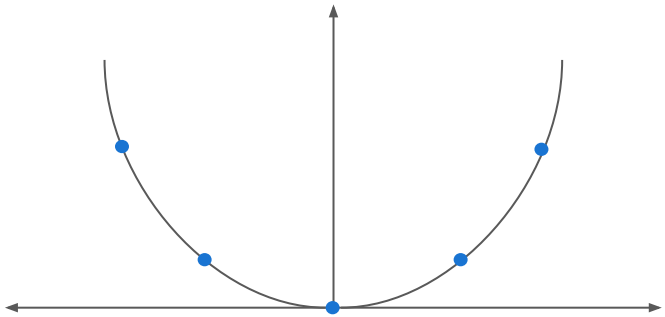
Differentiable

Not differentiable

Best-Fit Line Computation

- Use **squared values** instead

$$Error = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



- **Differentiable**

- **Accommodates** both **positive** and **negative** errors

- **Mathematically convenient** - differentiable

Best-Fit Line Computation

- Use **squared values** instead

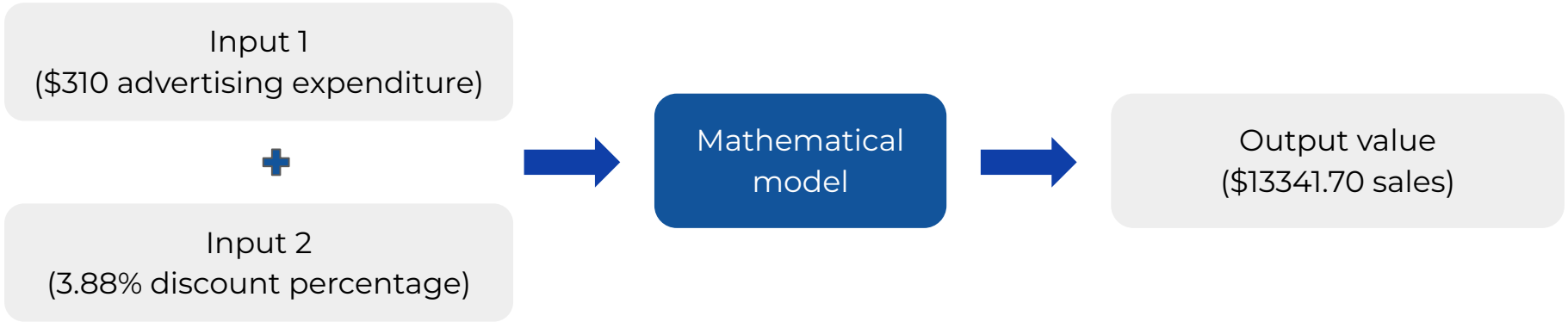
$$Error = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$Error = \frac{1}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$$

This is known as the **Method of Least Squares**

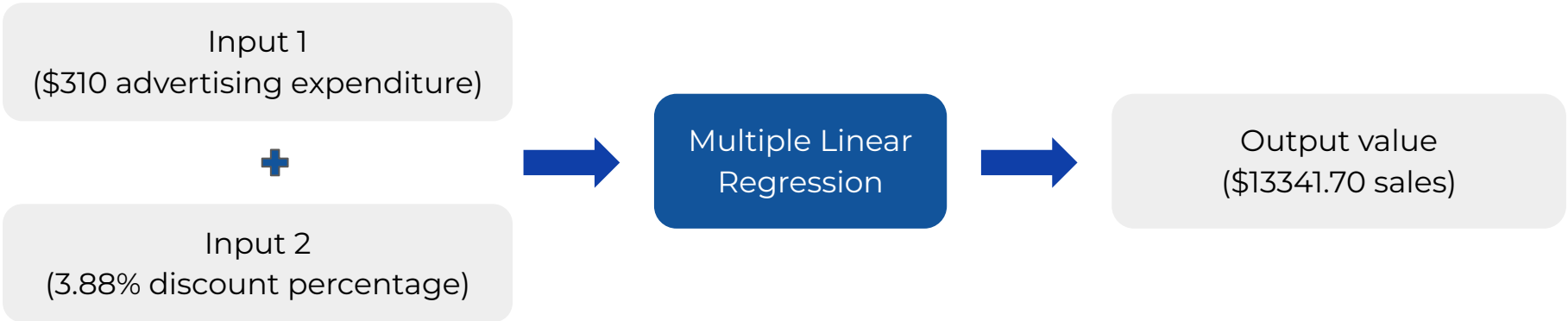
Multiple Linear Regression

- We have checked the relationship between sales and advertising expenditure
- What if there is another variable which can be used to predict the sales?



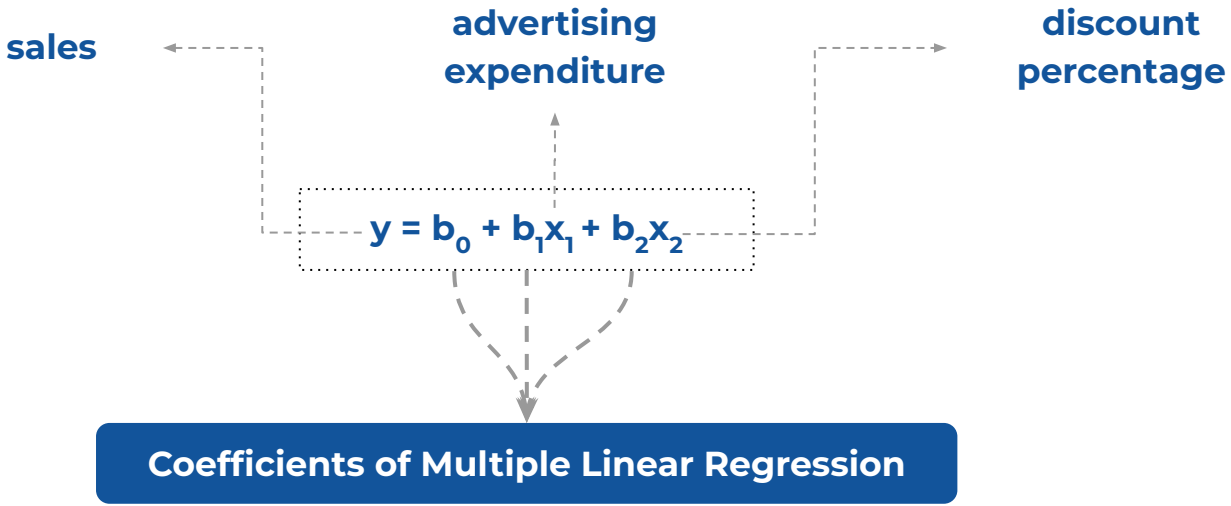
Multiple Linear Regression

- Multiple Linear Regression - **two or more explanatory** and **one response variable**
- Extension of Simple Linear Regression

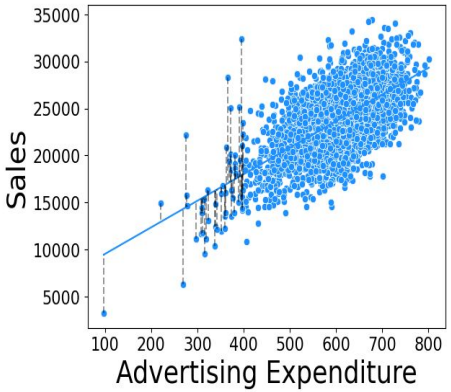


Multiple Linear Regression

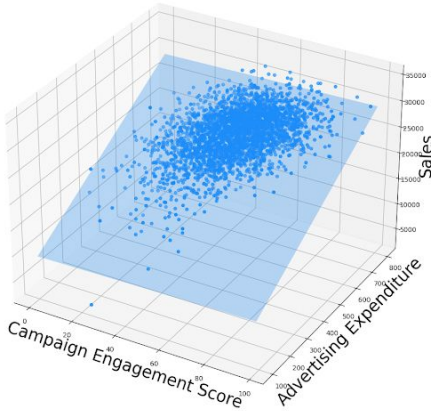
- Multiple Linear Regression equation - two explanatory variables



Multiple Linear Regression



$y = b_0 + b_1 x_1$



$y = b_0 + b_1 x_1 + b_2 x_2$

- For **one explanatory variable**, the **equation** was that of a **line**
- For **two explanatory variables**, the **equation** will be that of a **plane**

Coefficient Interpretation

- Consider the following model for our context

sales = 1.01 + 2.45 * advertising expenditure + 7.88 * discount percentage

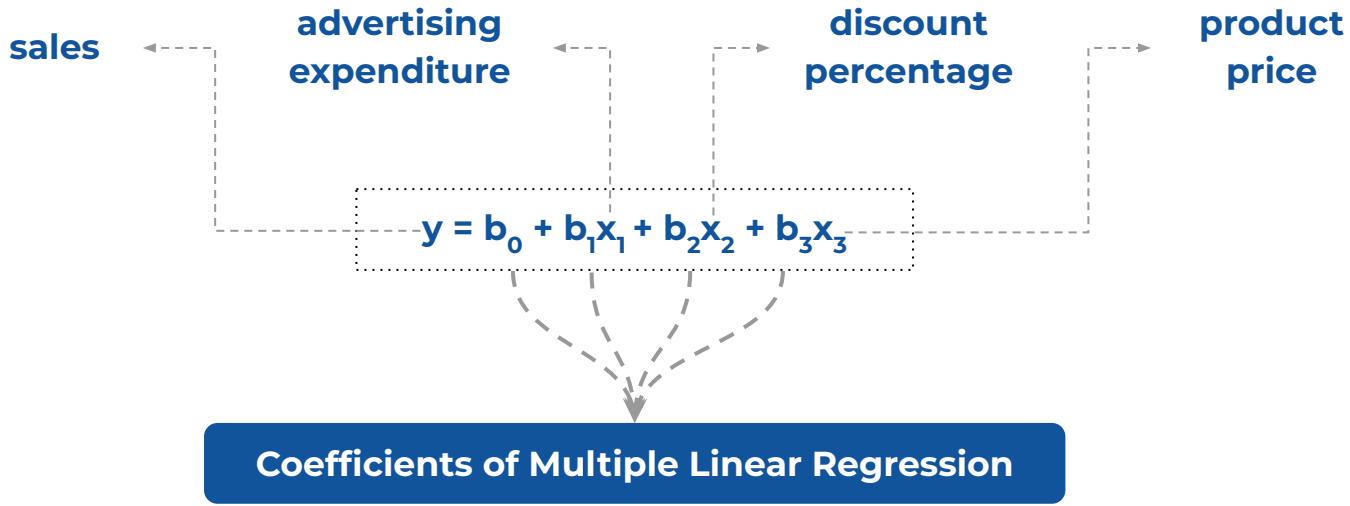
- For a **unit increase** in advertising expenditure, the sales will increase by **2.45 units**, provided all other variables are held constant

- For a **unit increase** in discount percentage, the sales will increase by **7.88 units**, provided all other variables are held constant

These **interpretations** are **valid ONLY IF** the **assumptions** of linear regression hold **true**

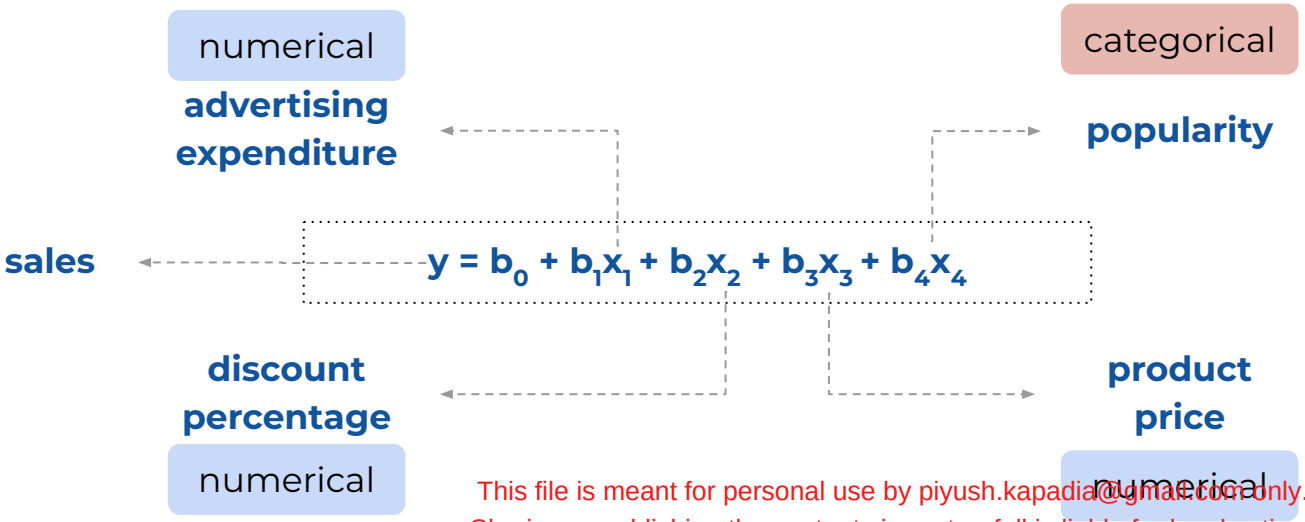
Multiple Linear Regression

- Multiple Linear Regression equation - more than two explanatory variables



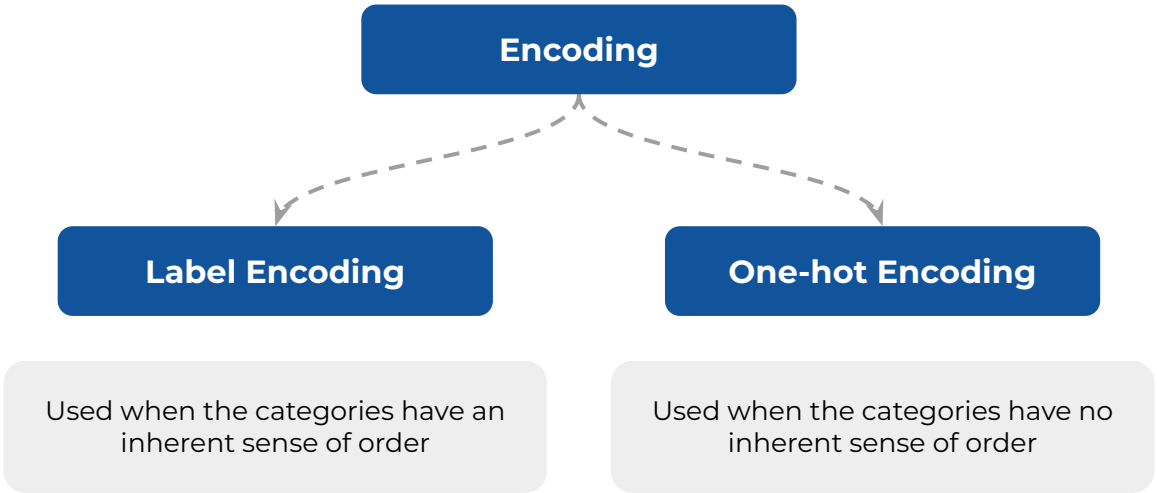
Categorical Variables in Regression

- So far we've worked with **numerical variables**
- But real-world data often contains **categorical variables**
- Consider the following case



Categorical Variables in Regression

- Categorical variables are not numbers - even if they might be represented by numbers
- Can’t be utilized directly in a linear regression model
- Need to be converted into a numerical format



Label Encoding

- Assigns a unique integer to each category
- Order of the integers represents the order of the categories

Popularity
Very Low
Low
Moderate
High
Very High

Label Encoding

Popularity
1
2
3
4
5

One-hot Encoding

- A new column is created for each category
- If the data point contains the category, corresponding column has value 1
- If the data point doesn't contain the category, corresponding column has value 0

Region
East
South
West
East
North

One-hot Encoding

Region_North	Region_East	Region_West	Region_South
0	1	0	0
0	0	0	1
0	0	1	0
0	1	0	0
1	0	0	0

Evaluation Metrics

- We have seen multiple models so far
- We don't know **'how well'** these models are **performing**
- Need to **evaluate** the models to gauge if they're performing 'well'
- **Model performance** is measured using **metrics**
- **Quantify** how well the model predictions align with the actual values

Evaluation Metrics

- An intuitive metric

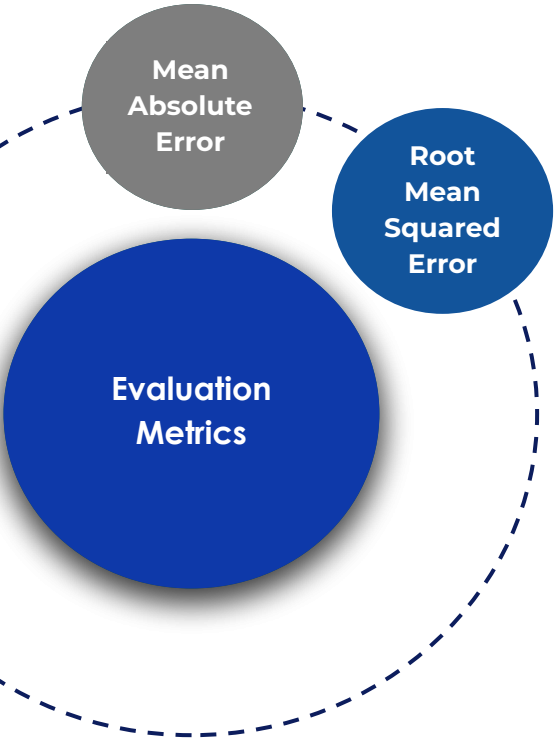
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Gives an idea of how much the model predictions deviate from the actual observations
- Relative to the range of the response

Evaluation Metrics

- Problem with considering the absolute value of errors is it doesn't penalize larger errors
- Needed to ensure that the model learns to do better when encountering edge cases

Evaluation Metrics

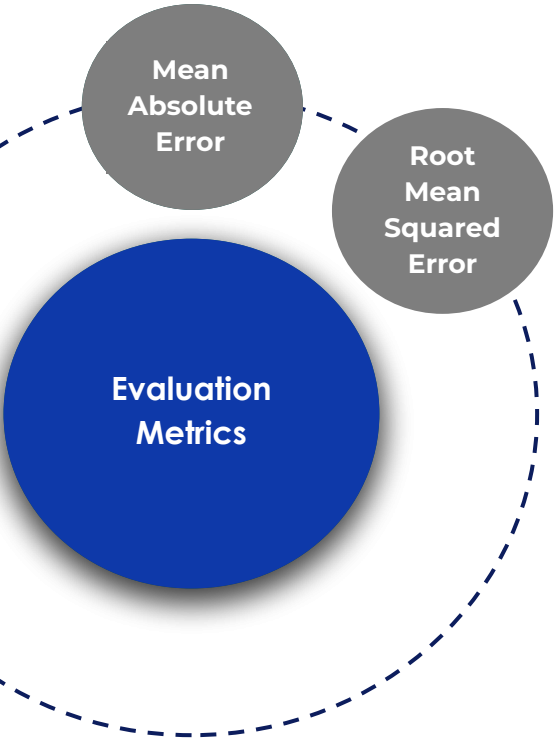


- Problem with considering the absolute value of errors is it doesn't penalize larger errors
- Needed to ensure that the model learns to do better when encountering edge cases

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

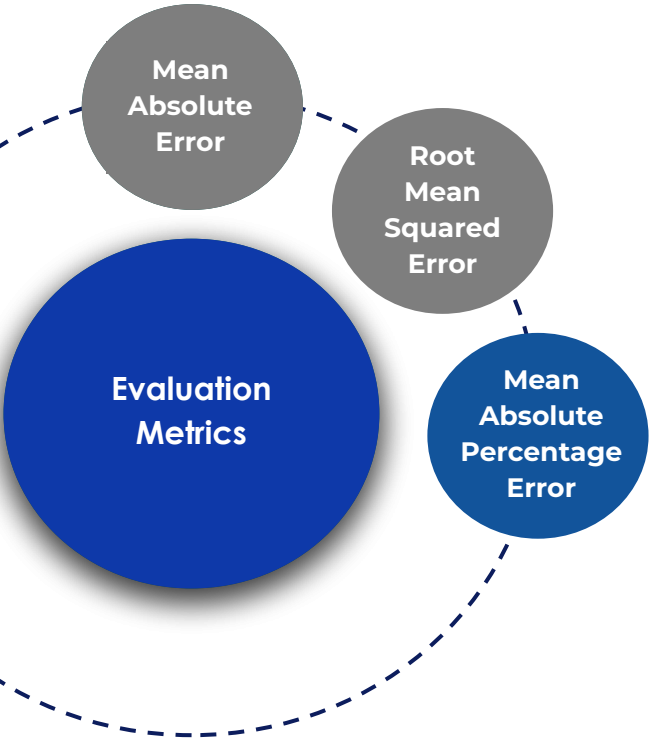
- Relative to the range of the response

Evaluation Metrics



- MAE and RMSE are relative to the scale of the response
- Cannot compare models across different data and scale of response value

Evaluation Metrics

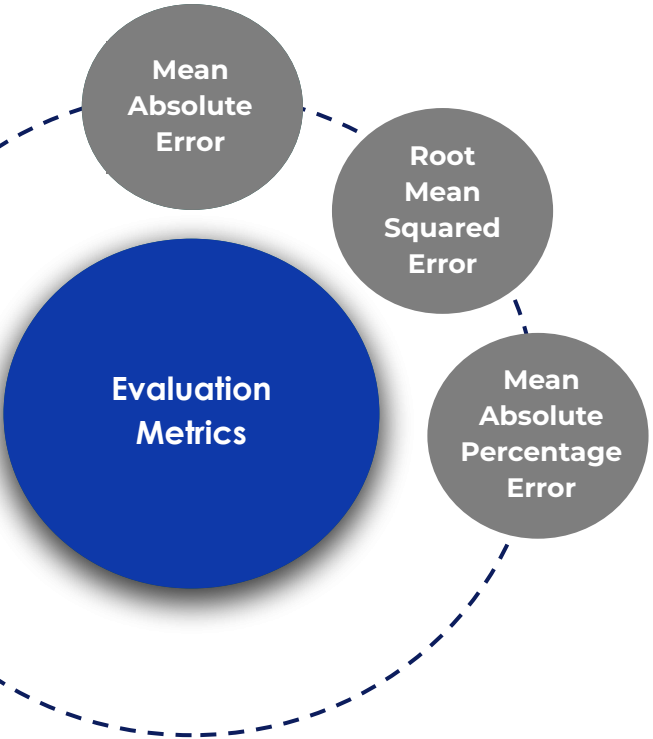


- MAE and RMSE are relative to the scale of the response
- Cannot compare models across different data and scale of response value

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100$$

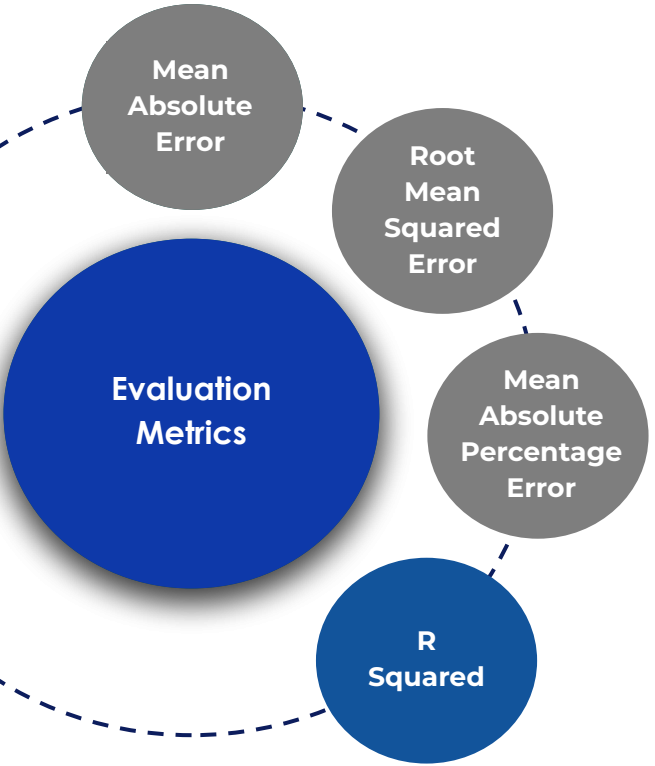
- Indifferent of the range of the response
- Needs to be adjusted when the actual value of the response is zero

Evaluation Metrics



- Previous metrics do not clearly quantify how well the model explains the variability in the data

Evaluation Metrics

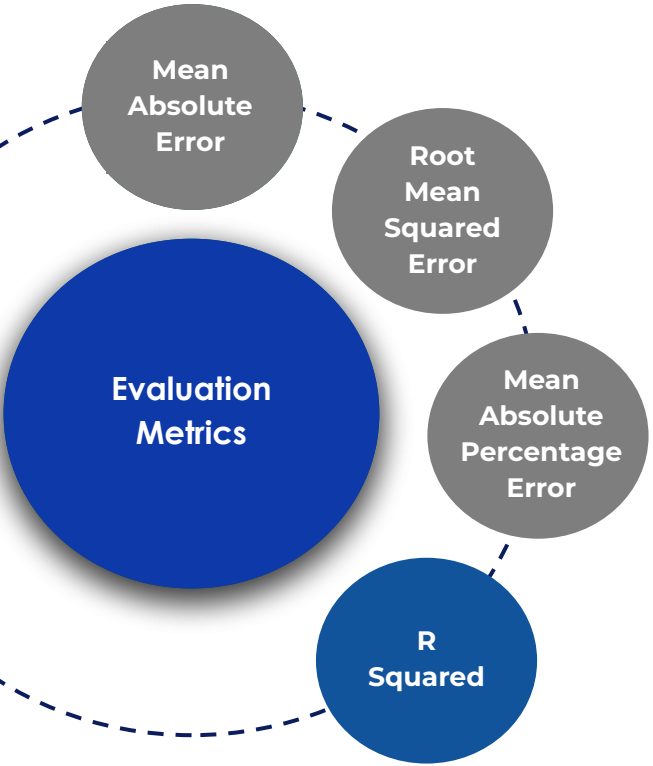


- Previous metrics do not clearly quantify how well the model explains the variability in the data

$$R^2 = 1 - \frac{\sum_{1=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{1=1}^{n}(y_i - \bar{y})^2}$$

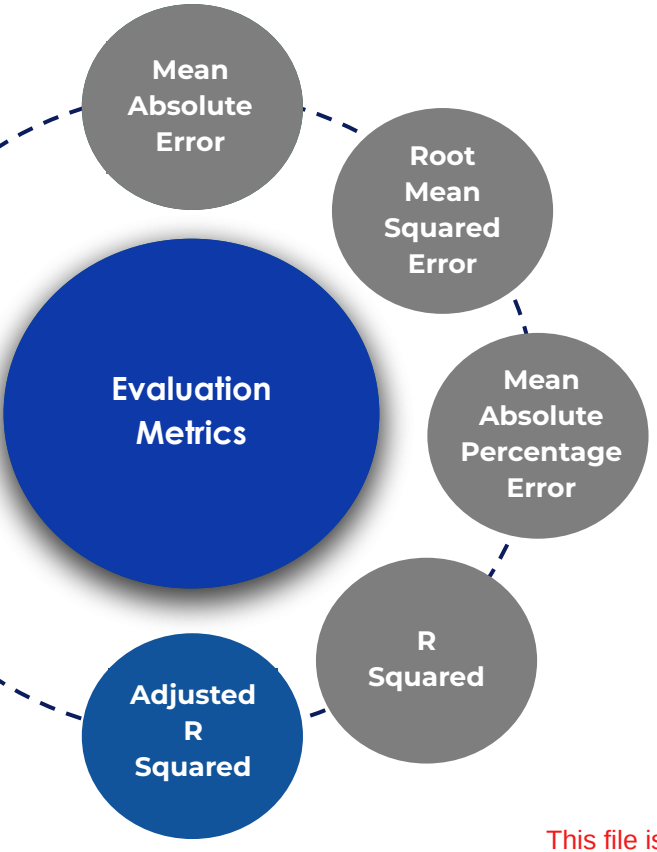
- Generally ranges between 0 and 1

Evaluation Metrics



- Tends to increase when adding more explanatory variables
- Does not account for the value addition from the added explanatory variables

Evaluation Metrics

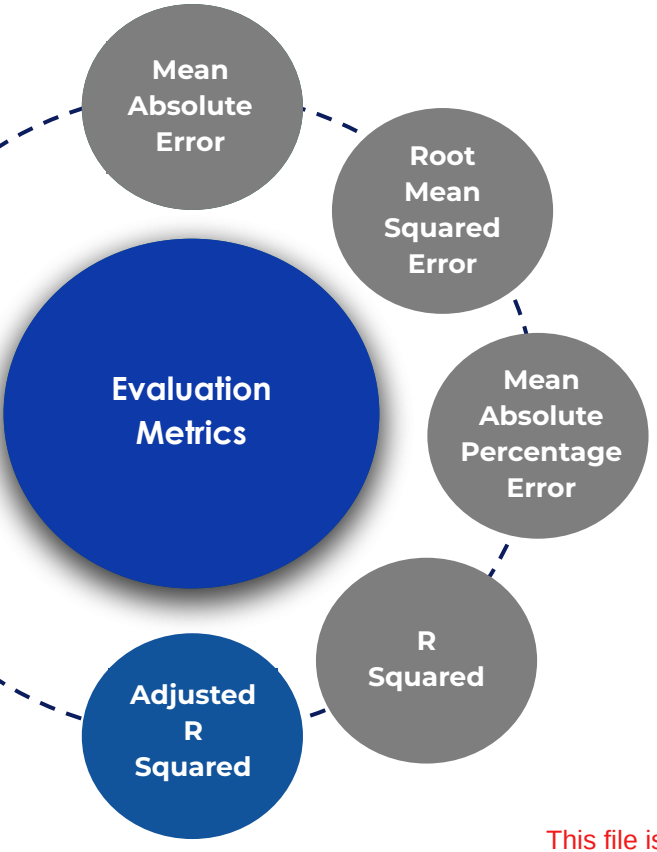


- Tends to increase when adding more explanatory variables
- Does not account for the value addition from the added explanatory variables

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2) * (n - 1)}{n - k - 1}$$

- Accounts for the number of explanatory variables in the model

Evaluation Metrics



- Gives a sense of which variables actually help in prediction and which ones do not
- Provides a balance between model fit and complexity (number of explanatory variables)

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2) * (n - 1)}{n - k - 1}$$

Summary

Here’s a quick recap of what we've learned:

- **Business Problem and Solution Space:** Identifies the specific problem that linear regression aims to solve and defines the scope of its application in business contexts.
- **Correlation and Linear Relationships:** Explores how correlation measures the strength and direction of linear relationships between variables, laying the foundation for understanding linear regression.
- **Simple Linear Regression:** Introduces the basic concept of simple linear regression, which models the relationship between a dependent variable and one independent variable using a straight line.

Summary

- **Multiple Linear Regression:** Expands on the concept of simple linear regression by incorporating multiple independent variables to predict a dependent variable, accommodating more complex relationships.
- **Categorical Variables in Regression:** Discusses strategies for encoding categorical variables in regression models to include qualitative data effectively in predictive analysis.
- **Evaluation Metrics for Regression:** Covers key metrics such as Mean Squared Error (MSE), R-squared, and others used to assess the accuracy and performance of regression models in predicting outcomes.

Learning Outcomes

You should now be able to:

- Explain how correlation measures the strength and direction of linear relationships, and apply this understanding to build simple linear regression models effectively.
- Gain proficiency in constructing and interpreting simple linear regression models to analyze and predict relationships between two variables.
- Develop multiple linear regression models to enable the prediction of business outcomes using multiple input variables.

Learning Outcomes

- Evaluate linear regression models using key metrics and implement strategies to enhance model performance and accuracy.
- Identify and apply linear regression techniques to solve various real-world business problems, leveraging its predictive capabilities across different domains.

Great Learning

Happy Learning !